

Information Fusion and Data Science

Series Editor: Henry Leung

Éloi Bossé
Galina L. Rogova *Editors*

Information Quality in Information Fusion and Decision Making



Information Fusion and Data Science

Series editor

Henry Leung, University of Calgary, Calgary, AB, Canada

This book series provides a forum to systematically summarize recent developments, discoveries and progress on multi-sensor, multi-source/multi-level data and information fusion along with its connection to data-enabled science. Emphasis is also placed on fundamental theories, algorithms and real-world applications of massive data as well as information processing, analysis, fusion and knowledge generation. The aim of this book series is to provide the most up-to-date research results and tutorial materials on current topics in this growing field as well as to stimulate further research interest by transmitting the knowledge to the next generation of scientists and engineers in the corresponding fields. The target audiences are graduate students, academic scientists as well as researchers in industry and government, related to computational sciences and engineering, complex systems and artificial intelligence. Formats suitable for the series are contributed volumes, monographs and lecture notes.

More information about this series at <http://www.springer.com/series/15462>

Éloi Bossé • Galina L. Rogova
Editors

Information Quality in Information Fusion and Decision Making



Editors

Éloi Bossé
IMT-Atlantique
Brest, France

McMaster University
Hamilton, Canada

Galina L. Rogova
The State University of New York
at Buffalo
Buffalo, NY, USA

ISSN 2510-1528

ISSN 2510-1536 (electronic)

Information Fusion and Data Science

ISBN 978-3-030-03642-3

ISBN 978-3-030-03643-0 (eBook)

<https://doi.org/10.1007/978-3-030-03643-0>

Library of Congress Control Number: 2019932814

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The subject of information quality has been considered by researchers and practitioners in many diverse fields such as organizational processes, management, product development, marketing, personal computing, health care, and publishing among others. At the same time, the problem of information quality in the fusion-based human-machine systems for decision-making has just recently begun to attract attention.

Information fusion is dealing with gathering, processing, and combining a large amount of diverse information from physical sensors (infrared imagers, radars, chemical, etc.), human intelligence reports, and information obtained from open sources (traditional such as newspapers, radio, TV, as well as social media such as Twitter, Facebook, and Instagram). That data and information obtained from observations and reports as well as information produced by both human and automatic processes are of variable quality and may be unreliable, of low fidelity, insufficient resolution, contradictory, and/or redundant. Furthermore, there is often no guarantee that evidence obtained from the sources is based on direct, independent observations. Sources may provide unverified reports obtained from other sources (e.g., replicating information in social networks), resulting in correlations and bias. Some sources may have malicious intent and propagate false information through social networks or even coordinate to provide the same false information in order to reinforce their opinion in the system.

The success of the information fusion processing depends on how well knowledge produced by the processing chain represents reality, which in turn depends on how adequate data are, how good and adequate are the models used, and how accurate, appropriate, or applicable prior and contextual knowledge is. The objective of this book is to provide an understanding of the specific problem of information quality in the fusion-based processing and address the challenges of representing and incorporating information quality into the whole processing chain from data to information to actionable knowledge to decisions and actions to support decision-makers in complex dynamic situations.

The book will emphasize a contemporary view on the role of information quality in fusion and decision-making and provide a formal foundation and implementation

strategies required for dealing with insufficient information quality in building fusion-based systems for decision-making. It offers contributions of experts discussing the fundamental issues, challenges, and the state of the art of computational approaches to incorporating information quality in information fusion processes to various decision support applications for real-life scenarios such as remote sensing, medicine, automated driving, environmental protection, crime analysis, intelligence, and defense and security. The book comprises two parts. Part one contains chapters devoted to models, concepts, and dimensions of information quality in information fusion. Part two includes chapters that describe the incorporation and evaluation of information quality in the fusion-based systems designed for various applications.

Hamilton, ON, Canada
Buffalo, NY, USA

Éloi Bossé
Galina L. Rogova

Addendum

This Addendum concerns a recent open access survey paper that the Editors of this book highly recommend to read as a supplement to Part I:

Dubois, D., Liu, W., Ma, J., & Prade, H. (2016). The basic principles of uncertain information fusion. An organised review of merging rules in different representation frameworks. *Information Fusion*, 32, 12–39. <https://doi.org/10.1016/j.inffus.2016.02.006>

The authors of the paper present a state-of-the art survey of information fusion rules and their properties across various theories of uncertainty. Reading of the paper will enrich the reader's background to fully benefit from applications in Part II.

The Editors
Éloi Bossé
Galina L. Rogova

Acknowledgement

The co-editors wish to express their gratitude to the authors whose work produced this book. We would also like to thank the Springer staff, Christopher Coughlin, HoYing Fan, Jeffrey Taub and Chandhini Kuppusamy for their support and patience throughout this process.

Contents

Part I Information Quality: Concepts, Models and Dimensions

1	Information Quality in Fusion-Driven Human-Machine Environments	3
	Galina L. Rogova	
2	Quality of Information Sources in Information Fusion	31
	Frédéric Pichon, Didier Dubois, and Thierry Denœux	
3	Using Quality Measures in the Intelligent Fusion of Probabilistic Information	51
	Ronald R. Yager and Frederick E. Petry	
4	Conflict Management in Information Fusion with Belief Functions ..	79
	Arnaud Martin	
5	Basic Properties for Total Uncertainty Measures in the Theory of Evidence	99
	Joaquín Abellán, Carlos J. Mantas, and Éloi Bossé	
6	Uncertainty Characterization and Fusion of Information from Unreliable Sources	109
	Lance Kaplan and Murat Sensoy	
7	Assessing the Usefulness of Information in the Context of Coalition Operations	135
	Claire Saurel, Olivier Poitou, and Laurence Cholvy	
8	Fact, Conjecture, Hearsay and Lies: Issues of Uncertainty in Natural Language Communications	155
	Kellyn Rein	
9	Fake or Fact? Theoretical and Practical Aspects of Fake News	181
	George Bara, Gerhard Backfried, and Dorothea Thomas-Aniola	

10	Information Quality and Social Networks	207
	Pontus Svenson	
11	Quality, Context, and Information Fusion	219
	Galina L. Rogova and Lauro Snidaro	
12	Analyzing Uncertain Tabular Data	243
	Oliver Kennedy and Boris Glavic	
13	Evaluation of Information in the Context of Decision-Making	279
	Mark Burgin	
14	Evaluating and Improving Data Fusion Accuracy	295
	John R. Talburt, Daniel Pullen, and Melody Penning	
Part II Aspects of Information Quality in Various Domains of Application		
15	Decision-Aid Methods Based on Belief Function Theory with Application to Torrent Protection	329
	Simon Carladous, Jean-Marc Tacnet, Jean Dezert, and Mireille Batton-Hubert	
16	An Epistemological Model for a Data Analysis Process in Support of Verification and Validation	359
	Alicia Ruvinsky, LaKenya Walker, Warith Abdullah, Maria Seale, William G. Bond, Leslie Leonard, Janet Wedgwood, Michael Krein, and Timothy Siedlecki	
17	Data and Information Quality in Remote Sensing	401
	John Puentes, Laurent Lecornu, and Basel Solaiman	
18	Reliability-Aware and Robust Multi-sensor Fusion Toward Ego-Lane Estimation Using Artificial Neural Networks	423
	Tran Tuan Nguyen, Jan-Ole Perschewski, Fabian Engel, Jonas Kruesemann, Jonas Sitzmann, Jens Spehr, Sebastian Zug, and Rudolf Kruse	
19	Analytics and Quality in Medical Encoding Systems	455
	John Puentes, Laurent Lecornu, Clara Le Guillou, and Jean-Michel Cauvin	
20	Information Quality: The Nexus of Actionable Intelligence	471
	Marco Antonio Solano	
21	Ranking Algorithms: Application for Patent Citation Network	519
	Hayley Beltz, Timothy Rutledge, Raoul R. Wadhwa, Péter Bruck, Jan Tobochnik, Anikó Fülöp, György Fenyvesi, and Péter Érdi	

22 Conflict Measures and Importance Weighting for Information Fusion Applied to Industry 4.0	539
Uwe Mönks, Volker Lohweg, and Helene Dörksen	
23 Quantify: An Information Fusion Model Based on Syntactic and Semantic Analysis and Quality Assessments to Enhance Situation Awareness	563
Leonardo Castro Botega, Allan Cesar Moreira de Oliveira, Valdir Amancio Pereira Junior, Jordan Ferreira Saran, Lucas Zanco Ladeira, Gustavo Marttos Cáceres Pereira, and Seiji Isotani	
24 Adaptive Fusion.....	587
Vincent Nimier and Kaouthar Benameur	
Index.....	607

About the Editors

Éloi Bossé received the B.A.Sc., M.Sc., and Ph.D. degrees from Université Laval, Québec City, QC, Canada, in 1979, 1981, and 1990, respectively, all in electrical engineering. In 1981, he joined the Communications Research Centre, Ottawa, ON, Canada, doing research on signal processing and high-resolution spectral analysis. In 1988, he was transferred to the Defence Research Establishment, Ottawa, to research on radar target tracking in multipath. In 1992, he moved to Defence Research and Development Canada Valcartier (DRDC Valcartier), Courcelette, QC, Canada, to lead a group of four to five Defence Scientists on information fusion and decision support. From 1993 to now, he held Adjunct Professor positions with several universities, such as Université Laval (Québec, CA), the University of Calgary (Alberta, CA), and McMaster University (Ontario, CA). He headed the Command and Control Decision Support Systems Section at DRDC Valcartier from 1998 to his retirement in 2011. Dr. Bossé represented Canada (as a DRDC Member) in numerous international research fora under the various cooperation research programs (NATO, TTCP, and bi- and trilaterals) in his area of expertise. Since 2011, he has been conducting some research activities under NATO Science for Peace and Security Programme, as a Researcher with the Mathematics and Industrial Engineering Department, Polytechnic of Montreal, Montreal, QC, Canada, and with the Department of Computer and Electrical Engineering Department at McMaster University and finally, as an Associate Researcher with IMT Atlantique, Plouzané, France (since 2010). In 2015, he founded Expertises Parafuse Inc., at Québec City, a consultant firm on Analytics and Information Fusion Technologies. He has published over 200 papers in journals, book chapters, conference proceedings, and technical reports. He has coauthored and coedited four to five books on information fusion.

Galina L. Rogova received her M.Sc. and Ph.D. in Moscow, Russia. She is a Research Professor at the State University of New York at Buffalo. She is recognized internationally as an expert in information fusion, machine learning, decision-making under uncertainty, and information quality, and lectured extensively on these topics. Dr. Rogova has worked on a wide range of defense and non-defense

applications such as situation and threat assessment, understanding of volcanic eruption patterns, computer-aided diagnosis, and intelligent transportation system, among others. Her research was funded by multiple government agencies as well as commercial companies. She published numerous papers and coedited seven books on information fusion and decision-making.

Part I

**Information Quality: Concepts, Models
and Dimensions**

Chapter 1

Information Quality in Fusion-Driven Human-Machine Environments



Galina L. Rogova

Abstract Effective decision making in complex dynamic situations calls for designing a fusion-based human-machine information system requiring gathering and fusing a large amount of heterogeneous multimedia and multispectral information of variable quality coming from geographically distributed sources. Successful collection and processing of such information strongly depend on the success of being aware of, and compensating for, insufficient information quality at each step of information exchange. Designing methods of representing and incorporating information quality into fusion processing is a relatively new and rather difficult problem. The chapter discusses major challenges and suggests some approaches to address this problem.

Keywords Information fusion · Quality ontology · Meta-data · Subjective quality · Quality control · Higher level quality

1.1 Introduction

Decision making in complex dynamic environments involves gathering and fusing a large amount of heterogeneous multimedia and multispectral information of variable quality and data rates coming from geographically distributed sources to gain knowledge of the entire domain. Data and information to be processed and made sense of are not limited to data obtained from traditional sources such as physical sensors (infrared imagers, radars, chemical, etc.), human intelligence reports, operational information, and traditional open sources (such as newspapers, radio, TV). Nowadays, when each person can be a sensor, a huge amount of information can be also obtained from opportunistic human sensors and social media such as Twitter, Facebook, and Instagram. Such complex environments call

G. L. Rogova (✉)
The State University of New York at Buffalo, Buffalo, NY, USA
e-mail: rogova@buffalo.edu

for an integrated fusion-based human-machine system, in which some processes are best executed automatically while for others judgment and guidance of human experts and end users are critical.

The problem of building such integrated systems is complicated by the fact that data and information obtained from observations and reports as well as information produced by both human and automatic processes are of variable quality. They may be uncertain, unreliable, of low fidelity, insufficient resolution, contradictory, and/or redundant or even fake. The success of decision making in complex fusion-driven human-machine environments depends on the success of being aware of, and compensating for, insufficient information quality at each step of information exchange. Thus quality considerations play an important role at each time when raw data (sensor readings, open source, database search results, and intelligence reports) enter the system as well as when information is transferred between automatic processes, between humans, and between automatic processes and humans.

There are multiple reasons for information deficiency in such environment. Physical sensors can be unreliable, broken, or improperly used, fusion models can be imperfect, opportunistic human sensors and human decision makers can have variable expertise, their mental model imperfect and their decisions usually affected by cognitive biases. Another source of information deficiency may be imperfection of domain knowledge and statistical information about the environment, which is unavoidable in many domains such as natural disasters or terrorist attack since even the same type of man-made or natural disasters are rarely exactly the same. There is an inevitable delay in data transmission in the dynamic distributed environment, and therefore information entering the system can be obsolete. Information obtained from social networks and opportunistic human sensors has additional deficiency since people may not have precise knowledge of the subject they are talking about or have malicious intent and provide misinformation. Furthermore, there is often no guarantee that evidence acquired from the sources is based on direct, independent observations. Sources may provide unverified reports obtained from other sources (e.g., replicating information in social networks), resulting in correlations and bias. In a more malicious setting, some sources may coordinate to provide similar information in order to reinforce their opinion in the system. In addition, the lack of proper consideration of context or context of insufficient quality may result in using inadequate or erroneous domain knowledge or inappropriate models and parameters for quality assessment.

Information fusion is a complex concept and its definition varies from one application field to another. At the same time, the authors of the majority of the information fusion papers consider obtaining better quality of information by information combination as one of the main goals of designing fusion systems. For example, Wald in [1] defined it as "...a formal framework, in which are expressed the means and tools for the alliance of data originating from different sources. It aims at gaining information of greater quality." Benoit and Huget in [2] state that "the aim of an Information Fusion System is to compute results of higher quality (with respect to some criteria to be defined)" while in [3] data fusion is defined as "a combination of multiple sources to obtain improved information;

in this context, improved information means less expensive, higher quality, or more relevant information.”

Although the subject of information and data quality has been receiving significant attention in the recent years in many areas including communication, business processes, personal computing, health care, and databases (see, e.g., [4–6]), the problem of information quality in the fusion-based human-machine systems for decision making has just recently begun to attract attention. The main body of the literature on information fusion concerns with building an adequate uncertainty model without paying much attention to the problem of representing and incorporating other quality characteristics into fusion processes.

There are many research questions related to the information quality problem in designing fusion-based systems including:

- What is the information quality ontology?
- How to assess information quality of incoming heterogeneous data as well as the results of processes and information produced by users (where do the numbers come from?).
- How to evaluate usability of information?
- How to combine quality characteristics into a single quality measure?
- How to evaluate the quality of the quality assessment of information and processing results?
- How to compensate for various information deficiencies?
- How do quality and its characteristics depend on context and its quality?
- How does subjectivity, i.e. user biases, affect information quality?

The remainder of this paper is an effort to establish a conceptual framework, in which these questions may be addressed.

1.2 Information Fusion and Information Quality

There are various fusion models considered in the literature, which are reviewed in multiple publications (see, e.g., [7–12]). One of the earliest and still most influential one is the JDL¹ [7], which in its initial form contained operations aggregating entities at four increasing abstraction levels: source preprocessing (Level 0), object refinement (Level 1), situation refinement and impact or threat refinement (Levels 2–3). It also included process refinement (Level 4), a meta-process monitoring the underlying information fusion processes by implementing quality control. Further a Level 5 [8] addressing the issues associated with the human components of the fusion process was introduced. The JDL model is a functional rather than process model since no specific order of processing is prescribed. In general, the data to be

¹JDL: Joint Directors of Laboratories, a US DoD government committee overseeing US defense technology R&D; the Data Fusion Group of the JDL created the original JDL Data Fusion Model.

input at a given level includes tokens and parameters derived from a common bus connected to all other fusion levels, databases, and human-computer interfaces. The model has since been revised in [9, 10] and further evolved to add at least one higher level, that of mission management [11], and the model development continues to this date [12].

Various other fusion processing model-structures were proposed. In 1997, Dasarathy [13] introduced some ideas associated with the notion that there were three general levels of abstraction in fusion processing: the data level, the feature level, and the decision level. Accordingly, he published a model that characterizes the processing at and across such levels. This model, while providing a useful perspective, is not as comprehensive in scope as the JDL model.

Over 1999–2000, Bedworth and O’Brien published their “Omnibus Model” [14] which combines aspects of the Observe-Orient-Decide-Act or “OODA” decision/control loop [15] with the “Waterfall” software development process [16]. The “Omnibus Model” makes aspects of feedback more explicit and is claimed to enhance a data fusion process description by combining a system-goal point of view with a task-oriented point of view. In an effort to connect information fusion with behavioral psychology, Endsley [17] developed a descriptive model of human situation awareness which proved highly effective in guiding design of interfaces supporting time-critical decision making. In this model, the agent perceives elements of the current situation, processes and comprehends them to form an assessment, then projects the future status based on that assessment and possible current actions. An action is decided and taken upon that basis, and the results of that action create a new state of the environment to be perceived, closing the loop. Rather than prescribed normative behaviors, Endsley’s model describes human capabilities and cognitive constraints on performance that should guide the design of decision support system to maximize the agent’s performance and to minimize the risk of human error.

Information in these models is not processed sequentially, but there is a feedback across and within levels. Successful processing of this information requires being aware of and compensating for insufficient information quality at each step of information exchange. At the same time, the existing models very rarely implicitly address inter- and intra-processing information quality. The only information fusion model implicitly incorporating information quality in inter-level processioning is the so-called Revision of the JDL model II [10], which considers two quality characteristics: reliability and inconsistency.

It is important to notice that good quality of input information does not, of course, guarantee sufficient quality of the system output and therefore the insufficient quality of the information may build-up from one sub process to the other. Figure 1.1 shows major points of information exchange and quality considerations in the fusion-based machine-human system, JDL model Level 4 (sensor and process refinement), which performs quality control.

The result of quality assessment and control depends on context and its quality. *Context* is represented by a set of contextual variables and their characteristics, while *quality of context* is defined as the quality of the assessment of these

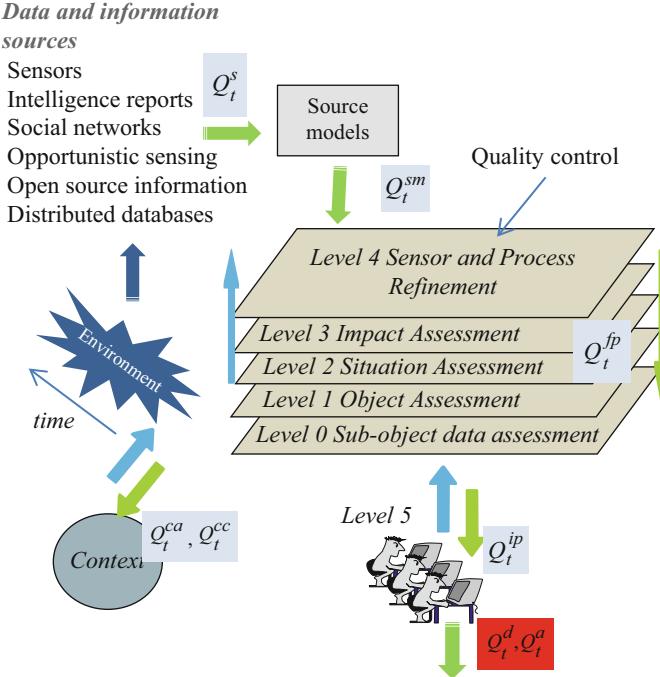


Fig. 1.1 Information flow in a fusion-based human-machine system. (Adapted from [18])

variables [19, 20]. Similar to the information utilized in the fusion processes, the information defining contextual variables can be obtained from available databases, observations, reports, traditional and social media, or as the result of various levels of information fusion. Of course, the quality of such information could be insufficient for a particular use: it might be uncertain, unreliable, irrelevant or conflicting.

Knowledge of the quality of this information and its effect on the quality of context characterization can improve contextual knowledge. At the same time, knowledge about a current context can improve the quality of observation and fusion results. Thus there are two interrelated problems concerning both information and context quality: imperfect information used in context estimation and discovery negatively impacts context quality while imperfect context characterization adversely affects the characterization of the quality of information used in fusion as well as the fusion results. That interrelationship represents one of the challenges of modeling and evaluating context quality and of using context in defining the quality of information used in fusion and the quality of fusion process results. A discussion of interrelations between quality and context is presented in [20].

In Fig. 1.1, Q_t^s - the quality of information sources, Q_t^{sm} -quality of source model output, Q_t^{fp} quality of information fusion processes at each level, Q_t^{ip} -quality of

presenting information to the user, Q_t^{ca} , Q_t^{cc} - quality of context estimation and context characteristics, respectively, and Q_t^{d} , Q_t^{a} - quality of decisions and actions, respectively.

1.3 Information Quality: Definitions and Metrics

There are several definitions of information quality available in the literature:

- “Quality is the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs” [21].
- “Quality is the degree, to which information has content, form, and time characteristics, which give it value to specific end users” [22].
- “Quality is the degree, to which information is meeting user needs according to external, subjective user perceptions” [23].
- “Quality is fitness for use” [24].

These definitions point to two different sides of information quality: “fitness for use” and “meta-data”. Information quality has to be useful to the users for performing their function and achieving their goals². Assessment of the usefulness of this information (“fitness for use”) is based on the “objective” measurable characteristics of information representing inherent properties of information (Fig. 1.2).

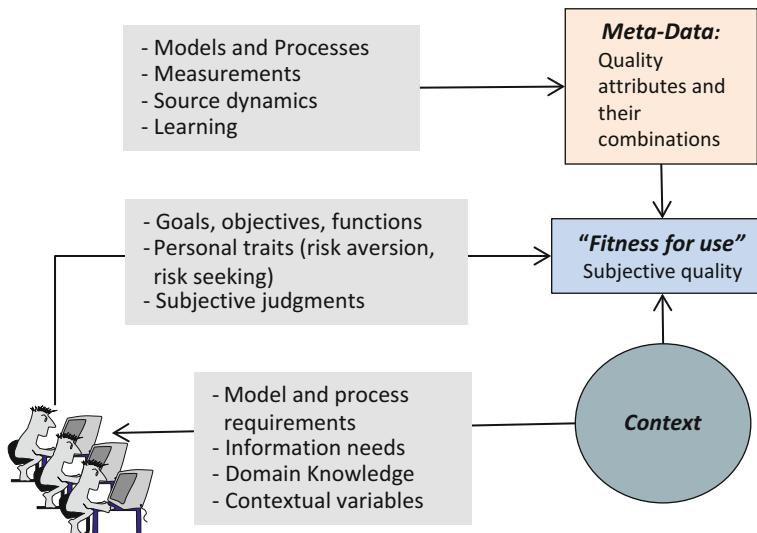


Fig. 1.2 Meta-data and “Fitness for use.” (From [20])

²In the machine-human system, context “users” can be either humans or automated agents and models.

Decision makers have to evaluate these characteristics and decide whether this information is meeting user needs according to external, subjective user perceptions. Thus information quality as “fitness for use” represents subjective or contextual quality and measures the level of satisfaction of users and process designers in relation to their specific goals, objective, and functions in a specific context. The inherent information characteristics are considered independently from information users and context. They give a value to subjective quality and are defined in this chapter as *objective quality* or *meta-data*. This consideration of two different sides of quality is similar to one considered for sensor data in [25] where subjective information quality is referred to as information quality and the objective one as volume of information. *Objective quality* is represented and measured by its attributes since “without clearly defined attributes and their relationships, we are not just unable to assess Information Quality (IQ); we may be unaware of the problem.” [26]. *Meta-data* can be obtained as model and process results from domain knowledge, learning, level of agreement between sources, source dynamic or direct measurements. It is important to notice that meta-data attributes and the method of their evaluation can differ for hard data produced by physical sensors and numerical models, and soft data coming from human sources. Some of the quality characteristics can be both objective and subjective while others represent subjective or contextual data and cannot be measured independently of users and context.

There have been multiple information quality classifications identifying quality attributes and assigning them into broad categories and relations. In [23], quality attributes are obtained by processing the result of a two-stage data consumer survey (so-called empirical approach well established in the marketing disciplines) to capture data consumers’ perspectives on data quality. Four classes of quality dimensions (intrinsic, contextual, representational, and accessibility) were identified as most important for data consumers in industry and government. The model in [26] is “closely parallel” to one in [23] and considers four attributes of quality as “fitness for use”: integrity, accessibility, interpretability, and relevance. In [27], six categories were enumerated: accessibility, accuracy, specificity, timeliness, relevance, and the amount of information for measuring information quality in decision making. In [28], representation of information quality is limited by information imperfection, a subcategory of information quality, which was classified into two general categories: uncertainty and imprecision. At the same time, there is no clear understanding of what dimensions define information quality from the perspective of information fusion process designers and how different dimensions defining information quality are interrelated. The information quality ontology introduced in this chapter represents an attempt to fill this gap.

The type of information exchange in the fusion-based human-system environment, as shown in Fig. 1.1, notes the three main interrelated categories of information quality considered in this paper [18, 29] (Fig. 1.3):

- Quality of *information source*
- Quality of *information content*
- Quality of *information presentation*

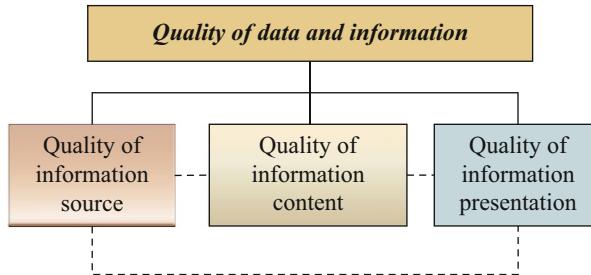


Fig. 1.3 Main quality characteristics and their relationships. (From [29])

As it can be seen in Fig. 1.1, each node of the fusion process (machine or human) can represent an information source and information recipient at the same time, and quality of information/decision/action of such node as a source of information transmitted to the next node depends on the quality of the information and the quality of the process represented by this node. Quality of information presentation is related to human nodes. The importance of considering information presentation as a component of overall quality stems from the fact that the quality of decisions and actions depends not only on incoming information, model outputs, and human mental models but also on the way information is presented to a user. The main quality characteristics are interrelated. Thus for example, the quality of a fusion process depends on the quality of the information source producing input information for this process. The next subsection will describe some components of the main information quality types shown in Fig. 1.3.

1.3.1 *Quality of Information Content*

There are five major attributes of the quality of information content: *accessibility*, *availability*, *relevance*, *timeliness*, and *integrity* or lack of imperfection, which represents a complex notion characterized by many other attributes (Fig. 1.4). Accessibility and availability refer to the ability of users to access this information effectively. Accessibility is related to the cost of obtaining this information, which can be measured by time required for accessing this information. While the cost of obtaining information represents an intrinsic quality of information, its “fitness for use” depends on specific user constraints and a specific context. Users have to compare the cost of obtaining this information with the benefits of its utilization. For example, accessibility of a particular piece of information can be considered good if the time required for obtaining this information is much smaller than the time available for making decisions based on this information. Availability is an important characteristic, which has a binary value since information can be either available or not. If availability is 0, all other attributes are irrelevant. Both

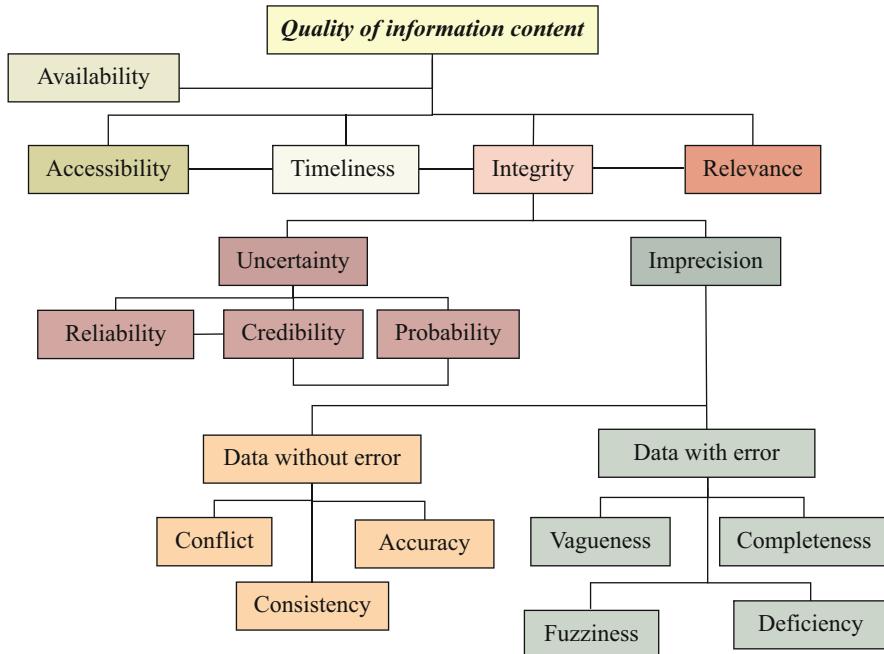


Fig. 1.4 Ontology of quality of information content. (Adopted from [28])

availability and accessibility are time dependent since if information is unavailable now to an agent, it doesn't mean it will always be unavailable to this or all other agents.

Timeliness is a subjective attribute of information and as an attribute of the content of information is different from *timeliness* of information presentation. It can be measured by utility of the information under consideration at the time it becomes available or by threshold satisfaction. Threshold satisfaction compares a time-dependent threshold defined for comparing the quality of information under consideration with the quality of either incoming information or information obtained by multiple intra- and inter-level fusion processes. The threshold and quality characteristics compared with the threshold depend on the context, uses' goals, objectives and functions.

One of the central subjective quality attributes is *relevance*, which determines which information has to be incorporated into fusion process at a particular time. While utilizing relevant information can potentially improve the result of any fusion node provided that information is reliable, utilization of irrelevant information can hamper the outcome of the fusion processes. Relevance is defined by many authors as relation to the matter at hand and therefore depends on context as well as goals and functions of decision makers and therefore relevance is not a property but “is understood as a relation; relevance is a tuple—a notion consisting of a

number of parts that have a relation based on some property or criteria”. The dynamics of context, goals, and functions of the decision makers in the dynamic environment make relevance a temporal attribute. Thus irrelevant information can become relevant later or relevant information can become obsolete at a certain time [30].

Quantification of the level of relevance traditionally is based on the following definition [31]: “On the basis of prior evidence e , a hypothesis h is considered, and the change in the likelihood of h due to additional evidence i is examined. If the likelihood of h is changed by the addition of i to e , i is said to be relevant to h on the evidence e ; otherwise it is irrelevant. In particular, if the likelihood of h is increased due to the addition of i to e , i is said to be positively relevant to h ; if the likelihood is decreased, i is said to be negatively relevant.” Usually relevance analysis processes qualifying relevance are based on two methods [32]: the Probability Covariance and the Mutual Information. Relevance also depends on the quality of information sources (sensor reliability, truthfulness, expertise, reputation, etc.).

Integrity or lack of imperfection of the content of information is the most studied category of information quality (see, e.g., [26, 28, 33–36]). In the context of a human-system integrated environment, imperfection will be defined as something that causes inadequacy or failure of decision making and/or actions. Motivated by [28], we consider here two major characteristics of imperfection: *uncertainty* and *imprecision*. *Uncertainty* “is partial knowledge of the true value of the data.” It arises from either a lack of information or as the result of deficiencies of both formal and cognitive models [28, 36]. It can be either objective and represented by *probabilities* reflecting relative frequencies in repeatable experiments or subjective and represented by *credibility* (believability) describing information which is not completely trustworthy. Uncertainty is the most studied in information fusion component of imperfection [23, 34–38]. A good review of various types of uncertainty is given in [36]. Another uncertainty characteristic, *reliability* (see, e.g., [33]), can be defined in two different ways. Reliability can characterize the level of agreement with ground truth and the quality of credibility estimation represented by beliefs or plausibility assigned to it [33]. It is usually represented by reliability coefficients, which measure adequacy of each belief model to the reality. Incorporation of reliability coefficients is important due to the fact that the majority of fusion operators presume that information sources are equally reliable, which is not always the case. Therefore it is necessary to account for variable information reliability to avoid decreasing in performance of fusion results.

Imprecision can be possessed by so called “information with or without error” [28]. Information without error can be approximate (lacking *accuracy*) or *conflicting* and *inconsistent*. *Accuracy* represents the degree to which data corresponds to characteristics of objects or situations. *Consistency* of a piece of information is usually measured when it is compared with some background knowledge, e.g., databases or knowledge obtained or inferred earlier considered in the context under consideration. Consistency of transient and background information is especially important for situation assessment in the dynamic environment since it can lead to discovery of new and unexpected situations or context discovery. *Conflict*

assumes several pieces of parallel or temporal information contradicting each other and it may occur if either these pieces of information have different reliability or they report on different objects or situations. Information with error can be *incomplete, deficient* (lacking important pieces, which may prevent its usage), *vague* (ambiguous), or *fuzzy* (not well defined).

Such characteristics of imprecision as *inaccuracy, fuzziness, and vagueness* are inherent to soft data that usually are expressed in natural language, which is ambiguous by its nature [39]. Besides, errors in information conveyed by humans can be influenced by cognitive biases and or even be intentional. The ontology of quality of information content adopted in this chapter and presented in Fig. 1.4 is inspired by the one introduced in [28].

There are multiple theories developed to deal with information uncertainty and imprecision. Uncertainty, for example, can be represented and reasoned about in the framework of the probability, Bayesian, belief, and interval probability theories while imprecision can be expressed in the framework of possibility and fuzzy set theories. Selection of one of these theories depends on many factors like, for example, existence of prior probability, type of information (soft, hard or both), whether the hypotheses about the state of environment under consideration are exhaustive, etc. Selection of these theories strictly depends on a context. For example, in highly uncertain open world environment required dealing with both hard and soft data characterized within the different uncertainty representation, the appropriate framework is the Transferable Belief Model (TBM) introduced in [40]. The TBM is a two-level model, in which quantified beliefs in hypotheses about an object or state of the environment are represented and combined at the *credal level* while decisions are made based on probabilities obtained from the combined belief by the *pignistic transformation* at the *pignistic level*. Dempster-Shafer beliefs [41] and probability and possibility [42] distributions can be expressed as belief structures within the framework of the TBM allowing representing both soft and hard information [43]. Beliefs are sub-additive, which permits for numerically expressing *uncertainty* and *ignorance*. Within the TBM, the unnormalized Dempster's rule can combine basic belief masses based on multiple pieces of evidence and allows for incorporation of belief reliability. Moreover, the TBM works under the open world assumption, i.e., it does not assume that the set of hypotheses under consideration is exhaustive. It also permits to represent conflict. These properties of the TBM have been successfully exploited in information fusion in general and in the crisis context specifically (see, e.g., [44–47]).

1.3.2 *Quality of Information Sources*

From the information quality point of view, we consider two types of information sources: *subjective* and *objective*. Quality of objective information sources such as sensors, models, automated processes is free from biases inherent to human judgment and depends only on how well sensors are calibrated and how adequate models

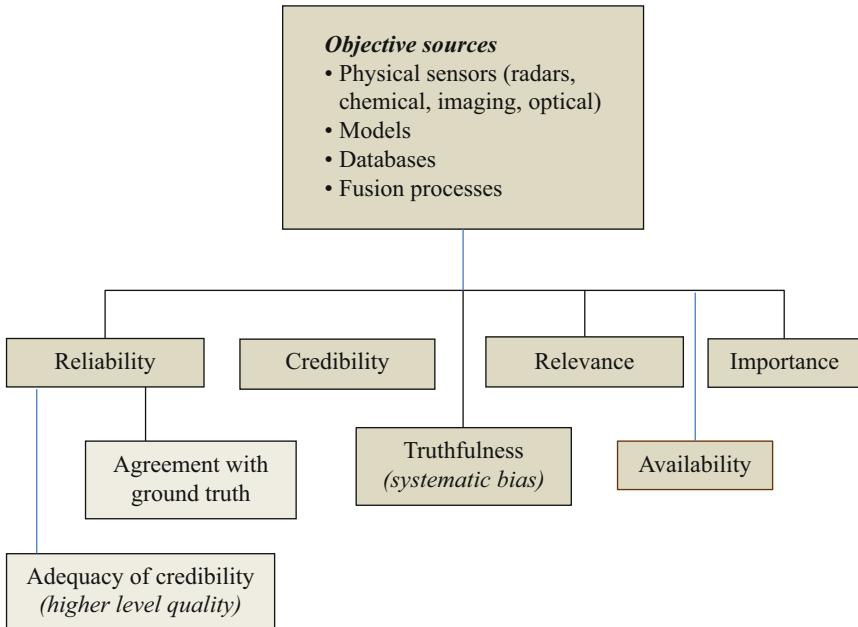


Fig. 1.5 Ontology of source quality. (Objective sources)

are. Objective sources usually deliver information in numerical form (“hard information”). Ontology of the quality of information sources is presented in Fig. 1.5.

The quality of objective sources is represented by *credibility*, *reliability*, systematic bias (*objective truthfulness*), *relevance* and *importance*. Reliability and credibility of objective information sources measure the correctness of the choice of a source model or fusion process and the parameters affecting their performance. They can represent *meta-data* of the source when considered independently of the user of the information and *subjective quality* when this information is considered in context and related to the functions and objective of a user. An objective source can be *irrelevant* if it reports information about attributes of an object or events different from the one under consideration, or it does not work properly, or is not designed to deal with the object or event under consideration.

Subjective sources such as human observers, members of social networks, intelligence agents, newspaper reporters, experts and decision makers, supply observations, subjective beliefs, hypotheses, and opinions about what they see or learn. These sources use subjective judgment that can be affected by imperfect mental models leading to all kind of biases of the observer. Subjective sources can have malicious intent and intentionally provide wrong information or can supply unverified facts obtained from a malicious agent. Information coming from subjective sources is usually represented in nonnumeric unstructured form (“soft information”). Ontology of the quality of subjective information sources is presented in Fig. 1.5.

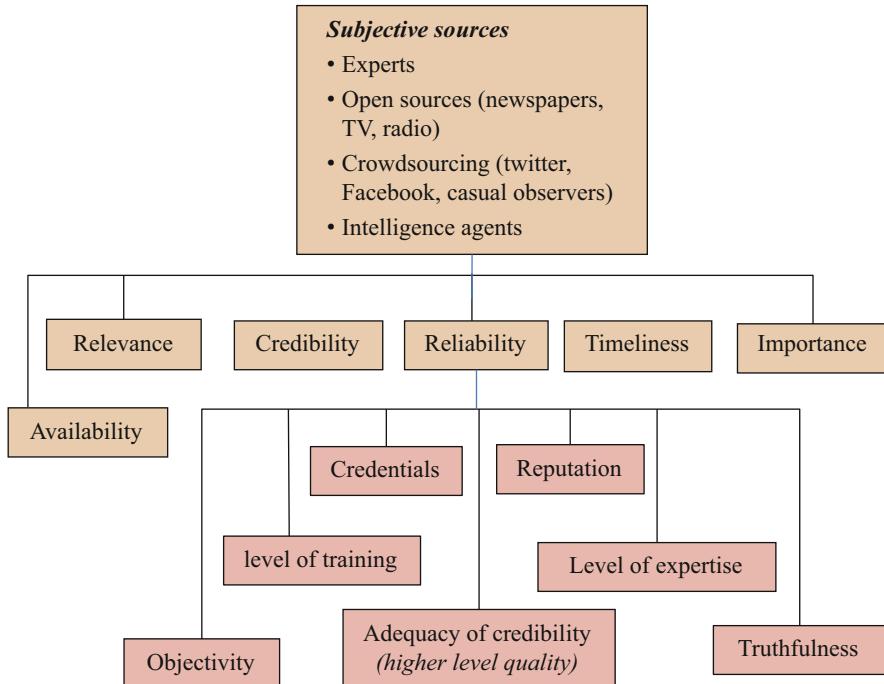


Fig. 1.6 Ontology of the quality of subjective sources

As it can be seen from Figs. 1.5 and 1.6, characteristics of the quality of information sources include *importance*, *credibility*, *reliability*, and *relevance*. Similar to other main quality characteristics, quality of presentation and contents, the source quality characteristics are not independent from each other. For example, importance of information sources depends on relevance and reliability. Selection of important sources requires prioritization (ranking) of them to select the most relevant and reliable to reduce conflict, computation complexity, or decision maker overload. Similar to relevance of information contents and representation, relevance of information sources is a relation and represents subjective characteristic only.

Reliability of subjective sources is measured by additional characteristics as compared to reliability of objective sources. These additional characteristics include *truthfulness*, *level of training and expertise*, *reputation and credentials*. Ideally, *reliability* of a source has to be evaluated by combining all these reliability characteristics. It is important to notice that credible information may not be reliable and reliable information may not be credible. The notion of reliability of subjective sources is related to the notion of *trust* used, for example, in the literature on network centric operations and information sharing (see, e.g., [48]). While there is no consensus in the literature about the notion of trust, following [49] we define trust here as “the psychological state in which (1) the trustor believes that the trustee behaves as expected in a specific context, based on evidence of the trustee’s

competence and goodwill; (2) the trustor is willing to be vulnerable to that belief'. Trustee here can be either an objective or subjective source. Competence and goodwill could be defined by the reliability characteristics shown in Fig. 1.6.

1.3.3 *Quality of Information Presentation*

The quality of information presentation affects perception of decision makers and end users and influences their actions, decisions, judgments, and opinions. Information has to be reliable and must be presented on time and in a way which makes it understandable, complete, and easy to interpret. Thus attributes of the quality of presentation are related to when, which, and how information is presented. The Ontology of quality of information presentation is shown in Fig. 1.7.

Presented information has to be useful for decision makers in performing their functions in a specific context. An overall quality of presentation as well as its characteristics represents subjective quality depending not only on the specific problem and context but also on personal traits of the decision makers, their cognitive ability, biases, and expertise. While some of these characteristics can serve as both *meta-data* and "fitness for use", for example, completeness and timeliness, the other can be ones of subjective quality only. As well as main quality characteristics, characteristics of the quality of presentation are not always independent from each other. Thus, for example, interpretability and understandability depend on each other since in order to interpret presented information one has to understand it and vice versa. Interpretability defines to what extent the users can understand information presented while understandability characterizes the level to which the user is able to follow the logic of automatic processes producing this information. It is important to mention that the quality of such information attributes as interpretability and understandability depends on the level of training and expertise of the user and can be high for one use and poor for the others.

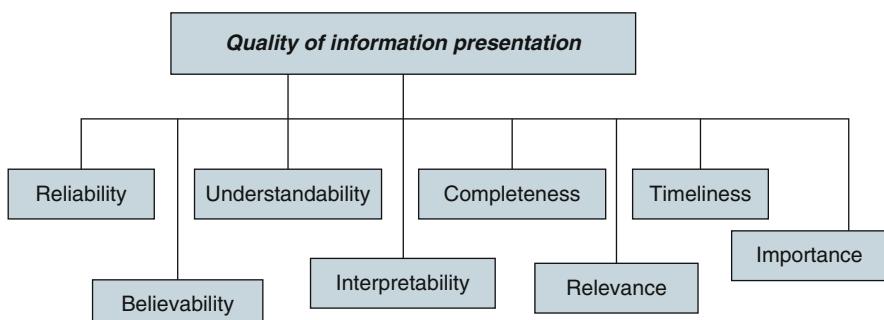


Fig. 1.7 Ontology of quality of information representation

Believability is one of the most important and complex characteristics of information presentation since if decision makers do not believe the information presented, they will not consider it while making decisions. In [23] *believability* is defined as “the extent to which data are accepted or regarded as true, real, and credible.” In this chapter, believability is considered as a function of provenance, its origin (sources) and subsequent processing history and it depends on relevance and reliability of the whole chain of historical information. At the same time, it is considered as “intrinsic,” not context dependent. Here we claim that believability is a subjective context-dependent attribute since information cannot be considered believable if it contradicts expectations provided by context under consideration.

Relevance is another subjective context-dependent characteristic of information presentation. The dynamics of context, goals, and functions of the decision makers in the dynamic environment makes relevance a temporal attribute: irrelevant information can become relevant later or relevant information can become obsolete at a certain time. While relevance is one of the characteristics of all main information quality dimensions (contents, sources, and presentation), relevance as characteristic of information presentation is more than the result of natural language processing or matching algorithms. Humans can find other relevant information that is not detected by an automatic system for a variety of reasons [50]. They provide a level of relevance by using their expertise, prior experience, ideas, and clues and therefore human-based relevance is extremely subjective.

There is a definite connection between cognitive effect of information, information processing time, and relevance [51]:

- The greater the cognitive effects, the greater the relevance is.
- The smaller the processing effort required for deriving these effects, the greater the relevance is.

In a human-machine system, consideration of *relevance* as a characteristic of the quality of information presentation is very important for increasing the information cognitive effect.

Timeliness is affected by two factors: whether the information is presented by the time it must be used and whether the presented dynamic information corresponds to the real world information obtained at the time of the presentation. Timeliness is inevitably affected by the communication delays.

Completeness is the ability of an information system to represent every meaningful state of the real world by representing all necessary information related to this state [52] and is defined by both the quantity and the number of information attributes presented. While presentation of complete information allows for more informed and effective decisions, it is not always beneficial to provide complete information since too much information can overwhelm the user and lead to fatigues and decision errors. Thus it is necessary to establish a trade-off between the level of completeness and the decision quality.

An important problem related to information presentation is whether it is beneficial to present the value of information quality along with information itself

[52, 53]. For example, it was found that the benefits of incorporating meta-data depend on the experience of the decision maker [54].

1.3.4 Higher Level Quality

Higher level quality measures how well the quality of information is assessed³. The needs for considering the higher level quality stem from the fact that the processes of assessing the value of the attributes of information quality have their limitations. For example, assessment of probability may need more observations or more time than may be available, and therefore the result is not completely reliable. In some cases, point-value probabilities cannot be estimated and are presented as intervals representing the *accuracy* of probability estimation. Another example is assessment of sensor reliability that produces a probability distribution over the reliability values. The ontology of information quality shown in Figs. 1.3, 1.4, 1.5, 1.6, and 1.7 may also serve as the basis for building an ontology for higher level quality.

Since comprehensive assessment of quality of information requires taking into consideration higher level quality, it needs establishing relations between attributes of quality at different levels. It is important to mention that while the quality of a source is always a higher level quality for quality of information content, some attributes of information content or an information source can serve as higher level quality for other attributes of information source or information content, respectively. For example, reliability of a model (a quality attribute characterizing an information source) of producing plausibility of a hypothesis is a higher level quality for plausibility assessment, which is an attribute characterizing information content. We can also consider a certain level of belief that a source of information is reliable as a higher level quality for the assessment of reliability of this source. Another example of higher level quality is ambiguity represented as an interval attached to a self-assessed credibility of a statement produced by a subjective source.

1.4 Assessing the Values of Quality Attributes: Where Do the Numbers Come From?

In order to compensate for insufficient information quality in a fusion-based human-machine system, it is necessary to be able to assess the values of quality attributes and combine these values into an overall quality measure.

³Usually this measure is referred to *uncertainty* only and is called “higher order uncertainty,” which is treated without relation to the other quality attributes. Here we define this measure for any quality characteristic and consider it with relation to other attributes.

The methods of assessing information quality represented by single attributes can be assessed by utilizing:

- A priori domain knowledge
- Measurements
- Outputs of models and processes
- Past experience
- A level of agreement between sources
- Source dynamics, i.e., the degree of consistency with prior reports for human observers or prior readings for sensors
- Expert opinions
- Credentials defined by interaction with other sources
- The difference between observed or computed values and expected ones in the context under consideration
- Utility of decisions

Selection of a particular method for defining the value of information quality depends on the attribute under consideration, and information available. Thus *a priori* domain knowledge about the context of the problem under consideration can provide quality values for many attributes, such as source reputation, expertise, level of training, information availability and accessibility, or sensor reliability and credibility. The methods utilizing contextual information have proven themselves very successful provided that known quality of contextual variables is good.

In [46], context-dependent credibility, reliability, and timeliness have been introduced for sequential decision making for threat assessment. The context there was represented by the time-dependent distance between an observed target and a sensor, and a time-dependent threshold on credibility. A case study involving an evidential learning system for sequential decision making designed for a quantitative evaluation of the described method showed the benefits of context considerations. Context as a source of reliability estimation has been reported in many publications (see, e.g., [33, 55–59]). The authors of [55] present two methods of defining source reliability based on contextual information. Both methods utilize subjective contextual information modeled by the theory of fuzzy events and used in connection with probability theory. In [56], expert knowledge is used to represent reliability by a possibility distribution defined on the sensor domain. The computed reliability coefficients are then used in a production system to determine the sources of satisfactory reliability to be used in combination. A method of defining relative reliability of model outputs expressed as beliefs and based on relative distance between bodies of evidence is proposed in [57]. The authors of [58] propose to base evaluation of intelligence information on the correlation between two pieces of information. In many cases, *a priori* domain knowledge does not directly contain values of quality attributes but includes certain information such as training examples, which can be exploited for learning these values [59].

Models and processed outputs can serve as a source of assessing integrity attributes of data obtained with these models such as relevance, level of conflict, fuzziness, ambiguity, credibility. In general, information can be considered relevant

for situation assessment if a change of its value affects the set of hypotheses about objects or situations under consideration, the levels of belief assigned to these hypotheses (e.g., reduces ignorance, strife or reduces/increases conflict), or values of utilities of a set of possible courses of action. Subjective judgment of human experts is used when there is no *a priori* domain knowledge of, e. g. probability of non-repeatable events, or when it is important to know the quality of information from the subjective point of view of an expert, e.g., the level of understandability of information. As it was mentioned above, consistency measures the correspondence of observed or inferred information with expected information obtained from context, databases, and domain knowledge.

While the value of any information quality attribute may be defined in various ways, the users and process designers are mostly interested in how good the quality is. Thus the subjective quality scores usually measure the level of satisfaction with the information under consideration in relation to the decision makers' goals and objectives or the purpose of the models or processes. Such level of satisfaction with objective quality attributes is based on the value of these attributes and can be expressed either in a linguistic form (e.g., good, fair, bad) or numerically by a number from the interval [0 1]. Some attributes, e.g., availability, have binary values (0 or 1) only. The level of satisfaction then can be represented and treated within a certain uncertainty theory (e.g., fuzzy, belief, or possibility).

One of the methods of subjective quality evaluation is a so-called threshold satisfaction that compares a particular quality attribute or a combination of attributes with a certain context-specific and, in many cases, subjective threshold [45, 46]. The difference between a threshold and the value of the quality attribute can represent the level of satisfaction with the level of quality of this attribute. This distance can be transformed into an uncertainty measure within an uncertainty theory under consideration, e.g., belief or possibility, and can measure a belief or possibility that the information quality represented by this attribute or the attribute combination is satisfactory.

In some cases, the quality score of a particular attribute may be defined by comparing a different attribute with a context-specific threshold. Thus, for example, if the reliability of source is lower than a certain threshold, information produced by this source may be considered irrelevant. In this case, the degree of information relevance can be defined by the function of the distance between the threshold and the source reliability. In the human-system environment, relevance can be defined by a human-in-the loop and be represented either by a number between 0 and 1 or in linguistic form (relevant, maybe relevant, irrelevant). It is important to mention that *subjective quality* scores should be considered along with the quality of presentation and the quality (e.g., level of expertise and objectivity), of the users and experts.

One of the most difficult for evaluation and at the same time very important quality attribute is *trust*, which is defined here as measure of reliability and intent of a subjective information source. The problem is that most often the information is coming from open sources such as social networks, or opportunistic sensors such as the bystanders, whose identity, intentions, and source of information may not be known. Trust can be partially based on the valuation of the information that

the source provides. For example, if any portion of that information is believed to be unlikely or inaccurate, the value of trust of the information sources can be reduced [60]. The believability of information is usually evaluated by an information recipient and in many cases subjective. Ideally, trust in a source has to be evaluated by combining all reliability characteristics.

It can be seen from the definition of different types of reliability that we can have either direct or indirect reliability. While direct reliability of hard data can be obtained from context, domain knowledge, and statistical information based on the previous experience/experiments, defining reliability of soft data is a more difficult problem. For example, sources of soft information cannot be trusted if they do not have incentives to tell the truth or enough knowledge about the context in which observations are made. Another problem is that the soft information is rarely characterized by direct reliability since in many cases it comes from a network of agents with variable reliability.

There are several issues to consider in modeling indirect trust such as (see, e.g., [61–65]):

- How sources such as social media can be manipulated
- How one should revise one's notions of trust based on the past actions of individuals
- Which of several competing sources of conflicting information one should trust
- How to take into account reliability of each individual
- What are the methods of propagating reliability through the “reliability network

To address the majority of these issues, it is necessary to consider information provenance (pedigree). Provenance as defined in [66] is information about entities, activities, and people involved in producing a piece of data or a thing to be used to form assessments about their quality, reliability, or trustworthiness. Provenance defines the origins of information and how and by whom this information is interpreted before entering the system. Provenance is used to construct a reliability or a trust network in which nodes represent information sources and links the level of reliability or trust between a pair of sources.

Figure 1.8 shows an information flow in a processing designed for obtaining reliability and trust of information coming from a member of a social network by a human decision maker. Provenance is represented by a reliability network in which nodes denote variable subjective sources of information, and directed links define reliability of information transferred between them. The result of reasoning about reliability of the node under consideration is presented to a decision maker who evaluates its trustworthiness. The final result of the trust estimation is a combination of the assessed reliability of the network under consideration, decision maker's opinion, and reliability of the decision maker himself such as his level of training and objectivity.

Implementation of the processing shown in Fig. 1.8 requires methods for defining relevance of information transmitted between nodes, relevance propagation involving a combination of trust in an individual node, a method of constructing the network, dealing with cycles in this network as well as methods for verifying

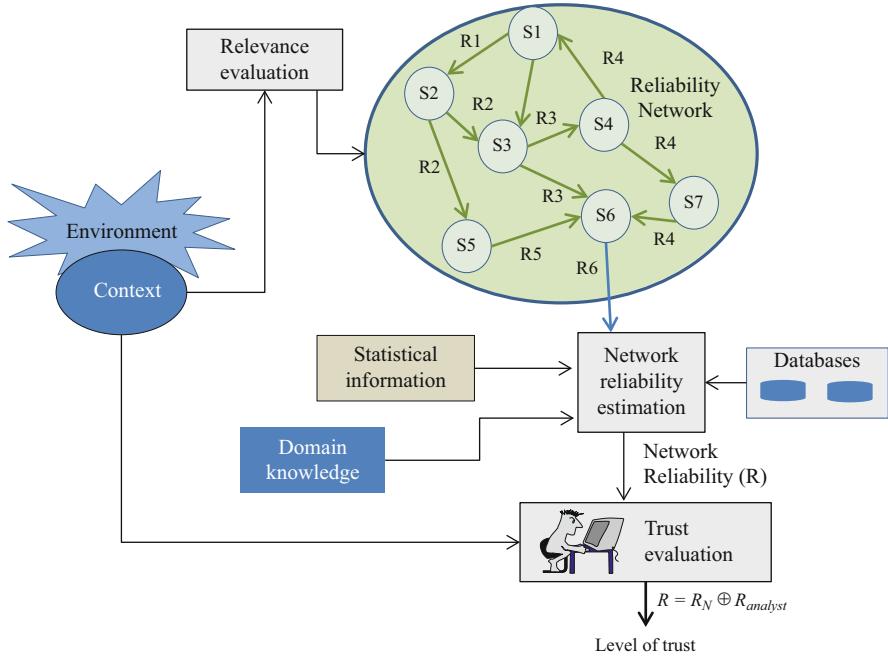


Fig. 1.8 Information flow in the process of obtaining social network reliability and trust evaluation

node independence. It also requires a criterion defining whether the reliability of information presented to an analyst is trustworthy enough to build an argument and assign reliability to it. A review of various models of reliability and trust evaluation is presented in [67]. If reliability is represented by probabilities or beliefs, reliability networks can be considered as belief networks and belief propagation is similar to causal reasoning in probabilistic networks [64, 68].

1.5 Overall Quality Measure

Depending on the context and user requirements, the overall quality measure may relate to a single attribute, a combination of several or all the attributes. Evaluation of the overall quality measure requires a method of combining different quality attributes into a composite measure. The subset of attributes considered depends on user goals, objectives, and functions as well as the purpose of model or process of interest. For example, a combination of credibility, reliability, and timeliness and a method of incorporating their combination into sequential decision making threat assessment have been introduced in [46]. In [47], reliability and credibility were combined; relevance and truthfulness as an overall quality measure were considered in [69]. Other combinations, such as completeness and understandability, completeness and

timeliness, can be also considered. In many cases, the overall quality represents a trade-off between different attributes, for example, between completeness and understandability, or credibility and timeliness. While designing an overall quality measure, one has to take into account the hierarchy of quality attributes, their possible different importance under different contexts and for different users, and the quality of the values assigned to them (higher level quality). The problem of combining several quality attributes into an overall quality measure is similar to the problem of multi-criteria decision making, which requires comparing several alternatives based on the values of the criteria under consideration. Different quality attributes can be considered and different criteria while alternatives are different values of an overall quality measure.

One of the possible ways of representing such unified quality is to consider a t-norm or weighted average of the quality scores defined for each component while the weights are nonnegative and their sum is normalized to unity. The weights representing a trade-off between the attributes under consideration are context specific and can be assigned by the users based on their needs and preferences. A more general representation of an overall quality measure can be obtained by training a neural network, which can serve as a tool for transforming vector of individual quality scores into a subjective unified quality score. The training of this neural network requires a training set of vectors of individual quality scores and an expert who provides the value of outputs. The drawback of this method is the possible cognitive biases of the expert that, however, can be overcome by involving several experts and combining their opinions. Another way of representing a unified quality of information is to consider utility of decisions or actions or total information based on this information. For example, a combination of credibility and timeliness can be measured by the increase of utility of decisions or actions.

If the quality attribute values are represented within the framework of an uncertainty theory, their combination can be obtained by the combination rule defined in this theory. For example, it is possible to use conjunctive combination of reliability and relevance expressed within the framework of the possibility theory. It is also possible to represent the unified quality measure as a belief network in which quality of single attributes is expressed within a belief theory [26]. This method of attribute combination is especially appropriate when values of single attributes are heterogeneous, i.e., expressed in different forms, e.g., point-value numbers, intervals, and linguistic values.

1.6 Quality Control

To ensure successful decision making and actions in the fusion-based machine-human system environment, it is important to deliver the right information at the right time. This requires not only awareness of the quality values but also applying quality control measures. Quality control should be implemented at each step of information exchange. Thus it is necessary to account for such quality attributes as

reliability, credibility, relevance, timeliness when raw data and information enter the fusion processes and delivered to the human users; quality of models and fusion results (credibility, reliability, timeliness, etc.) when information is transferred between and within fusion levels; quality of judgment such as trust, confidence when information is transferred between humans, reliability, credibility, timeliness, etc. when information is transferred between human and automatic module of the system.

Quality control can include (see, e.g., [18]):

- Eliminating information of insufficient quality from consideration
- Incorporating information quality into models and processing by:
 - Employing formal methods, e.g., methods of belief change to deal with inconsistency
 - Modifying the fusion processes to account for data and information quality
- Modifying the data and information by compensating for its quality before processing or presenting to the users
- Delaying transmission of information to the next processing level or to decision makers until it has matured as a result of additional observations and/or computations improving its associated information quality
- Combination of the strategies mentioned above

Selection of a particular quality control method depends on the type of quality attributes under consideration. For example, irrelevant information is usually eliminated from consideration while credibility is often explicitly incorporated into models and processes.

Elimination of information of insufficient quality requires a criterion to be used for deciding when the quality of information is not sufficient. In the majority of cases, such criterion is related to “fitness for use” and depends on context, user’s objective goals, task, and cognitive model. It involves consideration of a user-specific threshold, or information gain, or decision utility. For example, one can consider a time varying threshold on credibility of hypotheses about the state of the environment or accessibility of information measured by the difference between the cost of getting this information and its utility.

Elimination of information can be applied to a subset of multiple sources in order to avoid or mitigate conflict and possible decrease of confidence of the fusion result or to deal with computational complexity to decrease error caused by approximation. This quality control method can be also used to deal with the pedigree problem to avoid double counting decreasing credibility of the fusion results. The problem with pedigree of sources producing hard data (e.g., physical sensors in distributed fusion) received significant attention in the past and is usually approached by employing dynamic network analysis. This problem is becoming more complicated when more and more information is coming from human sensors, e.g., bystanders, Twitter, Facebook, etc. because of often unknown dependence between sources, rumors, the lack of prior information about quality of sources, and possible outdated information. One of the methods of selection of the least

dependent sources is suggested in [70]. This method exploits the notion of “close” and “non-close” source in a social graph to suppress rumors and chain messaging.

As it was described in the previous sections, methods of modifying information for making it totally or at least equally reliable before fusion processing often employ reliability coefficients, which measure the adequacy of the source of incoming information (either soft or hard), model and fusion processes outputs to the reality. Ignoring reliability of information can lead to counterintuitive results [71]. Information modification to account for its reliability before fusion is usually performed by discounting and using the discounted values in the fusion model (see, e.g., [33, 40–42, 58, 60, 72]). Reliability coefficients are also employed for modifying fusion processes.

Methods of compensating for reliability are reviewed in, e.g., [33]. Modification of possibility fusion by incorporating reliability in the case when (1) a subset of sources is reliable but we do not know which one and (2) only an order of the reliabilities of the sources is known but no precise values are known is described in [73].

Delaying the transfer of information demands a balance between timeliness and other characteristics of information quality, e.g., credibility, reliability, accessibility, relevance, etc. This method of quality control is based on the assumption that the quality may be improved over time when more information becomes available. At the same time, in many situations, waiting may result in unacceptable decision latency and unwanted, even catastrophic, consequences. Therefore, the cost of waiting for additional information has to be justified by obtaining the results of better quality and ultimately better decisions and actions. Thus, the timeliness of the decision is defined by a context-dependent balance between the waiting time and improved information quality. There may be several criteria to consider for dealing with the trade-off between decision latency and improved decision outcome. One of these criteria is the Maximum Expected Utility Principle [74]. According to this criterion, a new observation is justified if the difference between maximum expected utility with the new observation and without the new observation is greater than the cost of obtaining this new observation. The core difficulty of utilizing the Maximum Expected Utility Principle is a problem of finding the utility for each decision and utilizing this principle in the highly dynamic uncertain crisis environment. A more appropriate criterion in such environment is threshold satisfaction mentioned in the previous section, which compares a predefined situation, user goals, and domain knowledge based threshold with the quality of information obtained from the sensors (physical or human) or fusion processes. It is important to note that when the users of this information are human the quality threshold also depends on their personal traits, for example, of their level of risk tolerance. The selection of quality attributes to be combined with timeliness and a method of measuring the value of quality depend on a situation. Sensor management (e.g., physical sensors or members of a crisis management team) should be employed when it is either impossible or very costly to get information of acceptable quality with current set of sensor or sensor configuration.

1.7 Conclusions

This chapter has discussed major challenges and some possible approaches addressing the problem of data and information quality in the fusion-based human-machine information environment. In particular, this chapter presents an ontology of quality of information and identifies several potential methods of assessing the values of quality attributes, combining these values into an overall quality measure as well as possible approaches to quality control. Designing the methods of representing and incorporating information quality into fusion systems is a relatively new and a rather difficult problem and more research is needed to confront all its challenges. One of the main challenges is the growing number of opportunistic human sensors, a huge amount of unstructured information coming from social networks, context exploitation, interrelationship between quality of information and quality of context and their effect on fusion system performance, as well as growing importance of human-centric fusion and a related problem of modeling quality of information resulting from this.

References

1. L. Wald, *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolution* (Les Presses, Ecole des Mines de Paris, Paris, 2002)
2. E. Benoit, M-Ph. Huget, M. Patrice, and P. Olivier, Reconfiguration of a distributed information fusion system, Workshop on Dependable Control of Discrete Systems, Bari: Italie, HAL CCSd, Sci. (2009)
3. F. Castanedo, A review of data fusion techniques. *Sci. World J.* **2013**, 704504 (2013). <https://doi.org/10.1155/2013/704504>
4. Y. Lee, L. Pipino, J. Frank, R. Wang, *Journey to Data Quality* (MIT Press, Cambridge, 2006)
5. M. Helfert, Managing and measuring data quality in data warehousing, in *Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics*, pp. 55–65, 2001
6. S.E. Madnick, Y.W. Lee, R.Y. Wang, H. Zhu, Overview and framework for data and information quality research. *ACM J. Data Inf. Qual.* **1**(1), 2 (2009)
7. F. White, A model for data fusion, in *Proceedings of the 1st National Symposium on Sensor Fusion*, 1988
8. E.P. Blasch, S. Plano, Level 5: user refinement to aid the fusion process, in *Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications*, ed. by B. Dasarathy, Proceedings of the SPIE, vol. 5099 (2003)
9. A.N. Steinberg, C.L. Bowman, Rethinking the JDL data fusion model. In: *Proceedings of the MSS National Symposium on Sensor and Data Fusion*, vol. 1, June 2004
10. L. Llinas, C.L. Bowman, G.L. Rogova, A.N. Steinberg, E. Waltz, F. White, Revisions to the JDL Data Fusion Model II, in *Proceedings of the FUSION'2004-7th Conference on Multisource Information Fusion*, Stockholm, 2004
11. S. Schreiber-Ehle, W. Koch, The JDL model of data fusion applied to cyber-defense—A review paper, in *IEEE Workshop on Sensor Data Fusion: Trends Solutions Applications (SDF)* (2012), pp. 116–119
12. E. Blasch, A. Steinberg, S. Das, L. Llinas, C. Chong, O. Kessler, F. White, Revisiting the JDL model for information exploitation, in *Proceedings of the 16th International Conference on Information Fusion*, pp 129–136, 2013

13. B. Dasarathy, Sensor fusion potential exploitation- innovative architectures and illustrative applications. *IEEE Proc.* **85**(1), 24 (1997)
14. M. Bedworth, J. O'Brien, The omnibus model: a new model of data fusion? *IEEE Aerosp. Electron. Syst. Mag.* **15**(4), 30–36 (2000)
15. J. Boyd, *A Discourse on Winning and Losing* (Maxwell AFB Lecture, 1987)
16. M. Markin, C. Harris, M. Bernhardt, J. Austin, M. Bedworth, P. Greenway, R. Johnston, A. Little, D. Lowe, *Technology Foresight on Data Fusion and Data Processing* (The Royal Aeronautical Society, London, England 1997)
17. M. Endsley, Toward a theory of situation awareness in dynamic systems. *Hum. Factors J Hum Factors Ergon Soc* **37**(1), 32–64 (1995)
18. G. Rogova, Information quality in information fusion and decision making with applications to crisis management, in *Fusion Methodology in Crisis Management: Higher Level Fusion and Decision Making*, ed. by G. Rogova, P. Scott, pp. 65–86, (Springer, Cham, 2016)
19. T. Buchholz, A. Kupper, M. Schifflers, Quality of context information: what it is and why we need it, in *Proceedings of the 10th International Workshop of the HP Open View University Association (HPOVUA)*, vol. 200, Geneva, Switzerland, 2003
20. G. Rogova, L. Snidaro, Considerations of context and quality in information fusion, in *Proceedings of the 21st International Conference on Information Fusion*, (IEEE, Cambridge, UK, 2018), pp. 1929–1936
21. Standard 8402, 3, I, *International organization of standards*, 1986
22. J.A. O'Brien, G. Marakas, *Introduction to Information Systems* (McGraw-Hill/Irwin, New York City, US, 2005)
23. R. Wang, D. Strong, Beyond accuracy: what data quality means to data consumers. *J. Manag. Inf. Syst.* **12**, 5–34 (1996)
24. J.B. Juran, A.B. Godfrey, *Juran's Quality Handbook*, 5th edn. (McGraw-Hill, New York, 1988)
25. C. Bisidian, L. Kaplan, M. Srivastava, D. Thornley D. Verma, R. Young, Building principles for a quality of information, specification for sensor information, in: *Proceedings of the 12th International Conference on Information Fusion*, Seattle, WA, USA, pp. 1370–1377, 6–9 July 2009
26. M. Bovee, R.P. Srivastava, B. Mak, A conceptual framework and belief-function approach to assessing overall information quality. *Int. J. Intell. Syst.* **18**, 51–74 (2003)
27. C.A. O'Reilly III, Variations in decision makers' use of information source: the impact of quality and accessibility of information. *Acad. Manag. J.* **25**(4) (1982)
28. P. Smets, Imperfect information: imprecision – uncertainty, in *Uncertainty Management in Information Systems: From Needs to Solutions*, ed. by A. Motro, P. Smets, (Kluwer, Boston, 1997), pp. 225–254
29. G. Rogova, E. Bosse, Information quality in information fusion, in *Proceedings of the 13th International Conference on Information Fusion*, Edinburg, Scotland, July 2010
30. A. Y. Tawfik, E. M. Neufeld, Irrelevance in uncertain temporal reasoning, in *Proceedings of the Third International IEEE Workshop on Temporal Representation and Reasoning*, pp. 196–202, 1996
31. P. Gardenfors, On the logic of relevance. *Synthese* **37**(3), 351 (1978)
32. M.-S. Zhong, L. Liu, R.-Z. Lu, A new method of relevance measure and its applications, in *Proceedings of the IEEE Sixth International Conference on Advanced Language Processing and Web Information Technology*, (2007), pp. 595–600
33. G. Rogova, V. Nimier, Reliability in information fusion: literature survey, in *Proceedings of the FUSION'2004-7th Conference on Multisource- Information Fusion*, (2004), pp. 1158–1165
34. P. Bosc, H. Prade, An introduction to the fuzzy set and possibility theory-based treatment of flexible queries and uncertain or imprecise databases, in *Uncertainty in Information Systems: From Needs to Solutions*, ed. by A. Motro, P. Smets, (Kluwer, Boston, 1997), pp. 285–324
35. M. Smithson, *Ignorance and Uncertainty: Emerging Paradigms* (Springer, New York, 1989)
36. E. Bossé, J. Roy, S. Wark, *Concepts, models, and tools for information fusion* (Artech House, Norwood, 2007)

37. P. Krause, D. Clark, *Representing Uncertain Knowledge: An Artificial Intelligence Approach* (Kluwer Academic Publishers, Dordrecht, 1993)
38. G.J. Klir, M.J. Wierman, Uncertainty-based information, in *Studies in Fuzziness in Soft Computing*, vol. 15, 2nd edn., (Physica-Verlag, Heidelberg, New York, 1999)
39. V. Dragas, An ontological analysis of uncertainty in soft data, in *Proceedings of the 16th International Conference on Information Fusion*, Istanbul, Turkey, pp. 1566–1573, 2013
40. P. Smets, Data fusion in the transferable belief model, in: *Proceedings of the FUSION'2000-Third Conference on Multisource- Multisensor Information Fusion*, pp. 21–33, 2002
41. G. Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, 1976)
42. D. Dubois, H. Prade, *Possibility Theory: An Approach to Computerized Processing of Uncertainty* (Plenum, New York, 1988)
43. R. Yager, Conditional approach to possibility-probability fusion. *IEEE Trans. Fuzzy Syst.* **20**(1), 46–55 (2012)
44. F. Delmotte, P. Smets, Target identification based on the transferable belief model interpretation of Dempster-Shafer model. *IEEE Trans. Syst. Man Cybern. A* **34**, 457–471 (2004)
45. G. Rogova, P. Scott, C. Lollett, R. Mudiyarur, Reasoning about situations in the early post-disaster response environment, in *Proceedings of the FUSION'2006-9th Conference on Multisource Information Fusion*, (2006)
46. G. Rogova, M. Hadrazagic, M-O. St-Hilaire, M. Florea, P. Valin, Context-based information quality for sequential decision making, in *Proceedings of the 2013 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 2013
47. G. Rogova, Adaptive real-time threat assessment under uncertainty and conflict, in *Proceedings of the 4th IEEE Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, San Antonio, TX, 2014
48. H. Hexmoor, S. Wilson, S. Bhattacharam, A theoretical inter-organizational trust-based security model. *Knowl. Eng. Rev.* **21**(2), 127–161 (2006)
49. J. Huang, M.S. Fox, Trust judgment in knowledge provenance, in *Proceedings of the 16th International Workshop on Database and Expert Systems Applications (DEXA '05)*, 2005
50. T. Saracevic, Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II. *J. Am. Soc. Inf. Sci. Technol.* **58**(3), 1915–1933 (2006)
51. H.D. White, Relevance theory and citations. *J. Pragmat.* **43**(14), 3345–3361 (2011)
52. Y. Wang, R. Wang, Anchoring data quality dimensions in ontological foundations. *Commun. ACM* **39**(11), 86–95 (1996)
53. I. Chengalur-Smith, D. Ballou, H. Pazer, The impact of data quality information on decision making: an exploratory analysis. *IEEE Trans. Knowl. Data Eng.* **11**(6), 853–864 (1999)
54. C.W. Fisher, I. Chengalur-Smith, D.P. Ballou, The impact of experience and time on the use of data quality information in decision making. *Inf. Syst. Res.* **14**(2), 170–188 (2003)
55. S. Fabre, A. Appriou, X. Briottet, Presentation and description of two classification methods using data fusion based on sensor management. *Inf. Fusion* **2**, 47–71 (2001)
56. F. Kobayashi, F. Avai, T. Fucuda, Sensor selection by reliability based on possibility measure, in *Proceedings of the International Conference on Robotics and Automation*, Detroit, MI, pp. 2614–2619, 1999
57. W. Jiang, A. Zhang, Q. Yang A new method to determine evidence discounting coefficient. In *Advanced Intelligent Computing Theories and Applications with Aspects of Theoretical and Methodological Issues. ICIC 2008*. Lecture Notes in Computer Science, ed. by D.S. Huang, D.C. Wunsch, D.S. Levine, K.H. Jo, vol. 5226, (Springer, Berlin, Heidelberg, 2008)
58. J. Besombes, V. Nimier, L. Cholvy, Information evaluation in fusion using information correlation, in: *Proceedings of the 12th International Conference on Information Fusion*, Seattle, WA, pp. 264–269, July 2009
59. G. Rogova, J. Kasturi, Reinforcement learning neural network for distributed decision making, in *Proceedings of the Forth Conference on Information Fusion*, August 2001, Montreal, Canada

60. F. Pichon, D. Dubois, T. Denoeux, Relevance and truthfulness in information correction and fusion. *Int. J. Approx. Reason.* **53**(2), 159–175 (2012)
61. D.N. Walton, *Appeal to Expert Opinion: Arguments from Authority* (Penn State University Press, University Park, 1997)
62. J. Lang, M. Spear, S.F. Wu, Social manipulation of online recommender systems, in *Proceedings of the 2nd International Conference on Social Informatics*, 2010
63. Y. Wang, C.W. Hang, M.P. Singh, A probabilistic approach for maintaining trust based on evidence. *J. Artif. Intell. Res.* **40**(1), 221–226 (2011)
64. S. Parsons, K. Atkinson, Z. Li, P. McBurney, E. Sklar, M. Singh, J. Rowe, Argument schemes for reasoning about trust. *Argument Comput.* **5**(2–3), 160–190 (2014)
65. X. L. Dong, L. Berti-Equille, D. Srivastava, Integrating conflicting data: the role of source dependence, in *Proceedings of the 35th International Conference on Very Large Databases*, 2009
66. PROV-DM: the PROV data model, <https://www.w3.org/TR/prov-dm/>
67. A. Jøsang, R. Ismail, C. Boyd, A survey of trust and reputation systems for online service provision. *Decis. Support. Syst.* **43**(2), 618–644 (2007)
68. D. Koller, N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, Cambridge, MA, 2009)
69. H. Prade, A Qualitative Bipolar Argumentative view of trust, in *Scalable Uncertainty Management. SUM 2007*, Lecture Notes in Computer Science, ed. by H. Prade, V.S. Subrahmanian, vol. 4772, (Springer, Berlin, Heidelberg, 2007)
70. M. Uddin, M. Amin, H. Le, T. Abdelzaher, B. Szymanski, T. Nguyen, On diversifying source selection in social sensing, in *Proceedings of the 9th International Conference on Networked Sensing Systems (INSS)*, pp. 1–8, 2012,
71. R. Haenni, Shedding new light on Zadeh's criticism of Dempster's rule, in *Proceedings of the 7th International Conference on Information Fusion (FUSION2005)*, pp. 879–884, 2005
72. J. Schubert, Conflict management in Dempster-Shafer theory by sequential discounting using the degree of falsity, in ed. by L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay *Proceedings of IPMU'08*, Torremolinos (Malaga), pp. 298–305, 22–27 June 2008
73. D. Dubois, H. Prade, Combination of fuzzy information in the framework of possibility theory, in *Data Fusion in Robotics and Machine Intelligence*, ed. by M. A. Abidi, R. C. Gonzalez (Eds), (Academic Press, 1992), pp. 481–505
74. J. von Neuman, O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, 1947)

Chapter 2

Quality of Information Sources in Information Fusion



Frédéric Pichon, Didier Dubois, and Thierry Denœux

Abstract Pieces of information can only be evaluated if knowledge about the quality of the sources of information is available. Typically, this knowledge pertains to the source relevance. In this chapter, other facets of source quality are considered, leading to a general approach to information correction and fusion for belief functions. In particular, the case where sources may partially lack truthfulness is deeply investigated. As a result, Shafer's discounting operation and the unnormalised Dempster's rule, which deal only with source relevance, are considerably extended. Most notably, the unnormalised Dempster's rule is generalised to all Boolean connectives. The proposed approach also subsumes other important correction and fusion schemes, such as contextual discounting and Smets' α -junctions. We also study the case where pieces of knowledge about the quality of the sources are independent. Finally, some means to obtain knowledge about source quality are reviewed.

Keywords Dempster-Shafer theory · Belief functions · Evidence theory · Boolean logic · Information fusion · Discounting

This chapter is a shorter and revised version of [20].

F. Pichon (✉)

EA 3926, Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A),
Université d'Artois, Béthune, France
e-mail: Frederic.Pichon@univ-artois.fr

D. Dubois

CNRS, IRIT (UMR 5505), Université Paul Sabatier, Toulouse Cedex 09, France
e-mail: Didier.Dubois@irit.fr

T. Denœux

CNRS, Heudiasyc (UMR 7253), Sorbonne Universités, Université de Technologie
de Compiègne, Compiègne Cedex, France
e-mail: Thierry.Denoeux@utc.fr; tdenoeux@utc.fr

2.1 Introduction

A central problem in various kinds of information systems is to determine the correct answer to a given question, for instance, find the value of deterministic parameter x defined on a set \mathcal{X} of possible values, from information provided by one or many sources.

Pieces of information about x provided by several sources cannot be evaluated unless some *meta-knowledge* on the sources, i.e. knowledge about their quality, is available. Classically, such meta-knowledge relies on assumptions about the source relevance, where a relevant source is a source providing useful information about x . For instance, if the source is a sensor, then its response is typically relevant when it is in order and irrelevant when it is out of order.

In particular, if a source s is assumed to be relevant with probability p and s provides the testimony $x \in A$, then the information item $x \in A$ is considered not useful with probability $1 - p$. This is known as the *discounting* of a piece of information [25, p. 251], [26] in the context of the theory of belief function [1, 25, 30], and the resulting state of knowledge is represented by a simple support function [25]: the weight p is allocated to the fact of knowing only that $x \in A$ and the weight $1 - p$ is allocated to knowing that $x \in \mathcal{X}$ (it is the probability of knowing nothing from the source). If two sources, with respective independent probabilities of relevance p_1 and p_2 , both supply the information item $x \in A$, then one attaches reliability $p_1 + p_2 - p_1 p_2$ to the statement $x \in A$, since one should deduce that $x \in A$ whenever at least one of the sources is relevant; this is the result obtained by Dempster's rule of combination [1, 25]. This is actually the old problem of merging unreliable testimonies (see the entry "Probabilité" in [5]).

Beyond source relevance, it is proposed in this chapter to consider the case where knowledge on other facets of the quality of the sources is available. We start our study by adding the possibility to make assumptions about the truthfulness of the sources (Sect. 2.2). Then, we present a general framework able to deal with assumptions about various forms of source quality (Sect. 2.3). This study is conducted within the theory of belief functions and leads to a general approach to the fusion of belief functions. Related works, and especially practical means to apply our framework as well as relationships with some previous works, are reviewed in Sect. 2.4, before concluding in Sect. 2.5.

2.2 Relevance and Truthfulness

The reliability of a source is usually assimilated to its relevance. In this section, we assume that reliability also involves another dimension: truthfulness. A truthful source is a source that actually supplies the information it possesses. A source may be non-truthful in different ways. The crudest form of lack of truthfulness for a source is to declare the contrary of what it knows. It may also tell less or something different, even if consistent with its knowledge. For instance, a systematic bias of a

sensor may be regarded as a form of lack of truthfulness. In this section, however, we shall only assume the crudest form of non-truthfulness.

2.2.1 The Case of a Single Source

Consider the case where a single source s provides information about x and that this information is of the form $x \in A$, where A is a proper non-empty¹ subset of \mathcal{X} . If s is assumed to be irrelevant, whatever may be its truthfulness, the information it provides is totally useless and can be replaced by the trivial information $x \in \mathcal{X}$. In contrast, if s is assumed to be relevant and to tell the opposite of what it knows, then the actual information about x can be retrieved: one should replace $x \in A$ by $x \in \overline{A}$, where \overline{A} denotes the complement of A . Obviously, if s is assumed to be relevant and truthful, then one infers that $x \in A$.

Formally, let $\mathcal{H} = \{(R, T), (R, \neg T), (\neg R, T), (\neg R, \neg T)\}$ denote the space of possible states of the source with respect to its relevance and truthfulness, where R (resp. T) means that s is relevant (resp. truthful). Then, following Dempster's approach [1], the above reasoning can be encoded by the multivalued mapping $\Gamma_A : \mathcal{H} \rightarrow \mathcal{X}$ such that

$$\Gamma_A(R, T) = A; \quad (2.1)$$

$$\Gamma_A(R, \neg T) = \overline{A}; \quad (2.2)$$

$$\Gamma_A(\neg R, T) = \Gamma(\neg R, \neg T) = \mathcal{X}. \quad (2.3)$$

$\Gamma_A(h)$ interprets the testimony $x \in A$ in each state $h \in \mathcal{H}$ of s .

In general, the knowledge about the source relevance and truthfulness is uncertain. Specifically, each state $h \in \mathcal{H}$ may be assigned a subjective probability $\text{prob}(h)$ such that $\sum_h \text{prob}(h) = 1$. In such case, the information item $x \in A$ yields the state of knowledge on \mathcal{X} represented by a belief function in the sense of Shafer [25], with mass function $m^{\mathcal{X}}$ on \mathcal{X} defined by

$$m^{\mathcal{X}}(A) = \text{prob}(R, T); \quad (2.4)$$

$$m^{\mathcal{X}}(\overline{A}) = \text{prob}(R, \neg T); \quad (2.5)$$

$$m^{\mathcal{X}}(\mathcal{X}) = \text{prob}(\neg R) = \text{prob}(\neg R, T) + \text{prob}(\neg R, \neg T). \quad (2.6)$$

A mass function $m^{\mathcal{X}}$ is formally a probability distribution on the power set of \mathcal{X} (thus $\sum_{A \subseteq \mathcal{X}} m^{\mathcal{X}}(A) = 1$). The quantity $m^{\mathcal{X}}(A)$ represents the weight allocated to the fact of knowing only that $x \in A$; it does not evaluate the likelihood of event A like does the probability $\text{prob}(A)$.

¹We consider as a source any entity that supplies a non-trivial and non-self-contradictory input.

The testimony provided by the source may itself be uncertain. In particular, it may take the form of a mass function $m_s^{\mathcal{X}}$ on \mathcal{X} . Assume that s is in a given state h , then each $m_s^{\mathcal{H}}(A)$ should be transferred to $\Gamma_A(h)$, yielding the mass function denoted $m^{\mathcal{X}}(B|h)$ and defined by

$$m^{\mathcal{X}}(B|h) = \sum_{A:\Gamma_A(h)=B} m_s^{\mathcal{X}}(A), \quad \forall B \subseteq \mathcal{X}. \quad (2.7)$$

More generally, assume each state h has a probability $\text{prob}(h)$, then (2.7) implies that the state of knowledge on \mathcal{X} is given by the following mass function:

$$m^{\mathcal{X}}(B) = \sum_h m^{\mathcal{X}}(B|h) \text{prob}(h) = \sum_h \text{prob}(h) \sum_{A:\Gamma_A(h)=B} m_s^{\mathcal{X}}(A). \quad (2.8)$$

Assuming that $p = \text{prob}(R)$ and $q = \text{prob}(T)$, and that the relevance of a source is independent of its truthfulness, leads then to transforming the mass function $m_s^{\mathcal{X}}$ into a mass function denoted by $m^{\mathcal{X}}$ and defined by:

$$m^{\mathcal{X}}(A) = pq m_s^{\mathcal{X}}(A) + p(1-q) \bar{m}_s^{\mathcal{X}}(A) + (1-p) m_{\mathcal{X}}^{\mathcal{X}}(A), \quad \forall A \subseteq \mathcal{X}, \quad (2.9)$$

where $\bar{m}_s^{\mathcal{X}}$ is the *negation* of $m_s^{\mathcal{X}}$ [6], defined by $\bar{m}_s^{\mathcal{X}}(A) = m_s^{\mathcal{X}}(\overline{A})$, $\forall A \subseteq \mathcal{X}$, and $m_{\mathcal{X}}^{\mathcal{X}}$ the vacuous mass function defined by $m_{\mathcal{X}}^{\mathcal{X}}(\mathcal{X}) = 1$.

The discounting operation proposed by Shafer [25] to integrate the reliability of information sources is a special case of transformation (2.9), recovered for $q = 1$: it corresponds to a partially relevant source that is truthful. The negating operation [23] is also a special case recovered for $p = 1$, which corresponds to a partially truthful source that is relevant. In particular, the negation of a mass function is obtained for $p = 1$ and $q = 0$: it corresponds to a relevant source that is lying.²

Other forms of uncertain meta-knowledge about the source may be considered. In particular, it may be known only that the source state belongs to a subset H of \mathcal{H} . This happens, for instance, if the source is assumed to be either relevant or truthful but not both, i.e. $H = \{(R, \neg T), (\neg R, T)\}$. In such case, one should deduce that $x \in \Gamma_A(H)$, where $\Gamma_A(H)$ denotes the image of H under Γ_A defined as

$$\Gamma_A(H) = \bigcup_{h \in H} \{\Gamma_A(h)\}.$$

Such nonelementary assumptions are actually not so interesting, since $\Gamma_A(H) = \mathcal{X}$ if $|H| > 1$, where $|H|$ denotes the cardinality of H . Nonetheless, they are important in the case where multiple sources provide information items, which is the topic of the next section.

²The term “lying” is used as a synonym of “telling the negation of what is believed to be the truth”, irrespective of the existence of any intention of a source to deceive.

2.2.2 The Case of Multiple Sources

Interpreting pieces of information provided by two sources requires making assumptions about their joint state with respect to their relevance and truthfulness. Let \mathcal{H}_i denote the set of possible states of source s_i , $i = 1, 2$. The set of elementary joint state assumptions on sources is then $\mathcal{H}_{1:2} = \mathcal{H}_1 \times \mathcal{H}_2$ (we have $|\mathcal{H}_{1:2}| = 16$).

Let us first consider the simple case where each source s_i provides a crisp piece of information $x \in A_i$, $i = 1, 2$, and where the two sources are assumed to be in some joint state $\mathbf{h} = (h^1, h^2) \in \mathcal{H}_{1:2}$, with $h^i \in \mathcal{H}_i$ the state of s_i , $i = 1, 2$. The result of the merging of the information items $\mathbf{A} = (A_1, A_2) \subseteq \mathcal{X} \times \mathcal{X}$ provided by the sources depends on the assumption \mathbf{h} made about their behaviour and can be represented by a multivalued mapping $\Gamma_{\mathbf{A}} : \mathcal{H}_{1:2} \rightarrow \mathcal{X}$.

As one must conclude $x \in \Gamma_{A_i}(h^i)$ when s_i tells $x \in A_i$ and is in state $h_i \in \mathcal{H}_i$, $i = 1, 2$, it is clear that one must deduce that $x \in \Gamma_{A_1}(h^1) \cap \Gamma_{A_2}(h^2)$ when the sources are assumed to be in the joint state $\mathbf{h} = (h^1, h^2) \in \mathcal{H}_{1:2}$, i.e. we have

$$\Gamma_{\mathbf{A}}(\mathbf{h}) = \Gamma_{A_1}(h^1) \cap \Gamma_{A_2}(h^2).$$

Concretely, this means that

- if both sources are relevant and
 - they are both truthful, then one must conclude that $x \in A_1 \cap A_2$;
 - s_1 is truthful and s_2 is not, then one must conclude that $x \in A_1 \cap \overline{A_2}$;
 - s_2 is truthful and s_1 is not, then one must conclude that $x \in \overline{A_1} \cap A_2$;
 - they are both non-truthful, then one must conclude that $x \in \overline{A_1} \cap \overline{A_2}$;
- if source s_i is relevant and the other source is irrelevant, then one must conclude that $x \in A_i$ (resp. $x \in \overline{A_i}$) if s_i is truthful (resp. non-truthful);
- if both sources are irrelevant, then one must conclude that $x \in \mathcal{X}$, irrespective of the elementary assumption made on their truthfulness.

Nonelementary assumptions $H \subseteq \mathcal{H}_{1:2}$ can also be considered. Under such an assumption, one must conclude that $x \in \Gamma_{\mathbf{A}}(H) = \bigcup_{\mathbf{h} \in H} \{\Gamma_{\mathbf{A}}(\mathbf{h})\}$. Among the 2^{16} assumptions $H \subseteq \mathcal{H}_{1:2}$, only a few are interesting (since, for instance, as soon as H contains an elementary assumption such that both sources are nonrelevant, we have $\Gamma_{\mathbf{A}}(H) = \mathcal{X}$). In particular, by considering assumptions pertaining to the number of truthful and/or relevant sources, as well as to logical dependence between source states, it is possible to retrieve other Boolean binary connectives besides the four binary connectives $A_1 \cap A_2$, $A_1 \cap \overline{A_2}$, $\overline{A_1} \cap A_2$ and $\overline{A_1} \cap \overline{A_2}$ retrieved above.

Some interesting cases are as follows:

- Both sources are relevant, and at least one of them is truthful. This induces $x \in A_1 \cup A_2$. Note that this connective is also obtained by other assumptions, such as both sources are truthful, and exactly one of them is relevant.
- Both sources are relevant, exactly one of which is truthful, which yields $x \in A \Delta B$ (*exclusive or*).

- Both sources are relevant, and s_1 is truthful if and only if s_2 is so too. This results in $x \in (A \cap B) \cup (\overline{A} \cap \overline{B})$, which corresponds to the Boolean equivalence connective.

As a matter of fact, each logical connective can be retrieved from assumptions on the global quality of information sources, in terms of relevance and truthfulness. Accordingly, we will denote by \otimes_H the set-theoretic connective associated to the assumption H .

Suppose now that assumption $H \subseteq \mathcal{H}_{1:2}$ is made about the source quality and that s_i supplies an uncertain testimony in the form of a mass function $m_i^{\mathcal{X}}$, $i = 1, 2$. Assume further that the sources are independent, where independence means the following: if we interpret $m_i^{\mathcal{X}}(A_i)$ as the probability that s_i provides statement $x \in A_i$, then the probability that s_1 provides information item $x \in A_1$ and that s_2 provides conjointly information item $x \in A_2$ is $m_1^{\mathcal{X}}(A_1) \cdot m_2^{\mathcal{X}}(A_2)$. In such case, the state of knowledge on \mathcal{X} is represented by the following mass function:

$$m^{\mathcal{X}}(B) = \sum_{A: \Gamma_A(H)=B} m_1^{\mathcal{X}}(A_1) \cdot m_2^{\mathcal{X}}(A_2) \quad (2.10)$$

$$= \sum_{A: A_1 \otimes_H A_2=B} m_1^{\mathcal{X}}(A_1) \cdot m_2^{\mathcal{X}}(A_2), \quad (2.11)$$

which follows from the fact that if s_1 tells $x \in A_1$ and s_2 tells $x \in A_2$, then it is known under assumption H that $x \in A_1 \otimes_H A_2$.

Combination rule (2.11) encompasses the conjunctive rule [28] (the unnormilised version of Dempster's rule) and the disjunctive rule [6]. The former is retrieved by assuming that both sources are relevant and truthful and the latter by assuming that, e.g. both sources are relevant and at least one of them is truthful. Note also that if $A_1 \otimes_H A_2 = \emptyset$ for two sets A_1 and A_2 such that $m_1^{\mathcal{X}}(A_1) > 0$ and $m_2^{\mathcal{X}}(A_2) > 0$, then this inconsistency pertains to a disagreement between the testimonies provided by the sources and the assumption H . In such case, a solution consists in rejecting H and preferring an assumption compatible with the information provided by the sources. This will be discussed further in Sect. 2.4.1.

Assume now that the sources supply crisp testimonies of the form $x \in A_1$ and $x \in A_2$ but that the meta-knowledge regarding source quality is uncertain. Due to the interest of nonelementary assumptions, it seems useful to represent this uncertainty by a mass function, rather than a probability distribution, on $\mathcal{H}_{1:2}$. The merging of A_1 and A_2 under $m^{\mathcal{H}_{1:2}}$ results in a mass function on \mathcal{X} defined by

$$m^{\mathcal{X}}(B) = \sum_{H: A_1 \otimes_H A_2=B} m^{\mathcal{H}_{1:2}}(H). \quad (2.12)$$

Remark that $m^{\mathcal{H}_{1:2}}$ induces a probability distribution over the Boolean binary connectives attached to assumptions H :

$$p^{\mathcal{H}_{1:2}}(\otimes) = \sum_{H: \otimes_H = \otimes} m^{\mathcal{H}_{1:2}}(H). \quad (2.13)$$

In addition, let us stress that $m^{\mathcal{H}_{1:2}}$ may carry a form of independence, which we call *meta-independence* between sources, different from the independence between sources defined above. Indeed, information supplied by the sources may be independent from each other, but pieces of meta-knowledge about the source state may not be independent. Formally, meta-independence between sources may be modelled by assuming that $m^{\mathcal{H}_{1:2}}(H) = m^{\mathcal{H}_1}(H_1)m^{\mathcal{H}_2}(H_2)$ if $H = H_1 \times H_2$ and $m^{\mathcal{H}_{1:2}}(H) = 0$ otherwise, which corresponds to evidential independence [25] between frames \mathcal{H}_1 and \mathcal{H}_2 with respect to $m^{\mathcal{H}_{1:2}}$.

A more general case, extending the combination operations (2.11) and (2.12), is obtained when both the testimonies supplied by independent sources and the meta-knowledge about their quality are uncertain and represented by mass functions. This case induces the following mass function on \mathcal{X} :

$$m^{\mathcal{X}}(B) = \sum_H m^{\mathcal{H}_{1:2}}(H) \sum_{\mathbf{A}: A_1 \otimes_H A_2 = B} m_1^{\mathcal{X}}(A_1)m_2^{\mathcal{X}}(A_2) \quad (2.14)$$

$$= \sum_{\otimes} p^{\mathcal{H}_{1:2}}(\otimes) \sum_{\mathbf{A}: A_1 \otimes A_2 = B} m_1^{\mathcal{X}}(A_1)m_2^{\mathcal{X}}(A_2). \quad (2.15)$$

This approach can be formally extended to the case of dependent sources, using the setting of [3]. It can also be readily extended to the case of $n > 2$ sources that are partially truthful and relevant (see [22]). In theory, this latter extension may pose a computational issue (the belief function expressing meta-knowledge on the sources has a 2^{4^n} complexity in the general case). However, in practice, it can remain manageable as illustrated in [22], where special cases of the general combination rule (2.14) are considered. Besides the conjunctive and disjunctive rules, there are indeed other important combination schemes subsumed by (2.14). In particular, it is the case for the method, used in various approaches (see, e.g. [2, 14, 24, 31]), that consists in discounting sources and then combining them by the conjunctive rule, as shown in [20] and further discussed in [22]. The weighted average of belief functions and the combination rule corresponding to the assumption that r -out-of- n sources are relevant, with $1 \leq r \leq n$, are also included in (2.14), as shown in [20] and [22], respectively.

2.3 A General Model of Meta-knowledge

In the preceding section, meta-knowledge concerned the relevance and the crudest form of lack of truthfulness of sources of information. However, in some applications, the lack of truthfulness may take a more refined form. Moreover, knowledge about the source quality may even be different from knowing their relevance and

truthfulness. An approach to account for general source behaviour assumptions is thus necessary. Such an approach is proposed in this section.

2.3.1 The Case of a Single Source

Suppose a source \mathfrak{s} provides information about a parameter y defined on a set \mathcal{Y} of possible values and that this piece of information is of the form $y \in A$, for some $A \subseteq \mathcal{Y}$. Assume further that \mathfrak{s} may be in one of N elementary states instead of four as is the case in Sect. 2.2.1, i.e. we generalise the state space from $\mathcal{H} = \{(R, T), (R, \neg T), (\neg R, T), (\neg R, \neg T)\}$ to $\mathcal{H} = \{h_1, \dots, h_N\}$. Moreover, we are interested by the value taken by a related parameter x defined on a domain \mathcal{X} , and we have at our disposal some meta-knowledge that relate information item $y \in A$ supplied by \mathfrak{s} to an information of the form $x \in B$, for some $B \subseteq \mathcal{X}$, for each possible state $h \in \mathcal{H}$ of \mathfrak{s} . Namely, for each $A \subseteq \mathcal{Y}$, there is a multivalued mapping $\Gamma_A : \mathcal{H} \rightarrow \mathcal{X}$ prescribing, for each elementary assumption $h \in \mathcal{H}$, how to interpret on \mathcal{X} information item $y \in A$ provided by \mathfrak{s} . We also add the natural requirement that there exists $h \in \mathcal{H}$ such that $\Gamma_{A_1}(h) \neq \Gamma_{A_2}(h)$ for any two distinct subsets A_1 and A_2 of \mathcal{Y} . Incomplete assumptions $H \subseteq \mathcal{H}$ may also be considered, in which case information item $y \in A$ is interpreted as $x \in \Gamma_A(H) = \cup_{h \in H} \Gamma_A(h)$.

The setting of Sect. 2.2.1 is obtained as a particular case of this approach, by choosing $N = 4$, $y = x$ and, e.g. $h_1 = (R, T)$, $h_2 = (R, \neg T)$, $h_3 = (\neg R, T)$, $h_4 = (\neg R, \neg T)$, in which case we have, for all $A \subseteq \mathcal{X}$:

$$\begin{aligned}\Gamma_A(h_1) &= A, \\ \Gamma_A(h_2) &= \overline{A}, \\ \Gamma_A(h_3) &= \Gamma_A(h_4) = X.\end{aligned}\tag{2.16}$$

Example 1 provides another illustration of this approach.

Example 1 (Case $Y \neq X$, inspired from Janez and Appriou [12]) We are interested in finding the type x of a given road, with x taking its value in the set $X = \{\text{track, lane, highway}\}$. A source \mathfrak{s} provides information on this type, but it has a limited perception of the possible types of road and in particular is not aware of the existence of the type “lane”, so that it provides information on the space $Y = \{\text{track, highway}\}$. In addition, we know that this source discriminates between roads either using their width or their texture. If the source uses the road width, then when it says “track”, we may only safely infer that the type is “track or lane” since tracks and lanes have similar width, and when it says “highway”, we may infer “highway”. On the other hand, if the source uses the road texture, then when it says “track”, we may infer “track”, and when it says “highway”, we may only infer “highway or lane” since highways and lanes have similar textures.

This problem may be formalised using multivalued mappings Γ_{track} , Γ_{highway} , and Γ_Y from $\mathcal{H} = \{\text{width}, \text{texture}\}$ to \mathcal{X} defined as

$$\begin{aligned}\Gamma_{\text{track}}(\text{width}) &= \{\text{track}, \text{lane}\}, \\ \Gamma_{\text{track}}(\text{texture}) &= \{\text{track}\}, \\ \Gamma_{\text{highway}}(\text{width}) &= \{\text{highway}\}, \\ \Gamma_{\text{highway}}(\text{texture}) &= \{\text{lane}, \text{highway}\}, \\ \Gamma_Y(\text{width}) &= X, \\ \Gamma_Y(\text{texture}) &= X.\end{aligned}$$

□

The proposed approach also makes it possible to model refined forms of lack of truthfulness, as explained hereafter.

By taking a closer look at the non-truthful state $\neg T$ considered in Sect. 2.2.1, we can remark that it corresponds to assuming that source s tells the contrary of what it knows, whatever it is telling concerning each of the possible values $x_i \in \mathcal{X}$ that admits parameter x , since one must invert what s tells for each of these values. For instance, let $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ and suppose that s asserts $x \in A = \{x_1, x_3\}$, i.e. it tells that x_1 and x_3 are possible values for x and that x_2 and x_4 are not possible values for x . Then, if s is assumed to be in state $\neg T$, one must deduce that $x \in \overline{A} = \{x_2, x_4\}$, i.e. x_1 and x_3 are not possible values for x and x_2 and x_4 are possible values for x .

Accordingly, we may introduce the notion of the truthfulness of a source *for a value* $x_i \in \mathcal{X}$: a truthful (resp. non-truthful) source for a value $x_i \in \mathcal{X}$ is a source that tells what it knows (resp. the contrary of what it knows) for this value. Hence, state $\neg T$ corresponds to the assumption that a source is non-truthful for *all* values $x_i \in \mathcal{X}$. It is therefore a quite strong model of the lack of truthfulness of a source.

It seems interesting to consider more subtle forms of lack of truthfulness and in particular the assumption that a source s could be non-truthful only for *some* values $x_i \in \mathcal{X}$ (and truthful for all other values $x_i \in \mathcal{X}$), i.e. a kind of *contextual* lack of truthfulness. Let $B \subseteq \mathcal{X}$ be the set of values for which s is truthful, and \overline{B} the set of values for which it is not truthful. We will denote this state by ℓ_B (the state $\neg T$ corresponds then to the state ℓ_\emptyset , and T corresponds to $\ell_{\mathcal{X}}$). As shown in [23, Proposition 1], if s asserts $x \in A$ for some $A \subseteq \mathcal{X}$, and is assumed to be in state ℓ_B , for some $B \subseteq \mathcal{X}$, then one must deduce that $x \in (A \cap B) \cup (\overline{A} \cap \overline{B})$. We refer the reader to [23] for examples where such states ℓ_B may be relevant.

Considering the space of possible states ℓ_B , i.e. $\mathcal{H} = \{\ell_B | B \subseteq \mathcal{X}\}$, and a testimony $x \in A$ supplied by a source, the above transformation can then be encoded by the multivalued mapping $\Gamma_A : \mathcal{H} \rightarrow \mathcal{X}$ defined as:

$$\Gamma_A(\ell_B) = (A \cap B) \cup (\overline{A} \cap \overline{B}), \quad \forall B \subseteq \mathcal{X}.$$

A further refined model of lack of truthfulness can be obtained by being even more specific about the assumptions underlying the state $\neg T$, using the notions of positive and negative clauses [11, Chapter 8] told by the source. For instance, when s declares that x_1 is a possible value for x , this is a positive clause told by the source, and when s declares that x_2 is not a possible value for x , it is a negative clause. Accordingly, we may characterise the truthfulness of a source for each $x_i \in \mathcal{X}$, with respect to the *polarity* of the clauses it tells. Specifically, a source is said to be *positively* truthful (resp. non-truthful) for a value $x_i \in \mathcal{X}$, when it declares that x_i is a possible value for x and knows that it is (resp. it is not). Hence, when a source is assumed to be positively truthful (resp. non-truthful) for $x_i \in \mathcal{X}$ and declares that x_i is a possible value for x , then one must deduce that it is (resp. it is not). Similarly, a source is said to be *negatively* truthful (resp. non-truthful) for a value $x_i \in \mathcal{X}$, when it declares that x_i is a not possible value for x and knows that it is not (resp. it is). Hence, when a source is assumed to be negatively truthful (resp. non-truthful) for $x_i \in \mathcal{X}$ and declares that x_i is not a possible value for x , then one must deduce that it is not (resp. it is). Accordingly, state $\neg T$ corresponds to assuming that a source is positively *and* negatively non-truthful for *all* values $x_i \in \mathcal{X}$. In that case, we make two strong assumptions: the context (set of values) concerned by the lack of truthfulness is the *entire* frame, and *both* polarities are concerned by the lack of truthfulness.

This suggests again to consider states corresponding to weaker assumptions on the lack of truthfulness. Two states are particularly interesting, as shown in [23]. The first one, denoted p_B , corresponds to the assumption that a source is (positively and negatively) truthful for all $x_i \in B$ and positively non-truthful and negatively truthful for all $x_i \in \bar{B}$. Under such an assumption p_B , a testimony $x \in A$ is transformed into knowing that $x \in A \cap B$ [23, Proposition 2]. The second one, denoted n_B , corresponds to the assumption that a source is positively truthful and negatively non-truthful for all $x_i \in B$ and (positively and negatively) truthful for all $x_i \in \bar{B}$. A testimony $x \in A$ is transformed into $x \in A \cup B$ under this latter assumption [23, Proposition 3].

These states also fit our approach since, e.g. the transformations associated to states n_B can be represented by a multivalued mapping Γ_A from $\mathcal{H} = \{n_B | B \subseteq \mathcal{X}\}$ to \mathcal{X} such that $\Gamma_A(n_B) = A \cup B$, for all $B \subseteq \mathcal{X}$. Let us remark that states ℓ_B , p_B and n_B , with associated transformations $(A \cap B) \cup (\bar{A} \cap \bar{B})$ (logical equality), $A \cap B$ (conjunction) and $A \cup B$ (disjunction), given testimony $x \in A$, are particular cases of a more general model of truthfulness assumptions yielding all possible binary Boolean connectives between testimony A and context B , as detailed in [23].

The proposed approach can be further generalised to the case where both the information provided by the source and the meta-knowledge on the source are uncertain and represented by mass functions $m_s^{\mathcal{Y}}$ and $m^{\mathcal{H}}$, respectively. Since each mass $m_s^{\mathcal{Y}}(A)$ should be transferred to $\Gamma_A(H)$ under some hypothesis $H \subseteq \mathcal{H}$, the state of knowledge given $m_s^{\mathcal{Y}}$ and $m^{\mathcal{H}}$ is represented by a mass function defined by

$$m^{\mathcal{X}}(B) = \sum_H m^{\mathcal{H}}(H) \sum_{A: \Gamma_A(H)=B} m_1^{\mathcal{Y}}(A), \quad \forall B \subseteq X, \quad (2.17)$$

which generalises (2.8). Transformation (2.17) is referred to as *behaviour-based correction* (BBC) as it modifies, or *corrects* [17], the information supplied by the source given our knowledge on its behaviour.

As detailed in [20], the BBC procedure generalises a deconditioning method known as the method by association of highest compatible hypotheses [13], which itself generalises a familiar operation of Dempster-Shafer theory, called conditional embedding (or ballooning extension) [26].

As shown recently in [23], the BBC procedure applied with states n_B and p_B , respectively, can also be used to provide an interpretation to the contextual discounting and contextual reinforcement operations, which are two correction mechanisms introduced formally in [18]. In addition, when applied with state ℓ_B , it can be used to obtain a contextual version of the negating operation [23].

2.3.2 The Case of Multiple Sources

Consider now that two sources s_1 and s_2 provide information about y and that each source may be in one of N elementary states (those N states are the same for both sources), with \mathcal{H}_i the set of possible states of source s_i , $i = 1, 2$. Let $\mathcal{H}_{1:2} = \mathcal{H}_1 \times \mathcal{H}_2$ denote the set of elementary joint state assumptions on the sources.

Assume that each source s_i provides a crisp piece of information $x \in A_i$, $i = 1, 2$. For the same reason as in Sect. 2.2.2, if the sources are assumed to be in state $\mathbf{h} = (h^1, h^2) \in \mathcal{H}_{1:2}$, then we must conclude that $x \in \Gamma_{A_1}(h^1) \cap \Gamma_{A_2}(h^2)$, where Γ_{A_i} , $i = 1, 2$, are the mappings defined in Sect. 2.3.1. We can then define a multivalued mapping $\Gamma_{\mathbf{A}} : \mathcal{H}_{1:2} \rightarrow \mathcal{X}$, which assigns to each elementary hypothesis $\mathbf{h} \in \mathcal{H}_{1:2}$ the result of the fusion of the information items $\mathbf{A} = (A_1, A_2) \subseteq \mathcal{X} \times \mathcal{X}$, as follows:

$$\Gamma_{\mathbf{A}}(\mathbf{h}) = \Gamma_{A_1}(h^1) \cap \Gamma_{A_2}(h^2).$$

As in Sect. 2.2.2, incomplete assumptions $H \subseteq \mathcal{H}_{1:2}$ can be considered. In such case, one must conclude that $x \in \Gamma_{\mathbf{A}}(H) = \bigcup_{\mathbf{h} \in H} \{\Gamma_{\mathbf{A}}(\mathbf{h})\}$. More generally, suppose that the information supplied by the sources s_1 and s_2 and the meta-knowledge about their behaviour are uncertain and represented by mass functions $m_1^{\mathcal{Y}}$, $m_2^{\mathcal{Y}}$ and $m^{\mathcal{H}_{1:2}}$, respectively. If furthermore the sources are assumed to be independent, then the fusion of $m_1^{\mathcal{Y}}$ and $m_2^{\mathcal{Y}}$ given $m^{\mathcal{H}_{1:2}}$ results in the mass function on \mathcal{X} defined by

$$m^{\mathcal{X}}(B) = \sum_H m^{\mathcal{H}_{1:2}}(H) \sum_{\mathbf{A}: \Gamma_{\mathbf{A}}(H)=B} m_1^{\mathcal{Y}}(A_1) m_2^{\mathcal{Y}}(A_2), \quad (2.18)$$

which generalises (2.14). The combination (2.18) will be referred to as the *behaviour-based fusion* (BBF) rule.

Remark that this rule can be straightforwardly extended to the case of $n > 2$ sources [22]. In addition, it can be extended to the case where sources s_i , $i = 1, \dots, n$ provide information on different spaces \mathcal{Y}_i and have a different number N_i of elementary states, which may be faced in some problems.

In [27], Smets introduced a family of combination rules representing the set of associative, commutative and linear operators for belief functions with the vacuous mass function as neutral element. This family, called the α -conjunctions, depends on a parameter $\alpha \in [0, 1]$. Operators in this family range between two rules based on Boolean operators as α decreases: the conjunctive rule (for $\alpha = 1$) and the so-called equivalence rule [19] (for $\alpha = 0$) based on the logical equivalence operator. Smets did not provide any clear interpretation for these rules for the cases $\alpha \in (0, 1)$. As shown in [19], they are a particular case of the BBF rule: they correspond to assuming that either both sources tell the truth or they commit the same contextual lie ℓ_B , with some particular weight depending on α and B .

2.3.3 The Case of Meta-independent Sources

Interestingly, the BBC procedure and the BBF rule can be recovered by defining particular valuation networks [15], representing the available pieces of information, and by propagating uncertainty in these networks, as stated by Lemmas 1 and 2 with associated valuation networks shown in Fig. 2.1a, b, respectively.

Lemma 1 (BBC) Let $m^{\mathcal{H}}$ and $m_s^{\mathcal{Y}}$ be the mass functions and Γ_A , $A \subseteq \mathcal{Y}$, the mappings in (2.17). Let $\mathcal{Z} = 2^{\mathcal{Y}}$, and let z_A denote the element of \mathcal{Z} corresponding to the subset A of \mathcal{Y} , $\forall A \subseteq \mathcal{Y}$. For each $A \subseteq \mathcal{Y}$, let Γ_{z_A} be the multivalued mapping

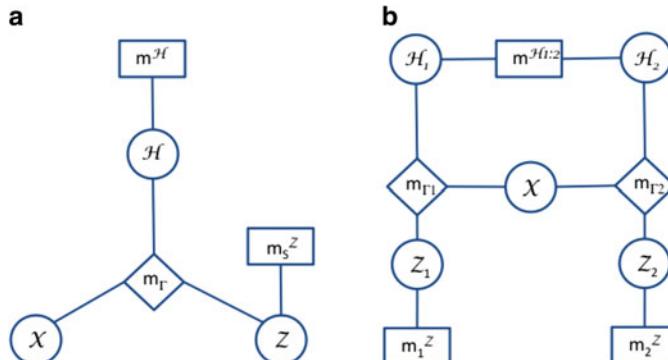


Fig. 2.1 Valuation networks corresponding to the BBC procedure (a) and the BBF rule (b)

from \mathcal{H} to \mathcal{X} such that $\Gamma_{z_A}(h) = \Gamma_A(h)$ for all $h \in \mathcal{H}$. Let $m_{\mathfrak{s}}^{\mathcal{Z}}$ be the mass function defined by $m_{\mathfrak{s}}^{\mathcal{Z}}(\{z_A\}) = m_{\mathfrak{s}}^{\mathcal{Y}}(A)$, for all $A \subseteq \mathcal{Y}$. Let $m_{\Gamma}^{\mathcal{H} \times \mathcal{Z} \times \mathcal{X}}$ be the mass function defined by $m_{\Gamma}^{\mathcal{H} \times \mathcal{Z} \times \mathcal{X}}\left[\bigcup_{h \in \mathcal{H}, z_A \in \mathcal{Z}} (\{h\} \times z_A \times \Gamma_{z_A}(h))\right] = 1$.

We have, for all $B \subseteq \mathcal{X}$,

$$\left(m_{\mathfrak{s}}^{\mathcal{Z}} \circledcirc m_{\Gamma}^{\mathcal{H} \times \mathcal{Z} \times \mathcal{X}} \circledcirc m^{\mathcal{H}}\right)^{\downarrow \mathcal{X}}(B) = m^{\mathcal{X}}(B),$$

where $m^{\mathcal{X}}$ is the mass function defined by (2.17) and where \circledcirc and \downarrow denote, respectively, the unnormalised Dempster's rule on product spaces and the marginalisation operation whose definitions are provided on [15, p.8] and [15, p.9], respectively.

Proof From Theorem 1 of [15], using deletion sequence \mathcal{Z}, \mathcal{H} , we obtain

$$\left(m_{\mathfrak{s}}^{\mathcal{Z}} \circledcirc m_{\Gamma}^{\mathcal{H} \times \mathcal{Z} \times \mathcal{X}} \circledcirc m^{\mathcal{H}}\right)^{\downarrow \mathcal{X}} = \left(\left(m_{\mathfrak{s}}^{\mathcal{Z}} \circledcirc m_{\Gamma}^{\mathcal{H} \times \mathcal{Z} \times \mathcal{X}}\right)^{\downarrow \mathcal{H} \times \mathcal{X}} \circledcirc m^{\mathcal{H}}\right)^{\downarrow \mathcal{X}}. \quad (2.19)$$

Let $m_{\mathfrak{s}\Gamma}^{\mathcal{H} \times \mathcal{Z} \times \mathcal{X}} := m_{\mathfrak{s}}^{\mathcal{Z}} \circledcirc m_{\Gamma}^{\mathcal{H} \times \mathcal{Z} \times \mathcal{X}}$. Equation (2.19) may then be rewritten

$$\left(m_{\mathfrak{s}}^{\mathcal{Z}} \circledcirc m_{\Gamma}^{\mathcal{H} \times \mathcal{Z} \times \mathcal{X}} \circledcirc m^{\mathcal{H}}\right)^{\downarrow \mathcal{X}} = \left(m_{\mathfrak{s}\Gamma}^{\mathcal{H} \times \mathcal{Z} \times \mathcal{X}} \downarrow \mathcal{H} \times \mathcal{X} \circledcirc m^{\mathcal{H}}\right)^{\downarrow \mathcal{X}}.$$

We have

$$m_{\mathfrak{s}\Gamma}^{\mathcal{H} \times \mathcal{Z} \times \mathcal{X}}(C) = \begin{cases} m_{\mathfrak{s}}^{\mathcal{Z}}(\{z_A\}) & \text{if } C = (\bigcup_{h \in \mathcal{H}, z_A \in \mathcal{Z}} \{h\} \times z_A \times \Gamma_{z_A}(h)) \cap (\mathcal{H} \times z_A \times \mathcal{X}), \\ 0 & \text{otherwise.} \end{cases}$$

For all $z_A \in \mathcal{Z}$ we have

$$\left[\left(\bigcup_{h \in \mathcal{H}, z_A \in \mathcal{Z}} \{h\} \times z_A \times \Gamma_{z_A}(h) \right) \cap (\mathcal{H} \times z_A \times \mathcal{X}) \right] \downarrow \mathcal{H} \times \mathcal{X} = \bigcup_{h \in \mathcal{H}} \{h\} \times \Gamma_{z_A}(h).$$

Hence, $m_{\mathfrak{s}\Gamma}^{\mathcal{H} \times \mathcal{Z} \times \mathcal{X}} \downarrow \mathcal{H} \times \mathcal{X}(B)$ for any $B \subseteq \mathcal{H} \times \mathcal{X}$ can be obtained by summing over all $z_A \in \mathcal{Z}$ such that $\bigcup_{h \in \mathcal{H}} \{h\} \times \Gamma_{z_A}(h) = B$:

$$\begin{aligned} m_{\mathfrak{s}\Gamma}^{\mathcal{H} \times \mathcal{Z} \times \mathcal{X}} \downarrow \mathcal{H} \times \mathcal{X}(B) &= \sum_{z_A \in \mathcal{Z}: \bigcup_{h \in \mathcal{H}} \{h\} \times \Gamma_{z_A}(h) = B} m_{\mathfrak{s}}^{\mathcal{Z}}(\{z_A\}) \\ &= \sum_{A \subseteq \mathcal{Y}: \bigcup_{h \in \mathcal{H}} \{h\} \times \Gamma_A(h) = B} m_{\mathfrak{s}}^{\mathcal{Y}}(A). \end{aligned}$$

Since $\exists h \in \mathcal{H}$ such that $\Gamma_{A_1}(h) \neq \Gamma_{A_2}(h)$ for any two distinct subsets A_1 and A_2 of \mathcal{Y} , we obtain

$$m_{\mathfrak{s}\Gamma}^{\mathcal{H} \times \mathcal{Z} \times \mathcal{X} \downarrow \mathcal{H} \times \mathcal{X}} \left[\bigcup_{h \in \mathcal{H}} (\{h\} \times \Gamma_A(h)) \right] = m_{\mathfrak{s}}^{\mathcal{Y}}(A), \quad \forall A \subseteq \mathcal{Y}.$$

The lemma follows then from Lemma 1 of [20]. \square

Lemma 2 (BBF) *Let $m^{\mathcal{H}_{1:2}}$ and $m_i^{\mathcal{Y}}$, $i = 1, 2$, be the mass functions in (2.18). For $i = 1, 2$, let $\mathcal{Z}_i = 2^{\mathcal{Y}}$, and let z_A^i denote the element of \mathcal{Z}_i corresponding to the subset A of \mathcal{Y} . For each $A \subseteq \mathcal{Y}$ and $i = 1, 2$, let $\Gamma_{z_A^i}$ be the multivalued mapping from \mathcal{H}_i to \mathcal{X} such that $\Gamma_{z_A^i}(h) = \Gamma_A(h)$ for all $h \in \mathcal{H}_i$. Let $m_i^{\mathcal{Z}_i}$ be the mass function defined by $m_i^{\mathcal{Z}_i}(\{z_A^i\}) = m_i^{\mathcal{Y}}(A)$, for all $A \subseteq \mathcal{Y}$. For $i = 1, 2$, let $m_{\Gamma_i}^{\mathcal{H}_i \times \mathcal{Z}_i \times \mathcal{X}}$ be the mass function defined by $m_{\Gamma_i}^{\mathcal{H}_i \times \mathcal{Z}_i \times \mathcal{X}} \left[\bigcup_{h \in \mathcal{H}_i, z_A^i \in \mathcal{Z}_i} (\{h\} \times z_A^i \times \Gamma_{z_A^i}(h)) \right] = 1$.*

We have, for all $B \subseteq \mathcal{X}$,

$$\left(m_1^{\mathcal{Z}_1} \circledcirc m_{\Gamma_1}^{\mathcal{H}_1 \times \mathcal{Z}_1 \times \mathcal{X}} \circledcirc m_2^{\mathcal{Z}_2} \circledcirc m_{\Gamma_2}^{\mathcal{H}_2 \times \mathcal{Z}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_{1:2}} \right)^{\downarrow \mathcal{X}}(B) = m^{\mathcal{X}}(B),$$

where $m^{\mathcal{X}}$ is the mass function defined by (2.18).

Proof From Theorem 1 of [15], using deletion sequence $\mathcal{Z}_1, \mathcal{Z}_2$, we obtain

$$\begin{aligned} & \left(m_1^{\mathcal{Z}_1} \circledcirc m_{\Gamma_1}^{\mathcal{H}_1 \times \mathcal{Z}_1 \times \mathcal{X}} \circledcirc m_2^{\mathcal{Z}_2} \circledcirc m_{\Gamma_2}^{\mathcal{H}_2 \times \mathcal{Z}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_{1:2}} \right)^{\downarrow \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{X}} \\ &= \left(m_1^{\mathcal{Z}_1} \circledcirc m_{\Gamma_1}^{\mathcal{H}_1 \times \mathcal{Z}_1 \times \mathcal{X}} \right)^{\downarrow \mathcal{H}_1 \times \mathcal{X}} \circledcirc \left(m_2^{\mathcal{Z}_2} \circledcirc m_{\Gamma_2}^{\mathcal{H}_2 \times \mathcal{Z}_2 \times \mathcal{X}} \right)^{\downarrow \mathcal{H}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_{1:2}}. \end{aligned} \tag{2.20}$$

For $i = 1, 2$, let $m_{i\Gamma}^{\mathcal{H}_i \times \mathcal{X}} := \left(m_i^{\mathcal{Z}_i} \circledcirc m_{\Gamma_i}^{\mathcal{H}_i \times \mathcal{Z}_i \times \mathcal{X}} \right)^{\downarrow \mathcal{H}_i \times \mathcal{X}}$. Equation (2.20) may then be rewritten

$$\begin{aligned} & \left(m_1^{\mathcal{Z}_1} \circledcirc m_{\Gamma_1}^{\mathcal{H}_1 \times \mathcal{Z}_1 \times \mathcal{X}} \circledcirc m_2^{\mathcal{Z}_2} \circledcirc m_{\Gamma_2}^{\mathcal{H}_2 \times \mathcal{Z}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_{1:2}} \right)^{\downarrow \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{X}} \\ &= m_{1\Gamma}^{\mathcal{H}_1 \times \mathcal{X}} \circledcirc m_{2\Gamma}^{\mathcal{H}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_{1:2}}. \end{aligned} \tag{2.21}$$

The transitivity of marginalisation [15] yields

$$\begin{aligned} & \left(\left(m_1^{\mathcal{Z}_1} \circledcirc m_{\Gamma_1}^{\mathcal{H}_1 \times \mathcal{Z}_1 \times \mathcal{X}} \circledcirc m_2^{\mathcal{Z}_2} \circledcirc m_{\Gamma_2}^{\mathcal{H}_2 \times \mathcal{Z}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_{1:2}} \right)^{\downarrow \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{X}} \right)^{\downarrow \mathcal{X}} \\ &= \left(m_1^{\mathcal{Z}_1} \circledcirc m_{\Gamma_1}^{\mathcal{H}_1 \times \mathcal{Z}_1 \times \mathcal{X}} \circledcirc m_2^{\mathcal{Z}_2} \circledcirc m_{\Gamma_2}^{\mathcal{H}_2 \times \mathcal{Z}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_{1:2}} \right)^{\downarrow \mathcal{X}}, \end{aligned}$$

from which we obtain, using (2.21),

$$\begin{aligned} & \left(m_1^{\mathcal{Z}_1} \circledcirc m_{\Gamma 1}^{\mathcal{H}_1 \times \mathcal{Z}_1 \times \mathcal{X}} \circledcirc m_2^{\mathcal{Z}_2} \circledcirc m_{\Gamma 2}^{\mathcal{H}_2 \times \mathcal{Z}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_{1:2}} \right)^{\downarrow \mathcal{X}} \\ &= \left(m_{1\Gamma}^{\mathcal{H}_1 \times \mathcal{X}} \circledcirc m_{2\Gamma}^{\mathcal{H}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_{1:2}} \right)^{\downarrow \mathcal{X}}. \end{aligned}$$

From the proof of Lemma 1, we have for $i = 1, 2$,

$$m_i^{\mathcal{H}_i \times \mathcal{X}} \left[\bigcup_{h \in \mathcal{H}_i} (\{h\} \times \Gamma_A(h)) \right] = m_i^{\mathcal{Y}}(A), \quad \forall A \subseteq \mathcal{Y}.$$

The lemma follows then from Lemma 2 of [20]. \square

Lemmas 1 and 2 are instrumental to show Theorem 1, which concerns meta-independent sources.

Theorem 1 *With meta-independent sources, it is equivalent to combine the uncertain information $m_1^{\mathcal{Y}}$ and $m_2^{\mathcal{Y}}$ by the BBF rule or to combine by the conjunctive rule each of these pieces of information corrected using the BBC procedure.*

Proof Let $m^{\mathcal{H}_1}$ and $m^{\mathcal{H}_2}$ represent meta-knowledge on the two sources s_1 and s_2 , respectively. Meta-independence of s_1 and s_2 is equivalent to $m^{\mathcal{H}_{1:2}} = m^{\mathcal{H}_1} \circledcirc m^{\mathcal{H}_2}$. Under this assumption, we have, with the same notations as above:

$$\begin{aligned} & m_1^{\mathcal{Z}_1} \circledcirc m_{\Gamma 1}^{\mathcal{H}_1 \times \mathcal{Z}_1 \times \mathcal{X}} \circledcirc m_2^{\mathcal{Z}_2} \circledcirc m_{\Gamma 2}^{\mathcal{H}_2 \times \mathcal{Z}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_{1:2}} \\ &= m_1^{\mathcal{Z}_1} \circledcirc m_{\Gamma 1}^{\mathcal{H}_1 \times \mathcal{Z}_1 \times \mathcal{X}} \circledcirc m_2^{\mathcal{Z}_2} \circledcirc m_{\Gamma 2}^{\mathcal{H}_2 \times \mathcal{Z}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_1} \circledcirc m^{\mathcal{H}_2}. \end{aligned}$$

From Theorem 1 of [15], using deletion sequence $\mathcal{Z}_1, \mathcal{Z}_2, \mathcal{H}_2, \mathcal{H}_1$, we obtain

$$\begin{aligned} & \left(m_1^{\mathcal{Z}_1} \circledcirc m_{\Gamma 1}^{\mathcal{H}_1 \times \mathcal{Z}_1 \times \mathcal{X}} \circledcirc m_2^{\mathcal{Z}_2} \circledcirc m_{\Gamma 2}^{\mathcal{H}_2 \times \mathcal{Z}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_1} \circledcirc m^{\mathcal{H}_2} \right)^{\downarrow \mathcal{X}} \\ &= \left(\left(m_1^{\mathcal{Z}_1} \circledcirc m_{\Gamma 1}^{\mathcal{H}_1 \times \mathcal{Z}_1 \times \mathcal{X}} \right)^{\downarrow \mathcal{H}_1 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_1} \right)^{\downarrow \mathcal{X}} \\ &\quad \circledcirc \left(\left(m_2^{\mathcal{Z}_2} \circledcirc m_{\Gamma 2}^{\mathcal{H}_2 \times \mathcal{Z}_2 \times \mathcal{X}} \right)^{\downarrow \mathcal{H}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_2} \right)^{\downarrow \mathcal{X}}, \end{aligned}$$

which, using (2.19), can be rewritten

$$\begin{aligned} & \left(m_1^{\mathcal{Z}_1} \circledcirc m_{\Gamma 1}^{\mathcal{H}_1 \times \mathcal{Z}_1 \times \mathcal{X}} \circledcirc m_2^{\mathcal{Z}_2} \circledcirc m_{\Gamma 2}^{\mathcal{H}_2 \times \mathcal{Z}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_1} \circledcirc m^{\mathcal{H}_2} \right)^{\downarrow \mathcal{X}} \\ &= \left(m_1^{\mathcal{Z}_1} \circledcirc m_{\Gamma 1}^{\mathcal{H}_1 \times \mathcal{Z}_1 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_1} \right)^{\downarrow \mathcal{X}} \circledcirc \left(m_2^{\mathcal{Z}_2} \circledcirc m_{\Gamma 2}^{\mathcal{H}_2 \times \mathcal{Z}_2 \times \mathcal{X}} \circledcirc m^{\mathcal{H}_2} \right)^{\downarrow \mathcal{X}}. \end{aligned}$$

The theorem follows then from Lemmas 1 and 2. \square

2.4 Related Works

The framework described in the previous sections for the correction and fusion of pieces of information is theoretical. It does not include any practical means to apply it and in particular means to obtain the meta-knowledge that it requires. This latter issue is discussed in Sect. 2.4.1. In addition, other approaches have been proposed to exploit meta-knowledge about sources. They are related to ours in Sect. 2.4.2.

2.4.1 Obtaining Meta-knowledge

Both the BBC procedure (2.17) and the BBF rule (2.18) require meta-knowledge on the sources, in the form of a mass function $m^{\mathcal{H}}$ on some space \mathcal{H} of assumptions about the sources, where the transformations of a source testimony associated to these assumptions are encoded by multivalued mappings from \mathcal{H} to \mathcal{X} . A central issue is thus to obtain such meta-knowledge. Two main situations can be distinguished with respect to this problem.

First, one may have some prior information about the sources. This information may take the form of data. In particular, one may have access to a confusion matrix counting the correct (crisp and precise) outputs of a source and its errors. As detailed in [8, 16], it is possible to estimate the relevance or the truthfulness of a source from such data. If one has access to the uncertain outputs of sources for some known objects, then one may search for the meta-knowledge that induces the least errors [7], and specifically, as shown recently in [23], the meta-knowledge associated with the contextual discounting, contextual reinforcement and contextual negating operations can be learnt efficiently. Prior information about the sources may also take the form of expert knowledge. For instance, schemes using such knowledge and relying on multicriteria aggregations have been proposed to evaluate the reliability of information sources [4, 21].

In the absence of prior information about the sources, a piece of meta-knowledge that induces a good tradeoff between specificity and consistency of the inferred knowledge about x can be selected [22]. To make this principle operational, it is proposed in [22] to consider an ordered collection $\mathbf{m}^{\mathcal{H}} = (m_1^{\mathcal{H}}, \dots, m_M^{\mathcal{H}})$ such that the piece of meta-knowledge $m_1^{\mathcal{H}}$ corresponds to the conjunctive rule and, for any $1 \leq j < M$, $m_j^{\mathcal{H}}$ induces a more specific knowledge about x than $m_{j+1}^{\mathcal{H}}$ whatever the source testimonies may be. Then, using the fact that $m_j^{\mathcal{H}}$ necessarily induces a less consistent knowledge on x than $m_{j+1}^{\mathcal{H}}$, one should test iteratively each $m_j^{\mathcal{H}}$ with $j = 1, \dots, M$, and select the first one for which a sufficient degree of consistency is obtained. As illustrated in [22], collection $\mathbf{m}^{\mathcal{H}}$ can be based on important fusion schemes, such as discount and combine, the r -out-of- n relevant sources assumption or the α -conjunctions. Besides, this general approach subsumes some classical fusion strategies and in particular sequential discounting approaches [22].

2.4.2 Other Modelling Approaches

We have already seen that the approach presented in Sect. 2.2.1 extends the discounting operation, which corresponds to the case where the source is known to be truthful, but has only a probability of being relevant. Smets [29] proposed a counterpart to this operation, in which the source is relevant but is not truthful, which is also clearly extended by our approach.

The approach described in Sect. 2.3.1, i.e. the BBC procedure, subsumes the ballooning extension and contextual correction mechanisms, as already mentioned. It is also more general than the partially relevant information sources model proposed by Haenni and Hartmann [10], as explained in [20].

An extension of the discounting operation was proposed in [17], in which uncertain meta-knowledge on a source s is quantified by a mass function $m^{\mathcal{H}}$ on a space $\mathcal{H} = \{h_1, \dots, h_N\}$ of possible states of the source, as is the case for the BBC procedure. The interpretation of states $h \in \mathcal{H}$ is given in this extension by transforms $m_h^{\mathcal{X}}$ of $m_s^{\mathcal{X}}$: if the source supplies the uncertain testimony $m_s^{\mathcal{X}}$ and is in state h , then our knowledge on x is represented by $m_h^{\mathcal{X}}$. As detailed in [20], this extension and BBC coincide in the special case where the mass function $m^{\mathcal{H}}$ on \mathcal{H} is Bayesian and $m_h^{\mathcal{X}}$ is defined from $m_s^{\mathcal{X}}$ using multivalued mappings Γ_A as: $m_h^{\mathcal{X}}(B) = \sum_{A: \Gamma_A(h)=B} m_s^{\mathcal{Y}}(A)$, for all $B \subseteq \mathcal{X}$. Nonetheless, the two models are not equivalent in general, and one should use one model or the other depending on the nature of available knowledge.

Finally, we may remark that alternatives to Dempster's rule, where intersection is replaced with other set operations, have already been considered in [6]. However, the approach of Sect. 2.2.2 is the first to provide an explicit interpretation for the resulting rules. In addition, the idea in Sect. 2.3.3 of using valuation networks to recover the BBC procedure and the BBF rule is inspired from similar approaches [9, 26] used to recover the discounting operation and the disjunctive rule.

2.5 Conclusions

In this chapter, a general approach to the correction and fusion of belief functions has been proposed. It integrates meta-knowledge about the sources of information, that is, knowledge about their quality. An important particular case where meta-knowledge concerns the relevance and the truthfulness of the sources has been deeply studied. It significantly extends Shafer's discounting operation and the unnormalised Dempster's rule, which deal only with source relevance. Various forms of lack of truthfulness have also been considered. Specifically, different definitions of the contextual non-truthfulness of a source have been introduced. With these definitions, contextual discounting and Smets' α -junctions can be seen as special cases of the proposed correction and fusion procedures, respectively. In addition, we have proved that these procedures can be recovered by propagating

uncertainty in some particular valuation networks. This result allowed us to show that, when the behaviours of the sources are independent, it is equivalent to combine the sources' information using our fusion procedure or to combine the pieces of information modified using our correction procedure by the unnormalised Dempster's rule. Finally, practical means to apply this general approach have been reviewed, making it a potentially useful tool for various problems, such as combining classifiers or handling intelligence reports.

References

1. A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* **38**, 325–339 (1967)
2. T. Denœux, A k -nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern.* **25**(5), 804–213 (1995)
3. S. Destercke, D. Dubois, Idempotent conjunctive combination of belief functions: extending the minimum rule of possibility theory. *Inf. Sci.* **181**(18), 3925–3945 (2011)
4. S. Destercke, P. Buche, B. Charnomordic, Evaluating data reliability: an evidential answer with application to a web-enabled data warehouse. *IEEE Trans. Knowl. Data Eng.* **25**(1), 92–105 (2013)
5. D. Diderot, J. le Rond d'Alembert (eds.), *Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers*. University of Chicago: ARTFL Encyclopédie Project (Spring 2016 Edition), ed. by R. Morrissey, G. Roe. <http://encyclopedie.uchicago.edu/>
6. D. Dubois, H. Prade, A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. *Int. J. Gen. Syst.* **12**(3), 193–226 (1986)
7. Z. Elouedi, K. Mellouli, P. Smets, Assessing sensor reliability for multisensor data fusion within the transferable belief model. *IEEE Trans. Syst. Man Cybern. B* **34**(1), 782–787 (2004)
8. Z. Elouedi, E. Lefèvre, D. Mercier, Discountings of a belief function using a confusion matrix, in *22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, vol 1, Oct 2010, pp. 287–294
9. R. Haenni, Uncover Dempster's rule where it is hidden, in *Proceedings of the 9th International Conference on Information Fusion (FUSION 2006)*, Florence (2006)
10. R. Haenni, S. Hartmann, Modeling partially reliable information sources: a general approach based on Dempster-Shafer theory. *Inf. Fusion* **7**(4), 361–379 (2006)
11. R.D. Huddleston, G.K. Pullum, *A Student's Introduction to English Grammar* (Cambridge University Press, Cambridge, 2005)
12. F. Janez, A. Appriou, Théorie de l'évidence et cadres de discernement non exhaustifs. *Traitement du Signal* **13**(3), 237–250 (1996)
13. F. Janez, A. Appriou, Theory of evidence and non-exhaustive frames of discernment: plausibilities correction methods. *Int. J. Approximate Reason.* **18**, 1–19 (1998)
14. J. Klein, O. Colot, Automatic discounting rate computation using a dissent criterion, in *Proceedings of the 1st Workshop on the Theory of Belief Functions (BELIEF)*, Brest (2010), pp. 1–6
15. J. Kohlas, P.P. Shenoy, Computation in valuation algebras, in *Handbook of Defeasible Reasoning and Uncertainty Management Systems: Algorithms for Uncertainty and Defeasible Reasoning*, ed. by D.M. Gabbay, Ph. Smets, vol. 5 (Kluwer, Dordrecht, 2000), pp. 5–39
16. E. Lefèvre, F. Pichon, D. Mercier, Z. Elouedi, B. Quost, Estimation de sincérité et pertinence à partir de matrices de confusion pour la correction de fonctions de croyance, in *Rencontres Francophones sur la Logique Floue et ses Applications (LFA)*, Cépaduès, Nov 2014, vol. 1, pp. 287–294

17. D. Mercier, T. Denœux, M.-H. Masson, Belief function correction mechanisms, in *Foundations of Reasoning Under Uncertainty*, ed. by B. Bouchon-Meunier, R.R. Yager, J.-L. Verdegay, M. Ojeda-Aciego, L. Magdalena. Volume 249 of Studies in Fuzziness and Soft Computing (Springer, Berlin, 2010), pp. 203–222
18. D. Mercier, E. Lefèvre, F. Delmotte, Belief functions contextual discounting and canonical decompositions. *Int. J. Approximate Reason.* **53**(2), 146–158 (2012)
19. F. Pichon, On the α -conjunctions for combining belief functions, in *Belief Functions: Theory and Applications*, ed. by T. Denœux, M.-H. Masson. Volume 164 of Advances in Intelligent and Soft Computing (Springer, Berlin/Heidelberg, 2012), pp. 285–292
20. F. Pichon, D. Dubois, T. Denœux, Relevance and truthfulness in information correction and fusion. *Int. J. Approximate Reason.* **53**(2), 159–175 (2012)
21. F. Pichon, C. Labreuche, B. Duqueroie, T. Delavallade, Multidimensional approach to reliability evaluation of information sources, in *Information Evaluation*, ed. by P. Capet, T. Delavallade (Wiley, Hoboken, 2014), pp. 129–156
22. F. Pichon, S. Destercke, T. Burger, A consistency-specificity trade-off to select source behavior in information fusion. *IEEE Trans. Cybern.* **45**(4), 598–609 (2015)
23. F. Pichon, D. Mercier, E. Lefèvre, F. Delmotte, Proposition and learning of some belief function contextual correction mechanisms. *Int. J. Approximate Reason.* **72**, 4–42 (2016)
24. J. Schubert, Conflict management in Dempster-Shafer theory using the degree of falsity. *Int. J. Approximate Reason.* **52**(3), 449–460 (2011)
25. G. Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, 1976)
26. P. Smets, Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *Int. J. Approximate Reason.* **9**(1), 1–35 (1993)
27. P. Smets, The α -junctions: combination operators applicable to belief functions, in *First International Joint Conference on Qualitative and Quantitative Practical Reasoning (ECSQARU-FAPR'97)*, Bad Honnef, 09 June 1997–12 June 1997, ed. by D.M. Gabbay, R. Kruse, A. Nonnengart, H.J. Ohlbach. Volume 1244 of Lecture Notes in Computer Science (Springer, 1997), pp. 131–153
28. P. Smets, The transferable belief model for quantified belief representation, in *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, ed. by D.M. Gabbay, P. Smets, vol. 1 (Kluwer, Dordrecht, 1998), pp. 267–301
29. P. Smets, Managing deceitful reports with the transferable belief model, in *8th International Conference on Information Fusion*, Philadelphia, July 2005. <https://doi.org/10.1109/ICIF.2005.1591953>
30. P. Smets, R. Kennes, The transferable belief model. *Artif. Intell.* **66**, 191–243 (1994)
31. Y. Yang, D. Han, C. Han, Discounted combination of unreliable evidence using degree of disagreement. *Int. J. Approximate Reason.* **54**(8), 1197–1216 (2013)

Chapter 3

Using Quality Measures in the Intelligent Fusion of Probabilistic Information



Ronald R. Yager and Frederick E. Petry

Abstract Our objective here is to obtain quality-fused values from multiple sources of probabilistic distributions, where quality is related to the lack of uncertainty in the fused value and the use of credible sources. We first introduce a vector representation for a probability distribution. With the aid of the Gini formulation of entropy, we show how the norm of the vector provides a measure of the certainty, i.e., information, associated with a probability distribution. We look at two special cases of fusion for source inputs, those that are maximally uncertain and certain. We provide a measure of credibility associated with subsets of sources. We look at the issue of finding the highest-quality fused value from the weighted aggregation of source-provided probability distributions.

Keywords Quality measures · Information fusion · Entropy · Credibility · Choquet integral

3.1 Introduction

The concept of quality informs many applications and disciplines. It is difficult to give a definition of quality that can cover the aspects of quality for so many areas. Terms such as worth, merit, value, etc. have often been synonymously used. There have been many general viewpoints on the meaning and concept of quality both abstractly and concretely. Abstractly this is an issue in philosophy such as the metaphysics of quality [1]. Concretely it has been extensively considered under the topic of quality control [2] as the issue ultimately of product quality.

R. R. Yager (✉)
Iona College, New Rochelle, NY, USA
e-mail: yager@panix.com

F. E. Petry
Naval Research Laboratory, Stennis Space Center, St. Louis, MS, USA
e-mail: fred.petry@nrlssc.navy.mil

Here we are interested in the quality of information [3, 4] that will be the product of the fusion process we present in this chapter. The use of fusion to combine data provided by multiple sources about the value of a variable is common in many applications [5]. One rationale for fusing probabilistic distributions provided by multiple sources is to improve the quality of the information to decision-makers [6]. Our interest here is looking at the problem of obtaining high-quality fused values. One aspect of this quality is a reduction in the uncertainty of the information. Unfortunately, combining probability distributions information does not always result in a probability distribution with less uncertainty, this particularly is the case when the data that are being fused are conflicting. In order to formally quantify the uncertainty associated with a probability distribution, we will use the concept of entropy. A second contributing factor to the association of quality with a fused value is that we have used quality sources of information; the more of these sources used, the more credible the results of the fusion process. In order to capture these criteria of a quality fusion, we introduce a measure of credibility associated with use of various subsets of the sources. Here we provide a quantification of the notion of a quality fusion based on the objective of providing fused values having little uncertainty based on a credible subset of the sources.

3.2 Vector Representation of Probability Distributions

Assume P_i is a probability distribution on the space $X = \{x_1, x_2, \dots, x_n\}$, where p_{ij} is the probability of the occurrence of x_j . Here, each $p_{ij} \in [0, 1]$ and $\sum_{j=1}^n p_{ij} = 1$. For our purposes in the following, we shall find it useful, at times, to represent a probability distribution as an n -dimensional vector $P_i = [p_{i1}, p_{i2}, \dots, p_{in}]$. Here the vector has the special properties that all its components lie in the unit interval and their sum is one. If P_i , for $i = 1$ to n , are a collection of probability distribution vectors then their weighed sum, $P = \sum_{i=1}^q w_i P_i$, is another vector whose components are $p_j = \sum_{i=1}^q w_i p_{ij}$. Furthermore, if the weights are standard weights, $w_i \in [0, 1]$ and $\sum_{i=1}^q w_i = 1$, then P is also a probability distribution vector.

Another operation on vectors is the dot or inner product, (see Bustince and Burillo [7]). If P_i and P_k are two probability vectors on the space X , then their dot product is $P_i \cdot P_k = \sum_{j=1}^n p_{ij} p_{kj}$. We emphasize that the dot product is a scalar value. Furthermore, in the case where P_i and P_k are probability distributions then $0 \leq P_i \cdot P_k \leq 1$. A special case of dot product is where P_i and P_k are the same, then $P_i \cdot P_k = \sum_{j=1}^n (p_{ij})^2$. For notational simplicity, at times when it causes no confusion, we shall simply use $P_i P_k$ for the dot product.

An important concept that is associated with this self dot product is the idea of the norm of the vector. In particular then norm

$$\|P_i\| = \sqrt{P_i P_i} = \left(\sum_{j=1}^n (p_{ij})^2 \right)^{1/2} \quad (3.1)$$

The norm is referred to as the Euclidean length of a vector. Because of the special properties of the probability distribution vector, $p_{ij} \in [0, 1]$ and $\sum_i p_{ij} = 1$, it can be easily shown that the maximal value of $\|P_i\|$ occurs when one $p_{ij} = 1$ and all other $p_{ij} = 0$. In this case, $\|P_i\| = 1$. Furthermore, in this case of a probability distribution vector the minimum value of $\|P_i\|$ occurs when all $p_{ij} = 1/n$ and this has the value

$$\|P_i\| = \left(\sum_{i=1}^n \left(\frac{1}{n} \right)^2 \right)^{1/2} = \left(\frac{1}{n} \right)^{1/2} = 1/\sqrt{n} \quad (3.2)$$

We note for the self dot product, $P_i P_i = \|P_i\|^2$ we have a maximal value of one and minimal value of $\frac{1}{n}$ when all $p_{ij} = \frac{1}{n}$. In the following, we shall benefit from the use of an illustration of the probability vector in the two-dimensional case as shown in Fig. 3.1.

If P_i and P_k are two probability vectors it is known [8] that the Cosine of the angle between them denoted θ_{ik} is expressed as

$$\cos \theta_{ik} = \frac{P_i P_k}{\|P_i\| \|P_k\|}$$

We note $\cos \theta_{ik}$ is the dot product of P_i and P_k divided by their respective norms. It is well known that if $\cos \theta_{ik} \in [0, 1]$, as is the case when P_i and P_k are probability distribution vectors, that $\cos \theta_{ik} \in [0, \pi/2]$. We further see that if $P_i = P_k$ then

$$\cos \theta_{ik} = \frac{P_i P_k}{\|P_i\| \|P_k\|} = \frac{P_i^2}{\|P_i\|^2} = \frac{P_i^2}{P_i^2} = 1 \quad (3.3)$$

Fig. 3.1 Angle between probabilistic vectors

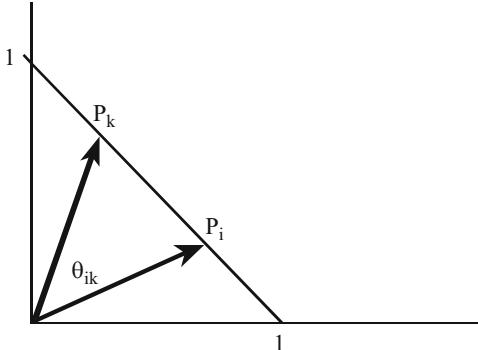
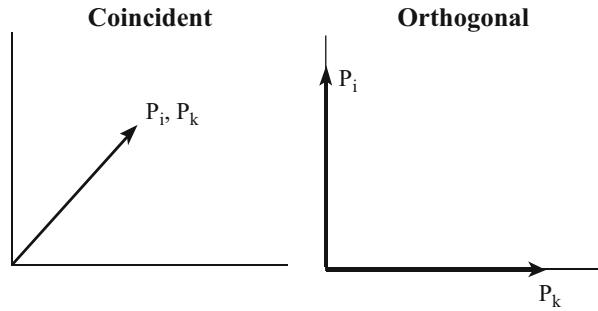


Fig. 3.2 Different relationships between probabilistic distributions



Thus if P_i and P_k are the same coincident, then $\cos\theta_{ik} = 1$. Furthermore, it is known in this case that $\theta_{ik} = 0$. At the other extreme is the case where P_i and P_k are orthogonal, $P_i \cdot P_k = \sum_{j=1}^n p_{ij} p_{kj} = 0$ here $\cos\theta_{ik} = \frac{P_i \cdot P_k}{\|P_i\| \|P_k\|} = 0$. We get in this case that $\theta_{ik} = \pi/2$. We note that in the case where P_i and P_k are orthogonal then $p_{ij} = 0$ when $p_{ik} \neq 0$ and $p_{ik} = 0$ when $p_{ij} \neq 0$. We illustrate these extremes of coincident and orthogonal distributions for the two dimensional case in Fig. 3.2.

We note in the n -dimensional case, a prototype example of orthogonality occurs when P_i has $P_{ij_1} = 1$ and P_k has $P_{kj_2} = 1$. Here they each completely support different outcomes. In [9] we suggested that $\cos\theta_{ik}$ can be used as measure of the degree of compatibility, Comp, between the two probability distributions, thus

$$\text{Comp}(P_i, P_k) = \frac{P_i \cdot P_k}{\|P_i\| \|P_k\|} \quad (3.4)$$

Here $\text{Comp}(P_i, P_k) \in [0, 1]$, and the closer to 1, the more compatible the probability distributions. Furthermore, $1 - \text{Comp}(P_i, P_k)$, denoted $\text{Conf}(P_i, P_k)$, can be seen as the degree of conflict between the two probability distributions. We note that if P_i and P_k are orthogonal then $\text{Comp}(P_i, P_k) = 0$ that $\text{Conf}(P_i, P_k) = 1$. On the other hand if P_i and P_k are 1-coincident, the same, then $\text{Comp}(P_i, P_k) = 1$ and $\text{Conf}(P_i, P_k) = 0$.

An interesting special case occurs when one of the distributions, P_i , has $P_{ij} = 1/n$ for all j . Here we previously noted $\|P_i\| = (1/n)^{1/2}$. Consider now $\text{Comp}(P_i, P_k)$ where P_i is this uniform probability distribution. Here $\text{Comp}(P_i, P_k) = \frac{P_i \cdot P_k}{\|P_i\| \|P_k\|}$. However in this case

$$P_i \cdot P_k = \sum_{j=1}^n p_{ij} p_{kj} = \frac{1}{n} \sum_{j=1}^n p_{kj} = \frac{1}{n} \quad (3.5)$$

and thus

$$\text{Comp}(P_i, P_k) = \frac{1/n}{\|P_k\| (1/n)^{1/2}} = \frac{(1/n)^{1/2}}{\|P_k\|} = \frac{1}{\sqrt{n}} \frac{1}{\|P_k\|} \quad (3.6)$$

Two special cases of P_k are worth commenting on. If P_k is a certain distribution, it has $p_{kj} = 1$ for one element, then $\|P_k\| = 1$ and $\text{Comp}(P_i, P_k) = \frac{1}{\sqrt{n}}$. If P_k is also a uniform probability distribution, all $p_{kj} = 1/n$; here then $\|P_k\| = \frac{1}{\sqrt{n}}$, and we get $\text{Comp}(P_i, P_k) = 1$.

3.3 Entropy, Certainty, and Information

In order to provide concrete measures of quality, we will consider in this section introducing entropy measures. This will allow assessments of the quality of our approaches to the fusion of probabilistic information. An important concept associated with a probability distribution on the space $X = \{x_1, x_2, \dots, x_n\}$ is the idea of entropy [10, 11]. The most common measure of entropy is the Shannon entropy. Here if P is a probability distribution on the space with p_j , the probability associated with x_j , then the Shannon entropy is $H(P) = -\sum_{j=1}^n p_j \ln p_j$. It is well known that the maximal entropy occurs when all $p_j = 1/n$ in which case $H(P) = \ln n$. The minimal entropy occurs for the case when one $p_j = 1$ and all other $p_j = 0$, in this case $H(P) = 0$. What is clear is that the entropy is measuring the uncertainty associated with the probability distribution, the more uncertainty the more entropy. The complimentary idea of entropy is certainty (or information). The smaller the entropy, the more information conveyed by a probability distribution. What should be clear is that for decision-making purposes we prefer distributions with smaller entropy as we have less uncertainty, more information.

While the Shannon entropy is the most well-known formulation of entropy, other formulations have been suggested [12–15], particularly in an attempt to overcome the complexity involved in working with the \ln . One successful alternative formulation is the Gini Entropy [16], which is much simpler to work with. The Gini entropy of the probability distribution P on X is

$$G(P) = 1 - \sum_{j=1}^n p_j^2 \quad (3.7)$$

The Gini entropy is coincidental with the Shannon entropy in the sense the $G(P)$ assumes its maximal value when all $p_j = 1/n$ and assumes its minimum value when one $p_j = 1$. While the Shannon entropy is based on the term $p_j \ln p_j$, the Gini entropy is based on the term p_j^2 .

In the following, we shall find it convenient to use the Gini entropy. Again we see that the bigger the value of $G(P)$, the more the uncertainty in the knowledge provided by the probability distribution. On the other hand the smaller $G(P)$ the more certainty in the knowledge provided by the probability distribution, it is common to refer to this as being more information. Thus to increase certainty or information we decrease the entropy.

What is very notable about the Gini formulation for entropy is that it very clearly indicates what is needed to decrease entropy, increase the certainty or information. In particular, by increasing $\sum_{j=1}^n p_j^2$ we decrease the entropy or uncertainty, we increase information. Furthermore, in using the vector representation of the probability distribution P the norm of P , $\|P\| = \left(\sum_{j=1}^n p_j^2\right)^{1/2}$. Thus we see that increasing the norm of the distribution serves to decrease the entropy, increase its information content. We note that $\|P\|$ takes its maximal value of 1 with $p_j = 1$ for some x_j , and all other $p_j = 0$. In addition, the min value of $\|P\|$ under the constraint that all $p_j \in [0, 1]$ and $\sum_{j=1}^n p_j = 1$ occurs when all $p_j = 1/n$.

Since our interest will be on obtaining probability distributions with more information, we shall focus on how the term $\sum_{j=1}^n p_j^2$ or its norm is affected by various operations. We shall refer to term $\|P\|^2$ as the NegEnt and semantically see it as a measure of information or certainty. Here again we note that $1/n \leq \|P\|^2 \leq 1$ and $1/\sqrt{n} \leq \|P\| \leq 1$.

Assume P and Q are two probability vectors on the space X and that the relation between these is

$$q_1 = p_1 - \alpha$$

$$q_2 = p_2 + \alpha$$

$$q_j = p_j, \text{ for } j = 3 \text{ to } n, \text{ for } \alpha \geq 0.$$

We see here

$$\sum_{j=1}^n q_j^2 = (p_1 - \alpha)^2 + (p_2 + \alpha)^2 + \sum_{j=3}^n p_j^2 \quad (3.8)$$

We note that since $(p_1 - \alpha)^2 = p_1^2 - 2\alpha p_1 + \alpha^2$ and $(p_2 + \alpha)^2 = p_2^2 + 2\alpha p_2 + \alpha^2$ we get

$$\begin{aligned} \sum_{j=1}^n q_j^2 - \sum_{j=1}^n p_j^2 &= 2\alpha(p_2 - p_1) + 2\alpha^2 \\ &= 2\alpha((p_2 - p_1) + \alpha) \end{aligned} \quad (3.9)$$

We observe that if $p_2 > p_1$ then $Q^2 > P^2$ and we have decreased the entropy, increased the certainty. Thus moving some amount of probability from an element with less probability to one of greater probability increases the norm, the certainty. We also observe that even if $p_1 = p_2$, then $Q^2 > P^2$. Thus if two elements have the same probability, moving some probability from one element to the other increases the norm and increases the certainty. If $p_2 < p_1$, the situation is more complex.

If $\alpha < |p_1 - p_2|$, then we decrease certainty, but if $\alpha > |p_1 - p_2|$, we increase the certainty; in this case we have moved enough probability to overcome the initial difference and essentially reversed the relationship.

We consider another related situation. Assume $P = [p_1, p_2, \dots, p_n]$ is a probability distribution such that two elements, p_1 and p_2 , share an amount of probability Δ . We ask what is the allocation of the probability Δ between p_1 and p_2 that results in the largest norm, the most certainty, and the smallest entropy. We see

$$P^2 = \sum_{j=1}^n p_j^2 = p_1^2 + p_2^2 + \sum_{j=3}^n p_j^2 \quad (3.10)$$

Assume we assign a to one of p_1 or p_2 and $\Delta - a$ to the other. In this case the

$$\begin{aligned} P^2 &= a^2 + (\Delta - a)^2 + \sum_{j=3}^n p_j^2 \\ &= a^2 + \Delta^2 - 2a\Delta + a^2 + \sum_{j=3}^n p_j^2 \\ &= 2a^2 + \Delta^2 - 2a\Delta + \sum_{j=3}^n p_j^2 \end{aligned} \quad (3.11)$$

Taking the derivative of P with respect to a , we get $\frac{\partial P^2}{\partial a} = 4a - 2\Delta$. Setting this to zero gives us $a = \Delta/2$. To find the maximum of P^2 , we must evaluate P^2 at $a = \Delta/2$ and the two end points $a = 0$ and $a = \Delta$.

a	P^2
0	$\Delta^2 + \sum_{j=3}^n p_j^2$
$\frac{\Delta}{2}$	$\frac{\Delta^2}{2} + \sum_{j=3}^n p_j^2$
Δ	$\Delta^2 + 0 + \sum_{j=3}^n p_j^2$

Thus we see that the maximal value of P^2 occurs when we give all the Δ to one of the components.

A concept closely related to entropy is the idea of Cross entropy [17, 18]. If P and Q are two probability distributions on the space $X = \{x_1, x_2, \dots, x_n\}$ then the standard definition of Cross entropy is $H(P, Q) = -\sum_{j=1}^n p_j \ln q_j$. We see that this definition is in the spirit of the Shannon measure of entropy. Here we can define a related measure in the spirit of the Gini index. In particular $G(P, Q) = 1 - \sum_{j=1}^n p_j q_j$. The larger the $G(P, Q)$, the more different the probability distributions; it is a kind of measure of difference between the distributions. We note here that using the vector representation we can express $\sum_{j=1}^n p_j q_j$ as the dot product PQ . Furthermore, we see that PQ is a kind of measure of relatedness of the two distributions, PQ is the negation of Gini measure of cross entropy. We note that as opposed to the Shannon-type measure, the Gini type measure is symmetric, $G(P, Q) = G(Q, P)$.

3.4 Information in Maximally Certain and Uncertain Distribution

In this section, we develop measures of conflict in the fusion of probability distributions. The concept of conflict is an important aspect of assessment of the quality of the fusion process. Consider now two probability distributions on X , $P = [p_1, p_2, \dots, p_n]$ and $Q = [q_1, q_2, \dots, q_n]$. Their linear aggregation $R = w_1P + w_2Q$ is a probability distribution when $w_1 + w_2 = 1$. Here, for each component r_j of R we have $r_j = (w_1p_j + w_2q_j)$. To calculate the information associated with R we calculate

$$\begin{aligned} \|R\|^2 &= \sum_{j=1}^n (w_1 p_j + w_2 q_j)^2 \\ &= \sum_{j=1}^n (w_1^2 p_j^2 + w_2^2 q_j^2 + 2w_1 w_2 p_j q_j) \end{aligned} \quad (3.12)$$

$$\|R\|^2 = w_1^2 \sum_{j=1}^n p_j^2 + w_2^2 \sum_{j=1}^n q_j^2 + 2w_1 w_2 \sum_{j=1}^n p_j q_j \quad (3.13)$$

$$\|R\|^2 = w_1^2 \|P\|^2 + w_2^2 \|Q\|^2 + 2w_1 w_2 P \cdot Q \quad (3.14)$$

where $P \cdot Q$ denotes the dot product of P and Q .

In passing we note that since $w_2 = 1 - w_1$, we have $w_1^2 + w_2^2 + 2(w_1 w_2) = 1$, and hence $\|R\|^2$ is a weighted average of components $\|P\|^2$, $\|Q\|^2$, and $P \cdot Q$. We also observe that since

$$\cos(P, Q) = \frac{P \cdot Q}{\|P\| \|Q\|} \quad (3.15)$$

we have

$$\|R\|^2 = w_1^2 \|P\|^2 + w_2^2 \|Q\|^2 + 2w_1 w_2 \|P\| \|Q\| \cos(P, Q) \quad (3.16)$$

where $\cos(P, Q)$ is the compatibility of P and Q . Thus the more compatible, the less conflict, the bigger the NegEnt, more information, contained in R . In the special case where $w_2 = w_1 = 1/2$, we have

$$\begin{aligned} \|R\|^2 &= \frac{1}{4} \|P\|^2 + \frac{1}{4} \|Q\|^2 + \frac{1}{2} P \cdot Q \\ &= \frac{1}{4} \|P\|^2 + \frac{1}{4} \|Q\|^2 + \frac{1}{2} \|P\| \|Q\| \cos(P, Q) \end{aligned} \quad (3.17)$$

Let us focus on this case where $w_1 = w_2 = 1/2$. If P and Q are two certain probability distributions, then $\|P\|$ and $\|Q\| = 1$, and there are two cases of interest. The first case is when they are completely compatible, $p_j = q_j = 1$ for some j . Here $\cos(P, Q) = 1$ and $\|R\|^2 = \frac{1}{4}(1) + \frac{1}{4}(1) + \frac{1}{2}(1) = 1$. On the other hand, if they

are completely conflicting, $p_j = 1$ and $q_k = 1$, then $\cos(P, Q) = 0$ and $\|R\|^2 = \frac{1}{4} + \frac{1}{4} + 0 = \frac{1}{2}$. Assume P and Q are both maximally uncertain distributions, they have $p_j = p_j = 1/n$ for all j . Here $\|P\|$ and $\|P\| = \|Q\| = (1/n)^{1/2}$ and since we have shown in this case that $\cos(P, Q) = 1$ then

$$\|R\|^2 = \frac{1}{4} \left(\frac{1}{n} \right) + \frac{1}{4} \left(\frac{1}{n} \right) + \frac{1}{2} \left(\frac{1}{n} \right)^{1/2} \left(\frac{1}{n} \right)^{1/2} 1 = \frac{1}{n}$$

If P is a completely certain distribution, $\|P\| = 1$, and Q is a completely uncertain distribution, $\|Q\| = (1/n)^{1/2}$, then $PQ = 1/n$ and $\|R\|^2 = \frac{1}{4}(1) + \frac{1}{4} \left(\frac{1}{n} \right) + \frac{1}{2} \left(\frac{1}{n} \right) = \frac{n+3}{4n}$. Let us now consider the case where we have t probability distributions P_1, \dots, P_t , where $P_i = [p_{i1}, \dots, p_{in}]$. Assume now $R = \frac{1}{t} \sum_{i=1}^t P_i$. Here then each component of R , $r_j = \frac{1}{t} \sum_{i=1}^t p_{ij}$. In this case

$$\begin{aligned} \|R\|^2 &= \sum_{j=1}^n r_j^2 = \sum_{j=1}^n \left(\frac{1}{t^2} \right) (p_{1j} + p_{2j} + \dots + p_{tj})^2 \\ &= \sum_{i=1}^t \left(\frac{1}{t^2} \right) (\|P_i\|)^2 + \sum_{i=1}^t \sum_{k=1, k \neq i}^t \frac{1}{t^2} P_i \cdot P_k \\ &= \sum_{i=1}^t \left(\frac{1}{t^2} \right) (\|P_i\|)^2 + \frac{2}{t^2} \sum_{i=1}^{t-1} \sum_{k=i+1}^t P_i \cdot P_k \end{aligned} \quad (3.18)$$

To understand how we obtained this last formula from $\sum_{j=1}^n \left(\frac{1}{t^2} \right) (p_{1j} + p_{2j} + \dots + p_{tj})^2$ we illustrate the situation for $t=4$. Consider the term $(p_{1j} + p_{2j} + p_{3j} + p_{4j})^2$. The value of the squared sum can be best viewed in using the following matrix:

	p_{1j}	p_{2j}	p_{3j}	p_{4j}
p_{1j}	$p_{1j}p_{1j}$	$p_{1j}p_{2j}$	$p_{1j}p_{3j}$	$p_{1j}p_{4j}$
	$p_{2j}p_{1j}$	$p_{2j}p_{2j}$	$p_{2j}p_{3j}$	$p_{2j}p_{4j}$
	$p_{3j}p_{1j}$	$p_{3j}p_{2j}$	$p_{3j}p_{3j}$	$p_{3j}p_{4j}$
	$p_{4j}p_{1j}$	$p_{4j}p_{2j}$	$p_{4j}p_{3j}$	$p_{4j}p_{4j}$
p_{2j}				
p_{3j}				
p_{4j}				

The value of the squared sum is equal to the sum of all the $4 \times 4 = 16$ terms in the matrix. We can consider the matrix as consisting of three parts, the main diagonal and the upper and lower triangles. The sum of the term on the main diagonal is $p_{1j}p_{1j} + p_{2j}p_{2j} + p_{3j}p_{3j} + p_{4j}p_{4j}$. If we calculate the sum of these over all $j = 1$ to

n we get

$$\sum_{j=1}^n p_{1j}^2 + \sum_{j=1}^n p_{2j}^2 + \sum_{j=1}^n p_{3j}^2 + \sum_{j=1}^n p_{4j}^2 = \sum_{i=1}^4 \|P_i\|^2 \quad (3.19)$$

Consider now the upper and lower triangle. First note that they consist of exactly the same elements. So to find the sum of elements in the upper and lower triangle, we need just calculate the sum of one of them and multiply by two. Consider the upper triangle

$$p_{1j}p_{2j} + p_{1j}p_{3j} + p_{1j}p_{4j}$$

$$p_{2j}p_{3j} + p_{2j}p_{4j}$$

$$p_{3j}p_{4j}$$

With a little thought we see that we can express this sum as $\sum_{i=1}^3 (\sum_{k=i+1}^4 p_{ij} p_{kj})$. If we now sum these over $j = 1$ to n we get $\sum_{j=1}^n (\sum_{i=1}^3 (\sum_{k=i+1}^4 p_{ij} p_{kj})) = \sum_{i=1}^3 (\sum_{k=i+1}^4 (\sum_{j=1}^n p_{ij} p_{kj}))$. However, we observe that $\sum_{j=1}^n p_{ij} p_{kj}$ to the dot product $P_i \cdot P_k$. Thus the sum of the elements in the upper triangle is $\sum_{i=1}^3 (\sum_{k=i+1}^4 P_i \cdot P_k)$. Doubling this we get the sum of upper and lower triangles, $2\sum_{i=1}^3 (\sum_{k=i+1}^4 P_i \cdot P_k)$. Thus we get

$$\|R\|^2 = \left(\frac{1}{t^2}\right) \sum_{i=1}^t (\|P_i\|)^2 + \frac{2}{t^2} \sum_{i=1}^{t-1} \sum_{k=i+1}^t P_i \cdot P_k \quad (3.20)$$

and now consider the calculation of $\|R\|^2$ for some special cases. First we consider the case where all t probability distributions are certain distributions but may be conflicting in that they may be focused on different x_j . In this case $\|P_i\|^2 = 1$ for all i . We now partition the probability distributions into groups of agreement. For simplicity we let g_j be the number of probability distributions that are focused on x_j , these distributions are compatible. We note that $\sum_{j=1}^n g_j = t$. We observe that if a pair of certain distributions are in agreement then $P_i \cdot P_k = 1$; otherwise $P_i \cdot P_k = 0$. We see that $\frac{2}{t^2} \sum_{i=1}^{t-1} \sum_{k=i+1}^t P_i \cdot P_k$ can be seen as being equal to the $\frac{2}{t^2}$ times number of pairs of distribution that are compatible. If g_j are focused on x_j , then there are $\frac{g_j!}{2!(g_j-2)!} = \frac{g_j(g_j-1)}{2} = \frac{g_j^2-g_j}{2}$ pairs of elements in agreement. Thus in this case of all certain probability distributions, we have

$$\|R\|^2 = \left(\frac{1}{t^2}\right)t + \frac{2}{t^2} \sum_{j=1}^n \frac{g_j(g_j - 1)}{2} = \frac{1}{t} + \frac{1}{t^2} \sum_{j=1}^n g_j(g_j - 1) \quad (3.21)$$

If all the elements agree on the same value, x_1 , then $g_1 = t$ and all other $g_j = 0$, and we get

$$\|R\|^2 = \frac{1}{t} + \left(\frac{1}{t^2}\right)t(t-1) = \frac{1}{t} + \frac{t^2-t}{t^2} = \frac{1}{t} + 1 - \frac{1}{t} = 1$$

Assume half the distributions agree on one value and the other half agree on a second value. Here we have $g_1 = t/2$ and $g_2 = t/2$ (we assume t even for simplicity) then we get

$$\|R\|^2 = \frac{1}{t} + \left(\frac{1}{t^2}\right)(2)\left(\frac{t}{2}\right)\left(\frac{t}{2}-1\right) = \frac{1}{t} + \frac{1}{t}\left(\frac{t}{2}-1\right) = \frac{1}{t}\left(1 + \frac{t}{2} - 1\right) = 0.5$$

If one third of the distributions agree on different values, $g_1 = g_2 = g_3 = t/3$ then

$$\|R\|^2 = \frac{1}{t} + \left(\frac{1}{t^2}\right)(3)\left(\frac{t}{3}\right)\left(\frac{t}{3}-1\right) = \frac{1}{t} + \frac{1}{t}\left(\frac{t}{3}-1\right) = \frac{1}{t}\left(1 + \frac{t}{3} - 1\right) = \frac{1}{3}$$

The value of $\|R\|^2$ for complex allocations of the g_j can be calculated using our formula $\|R\|^2 = \frac{1}{t} + \frac{1}{t^2} \sum_{j=1}^n g_j(g_j - 1)$.

Let us return to our formula $\|R\|^2 = \sum_{i=1}^t \left(\frac{1}{t^2}\right)(\|P_i\|)^2 + \frac{2}{t^2} \sum_{i=1}^{t-1} \sum_{k=i+1}^t P_i \cdot P_k$ and consider the situation where we have two classes of probability distributions. One being as in the preceding a certainty distribution, one of its components is one. The other is a pure uncertainty distribution; here all elements are $1/n$. First we note for P_i that has certainty; then $(\|P_i\|)^2 = 1$. On the other hand, for any P_i that is pure uncertainty, we have shown that $(\|P_i\|)^2 = 1/n$. Consider now the dot products. As in the preceding, if both P_i and P_k are certainties, then $P_i \cdot P_k = 1$ if they agree on the certainty element and $P_i \cdot P_k = 0$ if they disagree. Consider the situation when one of $P_i P_k$ is pure uncertainty, for example, P_i . If the other distribution P_k is a certainty, we have $P_i P_k = \sum_{j=1}^n p_{ij} p_{kj} = 1 \frac{1}{n} = \frac{1}{n}$. If the other distribution is all pure uncertainty, we get $P_i P_k = \sum_{j=1}^n p_{ij} p_{kj} = \frac{n}{n^2} = \frac{1}{n}$. Thus we see independent of the second probability distribution if one of P_i or P_k is pure uncertainty, then $P_i P_k = 1/n$.

Let us now consider the calculation of $\|R\|^2$ in mixed case. Assume t_1 of the distributions are pure certainty and t_2 are pure uncertainty. Furthermore, assume for $j = 1$ to n that g_j of the pure certainty distributions agree on the same value, x_j , here $\sum_{j=1}^n g_j = t_1$. In this case we get

$$\|R\|^2 = \frac{1}{t^2} \left(t_1 + \frac{t - t_1}{n} \right) + \left(\frac{2}{t^2}\right) \left(\frac{1}{n}\right) S1 + \left(\frac{2}{t^2}\right) S2 \quad (3.22)$$

where $S1 = \#$ of pairs containing at least a pure uncertainty distribution and $S2 = (\# \text{ of pairs of two pure certainty distributions in agreement})$. We have already calculated the last term in the preceding, that is

$$\frac{2}{t^2} S2 = \frac{1}{t^2} \sum_{j=1}^n g_j (g_j - 1) = \frac{1}{t^2} \sum_{j=1}^n g_j^2 - g_j \quad (3.23)$$

We must now calculate the number containing a pure uncertainty, $S1$. First we note given t distributions there are $(t)(t - 1)/2$ possible pairs. Given that there are t_1 distributions with pure certainty, then there are $(t)(t - 1)/2$ pairs of elements consisting of two pure certain distributions. From this we can conclude that $S1$, # of pairs containing at least one pure uncertainty distribution, is

$$S1 = \frac{(t)(t - 1)}{2} - \frac{(t_1)(t_1 - 1)}{2} = \frac{(t - t_1)(t + t_1 - 1)}{2} \quad (3.24)$$

$$\begin{aligned} \|R\|^2 &= \frac{1}{t^2} \left[t_1 + \frac{t-t_1}{n} + \frac{1}{n} ((t - t_1)(t + t_1 - 1)) + \sum_{j=1}^n g_j (g_j - 1) \right] \\ &= \frac{1}{t^2} \left[t_1 + \sum_{j=1}^n g_j (g_j - 1) + \frac{1}{n} ((t - t_1)(t + t_1)) \right] \\ &= \frac{1}{t^2} \left[t_1 + \sum_{j=1}^n g_j (g_j - 1) + \frac{1}{n} (t^2 - t_1^2) \right] \end{aligned} \quad (3.25)$$

We note in the special case where all the pure certain distributions agree, $g_1 = t_1$ and all other $g_j = 0$ we get

$$\begin{aligned} \|R\|^2 &= \frac{1}{t^2} \left[t_1 + t_1(t - 1) + \frac{1}{n} (t^2 - t_1^2) \right] \\ &= \frac{1}{t^2} \left[t_1^2 + \frac{1}{n} (t^2 - t_1^2) \right] \end{aligned} \quad (3.26)$$

3.5 Fusion of Probability Distributions

We now turn to our major interest, the fusion of multi-source probabilistic information. Here we will use the quality issues related to credibility measures to guide the multi-source fusion process. Assume V is a variable that takes its value in the space $X = \{x_1, \dots, x_n\}$. In the following, we let P_i be a probability distribution on X indicating the information provided by source i regarding the value of V . Here we let each $P_i = [p_{i1}, \dots, p_{in}]$. If we have t probability distributions, P_i for $i = 1$

to t , then the distribution basic uniform fusion of these is a probability distribution $R = \frac{1}{t} \sum_{i=1}^t P_i$. Each component of R is $r_j = \frac{1}{t} \sum_{i=1}^t p_{ij}$. We have previously shown the associated NegEnt is

$$\|R\|^2 = \frac{1}{t^2} \left[\sum_{i=1}^t \|P_i\|^2 + 2 \sum_{i=1}^{t-1} \sum_{k=i+1}^t P_i P_k \right] \quad (3.27)$$

The larger $\|R\|^2$ the more information provided by the fusion.

We note $\sum_{i=1}^{t-1} \sum_{k=i+1}^t P_i P_k$ consists of $\frac{(t)(t-1)}{2}$ terms. Thus we see total number of terms being combined is $t + \frac{2(t)(t-1)}{2} = t^2$. Thus this is a simple weighted average. Since each term is contained in $[0, 1]$ and since this is a simple weighted average, then $\|R\|^2 \in [0, 1]$. For this aggregation we can calculate the average conflict between the components. We first recall the degree of conflict between P_i and P_k is

$$\text{Conf}(P_i, P_k) = 1 - \cos \theta_{ik} = 1 - \frac{P_i P_k}{\|P_i\| \|P_k\|} \quad (3.28)$$

Given that there are t distributions being combined, then there are $\frac{(t)(t-1)}{2}$ distinct pairs of distributions. Thus the average conflict in this fusion is

$$\begin{aligned} \text{AveConf}(P_i, P_k) &= \frac{2}{(t)(t-1)} \sum_{i=1}^{t-1} \sum_{k=i+1}^t \left(1 - \frac{P_i P_k}{\|P_i\| \|P_k\|} \right) \\ &= 1 - \frac{2}{(t)(t-1)} \sum_{i=1}^{t-1} \sum_{k=i+1}^t \left(\frac{P_i P_k}{\|P_i\| \|P_k\|} \right) \end{aligned} \quad (3.29)$$

We note here since each $\text{Conf}(P_i, P_k) \in [0, 1]$, then $\text{AveConf}(P) \in [0, 1]$.

Let P_i for $i = 1$ to t , be a collection of probability distributions provided by multiple sources. Assume our purpose in combining these is to obtain a fused estimate for the value of V that gives us the most information about V . That is, we want our fused value to have a high NegEnt value, $\|R\|^2$. We can make some observations about $\|R\|^2$. Pairs of probability distributions that are nonconflicting, $P_i P_k$ large, tend to increase the NegEnt, the information, supplied by the aggregation. On the other hand, those pairs with small compatibility, $P_i P_k$ small, can tend to decrease the NegEnt. This reduction results from the fact that while pairs with small $P_i P_k$ may add a little to the sum of the $P_i P_k$, they affect the value $\frac{1}{t^2}$. We note that $t^2 = t+t(t-1)$, the number of probability distributions plus the number of pairs. Thus while a conflicting pair does not much affect the sum $\sum_{i=1}^{t-1} \sum_{k=i+1}^t P_i P_k$, it can reduce the $\|R\|^2$ because it is counted in the t^2 .

It appears that one approach to obtaining fused values that have high NegEnt is to only fuse the probability distributions that have a high compatibility. However, by just looking at any fusion consisting of a subset of the probability distributions we are clearly losing credibility, persuasiveness, in the fusion. In order to take this into account, we shall introduce a set measure that indicates the credibility of a

fusion based on a simple weighted average of a subset of probability distributions. If $Z = \{P_1, \dots, P_t\}$, the set of available probability distributions, then we define the set measure $\text{Cred} : 2^Z \rightarrow [0, 1]$ such that for any subset B of Z , $\text{Cred}(B)$ indicates the credibility of a fusion based on only the distributions in B . We see that natural properties to require of such a set measure are the following:

1. $\text{Cred}(\emptyset) = 0$,
2. $\text{Cred}(Z) = 1$ and
3. If $A \subseteq B$ then $\text{Cred}(A) \leq \text{Cred}(B)$ $A \subseteq B$.

Actually, more precisely, the credibility measure should be over the space of sources. However, since there is a one to one correspondence between a source and its provided probability distribution for simplicity of notation we shall continue to refer to this credibility as being over the set of probability distributions.

Let us look at some notable examples of credibility functions. One fundamental example is the case where $\text{Cred}(Z) = 1$ and $\text{Cred}(B) = 0$ for $B \neq Z$. Here the only fusion that has any credibility is the one using all of the supplied probability distributions. Another type of credibility function can be based on the requirement that “at least λ %” of the probability distributions are included in the fusion. Here $\text{Cred}(B) = 1$ if $\frac{\text{Card}(B)}{t} \geq \lambda$ and $\text{Cred}(B) = 0$ otherwise. Closely related is a credibility function that requires that “most” of the probability distributions are included in the fusion. Here we can represent “most” as a fuzzy subset M on the unit interval [19]. In this case, for any subset B of Z , we can obtain $\text{Cred}(B) = M(\text{Cred}(B)/Z)$ the membership grade of $\text{Cred}(B)/Z$ in the fuzzy subset M .

The preceding examples of credibility functions have not taken into account any distinction between the sources providing the probability distributions. Another type of credibility function can be obtained if we associate with each P_i a value $\alpha_i \in [0, 1]$ so that $\sum_{i=1}^t \alpha_i = 1$. Here α_i can be seen as some indication of the importance of P_i . Using these *importances* we can obtain $\text{Cred}(B) = \sum_{P_i \in B} \alpha_i$.

Another class of credibility functions can be obtained as follows. Let F_i for $i = 1$ to q be a collection of subsets of Z . We note that formally this collection need not satisfy any special requirements, that is, they don't have to be disjointed or cover the whole space Z . Further, we associate with each F_i a value $\alpha_i \in [0, 1]$ so that $\sum_{i=1}^q \alpha_i = 1$. We now can use this collection of subsets of Z to construct various kinds of credibility functions. Let $\text{Poss}(F_i/B) = 1$ if $F_i \cap B \neq \emptyset$ and $\text{Poss}(F_i/B) = 0$ if $F_i \cap B = \emptyset$. Using this we can obtain $\text{Cred}(B) = \sum_{i=1}^q \alpha_i \text{Poss}(F_i/B)$. Here we have associated with the collection of sources q categories and if a fusion contains an element from category F_i it get α_i points. Related to this is another credibility function using $\text{Cert}(F_i/B)$, which is defined so that $\text{Cert}(F_i/B) = 1$ if $F_i \subseteq B$ and $\text{Cert}(F_i/B) = 0$ if $F_i \not\subseteq B$. Using this we can obtain $\text{Cred}(B) = \sum_{i=1}^q \alpha_i \text{Cert}(F_i/B)$. Here again we have q categories and a fusion using B distributions; however, we get α_i credibility points if the fusion contains *all* the elements in a category F_i .

Many different types of functions can be constructed to reflect various complex relations regarding the credibility of subsets of probability distribution. We note

that we can use the Takagi-Sugeno [20] approach to fuzzy rule-based modeling to build credibility functions. As an example of this fusion approach, we consider a scenario of probability distributions of spatial locations in a search and rescue mission. Information from differing sources is common, but it is very important to make good decisions as to how to fuse such information for most likely locations due to both the need for timely rescue and associated search costs.

Specifically we assume there are three distributions, P_1 , P_2 and P_3 that have source information relative to 4 potential spatial locations (x_1, x_2, x_3, x_4) for the search. For example, the first two distributions might have been obtained from UAVs or a search plane. The third which differs somewhat was obtained from local officials who, from their previous rescue experiences, provide what they believe are the probabilities for the four search locations.

Now we examine these distributions determining their conflicts and the information (NegEnt) provided in the fusions. The distributions are:

$$P_1 : (.5, .2, .2, .1); P_2 : (.4, .3, .2, .1); P_3 : (.1, .2, .1, .6)$$

Then

$$\|P_1\| = \left(\sqrt{.34} \right) = .583; \|P_1\|^2 = .34$$

$$\|P_2\| = \left(\sqrt{.3} \right) = .547; \|P_2\|^2 = .3$$

$$\|P_3\| = \left(\sqrt{.42} \right) = .648; \|P_3\|^2 = .42$$

$$P_1 \cdot P_2 = .31; P_1 \cdot P_3 = .17; P_2 \cdot P_3 = .18$$

$$\text{So now we can calculate the conflict } \text{Conf}(P_i, P_j) = 1 - \frac{P_i \cdot P_j}{\|P_i\| * \|P_j\|}$$

$$\text{Conf}(P_1, P_2) = 1 - \frac{.31}{.583 * .547} = 1 - \frac{.31}{.318} = 1 - .975 = .025$$

$$\text{Conf}(P_1, P_3) = 1 - \frac{.17}{.583 * .648} = 1 - \frac{.17}{.378} = 1 - .450 = .550$$

$$\text{Conf}(P_2, P_3) = 1 - \frac{.18}{.547 * .648} = 1 - \frac{.18}{.354} = 1 - .508 = .452$$

These conflicts are compatible with our intuitions by examining the differences in the location probabilities particularly with respect to P_3 . Next we can examine the possible pairwise fusions of the distributions denoted by a distribution $R(i, j)$:

$$R(1, 2) = (.45, .25, .2, .1); \|R(1, 2)\|^2 = .316$$

$$R(1, 3) = (.3, .2, .15, .35); \|R(1, 3)\|^2 = .275$$

$$R(2, 3) = (.25, .25, .15, .35); \|R(2, 3)\|^2 = .27$$

So the distribution $R(1, 2)$ fusing the distributions with the least conflict provides the most information. The range for NegEnt values of fusions depends on the characteristics of the particular distributions. If we would fuse two identical distributions, the fusion information content would be the same as the original distribution. So a better comparison for fused distributions is the ratio of the NegEnt to the average of the NegEnt of the original distributions. If we do this, we have

$$\|R(1, 2)\|^2 = \frac{.316}{.32} = .988$$

$$\|R(1, 3)\|^2 = \frac{.275}{.38} = .724$$

$$\|R(2, 3)\|^2 = \frac{.27}{.36} = .75$$

This scaling reflects better the comparisons between the fused distributions. Also it reflects more appropriately the value for $R(2, 3)$ which has less conflict than $R(1, 3)$. Finally we can fuse all three distributions, $R(1, 2, 3)$

$$R(1, 2, 3) = (.333, .233, .166, .266); \|R(1, 2, 3)\|^2 = .264$$

and the ratio is

$$\|R(1, 2, 3)\|^2 = \frac{.264}{.353} = .748$$

So the ratio indicates that fusion of all three is roughly the same as $R(2, 3)$ but improved over $R(1, 3)$.

The distributions and their information content provide an initial basis for decisions on selections of search areas, but in the next section, 6, we discuss how to use credibility along with NegEnt for a final fusion. So in that section, we will revisit the above example using credibility.

3.6 On Weighted Average Fusion

Let us now look further at the multi-source fusion problem. Assume V is a variable taking its value in the space $X = \{x_1, \dots, x_n\}$ and we have a collection $Z = \{P_1, \dots, P_t\}$ of probability distribution type information about the value of V . In addition, we assume a credibility function Cred is providing information about the credibility of fused values using different subsets of Z .

Given a subset B of probability distributions from Z , we can calculate the associated fused value, P_B . In particular, if $|B|$ is the number of distributions in B , then $P_B = \frac{1}{|B|} \sum_{P_i \in B} P_i$. Our Cred measure now provides the credibility associated with the fusion based on the subset B , $\text{Cred}(B)$. Here $\text{Cred}(B) \in [0, 1]$. In addition, we can calculate the NegEnt value of the information associated with the fused value P_B

$$\|P_B\|^2 = \frac{1}{|B|^2} \left[\sum_{P_i \in B} \|P_i\|^2 + 2 \sum_{i=1}^{t-1} \sum_{\substack{k=i+1 \\ P_i \in B \\ P_k \in B}}^t P_i P_k \right] \quad (3.30)$$

Here also $\|P_B\|^2 \in [0, 1]$.

Our objective now is to obtain a fused value P_B that has both high values for $\text{Cred}(B)$ and $\|P_B\|^2$. Since the number of possible subsets of Z is not prohibitive, we can do an exhaustive search to find the best fusion. We first calculate for each subset B of Z its fused value P_B and its associated credibility $\text{Cred}(B)$ and $\text{NegEnt}(B)$, $\|P_B\|^2$. We must now compare the P_B based on their $\text{Cred}(B)$ and $\|P_B\|^2$. Here we are here faced with a multi-criteria decision problem.

We now introduce the idea of dominance. We say that a fusion based on subset B_1 dominates B_2 if $\text{Cred}(B_1) \geq \text{Cred}(B_2)$ and $\text{NegEnt}(B_1) \geq \text{NegEnt}(B_2)$ and at least one of these is a strictly greater than. We now remove all subsets that are dominated by some other subset. Our preferred fusion will be one of the nondominated fusions. We point out here the collection of nondominant fusion subsets have at least have one member with credibility equal one. This is true because the credibility of the fusion based on the whole set Z has credibility one and any subset B that dominates Z must have credibility one.

At this point we have a collection of non-dominated fusions where each fusion is determined from a subset of space Z ; we denote these as B_1, \dots, B_r . For each subset B_j , we have its associated fusion P_{B_j} and its NegEnt value, $\|P_{B_j}\|^2$, and its credibility, $\text{Cred}(B_j)$. Our objective is to use the information about the NegEnt value and credibility to select among these possible fusions, the P_{B_j} .

If there is only one non-dominated fusion, then this is our selected fusion. If there is more than one non-dominated fusion our procedure for adjudicating between these must involve the introduction of some subjective preference type information

from the responsible decision-maker. Here rather than dictate a best way, we shall suggest some ways a decision-maker can choose among the non-dominated fusions. We note other possibilities exist.

Here we consider how to choose the final distribution based both on the information measure and credibility in our search and rescue example. We will use two of the possible credibility functions from the previous section, C_1 based on the number of distributions in a subset and C_2 based on the idea of “most” distributions included in a subset.

The collection of relevant probability distributions is $\{P_1, P_2, P_3\}$. So there are seven subsets of Z to consider: $B_1 = \{P_1\}$, $B_2 = \{P_2\}$, $B_3 = \{P_3\}$, $B_4 = \{P_1, P_2\}$, $B_5 = \{P_1, P_3\}$, $B_6 = \{P_2, P_3\}$, and $B_7 = \{P_1, P_2, P_3\}$. For the first credibility function, C_1 , we use a threshold of at least two distributions included in a subset so we have

$$C_1(B_1) = C_1(B_2) = C_1(B_3) = 0$$

$$C_1(B_4) = C_1(B_5) = C_1(B_6) = C_1(B_7) = 1$$

Then for the second we use the idea of “most” as the fuzzy value M

$$M = \begin{cases} 0 & |B| = 1 \\ .7 & |B| = 2 \\ 1.0 & |B| = 3 \end{cases}$$

and so using M for the second credibility we have

$$C_2(B_1) = C_2(B_2) = C_2(B_3) = 0$$

$$C_2(B_4) = C_2(B_5) = C_2(B_6) = 0.7$$

$$C_2(B_7) = 1$$

Now to make our selection of which distribution to use we need to use the concept of dominance related to the credibility and information. We can express this again as the predicate $\text{Dom}(B_i, B_j)$:

$$\text{Dom}(B_i, B_j) = \left[(\text{Cred}(B_i) \geq \text{Cred}(B_j)) \wedge \left(\|B_i\|^2 \geq \|B_j\|^2 \right) \wedge \text{Cond}(>) \right] \quad (3.31)$$

where $\text{Cond}(>)$ is true only if at least one of the “ \geq ” is “ $>$.” If $\text{Dom}(B_i, B_j)$ is true, then B_i dominates B_j and so B_j can be removed from consideration. In order to

Table 3.1 A Summary of possible dominance combinations

Subset	NegEnt	C_1	C_2
B_1	.34	0	0
B_2	.3	0	0
B_3	.42	0	0
B_4	.316	1	0.7
B_5	.275	1	0.7
B_6	.27	1	0.7
B_7	.264	1	1.0

help in working through the possible dominance combinations for our example, we provide the summarizing Table 3.1.

We can assess Dom for each of the two credibility functions above to obtain the collection, ND , of non-dominated subsets. For the first credibility function we see that B_3 dominates B_1 and B_2 as all three have $C_1 = 0$ and B_3 has a strictly greater NegEnt. Then $B_1, B_2 \notin ND$. Likewise B_4 dominates B_5, B_6 and B_7 so $B_5, B_6, B_7 \notin ND$. Finally we have to consider the dominance relationship of B_3 and B_4 . Since $(C_1(B_3) = 0) < (C_1(B_4) = 1)$, B_3 cannot dominate B_4 . Also $\|B_4\|^2 < \|B_3\|^2$, so B_4 cannot dominate B_3 , and we have two non-dominated subsets $ND = \{B_3, B_4\}$.

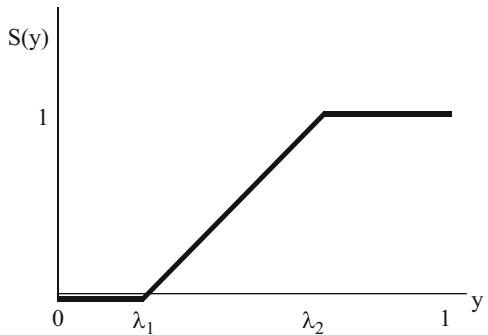
For this case of more than one non-dominating fusion we must utilize some subjective considerations to select the most appropriate one. A possible consideration is that the distribution P_3 obtained from the local officials might be thought of as more reliable than the fusion, B_4 , of the less effective sensor obtained probabilities P_1 and P_2 .

Finally we examine dominance relations for the second credibility function C_2 . As for C_1 , B_3 dominates B_1 and B_2 and B_4 dominates B_5 and B_6 , but not B_7 since $[C_2(B_4) = .7] < [C_2(B_7) = 1]$. So again the dominated subsets are eliminated: $B_1, B_2, B_4, B_5, B_6 \notin ND$. Now B_3 cannot dominate B_4 or B_7 due to credibility since $[C_2(B_3) = 0] < [C_2(B_4) = .7]$ and $[C_2(B_7) = 1.0]$. Also B_4 cannot dominate B_3 since $\|B_4\|^2 < \|B_3\|^2$ and B_7 cannot dominate either B_3 or B_4 since its NegEnt is less than either one. So for the credibility C_2 , $ND = \{B_3, B_4, B_7\}$ and again we must discuss how to select one of these. The considerations discussed for C_1 can apply again to B_3 and B_4 , but now B_7 has to be considered also. A criteria here might dictate that since B_7 has the highest credibility due to being the fusion of all three probabilities it should be chosen.

One observation we would like to make is that there appears to be some asymmetry between the two criteria, credibility and information content. In particular with regard to information content it would appear that we would like to obtain as much of this as possible, while with regard to credibility it may be that we want to have at least a certain level of credibility. Thus if we don't have some minimal degree of credibility any amount of information may not compensate. There appears a kind of priority here [21–23].

One approach here is for the responsible decision-maker to provide some minimal level of credibility, λ , and select the fusion with the maximal NegEnt value having at least this level of credibility. In anticipation of generalizing this idea we

Fig. 3.3 Degree of satisfaction for credibility values



could look at this approach a little more formally. Let $S : [0, 1] \rightarrow [0, 1]$ be a function such that for any degree of credibility y , $S(y)$ indicates the decision-maker degree of satisfaction with this level y of credibility. Here S should be monotonic and have $S(1) = 1$ and $S(0) = 0$. Using this function we can associate with the fusion based on a subset B of Z a value, $\text{Qual}(B) = S(\text{Cred}(B))\|P_B\|^2$. Here then we select as our fusion value the P_B so that its, $\text{Qual}(B)$ is maximal.

If we consider the function S so that $S(y) = 0$ for $y < \lambda$ and $S(y) = 1$ for $y \geq \lambda$, we get the previous method for selecting the preferred fused value. In this case

$$\text{Qual}(B) = \|P_B\|^2 \text{forCred}(B) \geq \lambda$$

$$\text{Qual}(B) = 0 \text{ forCred}(B) < \lambda$$

Once having introduced this type of S function, we can consider other formulas for S . One formula is shown in Fig. 3.3. Here if $\lambda_1 = \lambda_2 = \lambda$, then we get the previous case. If $\lambda_1 = 0$ and $\lambda_2 = 1$, then we get $\text{Qual}(B) = \|P_B\|^2 \text{Cred}(B)$, the product of credibility and information content.

It is also possible to use Zadeh's idea of computing with words [24] to determine the form of S . Here we can linguistically specify some requirements for our requested degree of satisfaction by the credibility. We can represent this as a fuzzy set of the unit interval and then represent S as this fuzzy set.

Another approach for obtaining the quality of a fusion is to use the idea of prioritized aggregation introduced by Yager [23]. Here the score associated with a fusion is a weighted sum of its credibility and NegEnt; however, here the weight associated with the criteria having the lower priority depends on the degree of satisfaction of higher-order criteria. As a result of this relationship, lack of satisfaction to the higher priority criteria can't be compensated for by very high satisfaction to the lower priority criteria. More formally, the score of the fusion based on the subset B is determined as $\text{Score}(B) = w_1 \text{Cred}(B) + w_2 \|P_B\|^2$. However, here $w_2/w_1 = \text{Cred}(B)/1$. Normalizing the weights so that $w_1 + w_2 = 1$ we get

$$\begin{aligned}\text{Score}(B) &= \frac{1}{1+\text{Cred}(B)} \text{Cred}(B) + \frac{\text{Cred}(B)}{1+\text{Cred}(B)} \|P_B\|^2 \\ &= \frac{\text{Cred}(B)}{1+\text{Cred}(B)} (1 + \|P_B\|^2)\end{aligned}\quad (3.32)$$

Here we see that if

$$\begin{array}{ll}\text{Cred}(B) & \text{Score}(B) \\ 0 & 0 \\ \frac{1}{2} & \frac{1}{3}(1 + \|P_B\|^2) \\ 1 & \frac{1}{2}(1 + \|P_B\|^2)\end{array}$$

More generally if $\text{Cred}(B) = n/d$ then $\text{Score}(B) = \frac{n}{d+n}(1 + \|P_B\|^2)$. Thus if $\text{Cred}(B) = \frac{1}{4}$ then $\text{Score}(B) = \frac{1}{5}(1 + \|P_B\|^2)$ and if $\text{Cred}(B) = \frac{3}{4}$ then $\text{Score}(B) = \frac{3}{7}(1 + \|P_B\|^2)$. If the credibility is 90% then $\frac{n}{d} = \frac{9}{10}$ and $\text{Score}(B) = \frac{9}{19}(1 + \|P_B\|^2)$.

3.7 Unequally Weighted Fusions

In the preceding, we assumed that starting with the set $Z = \{P_1, \dots, P_t\}$ we selected a subset B of Z and found our fusion by taking an equally weighted aggregation of the probability distributions in B . Here we suggest a more general approach based on a nonuniform weighted average of the elements in Z . So here we let $W = [w_1, \dots, w_t]$ be a set of weights where $w_i \in [0, 1]$ and $\sum_{i=1}^t w_i = 1$, and we obtain a fusion $P_W = \sum_{i=1}^t w_i P_i$. Here P_W is a probability distribution whose j^{th} component $P_W(j) = \sum_{i=1}^t w_i p_{ij}$.

We note here that the type of exhaustive search through all possible fusions to find the best fusion is not feasible here as there are too many possibilities. The preceding fusions obtained by using a simple average of the probability distributions in the subset B is a special case of using W obtained by assigning the weights as $w_i = 1/|B|$ for $P_i \in B$ and $w_i = 0$ for $P_i \notin B$.

In this new situation we see that

$$\begin{aligned}\|P_W\|^2 &= \sum_{j=1}^n \left(\sum_{i=1}^t w_i P_{ij} \right)^2 \\ &= \sum_{i=1}^t w_i^2 \|P_i\|^2 + 2 \sum_{i=1}^{t-1} \sum_{k=i+1}^t w_i w_k P_i P_k\end{aligned}\quad (3.33)$$

It can be shown that $\sum_{i=1}^t w_i^2 + 2 \sum_{i=1}^{t-1} \sum_{k=i+1}^t w_i w_k = 1$ and thus $\|P_W\|^2$ is a weighted average of the complements of the Gini entropies and the Gini cross entropies. Thus $\|P_W\|^2$ provides an indication of the information in a fusion based on using the weighted function $W = [w_1, \dots, w_t]$. Here $\|P_W\|^2 \in [0, 1]$, and the bigger the value, the more information and less uncertainty.

A second aspect in determining the quality of the fusion obtained by using the weighted value $W = [w_1, \dots, w_t]$ is the credibility of this fusion. In order to determine the credibility associated with the weight vector W , $\text{Cred}(W)$, our point of departure will be the given credibility measure $\text{Cred} : 2^Z \rightarrow [0, 1]$, which associates with each subset $B \subseteq Z$ a value $\text{Cred}(B) \in [0, 1]$. Using some ideas provided by Wang and Klir [25] about extending set measures, we can obtain from this a credibility function, $\widetilde{\text{Cred}}(W)$, which associates with each vector W a value in the unit interval indicating a credibility of a fusion using a weighting of P_i based on the weights in W . In order to accomplish this we take the Choquet integral of a function f on Z with respect to the measure Cred on Z [26–28]. Here f is defined as $f(P_i) = w_i / \max_j w_j$. Thus $f(P_i)$ is the weight assigned to P_i divided by the max weight in W . Parenthetically, f can be viewed as the membership function of a fuzzy subset of Z .

We obtain the Choquet integral of f with respect to the measure Cred as follows [25]. Let ind be an index function so that $\text{ind}(j)$ is the index of element in Z with the j^{th} largest value for f , it is essentially the element in Z with the j^{th} largest value for w_i . Here then $f(P_{\text{ind}(j)}) = w_{\text{ind}(j)} / \max_i w_i$ is the j^{th} largest value for $f(P_i)$. We now let $H_j = \{P_{\text{ind}(k)} | k = 1 \text{ to } j\}$, it is the subset of Z with the j largest values of f . We note it is also the subset of Z with the j largest weights. Using this we obtain via the Choquet integral

$$\begin{aligned}\widetilde{\text{Cred}}(W) &= \sum_{j=1}^t (\text{Cred}(H_j) - \text{Cred}(H_{j-1})) f(P_{\text{ind}(j)}) \\ &= \sum_{j=1}^t (\text{Cred}(H_j) - \text{Cred}(H_{j-1})) \frac{w_{\text{ind}(j)}}{\max_i w_i} \\ &= \frac{1}{\max_i w_i} \sum_{j=1}^t (\text{Cred}(H_j) - \text{Cred}(H_{j-1})) w_{\text{ind}(j)}\end{aligned}\tag{3.34}$$

Consider now our original situation in which we used a subset B of Z with uniformly weighted components to get our fused value. Here the Credibility of this is $\text{Cred}(B)$. Let us show that we get this same value in the more general framework of using of vector W . In this case, we have as previously noted that W is such that $w_i = 1/|B|$ for $P_i \in B$ and $w_i = 0$ for $P_i \notin B$. In this case, we get

$\max_i w_i = \max_i [w_i] = 1/|B|$. Here we see that $\text{Cred}(W)$ is equal to

$$\begin{aligned} \widetilde{\text{Cred}}(W) &= \frac{1}{\max[W]} \left[\sum_{j=1}^{|B|} (\text{Cred}(H_j) - \text{Cred}(H_{j-1})) w_{\text{ind}(j)} \right. \\ &\quad \left. + \sum_{j=|B|+1}^t (\text{Cred}(H_j) - \text{Cred}(H_{j-1})) w_{\text{ind}(j)} \right] \end{aligned} \quad (3.35)$$

Since $w_{\text{ind}(j)} = 1/|B|$ for $j = 1$ to $|B|$ and $w_{\text{ind}(j)} = 0$ for $j = |B| + 1$ to t we get

$$\widetilde{\text{Cred}}(W) = \frac{1}{\max[W]} \left[\sum_{j=1}^{|B|} (\text{Cred}(H_j) - \text{Cred}(H_{j-1})) \frac{1}{|B|} \right] \quad (3.36)$$

with $\max[W] = 1/|B|$ we see that $\widetilde{\text{Cred}}(W) = \left[\sum_{j=1}^{|B|} (\text{Cred}(H_j) - \text{Cred}(H_{j-1})) \right]$. Further, we see that $\widetilde{\text{Cred}}(W) = \sum_{j=1}^{|B|} (\text{Cred}(H_j) - \text{Cred}(H_{j-1})) = \text{Cred}(H_{|B|}) - \text{Cred}(\emptyset) = \text{Cred}(H_{|B|})$. Since $H_{|B|}$ is the set of elements in Z with that $|B|$ largest weights, $H_{|B|} = B$ we get as desired that $\widetilde{\text{Cred}}(W) = \text{Cred}(B)$ in this case. We can express $\widetilde{\text{Cred}}(W) = \text{ChoqCred}(f_W)$ where f_W is the function $f_W(P_i) = w_i / \max[W]$. We shall now look at little more carefully at this $\widetilde{\text{Cred}}(W)$ function.

First let us make the following notation comment. If B is any nonempty subset of Z , then we shall denote W_B as its the weighting vector with all $w_j = 1/|B|$ for $P_j \in B$ and $w_j = 0$ for $P_j \notin B$. We note here that $f(P_j) = 1$ for all $P_j \in B$ and $f(P_j) = 0$ for $P_j \notin B$.

Theorem: Assume W^B is any weighting vector that has weights of zero for any element not in B ; then $\widetilde{\text{Cred}}(W_B) \geq \widetilde{\text{Cred}}(W^B)$.

Proof: Let f_B and f^B be the associated function for each of these weighted vectors. Here we have $f_B(P_j) = 1$ for $P_j \in B$ and $f_B(P_j) = 0$ for $P_j \notin B$ also $f^B(P_j) \leq 1$ for $P_j \in B$ and $f^B(P_j) = 0$ for $P_j \notin B$. From this we see that $f_B(P_j) \geq f^B(P_j)$ for all P_j . From the monotonicity of the Choquet integral [22] we get

$$\widetilde{\text{Cred}}(W_B) = \text{ChoqCred}(f_B) \geq \text{ChoqCred}\left(f^B\right) = \widetilde{\text{Cred}}(W^B) \quad (3.37)$$

More generally we see the following

Theorem: Let $D \subseteq B$ be two subsets of Z . Let W^D be a weighting vector that has zero weights for any element not in D . Then $\widetilde{\text{Cred}}(W_B) \geq \widetilde{\text{Cred}}(W^D)$.

Proof: Here again we have $f_B(P_j) = 1$ for $P_j \in B$ and $f_B(P_j) = 0$ for $P_j \notin B$ while $f^D(P_j) \leq 1$ for $P_j \in D$ and $f^D(P_j) = 0$ for $P_j \notin D$, $P_j \in \overline{D}$ where $\overline{D} \cap B = \emptyset$. From this we get $f_B(P_j) \geq f^D(P_j)$ for all P_j . The result again follows from the monotonicity of the Choquet integral.

A corollary of the preceding is that if $D \subseteq B$, then $\widetilde{\text{Cred}}(W_B) \geq \widetilde{\text{Cred}}(W_D)$. This is of course the same as the monotonicity of the measure Cred , $\text{Cred}(B) \geq \text{Cred}(D)$ if $D \subseteq B$. If W is a weighting vector such the components in B are nonzero, while those that are not in B are zero, we say that W is based on B . So we have shown for any weight based on B the most credibility occurs if we uniformly assign the weights to the elements in B .

More generally from the monotonicity of the Choquet integral we see if $W = [w_1, \dots, w_t]$ and $\widehat{W} = [\widehat{w}_1, \dots, \widehat{w}_t]$ are two sets of weighting vectors so that $\frac{w_j}{\max[W]} \geq \frac{\widehat{w}_j}{\max[\widehat{W}]}$ for all j then $\widetilde{\text{Cred}}(W) \geq \widetilde{\text{Cred}}(\widehat{W})$. We note that while there exists some type of relationship between the choice of weighting vector W and $\widetilde{\text{Cred}}(W)$, the relationship between $\|P_W\|^2$ and the choice of the weights is more complex since $\|P_W\|^2 = \sum_{i=1}^t w_i^2 \|P_i\|^2 + 2 \sum_{i=1}^{t-1} \sum_{k=i+1}^t w_i w_k P_i P_k$.

An interesting characterizing feature we can associate with each P_i is $R_i = \sum_{k=1}^t P_i P_k$. We see R_i is the average cross-information. Clearly the bigger this

$$k \neq i$$

value, the more desirable the probability distribution P_i is for including in the weighted fusion. Similarly, the larger the $\|P_i\|^2$, the more desirable the P_i is for including in the weighted fusion. Here then given a set of probability distributions, $Z = \{P_1, \dots, P_t\}$ and a credibility measure Cred on Z , the question we are faced with is the problem of how to select the appropriate weights for the fusion given our interest in obtaining a fused value that is both informative and credible. While we have provided formulations for calculating the information, $\|P_W\|^2$, and credibility, $\widetilde{\text{Cred}}(W)$, given a weighting vector W , the process of obtaining the optimal weighted vector W is a difficult multi-criteria optimization problem. Given the apparent difficulty of solving this optimization problem, we shall look for some satisfying solution to this problem.

Here again we shall assume some minimal required level of credibility α and look for solutions with a large value of $\|P_W\|^2$ that have at least this minimal level of credibility. Let us look at the formulation for $\|P_W\|^2$

$$\|P_W\|^2 = \sum_{i=1}^t w_i^2 P_i^2 + 2 \sum_{i=1}^{t-1} \sum_{k=i+1}^t w_i w_k P_i P_k \quad (3.38)$$

We note that terms $P_i P_k$ and P_i^2 can be calculated off line and are independent of choice of W . Let us denote these as follows: $P_i P_k = m_{ik}$ and $P_i^2 = m_{ii}$. Thus here then

$$\|P_W\|^2 = \sum_{i=1}^t w_i^2 m_{ii} + 2 \sum_{i=1}^{t-1} \sum_{k=i+1}^t w_i w_k m_{ik} \quad (3.39)$$

and we have the constraints that $\sum_{i=1}^t w_i = 1$ and $w_i \in [0, 1]$. Here our objective is to try to get $\|P_W\|^2$ as large as possible. In addition,

$$\widetilde{\text{Cred}}(W) = \sum_{j=1}^t (\text{Cred}(H_j) - \text{Cred}(H_{j-1})) \frac{w_{\text{ind}(j)}}{\max_i [w_i]} \quad (3.40)$$

which we require to have a value of at least α .

A useful characterizing feature we can associate with the basic Cred measure on the space $Z = \{P_1, \dots, P_t\}$ is its Shapely index [29–31]. For any $P_j \in Z$ we define its Shapely index S_j as

$$S_j = \sum_{k=0}^{t-1} \left(\gamma_k \sum_{\substack{K \subset F_j \\ |K|=k}} (\text{Cred}(K \cup \{P_j\}) - \text{Cred}(K)) \right) \quad (3.41)$$

In the above, K is a subset of cardinality $|K|$, $F_j = Z - \{P_j\}$ and $\gamma_k = \frac{(n-k-1)!k!}{n!}$. It can be shown [31] that $S_j \in [0, 1]$ and $\sum_{j=1}^t S_j = 1$. This index can be seen as the average increase in “credibility” obtained by adding the element P_j to a set that doesn’t contain it. We note that it can be shown that if Cred is a simple additive measure with $\text{Cred}(\{P_j\}) = \alpha_j$, then $S_j = \alpha_j$, and if Cred is a cardinality based measure, then $S_j = 1/t$ for all j [31]. Using these Shapely index values, we can obtain an approximation to the credibility of a subset associated with a weighting vector $W = [w_1, \dots, w_t]$ in particular $\widetilde{\text{Cred}}(W) = \sum_{j=1}^t S_j \frac{w_j}{\max_i [w_i]}$.

We recall that for the case where W is related to a crisp subset B of X , $w_j = 1/|B|$ for all $P_j \in B$, and $w_j = 0$ for all $P_j \notin B$. Here then $\max_i [w_i] = 1/|B|$ and we have

$\frac{w_j}{\max_i [w_i]} = 1$ for $P_j \in B$ and $\frac{w_j}{\max_i [w_i]} = 0$ for $P_j \notin B$. From this we get $\widetilde{\text{Cred}}(B) = \sum_{x_j \in B} S_j = \text{Cred}(B)$. This is $\text{Cred}(B)$ for the special case of uniform weights.

Using this definition for $\widetilde{\text{Cred}}(W)$, we now formulate our problem of finding the most informative weighted fusion of the elements of Z given that we want a minimal credibility of α . In particular, the problem becomes:

Find w_1, \dots, w_t to maximize:

$$\|P_W\|^2 = \sum_{i=1}^t w_i^2 m_{ii} + 2 \sum_{i=1}^{t-1} \sum_{k=i+1}^t w_i w_k m_{ik} \quad (3.42)$$

such that:

$$\frac{1}{\max_i [w_i]} \sum_{i=1}^t w_i S_i \geq \alpha;$$

$$w_i \in [0, 1];$$

$$\sum_{i=1}^t w_i = 1$$

This is a nonlinear mathematical programming problem.

3.8 Conclusion

Our objective here was to obtain quality-fused values about the value of a variable from information provided by multiple sources in the form of probabilistic distributions. Here quality was measured by a lack of uncertainty in the fused value, more informative fused values, and the use of credible sources. We introduced a vector representation for a probability distribution, and using the Gini formulation for entropy, we showed how the norm of the vector provides a measure of the certainty, information, associated with a probability distribution. We looked at special cases of fusion for source inputs that were maximally uncertain and certain. We provided a measure of credibility associated with subsets of sources. We looked at the issue of finding the highest quality fused value from the weighted aggregations of source provided probability distributions.

Acknowledgements Ronald Yager was supported in part by an ONR grant award. Fred Petry would like to thank the Naval Research Laboratory's Base Program, Program Element No. 0602435N for their sponsorship.

References

1. R. Pirsig, *Zen and the Art of Motorcycle Maintenance: An Inquiry into Values* (HarperCollins, New York, 1999)
2. J. Evans, W. Lindsay, *The Management and Control of Quality* (South-Western College Publications, Cincinnati, 1999)
3. H. Miller, The multiple dimensions of information quality. *Inf. Syst. Manag.* **13**(2), 79–82 (1996)
4. J.-E. Mai, The quality and qualities of information. *J. Am. Soc. Inf. Sci. Technol.* **64**(4), 675–688 (2013)
5. D.L. Hall, C.Y. Chong, J. Llinas, M. Liggins, *Distributed Data Fusion for Network-Centric Operations* (CRC Press, Boca Raton, 2012)

6. P. Elmore, F.E. Petry, R.R. Yager, Comparative measures of aggregated uncertainty representations. *J. Ambient. Intell. Humaniz. Comput.* **5**, 809–819 (2014)
7. H. Bustince, P. Burillo, Correlation of interval-valued intuitionistic fuzzy sets. *Fuzzy Set. Syst.* **74**, 237–244 (1995)
8. M.R. Spiegel, D. Lipschutz, D. Spellman, *Vector Analysis* (John Wiley, Hoboken, 2009)
9. R.R. Yager, N. Alajlan, An intelligent interactive approach to group aggregation of subjective probabilities. Technical report MII-3502, Machine Intelligence Institute, Iona College, New Rochelle, NY, 2015
10. S. Kullback, *Information Theory and Statistics* (John Wiley and Sons, New York, 1959)
11. B. Buck, *Maximum Entropy in Action: A Collection of Expository Essays* (Oxford University Press, New York, 1991)
12. J. Aczel, Z. Daroczy, *On Measures of Information and their Characterizations* (Academic, New York, 1975)
13. R.R. Yager, K. Engemann, Entropy measures in sports. *Int. J. Syst. Meas. Decis.* **1**, 67–72 (1981)
14. R.R. Yager, V. Kreinovich, Entropy conserving transforms and the entailment principle. *Fuzzy Sets Syst.* **158**, 1397–1405 (2007)
15. D.G. Luenberger, *Information Science* (Princeton University Press, Princeton, 2006)
16. C. Gini, Variabilità e mutabilità,” Reprinted in Memorie di Metodologica Statistica, ed. by E. Pizetti, T. Salvemini (Libreria Eredi Virgilio Veschi, Rome, 1955)
17. P.-T. De Boer, D.P. Kroese, S. Mannor, R.Y. Rubinstein, A tutorial on the cross-entropy method. *Ann. Oper. Res.* **134**, 19–67 (2005)
18. R.Y. Rubinstein, D.P. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning* (Springer-Verlag, New York, 2004)
19. L.A. Zadeh, A computational approach to fuzzy quantifiers in natural languages. *Comput. Math. Appl.* **9**, 149–184 (1983)
20. T. Takagi, M. Sugeno, Fuzzy identification of systems and its application to modeling and control. *IEEE Trans. Syst. Man Cybern.* **15**, 116–132 (1985)
21. R.R. Yager, Modeling prioritized multi-criteria decision making. *IEEE Trans. Syst. Man Cybern. B* **34**, 2396–2404 (2004)
22. R.R. Yager, Prioritized aggregation operators. *Int. J. Approx. Reason.* **48**, 263–274 (2008)
23. R.R. Yager, On prioritized multiple criteria aggregation. *IEEE Trans. Syst. Man Cybern. B* **42**, 1297–1305 (2012)
24. L.A. Zadeh, Fuzzy logic = computing with words. *IEEE Trans. Fuzzy Syst.* **4**, 103–111 (1996)
25. Z. Wang, G.J. Klir, *Generalized Measure Theory* (Springer, New York, 2009)
26. G. Beliakov, A. Pradera, T. Calvo, *Aggregation Functions: A Guide for Practitioners* (Springer, Heidelberg, 2007)
27. E.P. Klement, R. Mesiar, E. Pap, A universal integral as common frame for Choquet and Sugeno. *IEEE Trans. Fuzzy Syst.* **18**, 178–187 (2010)
28. Z. Wang, R. Yang, K.-S. Leung, *Nonlinear Integrals and Their Applications in Data Mining* (World Scientific, Singapore, 2010)
29. J.L. Marichal, M. Roubens, “Entropy of discrete fuzzy measures,” in *Proceedings of Third International Workshop on Preferences and Decisions*, Trento, pp. 135–148, 2000
30. T. Murofushi, A technique for reading fuzzy measures (i): the Shapely value with respect to a fuzzy measure, in *Proceedings Second Fuzzy Workshop*, Nagaoka, Japan (in Japanese), pp. 39–48, 1992
31. R.R. Yager, On the entropy of fuzzy measures. *IEEE Trans. Fuzzy Sets Syst.* **8**, 453–461 (2000)

Chapter 4

Conflict Management in Information Fusion with Belief Functions



Arnaud Martin

Abstract In information fusion, the conflict is an important concept. Indeed, combining several imperfect experts or sources allows conflict. In the theory of belief functions, this notion has been discussed a lot. The mass appearing on the empty set during the conjunctive combination rule is generally considered as conflict, but that is not really a conflict. Some measures of conflict have been proposed, and some approaches have been proposed in order to manage this conflict or to decide with conflicting mass functions. We recall in this chapter some of them, and we propose a discussion to consider the conflict in information fusion with the theory of belief functions.

Keywords Belief functions · Conflict measure · Combination rule · Reliability · Ignorance · Decision

4.1 Introduction

The theory of belief functions was first introduced by [5] in order to represent some imprecise probabilities with *upper* and *lower probabilities*. Then [30] proposed a mathematical theory of evidence which is now widely used for information fusion. Combining imperfect sources of information leads inevitably to conflict. One can consider that the conflict comes from the non-reliability of the sources or the sources do not give information on the same observation. In this last case, one must not combine them.

Let $\mathcal{Q} = \{\omega_1, \dots, \omega_n\}$ be a frame of discernment of exclusive and exhaustive hypothesis. A mass function m , also called basic belief assignment (bba), is the mapping from elements of the power set $2^{\mathcal{Q}}$ (composed by all the disjunctions of \mathcal{Q}) onto $[0, 1]$ such that:

A. Martin (✉)
DRUID, Univ Rennes, CNRS, IRISA, Lannion, France
e-mail: Arnaud.Martin@univ-rennes1.fr

$$\sum_{X \in 2^{\Omega}} m(X) = 1. \quad (4.1)$$

A focal element X is an element of 2^{Ω} such that $m(X) \neq 0$. If the focal elements are nested, the mass functions is *consonant*. A simple mass function, noted A^w is given by:

$$\begin{cases} m(A) = w \\ m(\Omega) = 1 - w \end{cases} \quad (4.2)$$

This mass function allows to show that we can model an imprecise information (if A is a union of singletons ω_i) and an uncertain information (if $w > 0$). All nondogmatic mass functions (with $m(\Omega) > 0$) can be decomposed by a set of simple mass functions [30].

Constraining $m(\emptyset) = 0$ corresponds to a closed-world assumption [30], while allowing $m(\emptyset) \geq 0$ corresponds to an open world assumption [32]. Smets interpreted this mass on the empty set such as a non-expected hypothesis and normalizes it in the pignistic probability defined for all $X \in 2^{\Omega}$, with $X \neq \emptyset$ by:

$$\text{BetP}(X) = \sum_{Y \in 2^{\Omega}, Y \neq \emptyset} \frac{|X \cap Y|}{|Y|} \frac{m(Y)}{1 - m(\emptyset)}. \quad (4.3)$$

The pignistic probability can be used in the decision process such as a compromise between the credibility and the plausibility. The credibility is given for all $X \in 2^{\Omega}$ by:

$$\text{bel}(X) = \sum_{Y \subseteq X, Y \neq \emptyset} m(Y), \quad (4.4)$$

The plausibility is given for all $X \in 2^{\Omega}$ by:

$$\text{pl}(X) = \sum_{Y \in 2^{\Omega}, Y \cap X \neq \emptyset} m(Y) = \text{bel}(\Omega) - \text{bel}(X^c) = 1 - m(\emptyset) - \text{bel}(X^c), \quad (4.5)$$

where X^c is the complementary of X . Hence, if we note the decision function f_d that can be the pignistic probability, the credibility, or the plausibility, we choose the element $A \in 2^{\Omega}$ for a given observation if:

$$A = \underset{X \in \Omega}{\operatorname{argmax}} (f_d(X)). \quad (4.6)$$

The decision is made on the mass function obtained by the combination of all the mass function from the sources.

The first combination rule has been proposed by Dempster [5] and is defined for two mass functions m_1 and m_2 , for all $X \in 2^{\Omega}$, with $X \neq \emptyset$ by:

$$m_{\text{DS}}(X) = \frac{1}{1 - \kappa} \sum_{A \cap B = X} m_1(A)m_2(B), \quad (4.7)$$

where $\kappa = \sum_{A \cap B = \emptyset} m_1(A)m_2(B)$ is the inconsistence of the combination and generally called conflict. We call it here the *global conflict* such as the sum of all *partial conflicts*.

To stay in an open world, Smets [32] proposes the non-normalized conjunctive rule given for two mass functions m_1 and m_2 and for all $X \in 2^{\Omega}$ by:

$$m_{\text{Conj}}(X) = \sum_{A \cap B = X} m_1(A)m_2(B) := (m_1 \odot m_2)(X). \quad (4.8)$$

These both rules allow to reduce the imprecision of the focal elements and to increase the belief on concordant elements after the fusion. The main assumptions to apply these rules are the cognitive independence and the reliability of the sources.

Based on the results of these rules, the problem enlightened by the famous Zadeh's example [37] is the repartition of the global conflict. Indeed, consider $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and two experts opinions given by $m_1(\omega_1) = 0.9, m_1(\omega_3) = 0.1$, and $m_2(\omega_2) = 0.9, m_1(\omega_3) = 0.1$, the mass function resulting in the combination using Dempster's rule is $m(\omega_3) = 1$ and using conjunctive rule is $m(\emptyset) = 0.99, m(\omega_3) = 0.01$. Therefore, several combination rules have been proposed to manage this global conflict [23, 33].

As observed in [17, 25], the weight of conflict given by $\kappa = m_{\text{Conj}}(\emptyset)$ is not a conflict measure between the mass functions. Indeed, the conjunctive-based rules are not idempotent (as the majority of the rules defined to manage the global conflict): the combination of identical mass functions leads generally to a positive value of κ . Hence, new kind of conflict measures are defined in [25].

In Sect. 4.2, we recall some measures of conflict in the theory of belief functions. Then, in Sect. 4.3 we present the ways to manage the conflict either before the combination or in the combination rule. Section 4.4 presents some decision methods in order to consider the conflict during this last step of information process.

4.2 Modeling Conflict

First of all, we should not mix up conflict measure and contradiction measure. The measures defined by [13, 34] are not conflict measures, but some discord and specificity measures (to take the terms of [16]) we call contradiction measures. We define the contradiction and conflict measures by the following definitions:

Definition A *contradiction* in the theory of belief functions quantifies how a mass function contradicts itself.

Definition (C1) The *conflict* in the theory of belief functions can be defined by the contradiction between two or more mass functions.

Therefore, is the mass of the empty set or the functions of this mass (such as $-\ln(1 - m_{\text{Conj}}(\emptyset))$) proposed by [30]) a conflict measure? It seems obvious that the property of the non-idempotence is a problem to use this as a conflict measure. However, if we define a conflict measure such as $\text{Conf}(m_1, m_2) = m_{\text{Conj}}(\emptyset)$, we note that $\text{Conf}(m_1, m_\Omega) = 0$ where $m_\Omega(\Omega) = 1$ is the ignorance. Indeed, the ignorance is the neutral element for the conjunctive combination rule. This property seems to be followed from a conflict measure.

Other conflict measures have been defined. In [14], a conflict measure is given by:

$$\text{Conf}(m_1, m_2) = 1 - \frac{\mathbf{pl}_1^T \cdot \mathbf{pl}_2}{\|\mathbf{pl}_1\| \|\mathbf{pl}_2\|} \quad (4.9)$$

where \mathbf{pl} is the plausibility function and $\mathbf{pl}_1^T \cdot \mathbf{pl}_2$ the vector product in 2^n space of both plausibility functions. However, generally $\text{Conf}(m_1, m_\Omega) \neq 0$ that seems counter-intuitive.

4.2.1 Auto-conflict

Introduced by [28], the auto-conflict of order s for one source is given by:

$$a_s = \left(\bigodot_{j=1}^s m \right) (\emptyset). \quad (4.10)$$

where \odot is the conjunctive operator of Eq. (4.8). The following property holds $a_s \leq a_{s+1}$, meaning that due to the non-indempotence of \odot , the more masses m are combined with itself, the nearer to 1 κ is, and so in a general case, the more the number of sources is high, the nearer to 1 κ is. The behavior of the auto-conflict was studied in [25], and it was shown that we should take into account the auto-conflict in the global conflict in order to really define a conflict. In [36], the auto-conflict was defined and called the plausibility of the belief structure with itself. The auto-conflict is a kind of measure of the contradiction, but depends on the order s of the combination. A measure of contradiction independent on the order has been defined in [31].

4.2.2 Conflict Measure Based on a Distance

With the definition of the conflict (C1), we consider sources to be in conflict if their opinions are far from each other in the space of corresponding bbas. That suggests a notion of distance. That is the reason why in [25], we give a definition of the measure of conflict between sources assertions through a distance between their respective bbas. The conflict measure between 2 experts is defined by:

$$\text{Conf}(1, 2) = d(m_1, m_2). \quad (4.11)$$

We defined the conflict measure between one source j and the other $M - 1$ sources by:

$$\text{Conf}(j, \mathcal{E}) = \frac{1}{M-1} \sum_{i=1, i \neq j}^M \text{Conf}(i, j), \quad (4.12)$$

where $\mathcal{E} = \{1, \dots, M\}$ is the set of sources in conflict with j . Another definition is given by:

$$\text{Conf}(j, M) = d(m_j, \overline{m_M}), \quad (4.13)$$

where $\overline{m_M}$ is the bba of the artificial source representing the combined opinions of all the sources in \mathcal{E} except j .

A comparison of distances in the theory of belief functions is presented in [14]. We consider the distance defined in [15] as the most appropriate. This distance is defined for two basic belief assignments \mathbf{m}_1 and \mathbf{m}_2 on 2^{Ω} by:

$$d_J(m_1, m_2) = \sqrt{\frac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)^T \underline{D}(\mathbf{m}_1 - \mathbf{m}_2)}, \quad (4.14)$$

where \underline{D} is an $2^{|\Omega|} \times 2^{|\Omega|}$ matrix based on Jaccard dissimilarity whose elements are:

$$D(A, B) = \begin{cases} 1, & \text{if } A = B = \emptyset, \\ \frac{|A \cap B|}{|A \cup B|}, & \forall A, B \in 2^{\Omega}. \end{cases} \quad (4.15)$$

An interesting property of this measure is given by $\text{Conf}(m, m) = 0$. That means that there is no conflict between a source and itself (that is not a contradiction). However, we generally do not have $\text{Conf}(m, m_{\Omega}) = 0$, where $m_{\Omega}(\Omega) = 1$ is the ignorance.

4.2.3 Conflict Measure Based on Inclusion Degree and Distance

We have seen that we cannot use the mass on the empty set as a conflict measure because of the non-idempotence of the conjunctive rule. We also have seen that the conflict measure based on the distance is not null in general for the ignorance mass. The conjunctive rule does not transfer mass on the empty set if the mass functions are *included*. We give here some definitions of the inclusion.

Definition 1 (Strict inclusion) We say that the mass function m_1 is *included* in m_2 if all the focal elements of m_1 are included in *each* focal elements of m_2 .

Definition 2 (Light inclusion) We say that the mass function m_1 is *included* in m_2 if all the focal elements of m_1 are included in *at least one* focal element of m_2 .

Definition We note this inclusion by $m_1 \subseteq m_2$. The mass functions are *included* if m_1 is included in m_2 or m_2 is included in m_1 .

In [21], we propose a conflict measure base on five following axioms. Let note $\text{Conf}(m_1, m_2)$ a conflict measure between the mass functions m_1 and m_2 . We present hereafter essential properties that must verify a conflict measure.

1. Nonnegativity:

$$\text{Conf}(m_1, m_2) \geq 0 \quad (4.16)$$

A negative conflict does not make sense. This axiom is, therefore, necessary.

2. Identity:

$$\text{Conf}(m_1, m_1) = 0 \quad (4.17)$$

Two equal mass functions are not in conflict. This property is not reached by the global conflict, but seems natural.

3. Symmetry:

$$\text{Conf}(m_1, m_2) = \text{Conf}(m_2, m_1) \quad (4.18)$$

The conflict measure must be symmetric. We do not see any case where the non-symmetry can make sense.

4. Normalization:

$$0 \leq \text{Conf}(m_1, m_2) \leq 1 \quad (4.19)$$

This axiom may not be necessary to define a conflict measure, but the normalization is very useful in many applications requiring a conflict measure.

5. Inclusion:

$$\text{Conf}(m_1, m_2) = 0, \text{ if and only if } m_1 \subseteq m_2 \text{ or } m_2 \subseteq m_1 \quad (4.20)$$

This axiom means that if the focal elements of two mass functions are not conflicting (the intersection is never empty), the mass functions are not in conflict and the mass functions cannot be in conflict if they are included. This axiom is not satisfied by a distance-based conflict measure.

These proposed axioms are very similar to ones defined in [7]. If a conflict measure satisfied these axioms that is not necessary a distance. Indeed, we only impose the identity and not the definiteness ($\text{Conf}(m_1, m_2) = 0 \Leftrightarrow m_1 = m_2$). The axiom of inclusion is less restrictive and makes more sense for a conflict measure. Moreover, we do not impose the triangle inequality ($\text{Conf}(m_1, m_2) \leq \text{Conf}(m_1, m_3) + \text{Conf}(m_3, m_2)$). It can be interesting to have $\text{Conf}(m_1, m_2) \geq \text{Conf}(m_1, m_3) + \text{Conf}(m_3, m_2)$ meaning that an expert given the mass function m_3 can reduce the conflict. Therefore, a distance (with the property of the triangle inequality) cannot be used directly to define a conflict measure.

We see that the axiom of inclusion seems very important to define a conflict measure. This is the reason why we define in [21] a degree of inclusion to measure how two mass functions are included. Let the inclusion index: $\text{Inc}(X_1, Y_2) = 1$ if $X_1 \subseteq Y_2$ and 0 otherwise, where X_1 and Y_2 are two focal elements of m_1 and m_2 , respectively. According to Definitions 1 and 2, we introduce two degrees of inclusion of m_1 in m_2 . A strict degree of inclusion of m_1 in m_2 is given by:

$$d_{incS}(m_1, m_2) = \frac{1}{|\mathcal{F}_1| |\mathcal{F}_2|} \sum_{X_1 \in \mathcal{F}_1} \sum_{Y_2 \in \mathcal{F}_2} \text{Inc}(X_1, Y_2) \quad (4.21)$$

where \mathcal{F}_1 and \mathcal{F}_2 are the set of focal elements of m_1 and m_2 , respectively, and $|\mathcal{F}_1|, |\mathcal{F}_2|$ are the number of focal elements of m_1 and m_2 .

This definition is very strict, so we introduce a light degree of inclusion of m_1 in m_2 given by:

$$d_{incL}(m_1, m_2) = \frac{1}{|\mathcal{F}_1|} \sum_{X_1 \in \mathcal{F}_1} \max_{Y_2 \in \mathcal{F}_2} \text{Inc}(X_1, Y_2). \quad (4.22)$$

Let $\delta_{inc}(m_1, m_2)$ a degree of inclusion of m_1 and m_2 define by:

$$\delta_{inc}(m_1, m_2) = \max(d_{inc}(m_1, m_2), d_{inc}(m_2, m_1)) \quad (4.23)$$

This degree gives the maximum of the proportion of focal elements from one mass function included in another one. Therefore, $\delta_{inc}(m_1, m_2) = 1$ if and only if m_1 and m_2 are included, and the axiom of inclusion is reached for $1 - \delta_{inc}(m_1, m_2)$.

Hence, we define in [21], a conflict measure between two mass functions m_1 and m_2 by:

$$\text{Conf}(m_1, m_2) = (1 - \delta_{\text{inc}}(m_1, m_2))d_J(m_1, m_2) \quad (4.24)$$

where d_J is the distance defined by Eq. (4.14). All the previous axioms are satisfied. Indeed the axiom of inclusion is $1 - \delta_{\text{inc}}(m_1, m_2)$ and the distance d_J satisfied the other axioms. Moreover $0 \leq \delta_{\text{inc}}(m_1, m_2) \leq 1$, by the product of $1 - \delta_{\text{inc}}$ and d_J , all the axioms are satisfied.

For more than two mass functions, the conflict measure between one source j and the other $M - 1$ sources can be defined from Eqs. (4.12) or (4.13).

4.3 Managing Conflict

The role of conflict is essential in information fusion. Different ways can be used to manage and reduce the conflict. The conflict can come from the low reliability of the sources. Therefore, we can use this conflict to estimate the reliability of the sources if we cannot learn it on databases as proposed in [25]. Hence, we reduce the conflict before the combination, but we can also directly manage the conflict in the rule of combination as generally made in the theory of belief functions such as explained in [23, 33].

According to the application, we do not search always to reduce the conflict. For example, we can use the conflict measure such as an indicator of the inconsistence of the fusion, for example, in databases [2]. Conflict information can also be an interesting information in some applications such as presented in [29].

4.3.1 Managing the Conflict Before the Combination

The conflict appearing while confronting several experts' opinions can be used as an indicator of the relative reliability of the experts. We have seen that there exist many rules in order to take into account the conflict during the combination step. These rules do not make the difference between the conflict (global or local conflict) and the auto-conflict due to the non-idempotence of the majority of the rules. We propose here the use of a conflict measure in order to define a reliability measure, which we consider before the combination, in a discounting procedure.

When we can quantify the reliability of each source, we can weaken the basic belief assignment before the combination by the discounting procedure:

$$\begin{cases} m_j^\alpha(X) = \alpha_j m_j(X), \forall X \in 2^\Omega \setminus \{\Omega\} \\ m_j^\alpha(\Omega) = 1 - \alpha_j(1 - m_j(\Omega)). \end{cases} \quad (4.25)$$

$\alpha_j \in [0, 1]$ is the discounting factor of the source j that is, in this case, the reliability of the source j , eventually as a function of $X \in 2^\Omega$.

Other discounting procedures are possible such as the contextual discounting [26] or a discounting procedure based on the credibility or the plausibility functions [38].

According to the applications, we can learn the discounting factors α_j , for example, from the confusion matrix [18]. In many applications, we cannot learn the reliability of each source. A general approach to the evaluation of the discounting factor without learning is given in [10]. For a given bba, the discounting factor is obtained by the minimization on α of a distance given by:

$$\text{Dist}^{\alpha_j} = \sum_{A \in \Omega} (\text{BetP}_j(A) - \delta_{A,j})^2, \quad (4.26)$$

where BetP_j is the pignistic probability (Eq. (4.3)) of the bba given by the source j and $\delta_{A,j} = 1$ if the source j supports A and 0 otherwise.

This approach is interesting with the goal of making decision based on pignistic probabilities. However, if the source j does not support a singleton of Ω , the minimization on α_j does not work well.

In order to combine the bbas of all sources together, we propose here to estimate the reliability of each source j from the conflict measure Conf between the source j and the others by:

$$\alpha_j = f(\text{Conf}(j, M)), \quad (4.27)$$

where f is a decreasing function. We can choose:

$$\alpha_j = (1 - \text{Conf}(j, M)^\lambda)^{1/\lambda}, \quad (4.28)$$

where $\lambda > 0$. We illustrate this function for $\lambda = 2$ and $\lambda = 1/2$ on Fig. 4.1. This function allows to give more reliability to the sources with few conflict with the other.

Other definitions are possible. The credibility degree defined in [4] is also based on the distance given in Eq. (4.14) and could also be interpreted as the reliability of the source. However the credibility degree in [4] is integrated directly in the combination with a weighted average. Our reliability measure allows the use of all the existing combination rules.

4.3.2 Managing the Conflict in the Combination

According to the application, if we cannot reduce the conflict before the combination, we can do it by incorporating it into a combination rule. The choice of the combination rule is not easy, but it can be done by utilizing the conflict and the assumption on its origin. Indeed Dempster's rule can be applied if the sources are independent and reliable. Dempster's rule is given for S sources for all $X \in 2^\Omega$,

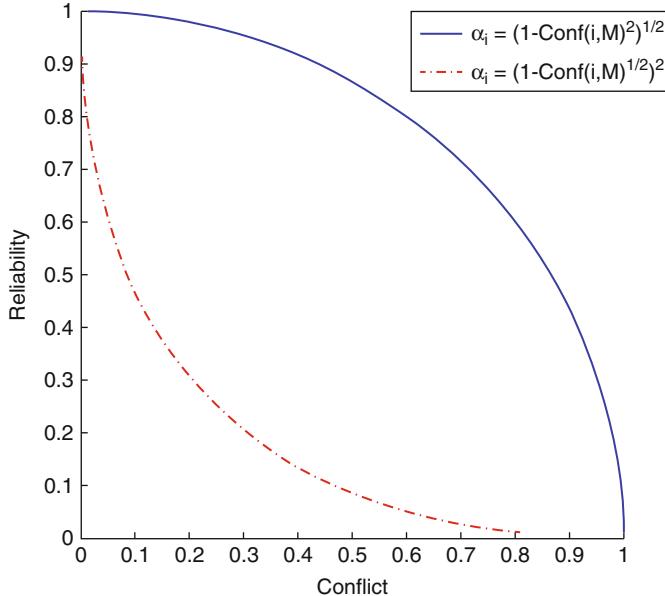


Fig. 4.1 Reliability of one source based on conflict of the source with the other sources

$X \neq \emptyset$ by:

$$m_{DS}(X) = \frac{1}{1 - m_{\text{Conj}}(\emptyset)} \sum_{Y_1 \cap \dots \cap Y_s = X} \prod_{j=1}^s m_j(Y_j) = \frac{m_{\text{Conj}}(X)}{1 - \kappa}, \quad (4.29)$$

where $\kappa = m_{\text{Conj}}(\emptyset)$, $Y_j \in 2^\Omega$ is the answer of the source S_j and $m_j(Y_j)$ the associated mass function. This normalization by $1 - m_{\text{Conj}}(\emptyset)$ hides the conflict, and so this rule is interesting only if we consider the closed world and if the sources are not highly conflicting.

If the assumption of independent and reliable sources is not reached, the application of Dempster's rule can produce some global conflict.

4.3.2.1 Conflict Coming from a False Assumption of Closed World

In the closed world, the elements of the frame of discernment are assumed exhaustive. If $m(\emptyset) > 0$, this mass can be interpreted such as another element, and so the assumption on the exhaustiveness of the frame of discernment is false. Hence, Smets [33] proposed the use of the conjunctive rule given for S sources for all $X \in 2^\Omega$ by:

$$m_{\text{Conj}}(X) = \sum_{Y_1 \cap \dots \cap Y_s = X} \prod_{j=1}^s m_j(Y_j). \quad (4.30)$$

Here, the sources must be cognitively independent and reliable, while the open world is considered. Hence, the mass on the empty set can be interpreted as another element unknown by the sources. In fact, in the proposed model by Smets, the conflict is transferred during the decision step by the pignistic probability [32], hiding the conflict to the end. This rule cannot be used in applications with a high value of κ .

The global conflict comes from the sum of the partial conflict. Hence, if the global conflict can be interpreted as an unknown element, all the partial conflict can be interpreted as many unknown elements. In that case we can keep the partial conflict in order to decide on these elements (see Sect. 4.4 for this consideration). Therefore the assumption of the exclusivity is considered here as false.

Under this assumption, the mass functions are no more defined on the power set 2^Ω but on the so-called hyper power set¹ D^Ω . Therefore the space Ω is closed by the union and intersection operators. This extension of the power set leads to a lot of reflections around this new expressiveness taking the name of DSmT (*Dempster-Shafer modified Theory*).

One can also consider a partial exclusiveness of the frame of discernment. Hence, we introduce the notation D_r^Ω in [19] in order to integrate some constraints on the exclusiveness of some elements of Ω and to reduce the hyper power set size. Under these assumptions, we define the PCR6 rule in [22], given by:

$$\begin{aligned} m_{\text{PCR6}}(X) &= m_{\text{Conj}}(X) \\ &+ \sum_{j=1}^s m_j(X)^2 \sum_{\substack{j'=1 \\ \bigcap_{j'=1}^{s-1} Y_{\sigma_j(j')} \cap X = \emptyset}}^{s-1} \left(\frac{\prod_{j'=1}^{s-1} m_{\sigma_j(j')}(Y_{\sigma_j(j')})}{m_j(X) + \sum_{j'=1}^{s-1} m_{\sigma_j(j')}(Y_{\sigma_j(j')})} \right), \end{aligned} \quad (4.31)$$

$(Y_{\sigma_j(1)}, \dots, Y_{\sigma_j(s-1)}) \in (D_r^\Omega)^{s-1}$

where σ is given by:

$$\begin{cases} \sigma_j(j') = j' & \text{if } j' < j, \\ \sigma_j(j') = j' + 1 & \text{if } j' \geq j, \end{cases} \quad (4.32)$$

This rule transfers the partial conflicts on the elements that generate it, proportionally to their masses. This rule has been used in many applications allowing for good results.

¹This notation is introduced by [8] and D come from the Dedekind lattice.

4.3.2.2 Conflict Coming from the Assumption of Source's Independence

If we consider some dependent sources, the conjunctive rule cannot be used. If we want to combine mass functions coming from dependent sources, the combination rule has to be idempotent. Indeed, if we combine two identical dependent mass functions (coming from different dependent sources), we have to obtain the same mass function without any global conflict.

The simplest way to obtain a non-idempotent rule is the average of the mass functions such given in [27] by:

$$m_{\text{Mean}}(X) = \frac{1}{S} \sum_{j=1}^S m_j(Y_j). \quad (4.33)$$

We showed the interest in a such rule in [28], but in that case the sources have to be assumed totally reliable. If the sources give high conflicting information, this rule can provide some errors in the decision process.

The cautious rule [6] could be used to combine mass functions for which independence assumption is not verified. Cautious combination of S nondogmatic mass functions m_j , $j = 1, 2, \dots, S$ is defined by the bba with the following weight function:

$$w(A) = \bigwedge_{j=1}^S w_j(A), \quad A \in 2^\Omega \setminus \Omega. \quad (4.34)$$

We thus have

$$m_{\text{Cautious}}(X) = \bigodot_{A \subsetneq \Omega} A^{\bigwedge_{j=1}^S w_j(A)}, \quad (4.35)$$

where $A^{w_j(A)}$ is the simple support function focused on A with weight function $w_j(A)$ issued from the canonical decomposition of m_j . Note also that \wedge is the min operator.

When the dependence/independence of the sources is estimated, another rule was proposed in [3].

4.3.2.3 Conflict Coming from Source's Ignorance

Another possible interpretation of the reason for the conflict is the ignorance of the sources. Indeed, if a source is highly ignorant, it should give a categorical mass function on Ω .

Therefore, [35] interprets the global conflict coming from the ignorance of the sources and transfers the mass on the total ignorance (i.e., on Ω) in order to keep the closed-world assumption. In the case of high conflict, the result of the fusion is the ignorance. This rule is given by:

$$\begin{aligned} m_Y(X) &= m_{\text{Conj}}(X), \forall X \in 2^{\Omega}, X \neq \emptyset, X \neq \Omega \\ m_Y(\Omega) &= m_{\text{Conj}}(\Omega) + m_{\text{Conj}}(\emptyset) \\ m_Y(\emptyset) &= 0. \end{aligned} \quad (4.36)$$

A source can also be ignorant not on all but only on some focal elements. Hence, [9] proposed a clever conflict repartition by transferring the partial conflicts on the partial ignorances. This rule is given for all $X \in 2^{\Omega}, X \neq \emptyset$ by:

$$m_{\text{DP}}(X) = \sum_{Y_1 \cap \dots \cap Y_s = X} \prod_{j=1}^s m_j(Y_j) + \sum_{\substack{Y_1 \cup \dots \cup Y_s = X \\ Y_1 \cap \dots \cap Y_s = \emptyset}} \prod_{j=1}^s m_j(Y_j), \quad (4.37)$$

where $Y_j \in 2^{\Omega}$ is a focal element of the source S_j , and $m_j(Y_j)$ the associated mass function. This rule has a high memory complexity, such as the PCR6 rule, because it is necessary to manage the partial conflict.

4.3.2.4 Conflict Coming from Source Reliability Assumption

If we have no knowledge of the reliability of the sources, but we know that at least one source is reliable, the disjunctive combination can be used. It is given for all $X \in 2^{\Omega}$ by:

$$m_{\text{Dis}}(X) = \sum_{Y_1 \cup \dots \cup Y_s = X} \prod_{j=1}^s m_j(Y_j). \quad (4.38)$$

The main problem of this rule is the loss of specificity after combination.

One can also see the global conflict $\kappa = m_{\text{Conj}}(\emptyset)$ such as an estimation of the conflict coming from the unreliability of the sources. Therefore, the global conflict can play the role of a weight between a conjunctive and disjunctive comportment of the rule such introduced by [12]. This rule is given for $X \in 2^{\Omega}, X \neq \emptyset$ by:

$$m_{\text{Flo}}(X) = \beta_1(\kappa)m_{\text{Dis}}(X) + \beta_2(\kappa)m_{\text{Conj}}(X), \quad (4.39)$$

where β_1 and β_2 can be defined by:

$$\begin{aligned} \beta_1(\kappa) &= \frac{\kappa}{1 - \kappa + \kappa^2}, \\ \beta_2(\kappa) &= \frac{1 - \kappa}{1 - \kappa + \kappa^2}. \end{aligned} \quad (4.40)$$

In a more general way, we propose in [23] to regulate the conjunctive/disjunctive comportment taking into consideration the partial combinations. The mixed rule is given for m_1 and m_2 for all $X \in 2^{\Omega}$ by:

$$m_{\text{Mix}}(X) = \sum_{Y_1 \cup Y_2 = X} \delta_1 m_1(Y_1) m_2(Y_2) + \sum_{Y_1 \cap Y_2 = X} \delta_2 m_1(Y_1) m_2(Y_2). \quad (4.41)$$

If $\delta_1 = \beta_1(\kappa)$ and $\delta_2 = \beta_2(\kappa)$, we obtain the rule of [12]. Likewise, if $\delta_1 = 1 - \delta_2 = 0$, we obtain the conjunctive rule and if $\delta_1 = 1 - \delta_2 = 1$ the disjunctive rule. With $\delta_1(Y_1, Y_2) = 1 - \delta_2(Y_1, Y_2) = \mathbb{1}_{\{\emptyset\}}(Y_1 \cap Y_2)$, we get back to the rule of [9] by taking into account partial conflicts.

The choice of $\delta_1 = 1 - \delta_2$ can also be made from a dissimilarity such as:

$$\delta_2(Y_1, Y_2) = \frac{|Y_1 \cap Y_2|}{\min(|Y_1|, |Y_2|)}, \quad (4.42)$$

where $|Y_1|$ is the cardinality of Y_1 . Jaccard dissimilarity can also be considered by:

$$\delta_2(Y_1, Y_2) = \frac{|Y_1 \cap Y_2|}{|Y_1 \cup Y_2|}. \quad (4.43)$$

Therefore, if we have a partial conflict between Y_1 and Y_2 , $|Y_1 \cap Y_2| = 0$, and the rule transfers the mass on $Y_1 \cup Y_2$. In that case $Y_1 \subset Y_2$ (or the contrary), $Y_1 \cap Y_2 = Y_1$, and $Y_1 \cup Y_2 = Y_2$; hence with δ_2 given by (4.42), the rule transfers the mass on Y_1 and with δ_2 given by (4.43) on Y_1 and Y_2 according to the ratio $|Y_1|/|Y_2|$ of cardinalities. In the case $Y_1 \cap Y_2 \neq Y_1, Y_2$ and \emptyset , the rule transfers the mass on $Y_1 \cap Y_2$ and $Y_1 \cup Y_2$ according to Eqs. (4.42) and (4.43). With such weights, the Mix rule considers partial conflict according to the imprecision of the elements at the origin of the conflict.

4.3.2.5 Conflict Coming from a Number of Sources

When we have to combine a many sources, the assumption of the reliability of all the sources is difficult to consider especially if the sources are human. The disjunctive rule (4.38) assumes that at least one source is reliable but a precise decision will be difficult to take. Moreover, the complexity of main rules managing the conflict in a clever way is too high such as the rules given by (4.41) and (4.31). That is the reason why we introduce in [39] a new rule according to the following assumptions:

- The majority of sources are reliable;
- The larger extent one source is consistent with others, the more reliable the source is;
- The sources are cognitively independent.

For each mass function m_j , we consider the set of mass functions $\{A_k^{w_j}, A_k \subset \Omega\}$ coming from the canonical decomposition. If group the simple mass functions $A_k^{w_j}$ in c clusters (the number of distinct A_k) and denote by s_k the number of simple mass functions in the cluster k , the proposed rule is given by:

$$m_{\text{LNS}} = \bigcirc_{k=1,\dots,c}^{1-\alpha_k+\alpha_k} (A_k) \prod_{j=1}^{s_k} w_j \quad (4.44)$$

where

$$\alpha_k = \frac{s_k}{\sum_{i=1}^c s_i}. \quad (4.45)$$

4.3.2.6 How to Choose the Combination Rule?

To answer the delicate question on which combination rule to choose, many authors propose a new rule. Of course, we could propose a *no free lunch theorem* showing that there is no a best combination rule.

To answer this question, we propose in [20] a global approach to transfer the belief. Indeed, the discounting process, the reduction of the number of focal elements, the combination rules, and the decision process can be seen such as a transfer of belief, and we can define these transfers in joint operator. However, it seems so difficult to propose a global approach which will be too general to be applied. In [20], we define a rule integrating only the reliability given for $X \in 2^\Omega$ by:

$$m(X) = \sum_{\mathbf{Y} \in (2^\Omega)^S} \prod_{j=1}^S m_j(Y_j) w(X, \mathbf{m}(\mathbf{Y}), \mathcal{T}(\mathbf{Y}), \boldsymbol{\alpha}(\mathbf{Y})), \quad (4.46)$$

where $\mathbf{Y} = (Y_1, \dots, Y_S)$ is the response vector of the S sources, $m_j(Y_j)$ the associated masses $(\mathbf{m}(\mathbf{Y}) = (m_1(Y_1), \dots, m_j(Y_s)))$, w is a weight function, $\boldsymbol{\alpha}$ is the matrix of terms α_{ij} of the reliability of the source S_j for the element i of 2^Ω , and $\mathcal{T}(\mathbf{Y})$ is the set of elements of 2^Ω , on which we transfer the elementary masses $m_j(Y_j)$ for a given vector \mathbf{Y} . This rule has been illustrated in a special case integrating the local reliability, but it seems even too complex to be easily applied.

Hence, the best way to choose the combination rule is to identify the assumptions that we can or we have to make and choose the adapted rule according to these assumptions.

However, we know that a given rule can provide good results in a context where the assumptions satisfy this rule. Hence, another way to evaluate and compare some rules of combination is to study the results (after decision) of the combined masses, e.g., on generated mass functions. In [28], from generated mass functions, we study the difference of the combination rules in terms of decisions. We showed that we have to take into account the decision process. We will present some of them in the next section in the context of conflicting mass functions.

4.4 Decision with Conflicting bbas

The classic functions used for decision-making such as the pignistic probability, the credibility, and the plausibility are increasing by sets inclusion. We cannot use these functions directly to decide on other elements than the singletons. When the assumption of exclusiveness of the frame of discernment is not made, such as in Eq. (4.31), we can decide on D_r^Ω . That can be interesting from the data mining point of view such as we show in [24].

The approach proposed by [1] has been extended to the consideration of D_r^Ω in [19] allowing to decide on any element of D_r^Ω by taking into account the mass function and the cardinality. Hence, we choose the element $A \in D_r^\Omega$ for a given observation if:

$$A = \operatorname{argmax}_{X \in D_r^\Omega} (m_d(X) f_d(X)), \quad (4.47)$$

where f_d is the considered decision function such as the credibility, plausibility, or pignistic probability calculated from the mass function coming from the result of the combination rule and m_d is the mass function defined by:

$$m_d(X) = K_d \lambda_X \left(\frac{1}{\mathcal{C}_M(X)^\rho} \right), \quad (4.48)$$

$\mathcal{C}_M(X)$ is the cardinality of X of D_r^Ω , defined by the number of disjoint parts in the Venn diagram, ρ is a parameter with its values in $[0, 1]$ allowing to decide from the intersection of all the singletons (with $\rho = 1$) until the total ignorance Ω (with $\rho = 0$). The parameter λ_X enables us to integrate the loss of knowledge on one of the elements X of D_r^Ω . The constant K_d is a normalization factor that guarantees the condition of Eq. (4.1). Without any constraint on D^Ω , all the focal elements contain the intersection of the singletons. One cannot choose the plausibility such as decision function f_d .

The choice of the parameter ρ is not easy to make. It depends on the size of Ω . According to the application, it can be more interesting to define a subset on which we want to take the decision. Hence, we can envisage the decision on any subset of D_r^Ω , noted \mathcal{D} , and Eq. (4.47) becomes simply:

$$A = \operatorname{argmax}_{X \in \mathcal{D}} (m_d(X) f_d(X)). \quad (4.49)$$

Particularly this subset can be defined according to the expected cardinality of the element on which we want to decide.

With the same spirit, in [11], another decision process is proposed by:

$$A = \operatorname{argmax}_{X \in \mathcal{D}} (d_J(m, m_X)), \quad (4.50)$$

where m_X is the categorical mass function $m(X) = 1$ and m is the mass function coming from the combination rule. The subset \mathcal{D} is the set of elements on which we want to decide.

This last decision process allows also a decision on imprecise elements of the power set 2^{Ω} and to control the precision of expected decision element without any parameter to fit.

4.5 Conclusion

In this chapter, we propose some solutions to the problem of the conflict in information fusion in the context of the theory of belief functions. In Sect. 4.2 we present some conflict measures. Today, there is no consensus in the community on the choice of a conflict measure. Measuring the conflict is not an easy task because a mass function contains some information such as auto-conflict we can interpret differently. The proposed axioms are a minimum that a conflict measure has to reach. In Sect. 4.3, we discuss how to manage the conflict. Based on the assumption that conflict comes from the unreliability of the sources, with a conflict estimation for each source, the best to do is to discount the mass function according to the reliability estimation (and so the conflict measure).

Another way to manage the conflict is the choice of the combination rule. Starting from the famous Zadeh's example, many combination rules have been proposed to manage the conflict. In this chapter, we present some combination rules (without exhaustiveness) according to the assumptions that the rules suppose. Hence, we distinguish the assumptions of open/closed world, dependent/independent sources, ignorant/not ignorant sources, reliable/unreliable sources, and few/many sources.

To end this chapter, when the assumption of exclusiveness of the frame of discernment is not made, and so when we postpone the matter of conflict to the decision, we present some adapted decision processes. These decision methods are also adapted to decide on some imprecise elements of the power set.

Of course, all the exposed methods here must be selected according to the application, to the possible assumptions, and to the final expected result.

References

1. A. Appriou, *Uncertainty Theories and Multisensor Data Fusion*, (ISTE Ltd, UK/Wiley, USA, 2014)
2. M. Chebbah, B. Ben Yaghane, A. Martin, Reliability estimation based on conflict for evidential database enrichment, in *Belief*, Brest (2010)
3. M. Chebbah, A. Martin, B. Ben Yaghane, Combining partially independent belief functions. *Decis. Support Syst.* **73**, 37–46 (2015)
4. L.Z. Chen, W.K. Shi, Y. Deng, Z.F. Zhu, A new fusion approach based on distance of evidences. *J. Zhejiang Univ. Sci.* **6A**(5), 476–482 (2005)

5. A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* **38**, 325–339 (1967)
6. T. Denœux, Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. *Artif. Intell.* **172**, 234–264 (2008)
7. S. Destercke, T. Burger, Revisiting the notion of conflicting belief functions, in *Belief*, Compiègne (2012)
8. J. Dezert, Foundations for a new theory of plausible and paradoxical reasoning. *Inf. Secur. Int. J.* **9**, 13–57 (2002)
9. D. Dubois, H. Prade, Representation and combination of uncertainty with belief functions and possibility measures. *Comput. Intell.* **4**, 244–264 (1988)
10. Z. Elouedi, K. Mellouli, Ph. Smets, Assessing sensor reliability for multisensor data fusion within the transferable belief model. *IEEE Trans. Syst. Man Cybern. B: Cybern.* **34**(1), 782–787 (2004)
11. A. Essaid, A. Martin, G. Smits, B. Ben Yaghlane, A distance-based decision in the credal level, in *International Conference on Artificial Intelligence and Symbolic Computation (AISC 2014)*, Sevilla (2014), pp. 147–156
12. M.C. Florea, A.-L. Joussetme, E. Bossé, D. Grenier, Robust combination rules for evidence theory. *Inf. Fusion* **10**, 183–197 (2009)
13. T. George, N.R. Pal, Quantification of conflict in Dempster-Shafer framework: a new approach. *Int. J. Gen. Syst.* **24**(4), 407–423 (1996)
14. A.-L. Joussetme, P. Maupin, Distances in evidence theory: comprehensive survey and generalizations. *Int. J. Approximate Reason.* **53**(2), 118–145, (2012)
15. A.-L. Joussetme, D. Grenier, E. Bossé, A new distance between two bodies of evidence. *Inf. Fusion* **2**, 91–101 (2001)
16. G.J. Klir, Measures of uncertainty in the Dempster-Shafer theory of evidence, in *Advances in the Dempster-Shafer Theory of Evidence*, ed. by R.R. Yager, M. Fedrizzi, J. Kacprzyk (Wiley, New York, 1994), pp. 35–49
17. W. Liu, Analyzing the degree of conflict among belief functions. *Artif. Intell.* **170**, 909–924 (2006)
18. A. Martin, Comparative study of information fusion methods for sonar images classification, in *International Conference on Information Fusion*, Philadelphia (2005)
19. A. Martin, Implementing general belief function framework with a practical codification for low complexity, in *Advances and Applications of DSmT for Information Fusion*, vol. 3 (American Research Press, Rehoboth, 2009), chapter 7, pp. 217–274
20. A. Martin, Reliability and combination rule in the theory of belief functions, in *International Conference on Information Fusion* (2009), pp. 529–536
21. A. Martin, About conflict in the theory of belief functions, in *Belief*, Compiègne (2012)
22. A. Martin, C. Osswald, Human experts fusion for image classification. *Inf. Secur.: Int. J. Spec. Issue Fusing Uncertain Imprecise Paradoxist Inf. (DSmT)* **20**, 122–143 (2006)
23. A. Martin, C. Osswald, Toward a combination rule to deal with partial conflict and specificity in belief functions theory, in *International Conference on Information Fusion*, Québec (2007)
24. A. Martin, I. Quidu, Decision support with belief functions theory for seabed characterization, in *International Conference on Information Fusion*, Cologne (2008)
25. A. Martin, A.-L. Joussetme, C. Osswald, Conflict measure for the discounting operation on belief functions, in *International Conference on Information Fusion*, Cologne (2008)
26. D. Mercier, B. Quost, T. Denœux, Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Inf. Fusion* **9**, 246–258 (2006)
27. C. Murphy, Combining belief functions when evidence conflicts. *Decis. Support Syst.* **29**, 1–9 (2000)
28. C. Osswald, A. Martin, Understanding the large family of Dempster-Shafer theory's fusion operators – a decision-based measure, in *International Conference on Information Fusion*, Florence (2006)
29. C. Rominger, A. Martin, Using the conflict: an application to sonar image registration, in *Belief*, Brest (2010)

30. G. Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, 1976)
31. F. Smarandache, A. Martin, C. Osswald, Contradiction measures and specificity degrees of basic belief assignments, in *International Conference on Information Fusion*, Boston (2011)
32. Ph. Smets, Constructing the pignistic probability function in a context of uncertainty. Uncertain. Artif. Intell. **5**, 29–39 (1990)
33. Ph. Smets, Analyzing the combination of conflicting belief functions. Inf. Fusion **8**(4), 387–412 (2007)
34. M.J. Wierman, Measuring conflict in evidence theory, in *IFSA World Congress and 20th NAFIPS International Conference*, vol. 3, no. 21 (2001), pp. 1741–1745
35. R.R. Yager, On the Dempster-Shafer framework and new combination rules. Inf. Sci. **41**, 93–137 (1991)
36. R.R. Yager, On considerations of credibility of evidence. Int. J. Approximate Reason. **7**, 45–72 (1992)
37. L.A. Zadeh, A mathematical theory of evidence (book review). AI Mag. **5**, 81–83 (1984)
38. C. Zeng, P. Wu, A reliability discounting strategy based on plausibility function of evidence, in *International Conference on Information Fusion*, Québec (2007)
39. K. Zhou, A. Martin, Q. Pan, Evidence combination for a large number of sources, in *International Conference on Information Fusion*, Xian (2017). The paper was accepted

Chapter 5

Basic Properties for Total Uncertainty Measures in the Theory of Evidence



Joaquín Abellán, Carlos J. Mantas, and Éloi Bossé

Abstract The theory of evidence is a generalization of the probability theory which has been used in many applications. That generalization permits to represent more different types of uncertainty. To quantify the total information uncertainty in theory of evidence, several measures have been proposed in the last decades. From the axiomatic point of view, any uncertainty measure must verify a set of important properties to guarantee a correct behavior. Thus, a total measure in theory of evidence must preserve the total amount of information or not to decrease when uncertainty is increased. In this chapter we review and revise the properties of a total measure of uncertainty considered in the literature.

Keywords Theory of evidence · Dempster-Shafer theory · Uncertainty-based information · Measures of uncertainty · Conflict · Non-specificity

5.1 Introduction

Uncertainty and information are two concepts intricately related to each other as two sides of the same coin. Uncertainty of information is a major dimension of information quality that is paramount to decision quality. Representation of

The work of the first two authors has been supported by the Spanish “Ministerio de Economía y Competitividad” and by “Fondo Europeo de Desarrollo Regional” (FEDER) under Project TEC2015-69496-R.

J. Abellán (✉) · C. J. Mantas

Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

e-mail: jabellan@decsai.ugr.es; cmantas@decsai.ugr.es

É. Bossé

IMT-Atlantique, Brest, France

McMaster University, Hamilton, Canada

e-mail: ebosse@gel.ulaval.ca

uncertainty is a crucial issue in many areas of science and engineering that is necessary for the transformation of information along the processing chain: data to information to knowledge to decisions and actions.

Shannon's entropy [25] has been the tool to quantify uncertainty in the classical theory of probability. That function verifies a set of desirable properties of probability distributions. It is well known that the probabilistic representation of information cannot deal with all different types of uncertainty; thus, an imprecise probability theory can be used (see [27]). Some of these theories are the Dempster-Shafer theory [8, 23], interval-valued probabilities [7], order-2 capacities [6], upper-lower probabilities [11, 26], or general closed and convex sets of probability distributions [12, 22, 27] also called credal sets.

One of the most used imprecise probability models is the Dempster-Shafer theory [8, 23], known as *theory of evidence*, which has been designed as an extension of the classical probability theory. In the theory of evidence, the available information is quantified by so-called basic belief assignment (bba) on a power set *frame of discernment*. This theory has been widely used in many areas.

Shannon's entropy has been used as the starting point for quantifying the uncertainty. It can be justified by different ways, but the most common one is the use of an axiomatic approach that consists of assuming a set of required basic properties that a measure of uncertainty should verify.

The analysis of the types of uncertainty that a bpa represents is an important aspect of the theory of evidence. With this theory, more types of uncertainty can be represented using a bpa than by a probability distribution in the probability theory [20]. In the theory of evidence, Yager [28] makes the distinction between two types of uncertainty called *discord* (or *randomness* or *conflict*) and *non-specificity*. The first one has been related to entropy and the second one to imprecision.

Klir and Wierman [20] presented a total uncertainty (TU) measure in the theory of evidence that has been justified by an axiomatic approach similar to the one in probability theory. They also attach to that TU, a set of five desired properties that TU must verify. Abellán and Masegosa [3] extended that set to add the important property of monotonicity as well as other behavioral properties related to TU.

The reminder of this chapter is structured as follows. In Sect. 5.2, we will introduce some necessary basic concepts and notations and some of the most important measures of uncertainty presented in the theory of evidence. Section 5.3 is devoted to the description of a set of the axiomatic properties necessary for total uncertainty measures in the theory of evidence and other important requirements for this type of measures. Section 5.4 presents conclusions.

5.2 Information Representation in Theory of Evidence

Let X be a finite set, considered as a set of possible situations, $|X| = n$, $\wp(X)$ the power set of X and x any element of X . The Dempster-Shafer theory (DST) of evidence (Dempster [8], Shafer [23]) is based on the concept of basic probability

assignment. A *basic probability assignment* (bpa), also called *mass assignment*, is a mapping $m : \wp(X) \rightarrow [0, 1]$, such that $\mathbf{m}(\emptyset) = 0$ and $\sum_{A \subseteq X} m(A) = 1$. A set A such that $m(A) > 0$ is called a *focal element* of m .

Let X, Y be finite sets, their product space $X \times Y$, and m a bpa on $X \times Y$. The marginal bpa on X , $m^{\downarrow X}$ (and similarly on Y , $m^{\downarrow Y}$), is defined in the following way:

$$m^{\downarrow X}(A) = \sum_{R|A=R^{\downarrow X}} m(R), \quad \forall A \subseteq X \quad (5.1)$$

where $R^{\downarrow X}$ is the set projection of R on X .

There are two functions associated with each basic probability assignment, a belief function, Bel , and a plausibility function, Pl : $Bel(A) = \sum_{B \subseteq A} m(B)$, $Pl(A) = \sum_{A \cap B \neq \emptyset} m(B)$. They can be seen as the lower and upper probability of A , respectively.

We may note that belief and plausibility are interrelated for all $A \in \wp(X)$: $Pl(A) = 1 - Bel(A^c)$, where A^c denotes the complement of A . Furthermore, $Bel(A) \leq Pl(A)$.

For each bpa on a finite set X , there exists a set of associated probability distributions p on X , of the following way:

$$K_m = \{p \mid Bel(A) \leq p(A), \quad \forall A \in \wp(X)\} \quad (5.2)$$

We note that $Bel(A) \leq p(A) \leq Pl(A)$. K_m is a closed and convex set of probability distributions, also called a credal set in the literature.

5.2.1 Measures of Uncertainty in DST

The classical measure of entropy (Shannon [25]) in probability theory is defined by the following continuous function:

$$S(p) = - \sum_{x \in X} p(x) \log_2(p(x)), \quad (5.3)$$

where $p(x)$ is the probability of value x . and \log_2 is normally used to quantify the value in bits, though \log and \log_2 are used in the literature interchangeably. The value $S(p)$ quantifies the only type of uncertainty presented of probability theory, and it verifies a large set of desirable properties (Shannon [25], Klir and Wierman [20]).

As it was mentioned above, Yager [28] makes the distinction between two types of uncertainty in the Dempster-Shafer theory. One is associated with cases where the information is focused on sets with empty intersections and the other one is

associated with cases where the information is focused on sets with cardinality greater than one. They are called *conflict* (or *randomness* or *discord*) and *non-specificity*, respectively.

The following function (5.4), introduced by Dubois and Prade [10], has its origin in classical Hartley measure (Hartley [15]) considered in classical set theory and in the extended Hartley measure, in possibility theory (Higashi and Klir [16]). It represents a measure of non-specificity associated with a bpa. It is expressed as follows:

$$I(m) = \sum_{A \subseteq X} m(A) \log(|A|). \quad (5.4)$$

$I(m)$ achieves its minimum, zero, when m is a probability distribution. The maximum, $\log(|X|)$, is obtained with $m(X) = 1$ and $m(A) = 0$, $\forall A \subset X$. It is showed in the literature that I verifies all the required properties for such a type of measure.

Many measures were introduced to quantify the conflict by utilizing a bpa (Klir and Wierman [20]). One of the most representative conflict functions was introduced by Yager [28]:

$$E(m) = - \sum_{A \subseteq X} m(A) \log Pl(A). \quad (5.5)$$

At the same time, this function does not verify all the required properties.

Harmanc and Klir [13, 14] proposed $S^*(m)$, a measure equal to the maximum of the entropy (upper entropy) of the probability distributions verifying $Bel(A) \leq \sum_{x \in A} p(x) \leq Pl(A)$, $\forall A \subseteq X$. This set of probability distributions is the credal set associated with a bpa, m , that we have noted as K_m in Eq. (5.2).

Harmanc and Klir [14] considered a function S^* as a total uncertainty measure in the Dempster-Shafer theory, i.e., as a measure that quantifies both conflict and non-specificity, but they do not separate both parts. Abellán, Klir, and Moral [5] have proposed upper entropy as an aggregate measure for more general theories than DST, separating coherently conflict and non-specificity. These parts can be also obtained in the Dempster-Shafer theory in a similar way.

$$S^*(m) = S_*(m) + (S^* - S_*)(m), \quad (5.6)$$

where $S^*(m)$ represents maximum entropy and $S_*(m)$ represents minimum entropy on the credal set K_m with a bpa m , while $S_*(m)$ coherently quantifying the conflict part and $(S^* - S_*)(m)$ its non-specificity part. That measure (5.6) has been successfully used in applications (see Abellán and Moral [4]).

Maeda and Ichihashi [21] proposed a total uncertainty measure on DST adding up Hartley generalized measure and upper entropy:

$$MI(m) = S^*(m) + I(m) \quad (5.7)$$

Jousselme et al. [17] presented a measure to quantify ambiguity (discord or conflict and non-specificity) in the Dempster-Shafer theory, i.e., a TU measure. This TU measure is based on the pignistic probability. Let m be a bpa on a finite set X , then the pignistic probability $Bet P_m$, of all the subsets A in X , is defined by:

$$Bet P_m(A) = \sum_{B \subseteq X} m(B) \frac{|A \cap B|}{|B|}. \quad (5.8)$$

For a singleton set $A = \{x\}$, we have $Bet P_m(\{x\}) = \sum_{x \in B} [m(B)/|B|]$. Hence, an ambiguity measure (AM) for a bpa m on a finite set X is defined as:

$$AM(m) = \sum_{x \in X} Bet P_m(x) \log(Bet P_m(x)), \quad (5.9)$$

i.e., the entropy of the $Bet P_m$ probability.

Recently, Shahpari and Seyedin [24] have presented a modification of AM, called MAM, to avoid the AM drawbacks identified in Klir and Lewis [19]. The function MAM uses a modified pignistic transformation:

$$MAM(m) = - \sum_{x \in X} MBet P_m(x) - \log(MBet P_m(x)), \quad (5.10)$$

Nevertheless, MAM presents also some mathematical shortcomings discussed in details in [2]) and briefly summarized below.

Shahpari and Seyedin showed that MAM coincides with AM on one-dimensional space, but in the case of two-dimensional space, they used a different definition for the pignistic distribution without providing the essential justification. Specifically, let X, Y be finite sets and m a b.p.a. on $X \times Y$. On $X \times Y$, MAM corresponds to the AM function, while using the probability distribution $MBet_{m_{XY}} = Bet_{m_{XY}}$ (MAM is the entropy of that probability distribution). In the case of two-dimensional space, Shahpari and Seyedin use the following function on each marginal b.p.a.:

$$MBet_{m_X}(x_i) = \sum_{B \in \wp(X), B \ni x_i} \sum_{A \in \wp(X \times Y), B = A \downarrow X} \frac{m_{XY}(A) \sharp(x_i \in A)}{|A|}, \forall x_i \in X \quad (5.11)$$

where $\sharp(x_i \in A)$ is the number of appearances of x_i in the set A and $|A|$ is the cardinality of A . Similarly they define the values $MBet_{m_Y}(y_i)$, $\forall y_i \in Y$.

To simplify, this expression can be reduced to the following one:

$$MBet_{m_X}(x_i) = \sum_{A \in \wp(X \times Y), A \downarrow X \ni x_i} \frac{m_{XY}(A) \sharp(x_i \in A)}{|A|}, \forall x_i \in X \quad (5.12)$$

Very recently, Deng [9] have presented a new uncertainty measure named *Deng entropy* that can be considered as a new composed measure, quantifying conflict and non-specificity.

This function, called E_d is defined as follow, for a bpa m on a finite set X :

$$E_d(m) = - \sum_{A \subseteq X} m(A) \log_2 \frac{m(A)}{2^{|A|} - 1}, \quad (5.13)$$

where it can be separated in two functions measuring the both types of uncertainty in DST:

$$E_d(m) = - \sum_{A \subseteq X} m(A) \log_2 m(A) + \sum_{A \subseteq X} m(A) \log_2 (2^{|A|} - 1), \quad (5.14)$$

where the first term quantifies the part of conflict and the second one the part of non-specificity of a bpa.

The E_d measure has been introduced to give more importance to the increase in uncertainty when the number of alternatives increases, i.e., on the non-specificity part. It is not agreed with the standard bounds of values for such type of measures. It can be observed that the upper bound for the part of conflict can be notably smaller than the one for the non-specificity part.

5.3 Basic Properties of Total Uncertainty Measures in TE with Discussion

We can find in Klir and Wierman [20] and in Klir [18] five requirements for a total uncertainty measure (TU) defined in the theory of evidence, i.e., for a measure capturing both conflict and non-specificity: *probabilistic consistency*, *set consistency*, *range*, *additivity*, and *subadditivity*. Recently, the property of *monotonicity* was added by Abellán and Masegosa [3].

As it has been mentioned in the previous section, the requirements of properties for a TU in evidence theory are based on the properties (listed as P1 to P5 below) verified for Shannon's entropy in probability theory and in classic set theory. Due to this fact, Klir and Wierman [20] extend the verification of those properties to the Dempster-Shafer theory. Properties P1 and P2 below assure us that when we are working in a more general framework such as DST and use the definitions of probability and classical set theories, then we end up with the classical uncertainty measures: Shannon entropy and the Hartley measure.

- (P1) **Probabilistic consistency** When all the focal elements of a bpa m are singletons, then the total uncertainty measure must be equal to the Shannon entropy: $TU(m) = \sum_{x \in X} m(x) \log m(x)$.
- (P2) **Set consistency** When there exist a set A such that $m(A) = 1$, then TU must collapse to the Hartley measure: $TU(m) = \log |A|$.

- (P3) **Range** The range of $TU(m)$ is $[0, \log |X|]$.
- (P4) **Subadditivity** Let m be a bpa on the space $X \times Y$ and m_X and m_Y its marginal bpas on X and Y , respectively; then TU must satisfy the following inequality: $TU(m) \leq TU(m_X) + TU(m_Y)$.
- (P5) **Additivity** Let m be a bpa on the space $X \times Y$ and m_X and m_Y their marginal bpas on X and Y , respectively. Let also assume that their marginals are not interactive ($m(A \times B) = m_X(A)m_Y(B)$, with $A \subseteq X$, $B \subseteq Y$ and $m(C) = 0$ if $C \neq A \times B$), then TU must satisfy the equality: $TU(m) = TU(m_X) + TU(m_Y)$.

The additivity and subadditivity properties guaranty that the total information is preserved. In the first property, it states that we do not add information in situations where a decomposition of the problem can be done, i.e., that decomposition should not imply an increase in information. The second one states that the total information obtained from two independent sources is preserved; when we join two independent problems, the total information is preserved. These two properties they are compatible with what happens in PT, where Shannon's entropy performs perfectly.

In the Dempster-Shafer theory, the information of a bpa can be contained by the information of another bpa. This situation allows us to consider the following property of Abellán and Masegosa [3]:

- (P6) **Monotonicity** when uncertainty increases or decreases, the measure of uncertainty must follow accordingly. Formally, let 2 bpas be on a finite set X , m_1 and m_2 , verifying that $K_{m_1} \subseteq K_{m_2}$, then: $TU(m_1) \leq TU(m_2)$.

The following example illustrates the need of this property.

Example 1 Let us first consider three pieces of evidence (e_1 , e_2 , and e_3) about the type of disease (d_1 , d_2 , or d_3) of a patient. To quantify the information available via a basic probability assignment, an expert uses the following bpa on the universal $X = \{d_1, d_2, d_3\}$:

$$e_1 \longrightarrow m_1(\{d_1, d_2\}) = 1/3,$$

$$e_2 \longrightarrow m_1(\{d_1, d_3\}) = 1/2,$$

$$e_3 \longrightarrow m_1(\{d_2, d_3\}) = 1/6.$$

Assume now that the expert find that the reasons not to consider d_3 in e_1 are false and that it is necessary to change his b.p.a to the following one:

$$e_1 \longrightarrow m_2(\{d_1, d_2, d_3\}) = 1/3,$$

$$e_2 \longrightarrow m_2(\{d_1, d_3\}) = 1/2,$$

$$e_3 \longrightarrow m_2(\{d_2, d_3\}) = 1/6.$$

In the example, we go from a first situation with a quantity of information, based on three pieces of evidence, to another more confused situation. It is logical that the second situation involves a greater level of uncertainty (less information). We have $Bel_2(A) \leq Bel_1(A)$ and $Pl_1(A) \leq Pl_2(A), \forall A \subseteq X$. Also, $K_{m_1} \subseteq K_{m_2}$, if K_{m_1} and K_{m_2} are the credal sets associated with m_1 and m_2 respectively.

It can be checked (see [1–3, 19]) that the MI , S^* , AM , MAM , and E_d functions verify the following sets of properties in DST:

MI : P1, P2, P4, P5, and P6.

S^* : P1, P2, P3, P4, P5, and P6.

AM : P1, P2, P3, and P5.

MAM : P1, P2, P3, and P5.

E_d : P1.

Hence, we see that as far only S^* satisfies all the proposed requirements.

5.3.1 Additional Requirements for Properties of Total Uncertainty Measures in the Theory of Evidence

In the paper of Abellán and Masegosa [3], other considerations about the behavior of a total uncertainty measure (TU) in evidence theory were considered. The examples of those additional requirements are:

- (A1) A TU should not have a too complex calculation.
- (A2) A TU must incorporate the two types of uncertainty coexisting in the evidence theory: conflict and non-specificity.
- (A3) A TU must be sensitive to changes of evidence, directly in its corresponding parts of conflict and non-specificity.
- (A4) The extension of a TU in the Dempster-Shafer theory to more general theories such as credal sets must be possible.

The evaluation of the significance of these requirements as well as the six (P1 to P6) others discussed in the previous sections in operational settings is left for future research.

5.4 Conclusions

This chapter has discussed one of the most important information quality characteristics: the uncertainty; its representation, and properties considered in the framework of the Dempster-Shafer theory of evidence. Specifically we have considered an axiomatic set of properties of a total uncertainty (TU) measure that ensure a mathematically consistent behavior when framed within the theory of evidence.

Five of them were introduced two decades ago; the last one, monotonicity, is much more recent. The properties intend to ensure that, in the theory of evidence, (i) the information is measured; (ii) the information is preserved when the union of independent systems is considered or that it does not increase when a disintegration of the original system occurs; and (iii) the uncertainty is measured. A measure of uncertainty that cannot meet the set of critical requirements (e.g., monotonicity, additivity, etc.) has a major drawback for its exploitation in operational contexts such as in analytics, information fusion, and decision support.

References

1. J. Abellán, Analyzing properties of Deng entropy in the theory of evidence. *Chaos Solitons Fractals* **95**, 195–199 (2017)
2. J. Abellán, E. Bossé, Drawbacks of uncertainty measures based on the pignistic transformation. *IEEE Trans. Syst. Man Cybern.: Syst.* **48**, 382–388 (2018)
3. J. Abellán, A. Masegosa, Requirements for total uncertainty measures in Dempster-Shafer theory of evidence. *Int. J. Gen. Syst.* **37**(6), 733–747 (2008)
4. J. Abellán, S. Moral, Upper entropy of credal sets. Applications to credal classification. *Int. J. Approximate Reason.* **39**, 235–255 (2005)
5. J. Abellán, G.J. Klir, S. Moral, Disaggregated total uncertainty measure for credal sets. *Int. J. Gen. Syst.* **35**(1), 29–44 (2006)
6. G. Choquet, Théorie des Capacités. *Ann. Inst. Fourier* **5**, 131–292 (1953/1954)
7. L.M. de Campos, J.F. Huete, S. Moral, Probability intervals: a tool for uncertainty reasoning. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2**, 167–196 (1994)
8. A.P. Dempster, Upper and lower probabilities induced by a multivaluated mapping. *Ann. Math. Stat.* **38**, 325–339 (1967)
9. Y. Deng, Deng entropy. *Chaos Solitons Fractals* **91**, 549–553 (2016)
10. D. Dubois, H. Prade, A note on measure of specificity for fuzzy sets. *BUSEFAL* **19**, 83–89 (1984)
11. T.L. Fine, Foundations of probability, in *Basics Problems in Methodology and Linguistics*, ed. by R.E. Butts, J. Hintikka (Reidel, Dordrecht, 1983), pp. 105–119
12. I.J. Good, Subjective probability as the measure of a non-measurable set, in *Login, Methodology and Philosophy of Science*, ed. by E. Nagel, P. Suppes, A. Tarski (Stanford University Press, California, 1962), pp. 319–329
13. D. Harmanec, G.J. Klir, Measuring total uncertainty in Dempster-Shafer theory: a novel approach. *Int. J. Gen. Syst.* **22**, 405–419 (1994)
14. D. Harmanec, G.J. Klir, Principle of uncertainty revisited, in *Proceedings of 4th Intern, Fuzzy Systems and Intelligent Control Conference*, Maui (1996), pp. 331–339
15. R.V.L. Hartley, Transmission of information. *Bell Syst. Tech. J.* **7**, 535–563 (1928)
16. M. Higashi, G.J. Klir, Measures of uncertainty and information based on possibility distributions. *Int. J. Gen. Syst.* **9**, 43–58 (1983)
17. A.L. Jousselme, C. Liu, D. Grenier, E. Bossé, Measuring ambiguity in the evidence theory. *IEEE Trans. Syst. Man Cybern.-Part A: Syst. Hum.* **36**(5), 890–903 (2006)
18. G.J. Klir, *Uncertainty and Information: Foundations of Generalized Information Theory* (Wiley, Hoboken, 2006)
19. G.J. Klir, H.W. Lewis, Remarks on “Measuring ambiguity in the evidence theory”. *IEEE Trans. Syst. Man Cybern.-Part A: Syst. Hum.* **38**(4), 995–999 (2008)
20. G.J. Klir, M.J. Wierman, *Uncertainty-Based Information* (Phisica-Verlag, Heidelberg/New York, 1998)

21. Y. Maeda, H. Ichihashi, A uncertainty measure with monotonicity under the random set inclusion. *Int. J. Gen. Syst.* **21**, 379–392 (1993).
22. I. Levi, *The Enterprise of Knowledge* (NIT Press, London, 1980)
23. G. Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, 1976)
24. A. Shahpari, S.A. Seyedin, Using mutual aggregate uncertainty measures in a threat assessment problem constructed by Dempster-Shafer network. *IEEE Trans. Syst. Man Cybern. Part A* **45**(6), 877–886 (2015)
25. C.E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423, 623–656 (1948)
26. P. Suppes, The measurement of belief (with discussion). *J. R. Stat. Soc. B* **36**, 160–191 (1974)
27. P. Walley, *Statistical Reasoning with Imprecise Probabilities* (Chapman and Hall, New York, 1991)
28. R.R. Yager, Entropy and specificity in a mathematical theory of evidence. *Int. J. Gen. Syst.* **9**, 249–260 (1983)

Chapter 6

Uncertainty Characterization and Fusion of Information from Unreliable Sources



Lance Kaplan and Murat Sensoy

Abstract Intelligent systems collect information from various sources to support their decision-making. However, misleading information may lead to wrong decisions with significant losses. Therefore, it is crucial to develop mechanisms that will make such systems immune to misleading information. This chapter presents a framework to exploit reports from possibly unreliable sources to generate fused information, i.e., an estimate of the ground truth, and characterize the uncertainty of that estimate as a facet of the quality of the information. First, the basic mechanisms to estimate the reliability of the sources and appropriately fuse the information are reviewed when using personal observations of the decision-maker and known types of source behaviors. Then, we propose new mechanisms for the decision-maker to establish fused information and its quality when it does not have personal observations and knowledge about source behaviors.

Keywords Subjective logic · Unreliable sources · Fusion of information · Quality of information · Uncertainty · Beliefs

6.1 Introduction

Decision-making requires the weighing of risk and benefits in light of uncertain information. While doing so, it is important to estimate the state of the world at sufficient certainty. For a specific decision-making task, this may boil down to estimating the values or a distribution of values for a number of state variables.

L. Kaplan (✉)

RDRL-SES-A, US Army Research Laboratory, Adelphi, MD, USA
e-mail: lkaplan@ieee.org; lance.m.kaplan.civ@mail.mil

M. Sensoy

Computer Science, Özyegin University, Istanbul, Turkey
e-mail: murat.sensoy@ozyegin.edu.tr

Let us consider an intelligent agent that needs to solicit help from a person in a specific organization. Instead of asking a random person in the organization, the agent should pick a person with a high probability of accepting and fulfilling the help request. Hence, for a given person, e.g., Joe, in the organization, the agent can compute a probability distribution over possible outcomes of the request. That is, in response to the help request, Joe may help, do nothing, or undermine. These three outcomes are possible values of the state variable about Joe’s behavior in response to the help request.

In this work, we adopt subjective logic [10], where opinions describe state variables. A state variable can take values from a domain. For instance, a state variable about Joe’s response when help is requested can take three values: help, do nothing, or undermine. Each of these values may lead to a binary proposition, such as “Joe helps when requested,” which can be either *true* or *false*. Instantiations of these propositions are observed and used to create opinions about Joe’s helpfulness.

The decision-maker may have past history of the number of times Joe helped, did nothing, and undermined the organization’s effort to form an opinion about Joe. From this history, the decision-maker can understand and account for the probability of Joe’s behavior for the upcoming mission. The more instances of Joe’s past behavior, the more certain the decision-maker is about these probabilities. In many cases, the uncertainty about Joe is too high to make a decision, and if time permits, the decision-maker should seek out more information about Joe.

In understanding Joe’s tendencies, the decision-maker may have limited experience with Joe and will need to seek reports from other sources about Joe. These sources may or may not provide truthful reports about their experiences with Joe. As a result, the fusion of these reports can lead to wrong probabilities describing Joe’s tendencies when help is requested. Furthermore, the decision-maker can become overconfident about these probabilities and make a poor decision.

To overcome these difficulties, the decision-maker needs to develop a trust behavior profile for its reporting sources to estimate how *trustful* and useful their reports are. Then, the decision-maker needs to properly fuse the reports in light of these profiles. It is desirable for the fused opinion about Joe to consistently represent an estimate of the ground truth probabilities of Joe’s tendencies and the uncertainty about these probabilities. In this work, the fused opinion is represented as an effective number of observations for which each value of a state variable, e.g., Joe’s action to help or not, is instantiated. The effective total number of observations represents the accuracy as a facet of the quality of the information, and it should relate to how close “on average” the estimated probabilities are to the ground truth values.

The development of the trust behavior profiles for the sources is updated as the decision-maker incorporates source reports about instantiations of different state variables. When the decision-maker has its own (limited) observations to form an initial opinion about the values of the state variable, it can leverage the consistency of its own opinion with a particular source’s report to update the source’s trust profile. In essence, the decision-maker is also its own ego-source. This chapter will review our recent research in trust estimation and fusion when the ego-source is available [15, 26, 27].

In many cases, the decision-maker will not be able to make observations about the values of state variables. This chapter will look at extensions of our previous work for this circumstance. Specifically, the conditions when trust estimation and fusion lead to and do not lead to information with a consistent quality of information characterization will be exposed.

This chapter is organized as follows. Section 6.2 reviews related work, and Sect. 6.3 provides the mathematical foundation to represent subjective opinions to represent distributions for the values of state variables. The trust estimation and fusion problem and corresponding models are presented in Sect. 6.4. Section 6.5 reviews recent solutions and demonstrates their effectiveness when an ego-source is available, and Sect. 6.6 extends the solutions for cases when the ego-source is unavailable. This section also demonstrates the effectiveness of the newly extended solutions. Finally, a discussion of results with concluding remarks is provided in Sect. 6.7.

6.2 Related Work

Fusing uncertain information from unreliable sources has drawn significant attention from the literature. It still stands as an important research problem with wide range of applications in many different domains [10]. There are a number of mathematical frameworks for modeling uncertainty and fusing uncertain information. One prominent example of such frameworks is the evidential theory proposed by Dempster and Shafer [29], where belief masses are assigned to possible outcomes of a proposition, i.e., subsets of a frame of discernment. There have been other approaches inspired from the work of Dempster and Shafer. Jøsang proposed subjective logic (SL), which is a probabilistic logic that explicitly takes uncertainty and belief ownership into account. It is used to model and reason with situations that involve uncertainty and incomplete knowledge. A subjective opinion represents assignment of belief masses to possible values of a state variable, and various logical/analytical operators are used to define a calculus over subjective opinions [7–10, 12, 13]. Each subjective opinion can be represented as a Dirichlet distribution over the values of a state variable, and operators defined over these opinions are performed over the underlying Dirichlet distributions. The statistical underpinning of SL makes it flexible and versatile for many domains and applications. For instance, Liu et al. used SL to compute reputation models of mobile ad hoc networks [18]. Oren et al. proposed to use SL to enhance argumentation frameworks with evidential reasoning [21]. Han et al. used SL for forensic reasoning over the surveillance metadata [5]. Sensoy et al. used it for determination of conflicts in and fusion of information from unreliable sources [25]. In this chapter, we also use Dirichlet distributions to represent and combine subjective opinions from unreliable sources.

Fusion of information from unreliable has been studied in the literature with different scenarios and assumptions. As a result of the rise of Internet, e-commerce

has been embraced by users. However, users do not only buy and sell on the Internet but also share their opinions, ratings, and experiences, e.g., through review sites. Therefore, initial work on the fusion of uncertain opinions focuses on propositions about the service quality of online vendors or service providers. A proposition is simply a state variable which can take only two values: *true* or *false*. However, these opinions are collected from unreliable sources, which may aim to mislead the decision-makers, e.g., online buyers. In these works, subjective opinions are usually represented as pairs of positive and negative number of interactions (or experiences) with the service providers (or vendor). Jøsang and Ismail proposed beta reputation systems (BRS) [11], where opinions about a proposition x such as “Bob provides good services” is modeled using beta distributions. Let p_x represent the probability that the proposition is *true*. A beta distribution is used to model the likelihood of each p_x value. Initially, before having any experience with Bob, the beta distribution is represented by parameters $\langle 1, 1 \rangle$, which corresponds to the uniform distribution. This means that p_x can be anything between 0 and 1 with equal probability. However, after having r good and s bad experiences with Bob, the beta distribution parameters are updated as $\langle r + 1, s + 1 \rangle$ using Bayesian update. In BRS, opinions about Bob are collected from a number of information sources, and these opinions are fused using Bayesian update, i.e., evidence aggregation. However, some malicious sources may disseminate misleading opinions.

Whitby et al. extended BRS to filter out misleading opinions provided by the malicious sources. This approach filters out those opinions that do not comply with the significant majority by using an *iterated filtering approach* [37]. Hence, this approach assumes that the majority of sources honestly share their opinions, i.e., liars are in the minority. The extended BRS does not assume that the decision-maker can use its observations to estimate the reliability of information sources. Because BRS is a simple trust-based fusion approach, it has been used in many domains, such as wireless sensor networks [6]. Bui et al. have proposed to use it to estimate trust in sensor readings in body area sensor networks [1]. Ganeriwal et al. proposed a reputation framework for high integrity sensor networks based on the BRS [4].

To avoid the need to rely on a majority of sources to be honest, some existing work assumes an ego-agent, i.e., a decision-maker may observe evidence about the ground truth using its own sensors. Hence, an ego-agent can evaluate the information sources by comparing its own observations against those reported by these sources. TRAVOS [31] is one such information fusion framework, which is similar to BRS in terms of representation and fusion of subjective opinions. However, TRAVOS keeps a history of opinions from information sources about propositions, such as the aforementioned proposition about Bob’s services. To measure the trustworthiness of a source, the decision-maker compares the source and ego opinions over multiple propositions to determine a beta distribution to describe the trust in each source.

Bayesian modeling has also been used to address fusion of subjective information from malicious sources. Regan et al. proposed BLADE [23] for reputation modeling of sources and fusion of their ratings in e-marketplaces. This model learns parameters of a Bayesian network to fuse subjective and possibly deceptive

information from unreliable sources with varying behavior. Most of the existing Bayesian approaches require at least some of the information sources to consistently share honest opinions. These approaches build models for the trustworthiness of information sources and exploit them while fusing their opinions. However, there are approaches, where fused opinions do not directly rely on the opinions from sources. For example, Teacy et al. proposed HABIT, which uses hierarchical Bayesian modeling for fusion of opinions from unreliable sources [32]. It does not directly estimate trustworthiness of information sources. Instead, it uses opinions from sources to measure similarity of the current proposition to past propositions. Then, using the computed similarities as weights, the fused opinion for the current proposition is computed as the weighted average of the decision-maker's opinions about the past propositions. This approach is robust to malicious behaviors, but it requires decision-maker to have accurate opinions for the past propositions similar to the current proposition.

Fact finders address the scenarios where the decision-maker cannot directly observe evidence about the ground truth and the sources only provide absolute claims. They try to identify truth among many conflicting claims without any prior knowledge or observation about the trustworthiness of the information sources. Unlike the previously mentioned approaches, fact-finding approaches assume that the truth is crisp and certain. That is, for a given state variable, only one of k mutually exclusive values can be true. Instantiation of the state variable with each of these values is called a claim. TruthFinder [39] defines trustworthiness of sources as a function of the confidences of the their claims and, conversely, defines the claim confidence as a function of the trustworthiness of the sources espousing them. Then, it iterates by calculating the confidence from the trustworthiness and vice versa. Pasternack and Roth [22] generalizes fact-finding by incorporating source-claim weights in the iterative equations to represent a degree of uncertainty in the observations of claims or in the belief of the sources in the claims.

Recently, Wang et al. formalized fact-finding as a maximum likelihood problem where the expectation-maximization (EM) algorithm [19] is used to estimate the reliabilities of the claims and the users at the same time iteratively [35]. This approach enables the formulation of Cramer-Rao bounds to establish the quality of the estimated reliabilities in terms of the structure of the source-claim network [34]. Furthermore, this approach has been applied for social sensing by estimating the reliability of information from the crowd for sensing situations and events. Specifically, the data from micro-blogging sites such as Twitter¹ has been used to detect social and environmental events earlier than traditional means [36]. The EM approach is further extended in [33] to incorporate the confidence of the users by incorporating weights into the iterative equations similar to [22].

In this work, we aim to exploit behaviors of unreliable sources while fusing their uncertain and possibly misleading opinions. In the literature, different types of information source behaviors are defined and studied [3, 17, 24, 40]. Yu and

¹<http://www.twitter.com>

Singh defined four major types of source behaviors over binomial subjective opinions: *honest*, *complementary*, *exaggerated positive*, and *exaggerated negative* [40]. Sources with *honest* behavior share their genuine opinion; on the other hand, the sources adopting non-honest behaviors transform their opinions before sharing. Sources with *complementary* behavior share the opposite of their genuine opinions, i.e., flipping its true opinion. A source with *exaggerated positive* behavior shares an opinion that is more optimistic than its genuine opinion. Similarly, a source with *exaggerated negative* behavior shares an opinion that is more pessimistic than its genuine opinion. The deception models of Yu and Singh received significant amount of attention from the literature. These models are also used in different domains and disciplines. Fung and Boutaba used these deception models for collaborative intrusion detection in networks [3]. In this setting, peers send feedback about the risk levels of a security alert to others.

The honest, complementary, and exaggeration behaviors require information source to know the truth about the state variable in question. However, an information source may still deceive the information requester without knowing the actual truth. In the *Encyclopedia of Deception* [17], fabrication is defined as another type of deception. In the case of fabrication, someone submits statements as truth, without knowing for certain whether or not it actually is true. Therefore, if a source makes up and shares an opinion without actually having any evidence about the proposition in question, then it would be fabricating. This kind of behavior is similar to randomly generating and sharing an opinion when requested.

6.3 Mathematical Preliminaries

A state variable is a random variable that takes on one value from a mutually exclusive set \mathbb{K} at each instantiation. There is a ground truth probability for each possible value to materialize. Given the observations that n^k instantiations of the variable are of value k for all $k \in \mathbb{K}$ are the result of sampling a multinomial distribution, the posterior knowledge about the distribution of the generating probability is the Dirichlet distribution:

$$f_{\beta}(\mathbf{p}|\mathbf{n}) = \frac{1}{B(\mathbf{n} + 1)} \prod_{k \in \mathbb{K}} \left(p^k \right)^{n^k}, \quad (6.1)$$

where

$$B(\mathbf{n} + 1) = \frac{\prod_{k \in \mathbb{K}} \Gamma(n^k + 1)}{\Gamma(\sum_{k \in \mathbb{K}} (n^k + 1))} \quad (6.2)$$

is the beta function and $\Gamma(\cdot)$ is the gamma function [16]. Throughout this chapter, the boldfaced variables are $|\mathbb{K}|$ dimensional vectors where their elements are non-

bold with a superscript representing the corresponding value in \mathbb{K} . Note that in (6.1), the probabilities are constrained to sum to one, i.e., $\sum_{k \in \mathbb{K}} p^k = 1$.

Subjective logic [10] connects the evidence \mathbf{n} to belief mass assignments as used in belief theories for reasoning under uncertainty such as Dempster-Shafer theory [29] and more recently the transferable belief model [30]. Specifically, the connection between the evidence \mathbf{n} and the beliefs (\mathbf{b}, u) is given by the following invertible mapping:

$$b^k = \frac{n^k}{W + \sum_{k \in \mathbb{K}} n^k} \quad \forall k \in \mathbb{K} \quad \text{and} \quad u = \frac{W}{W + \sum_{k \in \mathbb{K}} n^k}, \quad (6.3)$$

where the b^k 's are the beliefs for each value of the state variable and u is the remaining uncertainty. The beliefs and uncertainty are constrained to be nonnegative and sum to one. In (6.3), W is the prior weight, In this chapter, we set $W = |\mathbb{K}|$ and consider the uninformative uniform prior. The connection between beliefs and the Dirichlet distribution helps to define many of the operators in subjective logic, which distinguishes it from the prior belief theories by connecting it to second-order Bayesian reasoning.

It is well known that the expected value for the probabilities of the Dirichlet distribution is given by

$$m^k = \frac{n_k + 1}{\sum_{k' \in \mathbb{K}} (n^{k'} + 1)}, \quad (6.4)$$

and the variance is

$$\sigma^{2k} = \frac{m^k(1 - m^k)}{1 + \sum_{k' \in \mathbb{K}} (n^{k'} + 1)} \quad (6.5)$$

for $k \in \mathbb{K}$. In the context that an opinion about a state variable is given by \mathbf{n} , the mean given by (6.4) represents the information about, i.e., estimation of, the ground truth probabilities. Likewise, the variance given by (6.5) represents the derived quality of information. The smaller the variance, the higher the quality of the information. Note that the quality of information is proportional to the sum of evidences, i.e., $\sum_{k \in \mathbb{K}} n^k$. The derived quality of information is meaningful if it corresponds to the actual variance through (6.5). This will be discussed by examples throughout this chapter.

Subjective logic provided the inspiration for the fusion and trust characterization operators described in this chapter. The operators described here approximate Bayesian reasoning using the following framework. The input opinions about the state variables and source behaviors translate to Dirichlet distributions to describe the uncertainty about the corresponding appearance probabilities of the various values of these variables. Bayesian reasoning determines the exact output distribution for the appearance probabilities for fusion or discounting, and then this exact distribution is approximated by a Dirichlet distribution such that the

mean values match exactly and the variances match in the least squares sense. In other words, moment matching determines the Dirichlet approximation, and the corresponding Dirichlet parameters lead to the fused or discounted opinion.

6.4 The Source Estimation and Fusion Problem

In general, a decision-maker collects, over the course of his/her duties, reports from different sources about many different state variables. The decision-maker employs A unique sources and evaluates I variables. The decision-maker may or may not be able to form an initial opinion about each variable. We denote the opinion about the i -th state variable from the a -th source using a subscript as $\mathbf{n}_{i,a}$. When the a -th source does not have any observations about the i -th variable, it should report the vacuous opinion $n_{i,a}^k = 0$ for all $k \in \mathbb{K}$. The decision-maker may or may not be able to form an initial opinion about a state variable and acts as an ego-source. We index the ego-source as $a = 0$ and other sources as positive integers $a > 0$. For ease of illustration in the chapter, all of the I state variables are binary, i.e., their instantiations are propositions where $\mathbb{K} = \{+, -\}$ and $+$ and $-$ represent a positive and negative variable value, respectively. An example of such a proposition is that a particular vendor provides a satisfactory (+) or an unsatisfactory (-) transaction.

The sources do not necessarily correctly report their opinions based upon their individual observations. Many times, some sources intentionally lie and report opinions in direct conflict with other sources. The ultimate problem for the decision-maker is to form a fused opinion that portrays information about ground truth probabilities of the values of state variables consistent with the opinion's apparent quality of information. This fused opinion should represent higher quality of information than can be obtained from any smaller subset of sources.

To enable an effective solution to the fusion problem, we incorporate the beta model from [20]. Specifically, the behavior of a source is a state variable itself where the variable values are particular behaviors describing how the source transforms its truthful opinion into its reported opinion. While a large number of source behaviors may exist, we restrict the discussion in this chapter to the three well-studied behaviors from the literature [3, 17, 24, 40]: (1) good, (2) flipping, and (3) random. In the good behavior case, the a -th source accurately reports the number of positive and negative instantiations of state variables it observed. When the source exhibits flipping behavior, it exchanges the number of positive and negative instantiations. Finally in the random case, a source randomly selects the number of positive and negative instantiations to report independent of the actual numbers it observed. This chapter will examine the robustness of such a beta model by considering that the ground truth source behaviors are one of these three, but the fusion algorithms either account only for two behaviors (good and random) or all three. Clearly, performance drops when there is a model mismatch, and in real applications, one may want to incorporate a richer set of behavior models. In recent work, we developed methods to learn new behavior models using an ego-source [27]. These richer behavior

models are beyond the scope of this chapter. Nevertheless, the results in this chapter do provide insights about the impact of model mismatches.

The a -th source's behavior profile is the ground truth probabilities to exhibit each one of the three behaviors adopted while sharing its opinions. The decision-maker builds up an opinion about the behavior profile by determining the effective number of instances t_a^k that the a -th source exhibited behavior $k \in \{g, f, r\}$. After each time the decision-maker collects opinions from the different sources, it cannot directly determine which behavior each source actually followed. The next two sections describe methods to build up opinions about the source behaviors and then use these opinions to fuse the source reports. Due to the lack of direct observations, the behavior opinions t_a^k are not necessarily integers, which also means the fused opinions about the variables also need not take integer values.

To demonstrate the effectiveness of the methods presented in the next two sections, 100 sources reporting 1000 variables are simulated using the three behavior models. A given percentage of the source agents will be predominately good, flippers, and random, respectively. Predominately good sources report their true opinions about a particular variable with a probability of 0.7, and their flipping and random probabilities are 0.15. Similarly, the predominately flipping and random source exhibits their dominating behaviors with probability 0.7 and the other two behaviors with probabilities of 0.15. Predominately good sources can lie, albeit with a much smaller probability. In contrast a predominately flipping source can provide a truthful opinion. For the i -th state variable, the number of direct observations $N_{i,a}$ that the a -th source achieves for the variable's values is a random number drawn uniformly between 0 and 100. The underlying ground truth probability for the positive value p_i^+ of each of the state variables is sampled over the uniform distribution between 0 and 1. The a -th source's true opinion about the i -th variable is the result of $N_{i,a}$ draws from a Bernoulli process with probability p_i^+ . Each source then determines its behavior for the variable as a random multinomial draw using its behavior probabilities. If this draw selects the good behavior, the source reports its true opinion. If the draw selects the flipping behavior, the source swaps its $n_{i,a}^+$ and $n_{i,a}^-$ values. Otherwise the random behavior means that the source chooses the integer $n_{i,a}^+$ uniformly between 0 and $N_{i,a}$ and sets $n_{i,a}^- = N_{i,a} - n_{i,a}^+$.

If the decision-maker does not account for the various behaviors of the sources and assumes all the reported opinions for the i -th variable are correct, the fusion process is rather straightforward. The fusion operations make two weaker assumptions: (1) each reported opinion is statistically independent of the others (i.e., the observed evidence of the sources do not overlap), and (2) the prior distribution in light of no observed evidence is uniform (which is an uninformative prior). With these assumptions, it can be shown that the distribution of the fused opinion is a beta distribution given by

$$f_\beta(p|\mathbf{n}_{i,f}) \propto \prod_{a=1} f_\beta(p|\mathbf{n}_{i,a}) \quad (6.6)$$

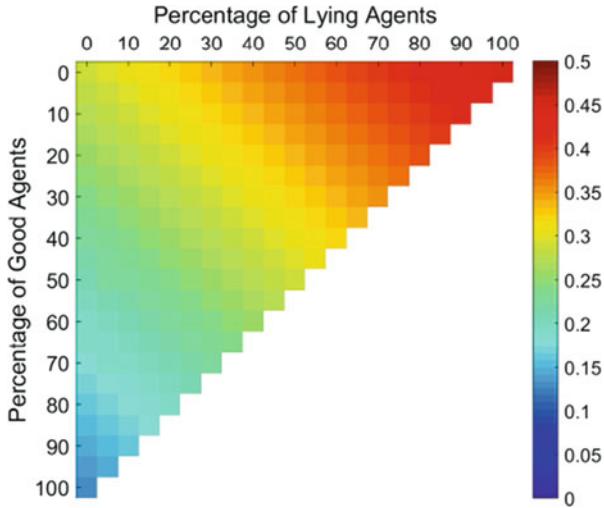


Fig. 6.1 RMSE of consensus fusion for various mixtures of predominately good, flipping, and random sources represented as a heat map

where $n_{i,f}^k = \sum_{a=1}^A n_{i,a}^k$ for $k \in \{+, -\}$. In other words, the fused opinion in evidence space is the output of the operator $\text{Consensus}(\mathbf{n}_{i,1}, \dots, \mathbf{n}_{i,A})$ that simply sums the evidence supplied by each source. Consensus fusion is one of the more commonly used operators in subjective logic [10].

When the consensus operator is applied over all simulated reports from 100 agents covering 1000 variables, the resulting root mean square error (RMSE) between the expected fused probability (see (6.4)) and the ground truth opinion for various mixtures of sources types is given in Fig. 6.1 as a heat map. When most of the sources are predominately good, the RMSE is fairly low at 0.13. As those sources are replaced by predominately random sources, the RMSE grows to about 0.3, which is consistent to a complete random guess from a uniform distribution. When most of the sources are predominately flipping, the RMSE grows to above 0.4 as the flipped reports are moving the estimated probabilities far from the ground truth. The predicted RMSE can be calculated as the root mean of the expected variance of the fused opinions given by (6.5). For all cases, the predicted RMSE is similar with a mean value of 0.0070, which is much smaller than the actual RMSE (even with 100% good sources). This is because consensus fusion assumes all source reports are good, which is not even true 30% of the time for good sources. Clearly, the behavior of the sources must be accounted for in the fusion process.

6.5 Fusion Using Behavior Estimates via Ego-Sources

When the decision-maker can directly observe instantiations of different state variables, it can act as its own ego-source. Assuming the decision-maker is competent and acting in its own best interest, the ego-source's opinion will always be good. Then, the decision-maker compares its opinions against those of the a -th source over a set of I variables to determine the source behavior profile. The procedure to determine the opinions about source behaviors was first derived in [15] for good and random behaviors (the two-mode model), and it can trivially be generalized for a finite set of known behaviors. Following the derivation in [15], the posterior distribution for the probabilities that the a -th source follows particular behaviors given the set of opinions from the ego-source about propositions and a -th source is

$$f(\mathbf{p}|\mathbf{t}_a) \propto \prod_i \left(\sum_{k \in \mathbb{K}} \text{Prob}(\mathbf{n}_{i,0}|\mathbf{n}_{i,a}, k) p^k \right), \quad (6.7)$$

where the likelihood of the a -th source exhibiting the k -th behavior when reporting its opinion about the i -th variable is

$$\begin{aligned} \text{Prob}(\mathbf{n}_{i,0}|\mathbf{n}_{i,a}, k) &= \int p^{n_{i,0}^+} (1-p)^{n_{i,0}^-} f_\beta(p|h^k(\mathbf{n}_{i,a})) dp, \\ &= \frac{B(h^k(\mathbf{n}_{i,a}) + \mathbf{n}_{i,0} + 1)}{B(h^k(\mathbf{n}_{i,a}) + 1)}, \end{aligned} \quad (6.8)$$

where $h^g(\mathbf{n}) = \mathbf{n}$, $h^f(\mathbf{n}) = [n^-, n^+]$, and $h^r(\mathbf{n}) = [0, 0]$ represent the accurate information that can be obtained from the source when it is known to employ good, flipping, or random behavior, respectively, for the given report. For the random behavior, the opinion is completely independent of the source's actual observation, and therefore the sources report is vacuous.

The source behavior characterization method approximates (6.7) by finding the Dirichlet distribution that matches the means of (6.7) and matches the variances as closely as possible (in the least squares sense). Closed form expressions for the means and variances of (6.7) are available because the distribution is a mixture of Dirichlets. However, the number of modes grows exponentially with respect to the number of variables I . In [15], a method is presented that updates a source behavior opinion by sequentially performing moment matching over one state variable at a time. It is shown in [15] that this sequential updating method is almost as accurate as the much more computationally complex method that incorporates all propositions at once. We refer to the operator $\mathbf{t}_a = \text{SourceBehavior}(\mathbf{n}_{1,0}, \mathbf{n}_{1,a}, \dots, \mathbf{n}_{I,0}, \mathbf{n}_{I,a})$ as the sequential method that approximates the Dirichlet distribution for the source behavior probabilities using the parameters \mathbf{t}_a as effective evidences of the source behaviors. In this chapter, while the simulated sources are randomly picking one of three behaviors for each

propositional report, the source characterization method is either employing the two-mode model, i.e., $\mathbb{K} = \{g, r\}$ or the three-mode model, i.e., $\mathbb{K} = \{g, f, r\}$. This allows understanding of the performance loss when the assumed model does not fully characterize the data.

In essence, the `SourceBehavior` operator calibrates the a -th source relative to the ego-source. Kaplan et al. [15] explain how the behavior opinions are updated based upon the consistency between the opinions of the ego-source and those of the a -th source. When both opinions represent similar probabilities and are supported by large evidence (i.e., small uncertainty), then the likelihood for the good behavior in (6.8) becomes very large relative to the other likelihoods. This actually means that the evidence for the good behavior is incremented by a number near 1 and the evidence for the other behaviors is slightly decremented. Similarly, if the probabilities are consistent with flipping, the evidence for flipping behavior is incremented by a number near 1. Otherwise when the probabilities are inconsistent to either good or flipping behaviors, the evidence for random behavior is incremented by a number near to 1. The increment of the behavior evidence update decreases as the uncertainty associated to either source opinion increases. When the ego-source's opinion becomes vacuous, i.e., $n_{i,0}^+ = n_{i,0}^- = 0$, the likelihoods for each of the behaviors become equal, and the update does not change any of the source behavior opinions. In other words, when the uncertainty of the reported propositions are low, the behavior update is comparable to directly observing which behavior the a -th source used in reporting the given proposition. As the uncertainty of either reported opinion grows, the increments to the source behavior evidence go to zero. The strength of the update depends on how much direct evidence the ego-source is able to observe.

Given the characterization of the behavior of the sources, the subjective logic method discounts each source's behavior followed by consensus fusion [14]. The discount operation in subjective logic, which originates from Dempster-Shafer theory [29], only considers the belief that the source provides a good report. Specifically, the function $\text{Discount}(\mathbf{t}, \mathbf{n}) = \frac{t^s + 1}{\sum_{k \in \mathbb{K}} (t^k + 1)} \mathbf{n}$ discounts the opinion based upon the expected probability of a good report. The reports of all sources for the i -th variable are discounted by `Discount` using their respective behavior profiles \mathbf{t}_a , and the outputs are passed through the `consensus` operator. In effect, the discount operator acts as a “soft” censor for sources.

Figure 6.2a shows the RMSE result s of subjective logic discounting and fusion. The error is clearly reduced as compared to consensus fusion. Specifically, the RMSE performance relies mostly on the percentage of predominately good sources. When the percentage of predominately good sources is 10%, the RMSE is about 0.2 (much lower than consensus alone), and this value decreases to 0.13 when all the sources are predominately good (comparable to consensus alone). Like consensus, the uncertainty associated to the fused opinion still greatly underpredicts the actual RMSE, where the predicted RMSE averages around 0.053 for the various mixtures of sources. The predicted RMSE is higher than that of simple consensus fusion

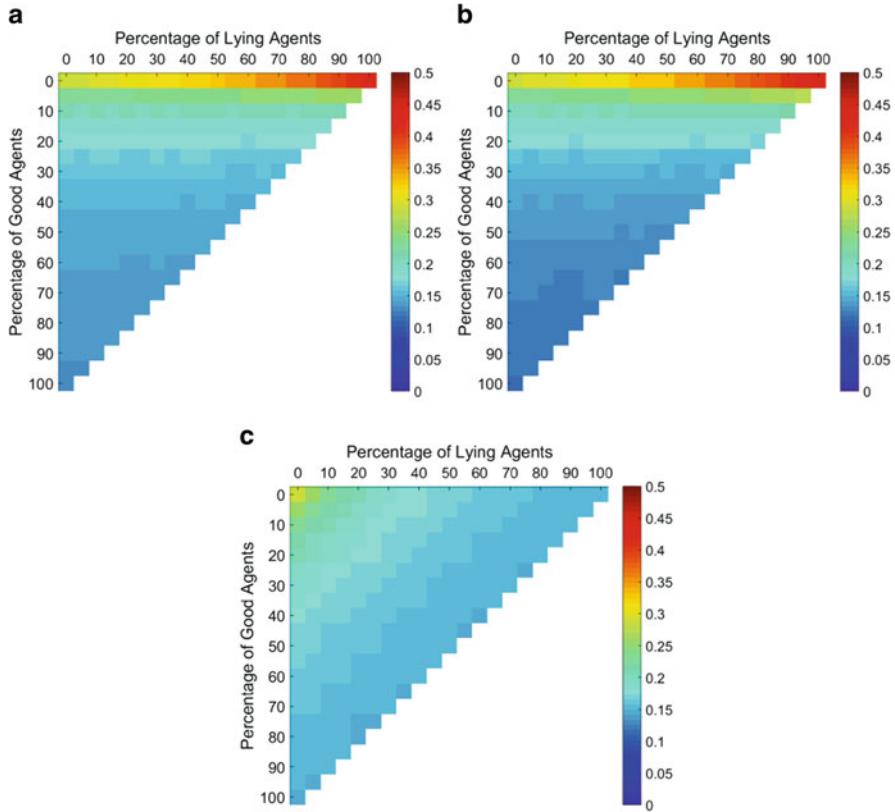


Fig. 6.2 RMSE of discounted fusion for various mixtures of predominately good, flipping, and random sources represented as a heat map: **(a)** subjective logic discounting followed by consensus fusion, **(b)** two-mode behavior discounting followed by consensus fusion, and **(c)** three-mode behavior discounting followed by consensus fusion

due to the discounting. It is greatest when the percentage of predominately flipping sources is 100% (0.13) and is smallest when the percentage of predominately good sources is 100% (0.026).

While the discount operator is intuitively appealing, it is ad hoc. By the two-mode beta model, the distribution for the probability of the i -th variable due to the a -th source's report is

$$f(p|\mathbf{t}_a, \mathbf{n}_{i,a}) \propto \frac{t_a^g + 1}{t_a^g + t_a^r + 2} f_\beta(p|[n_{i,a}^+, n_{i,a}^-]) + \frac{t_a^r + 1}{t_a^g + t_a^r + 2} f_\beta(p|[0, 0]). \quad (6.9)$$

The $\text{Discount}_2(\mathbf{t}_a, \mathbf{n}_{i,a})$ determines the discounted report as the evidence parameters of the beta distribution whose means and variances match the distribution in (6.9). This form of discounting was used in the TRAVOS trust and reputation model [31], and Eqs. (11)–(15) in [31] implement the Discount_2 operator.

Figure 6.2b shows the RMSE results when employing the Discount_2 operator followed by consensus. The results look very similar to SL discounting results. If one squints and look at the actual numbers, one will actually see a slight improvement using the two-mode discounting. Again, the predicted RMSEs are much lower than the actual errors. The predicted values are actually slightly larger than SL discounting, but not by much.

It is now natural to wonder about how the three-mode beta model can fair. In this case, the distribution for the probability of the i -th variable due to the a -th source's report is

$$f(p|\mathbf{t}_a, \mathbf{n}_{i,a}) \propto \frac{(t_a^g + 1)f_\beta(p|[n_{i,a}^+, n_{i,a}^-]) + (t_a^f + 1)f_\beta(p|[n_{i,a}^-, n_{i,a}^+]) + (t_a^r + 1)}{t_a^g + t_a^f + t_a^r + 3}. \quad (6.10)$$

Again, the method of moments can be employed to extract a discounted opinion by approximating (6.10) by a beta distribution. We refer the process as the $\text{Discount}_3(\mathbf{t}_a, \mathbf{n}_{i,a})$ operator. The actual operator is a special case of the joint discounting and consensus fusion operator described in [27] which will be discussed soon.

Figure 6.2c provides the RMSE results when employing Discount_3 before consensus. The results improve significantly over the two previous discounting operators as the number of predominately flipping sources increases. This is because the discounting operator actually to some extent “unflips” the reports from the flipping sources. Now, the performance of the fusion is primarily a function of the percentage of predominately random sources. With no random sources, the error is about 0.15 and grows to 0.29 when all sources are predominately random. Like the previous discounting operators, the predicted RMSE is much lower than the actual error. The predicted error is as low as 0.028 for no random source and is as high as 0.075 when all sources are predominately random.

The large gap between the predicted and actual fusion results for all the discounting methods indicated that more can be done. The discounted reports as given by the beta mixtures in (6.9) and (6.10) are poorly fitted by a single beta distribution. It is actually better to perform the fusion with the beta mixtures before finding an approximate beta distribution fit. Under the fairly general assumption that the prior on the distribution of values of propositions is uniform, the distribution after fusing the reports from all sources is

$$f(p|\mathbf{T}, \mathbf{N}_i) \propto \prod_{a=1}^A f(p|\mathbf{t}_a, \mathbf{n}_{i,a}), \quad (6.11)$$

where $f(p|\mathbf{t}_a, \mathbf{n}_{i,a})$ is given by (6.9) or (6.10) for the two-mode or three-mode behavior model, respectively. The operator $\mathbf{n}_{i,f} = \text{JointConDis}(\mathbf{t}_1, \mathbf{n}_{i,1}, \dots, \mathbf{t}_A, \mathbf{n}_{i,A})$ determines the fused opinion by selecting the opinion associated to the beta distribution that is determined through moment matching to (6.11). The distribution in (6.11) is a mixture of beta distributions, which leads to analytical expressions for

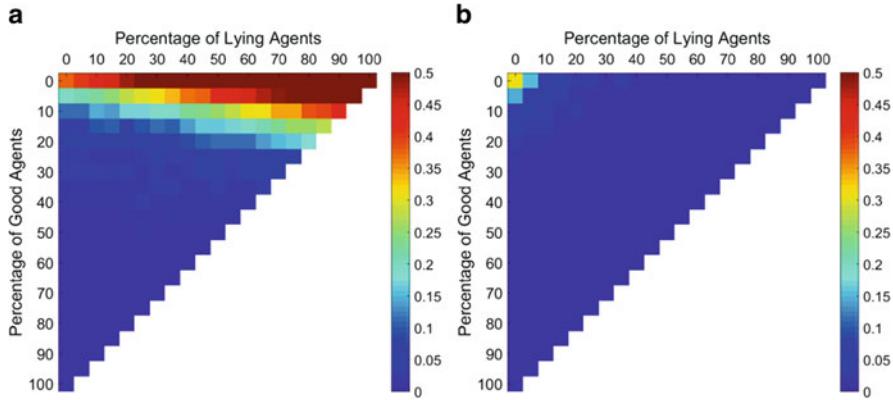


Fig. 6.3 RMSE of joint discounting and fusion for various mixtures of predominately good, flipping, and random sources represented as a heat map: (a) two-mode and (b) three-mode

the moments. However, the number of mixture components grows exponentially. A practical implementation of the `JointConDis` operator is presented in [27] that clusters similar components into a single component as more sources are integrated.

Figure 6.3a shows the RMSE results for two-mode `JointConDis`. The improvement over the previous discounting methods is obvious when the percentage of predominately good sources is 10% or greater. Otherwise, the reports from the predominately flipping agents can cohere in the fusion process and drown out the good reports. Therefore the error can become worse than the previous discounting methods when the percentage of predominately flipping sources is large. Now the predicted RMSE matches the actual RMSE as long as the percentage of predominately good agents dominates the flipping sources. Table 6.1 provides the actual and predicted RMSE numbers for various percentages of source types. The discrepancy between the actual and predicted values as more flipping sources are included is due to the fact that the two-mode model does not account for the flipping behavior.

Figure 6.3b shows the RMSE results for three-mode `JointConDis`. The error is now very small except for cases when the percentage of predominately random agents is 95% or more. Because the flipping behavior is modeled in the fusion approach, the reports of flippers can be “unflipped” so that predominately flipping sources are providing comparable information as predominately good agents, and the fusion method is able to exploit that information. Because the fusion method is modeling all the behaviors inherent in the synthesized sources, the predicted and actual errors are comparable except when all the sources are predominately random as provided in Table 6.2. It seems that the joint consensus fusion process that models all the source behaviors is able to achieve the lowest possible error and the fused opinion is able to represent the quality of information after fusion. The error is the lowest when none of the sources are predominately random. In such cases, the actual and predicted RMSE is 0.0075, which is slightly higher than the predicted RMSE

Table 6.1 RMSE and (predicted RMSE) of joint fusion and discounting using the two-mode model with an ego-source for various mixtures of predominately good, flipping, and random sources

% of good sources	% of flipping sources					
	0	20	40	60	80	100
0	0.3791 (0.1342)	0.4789 (0.1212)	0.5458 (0.0899)	0.5680 (0.0663)	0.5787 (0.0344)	0.5802 (0.0269)
20	0.0424 (0.0274)	0.0590 (0.0261)	0.0860 (0.0391)	0.1126 (0.0412)	0.1757 (0.0535)	—
40	0.0139 (0.0130)	0.0189 (0.0135)	0.0188 (0.0140)	0.0169 (0.0143)	—	—
60	0.0112 (0.0105)	0.0118 (0.0105)	0.0119 (0.0106)	—	—	—
80	0.0100 (0.0091)	0.0102 (0.0091)	—	—	—	—
100	0.0091 (0.0082)	—	—	—	—	—

Table 6.2 RMSE and (predicted RMSE) of joint fusion and discounting using the three-mode model with an ego-source for various mixtures of predominately good, flipping, and random sources

% of good sources	% of flipping sources					
	0	20	40	60	80	100
0	0.3124 (0.2004)	0.0284 (0.0246)	0.0113 (0.0112)	0.0095 (0.0094)	0.0083 (0.0083)	0.0076 (0.0075)
20	0.0281 (0.0252)	0.0117 (0.0113)	0.0092 (0.0094)	0.0085 (0.0083)	0.0075 (0.0075)	—
40	0.0110 (0.0119)	0.0093 (0.0094)	0.0081 (0.0083)	0.0075 (0.0075)	—	—
60	0.0096 (0.0094)	0.0085 (0.0083)	0.0077 (0.0075)	—	—	—
80	0.0081 (0.0083)	0.0075 (0.0075)	—	—	—	—
100	0.0075 (0.0075)	—	—	—	—	—

of standard consensus fusion as discussed in the previous section. This is because consensus fusion alone assumes all reports are honest, whereas the predominately honest and flipping sources still provide random reports 15% of the time, which is accounted for in joint consensus and discount fusion in (6.11).

6.6 Fusion Using Behavior Estimates Without Ego-Sources

The three-mode `JointConDis` is probably at the estimation limit when dealing with sources that probabilistically decide to manipulate their reported opinions. The problem is that it requires an ego-source to “calibrate” the source behavior profiles. This section investigates what is possible when an ego-source is unavailable. This occurs when the decision-maker does not have direct access to observe the values over different instantiations of the various variables. This section is inspired by the fact-finding work described in [36, 39].

It is interesting to look at the performance of the two-mode `JointConDis` when the source behavior opinion is vacuous, i.e., $t_a^g = t_a^r = 0$. Figure 6.4a shows the RMSE over the various mixtures of source types. Despite the lack of knowledge about the source behavior, the fusion still works well when the percentage of

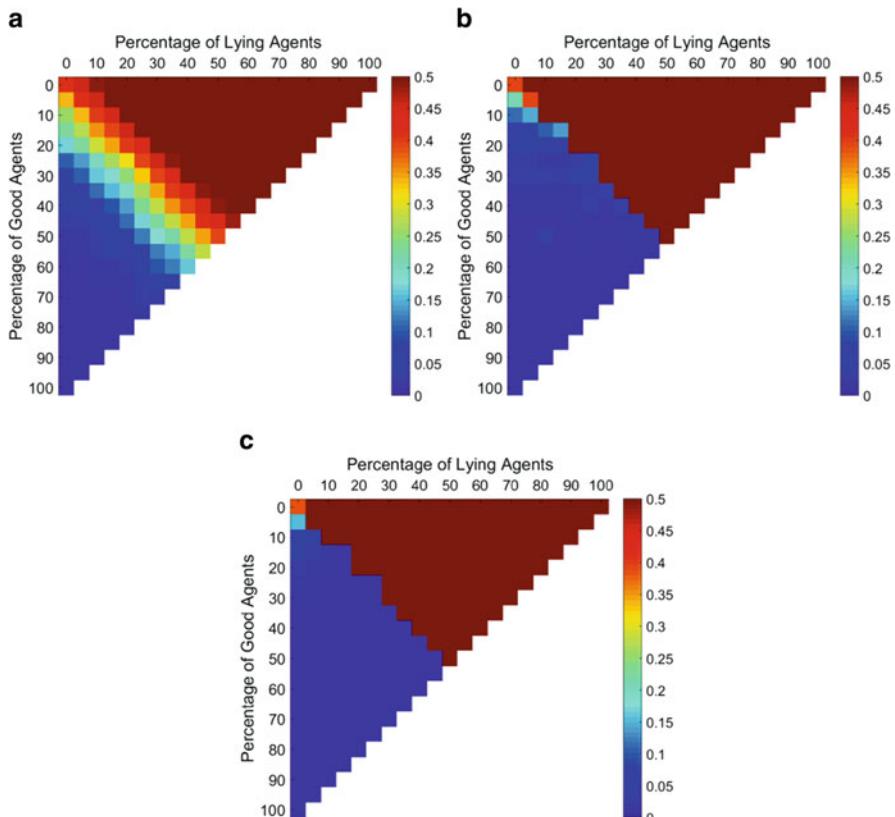


Fig. 6.4 RMSE without an ego-source for various mixtures of predominately good, flipping, and random sources represented as a heat map: (a) two-mode joint discounting and fusion using vacuous source behavior profiles, (b) two-mode fact-finding, and (c) three-mode fact-finding

Table 6.3 RMSE and (predicted RMSE) of joint fusion and discounting using the two-mode model without any ego-source for various mixtures of good, flipping, and random sources

% of good sources	% of flipping sources					
	0	20	40	60	80	100
0	0.4049 (0.0764)	0.5453 (0.0406)	0.5733 (0.0168)	0.5759 (0.0111)	0.5766 (0.0098)	0.5772 (0.0090)
20	0.1702 (0.0479)	0.4009 (0.0757)	0.5462 (0.0480)	0.5736 (0.0172)	0.5763 (0.0100)	—
40	0.0442 (0.0137)	0.1377 (0.0461)	0.3927 (0.0693)	0.5474 (0.0424)	—	—
60	0.0128 (0.0113)	0.0265 (0.0119)	0.1628 (0.0482)	—	—	—
80	0.0104 (0.0098)	0.0132 (0.0100)	—	—	—	—
100	0.0090 (0.0090)	—	—	—	—	—

predominately good sources is much larger than the percentage of predominately flipping sources. In fact, the RMSE is mostly a function of the difference of these two percentages. Table 6.3 provides the actual and predicted RMSE obtained by the fused opinions. When all the sources are predominately good, the match is very close. In this case, the good reported opinions are able to cohere in the JointConDis operation against the noncoherent random opinions. The match between the actual and predicted errors slowly deteriorates as the difference between the percentages of predominately good and flipping sources decreases. Once there are more flipping sources, the JointConDis is cohering to the flipped opinions, and the actual RMSE becomes large because the estimate is a flipped version of the ground truth. Overall, the performance of two-mode JointConDis with the vacuous behavior is not as good as using the ego-source-generated behavior profile, but it does significantly outperform the earlier discounting methods when the predominately good sources outnumber the flipping ones. This indicates that fusion without an ego-source to calibrate the sources is possible, but more can be done as we will now see.

The three-mode JointConDis operator using a vacuous source belief profile is ineffective because the distribution given by (6.11) is bimodal due to the modeling of the flipping behavior and the two modes are equiprobable in the absence of prior knowledge of the relative number of sources that are exhibiting good and flipping behaviors for the given variable. Fitting a single beta distribution to this bimodal distribution leads to a poor characterization of the fused opinion, and it is not clear which mode is representative of the ground truth and which mode is representative of the flipped ground truth.

The performance of the two-mode JointConDis operator using a vacuous source belief profile appears to provide a surprisingly good representation of the ground truth when the majority of sources are predominately good. The fact-finding

methods used for non-probabilistic propositions that have certain values either *true* or *false* [36, 39] provide inspiration to do more. The fact-finding methods alternate between estimating the trustworthiness of the sources given an estimate of the truth of their claims and estimating the truth of the claims given an estimate of the trustworthiness of the sources. In other words, the fused opinion that is the output of the two-mode *JointConDis* operator using the vacuous source behavior profile can serve as an initial surrogate for an ego-source opinion. Then, the *SourceBehavior* operator using a two-mode model provides an updated source behavior profile opinion for each source. Next, updated fused opinions are obtained using the two-mode *JointConDis* operator with the updated source behavior profile opinions, and then the *SourceBehavior* operator updates the source behavior profile opinions using the updated fused opinions. The process repeats until the source behavior profile opinions converge. The details of the two-mode fact-finding method is shown in Operator 1 as *FactFind*₂.

Operator 1 [$\mathbf{n}_{1,f}, \dots, \mathbf{n}_{I,f}, \mathbf{t}_1, \dots, \mathbf{t}_A$] = *FactFind*₂($\mathbf{n}_{1,1}, \dots, \mathbf{n}_{i,a}, \dots, \mathbf{n}_{I,A}$)

```

 $t_a^g = t_a^r = 0$  for  $a = 1, \dots, A$ 
 $t_a^{g'} = r_a^{r'} = 1$  for  $a = 1, \dots, A$ 
while  $\sum_{a=1}^A \|t_a' - t_a\|^2 > \epsilon$  do
     $t_a' = \mathbf{t}_a$  for  $a = 1, \dots, A$ 
    /* Use 2-mode source behavior model */
     $\mathbf{n}_{i,f} = \text{JointConDis}(\mathbf{n}_{i,1}, \mathbf{t}_1, \dots, \mathbf{n}_{i,A}, \mathbf{t}_A)$  for  $i = 1, \dots, I$ 
     $\mathbf{t}_a = \text{SourceBehavior}(\mathbf{n}_{1,f}, \mathbf{n}_{1,a}, \dots, \mathbf{n}_{I,f}, \mathbf{n}_{I,a})$  for  $a = 1, \dots, A$ 
end while

```

Figure 6.4b shows the RMSE of the two-mode fact-finding method. As long as the predominately good sources outnumber the predominately flipping agents, the RMSE is very low. Otherwise, the error is large because the flipped version of the ground truth prevails. Table 6.4 compares the actual and predicted RMSE values for various mixtures of source types. As long as the number of predominately good sources significantly outnumbers the other source types, the predicted and actual errors are comparable. The discrepancy between the actual and predicted errors when the predominately good sources are slightly in the majority indicates that more still can be done.

The two-mode fact-finding method does not correct for flipping behavior. A three-mode fact-finding method can do better, but without an initial estimate of the source behaviors, three-mode *JointConDis* suffers from the two-mode problem discussed earlier. Operator 2 describes the three-mode fact-finding operator *FactFind*₃ that alternates between joint discounting and fusion and source behavior characterization using the fused opinions as ego-source surrogates. It is initialized by *FactFind*₂ to determine initial evidence for the good behavior associated to each source. To this end, the second step in Operator 2 is actually transferring the belief in the random behavior into uncertainty. In setting up a three-

Table 6.4 RMSE and (predicted RMSE) of two-mode fact-finding for various mixtures of good, flipping, and random sources

% of good sources	% of flipping sources					
	0	20	40	60	80	100
0	0.3947 (0.1071)	0.5721 (0.0256)	0.5757 (0.0124)	0.5764 (0.0104)	0.5763 (0.0090)	0.5764 (0.0081)
20	0.0434 (0.0271)	0.5689 (0.0312)	0.5746 (0.0159)	0.5763 (0.0104)	0.5762 (0.0091)	—
40	0.0139 (0.0130)	0.0200 (0.0127)	0.5759 (0.0145)	0.5764 (0.0105)	—	—
60	0.0112 (0.0104)	0.0120 (0.0104)	0.0122 (0.0105)	—	—	—
80	0.0100 (0.0090)	0.0104 (0.0091)	—	—	—	—
100	0.0092 (0.0081)	—	—	—	—	—

mode belief, the beliefs in flipping and random behavior are set to zero, and the good behavior belief and uncertainty are transferred from the two-mode behavior opinion. This can be verified by the evidence to belief mapping given by (6.3). The rationale for moving the random behavior belief to uncertainty is because the two-mode fact-finding cannot distinguish between random and flipping behavior and only the belief in the good behavior is valid. Then, the fact-finding methods alternate between performing three-mode `JointConDis` to estimate fused opinions as surrogates for the ego-source opinions and three-mode `SourceBehavior` to update the source behavior profile opinions until convergence.

Operator 2 $[\mathbf{n}_{1,f}, \dots, \mathbf{n}_{I,f}, \mathbf{t}_1, \dots, \mathbf{t}_A] = \text{FactFind}_3(\mathbf{n}_{1,1}, \dots, \mathbf{n}_{i,a}, \dots, \mathbf{n}_{I,A})$

```

 $[\dots, \tilde{\mathbf{t}}_1, \dots, \mathbf{t}_A] = \text{FactFind}_2(\mathbf{n}_{1,1}, \dots, \mathbf{n}_{i,a}, \dots, \mathbf{n}_{I,A})$ 
 $t_a^g = \frac{3t_a^g}{2+t_a^r}$  for  $a = 1, \dots, A$ 
 $t_a^f = t_a^r = 0$  for  $a = 1, \dots, A$ 
 $t_a^g = t_a^f = t_a^{r'} = 1$  for  $a = 1, \dots, A$ 
while  $\sum_{a=1}^A \|\mathbf{t}'_a - \mathbf{t}_a\|^2 > \epsilon$  do
     $\mathbf{t}'_a = \mathbf{t}_a$  for  $a = 1, \dots, A$ 
    /* Use 3-mode source behavior model */
     $\mathbf{n}_{i,f} = \text{JointConDis}(\mathbf{n}_{i,1}, \mathbf{t}_1, \dots, \mathbf{n}_{i,A}, \mathbf{t}_A)$  for  $i = 1, \dots, I$ 
     $\mathbf{t}_a = \text{SourceBehavior}(\mathbf{n}_{1,f}, \mathbf{n}_{1,a}, \dots, \mathbf{n}_{I,f}, \mathbf{n}_{I,a})$  for  $a = 1, \dots, A$ 
end while

```

Figure 6.4c shows the RMSE of the three-mode fact-finding method. The boundary where the numbers of predominately good and flipping agents are comparable is sharper than that of the results from the two-mode fact-finding method. The actual

Table 6.5 RMSE and (predicted RMSE) of three-mode fact-finding for various mixtures of good, flipping, and random sources

% of good sources	% of flipping sources					
	0	20	40	60	80	100
0	0.3820 (0.1239)	0.5743 (0.0238)	0.5761 (0.0112)	0.5764 (0.0094)	0.5764 (0.0083)	0.5765 (0.0075)
20	0.0243 (0.0262)	0.5761 (0.0115)	0.5763 (0.0094)	0.5765 (0.0083)	0.5765 (0.0075)	—
40	0.0108 (0.0114)	0.0093 (0.0094)	0.5764 (0.0083)	0.5765 (0.0075)	—	—
60	0.0096 (0.0094)	0.0085 (0.0083)	0.0077 (0.0075)	—	—	—
80	0.0081 (0.0083)	0.0076 (0.0075)	—	—	—	—
100	0.0075 (0.0075)	—	—	—	—	—

and predicted RMSE numbers are provided in Table 6.5. When the predominately good sources outnumber the flippers, the three-mode fact-finding lowers the error as compared to two-mode fact-finding because it does explicitly correct for flipping behaviors. Furthermore, the agreement between the actual and predicted errors is maintained as long as the predominately good sources outnumber the flippers. It seems that the three-mode fact-finding is pushing the limits of what is possible for jointly performing source and fused opinion in the absence of an ego-source. Without the ego-agent, one must make the implicit assumption that lying sources are in the minority. This is true for state variables in general and is also true for traditional fact-finding methods that operate over crisp propositions [36, 39]. When the assumption is violated, the fact-finding method will fail. This seems to be a fundamental barrier when an ego-source is unavailable.

6.7 Discussion and Conclusions

This chapter demonstrates how to perform fusion of subjective opinions from possibly unreliable sources to estimate values of probabilistic state variables. Specifically, a subjective opinion about a state variable summarizes the evidence about the possible probabilities that the state variable takes one of K values. It encodes both the expected probabilities as the information and the amount of evidence that has been collected as the quality of the information. As shown in this chapter, the quality of information represents the spread (or difference) between the actual ground truth probabilities and the expectation information derived from the observations.

When the decision-maker (or its very trusted advisor) has direct observations about many of the state variables or propositions, the decision-maker can use the consistency of his/her observations and the corresponding reported opinions of a particular source to establish a source behavior profile opinion for that particular source. The decision-maker can achieve very effective fusion by accounting for its source behavior opinions of each source in conjunction with the reported opinions from each source. It is demonstrated that it is much more effective to perform the fusion in one shot rather than discounting each source's opinion by its corresponding behavior profile opinion. This is because a discounted opinion is unable to jointly capture the uncertainty of the reported opinion in light of the source's various behaviors.

Simulations of three (good, flipping, and random) source behaviors help to demonstrate the effectiveness of the various fusion methods. When the fusion method models all three behaviors, the fusion leads to a very tight estimate of ground truth that is well characterized by the ground truth. When the fusion method only models good and random behavior, the estimate of the ground truth is not as tight because the method is only able to censor (and not correct for) the flipping behavior, which is usually inconsistent with good behavior. The quality of information is still able to characterize the difference between the estimates and ground truth as long as the unmodeled flipping behavior does not become overly prevalent.

It is possible to perform fusion of a set of subjective opinions when the decision-maker does not have any direct observations to calibrate the behavior profiles of sources. In these cases, fused opinions act as a surrogate for the direct opinions so that inspired by fact-finding methods, one can iterate between fusion and source behavior estimation where the estimates progressively improve as long as the good behaviors occur more frequently than the bad behaviors. The fact-finding principle provides good estimates whose errors are well characterized by the quality of information as long a good behaviors occur more than flipping behaviors.

In general, sources can exhibit more than the three behaviors considered in this chapter. Nevertheless, the two-mode behavior model is a robust behavior model because the random behavior can capture source behaviors intended to move a fused estimate farther from the ground truth. The problem with the two-mode behavior model is that it does not allow fusion method to incorporate “bad” reports by implicitly transforming them into “good” reports. In essence, the two-mode behavior model only enables the fusion to censor (but not correct for) bad behaviors. The three-mode behavior model can correct for flipping behaviors, but it will not be able to correct for other unmodeled behaviors. The insight of the results in this chapter mean that in light of additional source behaviors, the three-mode fusion methods will still achieve good fusion with a meaningful quality of information characterization as long as the decision-maker has direct observations to calibrate the sources. It is just that the fusion performance could be improved by explicitly modeling the behaviors, and methods such as in [27] could be employed to learn new behaviors. Without the direct observation, the fact-finding methods will still be effective as long as the good behavior is the majority behavior exhibited in the

collections of reports. Furthermore, the three-mode fact-finding should still beat out two-mode fact-finding because it can use flipped reports as information.

The beta model to characterize source behavior is nice because it captures the idea that sources do not always lie or tell the truth. However, a clever and malicious source would try to be truthful as much as possible to build a good reputation and decide to lie at the moment that makes the decision-maker's organization the most vulnerable. This chapter does not present "the" technique to handle such a case, but this chapter does provide insights in the challenges to successfully deceive or protect from such deception. For instance, the malicious source can only be effective if he/she coordinates his/her lie with other sources and those sources are not drowned out by a larger group of good sources. Likewise, the decision-maker can use other stereotypical or profile information about the source, e.g., see [2, 28], along with risk/benefit analysis to build the sources reputation based upon its past forgone opportunities to cause harm in light of its likely affiliations. In other words, each proposition need not be considered equal in forming the source behavior profile. Furthermore, the fusion methods in this chapter assume independent sources, and as a result, they are vulnerable to coordinating sources. Understanding how the source profile information forms an influence network among sources can lead to better methods. For instance, social EM is a fact-finding method for binary propositions that is resilient to the "echo chamber" effect in social networks [38].

Specific applications will drive the exact source reputation and fusion system that is required. The methods presented in this chapter are generic. While they are not necessarily best for a particular scenario, such as a set of cooperating sources waiting for the exact right time to lie, the methods presented here can serve as the building blocks for customized systems. Overall, there are opportunities to design fusion systems to be resilient to conflicting and malicious sources. However, there are limitations to how resilient the system can be built. The chapter has identified some of these limitation and opportunities.

Acknowledgements Research was sponsored by the US Army Research Laboratory and was accomplished under agreement numbers W911NF-14-1-0199. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the US Army Research Laboratory or the US government. The US government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation hereon. Dr. Sensoy thanks the US Army Research Laboratory for its support under grant W911NF-14-1-0199 and The Scientific and Technological Research Council of Turkey (TUBITAK) for its support under grant 113E238.

References

1. V. Bui, R. Verhoeven, J. Lukkien, R. Kocielnik, A trust evaluation framework for sensor readings in body area sensor networks, in *Proceedings of the 8th International Conference on Body Area Networks, BodyNets '13* (ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, 2013), pp. 495–501

2. C. Burnett, T.J. Norman, K. Sycara, Stereotypical trust and bias in dynamic multiagent systems. *ACM Trans. Intell. Syst. Technol.* **4**(2), 26:1–26:22 (2013)
3. C. Fung, R. Boutaba, *Intrusion Detection Networks: A Key to Collaborative Security* (CRC Press, London, 2013)
4. S. Ganeriwal, L. Balzano, M. Srivastava, Reputation-based framework for high integrity sensor networks. *ACM Trans. Sens. Netw. (ToSN)* **4**(3), 15 (2008)
5. S. Han, B. Koo, A. Hutter, W. Stechele, Forensic reasoning upon pre-obtained surveillance metadata using uncertain spatio-temporal rules and subjective logic, in *Analysis, Retrieval and Delivery of Multimedia Content*, ed. by N. Adami, A. Cavallaro, R. Leonardi, P. Migliorati (Springer, New York, 2013), pp. 125–147
6. G. Han, J. Jiang, L. Shu, J. Niu, H.-C. Chao, Management and applications of trust in wireless sensor networks: a survey. *J. Comput. Syst. Sci.* **80**(3), 602–617 (2014)
7. A. Jøsang, A logic for uncertain probabilities. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **9**(3), 279–311 (2001)
8. A. Jøsang, The consensus operator for combining beliefs. *Artif. Intell. J.* **142**(1–2), 157–170 (2002)
9. A. Jøsang, Conditional reasoning with subjective logic. *J. Multiple-Valued Log. Soft Comput.* **15**(1), 5–38 (2009)
10. A. Jøsang, *Subjective Logic: A Formalism for Reasoning Under Uncertainty* (Springer, Cham, 2016)
11. A. Jøsang, R. Ismail, The beta reputation system, in *Proceedings of the Fifteenth Bled Electronic Commerce Conference e-Reality: Constructing the e-Economy*, Bled, June 2002, pp. 48–64
12. A. Jøsang, R. Hayward, S. Pope, Trust network analysis with subjective logic, in *Proceedings of the 29th Australasian Computer Science Conference*, Hobart (2006), pp. 85–94
13. A. Jøsang, J. Diaz, M. Rifqi, Cumulative and averaging fusion of beliefs. *Inf. Fusion* **11**(2), 192–200 (2010)
14. A. Jøsang, T. Ažderska, S. Marsh, Trust transitivity and conditional belief reasoning, in *6th IFIP WG 11.11 International Conference, IFIPTM 2012*, Surat, May 2012, pp. 68–83
15. L. Kaplan, M. Şensoy, S. Chakraborty, G. de Mel, Partial observable update for subjective logic and its application for trust estimation. *Inf. Fusion* **26**, 66–83 (2015)
16. S. Kotz, N. Balakrishnan, N.L. Johnson, *Continuous Multivariate Distributions*, vol. 1 (Wiley, New York, 2000)
17. T.R. Levine, *Encyclopedia of Deception* (SAGE Publications, Los Angeles, 2014)
18. Y. Liu, K. Li, Y. Jin, Y. Zhang, W. Qu, A novel reputation computation model based on subjective logic for mobile ad hoc networks. *Futur. Gener. Comput. Syst.* **27**(5), 547–554 (2011)
19. T.K. Moon, The expectation-maximization algorithm. *IEEE Signal Process. Mag.* **13**(6), 47–60 (1996)
20. T. Muller, P. Schweitzer, On beta models with trust chains, in *Proceedings of Trust Management VII: 7th IFIP WG 11.11 International Conference*, Malaga (2013), pp. 49–65
21. N. Oren, T.J. Norman, A. Preece, Subjective logic and arguing with evidence. *Artif. Intell.* **171**(10), 838–854 (2007)
22. J. Pasternack, D. Roth, Making better informed trust decisions with generalized fact-finding, in *IJCAI (Spatial Cognition*, Bremen, 2011), pp. 2324–2329
23. K. Regan, P. Poupart, R. Cohen, Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change, in *Proceedings of the 21st National Conference on Artificial Intelligence* (AAAI Press, Menlo Park, 2006), pp. 1206–1212
24. M. Şensoy, P. Yolum, Experimental evaluation of deceptive information filtering in context-aware service selection, in *International Workshop on Trust in Agent Societies* (Springer, Berlin/Heidelberg, 2008), pp. 326–347
25. M. Sensoy, G. de Mel, T. Pham, L. Kaplan, T.J. Norman, TRIBE: trust revision for information based on evidence, in *Proceedings of 16th International Conference on Information Fusion*, Istanbul (2013)

26. M. Sensoy, L. Kaplan, G. Ayci, G. de Mel, FUSE-BEE: fusion of subjective opinions through behavior estimation, in *18th International Conference on Information Fusion*, Washington, DC (2015), pp. 558–565
27. M. Sensoy, L. Kaplan, G. de Mel, T.D. Gunes, Source behavior discovery for fusion of subjective opinions, in *19th International Conference on Information Fusion*, Heidelberg (2016), pp. 138–145
28. M. Sensoy, B. Yilmaz, T.J. Norman, Stage: stereotypical trust assessment through graph extraction. *Comput. Intell.* **32**(1), 72–101 (2016)
29. G. Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, 1976)
30. P. Smets, The combination of evidence in the transferable belief model. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**(5), 447–458 (1990)
31. W.T.L. Teacy, J. Patel, N.R. Jennings, M. Luck, TRAVOS: trust and reputation in the context of inaccurate information sources. *Auton. Agents Multi-Agent Syst.* **12**(2), 183–189 (2006)
32. W.L. Teacy, M. Luck, A. Rogers, N.R. Jennings, An efficient and versatile approach to trust and reputation using hierarchical Bayesian modelling. *Artif. Intell.* **193**, 149–185 (2012)
33. D. Wang, C. Huang, Confidence-aware truth estimation in social sensing applications, in *12th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, Seattle, June 2015, pp. 336–344
34. D. Wang, L. Kaplan, T. Abdelzaher, C.C. Aggarwal, On credibility estimation tradeoffs in assured social sensing. *IEEE J. Sel. Areas Commun.* **31**(6), 1026–1037 (2013)
35. D. Wang, L. Kaplan, T.F. Abdelzaher, Maximum likelihood analysis of conflicting observations in social sensing. *ACM Trans. Sens. Netw. (ToSN)* **10**(2), 30 (2014)
36. D. Wang, T. Abdelzaher, L. Kaplan, *Social Sensing: Building Reliable Systems on Unreliable Data* (Morgan Kaufmann, Waltham, 2015)
37. A. Whitby, A. Jøsang, J. Indulska, Filtering out unfair ratings in Bayesian reputation systems. *ICFAIN J. Manag. Res.* **4**(2), 48–64 (2005)
38. S. Yao, S. Hu, S. Li, Y. Zhao, L. Su, L. Kaplan, A. Yener, T. Abdelzaher, On source dependency models for reliable social sensing: algorithms and fundamental error bounds, in *IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*, Nara (2016), pp. 467–476
39. X. Yin, J. Han, P.S. Yu, Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. Knowl. Data Eng.* **20**(6), 796–808 (2008)
40. B. Yu, M.P. Singh, Detecting deception in reputation management, in *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems* (ACM, New York, 2003), pp. 73–80

Chapter 7

Assessing the Usefulness of Information in the Context of Coalition Operations



Claire Saurel, Olivier Poitou, and Laurence Cholvy

Abstract This chapter presents the results of a study aiming at restricting the flow of information exchanged between various agents in a coalition. More precisely, when an agent expresses a need of information, we suggest sending only the information that is the most useful for this particular agent to act. This requires the characterization of “the most useful information.” The model described in this chapter defines a degree of usefulness of a piece of information as an aggregation of several usefulness degrees, each of them representing a particular point of view of what useful information might be. Specifically, the degree of usefulness of a piece of information is a multifaceted notion which takes into account the fact that it represents potential interest for the user with respect to his request, has the required security clearance level, can be accessed in time and understood by the user, and can be trusted by the user at a given level.

Keywords Useful information · Usefulness degree · Coalition

7.1 Introduction

In this, chapter, the problem of defining the usefulness of information is considered in the context of military coalition operations. Coalition operations require aggregation of information obtained from multiple systems since none of these systems alone provides enough information to achieve coalition goals. These systems are delivered and managed by different countries, which agree on combining their capabilities in order to create a global system which is more efficient than their own. Obviously, the individual systems have to coordinate in order to achieve the global tasks assigned to the coalition, and, for doing so, they have to exchange information. At the same time, utilizing all the information poses the risk that the volume of

C. Saurel · O. Poitou · L. Cholvy (✉)

ONERA, Toulouse, France

e-mail: claire.saurel@onera.fr; olivier.poitou@onera.fr; Laurence.cholvy@onera.fr

information flow increases drastically, leading to a loss of global efficiency. The problem of alleviating this risk is addressed in this chapter. The proposed solution is based on the following considerations: each agent (also called a user) who expresses a need of information (a request) has to be provided with only the information that is “the most useful to act.” This solution requires an understanding of the concept of information *usefulness* and the definition of *a degree of usefulness* of a piece of information.

The model proposed here is inspired by the Vector Model used in the field of information retrieval [1–3]. In the initial version of this model, any document in the information base is associated with a vector of attributes that are the most significant in a given corpus. These attributes are the terms that appear in the document, and the user request is associated with a vector of terms that describe his interests. The Vector Model approach uses the comparison between the semantic proximity of a request and a document represented by the proximity of their respective vectors. The assessment of the relevance of a document to a request is then done by using such distances as a cosine measure between their respective vectors, thus capturing a syntactic proximity of their contents. Some other distance functions can be defined, in order to specify different kinds of relevance concept. One advantage of such distance-based approaches is that they provide a numerical value of the similarity between documents. This level of similarity is interpreted as a degree of relevance of documents. An additional strength of the Vector Model is its implementation simplicity. These are the main reasons why the Vector Model approach has been used extensively. Multiple refinements of the basic model have been proposed by, for example, introducing the weighting factors, which take into account the size of the document or the request, and the frequency of the terms in the overall collection of documents considered. The most well-known weighting scheme is the TF-IDF (Term Frequency-Inverse Document Frequency factor) [3, 4] that introduces a concept of term importance. This scheme analyses a combination of the term frequency in the document with the term of its scarcity in the collection of documents in order to favor most discriminant terms.

The concept of *usefulness of information* is more than *the relevance of information*. Of course, to be considered as useful, a piece of information must be relevant to the request. For instance, if a user wants to know the condition of a main road, the condition of small roads in the area is not relevant, since it does not satisfy the user’s information need. But in addition to being relevant, the information must also be current, for example, telling the user the main road condition a month ago is not useful since it may have changed. In the same way, a measurement provided by a totally unreliable sensor is useless. Moreover, a given relevant, current, and reliable piece of information is useful to the user only if it can be read and understood timely. For instance, a 20-page document describing the condition of a main road is not useful to the user if the road condition has to be known immediately. Moreover, a document written in a language unknown to the user is not useful either. Thus, usefulness needs to be characterized by several attributes.

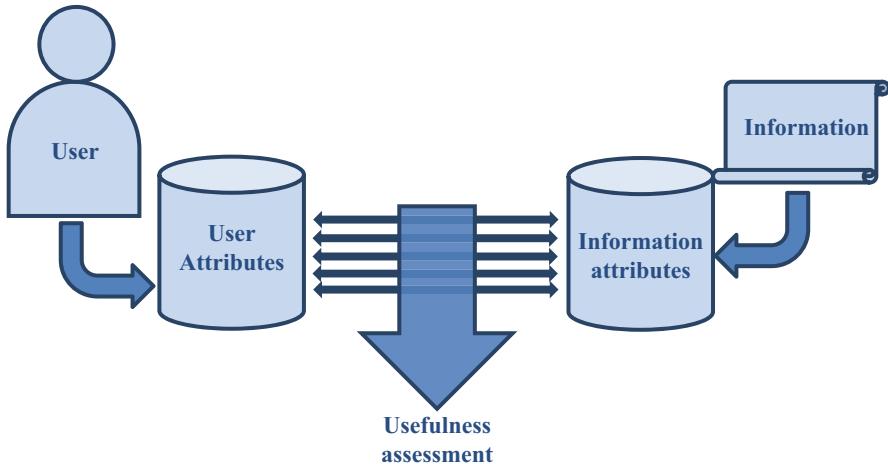


Fig. 7.1 Assessment of a global degree of usefulness for a piece of information

So, let us assume that a user has expressed a need for information by a request, and let us consider a piece of information as being a document, an image, a measurement provided by a sensor, etc. In order to estimate the usefulness of a piece of information to the user, we propose (Fig. 7.1) to:

- Model a user by a set of attributes that characterize the information need
- Model a piece of information by a set of attributes that characterize it
- Use attributes to compute a usefulness degree focused on a given point of view
- Aggregate the different usefulness degrees obtained from the different points of view to a global degree of usefulness

The rest of the chapter is organized as follows. Section 7.2 presents the various attributes used to model a user and the needs of information. Section 7.3 defines several usefulness degrees. Section 7.4 shows how to get a global usefulness degree by integrating various specific usefulness degrees. Section 7.5 presents an example illustrating the proposed approach. Finally, the chapter concludes in Sect. 7.6.

7.2 Modeling Information and Users

We consider that a piece of information is useful for the user if it is relevant to his/her request, temporally valid, sufficiently reliable, understandable by the user and if it can be read by the user in time. Consequently, we model information and users as vectors of attributes which deal with notions like topicality, time, reliability, languages.

7.2.1 Preliminaries

Let us introduce the following domains of values:

- $\text{Eval} = \{\text{A}, \dots, \text{E}\} \times \{1, \dots, 6\}$ is the set of all the possible information evaluation degrees as defined by STANAG 2511 (see Appendix A for a reminder of STANAG 2511). Moreover, we define an order on Eval as follows:

$$(X_1, Y_1) > (X_2, Y_2) \text{ iff } X_1 > X_2 \text{ or } (X_1 = X_2 \text{ and } Y_1 > Y_2)$$

- C is a finite set of access restriction (also called confidentiality) degrees associated with a total order $<$. For instance, $C = \{\text{restricted}, \text{confidential}, \text{secret}, \text{top secret}\}$ with $\text{restricted} < \text{confidential} < \text{secret} < \text{top secret}$.
- L is a set of values representing the different languages that are spoken in a coalition. For instance, $L = \{\text{English}, \text{French}, \text{German}\}$.
- F is a set of values representing the different formats representing the pieces of information which are exchanged in the coalition. For instance $F = \{\text{Word}, \text{pdf}, \text{L16}\}$.
- E is a finite set of values associated with a total preorder denoted \leq . For instance $E = \{\text{small}, \text{medium}, \text{strong}\}$ with $\text{small} \leq \text{medium} \leq \text{strong}$. Any value in E intends to represent the amount of effort needed to decode a given piece of information or the amount of effort a given user may spend to read a given piece of information.
- T is a finite set of topics related to pieces of information exchanged in the coalition.
- D is a set of integers that model dates.

7.2.2 Modeling Information

The different attributes that model a piece of information I , as well as their value domains, are given in Table 7.1.

Table 7.1 The information vector

Attribute names	Attribute values
$\text{Eval}(I)$	An element of Eval .
$C(I)$	An element of C
$L(I)$	A subset of L
$F(I)$	A subset of F
$E(I)$	An element of E
$T(I)$	A subset of T
$TV(I)$	$[d1, d2]$ where $d1$ and $d2$ belong to D
$\text{Emit}(I)$	An element of D

These attributes are further defined in the following way:

- $\text{Eval}(I) = \langle r, c \rangle$ is the evaluation degree of I according to STANAG 2511. The value r quantifies the reliability of the source that emitted I and the value c quantifies the credibility of I.
- $C(I)$ is the confidentiality level of I.
- $L(I)$ is the set of languages used to express I.
- $F(I)$ is the set of different formats that support I.
- $E(I)$ is a value in E that quantifies the amount of effort needed to read and analyze information I. It could be defined in a very simplistic way as the size of the support or in a smarter way with a linguistic analyzer.
- $T(I)$ is the set of topics that I is about.
- $TV(I)$, the temporal validity of I is the interval $[TV(I)_{\text{start}}, TV(I)_{\text{end}}]$ during which, I is valid.
- $\text{Emit}(I)$ is the date at which I is emitted.

7.2.3 Modeling Users

A user is a person who, at a given moment, formulates a query that expresses a need for information. The user is modeled by a set of attributes belonging to different domains. Some of these attributes depend on the query (for instance, the set of topics that the query is about, the deadline of the information needed), but some others only depend on the person (for instance, the languages that it can understand, its security clearance level etc.). All these attributes are used to characterize a useful piece of information.

Different attributes which model user U and their value domains are given in Table 7.2. One can notice that the first four attributes do not depend on the user's request.

These attributes are further defined in the following way:

- $\text{Eval}(U)$ defines the smallest evaluation degree of information that U is willing to accept. As a consequence, any piece of information whose evaluation degree is less than $\text{Eval}(U)$ will be considered as useless for U.

Table 7.2 A vector attributed of user U

Attribute names	Attribute values
$\text{Eval}(U)$	An element of Eval
$C(U)$	An element of C
$L(U)$	A subset of L
$F(U)$	A subset of F
$E(U)$	An element of E
$T(U)$	A subset of T
$TV(U)$	$[d1, d2]$ where $d1$ and $d2$ belong to D
$\text{Deadline}(U)$	An element of D

- $C(U)$ is the security clearance level of U . Any information whose confidentiality level is greater than $C(U)$ will be considered as nonuseful for U .
- $L(U)$ is the set of all the languages that U understands. As a consequence, any piece of information expressed in a language that does not belong to $L(U)$ will be considered as non-useful for U .
- $F(U)$ is the set of the representation formats U accepts. As a consequence, any piece of information coded in a format that does not belong to $F(U)$ will be considered as nonuseful for U .
- $E(U)$ is the amount of effort U can make to process answers. For instance, if $E(U)$ is low, then a big document requiring a huge effort will be considered as nonuseful for U .
- $T(U)$ is the set of the topics related to the request. As an example, if the request is related to *zone Z geography* (for instance, what is the length of that river? What is the height of that hill?), any information that does not concern the geography of zone Z (for instance, information about the population of Z) will be considered as useless for U .
- $TV(U)$ is an interval $[TV(U)_{\text{start}}, TV(U)_{\text{end}}]$ of dates during which information that U expects as answers is valid. For instance, if the request *what is the condition of that road?* is associated with the interval $[d1, d2]$, this means that U wants to know the road condition between dates $d1$ and $d2$.
- $\text{Deadline}(U)$ is the latest moment at which U wants to receive the requested information. Any information that cannot reach U before $\text{Deadline}(U)$ will be considered as useless.

7.3 Degrees of Usefulness

Let U be a user and I a piece of information, each of them being modeled by their 8 attributes. These attributes define several degrees of usefulness as listed below.

- The degree of usefulness of I for U with respect to attribute Eval is defined by:

$$\text{Usefulness}_1(I, U) = 1 \text{ iff } \text{Eval}(I) \geq \text{Eval}(U)$$

$$\text{Usefulness}_1(I, U) = 0 \text{ otherwise}$$

i.e., relative to attribute Eval , I is useful for U iff its evaluation degree is greater than the smallest evaluation degree required by U .

- The degree of usefulness of I for U with respect to attribute C is defined by:

$$\text{Usefulness}_2(I, U) = 1 \text{ iff } C(I) \leq C(U)$$

$$\text{Usefulness}_2(I, U) = 0 \text{ otherwise}$$

i.e., with respect to attribute C , the information I is useful for U iff its confidentiality level degree $C(I)$ is less than the confidentiality level degree of U :

$C(U)$. Any information whose confidentiality level degree is greater than $C(U)$ will not be sent to U because U is not authorized to read it (U does not have the required security clearance level).

- The degree of usefulness of I for U with respect to attribute L is defined by:

$$\text{Usefulness}_3(I, U) = 1 \text{ iff } L(I) \subseteq L(U)$$

$$\text{Usefulness}_3(I, U) = 0 \text{ otherwise}$$

i.e., with respect to attribute L , the information I is useful for U iff U understands all the languages used to describe I .

- The degree of usefulness of I for U with respect to attribute F is defined by:

$$\text{Usefulness}_4(I, U) = 1 \text{ iff } F(I) \subseteq F(U)$$

$$\text{Usefulness}_4(I, U) = 0 \text{ otherwise}$$

i.e., with respect to attribute F , the information I is useful for U iff U can process all the formats in which I is represented.

- The degree of usefulness of I for U with respect to attribute E is defined by:

$$\text{Usefulness}_5(I, U) = 1 \iff E(I) \leq E(U)$$

$$\text{Usefulness}_5(I, U) = 0 \text{ otherwise}$$

i.e., with respect to attribute E , the information I is useful for U iff the effort which is needed to process I is less than the effort U can provide.

- Two degrees of usefulness can be defined with respect to attribute T .

- The degree of appropriateness is defined by:

$$\text{Usefulness}_{6_1}(I, U) = \frac{|T(I) \cap T(U)|}{|T(U) \cup T(I)|}$$

- The degree of topical coverage is defined by:

$$\text{Usefulness}_{6_2}(I, U) = \frac{|T(I) \cap T(U)|}{T(U)}$$

- The degree of usefulness of I for U with respect to the attribute of temporal validity is informally defined by the amount of overlap between the two intervals; formally its definition is:

$$\text{Usefulness}_8(I, U) = 1 \text{ iff } \text{CurrentTime} + \text{CommunicationDuration}(I, U)$$

$$\leq \text{Deadline}(U) \quad 0 \text{ otherwise,}$$

- With respect to attribute Deadline, the degree of usefulness of I for U is defined by:

$$\text{Usefulness}_8(I, U) = 1 \text{ iff } \text{CurrentTime} + \text{CommunicationDuration}(I, U) \\ \leq \text{Deadline}(U) \quad 0 \text{ otherwise,}$$

where Current Time is the moment at which the evaluation is done and Communication Duration(I,U) is a value that represents a (worst case) transmission time of the information I to the user U if known (assumed to be null in case of having no knowledge on that point).

The usefulness degrees are defined in such a way that they have the same interpretation as their values are high or low, i.e., for a given piece of information, the higher the degree, the more useful the information is to the user. These degrees are defined either as ratios that take their value in a normalized interval [0,1] or as pseudo-Boolean values (0 or 1).

7.4 A Global Degree of Usefulness

This section presents the way we compute a global degree of usefulness. First, we assume that the user has expressed a preference order for the eight previous degrees of usefulness. This preference order reflects the relative importance given by the user to the particular criterion under consideration. Then, a global degree of usefulness is computed by aggregating the eight degrees while taking into account their relative importance.

7.4.1 *Choice of an Aggregation Function*

The choice of an aggregation function depends on several assumptions about the criteria to aggregate [5], among which:

- Separability of criteria:* Criteria are separable if there is no interaction between them. The lack of independence may produce a bias since they can be counted twice.
- The existence of absorbent attributes:* For a given piece of information, a Boolean usefulness dimension is absorbent if, whenever its value is null, the global aggregated usefulness value is null too, whatever a user preference order being defined between criteria.
- Compensability of the function:* The compensability is the ability to obtain a high global evaluation in spite of the low evaluation degrees of criterion to be aggregated (intuitively several good marks compensate a bad one). The compensability property is not always desirable, and one can prefer to have limited criteria compensability where compensation is hard or impossible.

Min or max aggregation methods only consider the numerical order between values of the criteria or dimensions to be aggregated. Therefore, they cannot capture any user preference order defined for the criteria. Most usual and simple additive aggregation methods, such as average, are totally compensatory. Compensation control may be performed by limiting the compensation function domain and adjusting the compensation inside this domain.

It is possible to introduce minimum score requirements to restrict the compensation function domain. This is useful to avoid the incorporation of absorbent attributes into the aggregation process.

Weighted min and max aggregation methods [6] address the area of compensation adjustment. They are not directly compensatory though still offering a linear compensation while only the gradient is changed. In addition, they introduce a vector of static numerical weights which could be considered as somewhat arbitrary. Some more complex weighted techniques try to make the compensation control more subtle, but in this family of functions, static numeric weights that are associated with each criterion (or criteria combination) are difficult to choose and control. Limits to these weighted approaches are certainly reached with Choquet integrals [7] and its 2^n weights for n criteria that potentially offer to choose one compensation linear gradient between each criterion. Weighted aggregation methods require techniques for computing weight values. These techniques are mostly heuristic and require the availability of experimental data to calibrate and validate their parameters. In industrial and even more in defense applications, such data are often confidential and not available. Moreover, most of the simple additive aggregation methods (such as the weighted average) ought to be avoided if criteria separability has not been verified.

In order to address the problems with the weighted aggregation methods in a new way, Yager has proposed prioritized aggregation functions [8]. One of Yager's main ideas is that, rather than attributing somewhat arbitrary numeric weights, the user should only provide a preference relation (i.e., a preorder) for the criteria. The numerical values of criteria weights are then no longer static and depend on each alternative (here, each piece of information). During evaluation of each alternative, the weight that is to be associated with each criterion will be computed based on the rank of the criteria class in the preference relation and the scores obtained in user preferred criteria classes. Specifically, a 1 value is given to the weight w_1 associated with the most preferred criteria class, and all criteria scores v are such that $0 \leq v \leq 1$. Then the next weight value is computed by the formula $w_{i+1} = w_i \times m_i$, where m_i is the lowest score of the alternative for all criteria in the criteria class indexed by i . Thus, an alternative gaining a low score for an important criteria, will result in small weights for less important criteria, hence reducing the impact of less preferred criteria scores on the overall evaluation score of the alternative.

Yager's approach may be particularly efficient from the compensability control point of view [9] —the compensation function is no more linear and becomes exponential—and may even offer absorbing elements opportunity. Indeed, if the score corresponding to a given criterion considered as absorbent is null, then all less preferred criteria will inherit a null weight which means that they will not

be taken into account at all (this is what is previously referred to as absorbency property). This is particularly powerful since applying this to the most preferred criteria requires considering them as necessary conditions, which must be fulfilled for the alternative to be considered (absorbency property).

The ability of Yager's algorithm described in [8] to overcome unwanted compensation effects is not the only advantage it provides. Other advantages are that it doesn't require either a lot of experimental input data or somewhat arbitrary numerical weights since it relies mostly on user's preferences between criteria. These advantages are particularly important for applications where a sufficient experimental data set is unavailable due, for instance, to confidentiality constraints that frequently appear in an industrial or military context and therefore to the needs of our application.

The following subsection describes the aggregation operator chosen for calculating the global usefulness degree, which is widely inspired by the operator introduced in [8].

7.4.2 *Definition of a Global Usefulness Degree*

Let us assume that the user expresses his preference by grouping the different degrees of usefulness in three priority classes. Class C1 comprises the degrees of usefulness corresponding to the attributes that are the most important to the user. The fulfillment of those degrees is considered as mandatory in our application: they correspond to absorbent attributes. Class C2 comprises the degrees of usefulness corresponding to the attributes whose importance is medium to the user. Class C3 comprises the degrees of usefulness corresponding to the attributes that are the least important to the user. The global degree of usefulness of I for U is then defined as follows:

$$\text{Usefulness}(I, U) = \sum_{i=1..3} p_i \times \text{Usefulness}_{C_i}(I, U) \text{ with :}$$

- $p_1 = 1$
- $\text{Usefulness}_{C_1}(I, U) = \prod_{j \in C_1} \text{Usefulness}_j(I, U)$.
- $p_2 = \min_{C_1} \text{Usefulness}_{C_1}(I, U)$
- $\text{Usefulness}_{C_2}(I, U) = \sum_{j \in C_2} \text{Usefulness}_j(I, U)$
- $p_3 = p_2 \times \min_{C_2} \text{Usefulness}_{C_2}(I, U)$
- $\text{Usefulness}_{C_3}(I, U) = \sum_{j \in C_3} \text{Usefulness}_j(I, U)$

In this definition, the degrees of class C1 are absorbing (i.e., as soon as at least one of these degrees equals to 0 then the global degree is bounded to be 0). This reflects their critical importance to the user.

7.5 Illustration

A Combat Search And Rescue (CSAR) scenario vignette, inspired by [10] has been chosen to illustrate the proposed approach. The CSAR vignette encompasses different agents with associated characteristics, roles, and missions. Among them are a helicopter bringing troops to a site, aircrafts ensuring air superiority or providing communication availability and observations, as well as deploying personnel to the ground with the AMI (i.e., FRIENDS) extraction mission itself. This scenario vignette offers different information request profiles ranging from a background long-term observation/intelligence task to a specific short-term mission. Of course, a well-equipped person can manage a large amount of information and would be far less demanding about information quality than a single person on the operation theater who does require a high degree of information usefulness.

A concrete application of our proposed approach is to add smart filtering as a part of a communication layer between agents of a coalition. To do so, a large number of parameters are to be selected, and users and information profiles have to be created and edited. This section first describes the tool that will help to prepare and to tune the filter. Then it shows how this filter could operate in a practical example.

7.5.1 The Usefulness Evaluation Tuning Tool

The main prototype that has been developed is a tool that can be used to configure the smart filter to come. It allows editing user profiles and requests as well as information profile, by giving a value to their different attributes and saving them as a whole to easily recall them on demand. Finally, it displays the detailed result of the evaluation of the selected information usefulness for the selected user. The main interface of this prototype (Fig. 7.2) is made of three columns: the first one describes the candidate destination user and its information need, the second column supports the description of the evaluated information, and the third column shows the result of the usefulness assessment.

Figure 7.3 zooms in the first column for a closer look on some of the main attributes of the candidate destination user profile.

The user profile attributes that can be found in this figure are the following:

- *Cotation trigger* is expressed with the classical double scale (B3—in the illustration). It indicates the minimum information evaluation degree expected by the user (see “Modeling Information” and “Modeling Users” sections).
- *Known languages* are the textual or vocal information language that can be handled by this user (see “Modeling Information” and “Modeling Users” sections).
- *Usable supports* are the communication means that are available to the user; it encompasses or has an impact on information media type, communication protocol, and equipment that can be used. It supports and extends the concept of representation format (see “Modeling Information” and “Modeling Users” sections).

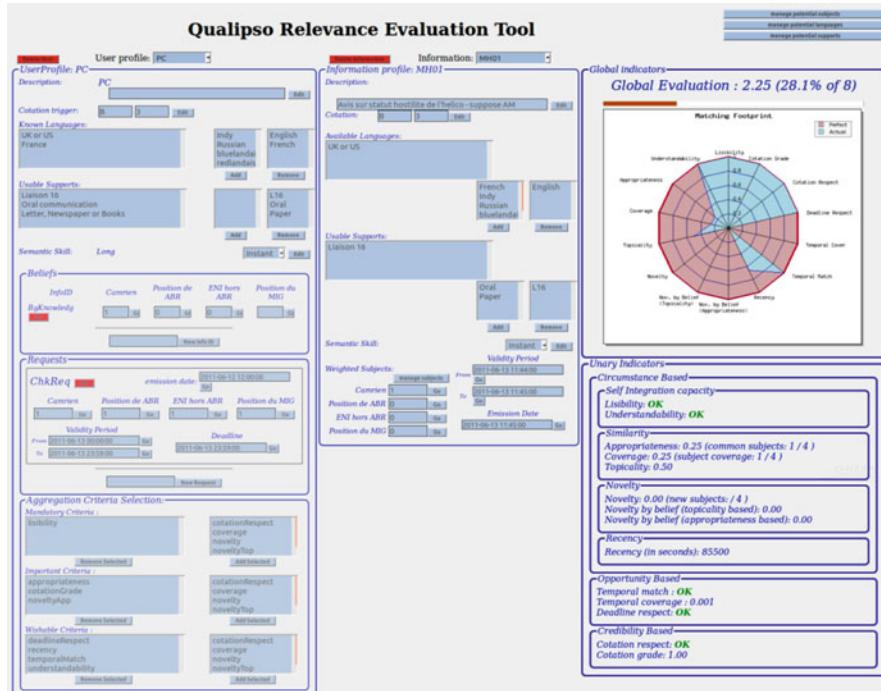


Fig. 7.2 Overview of the information usefulness assessment tool

Delete User

User profile: PC

UserProfile: PC

Description: PC

Cotation trigger: B 3

Known Languages:

- blueland
- UK or US
- France
- redland

Usable Supports:

- Liaison 16
- Oral communication
- Letter, Newspaper or Books

Semantic Skill: Long

Fig. 7.3 User profile tune-up window

The screenshot shows a software interface for managing information requests. At the top left, it says "Requests". Below that, there's a section for "ChkReq" with a "Del" button. To the right, it shows the "emission date: 2011-06-12 12:00:00" and a "Go" button. Below the date, there are four rows of data, each with a label and a "Go" button: "Camrien" (value 1), "Position de ABR" (value 1), "ENI hors ABR" (value 1), and "Position du MIG" (value 1). Underneath these rows, there's a "Validity Period" section with "From" and "To" fields, both set to "2011-06-13 00:00:00" and "2011-06-13 23:59:00" respectively, followed by "Go" buttons. To the right of the validity period, it says "Deadline" with a "2011-06-13 23:59:00" field and a "Go" button. At the bottom of the main form area, there's a "New Request" button.

Fig. 7.4 User information request description

- *Semantic skill* reflects the ability to extract specific information from an information set (time to read a long document for example as opposed to real-time reaction requirement). It reflects the concept of effort (see “Modeling Information” and “Modeling Users” sections).

Some user profiles are preset in advance and can be made available when needed. They have been given a name corresponding to their scenario counterpart and can be selected with the first field appearing in the figure. When doing so, all corresponding attributes are given a value accordingly. However, any modification remains possible to those settings in order to explore slightly different user profiles.

Then two windows are offered to describe the information request and the usefulness priorities to the user.

The first window (Fig. 7.4) gathers attributes describing the information request of the user with:

- An *emission date* is the date and time of the information request.
- A degree of interest for each subject/keywords of a preestablished list that reflects the attribute $T(U)$ in section “Modeling Users.”
- The *deadline* of the information request (will be used to decide if there is still an interest to send an information to this user or if it is too late) that reflects the attribute $Deadline(U)$ in section “Modeling Users”.
- The temporal interval on which information is desired (temporal validity (TV) in section “Modeling Users”).

The second window (Fig. 7.5) enables the user to tune up the aggregation algorithm by describing the user’s preference on the attributes. Each criterion is put into one of the priority classes: “mandatory,” “important,” and “desirable.” For example, every user would certainly put readability as “mandatory” meaning that it is useless to send him information that cannot be read(by lack of equipment or

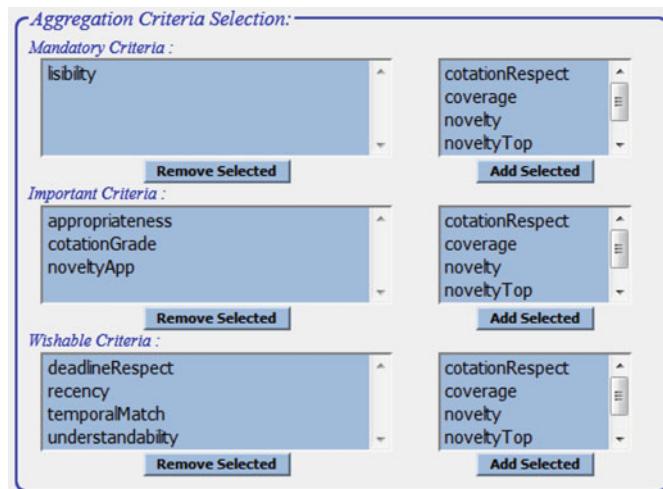


Fig. 7.5 Usefulness priorities tune up for the user

unknown language). The “mandatory” class is the one that is absorbent which means that no information is sent that does not meet all the criteria in this priority class.

Some usefulness dimensions may not be selected in any priority class as they are not considered relevant by the user. These dimensions are not considered as usefulness criteria and their corresponding usefulness degree will not be integrated into the global score (usefulness) computation.

The second column describes the information attributes (see Fig. 7.6).

Usefulness attributes are similar to those in the first column, describing:

- Contextual attributes (metadata) on the information evaluation: emission date, language, estimated associated semantic effort, and information technical support
- The contents of the information via the topic matching list and validity period

The third column displays the results of the evaluation using two windows. The upper window (Fig. 7.7) is showing the global, aggregated result in a numerical and graphical way.

In the upper part of Fig. 7.7, the numerical result is both given as a raw value as well as a ratio of the obtained score to the best possible score, i.e., the score that would obtain an information exactly matching the selected usefulness criteria with their associated priority (see Fig. 7.5). In the lower part of the figure, a graphical view relies on a radar diagram to give a more precise view of how the information fulfills or not the different usefulness criteria of the user.

The lower window (Fig. 7.8) provides finer details on the scores obtained by the information on each usefulness degrees.

Delete Information Information: MD01

Information profile: MD01

Description: Detection of an helicopter in the BlueLand **Edit**

Cotation: B [3] **Edit**

Available Languages:

UK or US	French Indy Russian bluelanda	English
----------	--	---------

Add Remove

Usable Supports:

Liaison 16	Oral Paper	L16
------------	------------	-----

Add Remove

Semantic Skill:

Instant **Edit**

Validity Period

Weighted Subjects: **manage subjects**

Camrien	1	Go
Position de ABR	0	Go
ENI hors ABR	0	Go
Position du MIG	0	Go

From: 2011-06-13 11:40 Go

To: 2011-06-13 11:41 Go

Emission Date: 2011-06-13 11:41 Go

Fig. 7.6 Information (support) description

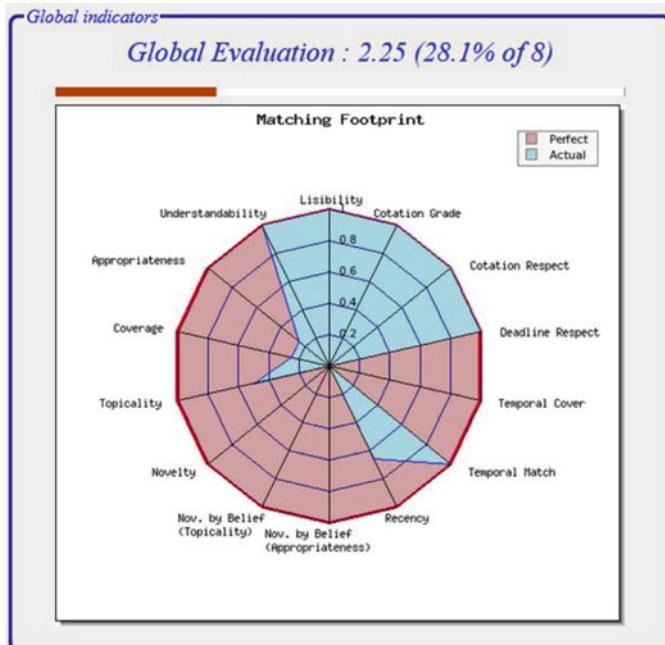


Fig. 7.7 Global usefulness assessment presentation



Fig. 7.8 Degree by degree usefulness assessment window

7.5.2 Illustration of the In-The-Loop Usage

The prototype described above should be seen as a tool that helps to tune the smart filtering mechanisms that could be embedded into a communication system in an operational setup. The design of this setup is beyond the scope of this chapter. However, this section offers some preliminary considerations.

A centralized architecture that receives all information and smartly redistributes it is neither realistic (terrain constraints from a military operation theater) nor the most theoretically efficient option (all emissions are done to a centralized equipment, while some of them should be avoided). In a distributed approach, the decision to send or not to send a message can be made by finding out whether the message is assigned a score above a given threshold and, in particular, whether all the intended receiver's mandatory criteria are met. On the receiver side, the degree of usefulness then helps to give emphasis to the most useful information.

Even in a situation that brings a high level of stress and does not let the user carefully take into account all information received, emphasizing the most important

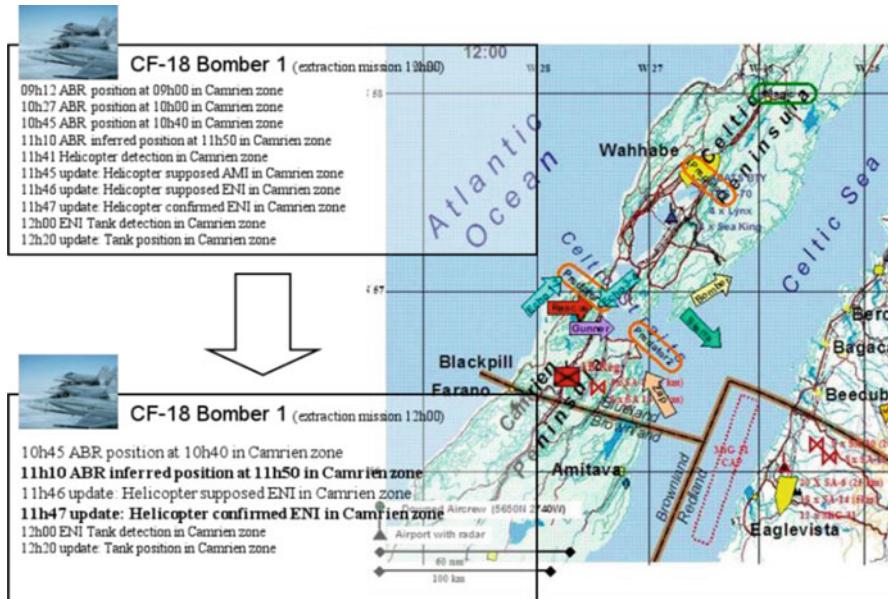


Fig. 7.9 Messages are filtered and emphasized depending on their usefulness for the user

ones improves the probability that the user will pay attention to this information. For instance, if the information is read from a screen, emphasizing its importance can be as simple as transposing the usefulness score in a different size or boldness or color as in Fig. 7.9. In this illustration, the upper list of messages is the one that would be displayed without any usefulness smart filtering, while the one below is an example of a usefulness smart filtering result: some messages have been discarded, and some put in large bold, while some have a very small font size.

Lastly, it is important to notice that this is clearly an over simplified way of exploiting the results of a usefulness evaluation of information. This evaluation should be integrated into a wider process of delivering the good information to the relevant person at the right time. More specifically, usefulness score should be an input to experts in human-machine interfaces from scientific domains such as ergonomics, cognition, etc. that deal with attention issues in a situation of high stress level (e.g., tunneling effect). They will be able to devise an algorithm that, from the usefulness score, select the medium (or media) and format(s) that should be used to communicate the information to the user (textual display like here, dedicated red alert flashing light, sounds, vibrations, etc.). This further step is outside of the scope of our work; it remains to be done in order to get experimental feedbacks and then improve our model for a given application, via, for example, operational mission simulations. Then it could lead to complementary investigations, such as the problem of planning useful information provision (inside a coalition during a mission), or the problem of the impacts of useful information availability on the mission performances.

7.6 Conclusion

This chapter presented a model for an assessment of a degree of usefulness of a piece of information. The model can be seen as an extension of the Vector Model approach used in retrieval of relevant information given a user's request. It is based on an aggregation of certain semantic measures of proximity between the attributes of the retrieved information and user request while each measure is capturing a usefulness dimension. Our approach describes some drawbacks that could be addressed in future work. For instance, the evaluation attribute values in the presented method have been manually and subjectively defined. Efforts might be dedicated to defining a process where these values can be formally computed. Another improvement might be to consider a taxonomy of topics that would enable some reasoning about them. A reasoning scheme can be added by taking into account equivalence and subsumption relations between topics rather than a direct syntactic comparison. Nevertheless, the work presented in this chapter contributes to the definition and understanding of a very important aspect of quality of information: its usefulness to users.

Acknowledgments This study was granted by DGA (Direction Générale de l'Armement) in the context of a Specific Agreement between France and Canada (AS 32).

A.1 Appendix A: STANAG 2511 Model for Information Evaluation

One model of Information Evaluation, widely used in military context, is STANAG 2511 [11–13]. According to this standard, the aim of information evaluation is to indicate the degree of confidence that may be placed on any item of information which has been obtained for intelligence. This is achieved by adopting an alphanumeric system of rating which combines a measurement of the reliability of the source of information with a measurement of the credibility of that information when examined in the light of existing knowledge. These two measures are defined as follows:

Reliability of the source is designated by a letter between A and F as defined below.

- A source is evaluated A if it is completely reliable. It refers to a tried and trusted source which can be depended upon with confidence.
- A source is evaluated B if it is usually reliable. It refers to a source which has been successfully used in the past, but for which there is still some element of doubt in particular cases.
- A source is evaluated C if it is fairly reliable. It refers to a source which has occasionally been used in the past and upon which some degree of confidence can be based.

- A source is evaluated D if it is not usually reliable. It refers to a source which has been used in the past but has proved more often than not unreliable.
- A source is evaluated E if it is unreliable. It refers to a source which has been used in the past and has proved unworthy of any confidence.
- A source is evaluated F if its reliability cannot be judged. It refers to a source which has not been used in the past.

Credibility of information is designated by a number between 1 and 6 signifying varying degrees of confidence as indicated below.

- If it can be stated with certainty that the reported information originates from another source than the already existing information on the same subject, then it is classified as “confirmed by other sources” and rated 1.
- If the independence of the source of any item of information cannot be guaranteed, but, if, from the quantity and quality of previous reports, its likelihood is nevertheless regarded as sufficiently established, then the information should be classified as “probably true” and given a rating of 2.
- If despite there being insufficient confirmation to establish any higher degree of likelihood, a freshly reported item of information does not conflict with the previously reported behavior pattern of the target, the item may be classified as “possibly true” and given a rating of 3.
- An item of information which tends to conflict with the previously reported or established behavior pattern of an intelligence target should be classified as “doubtful” and given a rating of 4.
- An item of information which positively contradicts previously reported information or conflicts with the established behavior pattern of an intelligence target in a marked degree should be classified as “improbable” and given a rating of 5.
- An item of information is given a rating of 6 if its truth cannot be judged.

References

1. G. Salton, *Automatic Information Organization and Retrieval* (McGraw-Hill, New York, USA, 1968)
2. J.M. Ponte, W.B. Croft, in *A Language Modeling Approach to Information Retrieval*, Research and Development in Information Retrieval. Proc. ACM-SIGIR (Melbourne, Australia, 1998)
3. G. Salton, M. J. McGill, *Introduction to Modern Information Retrieval* (McGraw-Hill, New York, USA, 1986)
4. G. Salton, A. Wong, C. Yang, A vector space model for automatic indexing. Commun. ACM **18**, 613–620 (November 1975)
5. M. Grabisch, P. Perny, Agrégation multi-critères, in *Logique floue, Principes, Aide à la Décision*, e, ed. by B. Bouchon-Meunier, C. Marsala (Eds), (Hermès, Paris, France, 2003), pp. 82–120
6. D. Dubois, H. Prade, Weighted minimum and maximum operations in fuzzy set theory. Inf. Sci. **39**, 205 (1986)
7. G. Choquet, Theory of Capacities, in *Annales de l'Institut Fourier tome 5*, (Imprimerie Durand, Luisant, France, 1953), pp. 131–295

8. R.R. Yager, Prioritized aggregation operators. *Int. J. Approx. Reason.* **48**, 263–274 (2008)
9. C. da Costa Pereira, M. Dragoni, G. Pasi, in *A prioritized “and” Aggregation operator for multidimensional Relevance Assessment*. R. Serra and R. Cucchiara (Eds): AI*IA 2009, LNAI 5883, pp 72–81, 2009, eds. by R. Serra, R. Cucchiara. AI*IA 2009: Emergent Perspectives in Artificial Intelligence. AI*IA 2009., 2009
10. Top Aces Consulting (Quebec) and Cirrus Research Associates (Ontario) for DRDCV, Recognized air picture (RAP) and vignette development CSAR scenario, October (2005)
11. NATO, STANAG 2511 Intelligence Reports, NATO Unclassified (2003)
12. L. Cholvy, V. Nimier, in *Information Evaluation in Fusion*, Symposium RTO/IST «Military Data and Information Fusion», Prague (2003)
13. E. Blasch and al, URREF reliability versus credibility in information fusion (STANAG 2511), in *Proc. of the 16th International Conference on Information Fusion, (FUSION 2013)*, Istanbul (2013)

Chapter 8

Fact, Conjecture, Hearsay and Lies: Issues of Uncertainty in Natural Language Communications



Kellyn Rein

Abstract Humans are very important sources of information for intelligence purposes. They are multi-modal: they see, hear, smell, and feel. However, the information which they relay is not simply that which they personally experience. They may pass on hearsay, they form opinions, they analyze and interpret what they hear or see or feel. Sometimes they pass on ambiguous, vague, misleading or even false information, whether intentional or not. However, whether imprecise or vague, when humans communicate information, they often embed clues in the form of lexical elements in that which they pass on that allows the receiver to interpret where the informational content originated, how strongly the speaker herself believes in the veracity of that information. In this chapter, we look at the ways in which human communications are uncertain, both within the content and about the content. We illustrate a methodology which helps us to make an initial evaluation of the evidential quality of information based upon lexical clues.

Keywords Opinions · Quality of information · Lies · Uncertain information · Natural language

8.1 Introduction

Information plays a central role in surviving and coping with today's world. Effective decision-making depends on the quality, completeness and trustworthiness of the information which decision-makers have at their disposal. Actionable intelligence for situational understanding is garnered both from human sources, in the form of text or speech, and from devices such as radar, acoustic arrays, ground sensors, video, etc., which report data on physical phenomenon in either digital or analog form. Both human and non-human sources may pass on data which is

K. Rein (✉)
Fraunhofer FKIE, Wachtberg, Germany
e-mail: kellyn.rein@fkie.fraunhofer.de

not fully accurate or trustworthy; however, there are some significant differences in the causes of the inaccuracies; understanding and analyzing these differences appropriately can improve the quality of the intelligence received.

A sensing device such as a video camera, a radar dish or a ground sensor is a neutral observer of events taking place within its scope. A video camera records light waves (and, if so equipped, also audio), a radar dish gathers data about movements with the range of its sweep, and a ground sensor documents vibrations detected near its location. Data from devices is always *historical*, that is, the physical events – motion, temperature, light, etc. – recorded by the sensor have actually taken place (fusion algorithms based upon data received from the sensor may project future states, but these are independent of the sensor itself). Sensor data is also *neutral* and *impartial*; the sensor has no vested interest in the meaning of its recordings. Furthermore, the type of data delivered by a device is always the same; the device records only certain types of physical phenomena. A thermometer does not measure motion, a motion detector does not measure temperature, an acoustic sensor does not record light waves.

In contrast, humans are multi-purpose sensors. We see, hear, feel, taste and smell, and we communicate what we have sensed using natural language. Our capabilities in each of these observational modes vary from person to person based upon a number of factors including physical condition (e.g., excellent or poor eyesight?), background knowledge in the phenomenon observed (e.g., trained expert or layperson?) as well as expectations or assumptions about the phenomenon which is being observed (is this normal or out of the ordinary?). In other words, a human often, intentionally or unintentionally, pre-process, interpret, speculate on, or self-filter their observations thereby modifying the observation before passing it on. Furthermore, humans relay information about events which they have not personally experienced or observed, for example, in the form of hearsay or in discussion of future events which have not yet taken place. In addition, humans pass on opinion, speculation, assumption, and inferences. But one of the most significant differences between a device and a human is that, whereas the device may provide bad data due to malfunction or environmental factors, humans can – and do – lie, distort, or otherwise misrepresent information (“fake news”). Human beings, as sources, are therefore problematic on several levels, which we will discuss in further depth in a later section in this chapter.

Regardless of the type of source from which the information comes, it is vitally important for decision-makers to have a realistic idea of how good that information is. Ideally, we would only use information which is complete and which has been definitively confirmed as factual. However, the reality is that this is seldom the case, and that we often need to make decisions based on incomplete and uncertain information. While using uncertain, incomplete, misleading or incorrect information as the basis for action can be a recipe for disaster, there are many times when incomplete and uncertain information is all that is available to decision-makers. Realistically assessed, even partial and uncertain information can be used to great effect. For example, in 2012, when President Barack Obama gave the order for an

assassination attempt on Osama Bin Laden at his hideout in Pakistan, the President himself rated the odds of actually finding Bin Laden there as “50–50” [1].

Device-derived data may be uncertain or unreliable for a variety of reasons. For example, sensors may be affected by environmental conditions such as heat, humidity, or light conditions thereby producing unreliable data. Devices may also malfunction and fail. However, devices may be tested and calibrated under various conditions, giving decision-makers important information about the overall reliability of a given source under different physical situations. Also, the algorithms which operate upon device-derived data have been a research focus for several decades now and are quite mature and well understood. While there is still much work to be done, on many levels, this work focuses on refinement, improvement, and tweaking of existing technologies. (There are many fine works which provide excellent coverage of the advances in this field.)

Human-derived data, in contrast, remains problematic, in part because of the factors mentioned previously but also because the information humans convey is delivered in natural language. Natural languages are flexible enough to deal with almost every aspect of the human experience and thus are powerful communication tools. At the same time, the flexible power of natural language information often makes it inherently uncertain, in part because natural language utterances are often ambiguous, vague, open to (mis)interpretation, or even incorrect. However, because uncertain information can play a vital role in intelligence analysis, in particular in the prediction of events which may happen in the future, we need to accept this uncertainty and find strategies to deal with it appropriately.

Thus, it is important to identify and understand how natural language information content itself is less than certain. Sometimes the uncertainty has to do with the content itself in the form of ambiguity (“I saw her duck”), vagueness (“down the road”) or imprecision (“some”, “tall”, “many”), which may be context-dependent. Often uncertainty is conveyed by the speaker in the form of lexical forms that express the writer’s stance toward the truth of the proposition in the sentence (“unlikely”, “possibly”), indicate the origin of non-observed information (“people say”, “I assume”), and other constructs such as modal verbs or future tense (“might”, “will be”). For intelligence purposes, it is important to differentiate between uncertainty *within* the proposition and uncertainty *about* the proposition.

Our focus in this chapter is to look briefly at human beings as sources of information, in particular, their motivations behind reporting. We then examine in more depth how and in which ways natural language is uncertain, as mentioned previously, and look at how these affect the quality of the reporting. Finally, making sense of the huge volume of natural language information which is generated on a daily basis requires some automatic preprocessing (e.g., text analytics) to locate potentially useful information. We will present a methodology identifying, evaluating, and weighting the evidentiality of textual information, with particular emphasis on lexical markers which the source used to convey the origin of the information being passed on, as well as their assessment of the quality of that information.

8.2 The Human as Sensor

It appears, from all this, that our eyes are uncertain. Two persons look at the same clock and there is a difference of two or three minutes in their reading of the time. One has a tendency to put back the hands, the other to advance them. Let us not too confidently try to play the part of the third person who wishes to set the first two aright; it may well happen that we are mistaken in turn. Besides, in our daily life, we have less need of certainty than of a certain approximation to certainty. – Remy de Gourmont, *Philosophic Nights in Paris* [2] p. 127

Human information sources cover the spectrum from trained intelligence personnel, police and emergency services to informants and prisoners of war to victims, refugees and local residents as well as open sources such as newspapers, government documents, blogs, and social media. Dragos and Rein [3] have described a number of influences on these information sources, including, but not limited to the following list:

Subjectivity As in the quote above, any event which is observed by two or more people will almost assuredly always result in different reporting of that event, partly because each observer will perceive and interpret according to internalized information. For example, an observer's understanding of the event observed will vary based upon factors such as their specific skill set, background knowledge, or emotional involvement in the event. An intelligence observer is trained to note specific details which may escape the notice of an untrained civilian; the untrained civilian, however, may have a different interpretation of events based upon their cultural or social background. As a result, there may be several different, perhaps conflicting, interpretations rather than one unique "truth." The well-documented unreliability of eyewitness testimony is partially due to subjectivity.

Intention As mentioned in the introduction, in contrast to devices, people often intentionally alter information they pass on through conscious efforts to fabricate, conceal, or distort evidence for a specific purpose. These modifications may be the omission of important details which would provide a different context ("cherry-picking") in an attempt to distort the truth, or they may outright lies, that is, false stories created specifically to deceive. Sometimes, false statements are embedded among true statements in order to hide their falsity or to create ambiguity and uncertainty. It should also be noted here that the intention of an informant who provides distorted information is not necessarily always malice or deception: he may, for example, provide intentionally false or misleading information because he believes that is what the hearer wishes to hear, in order to win favor or gain attention, or in order to protect himself from personal negative consequences. Regardless of whether the intent is malicious or self-serving, the intentional distortion affects the quality of the information.

Opinion In addition to filtering and interpreting their observations, humans also express opinions, make assumptions, and speculate about events. If those opinions or assumptions are based upon experience and solid background, they may be valid analyses. However, sometimes, the source offering an opinion lacks sufficient

background information or contextual knowledge, with the result that this opinion should be treated with skepticism. Both untrained individuals and trained intelligence officers will often offer/provide interpretation or speculation; in particular, intermingling factual information and personal impressions or speculation often makes it difficult to separate fact from speculation. Complicating the issue is that the level of competence of an individual observer in one area of expertise is no indicator or guarantee of that individual's competence in other areas. For example, a military observer who has been trained to accurately identify military equipment, vehicles, and insignia of an enemy may not have the background to recognize or understand the significance of cultural and social aspects of civilian behavior which may be critical in conflict situations. In other words, the same human who may be seen as reliable in certain types of observations is less competent in others; thus, the expertise of the source in the domain of the information should be factored in when judging information received.

Hearsay Another unique characteristic of human communication which contributes significantly to the uncertainty of information is hearsay. Rather than reporting on direct observations or personal opinions, the source passes along information or opinions which they (claim to have) received from a secondary source. Hearsay is problematic on a number of levels. One issue is that the originating source may have neglected to give the full context for that information; this loss of context may later result in erroneous interpretation or understanding of the information received. A variation on this is when the secondary sources adds their own context during the retelling; adding an interpretation can distort the original message. A second problem is that the individual may pass on hearsay that has gone through a succession of individuals: "my brother told me that his wife's sister's husband talked to a guy who attended the meeting and said . . ." Each link increases potential distortion, misunderstanding, (mis)interpretation, and distance from the original context, with the danger that the reported information bears little resemblance to the original.

Open sources such as online news, blogs, and social media can be very useful in certain operations such as crisis management. However, these sources can also be very problematic for a variety of reasons. Online news sources often reference and link to articles on other sites, often with accompanying commentary. Should, however, the original source print a retraction or correction, the sites which have commented on the original seldom update their own text to reflect the changes, leaving the incorrect information in place. The incredibly rapid propagation ("going viral") of incorrect or incomplete information by retweeting or re-posting from other sites can result in confusion or worse. Lastly, much information which has been debunked continues to be available indefinitely on the Internet, with a result that erroneous information continues to be "rediscovered" and perpetuated.

Hidden networks One important part of the information verification process in intelligence is to have confirmation from multiple independent sources: if the sources report the same thing, we are generally confident that the data we have

received reflects reality. However, because humans are not always neutral, impartial observers, and because they may (knowingly or unknowingly) pass on hearsay or disinformation, we can only apply this rule of thumb if we know that the sources are, in fact, independent. Individual sources may be interconnected on the basis of easily identifiable similarities (live in the same location, went to the same university, or similar attributes), social relationships (friendships, membership in groups) or other types of interactions (e.g., social media), not all of which may be known to or easily discoverable by intelligence. Regardless of their nature, the existence of undetected connections may affect the independence of the information provided by different sources. As a result, information received from assumed (but not actually) independent sources may assume a higher credibility ranking under false assumptions, degrading information credibility, as what appears to be independent reporting may actually be hearsay or coordinated deception.

In conclusion, while necessary to intelligence communities on both the military and the civilian sides, the human as a sensor is not unproblematic. Understanding the psychological, social and emotional components of human-derived information, as well as knowing the limits of expertise domains of the sources, assists in producing a realistic assessment of source reliability, which has an effect on the assessment of the quality of information.

However, our focus in this paper is on the quality of information from human sources; a determination on the reliability of a source is only one factor for determining that. In the following section, we will examine the second factor: uncertainty in natural language (text) communications.

8.3 Overview of Uncertainty in Text

As already briefly touched upon, the information (“signal”) delivered by a human source is in the form of spoken or written text. Some statements are precise, accurate historical accounts of actual, directly observed events. However, a great deal of human communication is neither precise nor accurate nor historical and therefore may be uncertain in one or more aspects.

Consider the following sentence:

1. *I think someone said there were some animals in the road.*

There are several ways in which this sentence is uncertain. The statement begins with *I think*, which expresses belief or opinion, not knowledge; the speaker is letting the listener know that there is some doubt about the veracity of the rest of the statement. *Someone said* indicates that the assertion is hearsay, and therefore second-hand information, which may or may not have been correctly understood by the speaker and therefore not an accurate reflection of what the original source had, in fact, said.

Within the assertion itself, there are also elements of uncertainty (*there were some animals in the road*). For example, we do not know how many *some* is, nor do we know what type of *animals* there were, nor even which *road* is being referred to (although this may have been determinable from the text surrounding this statement, i.e., from context knowledge).

Although all of these reflect uncertainty, the elements mentioned above differ as to the uncertainty that they project within the statement. In fact, there are two basic categories of detectable uncertainty which appear at the sentence level in written text or speech:

- Uncertainty *within* the content, including:
 - Imprecision
 - Vagueness
 - Ambiguity and polysemy
- Uncertainty *about* the content, including:
 - Modal verbs
 - Modal adverbs (including “words of estimative probability”)
 - Hearsay markers
 - “Mindsay” markers → belief, inference, assumption, etc.
 - Passive voice

While “hearsay” is a familiar term, “mindsay” may be new to many readers. This term appears in [4] and is used to describe information which is not a result of direct observation nor of passing on secondhand information but which is the product of someone’s mind, e.g., inference or belief.

The first category – uncertainty within the content – is important to applications such as information fusion, in which assertions containing imprecise or vague descriptions may be analyzed to determine whether these assertions refer to the same or to different objects. For example, suppose we are searching for an individual described by police as 5 feet 9 inches (175 cm) tall with light brown hair – when we receive a report about a tall, blond man, how likely is it that the report is about the suspect? At what height is a person “tall”? When is hair light enough to be considered blond rather than light brown? The potential ranges of values for vague or imprecise terms are often dependent upon a specific domain. We will discuss these in more depth in the following section.

Uncertainty about the content is a type of noncontent meta-information delivered by the speaker to the listener which provides important clues about both the original source of the information, for example, if the speaker is describing and transmitting second- or thirdhand information, and if the assertion is the result of opinion, belief, or a logical process.

In [5] we classified the uncertainty within the content as “inexactness” and the uncertainty about the content as “evidentiality” as shown in Fig. 8.1. The latter type of uncertainty provides clues about how strongly or weakly the content of the sentence may be considered as “evidence” of the state or event described.

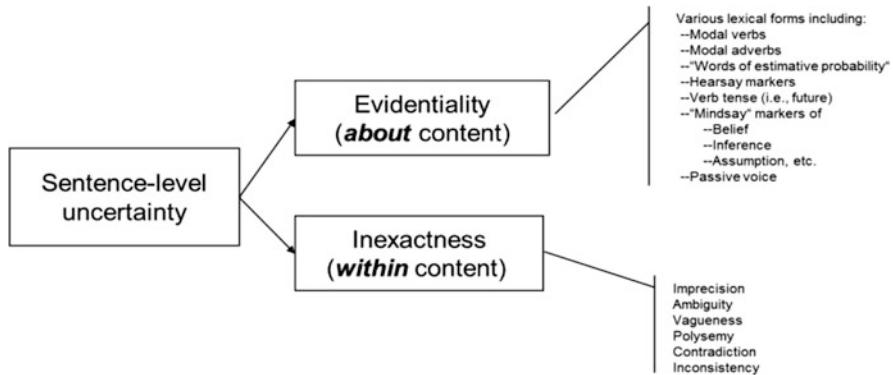


Fig. 8.1 Sentence-level uncertainty in natural language communications

In the following subsections, we will examine both types of uncertainty in more depth. We begin with uncertainty within the content, which is important for determining whether or not two or more reports refer to the same object or event. After that we will discuss the ways in which the speaker lets us know how strongly she believes what she says, communicates to us where the informational content came from (other human sources, logic or belief processes), which gives us an insight into how reliable the information content may be.

8.3.1 *Uncertainty Within the Content*

When humans communicate, we do more than convey basic facts. We express thoughts, hopes, and wishes, we speculate about the future, we pass on information that others have communicated to us, and we tell lies and half-truths to elicit cooperation, to be accepted as part of a group, to win approval from others, or to evade censure. Even when we are in fact passing on concrete information, we don't necessarily deliver it in clear, concise, and precise wording. We may use words that have multiple meanings or formulate our sentences so that they are ambiguous. There are a variety of ways in which the informational content ("signal") can be uncertain which we will examine in this subsection.

8.3.1.1 *Imprecision and Vagueness*

Human communication is often formulated in ways that obscure, however unintentionally, details that may be useful in the information fusion process.

Let's return to the informational assertion of the example used above:

1. *There were some animals in the road.*

Some is an imprecise number. The reader might possibly make some judgments on the range of numbers represented by *some*: it definitely means more than one, but the upper bound is unclear. Suppose the speaker would have used *a couple* as a descriptor; in general, this refers to a small number, which, while still greater than one, is most likely a very small number, say, two or three. Use of *a bunch* indicates more than *a couple*, whereas *many* would have been preferred if there were a large quantity. *Several* is usually understood to be more than *a couple* – perhaps five or six – but generally would not be considered to be *many*. However, *some* simply implies *multiple* without any further hint as to how many, so we can only guess.

Additionally, it is not clear what sort of animals these are: chickens? dogs? cows? camels? Quite possibly it was a mixture of different types (a sheepdog and a dozen sheep, for example). One type of animal would most likely not be intended here: a human. In that case, *some people* would have been used in place of *some animals*. However, the statement does not necessarily exclude the presence of a human: for example, when a herd of sheep are in the road, the shepherd is often somewhere in the vicinity as well, but the animals, not the shepherd, would likely be considered as some sort of anomaly worth mentioning. Similarly, a high level of precise detail may also be inaccurate: even if we were told that there were six brown Jersey cows in the road, it may well be that the observer neglected to let us know that there were two border collie dogs as well, or overlooked that one brown cow was, in fact, a Hereford and not a Jersey.

Correlating information about a shepherd and his dogs moving his cattle to new pasture with *some animals in the road* requires understanding of a variety of things, including that dogs and cattle are animals, that *some* means *multiple*.

Complicating things further, many vague or imprecise formulations may be context or domain dependent. For example, *large* is an adjective related to size of an object, but its exact (quantifiable) meaning is extremely domain dependent. There are many orders of magnitude difference in the numerical values indicated by *large* between a *large city*, a *large ship*, a *large dog*, and a *large molecule*.

Furthermore, even within the same domain, there may be variations due to other factors such as context information. For example, the phrase *a lot of people* will generate a different numerical range depending on expectation or physical factors such as facility size. If a smallish meeting room is filled to standing room only, it will be reported that the 50 persons attending the event were *a lot of people*. However, those same 50 persons would not be classified as *a lot of people* if they are the only occupants of a 400-seat auditorium. Likewise, 50 people in a 30,000-seat sport stadium would generate a comment more like *nobody was there*. Therefore, the decision about the numerical range represented relies on what we know about the location. Gross et al. [6] have discussed such problems and their resolution at some length.

8.3.1.2 Ambiguity and Polysemy

Statements may be ambiguous, that is, they may be open to more than one interpretation or have more than one possible meaning:

2. *Students hate annoying professors.*
3. *Sally gave Mary her book.*

In (2), it is unclear whether the students strongly dislike professors who irritate them or whether students try to avoid making their professors angry, perhaps in the hope of receiving a better grade in the course. In (3) we do not know whether Sally gave Mary her own (Sally's) book or whether Sally was returning Mary's book to her – or even if there may yet be another female involved. For example, it is possible that in a preceding sentence, the writer informs us that Jane is a well-known author and a friend of Sally's, and thus (3) tell us that Sally presented Mary with a copy of Jane's latest hit novel.

Another example of ambiguity is shown in (4):

4. *I saw her duck.*

There are two possible interpretations of (4); the first is that the female person referred to has lowered her head to avoid, say, a low-hanging branch and the second is that she keeps a waterfowl as a pet. This ambiguity stems from what is called *Polysemy*, the fact that the same word (label) can refer to two or more separate concepts. Other examples of polysemy are words such as *bank* which could be a financial institution, the side of a river, or a motion executed by a flying aircraft, to name just a few of the many meanings. Determining the intended meaning can generally be achieved by analysis of the surrounding text.

Regardless of the lack of detail in all of the examples above, there is nothing in any of these sentences to make us believe that the sentences are not true. However, very often sentences contain clues which the writer uses to signal to us there may be reason to doubt the veracity of the *content* contained in the sentence. These we will discuss in the following section.

8.3.1.3 Which Language?

Last, but not least, one of the most obvious problems is quite straightforward, indeed almost trivial: there are a multitude of spoken and written natural languages. According to the Linguistic Society of America, there are nearly 7000 distinct natural languages, of which some 230 are spoken in Europe [7].

A language such as English has a variety of regional variants with noticeable differences: the Irish playwright George Bernard Shaw is credited with commenting that England and America are two countries separated by a language in common. This is not simply a matter of pronunciation or even spelling – it is also a matter of vocabulary. The *biscuit* of a Brit is an American's *cookie* and an American's *biscuit* more akin to an unsweetened British *scone*. Here we have an instance of the same

word being associated with two different concepts based upon the variant of English being used – an example of the polysemy discussed in Sect. 8.3.1.2. There are also some lexical differences: if you ask an American about a *lorry* you will get a blank look – for her, the object referred to is a *truck*, and the word *lorry* does not exist in the American variant. Here we have two separate words for the same object – in essence, synonyms, but only in a cross-variant sense. Therefore, it is essential to know which version of English is being used.

In addition to regional language variants, there may be examples of polysemy which are domain-specific, that is, they may be very unique to certain subgroups of native speakers of that language, but not to all speakers. For example, *POV* within the US military is used as an abbreviation for *privately owned vehicle* (i.e., a soldier's own car), but within the fiction community, writers often use *POV* to mean *point of view*. Therefore, the meaning attached to the acronym will vary according to the context in which it appears. For outsiders, these need to first be deciphered or disambiguated.

Another issue, particularly with an internationally widely used language such as English, has to do with irregularities produced by nonnative speakers which can also cause misunderstanding and communication problems. This is often caused by what is known as “false friends” or words which appear to be identical in both languages but have quite different meanings. For example, native German speakers often misuse the English word *actual* (meaning *real* or *existing*) when they mean *current* (as in *current news*) because the German word *aktuell* has the latter meaning. Another variation is that non-native speakers borrow terms verbatim from another language but assign them a meaning not existing in the original language: an American uses *oldtimer* to refer to an old person, whereas a German using *oldtimer* means a *classic auto*.

Sometimes the issue is that languages do not map concepts one-to-one. In American English there are three words *pumpkin*, *squash* and *gourd* for a related family of fruits which map to only two words *Kürbis* and *Zierkürbis* in German. While *Zierkürbis* maps one-to-one onto *gourd*, *Kürbis* is used for both *pumpkin* and *squash*, leading to confusion for native speakers of English.

Yet another phenomenon is the invention of words which do not exist in the native version of the language. Ask a native speaker of any English variant what a *pullunder* is, and they will be puzzled, whereas a German speaker will describe to you a sleeveless sweater known as a *sweater vest* in the USA and Canada and as a *slipover* in England.

And finally, there is the issue of “code switching” [8] in which multilingual speakers use different languages within a single communication. It is not uncommon that a natural language acquires words or phrases from other languages, which then become standardized vocabulary for the acquiring language. Examples of this in English are *gestalt* and *angst* (from the German *Gestalt* and *Angst*), which are now fully integrated into the lexicon of the acquiring language and can no longer be considered “foreign.”

In contrast, “code-switching” means that the speaker shifts from one language to another within a sentence (i.e., a word or phrase) or between sentences in a single

communication. For example, it is not unusual within expatriate communities for speakers to insert words or phrases of a language from the host country into their mother tongue communications, particularly with other expatriates who likewise understand the foreign words. This can be problematic for automatic text analysis, which relies on vocabulary and grammatical structures of a single language and can cause confusion or uncertainty as to the meaning of the sentence, through uncertainty of individual words or phrases.

Clearly, there are a number of significant challenges to understanding the meaning of the information received from human sources. But regardless of whether we agree on how many *several* is or have an idea of what kind of animals were in the road, our ability to make use of that information as actionable intelligence depends upon how much we trust that information.

8.3.2 *Uncertainty About the Content*

In the preceding section, we used the following sentence as an example of ambiguity:

3. *Sally gave Mary her book.*

While our previous confusion about the ownership of the book continues, we can assume this event (the handing over of a book) actually took place. However, suppose the sentence read as follows:

5. It is possible that Sally gave Mary her book.

Now we are no longer certain as to whether indeed the event of Sally giving Mary a book occurred. There are multiple ways to view this. Perhaps there was no exchange of a book. Perhaps Mary did receive a book, but it was Georgina who gave it to her. Perhaps the players stayed the same, but it was Mary who gave Sally the book and not the other way around. The presence of “it is possible” changes the credibility of the event significantly. Natural languages are filled with a variety of different mechanisms which inject some uncertainty into the soft data they convey; analysis of these mechanisms will support fusion of soft data in that we may better assess the quality of the data which we are using.

8.3.3 “*Words of Estimative Probability*”

In his 1964 article, Sherman Kent of the United States Central Intelligence Agency relates the following anecdote about a conversation concerning an intelligence report on the possibility of a Soviet invasion of Yugoslavia:

A few days after the estimate [“NIE 29-51, “Probability of an Invasion of Yugoslavia in 1951”] appeared, I was in informal conversation with the Policy Planning Staff’s chairman. We spoke of Yugoslavia and the estimate. Suddenly he said, “By the way, what did you people mean by the expression ‘serious possibility’? What kind of odds did you have in mind?” I told him that my personal estimate was on the dark side, namely, that the odds were around 65 to 35 in favor of an attack. He was somewhat jolted by this; he and his colleagues had read “serious possibility” to mean odds very considerably lower. Understandably troubled by this want of communication, I began asking my own colleagues on the Board of National Estimates what odds they had had in mind when they agreed to that wording. It was another jolt to find that each Board member had had somewhat different odds in mind and the low man was thinking of about 20 to 80, the high of 80 to 20. The rest ranged in between. [9]

What makes this anecdote of particular interest is that the various individuals with whom Kent spoke were all intelligence analysts, that is, people who were working in the same domain (intelligence), who most likely had similar educational backgrounds and similar training for their jobs. In spite of our expectations that working in the same domain with a similar training background might result in some consistent understanding of words and phrases used within that working environment, this anecdote shows us that the understanding of such terms can be quite diverse.

Intrigued by Kent’s observations, another CIA analyst Richards Heuer [10] ran an informal study, asking a number of his CIA colleagues to assign a single probability to about 25 common uncertainty expressions used by the analysts. While the probabilities assigned by the analysts to some of the terms were clustered very closely (*betterthaneven, abouteven, highlyunlikely*), there were several which varied quite dramatically: *highlylikely* ranged more than 40% points, as did *improbable, probablynot* and *chancesareslight*, while the range for *probable* was from 25% to just over 90%.

A couple of decades later, Rieber [11] requested analysts in training at the CIA’s Kent School (named after Sherman Kent) to assign ranges of percentages rather than single values to a smaller number of hedges. Similar to Heuer’s informal study, the ranges of percentages vary from quite narrow to relatively large, but the ranges are not necessarily identical to those in the first chart, even for identical hedges. One can almost assume that giving the task of assigning probabilities for hedges to any random group of English-speakers will result in somewhat different numerical ranges.

In the half century since Kent’s initial work, the US intelligence community has continued to struggle to standardize the terminology which they used to assess situations, in order to reach a common understanding of the meaning of those terms. The current iteration at the time of this writing, the intelligence community has settled on a standard spectrum of words of estimative probability: *remote, veryunlikely, unlikely, evenchance, probably/likely, verylikely, and almostcertainly*.

The discussion above involves examples of uncertainty which are straightforward and immediately obvious to the reader. However, there are other, less obvious forms of uncertainty which are less obvious, often overlooked or ignored; we will examine these in the following sections.

8.3.3.1 Hedges and Evidential Markers

In general, when asked to consider markers of uncertainty in natural language utterances, the first group of words that comes to mind are the expressions (mostly modal adverbs) such as we have seen in the preceding subsection: *possibly*, *probably*, *likely*, etc. The next categories are often modal verbs, *might*, *could*, *may*, etc., followed by nouns such as *likelihood*, *possibility*, *probability*, and so on. Lexical verbs such as *suggest*, *assume*, *seem*, *guess*, etc. likewise convey uncertainty, as do adjectives such as *possible*, *probable*, *doubtful*, etc. The elements above are manifestations of uncertainty and are generally included in a group of lexical markers called *hedges*.

The term “hedge” is attributed to Lakoff [12] to mean any lexical or grammatical form which indicates “fuzziness” in natural language. Inspired by the mathematical theories of Zadeh, Lakoff defines a broad range of lexical and grammatical elements in natural languages which indicate any weakening of the formulation of propositions, which express vagueness or imprecision:

For me, some of the most interesting questions are raised by the study of words whose meaning implicitly involves fuzziness – words whose job is to make things fuzzier or less fuzzy. I will refer to such words as ‘hedges’. [12]

Since Lakoff’s first article, the definition of hedging which he proposed has shifted to focus more narrowly on expressions of uncertainty or commitment on the part of the speaker. Some researchers consider modals verbs (*could*, *should*, *might*, etc.) to be hedges, while others classify them differently (cf. Rein [12] for a more thorough discussion of this topic).

But as we have seen at the beginning of this section, hedges are not the only elements which signal uncertainty in text information. There are markers that indicate where the information contained in the sentence comes from when it is not a direct observation. These are called “evidential markers” which can be divided into two broad categories, *hearsay* and what Bednarek [4] refers to as *mindsay*.

Hearsay, that is, information which the writer has acquired from another source (not himself) is uncertain by nature in that we can never be certain that the writer has correctly and fully understood or recorded what the original source said (or indeed if there is only a single hearsay source rather than a chain) or that the unrelated context would cause us to view the information differently. As a result, we cannot be certain that the information that has been passed on is reliable.

Mindsay is information which comes not from a secondary or tertiary (external) source but from the primary source and which is based upon belief, speculation, and assumption rather than direct observation, that is, it is a product of some process in the primary source’s mind.

Take, for example, the following sentences:

6. *John is a terrorist.*
7. *The CIA have concluded that John is a terrorist.*
8. *I believe that John is a terrorist.*
9. *Mary thinks that John is a terrorist.*

In each of these sentences, the relationship pattern (“content”) of the sentence might produce the relation *John IS-A terrorist*. In (6) there are no lexical clues to indicate where the information came from, nor how credible the speaker considers the information to be. Thus, the basis of our decision as to whether John is, in fact, a terrorist must come from somewhere else, for example, previous knowledge of John’s activities or from our belief in the speaker’s truthfulness.

However, there are lexical clues contained in three of these four sentences which give us reasons to doubt on some level whether John is a terrorist. In (7) and (9), there are clear indicators of third-party information, i.e., *hearsay*, which may or may not have been repeated accurately by the writer.

Sentences (8) and (9) indicate belief, i.e., *mindsay*, rather than knowledge on the part of the sources. In (8), it is clear this is a first-person reporting of the speaker’s belief. Sentence (9) is particularly interesting because it is, in fact, ambiguous. In one interpretation, one could say that it contains both *hearsay* and *mindsay*: the writer informs us about something another person (Mary) has told him (*hearsay*) about her thoughts (*mindsay*) regarding John. A second interpretation could be that the writer expresses his own belief (*mindsay*) about what Mary thinks (also *mindsay*). Regardless of the interpretation, the strength of the assertion of John being a terrorist is weakened.

A single sentence may contain multiple clues as the veracity of the main proposition of the statement. For example, consider the variations on sentence (7):

10. *The CIA have concluded that John is probably a terrorist.*
11. *The CIA have concluded that John is most probably a terrorist.*

In (10) the addition of the adverb *probably* to (7) weakens the assertion of John being a terrorist, whereas in (11) adding *most* before *probably* strengthens the assertion as opposed to (10), but it still remains weaker than in (7) which contains no hedge. If requested, an English speaker would be able to identify and rank assertions from strongest to weakest according to the lexical clues the writer has left in the sentence.

Other factors that may be considered in the assessment of the strength or weakness of an uncertain proposition include variation such as whether in *hearsay* the original source is named (*the CIA*) as opposed to an unnamed source (*rumor has it*) or general knowledge (*it is widely accepted*). The relative strength of *mindsay* may also be determined via the verb used, e.g., *inferred* is stronger than *guessed*.

Most, if not all, decision-making models using natural language information will use some sort of mathematical weighting system based upon the perceived certainty or doubt about the veracity of the data which populates the model. Frajzyngier [13] comments, “the different manners of acquiring knowledge correspond to different degrees of certainty about the truth of the proposition.” Models designed for device-based information such as sensors, cameras, radar, etc. may use factors based upon testing, calibration, and the influence of environmental factors such as light, heat, or humidity to adjust reliability of readings or to fine-tune results. The human-generated information, in contrast, which comes in as text or speech is often assessed by a human, who uses her understanding of various factors including the background

knowledge domain of the information, heuristic or scientific models, or even just a “gut feeling” to evaluate the information and assign it some sort of credibility weight. In lieu of other information, lexical markers may also play a role in the assessment; the analyst may well assign a lower “truth value” to information tagged by the informant through hedges to be *doubtful* than to information considered as *highly likely*.

Whereas many hedges such as the words of estimative probability discussed in the preceding subsection tend to be relatively easy for humans to assign a numerical weight (for an overview cf. Rein, [12]), there is less research to be found on numerical weights for hearsay and mindsay markers. One could easily argue that weighting of the information from different types of sources in such a hierarchy is implicit. For example, direct perception (e.g., *I saw*) is often considered more reliable than hearsay (*he told me*); several authors including Goujon [14], Marin-Arrese [15], and Liddy et al.[16] have looked at such relative values..

8.3.3.2 Passive Voice, Depersonalization, Time, Etc.

Hedges and evidential markers are quite obvious indicators of uncertainty, even to nonlinguists. However, there are some more subtle ways in which uncertainty may appear.

In his discussion on hedges, Hyland [17] includes several other phenomena such as passive voice, conditionals (if-clauses), question forms, impersonal phrasing and time reference. Particularly in scientific writing, passive voice and impersonal phrasing are widely, almost universally, used, conveying an undertone of “but I might be wrong or have overlooked something.” With regard to impersonal phrasing, Hyland writes:

...the writer inevitably uses a wide range of depersonalized forms which shift responsibility for the validity of what is asserted from the writer to those whose views are being reported. Verb forms such as *argue*, *claim*, *contend*, *estimate*, *maintain* and *suggest* occurring with third person subjects are typical examples of forms functioning in the way, as are adverbials like *allegedly*, *reportedly*, *supposedly* and *presumably*.

Passive voice and impersonal phrasing, however, can also be used to express politeness, rather than uncertainty, which can only be determined by knowing some information about the context of the statement. Likewise, passive voice and impersonal phrasing can also sometimes be used in instances of differences in social ranking or power, in order not to offend. In such cases, these forms are not intended to create doubt about the veracity of the proposition, but to soften the impact of a message on the intended audience.

While time might not immediately spring to mind when considering expressions of uncertainty, it nevertheless plays a significant role and should therefore be discussed.

Any sentence which is formulated in the future tense is inherently uncertain, simply because the event or state which is described has not happened yet:

12. *Mary will be at the meeting next week.*

That is, of course, unless Mary decides not to go for some reason, her plane flight is cancelled due to snowfall, or she gets sick and lands in the hospital, or worse.

That being said, some future things are more certain than others:

13. *The next presidential election in the USA will take place in November 2020.*

Clearly, unless something unbelievably catastrophic happens, there is virtually no chance that the elections in the USA will not take place in the month and year named because of legal requirements in the voting laws, so this may be treated as a fact, rather than a possibility.

In other cases, it is a bit less clear. Take the case of routine, recurring behavior:

14. *Sam always attends the meetings of local political action group every Tuesday at 7 pm*

One can expect to find Sam at this meeting every Tuesday – unless, of course, Tuesday is a holiday and the meeting is cancelled or unless Sam, like Mary in (12), has come up with something to prevent his attendance. In other words, recurring behavior may be a good indicator of future behavior, but not a guarantee.

For intelligence purposes, information based upon future actions often plays a very significant role, particularly in preventative measures, but should nearly always be considered uncertain, until the expected date of that action has passed; at that point the event has either occurred or not occurred, and an update should be made to the knowledge base which is being used.

Now that we have examined a number of ways in which formulations may represent uncertainty about the veracity of the informational content in natural language, in the following section we will examine how we might exploit these to algorithmically generate initial credibility weights.

8.4 Using Lexical Clues for Credibility Weighting

When one admits that nothing is certain one must, I think, also admit that some things are much more nearly certain than others. – Bertrand Russell [18]

As previously discussed, when humans communicate with one another they transmit content information, but this content is often surrounded by additional, noncontent information from the speaker intended to convey the speaker's stance to that information.

Text analytics to extract actionable information from text utilize algorithms to locate and identify certain patterns which may identify objects or individuals, events, relationships, and other useful information. The weakness of these algorithms is that they seldom, if ever, take into account that some of those patterns may be couched in language that indicates those identified events or relationships may be questionable or false. In other words, for all intents and purposes, the information

extracted using text analytics is treated as “fact,” even though there may be clear evidence that they are not.

Thus, in order to be truly actionable information, a parallel process to text analytics should be carried out, in which evidence of uncertainty expressed in lexical forms, when found, is analyzed and utilized to assign an initial credibility rating.

We have discussed at some length those lexical elements the use of which indicates uncertainty about the propositional content delivered by a human source. However, not all lexical elements convey the same level of uncertainty: for example, speakers of English consider *probably* and *unlikely* to be more or less at opposite ends of a probability range (Sect. 8.3.3).

There have been many other studies done by linguists in papers such as Teigen and Brun [19], Brun and Teigen [20], Renooij and Witteman [21], Witteman et al. [22], and Ayyub et al. [23] in which single words or multi-word expressions have been evaluated and given numerical weights (probabilities, odds, etc.) by test subjects. There are variations among the studies in the values assigned to any of the expressions, from which one can draw the following conclusion: there are *no universally accepted* values. For a deeper discussion, cf. [5].

However, when taken as a group, what can be seen is that these elements may be relatively ordered along a scale from stronger to weaker (or higher to lower, or more true to less true, to name just a few possibilities).

For example, in general, English speakers would agree to the following ordering:

$$\textit{Unlikely} < \textit{probable}$$

even if they do not agree on the precise numerical values which they assign to these two words of estimative probability.

Very often these expressions are modified by other words, which can strengthen or soften their original meanings:

$$\textit{Highly unlikely} < \textit{unlikely} < \textit{probable} < \textit{very probable}$$

Negation, of course, has a dramatic effect on the ranking from weaker to stronger; we will discuss this in more depth later in Sect. 8.5.

But, as discussed previously, it is not simply the use of words of estimative probability which are indicative of uncertainty about the validity of information but also lexical markers for hearsay and mindsay.

For example, in general, English speakers would agree to the following ordering:

$$\textit{I saw} > \textit{I infer} > \textit{my neighbor told me}$$

As mentioned earlier, a number of researchers (Goujon [14], Marin-Arrese [15], etc.) have examined the relative strength or weakness of a proposition based upon such markers. DeHaan [23] proposed a cross-linguistic comparison of source evidentiality which reflects the previous ordering:

Sensory > inferential > quotative

While there has been research by linguists on this topic, there has been little attempt to assign any sort of (numerical) uncertainty weighting to these evidential markers. One can, however, argue that there is an implicit weighting based upon this hierarchy. There appears to be consistency in the rankings between different groups of people surveyed on these topics as documented by research, from which we can conclude that there seems to be some sort of universal scalar for the various elements which we may exploit for our purposes.

It should be clear from the discussion above that the assignment of numerical values (probabilities, odds) to the lexical and grammatical elements which are of interest to us is not easy. However, it can be very useful to assign uncertainty weights to propositions based upon these clues, especially when fusing uncertain information to use as the basis for informed decision-making.

There remains one more complication, namely, the fact that humans do not always use these elements in isolation. Rather, it is not uncommon that several different markers appear in a single sentence:

15. *I believe John told me that it is very possible that Mary will arrive on Sunday.*

How certain can we be that Mary will indeed arrive on Sunday, based upon this sentence? While different readers may, if requested, assign different probabilities to her arrival, in general one can say that each lexical marker (mindsight, hearsay, words of estimative probability) in this sentence collectively increases our uncertainty.

Natural languages are very flexible, allowing for an infinite combination of words. Thus, listing all possible combinations of these lexical markers and assigning each combination a value would be an arduous (and probably pointless) task, to say the least.

However, in [5] we have shown that it is possible to exploit certain types of lexical items to evaluate this chaining. For example, intensifiers may be used to weaken (*downtoners*) or to strengthen (*boosters*) the evidential weight of elements such as adverbs or adjectives. That is, use of the downtoner *somewhat* weakens *likely* in *somewhat likely*, and similarly the booster *very* will turn *likely* into the stronger *very likely*. If asked to arrange the resulting terms in order from weakest to strongest, speakers of English will generally arrive at the following relation:

Somewhat likely < likely < very likely

Not unexpectedly, there is the reverse effect when we use *somewhat* and *very* with the modal adverb *unlikely*:

Very unlikely < unlikely < somewhat likely

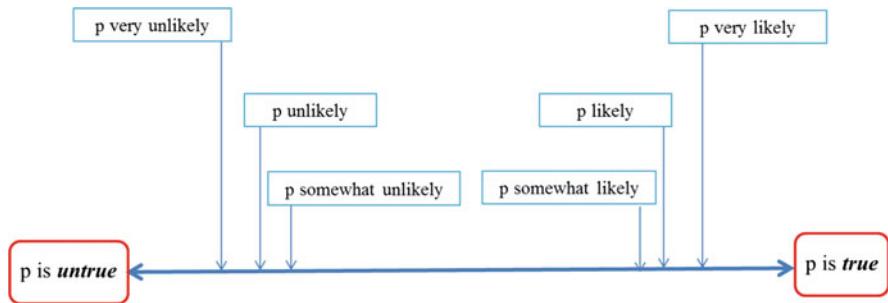


Fig. 8.2 Ranking of *unlikely* and *likely* modified by booster *very* and downtoner *somewhat* for a proposition p

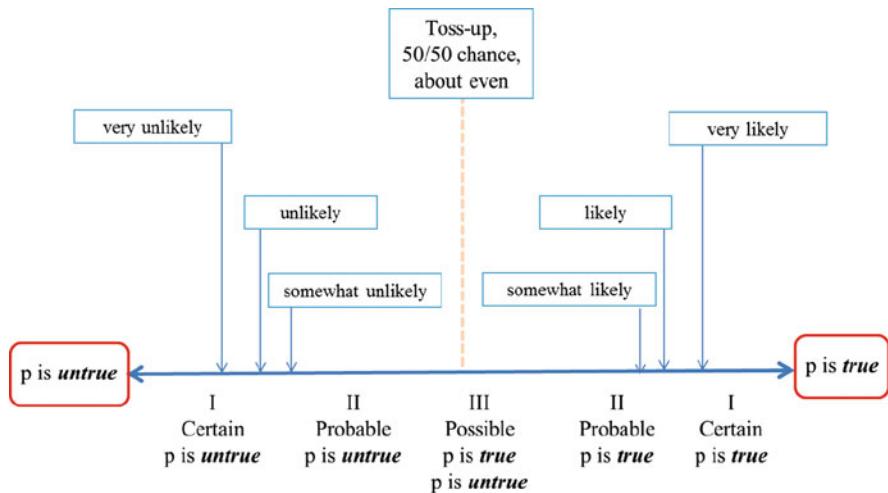


Fig. 8.3 Some modifications to Fig. 8.2 including labels and expressions for complete uncertainty

Figure 8.2 shows a relative placing of a proposition p modified by the above-mentioned hedges along a scale from p is *untrue* to p is *true* for any given proposition p .

But in both cases, we can say that the addition of *very* increased our certainty about a proposition p being either true or false: if something is *very likely* we are pretty certain it is true (or will happen); if something is *very unlikely*, we are quite certain that it is *not* true (or will *not* happen). When we are truly uncertain – the coin is still in the air – we cannot say we are more or less certain to believe p to be true or untrue and are therefore stuck in the middle between p being true and p being false (Fig. 8.3).

While many people view uncertainty on a scale ranging from uncertain to certain (i.e., equivalent to a “0–100 scale,” with 0 representing “uncertain” and 100 “certain”), it turns out that the scale is bipolar: the point of maximum uncertainty

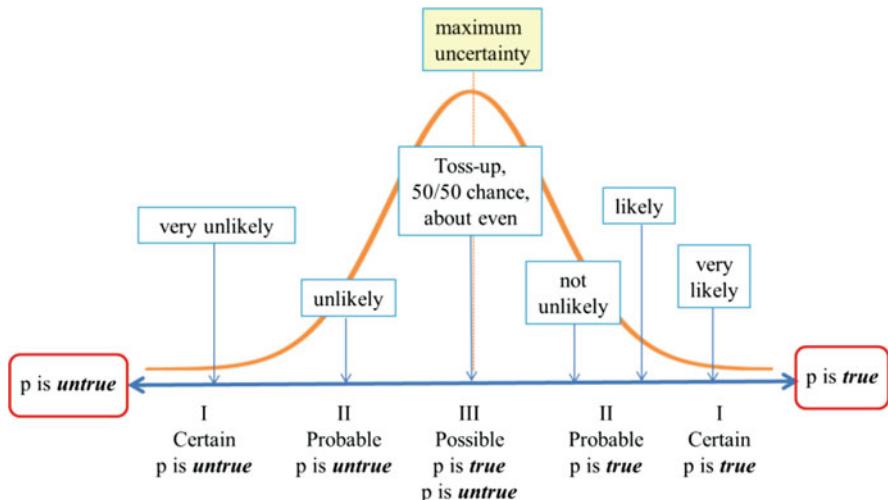


Fig. 8.4 Bipolar scale based on showing point of maximum uncertainty in the center

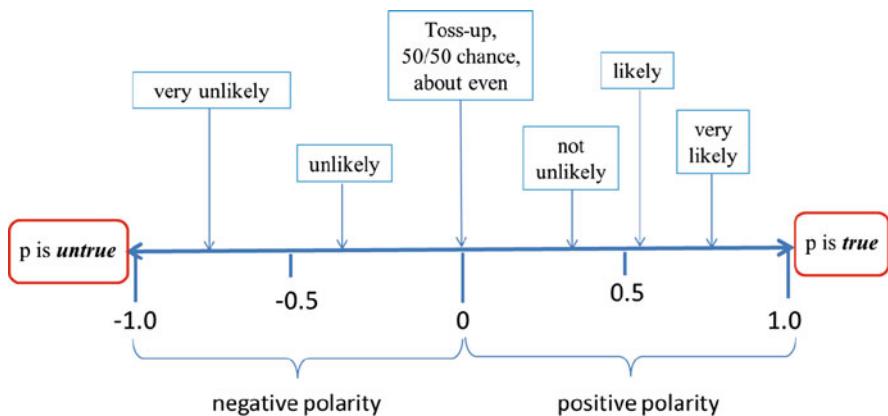


Fig. 8.5 A numerical scale for certainty p is untrue (-1.0) to certainty p is true (1.0), with the point of maximum uncertainty assigned the value 0

lies in the middle of the scale, while maximum certainty lies at both ends of the scale, as illustrated in Fig. 8.4.

Using this bipolarity as a basis, we can define a numerical scale in which the midpoint (the point of maximum uncertainty) is zero, while the end of the scale representing absolute certainty that p is true is assigned the value 1.0 and the end of the scale representing absolute certainty that p is untrue is assigned the value 1.0, as shown in Fig. 8.5.

We can then exploit this scale to help us to automatically determine relative evidentiality weights for chained and modified evidential markers. For this, we define the effect of a modifier on the original weight as follows:

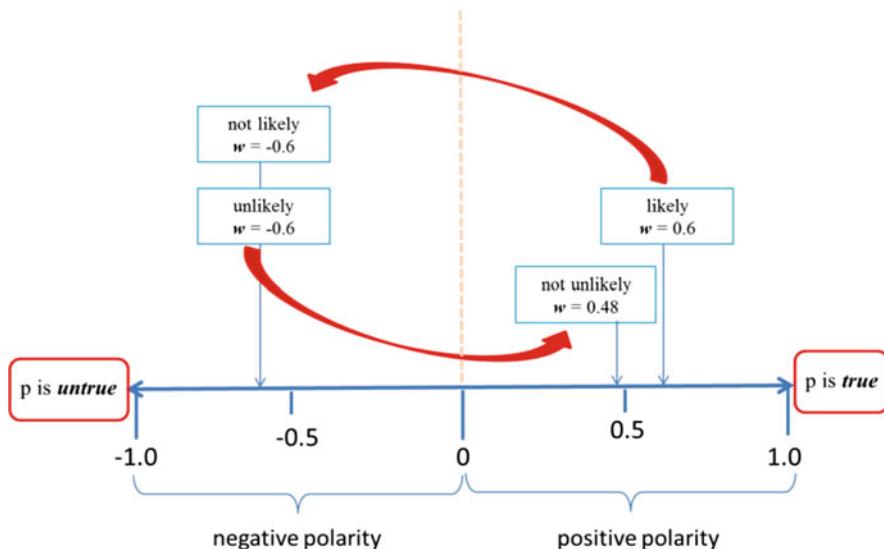


Fig. 8.6 Negation of words of estimative probability

$$w_{\text{modifiedhedge}} = w_{\text{original}} + p_{\text{original}} \times \text{effect}_{\text{modifier}} \times (1 - |w_{\text{original}}|)$$

where p_{original} is the polarity of the original hedge.

For example, suppose we assign the weight $w_{\text{likely}} = 0.6$ to *likely* and the weight $w_{\text{unlikely}} = -0.6$ to *unlikely*. If we have determined for our model that the adverb *very* amplifies (strengthens) that which it modifies by 0.3, then we can use the formula to obtain:

$$w_{\text{very_likely}} = 0.6 \times +(1) \times (0.3) \times (1 - |0 - 6|) = 0.72$$

$$w_{\text{very_unlikely}} = 0.6 \times +(-1) \times (0.3) \times (1 - |-0 - 6|) = -0.72$$

which places, as expected, *very likely* to the right of *likely* on the scale and *very unlikely* to the left of *unlikely*, as they appear in Fig. 8.5.

Negation can be effected by simply modifying the polarity, as shown in Fig. 8.6. Note that negating a negatively poled expression results in something less certain than its positively poled antonym, i.e., whereas *not likely* and *unlikely* are usually considered to be more or less equivalent, *not unlikely* is generally considered to be weaker than *likely*.

Similar to words of estimative probability as discussed above, lexical clues indicating hearsay or mindsay may be assigned values and modified as described above. A sample is shown in Fig. 8.7.

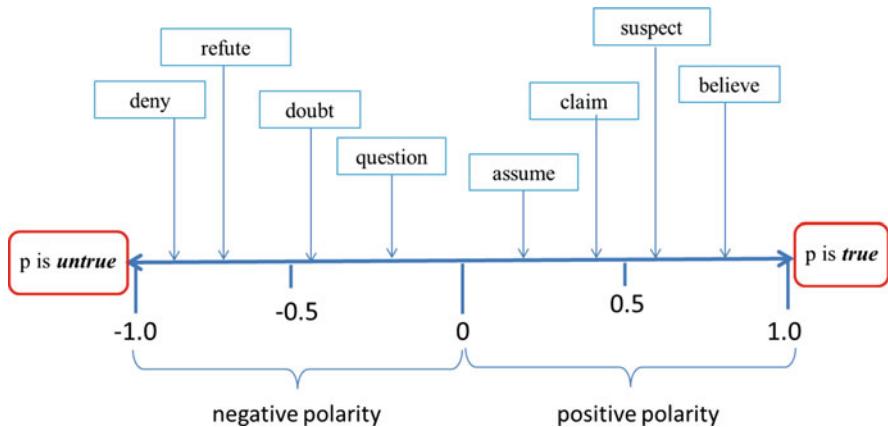


Fig. 8.7 Example of relative weightings of various verbs expressing uncertainty about propositional information

Thus, by evaluating the various multiple lexical expressions surrounding the content, we can come up with a rough estimation of the credibility of the information based upon clues the reporter has left us in the communication.

Once we have the ranking, we can map the results to existing scales such as a fuzzy scale using words of estimative probability or to a numerical scale such as percentages. Two simple examples appear in Fig. 8.8.

It must again be reiterated here that the values assigned to expressions, as well as the values to modifiers, are defined by the user; as mentioned earlier, there are no universal values, but there is a certain consistency in relative understanding of the values (meanings) of various expressions.

For a more detailed, in-depth discussion of the underlying research as well as a more detailed discussion of the algorithms, please refer to [5].

8.5 Summary

In this chapter we have discussed the need for evaluation of uncertainty in natural language information to give decision-makers a clear picture of the quality of that information. We discussed the various forms of uncertainty in natural language, categorized in two ways: uncertainty *within* the information which includes imprecision, vagueness, ambiguity, polysemy, and so on and uncertainty *about* the information, the constructs which appear in text that tell us whether the information is reliable or not, and in which way. We then briefly present the outlines of a concept which allows us to automatically generate evidential weights for information derived through text analytics by examining lexical clues concerning

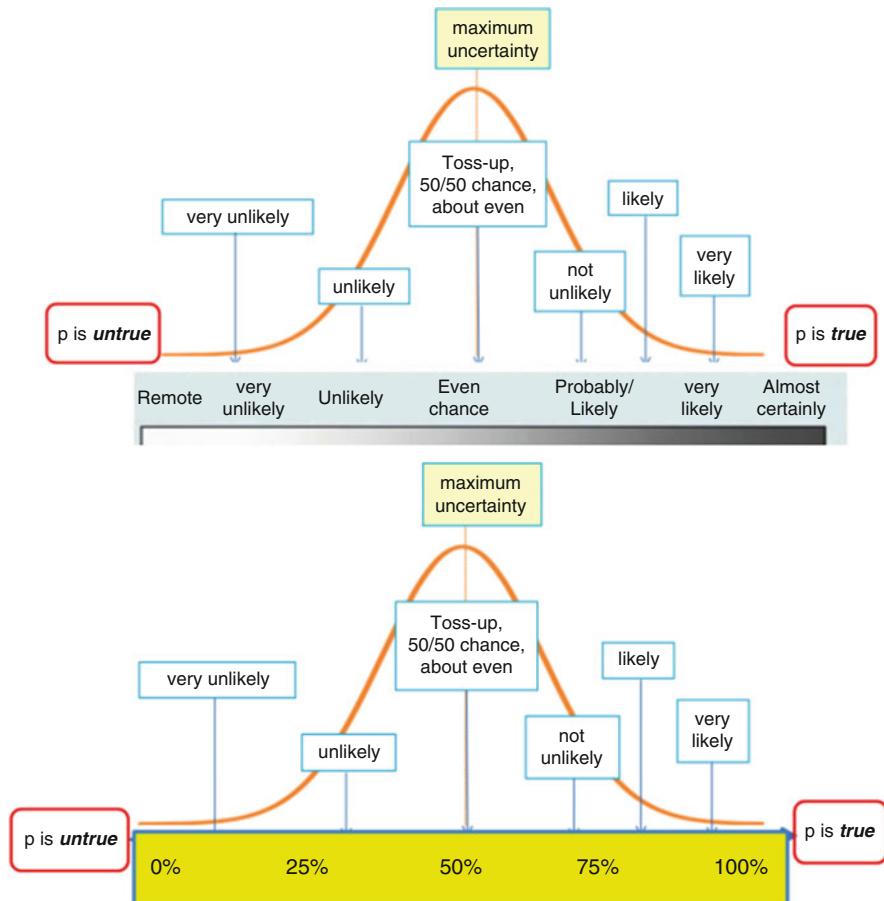


Fig. 8.8 Examples of mappings of relative rankings onto a scale of words of estimative probability (left) and percentages (right)

original source or stance toward the veracity of the information which speakers embed in their communications.

References

1. J.A. Gans Jr, “‘This is 50-50’: Behind Obama’s decision to kill Bin Laden”, *The Atlantic*, Oct 10, 2012,<https://www.theatlantic.com/international/archive/2012/10/this-is-50-50-behind-obamas-decision-to-kill-bin-laden/263449/>
2. R. de Gourmont, *Philosophic Nights in Paris* (J.W. Luce, Boston, 1920), p. 127
3. V. Dragos, K. Rein, “What’s in a message? Exploring dimensions of trust in reported information”, *Proceedings of Fusion 2016*, IEEE, 2016

4. M. Bednarek, *Evaluation in Media Discourse: Analysis of a Newspaper Corpus* (Continuum, London, 2006)
5. G.A. Gross, R. Nagi, D. R. KedarSambhoos S.C. Schlegel G.T. Shapiro, Towards Hard+Soft Data Fusion: Processing Architecture and Implementation for the Joint Fusion and Analysis of Hard and Soft Intelligence Data. Proceedings of Fusion 2012, pp. 955–962
6. <http://www.linguisticsociety.org/content/how-many-languages-are-there-world>
7. D. Claeser, D. Felske, S. Kent, Token level code-switching detection using Wikipedia as a lexical resource, in *Language Technologies for the Challenges of the Digital Age. GSCL 2017. Lecture Notes in Computer Science*, ed. by G. Rehm, T. Declerck, vol. 10713, (Springer, Cham, 2018)
8. K. Sherman, “Words of Estimative Probability”, Studies in Intelligence, Fall 1964, Central Intelligence Agency, <https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/sherman-kent-and-the-board-of-national-estimates-collected-essays/6words.html>, (1964)
9. R.J. Heuer Jr., “Psychology of intelligence analysis”, Center for the Study of Intelligence (1999)
10. S. Rieber, “Communicating Uncertainty in Intelligence Analysis”, http://citation.allacademic.com/meta/p100689_index.htmlf (2006)
11. G. Lakoff, Hedges: A study in meaning criteria and the logic of Fuzzy concepts. J. Philos. Logic **2**, 458–508 (1973). D. Reidel Publishing Co., Dordrecht, Holland
12. K. Rein, I believe it's possible it might be so.... exploiting lexical clues for the automatic generation of evidentiality weights for information extracted from English text. Universitäts- und Landesbibliothek Bonn, (2016). <http://hss.ulb.uni-bonn.de/2016/4471/4471.htm>
13. Z. Frajzyngier, Truth and the indicative sentence. Stud. Lang. **9**(2), 243–254 (1985)
14. B. Goujon, Uncertainty detection for information extraction, International Conference RANLP 2009, Borovets, Bulgaria, (2009), pp. 118–122
15. J.I. Marin-Arrese, “Epistemic Legitimizing Strategies, Commitment and Accountability in Discourse”. *{Discourse Studies}*, vol 13 (Sage Publications, 2011). <https://doi.org/10.1177/1461445611421360c>
16. E.D. Liddy, N. Kando, V.L. Rubin, “Certainty Categorization Model”, *The AAAI Symposium on Exploring Attitude and Affect in Text AAAI-EAAT*, vol 2004 (American Association for Artificial Intelligence, Stanford, 2004)
17. K.H. Hyland, *Hedging in Scientific Research Articles* (John Benjamins, Amsterdam/Philadelphia, 1998)
18. Russell, Bertrand, “Am I an Atheist or an Agnostic?”, Literary Guide Rationalist Rev. 64, 7, July, 1949, pp. 115–116
19. K.H. Teigen, W. Brun, Yes, but it is uncertain: Directions and communicative intention of verbal probabilistic terms. *Acta Psychologica* **88**, 233–258., Elsevier Science B.V. (1995)
20. W. Brun, K.H. Teigen, Verbal probabilities: Ambiguous, context-dependent, or both? *Organ. Behav. Hum. Decis. Process.* **41**(3), 390–404 (1988)
21. S. Renooij, C.L.M. Witteman, Talking probabilities: Communicating probabilistic information with words and numbers. *Int. J. Approximate Reason.* **22**(3), 169–195. Elsevier (1999)
22. C.L.M. Witteman, S. Renooij, P. Koele, *BMC Med. Inform. Decis. Mak.* **7**(13) (2007). BioMed Central Ltd, <http://www.biomedcentral.com/1472-6947/7/13>
23. B.M. Ayyub, G.J. Klir, *Uncertainty Modeling and Analysis in Engineering and the Sciences* (J. Chapman and Hall/CRC, Boca Raton, 2006)

Chapter 9

Fake or Fact? Theoretical and Practical Aspects of Fake News



George Bara, Gerhard Backfried, and Dorothea Thomas-Aniola

Abstract The phenomenon of fake news is nothing new. It has been around as long as people have had a vested interest in manipulating opinions and images, dating back to historical times, for which written accounts exist and probably much beyond. Referring to it as *post-truth* seems futile, as there's probably never been an era of truth when it comes to news. More recently, however, the technical means and the widespread use of social media have propelled the phenomenon onto a new level altogether. Individuals, organizations, and state-actors actively engage in propaganda and the use of fake news to create insecurity, confusion, and doubt and promote their own agenda – frequently of a financial or political nature. We discuss the history of fake news and some reasons as to why people are bound to fall for it. We address signs of fake news and ways to detect it or, to at least become more aware of it and discuss the subject of truthfulness of messages and the perceived information quality of platforms. Some examples from the recent past demonstrate how fake news has played a role in a variety of scenarios. We conclude with remarks on how to tackle the phenomenon – to eradicate it will not be possible in the near term. But employing a few sound strategies might mitigate some of the harmful effects.

Keywords Fake news · Fake news detection · Propaganda · Misinformation · Clickbait · Conspiracy

G. Bara
Zetta-Cloud, Cluj-Napoca, Romania

G. Backfried · D. Thomas-Aniola (✉)
SAIL LABS Technology, Vienna, Austria
e-mail: dorothea.aniola@gmx.at

9.1 A Brief History of Fake News

To be clear, *fake news* is nothing new. Intentionally circulating rumors and false information have been around for as long as humans have realized its influential power. What makes fake news seem *new* today, is the medium, speed and precision with which it can be distributed on a much larger scale by micro-targeting particular users. The term *fake news* has gained significant prominence during and after the 2016 US general election. This was partially due to *BuzzFeed News*'s story on teenagers in the Balkans who had intentionally spread false information in support of Donald Trump [13]. Their investigation revealed that at least 140 US politics websites were created and run by young Macedonians who had used Facebook adverts to spread fabricated news articles in favor of Trump. Their ultimate goal was not to help him win the election but to make money by generating lots of traffic to their websites [55]. According to several teenagers interviewed by BuzzFeed, Facebook was deemed as the best platform for spreading fake stories [55]¹. As false information and rumors have been around for centuries, the medium with which they can be disseminated today, namely, via social media, has made fake news look like a novel phenomenon, but, in fact, fake news is anything but new.

It is challenging to precisely pinpoint the origins of fake news. Following the definition of Allcott and Gentzkow [5] as “news articles that are intentionally² and verifiably false, and could mislead readers,” it could be argued that fake news has its origin somewhere after the invention of the printing press in 1439, which allowed news to be printed and thus shared on a much larger scale as in contrast to the preprinting press era, in which news spread from person to person. Soll [58] supports this argument by claiming that “Fake news took off at the same time that news began to circulate widely.” As there were plenty of news sources available, including official publications by political and religious authorities, a concept of journalistic ethics³ or objectivity was lacking, he argues [58]. This, in turn, enabled fake news to be distributed for various purposes, such as political or economic gains. One of the earliest examples of using fake news for profit was the often-cited example of the New York Sun in 1835, which ran stories about a British astronomer who had discovered, among other things, “blue-skinned goats” and “giant bat-like people” by using a newly invented telescope to view the moon. This story led to a massive surge in sales, and for a short time, *The New York Sun* became the world’s bestselling daily paper thanks to intentionally spreading fake news [60].

In a similar vein, Leetaru argues that fake news is as old as the news itself, but goes further and states that it became an inherent part of wartime propaganda,

¹Combining in a single incident two of the most prominent motivations for fake news: politics and money.

²One may also distinguish between intentional disinformation or unintentional misinformation, see below.

³See Lara Setrakian’s TED talk on the need of a *Hippocratic Oath* for journalists. https://www.ted.com/talks/lara_setrakian_3_ways_to_fix_a_broken_news_industry

especially during World War I and II [38, 39]. Leetaru bases his claims on an experiment with Google Books NGrams Viewer – an online search engine that plots the frequencies of words and phrases using a yearly count of n-grams found in printed books between 1800 and 2000. Using this search engine, Leetaru was specifically interested in how often the term *fake news* had appeared in books in the last two centuries [32]. He found the term was initially used at the start of World War I and reached its peak just before World War II. He explains the peak by noting that it is likely reflecting the rise of research into propaganda and its potential impact on societies [38, 39]. All in all, he argues, the phrase *fake news* originated somewhere around the beginnings of WWI and WWII but came to rapid prominence after the 2016 US general elections.

Others, such as Burkhardt, point to the preprinting press era as a new and potential origin of fake news⁴. However, Burkhardt carefully caveats there are no means to verify these claims. She continues by explaining that information was inscribed on materials, such as stone, clay and papyrus and was usually limited to the group leaders, who ultimately controlled this information. This control, in turn, gave them power over others as group members heavily relied on knowledge or information obtained from the leader [39]. So, intentionally spreading fake news to smear other group leaders, for example, could have been possible and would have probably worked, because “without the means to verify the claims, it’s hard to know whether the information was true or fake news” as Burkhardt argues [12].

Numerous examples of fake news can be found throughout the post-printing press era. American writer Edgar Allan Poe (1809–1849), for example, could be considered a *fake news author* as he is credited with writing at least six stories that all turned out to be hoaxes, most famously, the story of a balloonist who had crossed the Atlantic in 3 days [9]. The emergence of fake news is not limited to a specific century or period of time. In fact, numerous examples are documented, ranging from the mid-fifteenth century up to the present day.

As the means to communicate have developed, so has the way in which fake news has been circulated too. In the mass media era [12], for example, when radio broadcasting became widely available, fake news had also been around. One famous example is *The War of the Worlds* broadcast in 1938, which was a narrated adaptation of H. G. Wells’ novel *The War of the Worlds*. Even though the presenter emphasized at the beginning of the show that the following broadcast was a narrated novel, a mass panic followed as many people confused the novel with real news [68].

In sum, it can be stated that fake news has always been around for as long as humans have existed and realized its power of influencing others for economic, social, and political gains. It is still unclear when it emerged for the very first time, but many historians, journalists, and scholars point to the invention of the printing press, which gave rise to news media outlets and eventually to fake news. Numerous examples can be found throughout history, leading up to the 2016 US

⁴An early famous example is the battle of Kadesh, fought in 1274 BC and its presentation as a victory by Egyptian pharaoh Ramses II.

general elections and due to the pervasive nature of fake news, it is highly likely that rumors and false information will remain with us for the foreseeable future.

9.2 Why People Fall for Fake News

One reason why people fall for fake news is the fact that they perform poorly at fact-checking. Research from cognitive psychology has demonstrated that people are naturally bad fact-checkers and comparing known to unknown issues poses a significant challenge [29]. The so-called Moses Illusion, also known as the semantic illusion, serves as an illustrative example. This phenomenon was first identified in 1981 in a study that examined how meanings of individual words are combined to form a more global description of meaning [29]. The Moses Illusion occurs when people answer “two” in response to the question “How many animals of each kind did Moses take on the ark?” even though they know, it was Noah and not Moses who took animals onto the ark. The study found 81% of the participants ($N = 27$) did not notice the error in the question although they knew the correct answer [27]. Psychologists call this phenomenon *knowledge neglect* [40, 41].

So, why are humans so bad at noticing errors? There are several explanations for this, most of which stem from psychology. According to [29], two peculiar habits of psychology make humans miss these errors. First, their general bias to initially process and label new information as true, also referred to as *confirmation bias* [31]. Second, as long as new information is perceived as *close to the truth*, it will automatically be accepted. The underlying idea refers to the *good-enough model*, which presumes that to maintain conversations, “humans accept information that is good enough and just move on” [31]. In other words, small errors in sentences will be accepted in order to keep the conversation flowing.

Another reason why people believe fake news stories is a particular cognitive shortcut that occurs when users have to decide whether or not to share a story on their social media feed [14]. In general, the human brain uses cognitive shortcuts or simple heuristics⁵ to avoid information overload when making a decision. These shortcuts or heuristics are intuitive, short, and simple rules that subconsciously issue instructions as to what to do. In this respect, a research study examined the hypothesis that negative news stories are more likely to be retweeted on Twitter as in contrast to positive, nonnews tweets. The findings suggest that indeed, negative sentiment enhances *virality* in the news segment (but not in the nonnews segment) [34]. To put it simply, people tend to retweet negative news headlines without even challenging their credibility [14].

A further factor contributing to why people fall for fake news are so-called *echo chambers* [14]. An echo chamber in the context of media is a filter bubble around a user, in which a person is exposed to content that amplifies or reinforces existing

⁵For a good overview of heuristics, see Gigerenzer et al. [30].

beliefs. As users are free to choose whom they want to follow, or be friends with, or which web-sites and sources to read, it is up to the individual to decide what content they wish to see on their social media feed. The danger here is that users who only view content that reinforces existing political or social views are more likely to believe fake news, because people generally tend to favor information coming from within their social circles [20]. These social circles, however, can act as an echo chamber [53]. Therefore, users will most likely not challenge the authenticity of the news article or the source as they already trust their social circle (following a *source-based trust model* in that sense) [20].

In a longitudinal research study, researchers subjected the Facebook activity of 376 million English-speaking users to critical scrutiny by examining how Facebook users interacted with English-speaking news sources [54]. Their findings demonstrate that the higher the activity of a user, the more the user tends to focus on a small number of news outlets, which in turn, leads to distinct online community structures and strong user polarization. In other words, social media creates an echo chamber that provides the ideal environment for fake news to spread. In a similar vein, the authors conclude that the polarization of users online is probably the main problem behind misinformation [54].

Another reason why people fall for fake news, especially when it comes to images, is a humans' inability to detect photo forgeries. Considering that 3.2 billion images are shared each day on average [11] and highly sophisticated photo-editing tools are widely available, photo forgeries constitute a significant challenge. Images can sometimes form a critical part of a fake news story, making research in this area paramount. However, according to Nightingale et al. [45], there is a lack of research which specifically examines humans' performance in detecting photo manipulations. For this reason, Nightingale and colleagues conducted one of the first experimental studies of its kind, in which participants were presented both original and manipulated photos of real-world scenes to classify these accordingly and point out the manipulations. The results indicated that humans have an "extremely limited ability to detect and locate manipulations of real-world scenes" [45].

Communication and the sharing of information is known to be a key factor for the use of social media. Whereas the sharing of correct information may foster a better-informed society, the sharing of inaccurate and misleading information may have negative consequences [37]. Information quality – regarding the accuracy, truthfulness, or veracity of information on a particular medium or *perceived* information quality, an individual's *perception* of these factors – play a fundamental role in the sharing behavior of users and thus have a strong impact on the dissemination of fake news. High perceived information quality may encourage the sharing of messages whereas low information quality may lead to reduced willingness to share and increased consciousness about one's online reputation.

In sum, human psychology, communication patterns, echo chambers on social media, the inability to detect photo or video manipulations, and the perceived information quality of platforms are all critical factors that enable fake news to flourish. The good news is, however, there are techniques to spot fake news! But this will ultimately be up to the individual to decide whether or not he/she acknowledges

the existence of false information and rumors on social media. Most importantly, individuals will need to be ready to learn basic techniques to assess and verify the credibility of the news article they read. Efforts to increase digital literacy and use of social media in education and training, on all levels from governments to companies, platform providers, news-outlets, and individuals, should complement such efforts and ultimately lead to a more conscious manner of news consumption and distribution behavior.

9.3 Fake News: Practical Aspects

If you would ask someone what “fake news” was 10 years ago, they would probably indicate towards humorous satire websites such as *The Onion*⁶ or *The Daily Mash*⁷. Today, the definition of fake news is more nuanced, more diffuse, and a lot less humor-oriented and includes a wide spectrum of types of publishers and content distribution models.

The simplest categorization of fake news would be split into “mis-” and “dis”-information. Misinformation, which is not intentional and disinformation, which is intentional. Both are used to manipulate and influence opinions and can have a serious impact on topics and segments of society. Oftentimes, accurate and inaccurate facts and information are intertwined and twisted to provide slanted viewpoints and half-truths mixing fact and fiction [28]. The wide adoption of the Internet and of social media platforms, in particular, have made it very easy to create, distribute, and consume content in a rapid and massively parallel manner.

Misinformation, whether generated by unreliable sources of information or low-quality journalism, can have a significant impact on media consumers. Two relevant examples detailed further in this chapter, such as wrongly accusing BMW of rigging car emissions in the manner of Volkswagen, or ABC News causing the Dow Jones to fall by publishing a report wrongly accusing candidate Trump of colluding with Russia, demonstrate just how important it is to maintain a high-quality information verification process within any editorial newsroom.

Disinformation can take many forms and covers a wide spectrum of fake news: from the seemingly innocent satire websites to clickbait, source hacking, sponsored content, conspiracy theories, and even state-sponsored propaganda. What disinformation aims for is a significant impact in the information space, which can be obtained almost exclusively in the online battlegrounds where audiences are easy to reach and even act as content distributors themselves, while interacting with the content.

⁶www.theonion.com

⁷www.thedailymash.co.uk

9.4 Tackling Fake News

Analyzing the most recent prominent fake news scandals, from the Indonesian presidential elections smearing campaigns in 2015, the 2016 US Presidential elections, and the ongoing conspiracy theories that once in a while get published by mainstream media, recurring patterns into the techniques used to create fake news with most impact can be identified. These patterns have since been structured into identified signals presented in digital literacy programs and manuals (such as the ones put together by the *News Literacy Project*) and also implemented into algorithms used in academic or commercial software applications designed to spot online fake news, such as *Hoaxy*⁸ developed by the *Indiana University Open Networks Institute* or *TrustServista*⁹ developed by *Zetta Cloud*.

The fake news scandal that erupted after the 2016 US presidential elections has put a spotlight not only on social media platforms, such as Facebook or Google, accused of not taking action against the spread of disinformation but also on researchers and journalists, who were expected to bring expert-solutions to the problem. The main directions of addressing online fake news, mainly since late 2016, can be categorized as follows:

1. **Professional Fact-Checking**, when information is verified by media professionals. This approach provides an accurate and standardized method of information verification, which eliminates any bias and can lead to identical results even with different impartial verifiers. Notable examples of such professional fact-checking groups are [Snopes.com](#), *First Draft News* or [FactCheck.org](#). The disadvantage of this approach is that it is very time-consuming, it does not scale (considering the limitless size of the Internet), and most of the time it is a sluggish “post-mortem” analysis performed after the viral fake news content was distributed and has reached its goals.
2. **Crowdsourced Fact-Checking**, when verification is performed by non-professionals on virtual online platforms, using commonly agreed principles and standards of work. The idea of this approach is to be independent, to leverage “crowd wisdom” and to create a scalable model. In reality, it proved less successful than the professional fact-checking approach, with only a few such initiatives becoming known to the public (*Reddit*, *WikiTribune*), and results being challenged for being potentially partisan or being derailed by strong online trolling groups such as user boards [4Chan.org](#) or [8Chan](#)¹⁰.
3. **Automated Algorithmic Verification**, when content verification is performed in an unassisted, automated manner by a software program. This approach is being used, in combination with human review, by Facebook and Google, and also AI startups such as *Zetta Cloud*, *Factmata* or *Unpartial*. The idea behind this

⁸<https://hoaxy.iuni.iu.edu>

⁹www.trustservista.com

¹⁰8ch.net

approach is to successfully filter out most of the untrustworthy online content that matches certain statistical patterns, with potential to work in real-time, at scale, and a “good enough” fake news detection quality, similar to those of email spam filters.

The fake news types with most of the impact in the online space – clickbait, conspiracy, and propaganda – have common “red flags” that can be used to identify them as untrusted content, even if the content is in video, image, or text format. These “red flags” make up identifiable signals that a human content consumer with moderate digital literacy or critical reading skills could use; also, signals that could be implemented as software algorithms for the automatic detection on online disinformation, in real-time.

9.5 Algorithmic Fake News Detection

One of the areas of interest with regard to curbing the fake news phenomenon is the research of automatic fake news detection using Artificial Intelligence algorithms. Since the 2016 US presidential elections fake news scandal, more effort has been put into researching and developing tools that can detect fake news stories in the online space.

One such tool was launched in early 2017 by *Zetta Cloud*¹¹, a Romanian AI startup that received a grant from Google to develop a software prototype that can automatically detect fake news. The tool called *TrustServista*¹² was used for publishing a “News Verification Report” [63] detailing an approach to standardizing “red flags” for automatically detecting fake news with no human intervention. The report analyzed 17,000 online news articles (collected from news agencies, newspapers, blogs, magazines, and organization websites) in order to find patterns into how the quality, source referencing, and authorship of online news impact the trustworthiness of the produced content, outlining the current Text Analytics and Semantic Reasoning capabilities for the algorithmic detection of such content.

The final metric according to which the content was analyzed is called a “trustworthiness score” and indicates if a story is trustworthy or not, rather than classifying it as true or false or performing any checking of the facts found in the article. The trustworthiness score takes into account the following information:

Writeprint – Source and Author identification:

- Is it published by a known established publisher?
- Can the publisher’s credentials (ownership, editorial team) be verified?

¹¹www.zettacloud.com

¹²www.trustservista.com

- Does the article have a named, verifiable author with public social media profile?
- Does the publication have a history of political bias or allegiance?
- Is the web “footprint” verifiable (domain, hosting company, hosting location, advertising code, security certificate).

Writeprint – Content writing style:

- Is the article written in a balanced, non-sentimental way?
- Is the quality of the text consistent with the topic of the article?
- Does it contain actual and useful information?
- Does it mention the information sources it uses (including video/photo)?
- Were the images used before on the web in a different context?
- Does it contain strong emotions towards the subject or use of clickbait techniques?
- Does it use fraudulent images, out-of-context videos, claims that cannot be checked?
- Does it use anonymous or unnamed sources?

The information origin and context:

- Can the information be traced back to the original (Patient Zero) information?
- How trusted is the origin of information and other content sources referenced?
- What are other publishers and articles writing about the same topic?

In fact, these verifications which are performed automatically by software algorithms relying on text analytics, graph analytics, and statistical analysis provide a good baseline for the standardization and automation of fake news detection and analysis. Also, the impact of fake news stories can also be measured using standard metrics:

1. **Facebook or Twitter engagements**, such as likes, shares, comments, retweets. The higher these metrics are, the more people were reached by the content (have read at least the title) and have interacted with it, distributing it further in their networks by linking it, retweeting it, or commenting on it.
2. **Content engagement analytics**, typically available only to the content owners, showing the number of visits on their website or views on social media.
3. **References and citations**, either direct hyperlinks to the article’s URL from other websites or just mentions (“according to...”).

The advancements in Artificial Intelligence algorithms and processing power allows software programs to automatically extract key information from content with no human supervision: language detection, automatic summarization, named entity extractions, sentiment analysis, hyperlinks or references, clickbait detection, content classification, emotion extraction, semantic similarity comparison. All these metrics can be used to understand content in a similar way as a human does, with the goal of assessing its trustworthiness.

9.6 Fake News: “Red Flags”

When creating fake news, both the content and distribution channels are important. The content needs to be appropriate for the end goal: fringe websites will use clickbait techniques and rely on social media, state actors will create high-quality content and try to send the message into the mainstream, conspiracy theorists will use message boards and fora with “source hacking” techniques.

The most typical techniques or “red flags” for creating and distributing disinformation are:

- **Using clickbait writing style**, especially for the article title, with the main purpose of attracting attention and encouraging visitors to click on the hyperlink or distribute the content. Certain punctuation marks, heavy use of capitalized letters, words and expressions that trigger strong emotions, and the lack of actual information in the title (names, locations), the different elements of information presented in the body of the article in comparison with the title, are the markings of a “clickbait” content. The generation of content by every day users of social media platforms (UGC – user generated content), exhibiting large variation in information quality [3], poses additional and increasing challenges.
- **Republishing old stories** in order to enforce the *Anonymous or unverified sources*. As the Internet has democratized content production, information published by sources that have no contact details, no ownership information or anonymous authors, should be regarded as untrustworthy. Even if the publishers are anonymous, trust can be built up in time, if the information published is constantly verifiable and trustworthy.
- **Lack of factual information in the articles**. Opinion pieces, conspiracy theories, or low-quality journalism tend to exhibit a lack of information context – people, organizations, geographical, and time references – elements that typically constitute the factual information backbone of any news story.
- **Relying on unnamed sources**, such as unnamed officials, anonymous tipsters, or political insiders, is a technique used by established media organizations in order to either get ahead of the competition with breaking stories or to create propaganda stories. Although this information sourcing is accepted by most media organizations, it can be considered untrustworthy. These types of stories tend not to get verified afterward.
- **Lack of sources or using unreliable sources**. Since information is often created by news agencies or local newspapers, witnessing events first hand, the vast majority of other publication types typically rely on sources for their stories. The lack of sources referenced in the articles or referencing unreliable source constitutes a “red flag” leading to possible untrustworthiness. The generation viewpoints of certain online “echo chamber”-like groups are a known and potentially successful fake news technique, since the huge amount of news content produced every day ensures that the public cannot remember old stories and be assumed to react to the same story in the same way when confronted with republished news.

- **Source hacking**, a complex fake news creation technique specific to forums and message boards where anonymous users disguised as an insider, firsthand witness, or whistleblower information would publish fake information in the hope that journalists on the look for stories will pick up the information and publish it in the mainstream media. News organizations with established source verification and quality standards tend to keep away from such anonymous online information.
- **Use of manipulated images or wrongly attributed videos**, either as part of written news articles or standalone, which can be difficult to debunk, making it another successful fake news technique. The “red flags” for such content can only be obtained with specialized software and rarely verified using journalistic or critical reading techniques.
- **Different viewpoints** also make up an important verification process for any information. If other publishers are writing about the same topic, if there are different viewpoints on a specific matter, this is a sign that the story is not fake. Even if the stance on the topic differs from publisher to publisher, even if there is no consensus, witnessing multiple actors engaging with a story is a valuable verification starting point, this can be assumed to be the actual case.

The content red flags and engagement models can differ depending on the type of fake news. To summarize the 3 most popular types of written fake news content:

- **Clickbait**, which is usually created by non-established news websites and relies mostly on the writing style to obtain maximum audience reach, typically on social media platforms. This type of disinformation used either for political or commercial gains does not reference sources or does it only partially or in a deceptive way. The content does not contain sufficient factual information, rarely has a named and identifiable author, and the main focus is actually on the title written in such a way that it should motivate the reader to click on it (hence click-bait) or redistribute it without reading the full article. The content can also be overly negative or positive. Rarely cited or referenced outside the content producer’s own network, the aim of this content type is to “viralize” (becoming viral), to obtain a large number of views, likes, shares or retweets on social media platforms.
- **Conspiracy**, websites that publish mostly false information on different topics such as history, politics, religion, or health. Conspiracy websites tend to have a varied degree of popularity and present information that cannot be proved scientifically. The writing style can vary from clickbait-type content to high-quality content, resembling scientific papers. The Wikipedia page on conspiracy theories [69] lists 12 categories of conspiracy theories. Recently, message boards such as 4chan or 8chan have been successful at collaboratively creating and distributing conspiracy theories in the online space, one of the most famous users involved in these activities being *QAnon* [15].
- **Propaganda**, sponsored by various political organizations or even state-actors. Propaganda stories may span a large number of articles and are typically information-rich, displaying a high-quality writing style, and referencing a

known author in order to resemble high-quality journalism and not be labeled as clickbait or fake news. However, even if propaganda articles make use of multiple sources, they tend to mix trusted and verifiable information with untrusted, usually unnamed sources or other websites that are partisan.

The following examples detail the various types of fake news, from misinformation to disinformation, highlighting the various techniques of creating and distributing content, and the impact they had on the economy, society, and politics.

9.7 Examples

9.7.1 *The Volkswagen Emissions Scandal Spills over to BMW*

Even before the “fake news phenomena” made international headlines due to the 2016 US presidential elections, scandal, disinformation, misinformation, and unverified claims were plaguing the news ecosystem with considerable impact. One prominent example is how during the Volkswagen emissions scandal in the autumn of 2015, a German magazine managed to significantly impact the stock value of auto giant BMW after publishing an unverified claim that went viral.

On the September 18, 2015, the US Environmental Protection Agency (EPA) issued a Notice of Violation for Volkswagen AG, Audi AG, and Volkswagen Group of America, Inc., accusing the company group of installing “defeat” devices in certain car models that cheated on harmful emissions measurements. In fact, the EPA was accusing the VW Group of “rigging” their car emissions measuring devices by showing values that were lower than the real ones in order to conform to the strict US environmental regulations. The Notice of Violation was published on the EPA website [26] and was quickly picked up by the media the following day.

On September 19th, there were already several thousand online articles on this topic, according to Google Search, making it one of the biggest news stories of that week. One day later, VW CEO Martin Winterkorn confirmed the data published by the US EPA to be true. By September 21st, the VW stocks had plummeted by 20% [16] and a Wikipedia page on the “Volkswagen emissions scandal” [70] had been created.

While the media was busy trying to understand the magnitude of Volkswagen’s emissions rigging operation and also trying to assess the impact of this scandal on the auto industry worldwide, the German magazine *AUTO BILD* published an article claiming the rival BMW might have the same problem, stating that the BMW X3 model produced more than 11 times the European limit when tested by the International Council on Clean Transportation (ICCT). Published on September 24th under the title “*AUTO BILD exclusive: BMW diesel exhaust*

emissions exceed limits significantly" (original title: *AUTO BILD exklusiv: Auch BMW-Diesel überschreitet Abgas-Grenzwert deutlich*)¹³.

The article by AUTO BILD was picked up instantly and without proper verification by many publications and online blogs around the world, resulting in an immediate drop of BMW stock by nearly 10% [17]. BMW management denied any wrongdoing the same day and AUTO BILD modified the original article publishing a correctional statement: "*No indications of manipulation from BMW*" ("*Kein Indiz für Manipulation bei BMW*") (Auto [10]).

The mistake by AUTO BILD and its consequences can be regarded as a "textbook" example of disinformation – a type of fake news produced without intent. The effects of this disinformation by AUTO BILD were the same as for any other fake news, even if BMW was quick to deny the claim and the German auto magazine published a clarification only shortly afterwards. The information was quickly picked up by other media outlets and websites with no proper verification, mainly because AUTO BILD was regarded as a trustworthy news source, it spread rapidly in the online space, had significant impact (BMW stock crashing) and only a very small segment of those who picked up the information also published the correction afterwards.

9.7.2 Mainstream Media Is no Stranger to Fake News

If misinformation is produced when the editorial process does not enforce certain rules that have to do with content quality and source verification, disinformation is fake news created on purpose, whether by hyper-partisan blogs that use *clickbait*-techniques to spread hate speech, or false information about political adversaries, or by mainstream media actors using complex propaganda techniques in an effort to shape a wider audience's views on certain national or global topics.

One of the most frequent disinformation techniques employed by the media, especially when dealing with political topics, is using anonymous sources of information, referred to as "sources close to the matter," "experts that wish to remain unnamed," "undisclosed official sources," "confidants," or "people familiar with the case."

A relevant example of such disinformation carried out by the mainstream media is the alleged US arrest warrant for *Wikileaks* founder Julian Assange. On April 20, 2017, several US news publications reported the US Government was preparing charges for the arrest of *Wikileaks* founder Julian Assange. However, this information did not come from (nor was confirmed by) a named US official or institution. Leveraging statements from CIA Director Mike Pompeo and Attorney General Jeff Sessions on the *Wikileaks* investigation (dating back from 2010), the

¹³Article no longer available online, but archived copy available on presseportal.de: www.presseportal.de/pm/53065/3130280

claim that the US was preparing to charge Assange and *Wikileaks* was attributed to “US officials familiar with the matter” and “people familiar with the case.” The claim was never confirmed and remains so until today.

The information that the ‘US prepares charges to seek arrest of *Wikileaks*’ Julian Assange was first published by *CNN* [18] and then by *The Washington Post* [67] on April 20, 2017, just 13 min apart. The information was subsequently picked up by several other prominent news organizations, such as *Newsweek*, *Deutsche Welle*, *USA Today*, *The Verge*, *The Blaze*, *Russia Today*, and the *BBC*.

In fact, the information source cited by both *CNN* and the *Washington Post* was “US officials familiar with the matter” (*CNN*) and “according to people familiar with the case” (*Washington Post*). More than a year later, the US Government had not put forward any formal charges against Julian Assange, proving that either the sources were misinformed or were using disinformation. It is also possible that the publishers were either deceived by their unnamed sources or were just too quick to publish information that was impossible to verify.

It is commonly known that the use of unnamed sources for news stories can be accepted under certain conditions and circumstances. Frequently journalists have a choice between relying on unnamed sources and not publishing a story at all. Relying on unnamed sources that have produced reliable information in the past can be virtually risk-free. However, dealing with information that is not precise and lacks specifics can be an indication that the source is unreliable or even outright wrong. In such cases, journalistic quality standards should take precedence over being first and “breaking a story” before the competition, risking to produce fake news and lose the audience’s trust.

Another example of mainstream media-created “fake news” is that of *ABC News* claiming that presidential candidate Donald Trump instructed retired Lt. Gen. Michael Flynn to “contact the Russians,” an action that would be considered illegal in the USA. The story created insecurity in the financial markets and caused the Dow Jones to drop by 350 points.

On December 1, 2017, *ABC News* posted a message on Twitter regarding a journalistic investigation by senior journalist and anchor Brian Ross, gathering 25,000 retweets in the first hour:

“JUST IN: @BrianRoss on @ABC News Special Report: Michael Flynn promise full cooperation to the Mueller team and is prepared to testify that as a candidate, Donald Trump directed him to make contact with the Russians.”

This information, as part of the ongoing Russian collusion scandal, was aimed to confirm that candidate Donald Trump had colluded with Russia during the US Presidential Elections in 2016. After Michael Flynn was arrested and pleaded guilty to lying to the FBI on Friday, Ross reported that Flynn planned to tell authorities that President Trump had personally directed him to establish contact with Russia while he was running for president. This explosive claim, which suggested a new dimension to Robert Mueller’s investigation into Russian election interference, was picked up by many other journalists, and even caused a significant, temporary dip in the stock market, according to [NYmag.com](#) [46].

Brian Ross was suspended by ABC over the false Trump report the next day and apologized for the “serious error we made yesterday.” Journalist Brian Ross claimed that the mistake came from an unnamed source (referred to as “the confidant”) that “had actually told him that Trump had made the comments in question when he was the president-elect, not a candidate,” but after the original information had made it into his article.

The original ABC Twitter message containing the mistake was distributed 25,000 times before it was deleted and the referenced article [2] topped 100,000 Facebook interactions, as measured with www.sharedcount.com on December 11, 2017. More than 8500 news websites around the world picked up and republished the information, according to Google News search. Not all of the websites that picked up the story also published the ABC correction or the fact that author Brian Ross was suspended for his mistake. The quantifiable impact of this misinformation or disinformation was a significant drop of the Dow Jones by 350 points, an impact similar to the one created by AUTO BILD’s mistake regarding BMW’s emissions values.

9.7.3 Conflict Zones, the Playground for Propaganda

A very rigid approach on information verification could have a binary classification viewpoint: true or false. However, when dealing with information originating from conflict zones, as for example the war in Syria, the notion of “unverifiable information” comes into play, due to the multiple actors trying to control the information and the fact that on-the-ground journalistic information verification is nearly impossible. The Syrian conflict has been marked by constant propaganda and disinformation campaigns since its beginning in 2011 with all involved belligerents, whether Bashar al Assad and his allies, the Western Coalition, or the Islamic State.

One example from October 2015 is the news that a bombing by Russian warplanes of a hospital in the city of Sarmin, in the Idlib Governorate, Syria, resulted in the death of eight civilians. The article was published on the SOHR – Syrian Observatory for Human Rights – website www.syriahr.com (edited by Rami Abdurrahman, a Syrian national living in the UK) on October 2015 [57], but is not available anymore. On October 21, 2015, the information from the SOHR expanded into a full story for the AFP and subsequently was picked up by most major news agencies including DW, Radio Free Europe, Mail Online [21], NDTV [44], and Sky News, citing the SOHR website administrator, Rami Abdurrahman. On October 22, 2015, the information was then confirmed by the Syrian-Medical Society (Syrian American Medical Society or SAMS) by posting a picture of the alleged bombed hospital on Twitter. This was followed by a statement on their website [50] claiming the medical facility was a SAMS polyclinic in Sarmin, Idlib and that it was targeted by Russian airstrikes using air-to-surface missiles. The statement furthermore provided exact figures on the number of casualties, the SAMS

staff members killed in the attack and four pictures of the aftermath that are now missing from their website.

The state-funded news agency Russia Today published a statement by the spokesperson of the Russian Foreign Ministry on the matter the next day [48]. Maria Zakharova denied any Russian bombing of hospitals in Syria, claiming the report showed tremendous bias towards Russia's military efforts in Syria:

There are so-called mass media reports which allege that Russian aircraft bombed a field hospital in the Idlib Governorate in northwestern Syria and reportedly killed 13 people. I cannot say that these reports are written by journalists but their ingenuity delights.

[RT.com](#) then followed with an update on the story on October 29 [49], publishing a statement from Dominik Stillhart, director of operations at the International Committee of the Red Cross, (which has people on the ground in Syria), saying that he was unaware of any such incidents: "We've seen these reports as well, but in the absence of any firsthand information coming from our teams on the ground, I can neither confirm, nor deny these allegations."

The conclusion of Dominik Stillhart summarizes very well the way stories from conflict zones should be approached. As the source of this story is a statement from a Syrian activist (Rami Abdurrahman), followed by more information from a US-funded medical association (SAMS) featuring images that could not be verified, with no real follow-up on the story in order to find more evidence or witnesses and on-the-ground footage from a reliable independent source, it is actually impossible to say whether this story is true or false.

The same goes for another, more recent story regarding the Syrian conflict: the bombing of Syrian army positions by the US Army on May 24, 2018. The information was published by several media outlets, only to be disputed shortly afterwards. The source of information was traced back to the Reuters World Twitter account (twitter.com/ReutersWorld), citing the Syrian Arab News Agency (SANA):

U.S.-led coalition hits Syrian army positions: Hezbollah media unit <https://reut.rs/2IKBane>.

The tweet linked to the story that was updated quickly after the post on Twitter: "Syrian state media says U.S. hit army posts, U.S. denies" [47] tracing the information source to Syrian state media agency SANA that was citing "a military media unit run by Lebanon's Hezbollah." The information was denied by US military official Captain Bill Urban, a spokesman for the US Central Command as well as by Russian military sources cited by [RT.com](#).

The original post of Reuters was published on the SANA website [51] at 2018-05-23 23:40:42 and was deleted shortly thereafter. But not before being referenced by a number of Arab-language publications, including the Syrian Observatory for Human Rights and international news agencies. The SANA article, authored by "mohamad" only contained a short sentence (retrieved through search engine cache and translated from Arabic):

"Military source: Some of our military positions between Albuqmal and the intimacy of the dawn of the day to the aggression launched by the US coalition

aircraft in conjunction with the rallies of the terrorists of the organization advocated and limited damage to material”

One hour later, the same author (mohamad) published a more detailed piece on the same story [52] (translated from Arabic): “Within the framework of its support for the “daaish terrorists. The US coalition attacks on some of our military Prevlaka (rural) Deir Al-Zour.” The article used the same anonymous source as the previous article and mentioned that “The aggression after less than 24 hours to foil the Syrian Arab army forces and the generic attacked terrorists from “Daaish” on a number of points in the apparent fields Prevlaka (rural) Deir Al-Zour and the elimination of more than 10 of them some of them foreign nationalities and dozens of injured and the destruction of one of their vehicles.” It also specified that there were no injuries on the Syrian side.

The information was quickly picked up by Arab-language publications:

- **Al Majd** [7]. Citing SANA, claiming the loss of 25 civilian lives. The article has been deleted since.
- **Asharq Al Awsat** [1]. Using the information from SANA.
- **Alsumaria** [8]. Citing SANA, but mentioning the US did not confirm the attack. In addition, it mentioned that Syrian military sources claimed the attack targeted two sites near the T2 Atomic Energy facility, which is located near the border with Iraq, about 100 km west of the Euphrates River.
- **Al Bawaba** [4]. Identical to the AlSumaria article.
- **Sky News Arabia** [56]. The first Arab-language publication to mention that the source was actually Hezbollah.
- **Aljazeera** [6]. Mentioning that the Pentagon did not confirm the attack and attributing the information to the Syrian Government (SANA).

English-language publications that picked up the story and then published the Pentagon’s denial were: The Associated Press, North Jefferson News, *The Seattle Times*, *The Washington Post*, *New York Post*, Russia Today, *The Times of Israel*.

This story can be considered a successful source-hacking attempt orchestrated by Hezbollah. Publishing unverifiable information on the Syrian state news agency website SANA was carried out in order to be used by other Arab-language news websites and hopefully also by English mainstream news agencies, such as Reuters. Again, the unnamed Hezbollah source, the impossibility to verify the claim, and the risk of being just a daily piece of wartime propaganda did not prevent this story from reaching mainstream media in the English-speaking world although it was denied afterward by US and Russian military sources. It also provides an example of cross-language and multilingual efforts, making any verification and debunking process even more difficult.

Another example of how propaganda works in the Syrian conflict zone was debunked by the Atlantic Council’s Digital Forensics Lab¹⁴ in September 2017: “How Kremlin-backed and fringe media spread a false story claiming the U.S.-

¹⁴<https://medium.com/dfrlab>

led Coalition evacuated ISIS from the front lines.” The investigation focused on an article by Russian News Agency (Sputnik) trying to trace the information back to its original source [42].

Confirming the pattern, Sputnik’s information source is an anonymous “military and diplomatic” source, while another website – the Syrian-based Deirezor24.net [25] – cites “A D24 correspondent.” Although the information did not reach mainstream media, it was picked up by conspiracy media outlets such as [the-freethoughtproject.com](#), Al-Masdar News, The Fringe News, or The Duran.

One of the obvious patterns of disinformation in the Syrian conflict is that from the main belligerents – the Syrian government, the Russian military, and the US-led coalition – there is rarely any consensus on information among the three actors. When information is published by the Syrian Government, for example, it is quickly denied by the USA. When it is published by the USA and mainstream media, it is labeled as propaganda by the Russian officials and so on. In the end, the control of information on the Syrian conflict is just another facet of the way modern war – an information war – is conducted, with the public having great difficulty assessing the trustworthiness of information regarding this civil war which has been raging on since 2011.

9.7.4 Clickbait-Type Fake News for Political Gains

The term “fake news” came into global attention shortly after the US presidential elections in 2016 when the victory of Donald Trump was marred with accusations of his supporters using disinformation and clickbait techniques in the online space and more specifically on Facebook to target his opponent, Hillary Clinton, and the Democratic Party. Some statements even went so far as to claim that fake news might have gotten Donald Trump elected, as stated in an article by The Guardian [33] as “the most obvious way in which Facebook enabled a Trump victory has been its inability (or refusal) to address the problem of hoax or fake news.”

According to TechCrunch [61], the issue of spreading misinformation on the social media platform was real and was admitted by Adam Mosseri, VP of product management at Facebook, who conceded that “the company does need to do more to tackle this problem.” Later, Facebook founder Mark Zuckerberg denied that this phenomenon had an influence on getting Trump elected: “To think it influenced the election in any way is a pretty crazy idea” (statement made during the Techonomy conference in Half Moon Bay, Calif, cited by USA Today [66]). The culprits for this “fake news surge”, as it was called by the AI-startup Zetta Cloud [62], were anonymous supporters from all over the globe engaging in a sustained campaign to publish and distribute fake news on social media, with the aim of lowering the US public’s trust in the Democratic Party and candidate Hillary Clinton.

The claim that this fake news campaign was solely responsible for Trump’s victory was later corrected by a study pointing out that wide reach does not equal strong impact [24]. But the numerous sites that are still spreading fake news with

strong “clickbait” recipes represent the bulk of “fake news” sites known to the general public. Some examples of this type of hyper-partisan fake news show obvious patterns of how this type of disinformation can have such a wide reach in online communities.

The post “Damning Hillary Clinton Footage Leaks – The Truth is Out” [64] published on May 14, 2017, by TheTruthMonitor.com¹⁵ indicates another right-wing fake news story that is set out to “viralize” quickly and be instrumented as political propaganda. The title writing style aimed at generating strong emotions and a desire to click the headline (thus the term *clickbait*) containing misleading information is something that is commonly used when creating fake news.

The article, having as author “Adam Lott” (with no contact details or social media profile), trashes Hillary Clinton and her supporters in the context of Donald Trump firing FBI Director Comey while explaining the previous firing of state attorney Preet Bharara by using the precedent of Bill Clinton firing all of Bush’s state attorneys back in 1993. Even though the actual content of the article and the sources it references (a tweet from CNN supporting state attorney Preet Bhahara and a tweet from political commentator Dinesh D’Souza [65] do not present any “damning information” or any “footage leaks “regarding Hillary Clinton, this didn’t stop the article from going viral, gaining more than 10,000 Facebook engagements.

Another example is the targeting of other Democrats, such as America’s soon-to-be first female Muslim legislator, Ilhan Omar. In August 2016, the website “World Net Daily” [71] published a story accusing Ilhan Omar of being married to two men at the same time, including one who may be her brother. The story, published during Minnesota’s August 9, 2016, Democratic primary, won by 33-year-old Somali refugee Omar, originated from the blog of lawyer Scott Johnson¹⁶ and resulted in what the newspaper *Minnesota Post* [43] called a “five-day brush fire” causing serious issues for the candidate and resulting in a clarification: Ilhan Omar was married to only one man, who had changed name from Ahmed Hirsi to Ahmed Aden, hence the confusion.

Although this story was clarified in 2016, it reappeared 1 year later, on July 20, 2017. Several right-wing blogs supporting Donald Trump, such as dailyinfo.co, usanewspolitics.com, conservativefighters.com [19], TeaParty.org, and angrypatriotmovement.com, published a post with the exact same title: “JUST IN: Muslim Congresswoman Caught In SICK Crime. Should She Go To Prison?”

The posts, which together gathered close to 40,000 Facebook interactions, claim that Ilhan Omar married her own brother in a scheme to fool US immigration officials and supported refugee fraud in her state. The content of these posts picks up the year-old information from World Net Daily without the debunking and the clarifications, and “spices it up” with the information that Omar might go to jail. It employs the successful technique of taking old information that was debunked in the meantime and using it for the same purpose, knowing that people rarely read or

¹⁵ Site not online anymore.

¹⁶ www.powerlineblog.com

remember the corrections of fake stories. To accelerate the distribution speed, the clickbait title makes sure to be highly aggressive, suggesting that the scandal just happened and not mentioning the congresswoman's name.

Equally, on the other side of the US political spectrum, there is no reluctance to use clickbait titles to ensure a wide audience reach by mixing half-truths in order to obtain political gains. The liberal LA-based blog [yournewswire.com](#) has already gathered more than 106,000 Facebook interactions for an article published in May 23, 2018: "20 Priests Connected To Vatican Pedophile Ring Killed In Plane Crash" [72].

After a tragic plane crash in Cuba on May 19th, claiming the lives of 20 priests [23], Daily Mail authors link this information with a child sex abuse scandal from Chile involving the Catholic Church [22], stating that the priests perished in Cuba were pedophiles. Besides a "clickbait"-style title, the article uses a technique of merging disconnected information and ineptly "connects the dots" in order to create a fake news story that gathered significant traction on Facebook. YourNewsWire, a known fake news website, is published by Sean Adl-Tabatabai and Sinclair Treadway and has made the headlines in mainstream media for being an "alt-left agitator website" [35].

9.7.5 When Whole Countries Fall Victim to Fake News

Moving away from the English-language online space, fake news phenomena can become a nationwide critical problem in areas where digital literacy is low among the population. One relevant example is the Republic of Indonesia, with a diverse population of 261 million and a fake news problem that affects every part of its society.

From the smear campaign against Indonesian president Joko "Jokowi" Widodo in 2015 to Jakarta's Christian and ethnically Chinese governor Basuki Tjahaja Purnama, popularly known as Ahok (who was targeted by hardline Muslims, who successfully pressured the authorities to put him on trial for blasphemy in 2016), fake news is used in an organized and almost weaponized manner by political groups in Indonesia, resulting in mass protests, political crisis, and religious and ethnic hate.

In February 2018, the Indonesian government established a National Cyber and Encryption Agency (BSSN) with the goal of fighting the fake news phenomenon that has engulfed the country's political scene. A plan to establish the BSSN was put forth 4 years when president Joko Widodo took office, but the agency only started work in January 2018 after Major General Djoko Setiadi was installed as its leader [59]. In fact, current president Joko Widodo was the target of a smear campaign during the presidential elections of 2014 when a social media campaign was circulating rumors that the popular Jakarta governor was an ethnic Chinese and Christian – a sensitive issue in a Muslim-majority country with a low digital literacy level. Called "black campaigns" in Indonesia, these were aimed at hurting Widodo's electoral score in favor of more radical Muslim candidates. Pictures were

circulated of a fake marriage certificate claiming that Widodo is of Chinese descent and originally a Christian. The messages were spread over Blackberry message groups, WhatsApp, Facebook, and Twitter.

Featuring among the top five biggest users of Facebook and Twitter globally, Indonesia has faced regular and very effective fake news campaigns: from memes, like the one targeting social-media activist Ulin Yusron, to edited videos, like the one that got Jakarta's governor put on trial for blasphemy, to online-organized bounty hunts against more than 100 individuals who saw their personal information released online.

The Indonesian authorities are expecting a similar surge of fake news for the 2019 elections. Recently, according to the Jakarta Globe (Jakarta [36]), the group orchestrating most of these fake news campaigns, called the Muslim Cyber Army, was exposed and its members arrested. The group was accused of running an online army of Twitter bots, messaging groups, and fake accounts with the purpose of spreading fake news and smear campaigns against certain political actors while also fueling a state of paranoia and tension against the Christian and Chinese minorities.

9.8 Conclusion and Outlook

Fake news is an age-old phenomenon, from the disinformation campaigns in ancient empires, to the recent social-media campaigns aimed to influence election results. Along this timeline, there seems to be a balance between the skills and tools required for people to protect themselves from disinformation and the methods and distribution channels employed by the creators of fake news. When the distribution channel for fake news was the word-of-mouth, people would rely on their common sense and on the messenger to assess the trustworthiness of the information. Nowadays, digital literacy and the possibility to verify information in real-time across multiple sources (and some common sense) are the basic tools.

Even if the media agree on content creation models that enforce transparency and result in high-quality trusted content, even if online distribution channels automatically filter out or flag untrusted content using up-to-date algorithmic approaches, the Internet remains a place where anyone can create and distribute content easily. The importance of the human factor in the dissemination of fake news and the responsibility in the sharing of information cannot be stressed enough. Ultimately, the burden to judge the content rests with the audience, the content-consumers.

Digital literacy education programs are a must in the age of the Internet. A dedicated focus, both financial and academic, into the creation of software and tools to help content creators, distributors, and consumers to easily identify and filter-out untrustworthy content, is the technology backbone that will result in an applicable and scalable method to detoxify online information. Ethical issues such as possible censure and freedom of speech need to be taken into account from the start.

When it comes to being aware of what information to trust, the burden on the audience can be immense. Information quality and perceived information quality play a fundamental role, in particular on social media platforms, where the sharing of information is one of the key elements. The inclination to share rapidly and in an unreflected manner is only occasionally balanced by the fear of loss of reputation.

Continuous education about the techniques and patterns typically employed by producers of fake news, becoming immune to clickbait titles and resisting the tendency to immediately interact with and share content that has not been verified or that originates from untrustworthy sources, making a habit of cross-checking information from other sources, and using news database services and image verification tools – these and related measures and activities cause a heavy load and require knowledge and skills in order to stay “fake news free.” In sectors and societies where social media platforms and the Internet are still a *new thing*, this task can become nearly impossible.

Looking further into the AI future, conversational bots that can be easily mistaken for human users, deep fakes created with advanced image and video-processing technologies, the emergence of content that is created automatically – “robot journalism” – these new advancements will add even more complexity to the online information creation and dissemination landscape.

The solution needs to be one that will work in the long-term and is most likely to be a hybrid one: establishing digital literacy education as a must, from an early age, across as many educational systems as possible, combined with a continuous effort to create and deploy AI-based software that automate most of the tasks required to identify fake news and support consumers in their struggle.

The two approaches have little chances to be successful by themselves individually. Digital literacy educational programs are now mostly operated by non-governmental organizations; it will still take some years before they are adopted, adapted, and improved by governments and made part of the regular educational curricula. And even when they are, they will still need to keep up with the ever-new ways of creating and distributing “fake news.” On the other side, AI algorithms, even in their most advanced form, will probably never be as good as humans, when it comes to analyzing complex culture- and language-specific disinformation campaigns. But AI’s goal is not to replace human assessment, but rather provide a scalable method of addressing the most obvious forms of disinformation and to be able to quickly extract insights for humans to review, shortening investigation times. Time and scale are of essence and can be addressed by AI technologies. The two approaches, taken together, have the best chances of curbing the fake news surge in the online space, given that enough focus is given from society’s leading actors: governments, academia, and technology entrepreneurs. These efforts need to be accompanied and supported by further research into the detection of fake news, its dissemination, the role of information quality, and perceived information quality in such processes and how such insights could readily be incorporated into social media platforms to support users in their struggle. At the same time, commercial entities need to embrace such advances and provide tools and plug-ins allowing to put these results into practice.

References

1. Aawsat, Aawsat.com (2018) [سوریہ عسکریة موافق ع لی غارات پ شن الامیرکی الا تحالف](#) /US Coalition Launches Raids on Syrian Military Sites. Accessed 24 May 2018
2. ABC NEWS (2017) Flynn Prepared to Testify that Trump Directed him to Contact Russians about ISIS, Confidant Says. <https://abcnews.go.com/Politics/michael-flynn-charged-making-false-statements-fbi-documents/story?id=50849354>. Accessed 8 Aug 2018
3. E. Agichtein, C. Castillo, D. Donato, A. Gionis, G. andMishne, Finding high-quality content in social media, in *International Conference on Web Search and Data Mining*, (2008), pp. 183–193
4. AlBawaba, AlBawaba.com (2018) [السوری لج یشن موافق پ چصف "الا تحالف": الازور دبور](#) /Deir al-Zour: "Alliance" bombing sites of the Syrian army". Accessed 24 May 2018
5. H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election. *J. Econ. Perspect.* 31(2), 213 (2018)
6. AlJazeera, AlJazeera.net (2018) [السوریة ب الا بادیة الا نظام لموافق أمیرکی چصف](#) /US bombardment of the regime sites in the Syrian desert. Accessed 24 May 2018
7. AlMajd, AlMajd.net (2018) [الا تحالف ایدی على الحدّسکة ب روی سوریا مدنیا 25 اس تشهاد](#) /25 Syrian Civilians Were Killed in Al-Hasakah Countryside by the American Coalition. Accessed 24 May 2018
8. AlSumaria, AlSumaria.tv (2018) [الازور ب دبور الا سوری لج یشن موافق چصف الا تحالف: سانا](#) /SANA: Coalition Bombed Sites of the Syrian Army in Deir al-Zour. Accessed 24 May 2018
9. G. Arevalo (2018) The Six Hoaxes of Edgar Allan Poe, available at: <https://hubpages.com/literature/The-Six-Hoaxes-of-Edgar-Allan-Poe>. Accessed on 29 July 2018
10. B. Auto (2018) Kein Indiz für Manipulation bei BMW. <http://www.autobild.de/artikel/klarstellung-abgaswerte-bei-bmw-diesel-6920195.html>. Accessed 8 Aug 2018
11. Brandwatch, Content Statistics (2018) Available at: <https://www.brandwatch.com/blog/96-amazing-social-media-statistics-and-facts/>. Accessed 4 Aug 2018
12. J. Burkhardt (2017) History of Fake News (Chapter 1), in: Combating Fake News in the Digital Age, (Chicago, IL), p. 6
13. Buzzfeed (2016) How Teens in the Balkans are Duping Trump Supporters with Fake News. <https://www.buzzfeednews.com/article/craigsilverman/how-macedonia-became-a-global-hub-for-pro-trump-misinfo#.fu2okXaeKo>. Accessed on 10 Aug 2018
14. G.L. Ciampagalla, F. Menczer (2018) *Misinformation and biases infect social media, both intentionally and accidentally*, in: The Conversation UK, available at: <https://theconversation.com/misinformation-and-biases-infect-social-media-both-intentionally-and-accidentally-97148>. Accessed 4 Oct 2018
15. CNET, CNET.com (2018) What is QAnon? The Deep State vs. Trump conspiracy, explained. <https://www.cnet.com/news/what-is-qanon-deep-state-trump-conspiracy-explained>. Accessed 8 Aug 2018
16. CNN Money (2018) Volkswagen stock crashes 20% on emissions cheating scandal. <https://money.cnn.com/2015/09/21/investing/vw-emissions-cheating-shares/index.html>. Accessed 8 Aug 2018
17. CNN Money (2018) BMW shares plunge on report it broke pollution limit. <https://money.cnn.com/2015/09/24/investing/bmw-diesel-emissions-shares/index.html>. Accessed 8 Aug 2018
18. CNN (2018) Sources: US prepares charges to seek arrest of WikiLeaks' Julian Assange. <https://edition.cnn.com/2017/04/20/politics/julian-assange-wikileaks-us-charges/index.html>. Accessed 8 Aug 2018
19. Conservativefighter, Conservativefighter.com (2017) JUST IN: Muslim Congresswoman Caught In SICK Crime. Should She Go To Prison? <http://conservativefighters.com/news/just-muslim-congresswoman-caught-sick-crime-go-prison/>. Accessed 20 May 2017

20. CS181Journalism (2018) Journalism in the digital age. The Echo Chamber Effect. A Common Critique and its Implications for the Future of Journalism, available at: <https://cs181journalism2015.weebly.com/the-echo-chamber-effect.html>. Accessed on 5 Aug 2018
21. Daily Mail (2015) 13 Dead as Russia Strike Hits Syria Field Hospital: Monitor, <http://www.dailymail.co.uk/wires/afp/article-3283108/Russia-air-strike-field-hospital-kills-13.html>. Accessed 8 Aug 2018
22. Daily Mail (2018) All of Chile's 34 Bishops RESIGN over a Sex Abuse and Cover-Up Scandal after Crisis Meeting with the Pope. <http://www.dailymail.co.uk/news/article-5744503/Chilean-bishops-offer-resignation-Pope-abuse-scandal-statement.html>. Accessed 8 Aug 2018
23. DallasNews (2018) 20 Priests among Dead, 3 Survivors 'critical' in Havana Plane Crash, <https://www.dallasnews.com/business/airlines/2018/05/18/cuban-media-boeing-737-crashes-shortly-after-takeoff-104-passengers-aboard>. Accessed 8 Aug 2018
24. Dartmouth College (2018) Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign, <http://www.dartmouth.edu/~nyhan/fake-news-2016.pdf>. Accessed 8 Aug 2018
25. Deirezor24 (2018) With the Approaching of the Battle of DeirEzzor, the Coalition Evacuate Two of their Spies Who Operated as Daesh Commanders in Western DeirEzzor, Den. deirezor24.net/with-the-approaching-of-the-battle-of-deir-ezzor-the-coalition-evacuate-two-of-their-spies-who-operated-as-daesh-commanders-in-western-deir-ezzor/. Accessed 8 Aug 2018
26. EPA, EPA.gov (2015) <http://www3.epa.gov/otaq/cert/documents/vw-nov-caa-09-18-15.pdf>. Accessed 27 Oct 2015
27. T.D. Erickson, M.E. Mattson, From words to meaning: A semantic illusion. J. Verbal Learn. Verbal Behav. **20**, 540–551 (1981)
28. The Fake News Machine, How Propagandists Abuse the Internet and Manipulate the Public, 2018. https://documents.trendmicro.com/assets/white_papers/wp-fake-news-machine-how-propagandists-abuse-the-internet.pdf. Accessed 8 Aug 2018
29. L. Fazio (2018) Why you Stink at Fact-Checking, the Conversation UK, available at: <https://theconversation.com/why-you-stink-at-fact-checking-93997>. Accessed on 4 Aug 2018
30. G. Gigerenzer, P.M. Todd, ABC Research Group, *Simple Heuristics that Make Us Smart* (Oxford University Press, Oxford, 2000)
31. D.T. Gilbert, How mental systems believe. Am. Psychol. **46**(2), 107–119 (1991)
32. Google Books Ngram Viewer (2018) Example "fake news", available at: https://books.google.com/ngrams/graph?content=fake+news&year_start=1800&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t1%3B%2Cfake%20news%3B%2Cc0#t1%3B%2Cfake%20news%3B%2Cc0. Accessed 29 July 2018
33. The Guardian (2016) Facebook's failure: did fake news and polarized politics get Trump elected? <https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories>. Accessed 8 Aug 2018
34. L.K. Hansen, A. Arvidsson, F.A. Nielsen, E. Colleoni, M. Etter, Good friends, bad news – affect and Virality in twitter, in *Future Information Technology. Communications in Computer and Information Science*, eds. by J. J. Park, L. T. Yang, C. Lee, vol. 185, (Springer, Berlin/Heidelberg, 2011)
35. The Hollywood Reporter (2017) L.A. Alt-Media Agitator (Not Breitbart) Clashes with Google, Snopes. <https://www.hollywoodreporter.com/features/hollywoods-hidden-alt-media-firebrands-1041157>. Accessed 8 Aug 2018
36. J. Globe (2018) Police Arrest Four Members of 'Muslim Cyber Army'. <http://jakartaglobe.id/news/police-arrest-four-members-muslim-cyber-army/>. Accessed 8 Aug 2018
37. Koohikamali, M., and Sidorova, A. (2017).Information re-sharing on social network sites in the age of fake news. Inform Sci, 20, 215–235, Retrieved from <http://www.informingscience.org/Publications/3871> on 08/12/2018
38. K. Leetaru. (2017) Did Facebook's Mark Zuckerberg Coin the Phrase 'Fake News?' Available at: <https://www.forbes.com/sites/kalevleetaru/2017/02/17/did-facebooks-mark-zuckerberg-coin-the-phrase-fake-news/#748b02806bc4>. Accessed 29 July 2018

39. K. Leetaru, History of fake news (Chapter 1), in *Combating fake news in the digital age*, (Chicago, IL), pp. 5–9
40. E.J. Marsh, S. Umanath, Knowledge Neglect: failures to notice contradiction with stored knowledge, in *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational science*, ed. by D. N. Rapp, J. Braash (Eds), (Cambridge, MA 2013)
41. E.J. Marsh, A.D. Cantor, N.M. Brashier, Chapter three – Believing that humans swallow spiders in their sleep: False beliefs as side effects of the process that support accurate knowledge, in *Psychology of Learning and Motivation*, vol. 64, (2016), pp. 93–132
42. Medium, Medium.com (2018) Questionable Sources on Syria. <https://medium.com/dfrlab/questionable-sources-on-syria-36fcabddc950>. Accessed 8 Aug 2018
43. MinnPost, MinnPost.com (2016) What the Ilhan Omar story says about the media and political reporting in 2016. <https://www.minnpost.com/politics-policy/2016/08/what-ilhan-omar-story-says-about-media-and-political-reporting-2016>. Accessed 20 May 2017
44. NDTV (2015) 13 Dead as Russia Strike Hits Syria Field Hospital: Monitor. <https://www.ndtv.com/world-news/13-dead-as-russia-strike-hits-syria-field-hospital-monitor-1235011>. Accessed 8 Aug 2018
45. S.J. Nightingale, K.A. Wade, D.G. Watson, Can people identify original and manipulated photos of real-world scenes? *Cogn Res* 2(30), 1–21 (2017)
46. NYMAG.com (2017) ABC Suspends Brian Ross Over False Trump Report. <http://nymag.com/daily/intelligencer/2017/12/abc-suspends-brian-ross-over-false-trump-report.html>. Accessed 8 Aug 2018
47. Reuters (2018) Syrian State Media Says U.S. Hit Army Posts, U.S. denies. <https://www.reuters.com/article/us-mideast-crisis-syria-coalition/u-s-led-coalition-hits-syrian-army-positions-hezbollah-media-unit-idUSKCN1IO3HB>. Accessed 8 Aug 2018
48. Russia Today (2015) War of Words: Russian Foreign Ministry Calls out MSM Reports on Hospital Strike in Syria. <https://www.rt.com/news/319444-russia-hospital-media-ministry/>. Accessed 27 Oct 2015
49. Russia Today (2015) No Firsthand Info on Alleged Russian ‘airstrike’ on Hospital in Syria – Red Cross Top Executive. <https://www.rt.com/news/320046-stillhart-red-cross-hospital-russia/>. Accessed 29 Oct 2015
50. SAMS, Syrian American Medical Society, SAMS-USA.net (2015) Press Release: Two Medical Staff Killed in Russian Airstrikes on Sarmin. https://www.sams-usa.net/press_release/press-release-two-medical-staff-killed-in-russian-airstrikes-on-sarmin/. Accessed 27 Oct 2015
51. SANA, Syrian Arab News Agency (2018) <https://www.sana.sy/?p=758691>. Accessed 24 May 2018
52. SANA, Syrian Arab News Agency (2018) <https://www.sana.sy/?p=758694>. Accessed 24 May 2018
53. R. Saxena (2017) The social media “echo chamber”. Active social media users are self-segregated and polarised in news consumption. ArsTechnica, Available at: <https://arstechnica.com/science/2017/03/the-social-media-echo-chamber-is-real/>. Accessed 4 Aug 2018
54. Schmidt, A. L., Zollo, F., Del Vicario, M., Bessi, A., Scala, A., Caldarelli, G., Stanley, H. E. and Quattrociocchi, W. (2017) Anatomy of news consumption on Facebook, in: PNAS Volume 114 (12), p. 3035–3039
55. Silverman, C. and Alexander, L. (2016) How Teens in the Balkans Are Duping Trump Supporters with Fake News. Available at: <https://www.buzzfeednews.com/article/craigsilverman/how-macedonia-became-a-global-hub-for-pro-trump-misinfo>. Accessed 29 July 2018
56. Sky News Arabia (2018) لoyalists Killed by Coalition Strike East of Syria. Accessed 24 May 2018
57. SOHR, The Syrian Observatory of Human Rights (2015) <http://www.syriahr.com/en/2015/10/12-explosive-barrels-on-daraya-and-russian-airstrikes-on-idlib/>. Accessed 27 Oct 2015

58. J. Soll (2016) The Long and Brutal History of Fake News. Available at <https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535>. Accessed 29 July 2018
59. South China Morning Post (2018) Can Indonesia's New Cybercrime Unit Win Its War on Fake News? <https://www.scmp.com/week-asia/geopolitics/article/2132683/can-indonesias-new-cybercrime-unit-win-its-war-fake-news>. Accessed 8 Aug 2018
60. T. Standage (2017) The True History of Fake News. The History of Made-up Stories Explains why they Have Re-Emerged in the Internet Era. Available at: <https://www.1843magazine.com/technology/rewind/the-true-history-of-fake-news>. Accessed 29 July 2018
61. TechCrunch (2016) Facebook Admits it Must Do More to Stop the Spread of Misinformation on its Platform. <https://techcrunch.com/2016/11/10/facebook-admits-it-must-do-more-to-stop-the-spread-of-misinformation-on-its-platform/?ncid=rss>. Accessed 8 Aug 2018
62. TrustServista, TrustServista.com (2016) A Different Approach to Handling the Fake News Surge. <https://www.trustservista.com/2016/11/a-different-approach-to-handling-the-fake-news-surge/>. Accessed 8 Aug 2018
63. TrustServista (2018) TrustServista Publishes its First News Verification Report. <https://www.trustservista.com/2017/08/trustservista-publishes-its-first-news-verification-report/>. Accessed 8 Aug 2018
64. Truthmonitor, TruthMonitor.com (2017) Damning Hillary Clinton Footage Leaks – The Truth is Out. <https://www.truthmonitor.com/2017/05/damning-hillary-clinton-footage-leaks-the-truth-is-out/>. Accessed 20 May 2017
65. Twitter (2017) So Stop Crying Bharara: This is the letter Sessions got when he was booted by the incoming Clinton administration. <https://twitter.com/DineshDSouza/status/840687661148966915>. Accessed 20 May 2017
66. USA Today (2016) Mark Zuckerberg: Facebook fake news didn't sway election. <https://eu.usatoday.com/story/tech/news/2016/11/10/mark-zuckerberg-facebook-fake-news-didnt-sway-election/93622620/>. Accessed 8 Aug 2018
67. Washington Post (2017) National Security: Justice Dept. debating charges against WikiLeaks members in revelations of diplomatic, CIA materials. https://www.washingtonpost.com/world/national-security/justice-dept-debating-charges-against-wikileaks-members-in-revelations-of-diplomatic-cia-materials/2017/04/20/32b15336-2548-11e7-a1b3-faff0034e2de_story.html?utm_term=.fdf077127cfdf. Accessed 8 Aug 2018
68. Wikipedia (2018) The War of the Worlds (radio drama). Wikipedia entry, available at: [https://en.wikipedia.org/wiki/The_War_of_the_Worlds_\(radio_drama\)](https://en.wikipedia.org/wiki/The_War_of_the_Worlds_(radio_drama)). Accessed 29 July 2018
69. Wikipedia (2018) List of Conspiracy Theories. https://en.wikipedia.org/wiki/List_of_conspiracy_theories. Accessed 8 Aug 2018
70. Wikipedia (2018) Volkswagen emissions scandal. https://en.wikipedia.org/wiki/Volkswagen_emissions_scandal. Accessed 8 Aug 2018
71. World Net Daily (2016) Somali Muslim Candidate Denies Accusations of Bigamy. <https://www wnd com/2016/08/somali-muslim-candidate-denies-accusations-of-bigamy/>. Accessed 20 May 2017
72. YourNewsWire (2018) 20 Priests Connected To Vatican Pedophile Ring Killed In Plane Crash, <https://yournewswire.com/20-priests-pedophile-ring-plane-crash/>. Accessed 8 Aug 2018

Chapter 10

Information Quality and Social Networks



Pontus Svenson

Abstract Decision-making requires accurate situation awareness by the decision-maker, be it a human or a computer. The goal of high-level fusion is to help achieve this by building situation representations. These situation representations are often in the form of graphs or networks, e.g. they consist of nodes, edges and attributes attached to the nodes or edges. In addition to these situation representation networks, there can also be computational networks in fusion systems. These networks represent the computations that are being performed by the fusion system. Yet another relation between networks and fusion is that today much information comes from sources that are inherently organised as a network. The first example of this that comes to mind is the use of information from social media in fusion processes. Social media are also networks, where the links are formed by follow/reading/friend relations. There can also be implicit links between information sources that come from other It is vital for the fusion process and the ensuing decision-making to ensure that we have accurate estimates of the quality of various kinds of information. The quality of an information element has several components, for instance, the degree to which we trust the source and the accuracy of the information. Note that the source could be a high-level processing system itself: a fusion node that processed information from, e.g. sensors, and outputs a result. In this case, the quality determination must take account also of the way that the fusion node processed the data. In this chapter, we describe how social network analysis can help with these problems. First, a brief introduction to social network analysis is given. We then discuss the problem of quality assessment and how social network analysis measures could be used to provide quantitative estimates of the reliability of a source, based on its earlier behaviour as well as that of other sources.

Keywords Complex networks · Social network analysis · Information quality

P. Svenson (✉)
Kogma AB, Stockholm, Sweden
e-mail: pontus@kogma.se

10.1 Introduction

Decision-making requires accurate situation awareness by the decision-maker, be it a human or a computer. The goal of high-level fusion [1] is to help achieve this by building situation representations. These situation representations are often in the form of graphs or networks, e.g. they consist of nodes, edges and attributes attached to the nodes or edges.

In addition to these situation representation networks, there can also be computational networks in fusion systems. These networks represent the computations that are being performed by the fusion system. Yet another relation between networks and fusion is that today much information comes from sources that are inherently organised as a network. The first example of this that comes to mind is the use of information from social media in fusion processes. Social media are also networks, where the links are formed by follow/reading/friend relations. There can also be implicit links between information sources that come from other.

It is vital for the fusion process and the ensuing decision-making to ensure that we have accurate estimates of the quality of various kinds of information. The quality of an information element has several components, for instance, the degree to which we trust the source and the accuracy of the information. Note that the source could be a high-level processing system itself: a fusion node that processed information from, e.g. sensors, and outputs a result. In this case, the quality determination must take account also of the way that the fusion node processed the data. In this chapter, we describe how social network analysis can help with these problems. First, a brief introduction to social network analysis is given. We then discuss the problem of quality assessment and how social network analysis measures could be used to provide quantitative estimates of the reliability of a source.

10.2 Network Models

10.2.1 Simple Networks

In order to properly describe a general network or graph, two things are needed. First, we need a list of the *nodes* of the graph. The nodes can be named and have various properties associated to them, but for describing the graph, it is enough that they can be enumerated from 0 to $N - 1$. Second, we must know which nodes are connected to which. This is most easily thought of as a list of *edges* (i, j) that are connected. Each edge can have various properties associated to it (e.g. a weight w_{ij}). Most often, the graph is described using the *neighbour matrix* or *contact matrix* A_{ij} , whose entries are non-zero if and only if nodes i and j are linked. It is sometimes convenient to consider the graph as a function $\phi(i)$ that gives a list of the neighbours of node i .

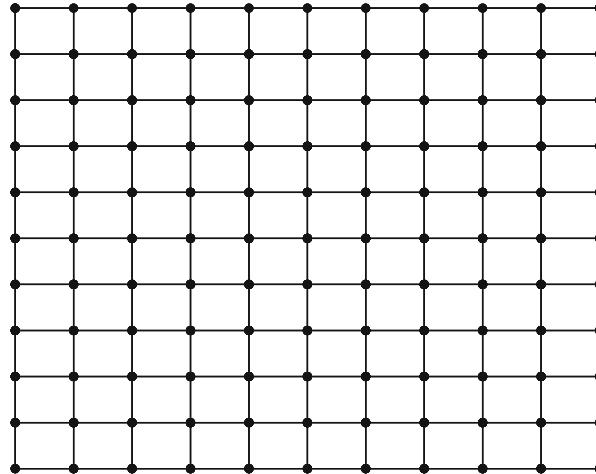


Fig. 10.1 A square lattice

Edges can be either directed (meaning, e.g. that an edge (i, j) can only transmit information from i to j) or undirected. A graph is connected if there is a chain of edges connecting any pair of nodes in it. A natural generalisation of graphs is to replace the edges by triples i, j, k or even n -tuples. Such structures are called hypergraphs. An important application of hypergraphs is to model sets of individuals that only interact in a group, never individually, for example, a group of rioters that meet at a certain place.

The simplest kinds of networks are regular, like the one shown in Fig. 10.1.

All regular lattices have some features in common. By looking at the graphs in Fig. 10.1, it is, for instance, apparent that these graphs are clustered, in the sense that if we remove one node, its neighbours will still have a short path between them. Another interesting characteristic of regular lattices is that the average distance between nodes is quite large. For a lattice with N sites in D dimensions,¹ it grows as $N^{1/D}$.

A natural extension of the regular lattice is to consider other graphs where all nodes are equivalent (i.e. have the same neighbourhood). The simplest example of such a graph is the complete graph with N nodes, K_N . This consists of N nodes where each node is connected to each of the other (so the graph has $\binom{N}{2}$ edges).

¹The simplest example to think of is \mathbb{Z}_l^D , where nodes are placed at integer coordinates and edges link nodes whose coordinates differ by ± 1 in exactly one dimension. Choose $l = N^{1/D}$ to get N nodes.

10.2.2 Classical Random Graphs

Traditionally, two different models of random graph processes have been used [2]. In the first model, $\mathcal{G}(N, p)$, each possible edge (i, j) is considered and included in the graph with a probability p . The other model, $\mathcal{G}(N, M)$, instead selects without replacement M of the $\binom{N}{2}$ possible edges. Note that these models are *not* completely equivalent. For the latter model, the graph is guaranteed to have exactly M edges, while the number of edges is a stochastic variable for the former. In the thermodynamic limit of $N, M \rightarrow \infty$, choosing

$$M = p \binom{N}{2}$$

gives graphs that should share all relevant properties. An important quantity characterising different random graphs is their *connectivity* or *average degree* γ , which measures the average number of neighbours that the nodes have. For random graphs with N nodes and M edges, this is given by $\gamma = 2\frac{M}{N} = p(N - 1)$ for the two ensembles.

Graph theory is a fascinating mathematical subject with many deep results; see, for instance, [2, 3]. One of the most interesting results is that there is a phase transition as the connectivity γ of a random graph grows. For small γ , the random graph consists of many isolated trees² of nodes. At $\gamma = 1$ this suddenly changes, and a giant component emerges. The size of the giant component scales linearly with the number of nodes, N . This percolating transition is somewhat surprising—note that the graph cannot be connected until it has a connectivity of at least $2(N - 1)/N$. Another important result is that the average path length between two nodes scales as $\log N$ for large N .

The random graph model, however, is not sufficient to describe many naturally occurring networks.

10.2.3 Small World Graphs

There are many different kinds of networks in nature. Perhaps the first that comes to mind is the social network of a society. Here each node represents a person, while there is an edge between two persons if they know each other. What does this graph look like? It is very unlikely that it would be a regular lattice—our acquaintances are not ordered in such a simple way. The social network however shares an important feature with regular lattices: they are clustered. Clustering means that there is a

²A tree is a connected graph without loops.

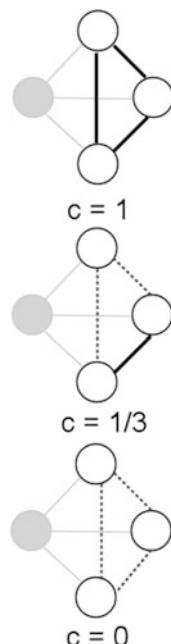
high probability that two neighbours of a given node also are direct neighbours themselves. An alternative way to think about it is to consider the average path length between two neighbours of a node i . Since both nodes are neighbours of i , this is obviously smaller than 2. If node i is now removed from the graph, we have to find a new shortest path between the nodes. If this new path length is still small, the graph is clustered. All regular lattices are obviously clustered, and social networks are clustered too: if person A knows persons B and C, there is a high probability that B and C will also know each other.

Real social networks are clustered in several ways: everybody's acquaintances can be divided into several distinct clusters, i.e. the people one knows from work all know each other, while the overlap between this group and one's neighbours often is zero. This can be modelled by allowing the network to have edges of different kinds or by superimposing several different networks on top of each other.

Mathematically, we can measure the degree of clustering in a graph by the *clustering coefficient*, C , defined as the average over all nodes of the local clustering coefficient C_i . For a given node i , consider its immediate neighbourhood, i.e. the set of nodes to which it is linked. The *local clustering coefficient* is now given by the fraction of all possible edges between nodes in the neighbourhood that actually appear in the graph. Figure 10.2 shows an example that should make the definition clear.

Another important feature of social networks is the so-called small world effect: when two strangers meet, it sometimes happens that the two people turn out to have mutual acquaintances.

Fig. 10.2 Examples of clustering coefficients. We wish to calculate the clustering coefficient of the grey node on the left, which has three neighbours (indicated by grey lines). The dark lines show the edges in the neighbourhood of the grey node that actually appear in the graph, while the dotted lines show all three possible such edges



The idea behind small world networks was first introduced by Milgram [4] in 1967. Milgram's experiment consisted of studying the path of letters addressed to a stockbroker in Pittsburgh. The letters were given to people in rural Nebraska with the rule that the current holder of the letter must hand it over to somebody with whom they were on a first-name basis. The average number of links in the chain of people between Nebraska and Pittsburgh was six, hence the term “six degrees of separation”. The number is of course not exact (a severe shortcoming of the experiment was that only one third of the letters were actually delivered!), but the phenomenon that people are linked via a small number of nodes has been verified by later, more careful experiments (e.g. [5]).

The small world effect has later been popularised by occurring in media, such as the movie “Six Degrees of Separation”. There are also various amusing games using the same concept, such as the web site <http://www.cs.virginia.edu/oracle/> where a user can find the distance between an arbitrary actor and Kevin Bacon. Actors here represent the nodes of the graph, and two actors are linked if they have participated in the same movie. It should be noted that the actors represented in the database are American and European ones. The network of actors in Indian movies, for instance, probably has few connections to this.

Another example is the Erdős numbers. Named after the famous mathematician Paul Erdős [6], these are defined recursively: Erdős has Erdős number 0; a person has Erdős number $n + 1$ if they have co-authored a paper with somebody who has Erdős number n (there are at least 507 persons with Erdős number 1; see the web site <http://www.oakland.edu/~grossman/erdoshp.html>).

Regular lattices do get shorter and shorter distances between nodes as the dimensionality increases (the diameter scales as $N^{1/d}$ for a d -dimensional lattice with N nodes, so it decreases if we increase d and keep N constant), but this is still too large to explain the small world effect. Instead, new graph models are needed.

A small world graph is intermediate between a regular lattice and a random graph—it has both clustering (like a regular lattice) and short maximum distances (like the random graph). It is constructed by considering in turn all the bonds (i, j) of a start graph (most often a regular lattice) and with some probability p replacing them with (i, k) , where k is a new, randomly chosen, node. So by changing the rewiring probability p , we can interpolate between the regular lattice and a random graph. An example of a small world obtained by rewiring the square lattice is shown in Fig. 10.3. Note that the small world for $p = 1$ differs slightly from a random graph, since all nodes are guaranteed to have a local connectivity of at least $\gamma/2$ where γ is the connectivity of the regular lattice. The distribution of connectivities is more broad for the small world with $p = 1$ than for the corresponding random graph.

The advance of the Internet and other communications networks has highlighted the need to be able to not only describe but also design networks that communicate efficiently. Efficiently there are two distinct meanings—the obvious one that a message from A to B should be transmitted along the shortest possible path and also an equally important one that the network should be fail-safe. If a node suddenly

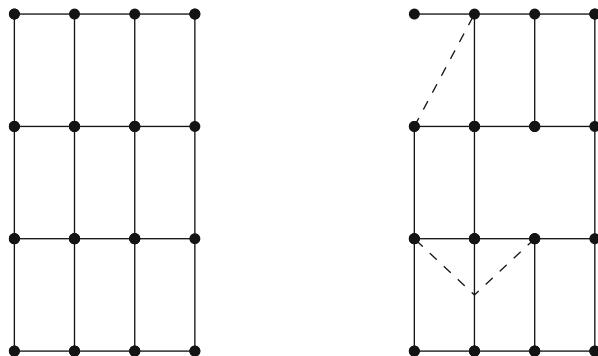


Fig. 10.3 This figure shows the construction of a small world starting from a 2D square lattice (left). In the right figure, two edges have been rewired and are shown as dashed lines

disappears, it should be possible to quickly find alternate paths between the rest of the nodes that don't involve the dead node. A very clear definition of small world behaviour in terms of *efficiency* has been given by Latora and Marchiori [7]. They measure the local efficiency as the time needed to communicate in the network, assuming unit velocity of signal propagation. The efficiency between two nodes is thus

$$\epsilon_{ij} = \frac{1}{d_{ij}} \quad (10.1)$$

where d_{ij} is the shortest distance between nodes i and j and $d_{ij} = \infty$ if there is no path between the nodes. The global efficiency is the average of this over all pairs of nodes in the graph. A high global efficiency corresponds to a small diameter of the graph. The local efficiency for a node i is calculated as an average of ϵ_{ik} over all neighbours k of i , and the total local efficiency of the graph is then the average of this over all nodes. The local efficiency is a measure of the fault tolerance of the network.

In addition to efficiency and clustering, there are a large number of measures that can be used to characterise a graph's properties. Many of these measures come from sociology and have been used to determine, e.g. the influence and power of individuals in different social networks. Others come from computer science or have been suggested by physicists.

A small world graph still has the same Poissonian distribution of node connectivities as random graphs. A different class of networks are the so-called scale-free graphs, which instead have a power law distribution.

10.3 Scale-Free and Growing Graphs

A network is called scale free if there is no characteristic length scale in it. In contrast to lattices, whose characteristic length scale is the lattice spacing, a scale-free graph divides its edges unequally among its nodes: the degree distribution follows a power law. This means that there are a few nodes (called hubs) that have very many edges, whereas most of the nodes have very few (Fig. 10.4). An important characteristic of scale-free networks is that while they are robust against accidental failures, they are very vulnerable to deliberate attacks against hubs.

A deterministic model for generating scale-free graphs has been introduced by Barabási and Ravasz [8]; this model generates the graph by iteratively replacing nodes with small graphs, in a manner similar to the construction of self-similar fractals.

There are many models of growing networks. In these models, one starts with a single node at time $t = 0$. In each new time step, a new node is added to the graph, and a new edge is created that connects this node to one of the older ones with a probability that depends on the connectivity of that node. If this probability is simply proportional to the node's connectivity (k), the model is reduced to the scale-free graph model of Barabasi and Albert [9]; see also [10]. It has been shown that the case where the probability is proportional to the connectivity is the *only* case which also leads to a power law for the distribution of connectivities in the entire graph [11]. If the probability is proportional to k^γ for any $\gamma \neq 1$, we get stretched exponential (if $\gamma < 1$) distributions or graphs where the majority of the edges share a common central node (for $\gamma > 1$).

The best example of a growing network is the Internet—each node that is added could be interpreted either as a new computer that is connected to it or a new web

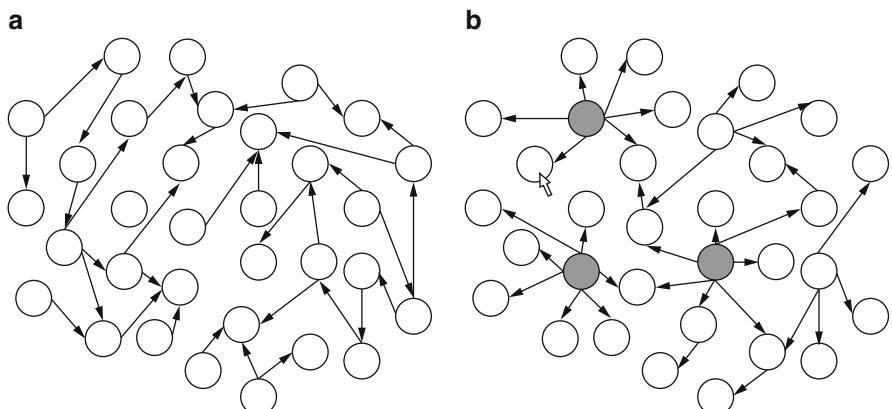


Fig. 10.4 This figure shows the difference between a random network and a scale-free network. Note the presence of hubs (nodes with many neighbours) in the right network. (a) Random network. (b) Scale-free network

site that is created. The edges created between this node and the older ones are then the hyperlinks that the addition of a new site entails. It turns out, however, that a more complicated model is needed to model the Internet; see below.

The so-called acquaintance networks, such as the Erdős graph, have been studied, among others, by Newman [12]. Such network has the characteristic that a small number of nodes have many edges. These nodes cannot be ignored when studying communication on such networks.

The crucial point of Newman's new model is that the probability distribution of the number of neighbours, which the neighbours of a specified node has, is not independent of that node. In social networks, a node with very few neighbours is likely to be linked to other nodes that also have few neighbours, while a node with many neighbours has similar nodes among its neighbours. Clustering is important for this calculation. If we know both the degree distribution³ and the clustering coefficient of a network, it is possible to calculate the number of neighbours at distance two from a given node. This is important to know when conducting research on social networks.

For example, we might be interested in how a network of terrorists grows in a country.

A very interesting approach to the problem of analysing growing networks has been introduced by Kleinberg and co-authors [13]. Perhaps the most prominent example of a growing network today is the Internet. The paper examines data representing routers on the Internet for a 2-year period and finds that the link density of the network increases with time, i.e. that the number of edges $e(t)$ is related to the number of nodes $n(t)$ by a power law:

$$e(t) \propto n(t)^a, \quad (10.2)$$

with $a = 1.18$.

Having such a relation between $e(t)$ and $n(t)$ means that the connectivity $\gamma = \frac{e}{n}$ is time-dependent, in sharp contrast to most models of graphs and networks. The authors find a similar relation (but with different a s) also for three different kinds of citation networks. In addition to the super linear scaling of edges with nodes, they also find that the average distance (also, somewhat non-standard, referred to as the effective diameter) between nodes *decreases* as a function of time. Recall that the average distance for a random graph grows as the logarithm of the number of nodes. While the two observations at first glance seem to be related, the authors note that it is possible to construct graphs that satisfy one of them but not the other. For large graphs, it is not practical to compute the diameter exactly. Instead, various approximate algorithms can be used. Since the shrinking diameter property is very surprising, the authors check that their result is robust by using several different such approximations to calculate the diameter. They also check for errors due to

³I.e. the probability that a node has a certain number of neighbours

the presence of a giant component; the shrinking diameter is present also if the calculation is restricted to just the giant component.

The paper presents several simple models that exhibit the densification property and also gives one model, the *forest fire model*, that possesses both it and also displays shrinking effective diameters.

The forest fire starts from a graph with just one node and then adds one additional node at each time step. At time t , let G_t be the current graph and v the added node. A node $w \in G_t$ is now selected randomly, and an edge $v \rightarrow w$ is formed. Most graph models would now continue by selecting another w and possibly adding an edge to it. In contrast, the forest fire model selects a random number of the nodes in G_t that were linked to w and adds edges from v to these nodes. This process is then repeated recursively for each of those nodes. (The process terminates if it reaches a node that has already been encountered. It is also possible to distinguish between out- and in-links when selecting the neighbours of w ; see the paper [13] for details.) Intuitively, the graph is generated in a similar way as friendships are formed: a newcomer finds one friend and with a certain probability becomes friends also with the friend's friends and so on. The name forest fire comes from a certain similarity to lattice cellular automata models used for studying forest fires. A natural extension of the model is to select several starting points w at each time step.

10.4 Using Social Network Analysis to Quantify Quality

Social network analysis [14–16] was introduced by sociologists as a means of analysing communications and relations in groups. It is a quite mature area of research, which has produced a large set of different measures to be used when analysing a network. Here we can only discuss some of the various measures that can be used to quantify the reliability of a network node. A more extensive list can be found in [17].

The simplest measures that can be applied to a network simply measure the number of connections that each node has. Another approach is to determine the minimum distance to other nodes in the network, i.e. determining how central the node is in the network. For some networks (such as the scale-free networks), this can give a reasonably good measure of the reliability of a node. However, for many networks, it gives quite misleading results.

A class of more advanced measures instead look at the flow in the network. In some networks, all links are associated with a maximum capacity that can be transported along it. This is the case, for example, for communications networks—the bandwidth imposes limits on the amount of data that can be sent along a link. For other networks, the flow algorithms assume that the capacity of each link is equal to 1.

The simplest flow-oriented measure simply determines the shortest paths between all nodes in the network. A node's reliability is then given by the number of such shortest paths that pass it. This measure is called the betweenness centrality measure.

Betweenness centrality, however, can also give misleading results. By focusing on only the shortest path, betweenness misses many cases where there are several short paths between nodes. An improved measure is the max flow centrality measure, which determines all the possibly paths between all the nodes in the network. Each node is then ranked according to the total amount of flow that passes through it.

Yet another way of characterising a social network is to look at the community structure in it. A community is loosely defined as a part of the network whose nodes have more connections within themselves than to nodes that are outside it. It is related to the concept of a clique, a maximally connected subgraph, but differs since it does not require full connectivity. (It must be mentioned that the exact definition of a community is of course application-dependent.) Several fast algorithms for determining the community structure of a network has been published [18–22]. Such algorithms could be used when analysing, for example, networks of data sources in order to determine which of them are independent from each other—and hence which should and should not be trusted.

Reliability and trustworthiness of a source can also be determined by computing the *social position* of it in the network; see [23] and [24] for a discussion.

10.5 Discussion

Determining the reliability of a source can be helped by determining the network structure of which the source is a part. This network can come either from the computation network in which it is included (see [25] for an example of how such networks can be modelled) or, if the source is, for instance, a social media component, from the communication network. We have given an introduction to some different network models and briefly described some social network analysis measure that could be used to determine the reliability of a source.

References

1. D.L. Hall, J. Llinas (eds.), *Handbook of Multisensor Data Fusion* (CRC Press, Boca Raton, 2001)
2. B. Bollobás, *Random Graphs* (Academic, New York, 1985)
3. B. Bollobás, *Graph Theory: An Introductory Course* (Springer, New York, 1979)
4. S. Milgram, The small world problem. *Psychol. Today* **2**, 60 (1967)
5. C. Korte, S. Milgram, Acquaintance linking between white and negro populations: application of the small world problem. *J. Pers. Soc. Psychol.* **15**, 101 (1970)

6. P. Hoffman, *The Man Who Loved Only Numbers* (Hyperion, New York, 1998)
7. V. Latora, M. Marchiori, Efficient behaviour of small-world networks. *Phys. Rev. Lett.* **87**, 198701 (2001)
8. A.-L. Barabási, E. Ravasz, Deterministic scale-free networks. eprint cond-mat/0107419
9. A. L. Barabási, R. Albert, Emergence of scaling in random networks. *Science* **286**, 509 (1999)
10. R. Albert, A.-L. Barabási, Statistical mechanics of complex networks. eprint cond-mat/0106096
11. P.L. Krapivsky, S. Redner, F. Leyvraz, Connectivity of growing random networks. *Phys. Rev. Lett.* **85**(21), 4629 (2000)
12. M.E.J. Newman, Ego-centered networks and the ripple effect. eprint cond-mat/0111070
13. J. Leskovec, J. Kleinberg, C. Faloutsos, Graphs over time: densification laws, shrinking diameters and possible explanations, in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago (2005), p. 177
14. J.P. Scott, *Social Network Analysis: A Handbook*, 2nd edn. (SAGE Publications, London, 2000)
15. S. Wasserman, K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, 1994)
16. P. Svenson, Complex networks and social network analysis in information fusion, in *2006 9th International Conference on Information Fusion*, Florence (2006), pp. 1–7
17. L. da F. Costa, F.A. Rodrigues, G. Travieso, P.R. Villas Boas, Characterization of complex networks: a survey of measurements (2005). eprint cond-mat/0505185. <http://www.arxiv.org/abs/cond-mat/0505185>
18. M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113 (2004)
19. M.E.J. Newman, Fast algorithm for detecting community structure in networks. *Phys. Rev. E* **69**, 066133 (2004)
20. A. Clauset, Finding local community structure in networks. *Phys. Rev. W* **72**, 026132 (2005)
21. A. Clauset, M.E.J. Newman, C. Moore, Finding community structure in very large networks. *Phys. Rev. E* **70**, 066111 (2004)
22. J. Dahlin, P. Svenson, Ensemble approaches for improving community detection (2013). Preprint arxiv:1309.0242
23. J. Brynielsson, J. Hogberg, L. Kaati, C. Mårtenson, P. Svenson, Detecting social positions using simulation, in *Proceedings 2010 International Conference on Advances in Social Networks Analysis and Mining*, Odense (2010)
24. J. Brynielsson, L. Kaati, P. Svenson, Social positions and simulation relations. *Soc. Netw. Anal. Min.* **2**(1), 39–52 (2012)
25. M. Jändel, P. Svenson, R. Johansson, Fusing restricted information, in *Proceedings of the 17th International Conference on Information Fusion*, Salamanca (2014)

Chapter 11

Quality, Context, and Information Fusion



Galina L. Rogova and Lauro Snidaro

Abstract Context has received significant attention in recent years within the information fusion community as it can bring several advantages to fusion processes by allowing for refining estimates, explaining observations, constraining computations, and thereby improving the quality of inferences. At the same time, context utilization involves concerns about the quality of the contextual information and its relationship with the quality of information obtained from observations and estimations that may be of low fidelity, contradictory, or redundant. Knowledge of the quality of this information and its effect on the quality of context characterization can improve contextual knowledge. At the same time, knowledge about a current context can improve the quality of observations and fusion results. This chapter discusses the issues associated with context exploitation in information fusion, understanding and evaluating information quality in context and formal context representation, as well as the interrelationships among context, context quality, and quality of information. The chapter also presents examples of utilization of context and information quality in fusion applications.

Keywords Context and information fusion · Context quality · Information quality in context · Context discovery

This chapter is an extended and revised version of [1].

G. L. Rogova (✉)
The State University of New York at Buffalo, Buffalo, NY, USA
e-mail: rogova@buffalo.edu

L. Snidaro
Department of Mathematics, Computer Science, and Physics, University of Udine, Udine, Italy

11.1 Introduction

The beginning of the 1990s marks the start of a significant interest in context by researchers in computer science, especially in view of widespread adoption of context-aware smart mobile devices. An area that has lately shown a rapidly escalating interest in context is information fusion (IF) (see, e.g., [1–4]). IF is a formal framework to exploit heterogeneous data and information originated from different sources as well as a priori models to obtain information of better quality and provide a better understanding of the phenomenon under consideration. Context can bring several advantages in achieving these IF goals, such as refining estimates, explaining observations, and constraining processing. In addition, context allows for improving the associability between problem-space knowledge and models and observational data, increasing fusion performance. The a priori models can better fit data exploiting the semantics provided by context [3].

Therefore, as discussed in [5], an integrated approach to context exploitation should extend problem modeling with representations of domain knowledge, which describe situations and possible entities' interactions and relations in the observed scenario. The role of context can be very relevant to the following considerations:

- Context can be used both to transform source data into information and knowledge [6, 7] and to acquire knowledge [8, 9].
- Context may provide information about the conditions of data and information¹ acquisition, and it may constrain and influence the reasoning about objects and situations of interest.
- Context is an important source of semantics, providing means to bind data and models, therefore increasing the capability of inferences performed at higher levels.
- Context can be used to adjust the parameters of the algorithms (e.g., tracking, event detection, etc.), thus making this information vital to foster the adaptivity of the system.
- Different types of context from coarse to refined, can be injected at different stages of the fusion process.
- Context is the only source that can resolve ambiguities in natural language (phonetic, lexical, syntactic, semantic, pragmatic).

In addition, the development of context-based fusion systems is an opportunity to improve the quality of the fused output and provide domain-adapted solutions. At the same time, information defining a context can be obtained from available non-compatible databases, observations, the result of sensor fusion, intelligence reports, mining of available information sources (e.g., traditional and social media), or as

¹While we recognize the difference between data and information, we will generally use these terms interchangeably throughout this chapter.

a result of various levels of information fusion processes. Of course, the quality of any such information, as well as the inference processes for obtaining it, may be insufficient for a particular use: it might be uncertain, unreliable, irrelevant, or conflicting.

Knowledge of the quality of observations and other sources of information used for inferencing and its effect on the overall quality of context characterization can improve contextual knowledge. There are two interrelated problems concerning both information and context quality: imperfect information used in context estimation and discovery negatively impacts context quality while imperfect context characterization adversely affects the characterization of quality of information used in the fusion processes, as well as the fusion results. This interrelationship represents one of the challenges of modeling and evaluating context quality and of using context in defining the quality of information used in fusion and the quality of fusion process results. Solving this and other problems related to effective context exploitation in fusion requires consideration of:

- Models of information quality and context
- Interrelationships among quality of context and quality of information
- Context dynamics and the role of information quality in context discovery
- Method of controlling the quality of context

The remainder of this chapter is devoted to discussions of these issues.

11.2 Context in Information Fusion

Only in recent years, it has been recognized in the fusion community that context represents an important source of information for the fusion process [3] and its full potential is thus far from being tapped. In order to use context effectively, we must understand both what context is and how it can be used [5]. Context has many facets that sometimes lead to defining it based on certain narrow characteristics of the specific problem being addressed. For example, in [9], context is defined as objects, location, and identities of nearby people and objects. In other works [10], it is considered as a computable description of the terrain elements, the external resources, and the possible inferences that are essential to support the fusion process; while in [11], context is represented by the operational knowledge.

A definition that does allow for better understanding of context – and therefore one that is more appropriate for formalizing and utilizing context in building IF processes – was introduced in [12] and further considered in [4, 13]. It assumes two paradigms: “context-of” (CO) and “context-for” (CF):

- CO: We can have certain expectations based on situations; e.g., “in the *context of* volcano eruptions, we would expect volcanic ash plume causing air traffic disruption”;

- CF: Alternatively, we can assess items of interest – whether individual entities or situations – in context: “the weather provides a *context for* understanding the volcanic eruption plume direction and height”.

Therefore, context characterizes a relevant situation, i.e. a situation that can be used to provide information either (a) to condition expectations (CO) or (b) to improve the understanding of a given inference or management problem (CF). A relevant situation can be characterized by “context variables” that are used for evaluation of “problem variables,” a set of variables of concern in the given problem. This definition immediately points out to the relationship between context and information quality (IQ) since it requires the understanding and methods of evaluation of relevance, one of the important IQ characteristics (see, e.g., [14–16]). A context variable can be called relevant if the values of problem variables, decision, or action under consideration change with the value of the context variable. If the problem variables are objects, object attributes, or relations, relevance can be defined in terms of mutual information of the problem and the candidate context characteristic. If the problem variables are situational items, we can call a contextual characteristic relevant if a change of its value affects the uncertainty distribution of hypotheses about these situational items and, therefore, decisions and actions. Relevance of a context variable can be also defined by decision makers based on their information needs.

Consideration of CO and CF provides for complex hierarchical relationships among characteristics of problem variables and context variables. It also offers a clear understanding of relationships between context and situations. Reasoning about entities and relationships while considering them as problem variables within a certain context corresponds to reasoning about situations. Such reasoning produces an answer to the question, “what is going on in the part of the environment corresponding to the reference item(s) within a specific context.” Therefore, we can define *context* as a meta-situation (situation of a higher level of granularity), comprising a set of relationships involving context variables: $C = (PV_i, CV_i, R_i)$ where PV_i and CV_i are problem and context variables respectively and R_i are relationships between various problem variables, various context variables, and problem and context variables. Modeling context for decision making is generally reduced to the following problem [12]: “Given an entity of interest (a physical object, a situational item, and an event) what context or a sequence of contexts can be formed, such that a task about this entity can be accomplished.” Since the values of problem and context variables can be of variable quality, it is necessary to incorporate their quality into context modeling. Context variables can serve as problem variables when they represent reference items for a different context. It is clear that the quality of estimation of context variables directly affects the quality of estimation of problem variables and vice versa. Figure 11.1 shows some important relationships between context and fusion processes.

The context engine here interacts with and supports fusion at all levels by [17]:

- Representing an initial overall context under considerations
- Establishing relevance, thereby constraining ontology of the domain, observations, rules, and statistics

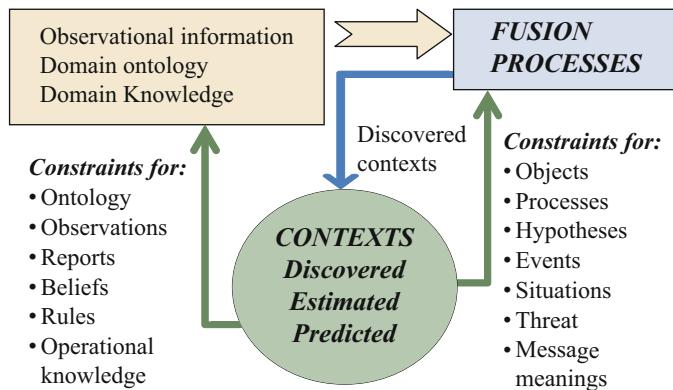


Fig. 11.1 Interaction between context and information fusion processes

- Providing the fusion processes with constraints on relationships among objects, situations, hypotheses, arguments, beliefs, and preferences
- Supporting situation and threat discovery
- Constraining the feasible meanings of messages, thereby facilitating effective communications among actors
- Improving the quality of transient incoming information and thereby fusion results

11.3 Quality of Information and Context

11.3.1 Objective and Subjective Information Quality

Quality of information represents “information about information.” The need for considering quality of incoming data and information as well as the results of the fusion processes at all fusion levels stems from the fact that the information models and fusion techniques have their limitations. These limitations come from imperfect domain knowledge, the difficulty of finding adequate information models and their parameters, the lack or insufficient amount of additional information that may be required for quality assessment, the subjectivity of quality evaluation performed by experts and end users, as well as insufficient quality of context and even consideration of a wrong context.

There are multiple definitions of IQ such as “Quality is the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs” [18], “fitness for use” [19], etc. It can be seen from these definitions that quality is measured in terms of potential and actual benefits to the users that can be humans or automatic processes. The assessment of the “fitness for use” is based on the characteristics of information corresponding to inherent properties of information

(meta-data) and their values. The inherent information characteristics constitute objective quality. Meta-data is represented and measured by its attributes and their combination since without clearly defined attributes and their relationships, we are not just unable to assess IQ; we may be unaware of the problem [19].

IQ characterizing “fitness of use” will be defined here as *subjective* quality. *Subjective quality* is the level of satisfaction of users and process designers in relation to their specific goals, objective, and functions in a specific context. Meta-data considered independently from information users and context gives value to subjective quality. The introduction of two different types of quality is similar to one introduced for sensor data in [20] where subjective IQ is referred to as *information quality* and objective one as *volume* of information. At the same time, the term *subjective quality* is more appropriate for dealing with mix initiative fusion systems that assume collaboration and information exchange between automatic processes and humans. Meta-data provide data consumers with a basis for deciding whether and to which degree the data under consideration fits their needs and can be utilized. Values of meta-data can result from models and information processing, learning, source dynamics, measurements. Subjective quality is provided by context and goals, objectives, and function as well as personal traits of decision makers (Fig. 11.2).

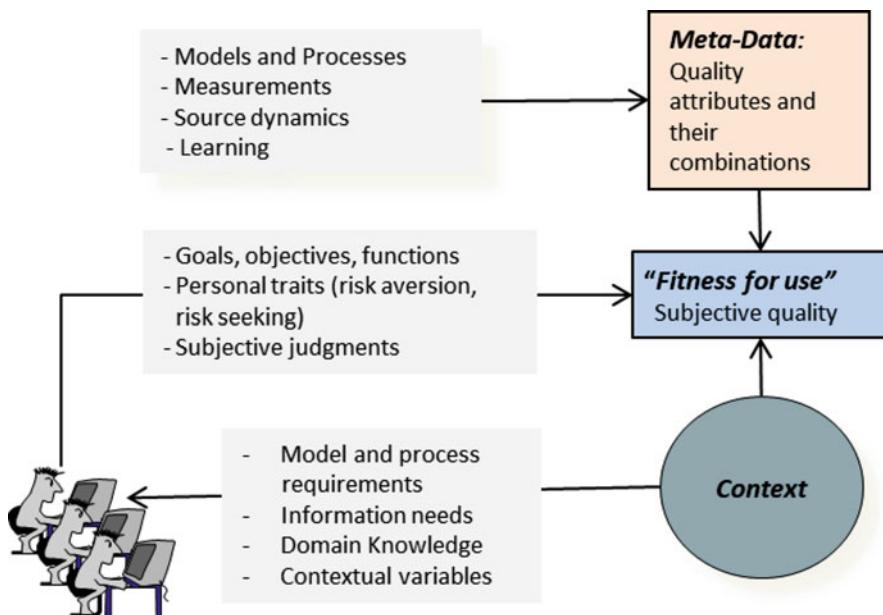


Fig. 11.2 Meta-data and subjective quality

11.3.2 *Quality Characteristics*

There are multiple quality characteristics considered in the literature (see, e.g., [14, 20, 21–25]). An ontology of IQ characteristics is presented in [14, 24, 25] where three major interrelated categories of quality are considered and further decomposed:

- Quality of information source
- Quality of information content
- Quality of information presentation

There are two types of information sources: objective and subjective. While objective sources are represented by physical sensors and models, subjective sources are sources that produce human-generated information. Thus subjective sources include newspapers, TV, the Internet, members of social networks, and opportunistic sources. The quality of objective sources includes such characteristics as reliability, relevance, and credibility. The quality of subjective sources comprises objectivity, intent, level of expertise reputation, reliability and confidence. Nowadays in view of widespread adoption of smart mobile devices when every person may be a sensor, designing methods of evaluation of the quality of subjective, human-originated information is becoming especially important. Context considerations for their evaluation are one of the main sources of the estimation of their values.

Quality of information presentation is related to the problem of when, how, and which information to represent. It includes understandability, completeness, interpretability, timeliness, relevance, believability, and trust. The majority of context-dependent quality characteristics are subjective, and their values are defined by the objectives, functions, mental models and beliefs, cognitive biases, and variable expertise of the human-in-the loop. Context can significantly improve the quality of information presentation, such as understandability and interpretability, since it provides decision makers with expectations of the meaning of the information presented.

Quality of information content is decomposed in timeliness, relevance, integrity, and importance. Integrity includes uncertainty such as credibility, probability, reliability; and imprecision such as accuracy, consistency, conflict, completeness, fuzziness, and vagueness.

The overall IQ may relate to a single quality attribute or a combination of several or all the attributes. The decision of which attributes to consider or combine depends on context and the problem at hand. For example, in an object recognition problem, an overall objective quality of the recognition result can be a combination of beliefs in a hypothesis and the reliability of the source. The value of subjective quality of this result corresponds to the user satisfaction with the objective quality values and can be evaluated, for example, by comparison of the comparing subjective quality with a context-specific threshold. Combining multiple quality attributes can be performed by employing various methods. For example, attributes selected for combination can be considered as an input into a neural network with a binary output

“1” for good and “0” for bad quality or with a vector of linguistic values such as very good, good, bad, very bad. This neural network is trained by one or multiple experts who consider specific objective and functions in a context under consideration. The overall quality measure can be also a weighted sum of the quality attributes under consideration, in which weights represent the importance of attributes, a subjective quality measure. The same value of the most characteristics mentioned above can represent both meta-data when it is considered by itself and subjective quality when considered in relation to uses objective in a specific context. For example, timeliness can be either a number between the actual and expected arrival time of the information or measure of the usefulness of this information for the user’s decision. Another example is uncertainty of a hypothesis, which represents meta-data when its value is obtained as the result of modeling, but it becomes subjective quality when an agent considered this value sufficient for making a decision based on it.

At the same time, certain quality characteristics cannot be considered outside context and always represent subjective quality. Main attributes that represent subjective quality only include accessibility, trust, importance, understandability, and relevance. Some of these attributes are particularly significant for IF applications and will be discussed in the following.

11.3.2.1 Accessibility

Accessibility, which can be measured by the cost of obtaining information, is a binary characteristic depending on context. It also depends on another subjective quality characteristic – importance. Depending on how important to the objectives and functions of the decision makers in the context under consideration, they are willing or not to accept the cost of obtaining a piece of information, which gives the value of the accessibility 0 or 1.

11.3.2.2 Trust

Another important subjective/contextual quality attribute is trust. There are multiple definitions of trust, and there is no consensus among theorists on how to define trust. At the same time, most approaches rely on some version of the concept proposed in [26], in which trust is defined as “a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another.” As it is stated in [27], “trust must be viewed as a layered notion in its basic meaning of subjective trust,” and trust is “a belief, attitude, or expectation concerning the likelihood that the actions or outcomes of another individual, group or organization will be acceptable or will serve the actor’s interests” [28]. Having this in mind, we define trust as a *subjective* level of belief of an agent (either human or computational) that information he is using can be admitted in the system. Utilization of the notion of trust and its management “aims at supporting

decision making in the presence of unknown, uncontrollable and possibly harmful entities” [29]. Trust in information can be defined by a user as a combination of belief in the context in which the information is produced, reliability, believability, and reputation of the source (physical, human sensors, or processing results), and information presentation, which then alone or in combination of other meta-data characteristics, will serve for establishing the level of trust [30].

11.3.2.3 Reliability

The notion of trust is directly connected to the notion of reliability since unreliable information cannot be trusted. Similar to trust, reliability is a subjective quality. Reliability of hard data is defined by applicability of a sensor producing the data in a specific context obtained from domain knowledge; statistical information corresponding to sensor performance and applicability of the sensor model in the context under consideration. Defining reliability of soft information is more complicated. For example, sources of soft information can be unreliable if they do not have incentives to tell the truth or enough knowledge about the context, in which observations are made. Another problem is that soft information is rarely characterized by direct reliability since in many cases it comes from a network of agents with variable and often unknown reliability, for example, from social networks. In order to assume that sources of soft information are reliable, it is important to take into account their characteristics (education, experience, prior and tacit knowledge, history of judgments) and understanding of context.

11.3.2.4 Relevance

One of the central subjective quality attributes is relevance, which can characterize quality of information content, information source, and information presentation, and strongly depends on a specific context as well as goals, functions, and expectations of decision makers. As it was mentioned above, in context-aware IF, relevance also plays an important role in selecting context variables for estimating problem variables. According to the definition of the Merriam-Webster dictionary, information is relevant if it has “significant and demonstrable bearing on the matter at hand.” Therefore, relevance is not a property but “is understood as a relation; relevance is a tuple – a notion consisting of a number of parts that have a relation based on some property or criterion” [31]. Formally, a tuple representing relevance is $(\{P_i\}, R, \{Q_j\}, S, C)$, where $\{P_i\}$ and $\{Q_j\}$ are sets of both tangible and intangible objects, R is a criterion defining relevance of these sets (e.g., utility), and S is a measure of the relevance strength. If $S = 0$, $\{P_i\}$ and $\{Q_j\}$ are not related, if $S = 1$ they are completely related. In the uncertain environment, relevance is not binary and $S \in [0, 1]$. Following Walton [32] we can consider contextual variables relevant if they “having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than

it would be without” these variables. Another criterion for considering a particular context variable is the increase in information or achieving a higher action utility as the result of using that variable for estimation and/or inferencing. Action utility includes consideration of benefits of a particular action and its cost that may involve costs of data collection, communication, and processing, as well as the cost of lost opportunity.

There are several context-dependent questions to be answered before a piece of information can enter a system:

- Is it relevant to the task or purpose of the processes? To which extent?
- Is the level of relevance enough to justify the use of this information?
- How reliable or trustworthy is the source of the information?
- Is the information timely enough?

There is a need to distinguish informational irrelevance from causal one [33]. Informational irrelevance means that “ X is independent of Y given context C .” Causal irrelevance means that “ X is causally irrelevant to Y in context C .”

The problem of selecting relevant context variables is complicated by the fact that relevance is often time-dependent: information can be relevant at a certain time but becomes irrelevant later and vice versa. For example, relevance of the information content depends on its timeliness, e.g., information arriving too late is irrelevant. Hence, relevance has to be evaluated along with other characteristics of IQ. Since there may be multiple analysts/automatic process sub functions, relevance of information needs to be evaluated at each step of agent interaction and according to their function requirements and context. For example, since threat is characterized as an integrated whole of threat, opportunity, and capability, relevance of incoming information has to be evaluated separately for each threat component.

11.3.3 *Quality of Context*

The quality of information considered for reasoning about problem variables or evaluation of the effectiveness and efficiency of the results of such reasoning strongly depends not only on context itself but also on the quality of context. *Quality of context* is defined in [34] as “any information describing the quality of information that is used as context information.” Another definition is given in [35], where *quality of context* is defined as “any inherent information that describes context information and can be used to determine the worth of information for a specific application.” Therefore, alike the definitions specifying IQ, these definitions specify different types of context quality with the former referring to objective measures of quality, while the latter characterizes both objective and subjective quality, which uses values of objective quality to measure the “fitness of use,” i.e., the degree to which context satisfies the needs of a particular application.

Similar to [36], here we define the quality of a context as the degree to which it satisfies the needs of an application, expressed as a function of quality of

the context variables defining the context. Similar to the subjective quality of information that has to satisfy users' needs in a particular context, the degree to which context satisfies the needs of a particular application can be represented either by a single quality characteristic or by a combination of characteristics. Selection of a characteristic used for evaluating the quality of context variables, or their combination, depends on the application.

The information defining a context is obtained from available databases, observations, represents the result of sensor fusion, received reports, mining of available information sources (e.g., traditional and social media), or from various levels of IF. Of course, the quality of any such information, as well as the inference process for obtaining it, could be insufficient for a particular use: it might be uncertain, unreliable, irrelevant, or conflicting. Knowledge of the quality of this information and its effect on the quality of context characterization can improve contextual knowledge and allows for discovery of a new context. At the same time, knowledge about a current context can improve the quality of observation and fusion results.

Figure 11.3 shows the interrelationships among the quality of incoming "soft" information, which is usually coming from human sensors and is expressed in natural language, and "hard" numerical data obtained from traditional "physical" sensors and fusion processes; quality of context and important quality characteristics influencing them. As shown in Fig. 11.3, fusion processes designed to estimate problem variables can use direct estimations and selected contextual information weighted for accuracy, reliability, consistency, and relevance. Indices are shown for contexts, both CO and CF, to stress that relevant contexts are often dynamic. Fusion outputs can be estimates at any fusion level (i.e. of objects, situations, impacts, etc.). Context consideration can improve the results of fusion products by taking into account the quality of input information (e.g., reliability of observations and reports) as well as the quality of interim results of the processes involved in fusion. For instance, a CO can serve for selecting relevant observations and provide expectations for sensor and process management. A CF can, for example, be used to improve fusion results by incorporating context-based reliability into sources' predictive uncertainty models such as probability or belief.

Selection of variables characterizing CF can be based on constraining domain-specific ontology of context variables and their relationships with problem variables to exploit the relevance of context variables to problem variables and their consistency. CF has to be relevant to both CO and problem variables. It is considered relevant if [13]:

- Reduces uncertainty and increases the accuracy of fusion results
- Improves the utility of information (e.g., of refining the value of a problem variable) and ultimately of decisions and actions based on fusion results utilizing this information
- Decreases information conflict

Selecting of context variables assumes that we can determine the ambient CO for given problem variables or inference problem. In some cases, this context is defined (declared or estimated). However, in other cases, context can be unknown

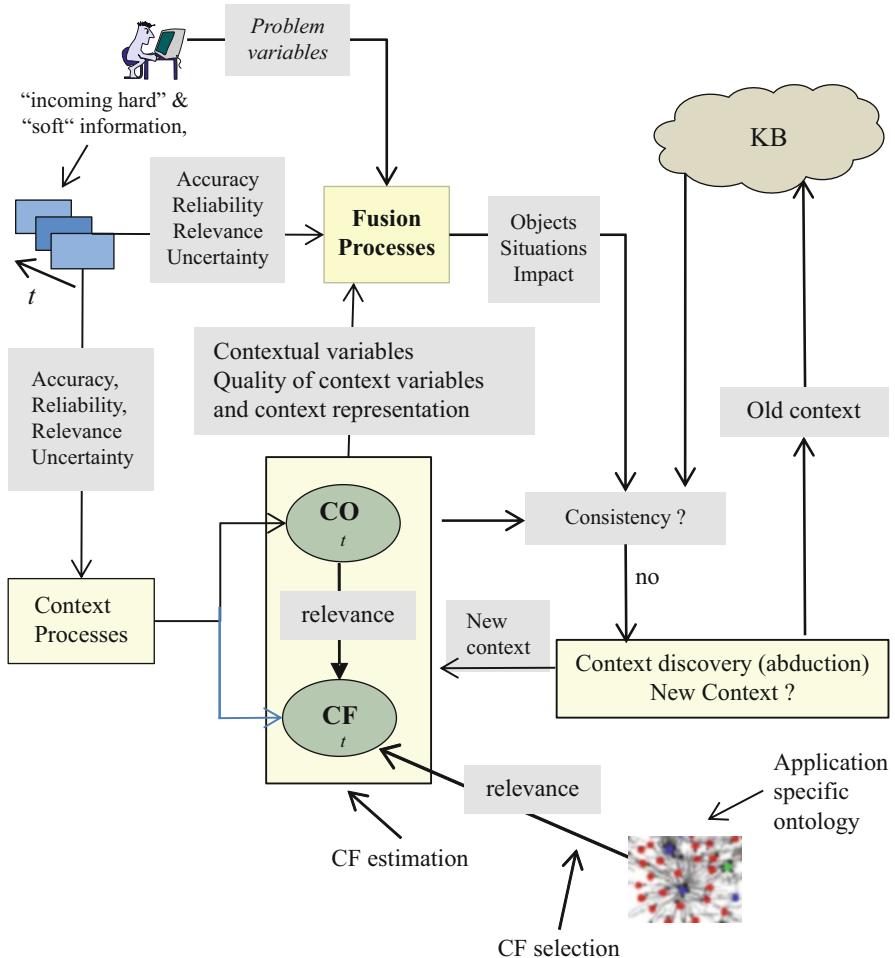


Fig. 11.3 Interrelationships among the quality of observations and reports, fusion processes, quality of context, CO and CF, and important quality characteristics influencing them

or different from what was expected since the expectation can be based on mistaken assumptions. Thus the set of beliefs characterizing current context have to be changed. This often happens in highly dynamic environments, in which observations, situational items, and relationships constantly change, and therefore, relevant context needs to be discovered. Discovery of underlying new context can be initiated based on another important characteristic of context quality: *consistency*. New context can be manifested by new observations or fusion results that are inconsistent with the currently assumed context characteristics, for example, contained in the knowledge base. The major problems here are:

- how and when to decide whether inconsistency exists;
- what is the source of this inconsistency; and
- whether the currently assumed context is no longer relevant.

Context consistency is evaluated based on the comparison of the characteristics and behavior of problem variables based on the observed or estimated data and information with the ones that are defined by contextual knowledge, which includes both CO and CF. Inconsistency can be the result of such factors as poorly characterized observations and reports, characteristics and behavior of problem variables based on these observations and evaluated within the current context, domain knowledge about current context, insufficient quality of context characteristics; e.g., consideration of irrelevant or unreliable context variables, or the fact that the earlier defined context has changed.

One unavoidable problem related to the quality of observations is the problem of time delay affecting the timeliness of decisions. The time delay is a combination of time required for communication and information processing. To mitigate this problem, it might be necessary to consider projected time characteristics and behavior of situational items and characteristics of CF. At the same time, incorporation of the time delay into estimation of reference and context variables represents one of the challenges of context exploitation.

Discovery of the source of this inconsistency can be performed by abductive reasoning (so-called reasoning for best explanations). This abductive process of reasoning from effect to cause requires [37, 38]:

- Constructing or postulating possible hypotheses explaining inconsistency
- Computing plausibility of these hypotheses
- Selecting the most plausible hypothesis from among these

The result of abductive reasoning can improve inferencing in different ways. It can lead to the discovery of new, sometime hidden context, which in turn can improve the estimation of problem variables. It can also lead to the discovery of the fact that the inconsistency was the result of poor quality of observations, reports, their processing, or insufficient quality of certain context characteristics, which were estimated or held in the knowledge base. This discovery can require reevaluation of this information.

11.3.4 *Context Representation and Information Quality*

The method of evaluating context quality depends on the context representation selected. There are several context models considered in the literature [39]. There are the following ones that appear to be appropriate for IF such as the *key-value* [40], *ontology-based* [11, 41, 42], and *logic-based* [43–46].

The simplest models are the *key-value models* [40], in which context is represented by values of context variables (e.g., location) under consideration obtained

by measurements or as the results of the matching processes and/or provided by humans. The quality of context modeled by *key-value models* depends on the quality of context attributes considered. To incorporate context quality into this model, each variable can be represented as a tuple comprising the value of the context variable under consideration and the values of quality characteristics pertinent to this variable. For example, if we consider location coordinates as a context variable, the tuple might also include the accuracy of the coordinate estimation and the level of relevance of location coordinates to the problem variables under consideration. The overall context quality will be represented by the combination of all the quality value estimated for each characteristic. The key-value models are easy to manage and may suffice for CF representation in low-level fusion [47] but lack capabilities for complex structuring necessary for representing CO, which involves representation of not only context attributes but also objects, their characteristics, and interrelationships. More general models of context are similar to the ones used for situation assessment and include *ontology-based* and *logic-based* models.

Ontology is an established framework for knowledge representation and for reasoning about situations. Since contexts are considered as meta-situations, ontology-based models offer an appropriate way of their modeling. These models provide a high degree of rigor in specifying core concepts, sub-concepts, facts, and their inter-relationships to enable realistic representation of contextual knowledge. An ontology of a specific context requires a relevance-based constraining domain ontology of the context under consideration. This can be done by considering context variables, relationships between them, and inference rules characterizing this particular context while incorporating quality of these variables and their relationships. In a dynamic environment, contextual information and its quality can rapidly change, and for that reason, a context ontology requires constant instantiation. Methods of measuring the overall quality of context represented by ontology are similar to methods of ontology evaluation considered in the literature (see, e.g., [48, 49]). Evaluation measures of ontology quality proposed in the literature and appropriated for context quality evaluation allow for assessing syntactic and semantic aspects [48]. Among the most important measures of the quality of context are completeness, accuracy, reliability, expandability, consistence, and scalability.

Logic-based models define contexts in terms of facts, expressions, and rules. McCarthy [43] introduces a formalization of logic-based models built on a relation *ist*(*c,p*) that reads as “proposition *p* holds in the context *c*.” The *ist* concept can be interpreted as validity: *ist*(*c,p*) is true if and only if the proposition *p* is true in context *c*. McCarthy’s context formalization includes two main forms of expression:

- *c'*: *ist*(*c,p*) means that *p* is true in context *c*, and this itself is stated in a context of a higher level of granularity (“outer context”) *c'*;
- *value*(*c,e*) defines the value of term *e* in the context *c*, which means that context *c* defines some values for *e*.

The quality of context representation here is defined by the quality of the expressions mentioned above. Thus we need to incorporate quality into the model expressions. For example, to use $ist(c,p)$ for making assertions about uncertain situational items, $bel(a,ist(c,p))$, can be used to represent an agent a 's belief that proposition p is true in the context c . which can be obtained as the result of reasoning under uncertainty. Since the knowledge about context is also uncertain, the reasoning process has to take into account quality of context and problem variables.

Another context representation framework of this type is situation theory [45, 46]. Situation theory represents units of information as *infons*, which are denoted as $\sigma = (R, a_1, \dots, a_n, i)$, where R is an n -place relation and a_1, \dots, a_n are state variables that range over entities of types appropriate for a given relation R . In “classical” situation theory, i is a binary variable, which is equal to 1 if a relationship $R(a_1, \dots, a_n)$ holds, 0 otherwise. Context representation by situation theory requires incorporation of the quality of contextual variables and inferred relations between them. This can be achieved by replacing binary relations with a value between 0 and 1 defining belief that the relationship holds [48] and the value of context variables by tuples that comprise the variable values and meta-data characterizing them.

11.4 Context Quality Control

As we can see from the previous sections, context is important for improving the quality of problem variables. At the same time, since context is obtained from observational information and reasoning, it can be imperfect. Thus it is necessary to implement quality control of contextual variables.

Similar to the methods of quality control of problem variables discussed in [14, 25], methods of quality control of context variables include:

- Eliminating from consideration any context variables of insufficient quality;
- Explicitly incorporating estimation of the quality of context variables into models and fusion processes; and
- Delaying a decision concerning poor quality of information while hoping that quality will improve with additional observations and/or computations.

Thus, elimination of context variable of insufficient quality can be based, for example, on CO-specific time-dependent relevance or reliability of CF characteristics. This type of quality control is important for improving the quality of information presentation since irrelevant or unreliable context characteristics can decrease context understandability and lead to increased fatigue and distrust in presented information.

Explicitly incorporation of quality of context variables into reasoning about problem variables or CF estimation invokes considerations of the second level of uncertainty such as reliability coefficients. Utilization of reliability coefficients allows for resolving conflict, which is especially important for fusion of credibility

since by definition, combination of credibility represented either by probability or beliefs requires that probability (beliefs) will be equally reliable.

Delaying decisions on insufficient quality of observations defining context variables can be utilized when more observations can be available or when computation time increases. These quality control methods are exploited to support time-critical decisions and swift actions in the uncertain environment, in which it is not realistic to obtain optimal solutions. They allow monitoring the progress in problem solving and offering a tradeoff between time and other quality attributes defining decision quality.

Following [50], we can define desirable properties of these quality control methods:

- *Measurable quality*: The quality of an approximate result can be determined precisely.
- *Recognizable quality*: The quality of an approximate result can easily be determined at run time (i.e., within a constant time).
- *Monotonicity*: The quality of the result is a nondecreasing function of time and input quality. Since the quality of input is generally not a nondecreasing function of time, it is possible to guarantee monotonicity by simply taking into account the best rather than the last generated result.
- *Consistency*: The quality of the accumulated over time result is correlated with computation time and input quality.
- *Interruptibility*: The process can be stopped at any time and provide some answer.

In general, selection and implementation of a specific control measure or their combination depend on the application and quality attributes under consideration.

11.5 Context and Information Quality in Fusion Applications

11.5.1 City Traffic Tracking

To show the importance of IQ in context utilization, let us, for example, consider Fig. 11.4 depicting a traffic monitoring system with optical sensors. The sensors can be easily tricked into generating false targets by reflective surfaces.

As discussed in [51], knowledge of the map of the city constitutes context allowing the system to prune or discount the measurements that seem to be originating from inside a building. However, the lack of knowledge about the reflective surface and thus the incompleteness of the contextual information do not allow to explain the behavior of the system.

In the target tracking domain, knowledge of CI can be exploited, for example, to constrain measurements or choose an optimal motion evolution model (e.g., constant liner velocity, curvilinear, accelerated motion, etc.). It can be done by generally yielding higher estimation accuracy with increased confidence (reduced

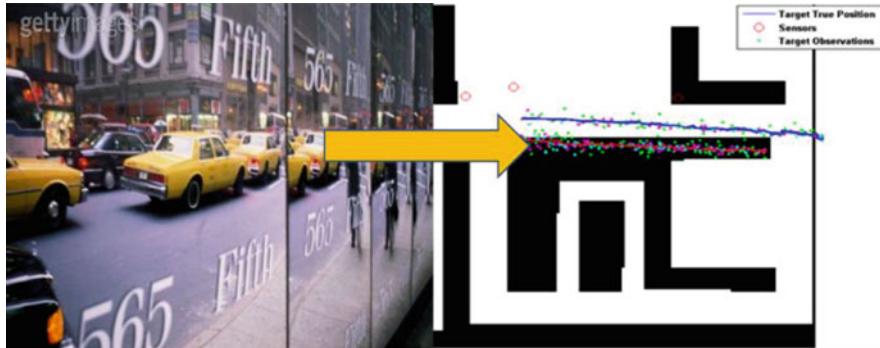


Fig. 11.4 Tracking of city traffic by video sensors. False detections are generated by a reflective surface and false tracks could be generated

error covariance). However, relying on imprecise or incorrect CI can negatively tune the system to react to conditions that do not reflect the reality of the environment. Thus insufficient context quality introduces biases and leads to the generation of erroneous estimates that can adversely affect the overall performance.

As per Fig. 11.3, the whole quality assessment process works in a continuous loop that leverages the availability of multiple data/information for reaching fused estimates of the problem variables while assessing the quality of the input and contextual information. For example, the false track in Fig. 11.4 could be the result of low-quality data and/or contextual information as well as processing errors:

- Noisy sensor data?
- Error-prone processing (e.g., calibration errors/wrong filtering model)?
- Imprecise/outdated context map?

Fusing multiple observations from multiple sensors (e.g., observing the scene from different angles) could rule out the possibility that the false track was generated by noisy sensory data. Again, multiple readings and trajectory analysis could rule out the possibility of a bias in sensor calibration (e.g., a trajectory generated by actual vehicles appears to be correctly generated within the road width). The observation that the false tracks are mostly generated during certain hours of the day and/or while certain weather conditions might be learned and accommodated as additional contextual information that will be used by the system to explain and rule out false tracks. Fusing multiple observations can give good results only if they are reliable. Considerations of context can be used for learning this reliability.

Figure 11.5 shows the flow of some of the processing and quality assessment steps necessary for vehicle tracking. In particular, the urban scenario depicted in Fig. 11.4 acts as CO and provides the setting that will provide the expectations for the behavior of the traffic. The hypotheses generated by the behavior understanding module can, for example, be checked against those expected for the current scenario (e.g., normal patterns of activity). This would rule out unlikely ones, given the

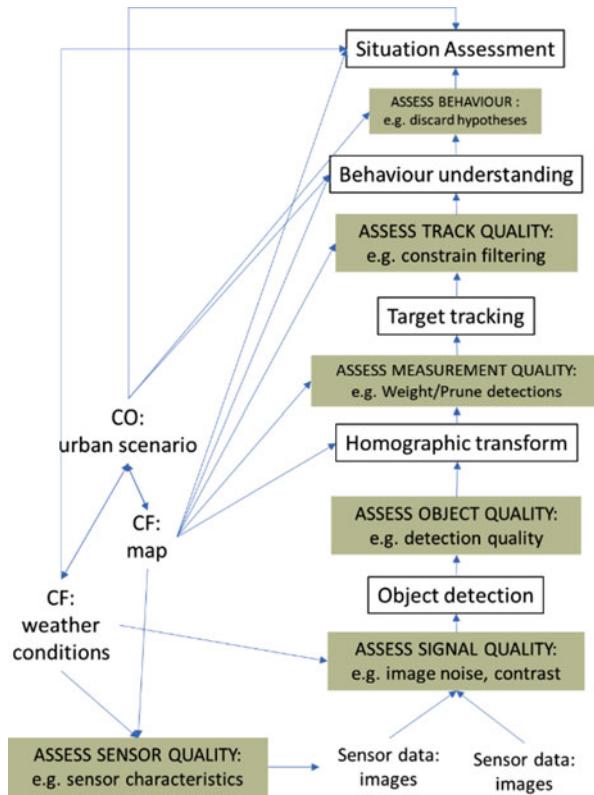


Fig. 11.5 Processing and quality assessment steps for the tracking scenario shown in Fig. 11.4

current CO. Specific elements of the scenario, such as the weather conditions and the map of the city, can be used as CF variables that refine expectation of the normal behavior and assist in quality assessment of sensor performance and tracking results. Sensor quality could be a function of the current weather conditions (e.g., infrared cameras should be preferred for low light conditions) as well as the actual positioning (e.g., the closest sensors to the target should be preferred).

The city map is a CF that is directly exploited both in transforming sensor detections into map plane points (homographic transform [51]) and into assessing their reliability (e.g., weighting the measurements depending on the likelihood of belonging to a certain element of the map, e.g., the road for vehicles, curbs, and zebra crossings for pedestrians).

Context quality should also undergo continuous assessment. For example, accumulating target trajectories over time can produce an activity model that differs from the expected one and thus leading to the discovery of a new context (e.g., an accident, etc.).

11.5.2 Maritime Surveillance

The following example describes the interplay between context, fusion, and IQ in a maritime surveillance scenario designed to identify suspicious activities. Unlike the example considered in the previous section, the focus here is on situation assessment and reasoning processes for constructing a dynamic surveillance picture for piracy threat detection and recognition [52, 53].

Maritime surveillance systems receive a huge amount of highly heterogeneous input often uncertain, unreliable, and irrelevant: from device-generated data such as radar contacts and AIS information to human-generated information such as coast guard reports. Given the relatively slow speeds and large distances typical of vessels, maritime surveillance systems generally have the time to acquire and reason on many different pieces of information received as streaming input or mined in a knowledge base.

The process of building a dynamic surveillance picture must include all fusion levels. Lower-level fusion combines data and information to provide unified and reliable tracking and recognition of all interesting entities by estimating the main problem variables (ship tracks, IDs) along with their relationships in order to assess the current situation. Higher-level fusion transforms information about vessel identity and tracks into knowledge about potentially suspicious behavior to aid the operator in recognizing this behavior as threat or false alarm. The process of reasoning for detection and recognition of threat and the role of context and IQ are shown in Fig. 11.6. The focus there is on two very important quality characteristics for situation and threat assessment: relevance and consistency.

As shown in Fig. 11.6, maritime suspicious activity is considered in the *context* of the geophysical and geopolitical world situation (e.g., the situation in Somalia, relations between different countries in the Malacca Strait, and recent reports on increased piracy attacks there), which provide expectations about the possible threat from pirates. The *context* under consideration comprises maritime traffic rules (general and for special vessels), weather conditions, and maps and is used for better situation understanding by utilizing characteristics of normal situations as opposite to ones specific to pirate behavior.

Before information is allowed into the fusion system, a relevance check is performed in a sifting process [5] in order to channel only relevant pieces of data and information to the fusion process. Context is the key in this selection process along with the given system requirements/mission [3]. Given the current context and mission requirements, input data and information are weighted accordingly. For example, sea-lanes regulations (CF) might be particularly enforced in areas plagued by piracy (CO) so that trajectories of non-abiding vessels might be promptly detected.

After the fusion processes are performed, it is of paramount importance to estimate the consistency of the data/information provided by the available input sources and fusion processes, and contextual information. As already mentioned in previous sections, it could, in fact, be the case that inconsistency is between the hypothesized

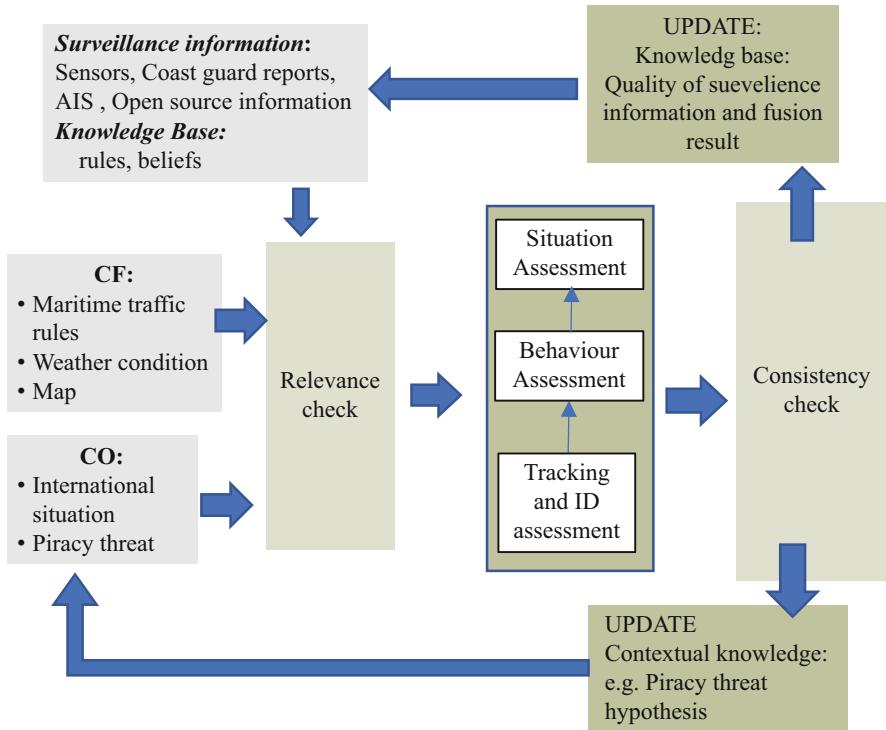


Fig. 11.6 The process of reasoning for detection and recognition of threat and the role of context and IQ

situation and the observed objects, relations, situational items, and their behaviour. Sensor/source observations could be of low accuracy or even purposely wrong as in the case of AIS spoofing, fusion processes can result in information of insufficient quality. It could also be the case that contextual information is inaccurate due to poor observability or unreliable information about pirates' intentions. The problem here, as it was mentioned in previous sections, is how and whether to change the state of the knowledge (e.g., threat/no threat) due to the arrival of new information if the latter contradicts prior knowledge in the context under consideration. The process of resolving the inconsistency involves reasoning for best explanations (abduction) to determine the sources of the contradiction such as possible insufficient quality of contextual knowledge, observations, fusion processes or current hypotheses about the situation. A continuous process of assessing and updating both the quality of the fused result, input data/information and contextual information takes place in order to find the best hypothesis explaining the data and/or rule out the discrepancies in the observations.

11.6 Conclusions

This chapter has examined the problems of context exploitation in information fusion and the interrelationships among context quality, the quality of observations, and the results of fusion processes. It discusses the role of context and information quality in information fusion and provides a definition of context quality and incorporation of quality into context representation, as well as methods of quality control along with desirable properties of these methods. The chapter also presents examples of consideration of context and information quality in the problems of city traffic monitoring and maritime surveillance for piracy threat detection.

“Fitness for use” of the incoming information as well as the performance and effectiveness of fusion processes depend on correct context selection and successful estimation of the quality of contextual variables and their relevance to the purpose of fusion. The information defining a context can be obtained from available databases, observations, the result of sensor fusion, received reports, mining of available information sources (e.g., traditional and social media), or various levels of information fusion. Of course, the quality of any such information, as well as the inference process for obtaining it, could be insufficient for a particular use: it might be uncertain, unreliable, irrelevant, or conflicting. Knowledge of the quality of this information and its effect on the quality of context characterization can improve contextual knowledge.

There are two interrelated problems concerning both information and context quality: imperfect information used in context estimation and discovery negatively impacts context quality, while imperfect context characterization adversely affects the characterization of quality of information used in fusion as well as the fusion results. This interrelationship represents one of the challenges of modeling and evaluating context quality and of using context in defining the quality of information used in fusion and the quality of fusion process results. A solution for this relatively new and important problem of interrelationships among information quality and quality of context is necessary for improved fusion performance. This and the problem of context discovery in uncertain dynamic environments pose a significant challenge for information fusion.

References

1. G. Rogova, L. Snidaro, Considerations of context and quality in information fusion, in *Proceedings of the 21st International Conference on Information Fusion*, (IEEE, Cambridge, UK, 2018), pp. 1929–1936
2. L. Snidaro, J. García, J. Llinas, E. Blasch (eds.), *Context-Enhanced Information Fusion – Boosting Real-World Performance with Domain Knowledge, Advances in Computer Vision and Pattern Recognition Series* (Springer International Publishing, Switzerland, 2016)
3. L. Snidaro, J. Garcia, J. Llinas, Context-based information fusion: A survey and discussion. *Inf. Fusion* **25**, 16–31. (Elsevier, Amsterdam, Netherlands, 2015)

4. A.N. Steinberg, G. Rogova, Situation and context in data fusion and natural language understanding, in *Proceedings of the 11th International Conference on Information Fusion*, (IEEE, Cologne, 2008), pp. 1868–1875
5. L. Snidaro, I. Visentini, J. Llinas, G.L. Foresti, Context in fusion: Some considerations in a JDL perspective, in *Proceedings of the 16th International Conference on Information Fusion*, (IEEE, Istanbul, 2013), pp. 115–120
6. J.R. Anderson, *Language, Memory and Thought* (Routledge, Erlbaum, Hillsdale, 1976)
7. P. Brézillon, J.-Ch. Pomerol, Misuse and nonuse of knowledge-based systems: The past experiences revisited, in *Implementing Systems for Supporting Management Decisions*, ed. by P. Humphreys, L. Bannon, A. McCosh, P. Migliarese, J. Ch. Pomerol, (Chapman and Hall, London, 1996), pp. 44–60
8. A. Dey, Understanding and using context. Pers. Ubi. Comp. **5**(1), 4–7 (2001)
9. B. Schilit, M. Theimer, Disseminating active map information to Mobile hosts. IEEE Netw. **8**(5), 22–32 (1994)
10. K. Sycara, R. Glinton, B. Yu, J. Giampapa, S. Owens, M. Lewis, C. Grindle, An integrated approach to high-level information fusion. Inf. Fusion **10**, 25–50. (Elsevier, Amsterdam, Netherlands, 2009)
11. J. Gómez-Romero, M.A. Serrano, J. García, J.M. Molina, G. Rogova, Context-based multi-level information fusion for harbor surveillance. Inf. Fusion **21**, 173–186. (Elsevier, Amsterdam, Netherlands, 2015)
12. L. Gong, Contextual modeling and applications, in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 1, (IEEE, Waikoloa, HI, 2005), pp. 381–386
13. G. Rogova, A. Steinberg, Formalization of “Context” for Information Fusion, in *Context-Enhanced Information Fusion - Boosting Real-World Performance with Domain Knowledge*, Advances in Computer Vision and Pattern Recognition Series, ed. by L. Snidaro, J. García, J. Llinas, E. Blasch, (Springer International Publishing, Switzerland, 2016), pp. 27–44
14. G. Rogova, E. Bosse, Information quality in information fusion, in *Proceedings of the 13th International conference on Information Fusion*, (IEEE, Edinburgh, Scotland, 2010)
15. J. Llinas, G. Rogova, K. Barry, J. Scrofani, Reexamining computational support for intelligence analysis: A functional design for a future capability, in *Proceedings of the SPIE DEFENSE + SECURITY Conference*, vol. 10653, (SPIE, Orlando, 2018)
16. F. Pichon, D. Dubois, T. Denœux, Relevance and truthfulness in information correction and fusion. Int. J. Approx. Reason. **53**(2), 159–175. (Elsevier, Amsterdam, Netherlands, 2012)
17. A. Steinberg, G. Rogova, System-level use of contextual information, in *Context-Enhanced Information Fusion – Boosting Real-World Performance with Domain Knowledge*, Advances in Computer Vision and Pattern Recognition series, ed. by L. Snidaro, J. García, J. Llinas, E. Blasch, (Springer International Publishing, Switzerland, 2016), pp. 157–184
18. Standard 8402, 3. I, International Organization of Standards, 1986
19. J. Juran, A. Godfrey, *Juran's Quality Handbook*, 5th edn. (McGraw-Hill, New York, 1988)
20. M. Bovee, R.P. Srivastava, B. Mak, A conceptual framework and belief-function approach to assessing overall information quality. Int. J. Intell. Syst. **18**, 51–74. (Wiley Periodicals, Inc., Hoboken, NJ, 2003)
21. C. Bisidikian, L. Kaplan, M. Srivastava, D.V. Thornley, R. Young, Building principles for a quality of information , specification for sensor information, in *Proceedings of the 12th International Conference on Information Fusion*, (IEEE, Seattle, 2009), pp. 1370–1377
22. R. Wang, D. Strong, Beyond accuracy: What data quality means to data consumers. J. Man. Inf. Syst. **12**(4), 5–34 (1996)
23. M. Helfert, Managing and measuring data quality in data warehousing, in *Proceedings of the 17th World Multiconference on Systemics, Cybernetics and Informatics*, vol. 38 (2001), pp. 55–65
24. Ph. Smets, Imperfect information: Imprecision – uncertainty, in *Uncertainty Management in Information Systems: From Needs to Solutions*, ed. by A. Motro, Ph Smets (Kluwer, Boston, 1997), pp. 225–254

25. G. Rogova, Information quality in information fusion and decision making with applications to crisis management, in *Fusion Methodology in Crisis Management: Higher Level Fusion and Decision Making*, ed. by G. Rogova, P. Scott (Springer International Publishing, Switzerland, 2016), pp. 65–86
26. D.M. Rousseau, S.B. Sitkin, R.S. Burt, C. Camerer, Not so different after all: A cross discipline view of trust. *Acad. Manag. Rev.* **23**(3), 393–404 (1998)
27. C. Castelfranchi, R. Falcone, *Trust Theory: A Socio-Cognitive and Computational Model* (Wiley, Chichester, 2010)
28. S.B. Sitkin, N.L. Roth, Explaining the limited effectiveness of legalistic “remedies” for trust/distrust. *Organ. Sci.* **4**(3), 367–392 (1993)
29. C.F. Chang, P. Harvey, A. Ghose, Combining credibility, in *A Source Sensitive Argumentation System*, ed. by G. Antoniou et al., in *Proceedings of the, 4th Hellenic Conference on AI*, SETN 2006, Heraklion, Crete, Greece, 18–20 May 2006
30. L. Bonelli, S. Felletti, F. Paglieri, From argumentative crisis to critical arguments: How to argue in the face of danger, in *Fusion Methodology in Crisis Management: Higher Level Fusion and Decision Making*, ed. by G. Rogova, P. Scott, (Springer, International Publishing, Switzerland, 2016), pp. 365–381
31. T. Saracevic, Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II. *J. Am. Soc. Inf. Sci. eTech.* **58**(3), 1915–1933 (2006)
32. D. Walton, *Relevance in Argumentation* (Routledge, Oxfordshire, UK, 2003)
33. D. Galles, J. Pearl, Axioms of causal relevance. *Artif. Intell.* **97**, 9–43 (1997)
34. T. Buchholz, A. Kupper, M. Schifflers, Quality of context information: What it is and why we need it, in *Proceedings of the 10th International Workshop of the HP OpenView University Association (HPOVUA)*, vol. 200, (Infonomics-Consulting, Stuttgart, Germany, Geneva, 2003)
35. M. Krause, I. Hochstatter, Challenges in Modelling and Using Quality of Context (QoC), in *Mobility Aware Technologies and Applications*, eds. by T. Magedanz, A. Karmouch, S. Pierre, I. Venieris, MATA2005, Lecture Notes in Computer Science, vol. 3744, (Springer, 2005), pp. 324–333
36. P. Brezzillon, Context in problem solving: A survey. *Knowl. Eng. Rev.* **14**(1), 47–80 (Cambridge University Press, Cambridge, England, 1999)
37. P. Thagard, C.P. Shelley, Abductive reasoning: Logic, visual thinking, and coherence, in *Logic and Scientific Methods*, ed. by M.-L. Dalla Chiara et al., (Kluwer, Dordrecht, Netherlands, 1997), pp. 413–427
38. J. Josephson, On the Logical Form of Abduction, AAAI Spring Symposium Series: Automated Abduction, (1990) pp. 140–144
39. T. Strang, C. Linnhoff-Popien, A context modeling survey, in *Proceedings of the First International Workshop on Advanced Context Modelling, Reasoning and Management at UbiComp 2004*, (Nottingham, 2004), pp. 34–41
40. B. Schilit, N. Adams, R. Want, Context-aware computing applications, in *IEEE Workshop on Mobile Computing Systems and Applications*, (IEEE, Santa Cruz, 1994)
41. M. Kokar, C.J. Matheus, K. Baclawski, Ontology-based situation awareness, *Inf. Fusion* **10**, 83–98, (Elsevier, Amsterdam, Netherlands, 2009)
42. J. Gómez-Romero, F. Bobillo, M. Delgado, Context representation and reasoning with formal ontologies, in *Proceedings of the Activity Context Representation Workshop in the 25th AAAI Conference*, (AAAI, San Francisco, 2011)
43. J. McCarthy, Notes on formalizing context, in *Proceedings of the Thirteenth International Joint Conference in Artificial Intelligence*, (Morgan Kaufmann Publisher, Chambery, 1993), pp. 555–560
44. M. Akman, M. Surav, The use of situation theory in context modeling. *Comp. Intell.* **13**(3), 427–438 (1997)
45. J. Barwise, J. Perry, The situation underground, in *Working Papers in Semantics*, vol. 1, (Stanford University Press, Redwood City, California, 1980)

46. K. Devlin, *Logic and Information* (Press Syndicate of the University of Cambridge, Cambridge, England, 1991)
47. J. George, J.L. Crassidis, T. Singh, Threat assessment using context-based tracking in a maritime environment, in *Proceedings of the 12th International Conference on Information Fusion*, (IEEE, Seattle, 2009), pp. 187–194
48. A. Burton-Jones, V.C. Storey, V. Sugumaran, P. Ahluwalia, A semiotic metrics suite for assessing the quality of ontologies. *Data Knowl. Eng.* **55**(1), 84–102 (2005)
49. S. Mc Gurk, C. Abela, J. Debattista, Towards ontology quality assessment, in *Proceedings of the 4th Workshop on Linked Data Quality (LDQ 2017)*, (Portorož, 2017)
50. S. Zilberstein, Using Anytime Algorithms in Intelligent Systems. *AI Magazine* **17**(3), 73–83 (AAAI, 1996)
51. I. Visentini, L. Snidaro, Integration of contextual information for tracking refinement, in *Proceedings of the 14th International Conference on Information Fusion*, vol. 5–8, (IEEE, Chicago, 2011)
52. L. Snidaro, I. Visentini, K. Bryan, Fusing uncertain knowledge and evidence for maritime situational awareness via Markov Logic Networks. *Inf. Fusion* **21**, 159–172 (2015). <https://doi.org/10.1016/j.inffus.2013.03.004>
53. G. Rogova, J. Herrero, Contextual Knowledge and information fusion for Maritime Piracy surveillance, in *Prediction and Recognition of Piracy Efforts Using Collaborative Human-Centric Information Systems*, ed. by E. Shahbazian, G. Rogova, E. Bosse, (IOS Press, Amsterdam, Netherlands, 2013)

Chapter 12

Analyzing Uncertain Tabular Data



Oliver Kennedy and Boris Glavic

Abstract It is common practice to spend considerable time refining source data to address issues of data quality before beginning any data analysis. For example, an analyst might impute missing values or detect and fuse duplicate records representing the same real-world entity. However, there are many situations where there are multiple possible candidate resolutions for a data quality issue, but there is not sufficient evidence for determining which of the resolutions is the most appropriate. In this case, the only way forward is to make assumptions to restrict the space of solutions and/or to heuristically choose a resolution based on characteristics that are deemed predictive of “good” resolutions. Although it is important for the analyst to understand the impact of these assumptions and heuristic choices on her results, evaluating this impact can be highly nontrivial and time-consuming. For several decades now, the fields of probabilistic, incomplete, and fuzzy databases have developed strategies for analyzing the impact of uncertainty on the outcome of analyses. This general family of uncertainty-aware databases aims to model ambiguity in the results of analyses expressed in standard languages like SQL, SparQL, R, or Spark. An uncertainty-aware database uses descriptions of potential errors and ambiguities in source data to derive a corresponding description of potential errors or ambiguities in the result of an analysis accessing this source data. Depending on the technique, these descriptions of uncertainty may be either quantitative (bounds, probabilities) or qualitative (certain outcomes, unknown values, explanations of uncertainty). In this chapter, we explore the types of problems that techniques from uncertainty-aware databases address, survey solutions to these problems, and highlight their application to fixing data quality issues.

O. Kennedy (✉)

Department of Computer Science and Engineering, University at Buffalo,
SUNY, Buffalo, NY, USA
e-mail: okennedy@buffalo.edu

B. Glavic

Computer Science, Illinois Institute of Technology, Chicago, IL, USA
e-mail: bglavic@iit.edu

Keywords Databases · Probabilistic databases · Incomplete databases · Incomplete information · Uncertain data

12.1 Introduction

Data quality is increasingly relevant to all facets of data management, from small-scale personal sensing applications to large corporate or scientific data analytics. In these and many other settings, sources of uncertainty include inaccuracies and failures of sensors, human data entry errors, systematic errors during fusion of data from heterogeneous sources, and many others. A prevalent problem when identifying data quality issues is that typically available information is insufficient to determine with certainty whether a fluke detected in the data is a quality issue or simply an unusual fact. For example, a high blood pressure value may be due to a medical problem or may as well be a measurement error – without further information it is impossible to distinguish between these two cases. Even if a quality issue can be detected with certainty, this does not imply that we can find a unique or even any correct resolution for this issue. For instance, consider an employment dataset with records storing for each employee a unique identifier like a social security number (SSN), their name, and the department they work in. A single person might occur multiple times in this dataset if the person works for more than one department. If we assume that unique identifiers are truly unique, one type of data quality issue that is easy to detect (for this dataset) is when multiple records corresponding to the same identifier have different names.¹ While this condition is easy to check, it is not straightforward to repair a dataset with records that share identifiers but have conflicting names. For example, assume that our dataset contains two conflicting records (777-777-7777, Peter Smith, Sales) and (777-777-7777, Bob Smith, Marketing). These records indicate that a person with SSN 777-777-7777 is working in both sales and marketing and is either named Peter Smith or Bob Smith. The correct way to fix this conflict depends on what caused the error and requires us to have additional information that is not readily available. If we assume that two people with the same SSN must have the same name, then there are four ways we could fix the error: (1) We could assume that one of the records was created mistakenly and delete it; (2) We could assume that our analysis will be minimally affected by this error and leave the records as-is; (3) We could assume that the name attribute of one (or both) of the records is incorrect and update them accordingly; and/or (4) We could assume that the SSN

¹In database terminology, we would say that a functional dependency $id \rightarrow name$ holds for the dataset, i.e., the “id” value of a record determines its name value. Or put differently, there are no two records that have the same SSN, but a different name. The intricacies of functional dependencies are beyond the scope of this paper. The interested reader is referred to database textbooks (e.g., [1]). Furthermore, see, e.g., [12] for how constraints like functional dependencies are used to repair data errors.

attribute of one (or both) of the records is incorrect, and update them accordingly. In the absence of further information, there is no way for us to determine what the correct fix for the problem is.

A typical approach for resolving this kind of ambiguity is to correlate the information in our dataset with externally available sources of high-quality data, often called master data, to determine what fix to apply. For instance, we may have access to a reliable list of what name corresponds to which SSN. Such information is useful, because it allows us to make SSNs consistent with names. However, it is not sufficient for figuring out whether the issue arose from incorrect name(s) or from incorrect SSN(s) and, thus, whether the correct solution is to repair SSN or name value(s). This example demonstrates that even trivial data quality issues can have hard- or impossible-to-determine resolutions.

In a typical analytics workflow, problems like this are addressed at the start: a task we call *data curation*. Common approaches to data curation include: (1) If redundant data sources can be found, the curator can manually repair errors; (2) If records are irreparable, the curator may remove them from the dataset outright; (3) If removing records with errors would create bias or leave too few records, the curator can apply heuristic techniques (e.g., imputation) to repair errors. (4) If the errors are deemed to not affect the analysis, the curator may simply choose to not address them. For example, we might train a classifier to find records in our master data most similar to Peter and Bob Smith, and then use it to fix the name and/or SSN fields of our erroneous data set.

A common assumption is that once curation is complete, the data is “correct,” at least to the point where specific analyses run over it will produce valid results. However, such assumptions can produce misleading analytical results. The most straightforward of these are the consequences of incorrect choices during curation, for example, selecting a classifier that is inappropriate for the available data. However, there is a more subtle possibility: The data being curated might be insufficient or of insufficient quality – regardless of curation technique – to support some or all of the analyst’s goals.

Existing data management systems, Relational Databases, Graph Databases, Spark, NoSQL Systems, R, and others, are built on the assumption that data is exact. These systems cannot distinguish low-confidence information derived of incomplete data or heuristic guesswork from high-confidence information grounded in verifiable data fused from multiple redundant data sources. In extreme cases, such misinterpretations ruin lives. For example, credit report errors can severely limit a person’s access to financial services and job opportunities.

Example 1 Consider an example bank’s loan database shown in Fig. 12.1. Table `customers` stores the SSN, name, income, and assets for each customer. Furthermore, we record for each customer the number of mortgages this customer has and whether they own property or not. Table `applications` stores loan applications (a customer’s SSN and the amount of loan they requested). The bank uses the following formula to determine the maximum amount $Max\$$ they are willing to loan to a customer.

<u>customers</u>					
SSN	name	income	assets	numMortgages	ownsProperty
777-777-7777	Alice Alison	\$60,000	\$200,000	0	NULL
333-333-3333	Bob Bobsen	\$102,000	\$15,000	0	no
111-111-1111	Peter Petersen	NULL	\$90,000	1	yes
555-555-5555	Arno Arnoldson	\$95,000	\$30,000	0	yes

<u>applications</u>	
SSN	loanAmount
777-777-7777	\$90,000
111-111-1111	\$10,000

Fig. 12.1 Example for resolving a data quality issue and how this affects the trustworthiness of analysis results. The input data contains missing values and needs to be imputed

$$Max\$ = 0.3 \cdot income + 0.2 \cdot assets - 10,000 \cdot numMortgages + 40,000$$

ownsProperty

The bank might use a SQL query such as the one shown below to determine which loan applications it should grant and which applications will be rejected.

```
SELECT SSN, name, CASE WHEN maxAllow >= requestedAmount
                      THEN 'yes' ELSE 'no'
                 END AS grantLoan
FROM (SELECT 0.3 * income + 0.1 * assets - 10000 * numMortgages
      + CASE WHEN ownsProperty THEN 20000 ELSE 0 END AS
      maxAllow, loanAmount AS requestedAmount
     FROM customer c, applications a WHERE c.SSN = a.SSN) sub
```

For readers unfamiliar with SQL, this query (1) retrieves the customer information for each loan request, then (2) computes $Max\$$ based on the customer information, and (3) returns the SSN and the name of the customer requesting the loan as well as a column *grantLoan* whose value is *yes* if the amount requested is lower than $Max\$$ or *no* otherwise.

Unfortunately, the customers table contains missing values (in SQL the special value **NULL** is used to denote that the value of an attribute is missing). For example, we do not know whether Alice owns property or not, or what Peter's income is. These missing values prevent us from determining whether Alice's loan should be approved or not. As mentioned above, one way to resolve this issue is to train a classifier, such as a decision tree. The classifier is trained on records that have all their attribute values. Given the values of the other attributes (SSN, name, income, assets, and #mortgages) for a record, the trained classifier predicts the missing value of the *ownsProperty* attribute.

Assume that the classifier predicts that the value of *ownsProperty* for Alice's record is yes with 40% probability and no with 60% probability. The **NULL** value is replaced with *no*, the value with the highest probability. If the dataset curated in this

Classifier Result for Alice		Query Result		
ownProperty	probability	SSN	name	grantLoan
yes	0.4	777-777-7777	Alice Alison	no
no	0.6	111-111-1111	Peter Peterson	yes

way is used to determine Alice’s eligibility to receive the loan, then the bank would deny her loan application – even if she does in fact own property and should be eligible. In this case, the bank does not have enough information to decide whether to give Alice a loan, but by treating imputed missing values as fact, this lack of information is obscured.

As we will explain in more detail in the remainder of this chapter, uncertainty-aware approaches to data management can expose this lack of information. In the example, this would allow the bank’s computer to produce a third response, “I don’t know,” exposing the uncertainty in the analytical result. Depending on what technique for uncertain data management is applied, it might also be feasible to determine exactly what information is necessary to produce a definitive answer to the bank’s question.

Uncertain data management techniques can help to expose implicit biases introduced during data curation, as well as the impact that these biases have on the quality and trustworthiness of analytical results. In this chapter, we provide an introduction to uncertain data management techniques developed by the database community and their use in data curation and analysis. While there are excellent surveys on probabilistic data management, e.g., [39] and [8], these surveys aim to describe the technical challenges involved. In contrast, we give a more goal-oriented introduction to the topic targeted at helping practitioners identify suitable techniques for their needs and to clarify the connection to practical applications of these techniques for data quality.

Making uncertainty a first-class primitive in data management has been the goal of the so-called *uncertain* and *probabilistic query processing* [39] or *PQP*. There have been significant efforts in this space, aiming to produce efficient algorithms for some of the most computationally challenging tasks involved in managing uncertain data [7]. While great progress has been made, some tasks in *PQP* are too computationally intensive to be widely applied. Thus, having practical applications in mind, we also survey lightweight approaches which produce less detailed and/or approximate descriptions of uncertainty, at the benefit of reduced computational complexity. Though they are approximations, these simpler representations of uncertainty are also often easier to interpret than more complex models.

The remainder of this chapter is organized as follows. We introduce core concepts of uncertain data management in Sect. 12.2. We then cover data models that encode uncertainty in one form or another in Sect. 12.3. Section 12.4 covers methods for computing the results of queries and evaluating general programs over uncertain data. Afterward, we discuss methods for presenting uncertainty to end users in Sect. 12.5. Finally, we conclude in Sect. 12.6.

12.2 Core Concepts: Incomplete and Probabilistic Databases

We begin with the broadest assumptions about the workflows applied by an analyst, before narrowing our scope to particular models of computation (mostly declarative query languages) when discussing specific solutions. Specifically, we assume that the analyst has specified her analysis as a program Q expressed in some suitable form such as an imperative programming language (e.g., Python or C) or a database query language (e.g., SQL or Spark). When applied to some dataset D , this program produces a result $Q(D)$. For instance, for our running example, the program Q is the SQL query shown in Fig. 12.1, D is the cleaned version of the loan dataset (after imputing the missing `ownsProperty` value of Alice’s record), and $Q(D)$ is the query result shown in Fig. 12.1.

12.2.1 Possible World Semantics

A well-established model for uncertainty in such an analysis is the so-called possible world semantics. Under the possible worlds semantics, we forsake the existence of single deterministic dataset and instead consider a set of *possible* datasets, the possible worlds. Each possible world is a dataset that could exist under certain assumptions. Formally, an uncertain dataset is a (potentially infinite) set of datasets \mathcal{D} . We refer to \mathcal{D} as an *incomplete database* and its members as *possible worlds*.

Example 2 Continuing with our running example, recall that we used a classifier to impute the missing `ownsProperty` attribute of Alice’s record. In the previous example, we did pick the most likely value predicted by the classifier (no) and saw how this incorrect choice caused Alice’s loan to be rejected. The heuristic underlying this choice is the classifier; We assume that it will always pick the correct replacement value – in general a quite strong assumption. Instead we can model the output of the missing value imputation as a set of possible worlds $\mathcal{D} = \{D_1, D_2\}$. In this example, there are two worlds that are shown in Fig. 12.2: either Alice does not own property (possible world D_1 shown on top) or Alice does own property (possible world D_2 shown on the bottom). In this possible world model, it is evident that we are uncertain about whether Alice owns property or not, as the only difference between the two possible worlds is the `ownsProperty` attribute value for Alice’s record.

When an analysis program Q is evaluated over an incomplete database \mathcal{D} , the result is not one but a set of possible results – the set of all results that could be obtained by evaluating Q in some possible world:

$$Q(\mathcal{D}) := \{ Q(D) \mid D \in \mathcal{D} \}$$

Possible World D_1 ($p = 0.6$) customers						
SSN	name	income	assets	#mortgages	ownsProperty	
777-777-7777	Alice Alison	\$60,000	\$200,000	0	no	
333-333-3333	Bob Bobson	\$102,000	\$15,000	0	no	
111-111-1111	Peter Petersen	NULL	\$90,000	1	yes	
555-555-5555	Arno Arnoldson	\$95,000	\$30,000	0	yes	

Possible World D_2 ($p = 0.4$) customers						
SSN	name	income	assets	#mortgages	ownsProperty	
777-777-7777	Alice Alison	\$60,000	\$200,000	0	yes	
333-333-3333	Bob Bobson	\$102,000	\$15,000	0	no	
111-111-1111	Peter Petersen	NULL	\$90,000	1	yes	
555-555-5555	Arno Arnoldson	\$95,000	\$30,000	0	yes	

Fig. 12.2 Possible worlds representation of the cleaned loan dataset created based on the possible imputed values for the missing ownsProperty value predicted by the classifier

That is, the result of evaluating an analysis program Q over an incomplete database is itself an incomplete database.

Example 3 Evaluating our loan query over the incomplete database from Fig. 12.2, we get two possible results shown below. We are sure that Peter should be granted a loan, but there are two possible outcomes for Alice's loan application. If Alice owns property (possible world D_2), then we would consider her to be eligible for this loan (possible result $Q(D_2)$). Otherwise, (possible world D_1) we should reject her loan application $Q(D_1)$.

$Q(D_1)$			$Q(D_2)$		
SSN	name	grantLoan	SSN	name	grantLoan
777-777-7777	Alice Alison	no	777-777-7777	Alice Alison	yes
111-111-1111	Peter Peterson	yes	111-111-1111	Peter Peterson	yes

12.2.2 Certain and Possible Records

It is often useful to reason about a dataset D as a collection of records $r \in D$ or a table. When using this model, we also treat analytical results as sets of records $r \in Q(D)$. We then may want to reason about what facts (records) we know for certain to be true, what facts are potentially true, and which facts are known to be untrue. For instance, in our running example, we know with certainty that Peter's

loan should be granted, since regardless of which possible world represents the true state of the world, the loan application will be granted. For incomplete databases, we say that a record is *certain* if the record appears in every possible world. Note that, because the result of analysis over an incomplete database is again an incomplete database, we can apply the same concept to analysis results. A record is certainly in the result of an analysis if it is present in the result irrespective of which possible world correctly describes the true state of the world. Formally, a record r is *certain* in an incomplete database \mathcal{D} if $\forall D \in \mathcal{D} : r \in D$. We use $\text{certain}(\mathcal{D})$ to denote the set of all certain records in \mathcal{D} . Analogously, the certain answers to a question $Q(\mathcal{D})$, which we write as $\text{certain}(Q(\mathcal{D}))$ are the set of all certain records in the incomplete result $Q(\mathcal{D})$.

$$r \text{ is certainly in } \mathcal{D} := \forall D \in \mathcal{D} : r \in Q(D)$$

$$\text{certain}(Q(\mathcal{D})) = \{ r \mid \forall D \in \mathcal{D} : r \in Q(D) \}$$

Conversely, we say that a record is *possible* if the record appears in at least one possible world (resp., possible result). The possible answers of a question ($\text{possible}(Q(\mathcal{D}))$) are defined symmetrically.

$$r \text{ is possibly in } \mathcal{D} := \exists D \in \mathcal{D} : r \in Q(D)$$

$$\text{possible}(Q(\mathcal{D})) = \{ r \mid \exists D \in \mathcal{D} : r \in Q(D) \}$$

Example 4 For instance, $(111-111-1111, \text{Peter Peterson}, \text{yes})$ is a certain answer in our running example since it is in the result for every possible world, but $(777-777-7777, \text{Alice Alison}, \text{yes})$ is not since this record is not in $Q(D_1)$. Conversely, $(777-777-7777, \text{Alice Alison}, \text{yes})$ is possible while $(777-777-7777, \text{Alice Peterson}, \text{yes})$ is not.

12.2.3 Multisets

Many database systems use *bags* (multisets) of records to represent data. That is, bag databases allow for multiple records that have all attribute values in common. For example, if we allow multiple loan applications by the same person, then it may be possible that two applications from the same person are requesting the same amount, i.e., are duplicates.

Example 5 Consider the bag table *applications* shown below. Here both Alice (SSN 777-777-7777) and Peter (SSN 111-111-1111) have submitted two applications. Alice's applications are both for a loan of \$90,000, i.e., there is a duplicate of the record $(777-777-7777, \$90,000)$ in the applications table.

SSN	loanAmount
777-777-7777	\$90,000
777-777-7777	\$90,000
111-111-1111	\$10,000
111-111-1111	\$3,000

We write $D[r]$ (resp., $Q(D)[r]$, $R[r]$) to denote the number of occurrences or the multiplicity of record r in database D (resp. result $Q(D)$ or record set R). We can now define the multiplicities of an uncertain record as a set of multiplicities for each possible world.

$$\mathcal{D}[r] = \{ D[r] \mid D \in \mathcal{D} \}$$

Libkin et al. [19] generalize the notion of certain and possible results to reason about bounds on these multisets (or bags).

$$\mathbf{certain}(\mathcal{D}[r]) = \min(\mathcal{D}[r]) \quad \mathbf{possible}(\mathcal{D}[r]) = \max(\mathcal{D}[r])$$

Observe that when a multiset \mathcal{D} encodes a set (i.e., when $\mathcal{D}[r] \in \{0, 1\}$), multiset possible and certain answers behave as their set-based counterparts.

12.2.4 Incorporating Probability

An incomplete database may be supplemented with a probability measure $P : \mathcal{D} \rightarrow [0, 1]$ over the set of possible worlds requiring that $\sum_{D \in \mathcal{D}} P(D) = 1$. The pair (\mathcal{D}, P) of incomplete database and probability measure is called a probabilistic database [16]. When a question Q is asked of an probabilistic database (\mathcal{D}, P) , we can derive a marginal distribution over the set of possible results $\{R_i\}$:

$$p[Q(\mathcal{D}) = R_i] := \sum_{D \in \mathcal{D} \text{ s.t. } Q(D) = R_i} P(D)$$

When the result is a set (resp., bag) of records, we derive the marginal probability of any possible record in the result similarly (where $M \in \mathbb{N}$ is a record multiplicity).

$$p[r \in Q(\mathcal{D})] := \sum_{D \in \mathcal{D} \text{ s.t. } r \in Q(D)} P(D)$$

$$p[Q(\mathcal{D})[r] = M] := \sum_{D \in \mathcal{D} \text{ s.t. } Q(D)[r] = M} P(D)$$

We will sometimes use $p(r)$ as a notational shortcut for $p[r \in Q(\mathcal{D})]$ when Q and \mathcal{D} are understood from the context.

Example 6 Recall that the classifier we have trained predicts that Alice owns property with 40% probability and that she does not own property with 60% probability. Thus, possible world D_1 has probability $p(D_1) = 0.6$ and possible world D_2 has probability $p(D_2) = 0.4$. Then based on this probability distribution of the input possible worlds, the probabilities of the two possible query results are determined. The probability of the result where Ann's loan application is rejected is 0.6, while the one where Ann's loan is granted is 0.4. Given this result, we can compute the marginal probability of an analysis result record, i.e., the likelihood that this particular record exists in the result as the sum of the probabilities of possible worlds containing this record. Let r_1 be the record corresponding to the loan granted to Peter. Since this record appears in both possible worlds, we get $p(r_1) = p(D_1) + p(D_2) = 0.4 + 0.6 = 1$. That is, the record has probability 1.

Observe that there is a strong connection between the concepts of certain/possible records and the marginal probability of a record. Given a record r , we have:

$$p[r \in Q(\mathcal{D})] = 1 \Leftrightarrow r \in \mathbf{certain}(Q(\mathcal{D}))$$

$$p[r \in Q(\mathcal{D})] > 0 \Leftrightarrow r \in \mathbf{possible}(Q(\mathcal{D}))$$

Records which are certain (occur in every possible world) must have marginal probability 1 (the probabilities of all possible worlds have to sum up to 1), while records that are possible must have a nonzero probability (they occur in at least one possible world).

12.3 Uncertainty Encodings

Possible worlds provide us with a convenient, intuitive model for uncertain data processing that is independent on the choice of language for expressing analysis tasks. However, the full set of possible worlds (and the corresponding probability distributions) may be extremely large (or even infinite). This may not be obvious from our running example, but observe that in our example there was only one missing value to impute and we only had two possible replacements for this missing value. Consider a moderately sized version of our running example dataset with 100k records and assume that there 100 missing `ownsProperty` values. Then there are 2^{100} possible worlds – one for each choice of values for these 100 missing values. To make analysis of uncertain data feasible or even to store such a large number of possible worlds, a more compact representation is needed. A variety of uncertain data encodings have been developed that compactly represent sets of possible worlds. We now focus our discussion exclusively on representations of collections (sets or bags) of records. Broadly, we categorize encodings of uncertain collection datasets into two classes: (1) *Lossless* encodings, which exploit

independence properties to factorize the set of possible worlds, and (2) *Sampled* encodings, which represent a finite subset of the possible worlds at the cost of information loss.

12.3.1 Lossless, Factorized Encodings for Incomplete Data

Independence can be exploited to create compact, factorized representations of uncertain data. For example, consider a record r_1 that is present in exactly half of the possible worlds of \mathcal{D} , where the remaining possible worlds (modulo r_1) are identical. Writing \mathcal{D}' to denote the possible worlds without r_1 ($\mathcal{D}' = \{ D \mid D \in \mathcal{D}, r_1 \notin D \}$), we can call record r_1 independent iff

$$\mathcal{D} = \mathcal{D}' \cup \{ \{r_1\} \cup D' \mid D' \in \mathcal{D}' \}$$

Intuitively, the above condition checks that the choice of including r_1 or not has no effect on the inclusion of any other possible records. A natural consequence of r_1 's independence is that the possible worlds encoding of \mathcal{D} need to store two full copies of \mathcal{D}' . Lossless encodings exploit such redundancy caused by independence to create more compact representations where this redundancy is factorized out. We will outline three lossless encodings: Tuple-Independent Databases, Disjoint-Independent Databases, and C-Tables, each a more expressive generalization of the previous one. For instance, in our running example, the record encoding that Peter's loan is granted appears in both possible worlds while Alice's record, even though it also appears in both possible worlds, occurs with different `ownsProperty` values. Based on this observation a more compact representation of D_1 and D_2 from our running example is as a single database containing both Peter's and Alice's record, but to record separately that there are two options (yes and no) for the `ownsProperty` attribute value for Alice's record.

12.3.1.1 Tuple-Independent Databases

We first consider an encoding that is only applicable if the following simple, but strong condition holds: all records are independent of one another. An incomplete database that satisfies this constraint is called a tuple-independent incomplete database [4, 40]. Note that *tuple* is the formal term used to denote records in a database. A tuple-independent incomplete database can be represented as a collection of records $\mathbb{D}_{TI} \in \text{dom}(r) \times \mathbb{B}$, where each record is annotated with a boolean attribute that indicates whether (or not) it is certain. Note that here $\text{dom}(r)$ denotes the domain of records, i.e., the set of all possible records. That is:

$$\mathbf{certain}(\mathbb{D}_{TI}) = \{ r \mid \langle r, \text{TRUE} \rangle \in \mathbb{D}_{TI} \}$$

$$\mathbf{possible}(\mathbb{D}_{TI}) = \{ r \mid \langle r, \text{FALSE} \rangle \in \mathbb{D}_{TI} \}$$

The incomplete database \mathcal{D} represented by \mathbb{D}_{TI} is defined as the set of all databases that are subsets of the set of all possible records and supersets of the set of certain records.

$$\mathcal{D} = \{ D \mid D \subseteq \text{possible}(\mathbb{D}_{TI}) \wedge D \supseteq \text{certain}(\mathbb{D}_{TI}) \}$$

The second requirement ($D \supseteq \text{certain}(\mathbb{D}_{TI})$) is based on the fact that certain records appear in every possible world. Thus, any possible world represented by \mathbb{D}_{TI} is a superset of the set of certain records as well.

Note that this type of factorization cannot be used to compress the incomplete database from our running example, because the two versions of Alice's record are not independent of each other (a possible world including the tuple recording that Alice owns property cannot also contain the tuple recording that Alice does not own property).

Example 7 To illustrate tuple-independent databases consider a modified version of our running example where Alice's record only exists if she owns property. This modified version of our running example can be represented as a tuple-independent incomplete database (Peter's record is certain while the tuple storing that Alice does own property is only possible) as shown below:

SSN	name	grantLoan	isCertain
777-777-7777	Alice Alison	yes	—
111-111-1111	Peter Peterson	yes	TRUE

When using probabilities, we make a similar assumption that the probabilities of individual tuples are independent. Accordingly, for any probabilistic database $\langle \mathcal{D}, P \rangle$ fulfilling this condition, we can define a tuple-independent encoding $\mathbb{D}_{TIP} \in \text{dom}(r) \times [0, 1]$ by annotating each tuple with its probability instead of a boolean value:

$$\mathbb{D}_{TIP} = \{ \langle r, P(r \in \mathcal{D}) \rangle \mid r \in \text{possible}(\mathbb{D}_{TI}) \}$$

In the probabilistic representation, certain answers are those records with a probability of 1, and possible answers are any records with a nonzero probability.

$$\text{certain}(\mathbb{D}_{TIP}) = \{ r \mid \langle r, 1.0 \rangle \in \mathbb{D}_{TIP} \}$$

$$\text{possible}(\mathbb{D}_{TIP}) = \{ r \mid \langle r, p \rangle \in \mathbb{D}_{TIP} \wedge (p > 0) \}$$

Observe that a tuple-independent incomplete or probabilistic database with n tuples represents up to 2^n possible worlds: for each tuple we can either choose the tuple to be included in the possible world or not and all these choices are independent of

each other. There are n boolean decisions resulting in 2^n possible options. That is, a tuple-independent incomplete or probabilistic database can be exponentially more concise than its possible world representation.

12.3.1.2 Disjoint-Independent Databases

The tuple-independent model assumes that records are entirely independent but says nothing about the contents of those records. It is assumed that each record is identical in all possible worlds where it appears. The disjoint-independent model of incomplete databases (sometimes called x-tuples) generalizes tuple-independent databases by allowing a record to take multiple forms in different possible worlds. Under this model, an x -tuple \mathbf{r} is simply a set of records $\{r_1, \dots, r_N\}$ called its *instantiations*. We say that \mathbf{r} is disjoint-independent in database \mathcal{D} iff we can define a subset of the possible worlds $\mathcal{D}' = \{D \mid (D \in \mathcal{D}) \wedge (\mathbf{r} \cap D = \emptyset)\}$ that do not contain elements of \mathbf{r} , such that \mathcal{D} can be defined as the cartesian product of \mathbf{r} and \mathcal{D}' . That is, \mathbf{r} is disjoint independent if:

$$\mathcal{D} = \{D \cup \{r\} \mid D \in \mathcal{D}' \wedge r \in \mathbf{r}\} \text{ or } \mathcal{D} = \{D \cup \{r\} \mid D \in \mathcal{D}' \wedge r \in \mathbf{r}\} \cup \mathcal{D}'$$

Note the two definitions: In the former case, *some* instantiation of \mathbf{r} appears in all possible worlds, while in the latter some set of possible worlds \mathcal{D}' do not contain any instantiation of \mathbf{r} . Accordingly, in the former case \mathbf{r} is certain, while in the latter it is merely possible. Note that if an x-tuple \mathbf{r} is certain in a disjoint-independent database, this x-tuple may have different values in different possible worlds. Applying our previous definition of certainty which requires that a record r appears in all possible worlds to disjoint-independent databases, a record r is certain if there exists an x-tuple $\mathbf{r} = \{r\}$.

Observe that disjoint-independent databases generalize tuple-independent databases in the following sense, a tuple-independent database can be modeled as a disjoint-independent database where each x-tuple has a single instantiation $\mathbf{r} = \{r\}$ for certain records and $\mathbf{r} = \{r, \perp\}$ for records that are not certain. However, the opposite direction does not hold, and there are disjoint-independent databases that cannot be represented as tuple-independent incomplete databases, e.g., a database with two possible worlds $D_1 = \{r_1\}$ and $D_2 = \{r_2\}$ where $r_1 \neq r_2$ can be encoded using a single x-tuple $\mathbf{r} = \{r_1, r_2\}$. However, this incomplete database cannot be encoded as a tuple-independent database.

As before, we define an encoding $\mathbb{D}_{DI} \in 2^{2^{\text{dom}(r) \cup \{\perp\}}}$ for any database known to contain exclusively disjoint-independent records. Specifically, the encoding is a collection of x-tuples (sets of records). The presence of a special distinguished value \perp in an x-tuple indicates that the x-tuple is not certain. Hence, the incomplete database \mathcal{D} corresponding to \mathbb{D}_{DI} is defined as follows (using \prod to denote a cartesian product of sets).

$$\mathcal{D} = \prod_{\mathbf{r} \in \mathbb{D}_{DI}} \{ \hat{r} \mid r \in \mathbf{r} \} \quad \text{where} \quad \hat{r} = \begin{cases} \{\} & \text{if } r = \perp \\ \{r\} & \text{otherwise} \end{cases}$$

Example 8 Our original running example database can be encoded as a disjoint-independent database as shown below consisting of two x-tuples \mathbf{r}_1 and \mathbf{r}_2 where \mathbf{r}_1 (Peter's record) is certain and has one instantiation r_{1_1} while \mathbf{r}_2 has two instantiations (Alice owns property or not) r_{2_1} and r_{2_2} :

record	instantiation	SSN	name	grantLoan
r_1	r_{1_1}	111-111-1111	Peter Peterson	yes
r_2	r_{2_1}	777-777-7777	Alice Alison	yes
	r_{2_2}	777-777-7777	Alice Alison	no

The disjoint-independent encoding of incomplete databases can be extended to also support probabilistic databases. A disjoint-independent probabilistic database $\mathbb{D}_{DIP} : 2^{2^{\text{dom}(r)}} \rightarrow (2^{\text{dom}(r)} \rightarrow [0, 1])$ is a function that map every possible x-tuple \mathbf{r} (a subset of $\text{dom}(r)$) to a probability mass function over \mathbf{r} (a function associating a probability from $[0, 1]$ with individual instantiations of x-tuple \mathbf{r}). Note that this does not mean that we require that every possible x-tuple \mathbf{r} exists. For nonexisting x-tuples, we would set the probability of all of its instantiations r_i to 0. While the domain $\text{dom}(r)$ of the function \mathbb{D}_{DIP} that encodes a disjoint-independent probabilistic database may be infinite, \mathbb{D}_{DIP} is finitely representable as long as each possible world is finite, which is typically assumed to be the case. A finite representation is achieved by only recording the output of function \mathbb{D}_{DIP} for input records with a total probability mass larger than 0. We use $p_{\mathbf{r}}$ to denote the probability distribution associated by \mathbb{D}_{DIP} to an x-tuple \mathbf{r} , i.e., $p_{\mathbf{r}} = \mathbb{D}_{DIP}(\mathbf{r})$. Observe that, we have eliminated the need for a distinguished element \perp to denote the x-tuple's absence by allowing its probability mass function to sum to less than 1. That is, we define:

$$P[\mathbf{r} \in \mathcal{D}] = \sum_{r \in \mathbf{r}} p_{\mathbf{r}}(r) \quad \text{and} \quad P[\mathbf{r} = r] = p_{\mathbf{r}}(r)$$

These probability mass functions give us natural definitions for both certain and possible records:

$$\begin{aligned} \mathbf{certain}(\mathbb{D}_{DIP}) &= \left\{ \mathbf{r} \mid \sum_{r \in \mathbf{r}} p_{\mathbf{r}}(r) = 1 \right\} \\ \mathbf{possible}(\mathbb{D}_{DIP}) &= \left\{ \mathbf{r} \mid \sum_{r \in \mathbf{r}} p_{\mathbf{r}}(r) > 0 \right\} \end{aligned}$$

Again, for a record r to be certain in the sense we defined for possible world semantics, the singleton $\mathbf{r} = \{r\}$ has to be mapped to the probability distribution $p_{\mathbf{r}}(r) = 1$ by \mathbb{D}_{DIP} .² While the representation of an incomplete database as a tuple-independent database is unique, this does not hold for disjoint-independent database since the same instantiation may occur in different tuples.

12.3.1.3 C-Tables

Like the tuple-independent model, the disjoint-independent model cannot capture arbitrary correlations between records. Originally proposed by Imielinski and Lipski, the C-Tables model [21] allows incomplete databases to be factorized, without requiring that their records be independent or otherwise uncorrelated. We use \mathbb{V} to denote an alphabet of variable symbols $v \in \mathbb{V}$. A C-Tables \mathbb{D}_C is a collection of records $\langle r, f_r(v_1, \dots, v_N) \rangle \in \mathbb{D}_C$, where every record r is annotated with a boolean expression f_r over variables $v_1 \dots v_N \in \mathbb{V}$. This expression is sometimes termed a *local condition*. The set of possible worlds defined by a C-table is based on assignments $\alpha : \mathbb{V} \rightarrow \mathbb{B}$ of boolean values to each variable. We use \mathcal{A} to denote the set of all such assignments. Specifically, for each assignment $\alpha \in \mathcal{A}$, there exists a possible world D_α defined by this assignment as the set of records in \mathbb{D}_C that are annotated with an expression that evaluates to true after replacing variables with their values assigned by α .

$$D_\alpha = \{ r \mid \langle r, f_r(v_1, \dots, v_N) \rangle \in \mathbb{D}_C \wedge f(\alpha(v_1), \dots, \alpha(v_N)) \}$$

Given some boolean expression $F(v_1, \dots, v_K)$, termed the *global condition*,³ the full set of possible worlds is defined by the set of assignments that cause F to evaluate to true:

$$\mathcal{D} = \{ D_\alpha \mid \alpha \in \mathcal{A} \wedge F(\alpha(v_1), \dots, \alpha(v_K)) \}$$

C-tables are more expressive than the tuple-independent and disjoint-independent databases. In fact, any finite set of possible worlds can be encoded as a C-table.⁴

²The reader may wonder whether it is possible to encode a certain record r as multiple x -tuples that all have r as an instantiation and where for each such x -tuple \mathbf{r} , we have $p_{\mathbf{r}}(r) < 1$. However, recall that x -tuples are assumed to be independent of each other. Thus, there would exist a possible world with a nonzero probability that does not contain r constructed by choosing an instantiation $r' \neq r$ or no instantiation for every x -tuple \mathbf{r} with $r \in \mathbf{r}$.

³Note that global conditions are not strictly necessary for expressive power, but they may allow for a more compact/convenient representation of a probabilistic database.

⁴Consider an incomplete database \mathcal{D} with 2^n possible worlds $D_1 \dots D_{2^n}$. (the construction has to be modified slightly if the number of possible worlds is not a power of 2). Then we use n variables: v_1, \dots, v_n . An assignment to these variables is interpreted as a number i in binary identifying one possible world D_i . For example, if there are $4 = 2^2$ possible worlds, then we would use

Probabilistic C-Tables This representation admits a straightforward extension to the probabilistic case, originally proposed by Green et al. [16]. This approach defines a probability distribution $P : \mathcal{A} \rightarrow [0, 1]$ over the space of assignments.⁵ A probabilistic C-Table (or PC-Table) is defined as a pair of database and probability distribution $\mathbb{D}_{PC} = (\mathbb{D}_C, P)$. Hence, the probability of a database and record can be defined as typical for possible worlds semantics:

$$P[D_\alpha \in (\mathbb{D}_C, P)] = P(v) \quad P[r \in (\mathbb{D}_C, P)] = \sum_{\alpha \in \mathcal{A}: r \in D_\alpha} P(\alpha)$$

The distribution P can be encoded using any standard approach for compactly encoding multivariate distributions, such as a graphical model [34].

Example 9 Continuing the running example, we can model the analysis result as a C-Table. There is one uncertain decision that affects the set of possible worlds: Whether or not Alice owns property. We define a single boolean variable v_1 to denote the outcome of this decision. Records in the C-Table encoding the result are annotated with boolean expressions ϕ over $\mathbb{V} = \{v_1\}$:

Query Result (Simple C-Table)			
SSN	name	grantLoan	ϕ
777-777-7777	Alice Alison	no	v_1
777-777-7777	Alice Alison	yes	$\neg v_1$
111-111-1111	Peter Peterson	yes	T

The two possible assignments $\{v_1 \mapsto \mathbf{T}\}$ and $\{v_1 \mapsto \mathbf{F}\}$ define the two possible worlds. A separately provided (joint) distribution over the variable(s) in \mathbb{V} assigns a probability to each possible world.

$$p(\alpha) = \begin{cases} 0.6 & \text{if } \alpha = \{v_1 \mapsto \mathbf{T}\} \\ 0.4 & \text{if } \alpha = \{v_1 \mapsto \mathbf{F}\} \end{cases}$$

Each possible world contains only records with boolean expressions that are true under the corresponding assignment. Hence the first two rows (with conditions v_1 and $\neg v_1$, respectively) are mutually exclusive.

two variables v_1 and v_2 , and the assignment $v_1 \mapsto \mathbf{T}$ and $v_2 \mapsto \mathbf{F}$ represents the possible world $1 \cdot 2^1 + 0 \cdot 2^0 = 2$. The database constructed contains all records that are possible in \mathcal{D} . For an assignment α , let $n(\alpha)$ denote the number encoded by α . Then the local condition for record r is

$$\bigvee_{\alpha: r \in D_n(\alpha)} \bigwedge_{j: \alpha(v_j) = \mathbf{T}} v_j.$$

⁵Note that [16] used per variable distributions which is less general.

Non-Boolean Variables and Assignments For C-Tables to efficiently model a disjoint-independent database, it is necessary to create local conditions that alternate between mutually exclusive options. Such conditions can be modeled with boolean formulas, as in Example 9.⁶ However, it is often both convenient and more efficient to express alternatives with a single integer- or real-valued variable. In this form, records are still annotated with boolean expressions, albeit over comparisons ($=$, $<$, \leq , etc.) over variables.

Example 10 Continuing the example, we could express the same result using integer-valued variables. The result C-Table and corresponding distribution are as follows:

Query Result (Integer-Valuation C-Table)			
SSN	name	grantLoan	ϕ
777-777-7777	Alice Alison	no	$(v_1 = 1)$
777-777-7777	Alice Alison	yes	$(v_1 = 2)$
111-111-1111	Peter Peterson	yes	T

$$p(\alpha) = \begin{cases} 0.6 & \text{if } \alpha = \{ v_1 \mapsto 1 \} \\ 0.4 & \text{if } \alpha = \{ v_1 \mapsto 2 \} \end{cases}$$

Taking the process even further, we can replace attributes with placeholders (often called “labeled” nulls or Skolem terms) indicating that their value is to be given by the valuation. The result is an even more compact representation, as records that were previously conditionally in the result can now be treated as certain.

Example 11 Having two assignments with nonzero probability, $\{ v_1 \mapsto 'no' \}$ and $\{ v_1 \mapsto 'yes' \}$, we can replace the result C-table as follows:

Query Result (General C-Table)			
SSN	name	grantLoan	ϕ
777-777-7777	Alice Alison	v_1	T
111-111-1111	Peter Peterson	yes	T

Observe that the `grantLoan` attribute of Alice’s record has been replaced by a placeholder. This value gets filled in by the assignment that defines each possible world.

This generalized form of assignments and the use of variable-valued attributes were originally proposed by Imielinski and Lipski as part of the original C-Tables formalism [21]. It has been used successfully by several systems, most notably ORCHESTRA [18, 22] and Pip [26]. In fact, as we will discuss in Sect. 17, Pip [26]

⁶Note that more than two options can be modeled by multiple boolean variables. For example, four alternatives can be modeled with annotations $v_1 \wedge v_2$, $\neg v_1 \wedge v_2$, $v_1 \wedge \neg v_2$, and $\neg v_1 \wedge \neg v_2$, respectively.

further generalizes this model by allowing symbolic expressions (formulas) over variables as attribute values.

12.3.1.4 U-Relations and World-Set Decompositions

Antova et al. proposed a more backward-compatible implementation of C-Tables called U-Relations [20]. A U-Relation is a database table that encodes a C-Table under the following restrictions: (1) No variable-valued attributes. (2) Local conditions must be pure conjunctions. (3) Atoms of local conditions must be equality comparisons between a variable and a value. To support arbitrary Boolean expressions, multiple copies of a record are allowed in the U-Relation, each annotated by a different local condition. When the full expression is needed, copies of each record are grouped and their local conditions are combined by disjunction (into disjunctive normal form).

To encode a U-Relation in a classical relational database, we first determine the maximal number of conjunctive clauses in any record. We then add twice as many integer-valued annotation fields to the record. Half are used to identify variables, while the other half identify the values.

Example 12 To encode the C-Table from Example 10, we would first see that there is at most one conjunctive atom in any local condition in the record. We add two new annotation attributes to the record: `var1` identifying the variable and `val1` identifying the corresponding value in the equality.

Query Result (U-Relation)				
SSN	name	grantLoan	var1	val1
777-777-7777	Alice Alison	no	1	1
777-777-7777	Alice Alison	yes	1	2
111-111-1111	Peter Peterson	yes	0	0

The special variable v_0 is always equal to 0, so unused fields can be filled in by the tautological expression $v_0 = 0$, as in the third row of the U-Relation.

World-Set Decompositions Without variable-valued attributes, U-Relations introduce significant redundancy as attributes that do not vary between possible worlds are still repeated. To mitigate this redundancy, Antova et al. propose [3] using a columnar data layout [38] in a strategy that they call world-set decompositions. Specifically, each record is assigned a unique identifier (e.g., a key attribute or database ROWID), while columns are stored independently.

Example 13 Using world-set decompositions with `SSN` as a row identifier, the query result from the prior example would be decomposed into two separate tables:

Observe that there is no uncertainty in the decomposed Name table, and there are no longer two copies of Alice's name being stored.

Query Result.Name		Query Result.grantLoan			
SSN	name	SSN	grantLoan	var1	val1
777-777-7777	Alice Alison	777-777-7777	no	1	1
111-111-1111	Peter Peterson	777-777-7777	yes	1	2
		111-111-1111	yes	0	0

12.3.2 Lossy, Sampling-Based Encodings for Incomplete Data

Certain types of analysis – what are called unsafe queries [7] – cannot be performed both efficiently and correctly on lossless encodings of probabilistic data. In such cases, results can be approximated by using Monte-Carlo methods. The most straightforward way to accomplish this is to select some finite set of samples $\hat{\mathcal{D}} \subset \mathcal{D}$ from the set of possible worlds, uniformly for incomplete databases or according to $P(D = \mathcal{D})$ for probabilistic databases. Analytical questions are evaluated in parallel on all possible worlds from the sample set:

$$\hat{Q}(\mathcal{D}) = \left\{ Q(D) \mid D \in \hat{\mathcal{D}} \right\}$$

Because of the sampling process, sampling-based encodings do not distinguish between incomplete and probabilistic databases. However, we observe that many statistical measures that might be computed over the set of results (e.g., the expectation) have no meaning for incomplete databases. We introduce two approaches to encoding sets of samples: (1) World-Annotated databases, which admit a more computationally efficient implementation using classical relational databases, and (2) Tuple Bundles, which encode uncertain data more compactly.

Note that both **certain** and **possible** are ill defined on samples. By definition the certain records are a subset of records across all samples, and the possible records are a superset.

$$\text{certain}(\hat{\mathcal{D}}) \subseteq \bigcap_{D \in \hat{\mathcal{D}}} D \quad \text{possible}(\hat{\mathcal{D}}) \subseteq \bigcup_{D \in \hat{\mathcal{D}}} D$$

However, these are only bounds on the sets of certain and possible records.

12.3.2.1 World-Annotated Sample Sets

To store a set of N samples, our first, naive approach creates a single database $\mathbb{D}_{WA} \in \text{dom}(r) \times [1, N]$, annotating each record with the index of the sample it appears in. Accordingly, the sample set is defined by demultiplexing the records.

$$\hat{\mathcal{D}} = \{ \{ r \mid \langle r, i \rangle \in \mathbb{D}_{WA} \} \mid i \in [1, N] \}$$

12.3.2.2 Tuple Bundles

The size of \mathbb{D}_{WA} is generally linear in the number of samples (i.e., $O(N)$). Unsurprisingly, the computational cost of analysis typically scales linearly as well. As with lossless encodings, eliminating redundancy can create more compact and efficient representations. One approach to eliminating redundancy is a type of record called a tuple bundle, originally proposed by Jampani et al. [23]. We assume that a record $r = \langle a_1, \dots, a_K \rangle$ is defined by k attribute values a_i . Accordingly, a tuple bundle $\mathbf{r} = \langle \mathbf{a}_1, \dots, \mathbf{a}_K, \phi \rangle$ is defined by a set of attributes \mathbf{a}_j and a sample-vector ϕ . Each attribute may either be a single value or a vector of size N .

$$\mathbf{a}_j = \begin{cases} a_j \\ (\langle a_{1,j}, \dots, a_{N,j} \rangle) \end{cases}$$

In the first case, the value a_j is constant across all samples, while the latter case defines explicit values for the attribute $a_{i,j}$ in each sample. The sample vector $\phi \in \mathbb{B}^N$ is a vector of B boolean values (bits) $\phi[i]$. Bit j being set to true (resp., false) indicates that the record is present in (resp., absent from) sample j . The corresponding sample set is defined by filtering on ϕ and plugging in attribute values.

$$\hat{\mathcal{D}} = \{ \{ \langle a_{i,1}, \dots, a_{i,K} \rangle \mid ((\mathbf{a}_1, \dots, \mathbf{a}_K, \phi) \in \mathbb{D}_{TB}) \wedge \phi[i] \} \mid i \in [1, N] \}$$

$$\text{where } a_{i,j} = \begin{cases} a_j & \text{if } \mathbf{a}_j = a_j \\ a_{i,j} & \text{otherwise} \end{cases}$$

12.4 Computing with Uncertain Tabular Data

Assume that we are given an encoding \mathbb{D} (resp., \mathbb{D}_P) that corresponds to an incomplete (resp., probabilistic) database \mathcal{D} (resp., (\mathcal{D}, P)). We want to compute the answer to a question $Q(\mathcal{D})$. However, answering this question directly on \mathcal{D} using possible worlds semantics is impractical, as the number of worlds is usually large. In this section we discuss techniques for computing answers more efficiently by directly manipulating the encodings \mathbb{D} .

12.4.1 Relational Algebra

Queries over tabular data are expressed through a range of different languages: SQL, SparQL, R, Spark, and others. To streamline our discussion, we focus on a tabular data processing language called relational algebra. Relational algebra is comparatively straightforward to reason about and also captures the core data

Operator	Notation	SQL
Table	R	<code>SELECT [DISTINCT] * FROM R;</code>
Projection	$\pi_{A,B,\dots}(R)$	<code>SELECT A, B, ... FROM R;</code>
Selection	$\sigma_\psi(R)$	<code>SELECT * FROM R WHERE \psi;</code>
Product	$R \times S$	<code>SELECT * FROM R, S;</code>
Union	$R \cup S$	<code>SELECT * FROM R UNION [ALL] SELECT * FROM S;</code>
Aggregate	$\gamma_{A,\dots,M \leftarrow \alpha_M,\dots}(R)$	<code>SELECT A, ..., \alpha_M AS M, ... FROM R GROUP BY A, ...;</code>

Fig. 12.3 Relational algebra

manipulation functionality of each of these other languages. Before we discuss evaluation techniques for uncertain data, we first present a short overview of normal relational algebra.⁷ We then introduce strategies for evaluating relational algebra expressions over encoded uncertain databases. We follow the focus on probabilistic databases exhibited by most of the work on uncertain databases but also note when probabilistic techniques apply to incomplete databases as well. We also focus on lossless encodings, as query evaluation over lossy encodings is a straightforward extension of classical query evaluation [23].

Relational algebra concerns itself with sets (resp., bags) of records called tables (R, S, T, \dots). An individual record $r \in R$ is a set of attribute/value pairs $r = \langle A : v_A, B : v_B, \dots \rangle$, and we assume that all records in a table have identical sets of attributes. We refer to this set of attributes as the table's schema $sch(R)$. Relational algebra, as we use it, defines six operators, summarized in Fig. 12.3: Input Tables, Projection, Selection, Product, Union, and Aggregation. Apart from the table operator, each operator takes the output of one or more other operators as input and produces an output that may be saved as a table or passed to another operator. Hence, operators can be linked together to express complex computations.

Projection transforms each record of its input, producing records with attributes given by a set of target columns (A, B, \dots). When working with sets, projection also ensures that the output is free of duplicates.

Selection filters its input down to records that satisfy the condition ψ .

Product pairs every record in one of its inputs with every record in the other. The combination of Selection and Product operators ($\sigma_\psi(R \times S)$) is often called a Join ($R \bowtie_\psi S$).

Union merges records from two input tables. If working with sets, Union also ensures that there are no duplicates in the outputs.

(Group-By) Aggregation creates groups of records according to a list of attributes (A, B, \dots). Records in each group are summarized by one or more aggregate functions ($\alpha_i \in \{ \text{SUM}, \text{COUNT}, \text{AVG}, \text{MIN}, \dots \}$), and one record per group is returned. If no grouping attributes are given, aggregation treats its entire input as a single group.

⁷For a more thorough introduction, we refer the interested reader to a textbook by Garcia-Molina et al. [14].

Original Operator	Probabilistic Implementation
$\pi_{A,B,\dots}(R)$	<code>SELECT A, B, ..., 1 - PROD(1-ϕ) FROM R GROUP BY A, B, ...;</code>
$\sigma_\psi(R)$	<code>SELECT * FROM R WHERE ψ;</code>
$R \times S$	<code>SELECT *, R.ϕ * S.ϕ AS ϕ FROM R, S;</code>
$R \cup S$	<code>SELECT *, 1 - PROD(1-ϕ) FROM (SELECT * FROM R UNION SELECT * FROM S) GROUP BY *;</code>

Fig. 12.4 Extensional evaluation implemented in SQL

12.4.2 Extensional Evaluation

We first consider the tuple-independent model [4, 39, 40]. Recall that a tuple-independent probabilistic database is encoded by annotating each record with an extra field $\phi \in [0, 1]$ denoting the independent probability of this record being in any given possible world. Naively, we might try modifying relational algebra operators to preserve these annotations, a strategy called “Extensional” evaluation [39, 40] of relational algebra. That is, for each operator, we define an evaluation strategy that ensures that each output row is annotated with the independent probability of the result. Figure 12.4 illustrates this strategy for queries evaluated over sets. We next discuss these operators – we omit Aggregation for the moment.

Projection For projection, we eliminate duplicate rows using a group-by query.

Each resulting record exists if any records that share projection attributes exist. Assuming that each row in the input is independent, the corresponding probability is a disjunction of independent events $(1 - (1 - p(t_1)) \cdot (1 - p(t_2)) \dots)$.

Selection Selection has no impact on probabilities of records. Records that are filtered out are excluded from the result regardless of their probability. Records that are not filtered out appear in the result with their original probabilities.

Product For a row to appear in the output of a product, it must have resulted from one row in each of the product’s inputs. Assuming that the inputs are independent, the probability of each output row can be computed as a conjunction of two independent events $(p(t_1) \cdot p(t_2))$.

Union Union itself does not affect the probability of its inputs. However, during duplicate removal, union may need to merge record probabilities. It does this in the same way as during duplicate removal for projection.

Modifying a relational algebra expression to maintain the probability annotation attribute ϕ through extensional evaluation adds minimal computational overhead. However, extensional evaluation has several serious limitations that all stem from its use of the tuple-independent model to represent state in between operators. First, the tuple-independent model cannot efficiently represent the outputs of the aggregate operator: The size of the output grows exponentially with the size of the input. Second, even if the input to a relational algebra expression is independent, the expression may introduce correlations between rows of output.

Example 14 Continuing our running example, consider the loan approval table and another table of homes available for purchase scraped from websites and government records. Scraping is an imprecise process, and the record at *45 Bassett* may not actually represent a home available for purchase.

possible(forSale)		
address	price	ϕ
123 Acacia	200k	1.0
45 Bassett	150k	0.9

possible($Q(\mathcal{D}) \times \text{forSale}$)					
SSN	name	grantLoan	address	price	ϕ
777-777-7777	Alice Alison	yes	123 Acacia	200k	0.4
777-777-7777	Alice Alison	yes	45 Bassett	150k	0.36
777-777-7777	Alice Alison	no	123 Acacia	200k	0.6
777-777-7777	Alice Alison	no	45 Bassett	150k	0.54
111-111-1111	Peter Peterson	yes	123 Acacia	200k	1.0
111-111-1111	Peter Peterson	yes	45 Bassett	150k	0.9

Consider the product of this new table with the result table from our running example (obtained via Extensional evaluation). There are two types of correlations in the result records. The presence of the second, fourth, and sixth possible records depends on whether or not the record for *45 Bassett* is present in *forSale*. Meanwhile the first two records must always appear together and are mutually exclusive with the third and fourth result records. The records are not independent, and the measure ϕ annotating each record can no longer be used to compute the probability of worlds containing the record. Note, however, that for this example the probabilities annotating each record do correspond to the *marginal* probability of that record being in the result.

Relational algebra introduces correlations, and program outputs are not guaranteed to be tuple-independent. Thus, it is possible that the resulting probability annotations will be meaningless. Fortunately that is not always the case. Dalvi and Suciu identified a particular class of relational algebra expressions termed “safe” [7, 39], as well as a procedure for (1) rewriting expressions into equivalent safe forms or (2) determining that there is no equivalent safe expression. When a safe expression is evaluated using the extensional rules, every output record will be annotated with the record’s confidence (marginal probability of being in the result). Extensional evaluation also has value for unsafe expressions. As shown by Gatterbauer and Suciu [15], extensional evaluation can be used to establish bounds on the actual confidence values.

12.4.3 Intensional Evaluation

For computing with unsafe relational algebra expressions, we need an evaluation strategy that takes into account potential inter-row correlations. Our next approach, called “Intensional” evaluation, uses C-Tables as an underlying data representation. Recall that records in C-Tables are annotated with a boolean formula ($f_r(v_1, \dots, v_N)$), parameterized by variable symbols (v_i). Possible worlds are defined by assignments of values to variables; Records are included in a result in worlds that assign values that satisfy the boolean expression.

Intensional query evaluation [21, 39] is closely related [17] to the provenance (i.e., lineage or pedigree) of query answers. Under intentional evaluation, each operator annotates output records with the conditions that need to hold on the assignment for the record to be in the result. Hence, correlations are explicitly captured in the query results as variables that appear repeatedly in a formula or across the annotations of multiple records.

For Intensional evaluation (a possible implementation is shown in Fig. 12.5), operators follow a virtually identical pattern to extensional evaluation, except that the resulting annotation ϕ is a boolean formula. In the probabilistic database literature, this type of formula is often called Lineage. Once the result is computed, the problem of computing marginal probabilities becomes one of simple inference. For each tuple, we are given boolean formula and a distribution over binary variables appearing in the formula. We need to compute the marginal probability of this boolean formula being true.⁸

The problem of inference has been well studied in the general context [28]. The specific problem of computing marginals under constraints belongs to the general problem of counting solutions to boolean formulas. Exact solutions are exponential in the size of the input (complexity class #P [33]), and numerous approximation schemes have been developed. However, probabilistic databases admit several specializations of general techniques. We next discuss several of these.

Original Operator	Probabilistic Implementation
$\pi_{A,B,\dots}(R)$	<code>SELECT A, B, ..., BOOLEAN_OR(ϕ) FROM R GROUP BY A, B, ...;</code>
$\sigma_\psi(R)$	<code>SELECT * FROM R WHERE ψ;</code>
$R \times S$	<code>SELECT *, BOOLEAN_AND(R.ϕ, S.ϕ) AS ϕ FROM R, S;</code>
$R \cup S$	<code>SELECT *, BOOLEAN_OR(ϕ) FROM (SELECT * FROM R UNION SELECT * FROM S) GROUP BY *;</code>

Fig. 12.5 Intensional evaluation implemented in SQL

⁸Observe that a binary version of this problem can be applied in the case of incomplete databases. A tuple is certain if its local condition is implied by the global condition.

KLM Estimators It is well known that any relational algebra expression which exclusively uses the operators we have introduced here can be rewritten into a normal form which consists of a *union of conjunctive queries* (or *UCQ* for short). A *conjunctive query* (*CQ*) is a query without union which consists of a projection over the result of a selection which in turn is applied to the result of zero or more cross-products. After such a rewriting, it is trivial to see that Boolean formulas annotating results are guaranteed to be in disjunctive normal form, because the final union will connect the formulas produced by individual CQs through OR, while the formulas produced by a CQ are conjunctions (hence the name). As observed by Olteanu et al. [32], this makes a form of Gibbs sampling proposed by Karp, Luby, and Madras [24] ideal for probabilistic databases. The KLM scheme begins with a disjunction $C_1 \vee C_2 \vee \dots \vee C_N$ of conjunctive clauses C_i . It initially assumes that each conjunctive clause is disjoint:

$$p(C_1 \vee C_2 \vee \dots \vee C_N) = p(C_1) + p(C_2) + \dots + p(C_N)$$

This assumption is an overestimate, as variable assignments that satisfy two (or more) of the conjunctive clauses are counted twice (or more). The scheme then attempts to derive a corrective factor by repeatedly sampling clauses at random and computing the expected number of clauses that a satisfying assignment for the clause would also satisfy. An approach by Dagum et al. [6] improves on this approach by bounding the number of samples required to estimate record confidence within desired $\epsilon - \delta$ bounds.

Anytime Approximation Another distinction to be found in this setting is that data analytics are often interactive processes. The process of approximating confidence values can also be made interactive, allowing the analyst to decide on-the-fly when a result is “accurate enough” before terminating the process. One such approach, proposed by Olteanu et al. [13], alternates between using an approximate estimator like KLM and repeated refinement of the Boolean formula toward one consisting exclusively of disjoint clauses through Shannon expansion. Given enough time, this approach eventually converges to an exact value for a record’s confidence.

Top-K Estimation One particular specialization of probabilistic databases that is of interest is finding the most likely records in the output of a relational algebra expression [30, 35]. For example, given uncertain inputs describing existing findings of protein-protein interactions, we might wish to predict other likely interactions [10]. In general, this problem can be framed as the task of finding the K most probable records (the Top-K records) from the result. In this case, our computational job is easier, as exact probabilities are not required. We only need a sufficient approximation of each probability to decide whether the record belongs in the Top-K or not. One family of approaches proposes using early cutoffs in approximations [35, 37]. A related approach by Gatterbauer and Suciu uses intensional evaluation to establish bounds on the probability of a record [15], allowing for early cutoffs as well. A final approach by Li et al. [30] attacks a further specialization aiming at the “best” results. Here, the notion of “best” is formalized

by one of several different strategies for combining a user-provided ranking function over result records with the probability of records being in the result.

Attribute-Level Uncertainty Thus far, most of our discussions have centered around record-level uncertainty: the presence of a specific record with a specific set of attribute/value pairs in the result set. However in many situations, it is not the record that is uncertain but rather one of the values of an attribute of that record. For example, when computing an aggregate over an incomplete or probabilistic table, aggregate values are likely to be uncertain, while the groups to which they belong need not be. Although most work on probabilistic databases focuses on record-level uncertainty, several efforts have attempted to encode uncertainty appearing in attributes. The original Imielinski and Lipski formalization of C-Tables [21] allows variable symbols to appear in place of values in tabular data – variable assignments with non-boolean values as well are used to assign values to these variables. Notably, this means that selection can modify the formula annotating selected records. Singh et al. [36] allows attributes to take values defined by normal distributions. Antova et al. [3] propose a strategy that fragments records in tables, replacing the table with a set of tables, one per field. Finally, Kennedy et al. [25] builds on the C-Tables formulation to construct formulas for the values of uncertain attribute fields, which can be evaluated after variables are replaced by an assignment.

12.4.4 Virtual C-Tables

A recently proposed evaluation strategy based on C-Tables instead *virtualizes* uncertainty. This approach uses a generalization of C-Tables called *Virtual C-Tables* (or *VC-Tables* for short). We first introduce this generalization and then explain the evaluation strategy that utilizes it.

Recall that in a C-Table, the attribute values of a tuple are constants or variables from a set \mathbb{V} and the local condition of a tuple is a boolean formula over comparisons between variables and constants. While C-Tables are quite powerful, there still exist operators whose result are complicated or impossible to express as C-tables.

Example 15 Assume we want to require customers to pay an application fee for every loan application they make that is 1% of the requested loan amount and that we want to automatically accept loan applications if this fee is higher than \$10,000. Using the `applications` table (attributes `SSN` and `loanAmount`), we can determine the set of loans that will be automatically approved as follows:

```
SELECT SSN, loanAmount, loanAmount * 0.01 AS fee
FROM applications
WHERE loanAmount * 0.01 > 10000.0;
```

Consider the C-Table `applications` shown below where we do not know what the loan amount for the customer with `SSN 111-777-2222` which is encoded by setting the value of attribute `loanAmount` for this record to a variable, say v_1 .

SSN	loanAmount	ϕ
777-777-7777	200,000	T
111-777-2222	v_1	T

The loan for the customer with SSN 111-777-2222 will be automatically approved if $v_1 * 0.01$ (the fee) is larger than \$10,000. In this case the fee attribute of the result record has to be set of $v_1 * 0.01$. This correlation between the loanAmount and fee attribute is hard to express in a standard C-Table since there is no support for arithmetic operations. In fact, it can only be represented if the domain of the loanAmount is finite. In this case, we would have to represent each possible loanAmount and fee pair that fulfills the condition as a separate tuple. Some of these tuples are shown below.

SSN	loanAmount	fee	ϕ
111-777-2222	1,000,001	10,000.01	$v_1 = 1,000,001$
111-777-2222	1,000,002	10,000.02	$v_1 = 1,000,002$
111-777-2222	1,000,003	10,000.03	$v_1 = 1,000,003$
...

VC-Tables overcome this limitation of C-Tables by allowing attribute values and inputs to comparisons in local and global conditions to be symbolic expressions using arithmetic operators, conditionals (if-then-else), constants, and variables. Possible worlds are still defined over variable assignments. The only difference is that the attribute values of a tuple in a possible world are determined by evaluating the symbolic expressions of the tuple under the assignment α corresponding to the possible world. For details of the formal definition of these expressions, see [26, 41].

Example 16 Continuing with the previous example, we can compactly represent the query result as a VC-Table as follows:

SSN	loanAmount	fee	ϕ
111-777-2222	v_1	$v_1 * 0.01$	$v_1 * 0.01 > 10000$

For example, for the assignment $v_1 = 1,500,000$, we get the possible world:

SSN	loanAmount	fee
111-777-2222	1,500,000	15,000

We can extend VC-Tables to support probabilistic databases by defining a probability distribution over possible variable assignments.

Having introduced VC-Tables, we now explain how the Mimir PQP middleware [31, 41] uses such encodings to virtualizes PQP. Mimir rewrites probabilistic queries into deterministic queries over a deterministic encoding of uncertain data such that the rewritten queries faithfully preserve the semantics of the probabilistic queries. To accomplish this, Mimir uses an extended form of relational algebra called *variable-generating relational algebra* or *VG-RA* [25].

Using VG-RA, uncertainty is introduced into deterministic data through queries. A VG-RA expression defines VC-Tables by allowing expressions to generate variable symbols through special functions called variable generating or VG-Terms. A VG-Term denoted *VG(·)* can appear in any boolean or arithmetic expression in any Project, Select, or Aggregate operator in a VG-RA query. The input of a VG-Term controls the name of the variable generated by VG-RA for a given input record. For instance, a VG-Term *VG(name)* would return a unique variable for each *name* attribute value from the input of the operator where the VG-Term is used. The result of a VG-RA expression is an incomplete database – a VC-Table. Essentially, once a variable is introduced by a VG-Term, expressions involving this VG-Term are evaluated symbolically. VG-RA allows the generated variables to be associated with a (potentially joint) distribution over possible assignments (we do not show the language constructs for this here).

Example 17 Recall the *customers* table from Fig. 12.1. Although this table is missing several values, there is no uncertainty about this fact. Mimir allows users to create a cleaned “view” over the data; For the *customers* table, Mimir would use the following query:

```
SELECT SSN, name, assets, numMortgages,
    CASE WHEN income IS NULL THEN VG('income', RID)
        ELSE income END AS income,
    CASE WHEN ownsProperty IS NULL THEN VG('ownsProperty', RID)
        ELSE ownsProperty END AS ownsProperty
FROM customers;
```

For each of the two attributes with missing values, Mimir replaces **NULL** by using the SQL fragment

```
CASE WHEN x IS NULL THEN VG('x', RID) ELSE x
```

If the value is present, the fragment leaves it unchanged. If it is **NULL**, the fragment replaces it with a variable created by the VG Term *VG('x', RID)*, where RID uniquely identifies each row of the table. By being keyed on the attribute name, as well as a unique row identifier, one fresh variable is instantiated for every column and row. Independently, Mimir trains a model on the same data, uses interpolation, or any other imputation technique. The resulting models are then linked to the new variables. For instance, omitting the probability distributions over assignments, the result of the query shown above over the database from Fig. 12.1 (assuming an additional attribute **RID** as supported by many database systems) would be:

In most probabilistic databases, queries are assumed to be deterministic, and the data is nondeterministic (i.e., $\mathcal{Q}(\mathcal{D})$). Conversely, in Mimir the reverse is true, as all uncertainty is introduced through VG-Terms (i.e., as part of $\mathcal{Q}(\mathcal{D})$). In addition

RID	SSN	name	income	assets	numM	ownsP	ϕ
1	777-777-7777	Alice Alison	\$60,000	\$200,000	0	$v \cdot \text{ownsProperty}', 1$	T
2	333-333-3333	Bob Bobsen	\$102,000	\$15,000	0	no	T
3	111-111-1111	Peter Petersen	$v \cdot \text{income}', 3$	\$90,000	1	yes	T
4	555-555-5555	Arno Arnoldson	\$95,000	\$30,000	0	yes	T

to the other benefits in terms of compact representation of the result of arithmetic expression brought by VC-Tables, this allows an evaluation strategy to be chosen at query time. This is important in practice since different evaluation strategies and approximations may exhibit vastly different performances, and performance may be affected significantly by the structure of the query that is evaluated. Thus, allowing the strategy to be chosen per query is critical to trade performance against accuracy of the result using the techniques introduced in this section and Sect. 12.5.

12.5 Presenting Uncertain Tabular Data

We next survey techniques for presenting uncertain tabular data to users [29], develop a taxonomy of presentation strategies, and relate these strategies to algorithms for computation over uncertain data. That is, we discuss techniques that allow us to communicate to a user the set (resp., distribution) of possible worlds represented by an incomplete database \mathcal{D} (resp., probabilistic database (\mathcal{D}, P)) or an encoding \mathbb{D} thereof (resp., \mathbb{D}_P).

12.5.1 Tuple Identity

We have introduced two forms of uncertainty: record-level uncertainty (is a record part of a result or not) and attribute-level uncertainty (what is the value of a specific attribute). There is a tension between these two forms of uncertainty: At what point are the attributes of two records from different possible worlds sufficiently similar to be considered the same record? If they are considered the same record, then (with respect to the two possible worlds) we have one certain record with uncertain attributes. Conversely, if the records are different, then we have two uncertain records, with no uncertainty about their attributes.

To resolve this tension, different uncertainty management systems define – often indirectly – a record identity function $id(r)$. $id(r)$ assigns to every record an identifier that is unique within a possible world of the database \mathcal{D} . The primary role of identifiers is to gather instances of the same records from different possible worlds while allowing the precise definition of record “sameness” to vary based on the needs of the representation. Record identifiers allow us to define, for a particular possible world $D \in \mathcal{D}$, the set of identifiers appearing in the possible world. We

refer to the set of possible worlds that contain a tuple with the same identifier as a tuple r as the support of r and denote it as $\text{sup}(r, \mathcal{D})$.

$$\text{ids}(D) = \{ id(r) \mid r \in D \} \quad \text{sup}(r, \mathcal{D}) = \{ D \mid D \in \mathcal{D} \wedge id(r) \in \text{ids}(D) \}$$

The support of r , in turn, gives us definitions for possible and certain records.

$$r \text{ is certainly in } \mathcal{D} := [\text{sup}(r, \mathcal{D}) = \mathcal{D}]$$

$$r \text{ is possibly in } \mathcal{D} := [|\text{sup}(r, \mathcal{D})| \geq 1]$$

Likewise, we can define the set of possible values of a record r 's attribute A as the set of values of A in all records with the same identity.

$$\text{possible}(r[A] \in \mathcal{D}) := \{ r'[A] \mid id(r) = id(r') \wedge r' \in D \wedge D \in \mathcal{D} \}$$

This in turn allows us to say that a record r 's attribute A is certain if and only if it has exactly one possible value, or that it is bounded if its possible values satisfy some constraint.

$$r[A] \text{ is certain in } \mathcal{D} := [|\text{possible}(r[A] \in \mathcal{D})| = 1]$$

$$r[A] \text{ is bounded by } [\ell, h] \text{ in } \mathcal{D} := \forall a \in \text{possible}(r[A] \in \mathcal{D}) : \ell \leq a \leq h$$

If the set of possible worlds has an associated probability measure $p(\mathcal{D})$, we can define the confidence of a record as the marginal probability over all possible worlds with a record with the same identifier.

$$\text{conf}(r \in \mathcal{D}) := \sum_{D \in \text{sup}(r)} p(D)$$

We call any subset of records $S \subseteq \bigcup_{D \in \mathcal{D}} D$ a *summary* if no pair of records in S share an id. We call a summary *complete* if every identifier in \mathcal{D} is represented. That is, S is a complete summary of \mathcal{D} if and only if:

$$S \subseteq \bigcup_{D \in \mathcal{D}} D \quad \wedge \quad \forall r_1 \neq r_2 \in S : id(r_1) \neq id(r_2) \quad \wedge \quad \text{ids}(S) = \bigcup_{D \in \mathcal{D}} \text{ids}(D)$$

12.5.2 Compact Encodings of Possible Worlds

With this terminology in place, we are now ready to describe the space of the representational schemes used by existing uncertain and probabilistic database systems. A specific representation is the result of three categories of representational

Alg. Family	Example Systems	Equivalence	Filtering	Statistics
Monte Carlo Chase	MCDB [23], Jigsaw [27] Data Exchange [11]	Implicit Set	Samples Certain	Agg, Conf. None
Local Condition Pruning	MayBMS [3, 20], Orchestra [18, 22] Top-K [37], Dissociation [15]	Set	Possible	Enum, Conf.
PC-Tables VC-Tables Model	PIP [26], Orion [36] Mimir [41, 31] Velox [5], MauveDB [9]	Implicit Implicit Implicit	Possible Best Guess Possible	Agg, Enum, Conf Taint, Top-K Prior Conf

Fig. 12.6 PQP systems in terms of the representational semantics used to communicate uncertain tables

features: (1) **Equivalence** or how the scheme decides which tuples to place in a specific tuple group, (2) **Filtering** or what subset of the complete summary relation the scheme incorporates, (3) **Statistics** or how the scheme summarizes properties of each tuple group. When appropriate, we also distinguish between **tuple-level** and **attribute-level** statistics. Figure 12.6 illustrates how existing PQP schemes relate to these features.

12.5.2.1 Record Equivalence: Assigning Identifiers

Record identifiers eliminate redundancy in the summary by allowing us to represent certain forms of conflicts through attribute-level uncertainty. To date, existing probabilistic and incomplete database systems have adopted one of two approaches to identifiers that we call Set and Implicit identity. The vast majority of literature on probabilistic databases (e.g., [4, 7, 13, 15, 16, 32, 37, 39]) ignores attribute level uncertainty. In this approach, which we term Set-identity, the entire record is used as an identifier. Under Set-identity, two records have the same identity if and only if the values of their attributes are identical.

Conversely, systems [23, 26, 36, 40, 41] that support attribute-level uncertainty typically give each tuple an implicit identifier. In this approach, which we term Implicit identity, each record in input tables is assigned a unique identifier (analogous to the ROWIDs of popular database systems). This identifier is propagated through relational algebra expressions in a manner that mimics database provenance [17] (i.e., lineage or pedigree) as illustrated in Fig. 12.7. Projection and selection preserve the identity of a record. Product deterministically derives a new identifier for each output record from the identifiers of the records used to derive it. Union deterministically derives new identifiers for each output record based on the input record’s identifiers and which side of the union it came from.⁹ Aggregates

⁹This prevents repeated identifiers if a record appears on both sides.

Operator	Notation	SQL
Table	R	<code>SELECT * , ROWID FROM R;</code>
Projection	$\pi_{A,B,\dots}(R)$	<code>SELECT A, B, ... , ROWID FROM R;</code>
Selection	$\sigma_\psi(R)$	<code>SELECT * FROM R WHERE \psi;</code>
Product	$R \times S$	<code>SELECT *, MERGE_IDS(R.ROWID, S.ROWID) AS ROWID FROM R, S;</code>
Union	$R \cup S$	<code>SELECT *, MERGE_IDS(R.ROWID, '1') AS ROWID FROM R UNION [ALL] SELECT *, MERGE_IDS(S.ROWID, '2') AS ROWID FROM S;</code>
Aggregate	$\gamma_{A,\dots,M \leftarrow \alpha_M,\dots}(R)$	<code>SELECT A, ..., \alpha_M AS M, ... MERGE_IDS(A, B, ...) AS ROWID FROM R GROUP BY A, B, ...;</code>

Fig. 12.7 Deriving implicit row identifiers

produce entirely new records. We identify the resulting records by deriving an identifier from the grouping attributes or using a default attribute if there are no grouping attributes.

Example 18 Consider the `customers` table in the two possible worlds D_1 and D_2 of Fig. 12.2. Set-identity schemes (e.g., U-Relations [2] or semiring annotations [16]) assign identity based on record values. The records for Bob, Peter, and Arno are each assigned one identifier across both worlds, while each of the two possible records for Alice is assigned its own identifier. Accordingly, records for Bob, Peter, and Arno are certain, while the other records are only possible.

Conversely, implicit-identity schemes (e.g., MCDB [23], Pip [26], or Mimir [31, 41]) might use a key attribute like `SSN` as an identifier for the row. In such schemes, all four records are certain and only Alice's `ownsProperty` attribute is uncertain.

12.5.2.2 Filtering Uncertain Records

It is often helpful to further summarize possible relations by filtering out low-importance record identities. The most general approach to filtering is based on a record's support. Most incomplete database systems intended for Data Exchange, Cleaning, or Fusion (e.g., [11]) return only certain tuples of a query result. Conversely, many probabilistic database systems present the complete set of possible results [2].

Certain and possible results represent two extremes of a spectrum. The former may omit potentially valuable information, while the latter might produce many records, overwhelming the user. The search for an intermediate “sweet spot” has led to the emergence of a variety of semantics that is a superset of certain answers and a subset of possible answers. (1) **sample** filtering [23, 27] includes a union of all records from a (lossy) sampled set of possible worlds. (2) **threshold** filtering includes all records with a support or confidence that larger than a threshold size or probability. (3) **top-k prior** filtering [31, 41] includes all records that appear in the results for the k most likely *possible worlds*. We call the special case of the top

1 prior **best guess** filtering, as these are the results from the most likely possible world. Finally, (4) **top-k posterior** results include only the top-k records, as ranked by the confidence or support of the *result itself* [10, 30].

Note the distinction between the two top-k filtering strategies: top-k prior filters before marginalizing while top-k posterior filters afterward. The top-k posterior filtering strategy is particularly appropriate for settings where the user is searching for the most likely result. For example a user examining a medical diagnostic query result is probably interested in the most likely diagnosis. We note that although this is appropriate for such specialized use cases like diagnostic or recommender systems, such representations can lead to a confusing proliferation of semantics [30].

12.5.2.3 Statistics for Uncertain Attributes

The final point to consider when designing a summary representation is how to represent the records themselves. For this, we need to convey both record- and attribute-level uncertainty.

Record-level statistics Uncertain result records do not appear in all possible worlds. When communicating this information to the user, we effectively wish to communicate some features of the record's support. Although some schemes are capable of **enumerating** the set of all worlds that a record appears in, this capability is typically expensive. Rather, most PQP schemes compute or approximate a record's **confidence** [7, 13, 15]. A simpler approach uses **taint** annotations [29, 31, 41] to differentiate between certain and possible records.

Attribute-level statistics Unless set-identity is being used, records may have uncertain attributes as well. Most probabilistic and incomplete database schemes assume that we are only interested in summarizing individual attributes and not more general properties. The most common strategy is to construct **aggregate** summaries of the attribute's values across the record. Histograms [23], expectations [23, 26], confidence bounds [27], or hard bounds have been used as aggregate summaries. Another approach is to summarize an attribute of a record by one or a set of possible values [10, 41]. Any record filtering strategies can be leveraged to decide which possible values to include, e.g., top 1 prior [41] or top-k posterior [10, 37].

12.6 Conclusions

In this paper, we explain how uncertainty arises in detection and resolving of data quality issues, and how this uncertainty may cause data quality issues in analysis results which often remain undetected. Such untrustworthy analysis results can in turn have severe adverse real world effects such as unfounded scientific discoveries, financial damages, or even affect people's physical well-being (e.g.,

medical decisions based on data with low quality). The main purpose of this work is to give an overview of uncertain data management techniques and raise awareness of how these techniques can be applied to explain how heuristic resolutions to data quality problems affect the quality and trustworthiness of analysis results.

References

1. S. Abiteboul, R. Hull, V. Vianu, *Foundations of Databases* (Addison-Wesley, Reading, 1995)
2. L. Antova, C. Koch, On apis for probabilistic databases, in *QDB/MUD*, Auckland (2008), pp. 41–56
3. L. Antova, C. Koch, D. Olteanu, MayBMS: managing incomplete information with probabilistic world-set decompositions, in *ICDE*, Istanbul (IEEE Computer Society, 2007), pp. 1479–1480
4. J. Boulos, N.N. Dalvi, B. Mandhani, S. Mathur, C. Ré, D. Suciu, MYSTIQ: a system for finding more answers by using probabilities, in *SIGMOD Conference* (ACM, New York, 2005), pp. 891–893
5. D. Crankshaw, P. Bailis, J.E. Gonzalez, H. Li, Z. Zhang, M.J. Franklin, A. Ghodsi, M.I. Jordan, The missing piece in complex analytics: low latency, scalable model management and serving with velox, in *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research*, Asilomar, 4–7 Jan 2015, Online Proceedings (2015). www.cidrdb.org
6. P. Dagum, R.M. Karp, M. Luby, S.M. Ross, An optimal algorithm for Monte Carlo estimation. *SIAM J. Comput.* **29**(5), 1484–1496 (2000)
7. N. Dalvi, D. Suciu, The dichotomy of probabilistic inference for unions of conjunctive queries. *J. ACM* **59**(6), 30:1–30:87 (2013)
8. G.V. den Broeck, D. Suciu, Query processing on probabilistic data: a survey. *Found. Trends Databases* **7**(3–4), 197–341 (2017)
9. A. Deshpande, S. Madden, MauveDB: supporting model-based user views in database systems, in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, SIGMOD'06*, New York (ACM, 2006), pp. 73–84
10. L. Detwiler, W. Gatterbauer, B. Louie, D. Suciu, P. Tarczy-Hornoch, Integrating and ranking uncertain scientific data, in *ICDE*, Shanghai (2009), pp. 1235–1238
11. R. Fagin, P.G. Kolaitis, R.J. Miller, L. Popa, Data exchange: semantics and query answering. *Theor. Comput. Sci.* **336**(1), 89–124 (2005). Database Theory
12. W. Fan, Dependencies revisited for improving data quality, in *Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2008*, Vancouver, 9–11 June 2008, ed. by M. Lenzerini, D. Lembo (ACM, 2008), pp. 159–170
13. R. Fink, J. Huang, D. Olteanu, Anytime approximation in probabilistic databases. *VLDB J.* **22**(6), 823–848 (2013)
14. M. Garcia, J. Ulman, J. Wisdom, *Database Systems: The Complete Book* (Prentice Hall, Upper Saddle River, 2002)
15. W. Gatterbauer, D. Suciu, Dissociation and propagation for approximate lifted inference with standard relational database management systems. *VLDBJ* **26**(1), 5–30 (2017)
16. T. Green, V. Tannen, Models for incomplete and probabilistic information, in *Current Trends in Database Technology – EDBT 2006*, ed. by T. Grust, H. Höpfner, A. Illarramendi, S. Jablonski, M. Mesiti, S. Müller, P.-L. Patranjan, K.-U. Sattler, M. Spiliopoulou, J. Wijsen. Lecture Notes in Computer Science, vol. 4254 (Springer, Berlin/Heidelberg, 2006), pp. 278–296
17. T.J. Green, G. Karvounarakis, V. Tannen, Provenance semirings, in *PODS*, Beijing (2007), pp. 31–40

18. T.J. Green, G. Karvounarakis, N.E. Taylor, O. Biton, Z.G. Ives, V. Tannen, ORCHESTRA: facilitating collaborative data sharing, in *SIGMOD*, Beijing (2007), pp. 1131–1133
19. P. Guagliardo, L. Libkin, Making SQL queries correct on incomplete databases: a feasibility study, in *PODS* (ACM, New York, 2016), pp. 211–223
20. J. Huang, L. Antova, C. Koch, D. Olteanu, MayBMS: a probabilistic database management system, in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, SIGMOD'09*, New York (ACM, 2009), pp. 1071–1074
21. T. Imielinski, W. Lipski, Jr., Incomplete information in relational databases. *J. ACM* **31**(4), 761–791 (1984)
22. Z.G. Ives, T.J. Green, G. Karvounarakis, N.E. Taylor, V. Tannen, P.P. Talukdar, M. Jacob, F. Pereira, The orchestra collaborative data sharing system. *SIGMOD Rec.* **37**(3), 26–32 (2008)
23. R. Jampani, F. Xu, M. Wu, L.L. Perez, C. Jermaine, P.J. Haas, MCDB: a Monte Carlo approach to managing uncertain data, in *SIGMOD*, Vancouver (2008), pp. 687–700
24. R.M. Karp, M. Luby, Monte-Carlo algorithms for enumeration and reliability problems, in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, Berkeley, vol. 0 (1983), pp. 56–64
25. O. Kennedy, The PIP MayBMS plugin. <http://maybms.sourceforge.net>
26. O. Kennedy, C. Koch, PIP: a database system for great and small expectations, in *ICDE*. (IEEE Computer Society, Piscataway, 2010), pp. 157–168
27. O.A. Kennedy, S. Nath, Jigsaw: efficient optimization over uncertain enterprise data, in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, SIGMOD'11*, New York (ACM, 2011), pp. 829–840
28. D. Koller, N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, Cambridge, 2009)
29. P. Kumari, S. Achmiz, O. Kennedy, Communicating data quality in on-demand curation, in *QDB* (2016)
30. J. Li, B. Saha, A. Deshpande, A unified approach to ranking in probabilistic databases. *pVLDB* **2**(1), 502–513 (2009)
31. A. Nandi, Y. Yang, O. Kennedy, B. Glavic, R. Fehling, Z.H. Liu, D. Gawlick, Mimir: bringing ctables into practice. Technical report, ArXiv (2016)
32. D. Olteanu, J. Huang, C. Koch, Approximate confidence computation in probabilistic databases, in *ICDE* (IEEE Computer Society, Piscataway, 2010), pp. 145–156
33. C.H. Papadimitriou, *Computational Complexity* (Wiley, Reading, 2003)
34. S. Parsons, *Probabilistic Graphical Models: Principles and Techniques by D. Koller, N. Friedman* (MIT Press), 1231pp. \$95.00, ISBN 0-262-01319-3. *Knowl. Eng. Rev.* **26**(2), 237–238 (2011)
35. C. Ré, N.N. Dalvi, D. Suciu, Efficient top-k query evaluation on probabilistic data, in *ICDE* (IEEE Computer Society, Los Alamitos, 2007), pp. 886–895
36. S. Singh, C. Mayfield, S. Mittal, S. Prabhakar, S. Hambrusch, R. Shah, Orion 2.0: native support for uncertain data, in *SIGMOD*, Vancouver (2008), pp. 1239–1242
37. M.A. Soliman, I.F. Ilyas, K.C. Chang, Top-k query processing in uncertain databases, in *ICDE* (IEEE Computer Society, Los Alamitos, 2007), pp. 896–905
38. M. Stonebraker, D.J. Abadi, A. Batkin, X. Chen, M. Cherniack, M. Ferreira, E. Lau, A. Lin, S. Madden, E.J. O’Neil, P.E. O’Neil, A. Rasin, N. Tran, S.B. Zdonik, C-store: a column-oriented DBMS, in *VLDB* (ACM, New York, 2005), pp. 553–564
39. D. Suciu, D. Olteanu, C. Ré, C. Koch, *Probabilistic Databases*. Synthesis Lectures on Data Management. (Morgan & Claypool Publishers, San Rafael, 2011)
40. J. Widom, Trio: a system for integrated management of data, accuracy, and lineage. Technical Report (2004)
41. Y. Yang, N. Meneghetti, R. Fehling, Z.H. Liu, O. Kennedy, Lenses: an on-demand approach to ETL. *Proc. VLDB Endow.* **8**(12), 1578–1589 (2015)

Chapter 13

Evaluation of Information in the Context of Decision-Making



Mark Burgin

Abstract In a broad sense, solving a problem can be treated as deciding or making a decision what the solution to this problem is. In particular, decision-making with respect to a question means finding an answer to the question. Thus, solution of any problem can be treated as decision-making. However, traditionally decision-making is understood as making a choice from a set of alternatives, which are usually alternatives of actions. Here we consider decision-making in the traditional form exploring the role and features of information in this process. In section “The Process of Decision-Making”, we consider existing models and elaborate a more detailed model of decision-making. In section “Properties of Information and Their Evaluation”, we demonstrate that each stage and each step of decision-making involve work with information—information search, acquisition, processing, evaluation, and application. Evaluation of information is especially important for decision-making because utilization of false or incorrect information can result in wrong and even disastrous decisions. We show how to evaluate quality of information in the context of decision-making, what properties are important for information quality, and what measures can be useful for information evaluation. The obtained results are aimed at improving quality of information in decision-making by people and development of better computer decision support systems and expert systems.

Keywords Information · Decision-making · Evaluation model · Process · Action · Selection

M. Burgin (✉)
University of California, Los Angeles, Los Angeles, CA, USA
e-mail: mburgin@math.ucla.edu

13.1 Introduction: Information as an Indispensable Resource in Decision-Making

Decision-making is information processing. Consequently, information plays a crucial role in decision-making. Information comes as input to a decision-making system, such as a person, robot, or a software system, is processed by this system, and is given as an output of this process. Even when the best choice is known to the decision-maker in advance or when the outcomes of the decision are not sufficiently important, information is basic for the process of decision-making. It is even truer for decision-making in uncertain complex situations when it is necessary to make difficult and important decisions with high risks and incomplete information.

Thus, it is essential to have high-quality information for making good decisions, and because we never have perfect, e.g., complete, information, it is necessary to evaluate information that we have, obtain, and produce.

To analyze the role and evaluate characteristics of information in decision-making, it is necessary to have a sufficiently adequate and detailed model of decision-making. We discuss existing models and elaborate a more detailed model of decision-making in the next section.

The primary goal of information evaluation is improving information quality. Another goal is evaluation of the made decision because features of information used in decision-making influence the quality of the made decision [48, 50].

The chapter is organized in the following way. In Sect. 13.2, going after Introduction, we consider existing models and elaborate a more detailed model of decision-making. In Sect. 13.3, information evaluation in decision-making is explored. We base this exploration on the general theory of information, the most detailed presentation of which is given in the book (Burgin, 2010) [1]. The last section contains conclusions.

13.2 The Process of Decision-Making

There are three basic types of decision-making models:

1. *Operational models* describe the process of decision-making.
2. *Descriptive models* bring about properties and relations of decision-making.
3. *Impact models* show what factors influence decision-making.

There are three basic classes of factors that influence decision-making:

1. *External factors* reflect what influences decision-making from the decision-maker's environment.
2. *Internal factors* reveal what traits of the decision-maker influence decision-making.
3. *Relational factors* expose relations between external factors and internal factors.

Here we are mostly interested in operational models because to evaluate information in a decision-making process, we need a detailed description of such a process. Using a well-structured decision-making process helps in making more purposeful, thoughtful, and careful decisions by obtaining and organizing pertinent information, examining alternative choices, and finding best possible route to take. This tactic increases the chances for making an optimal possible decision.

Researchers in the area of decision-making elaborated different structural representations of decision-making processes. One of the first models of decision-making was introduced by Herbert Simon [2]. It has three stages:

1. *Intelligence* deals with problem identification and information gathering for the problem.
2. *Design* involves generation of possible alternative solutions to the problem under consideration.
3. *Choice* is selection of the “best” solution from the possible alternative solutions using some relevant criterion.

Later, this model was developed and utilized for building computer decision support systems [3].

In his model, Raiffa employed the following stages for decision-making [4]:

1. List the viable options available to you for gathering information, for experimentation, and for action.
2. List the events that may possibly occur.
3. Arrange in chronological order the information you may acquire and choices you may make as time goes on.
4. Decide how well you like the consequences that result from the various courses of action open to you.
5. Judge what the chances are that any particular uncertain event will occur.

Another operational model of decision-making is proposed by Thakur [5]:

1. Identification and structuring of problem/opportunity
2. Putting the problem/opportunity in context
3. Generation of alternatives
4. Choice of the best alternative

Ryan and Shinnick [6] describe the structure of decision-making in the following way:

Step 1: Identify the decision.

When there is a need in making a decision, it is very important to clearly characterize the nature of the required decision.

Step 2: Gather relevant information.

Collect some pertinent information before you make your decision: what information is needed, the best sources of information, and how to get it. This step involves both internal and external “work.” Some information is internal: you’ll

seek it through a process of self-assessment. Other information is external: you'll find it online, in books, from other people, and from other sources.

Step 3: Identify the alternatives.

As you collect information, you will probably identify several possible paths of action, or alternatives. You can also use your imagination and additional information to construct new alternatives. In this step, you will list all possible and desirable alternatives.

Step 4: Weigh the evidence.

Draw on your information and emotions to imagine what it would be like if you carried out each of the alternatives to the end. Evaluate whether the need identified in Step 1 would be met or resolved through the use of each alternative. As you go through this difficult internal process, you'll begin to favor certain alternatives: those that seem to have a higher potential for reaching your goal. Finally, place the alternatives in a priority order, based upon your own value system.

Step 5: Choose among alternatives.

Once you have weighed all the evidence, you are ready to select the alternative that seems to be best one for you. You may even choose a combination of alternatives. Your choice in Step 5 may very likely be the same or similar to the alternative you placed at the top of your list at the end of Step 4.

Step 6: Take action.

You're now ready to take some positive action by beginning to implement the alternative you chose in Step 5.

Step 7: Review the made decision and its consequences.

In this final step, consider the results of your decision and evaluate whether or not it has resolved the need you identified in Step 1. If the decision has not met the identified need, you may want to repeat certain steps of the process to make a new decision. For example, you might want to gather more detailed or somewhat different information or explore additional alternatives.

In the procedural system GOFER, the name of which is an acronym for five decision-making steps, the following structure is identified [7, 8]:

1. Goals clarification, which involves a survey of values and objectives
2. Options generation, which demands to consider a wide range of alternative actions
3. Facts-finding, which means a search for information
4. Consideration of effects, which requires weighing the positive and negative consequences of the options
5. Review and implementation, which includes elaboration of a plan, how to implement the options, and implementing them

In the DECIDE model of decision-making, there are six steps [9]:

1. Define the problem.
2. Establish and enumerate all the criteria (constraints).
3. Consider and collect all the alternatives.
4. Identify the best alternative.
5. Develop and implement a plan of action.
6. Evaluate and monitor the solution and examine feedback when necessary.

Note that not all of these steps allow realization. For instance, in many problems, it is impossible to collect all alternatives—some of possible alternatives can stay unknown to the decision-maker. Besides, it is not always possible to identify the best alternative because there are situations in which obtaining the best alternative demands performing infinitely many operations.

Brown divides the decision-making process into seven steps [10]:

1. Outline your goal and outcome.
2. Gather data.
3. Develop alternatives (i.e., brainstorming).
4. List pros and cons of each alternative.
5. Make the decision.
6. Immediately take action to implement it.
7. Learn from and reflect on the decision.

In Bezerra, et al. [11], a decision-making process is represented in the form of six consecutive stages:

1. Perception activity
2. Mental representation
3. Data processing
4. Problem solving
5. Choice of solution
6. Decision making to act

Here we put forward the following stratification of decision-making processes, which constitutes the SDM model of decision-making:

1. Evaluation of the situation
2. Identification of the problem
3. Structuration of the problem
4. Determination of the context of a problem (what are initial conditions, what are accessible means and other resources, what are constraints, and so on)
5. Evaluation of the problem
6. Generation/identification of possible actions aimed at solving the problem
7. Evaluation of possible actions, which in some cases coincide with possible solutions
8. Identification of tentative outcomes of possible actions, which, in a general case, include possible solutions

9. Evaluation of tentative outcomes of possible actions
10. Selection of an optimal possible action (solution) with the goal to determine which course of actions is the best way to achieve the objective of decision-making
11. Development of a plan (algorithm) for realization of the chosen action
12. Taking action to solve the problem
13. Evaluation of the action
14. Evaluation of the results of action
15. Evaluation of the new situation
16. Learning from and reflecting on the made decision

It is possible to limit a decision-making process only by steps from 1 through 10. However, if it is necessary to have a complete picture of the process giving insight for future activities, we need all sixteen steps.

Many people can utilize considered steps of decision-making without realizing it, but gaining a clearer understanding of best practices can improve the efficiency of made decisions.

An important feature of the SDM model of decision-making is that it takes into account that, in general, actions can be composite consisting of other actions [12].

Note that some considered above stages (steps) in decision-making are often overlooked. For instance, learning from and reflecting on the made decision, which includes evaluation of the made decision for effectiveness, is an important stage in the decision-making process. This would allow making a better decision next time.

It is important to understand that one decision-making process can include other decision-making processes. Here is one of such situations. Often when decision-making is a thorough detailed process, action is complex and it is necessary to have a plan or even an algorithm of such an action. This involves another decision-making determining or elaborating a plan (algorithm) for the chosen action. In turn, elaboration of a plan (algorithm) requires identifying what resources are required for the action, how it must be organized, and what are the constraints delimiting its execution. For instance, very often temporal constraints are essential [13].

Group decision-making brings additional stages related to group organization, orientation, and correlation. For instance, Pijanowski [14] describes eight stages of group decision-making:

Establishing community by creating and nurturing the relationships, norms, values, and procedures that will influence how problems are understood and communicated.

Perception involves recognition that a problem exists.

Interpretation demands identification of competing explanations for the problem and evaluation of the drivers behind those interpretations.

Judgment as sifting through various possible actions or responses and determining which of them is more justifiable.

Motivation requires examination of the competing commitments, which may distract from an optimal course of action, and then prioritization and commitment to the values that were accepted when community (group) was created on the first stage.

Action means following through with action that supports the more justified decision.

Reflection in action.

Reflection on action.

Researchers found that establishing common values and norms in a group improves the quality of decisions, while the majority of opinions, which are called consensus values and norms, do not [15].

According to Aubrey Fisher (cf., Griffin, [16]), there are four stages or phases necessary for group decision-making:

Orientation goes when members of the group meet for the first time and start to get to know each other.

Conflict emerges once group members are becoming accustomed with each other, which sometimes involves disputes, little fights, and arguments.

Emergence when the group begins to clear up vague opinions by talking about them.

Reinforcement happens when members of the group finally make a decision and provide justification for it.

This shows that group decision-making is a more complex process in comparison with an individual decision-making. Thus, it is even more important to base group decision-making on an adequate and efficient model of decision-making.

13.3 Properties of Information and Their Evaluation

Information plays an important role on all stages of any decision-making process. To examine this role and describe evaluation of information in this process, we use the SDM model of decision-making described in the previous section. The reason for this choice is that assuming each step of decision-making involves work with information, the SDM model encompasses other known models of decision-making providing for better analysis and evaluation of information utilized on each stage of decision-making processes.

The first stage in decision-making is evaluation of the current situation. Naturally, evaluation of the situation demands gathering a sufficient amount of relevant information so that it would be possible to better understand what is going on and make thoughtful, informed decisions that have a positive impact. This requires making a value judgment, finding what information is necessary and what information might be useful for improved discernment of the current situation. Necessary information has to be evaluated by appraising its accessibility, cost of acquisition, and importance for understanding the situation. Let us consider these properties.

Accessibility of information reflects possibility or probability of obtaining this information. Taking the classical probability in consideration, we know that when the probability of obtaining is equal to 1, then it is possible to get this information.

For instance, if we have some portion of information, then the probability of obtaining it is naturally equal to 1. When the probability of obtaining is between 0 and 1, then there are chances that it might be impossible to get this information. Probability equal to 0 means complete impossibility of obtaining information.

However, probabilistic decision-making needs considering signed probability, which takes values between -1 and 1 . Negative probability indicates that obtaining some information can be dangerous and have a negative impact.

Signed probabilities, such as symmetric probability or combined probability, have become very popular in physics (cf., e.g., Dirac, 1974 [17]; Feynman, 1987 [18]; Ferrie and Emerson, 2008 [19]; Kronz, 2009 [20]). Negative probabilities also came to economics and finance (Duffie and Singleton [21]; Forsyth, et al. [22]; Haug [23], Burgin and Meissner [24, 25]). In addition, negative probabilities were used in social and behavioral sciences (cf., e.g., de Barros et al. [26, 27]). Axiomatics for symmetric probability and combined probability, which include negative values, is developed in Burgin [28, 47], while mathematically based interpretations are given in Burgin [46], Abramsky and Brandenburger [31], and Noldus [32].

Note that probabilities used in decision-making are, as a rule, subjective. This means that they are not sufficiently exact. That is why it is often reasonable to use possibility measures instead of probability measures in evaluation of properties of information because possibility measures are more general.

A *possibility measure* in X is a partial function $\text{Pos}: \mathbf{P}(X) \rightarrow [0, 1]$ that is defined on a subset \mathbf{A} from $\mathbf{P}(X)$ and satisfies the following axioms [33, 34]:

- (Po1) $\emptyset, X \in \mathbf{A}$, $\text{Pos}(\emptyset) = 0$, and $\text{Pos}(X) = 1$.
- (Po2) For any A and B from \mathbf{A} , the inclusion $A \subseteq B$ implies $\text{Pos}(A) \leq \text{Pos}(B)$.
- (Po3) For any system $\{A_i; i \in I\}$ of sets from \mathbf{A} ,
 $\text{Pos}(\bigcup_{i \in I} A_i) = \sup_{i \in I} \text{Pos}(A_i)$.

Note that similar to probability, it might be useful to consider possibility measures, which also take negative values in the form of a function $\text{Pos}: \mathbf{P}(X) \rightarrow [-1, 1]$.

Cost of information acquisition can be measured by a probability distribution or even by a fuzzy set on the set of possible costs.

To estimate importance of information for understanding the situation, it is possible to use necessity measures.

A *quantitative necessity measure* in X is a function $N: \mathbf{P}(X) \rightarrow [0, 1]$ that satisfies the following axioms [35].

- (Ne1) $N(\emptyset) = 0$, and $N(X) = 1$.
- (Ne2) For any A and B from $\mathbf{P}(X)$, $N(A \cap B) = \min \{N(A), N(B)\}$.

At the same time, importance is better assessed by a signed measure, which is not bounded from above or from below taking both positive and negative values. Information that makes understanding of the situation harder has negative importance. In a similar way, information that misleads or confuses the decision-maker also has negative importance.

The next thing to do after demarcation and evaluation of the necessary information is to decide where it is possible to get this information or, at least, some part of it, and the possible procedures to employ for this purpose. Considering information sources, it might be also useful to evaluate them estimating their relevance, reliability, accessibility, and information abundance. For this purpose, it is possible to use methods and measures developed and employed for software evaluation in Burgin and Debnath [36] because software is a symbolic representation of information.

The third step in the first stage is information acquisition and processing obtained information into knowledge about the situation.

To do this properly, we need to evaluate obtained information, estimating such properties as relevance, reliability, adequacy, completeness, usefulness, correctness, exactness, and importance of the obtained information.

In this context, *relevance* means the degree to which the obtained information is connected to the evaluated situation.

Reliability is to estimate if utilization of the obtained (considered) information will give positive results.

Adequacy means how well the obtained information represents the evaluated situation. Usually, it is possible to decompose adequacy into completeness, correctness, and exactness.

Completeness estimates whether the obtained information allows finding all main features of the evaluated situation.

Correctness appraises whether the obtained information does not distort the evaluated situation.

Exactness means the precision with which the obtained information represents the real situation.

Note that obtained information can be relevant, reliable, adequate, complete, correct, exact, and important but useless because, for example, the decision-maker does not understand or does not know how to use this information. For instance, the history of mathematics tells us that when information about negative numbers came to Europe from the East where negative numbers were utilized for centuries, many European mathematicians, such as Chuquet and Maseres, did not understand this information and dismissed their sensibility rejecting negative numbers until the seventeenth century and referring to them as “absurd” or “meaningless.” Even in the eighteenth century, it was a common practice to ignore any negative results derived from equations, on the assumption that they were meaningless [37]. As a result, information about negative numbers was useless to those who did not understand them although this information was relevant, reliable, adequate, complete, correct, exact, and important. Similarly, information about irrational numbers and later imaginary numbers was firstly rejected by many mathematicians with the same consequences [38].

The second stage of decision-making is identification of the problem. According to the methodology of science, a *problem* is absence of something, e.g., of knowledge or information, and explicit representation of this absence [39].

Thus, the second stage also demands collection of additional information and can be decomposed into three steps: (1) finding what information is necessary for problem identification and to what extent; (2) delineation of possible sources of the necessary information and procedures for information acquisition; and (3) information acquisition and processing obtained information into knowledge about the problem.

Evaluation of information in these steps is similar to evaluation of information on the first stage, which is described above. We use the same properties of information and information sources as well as the same procedures of evaluation.

All other stages of decision-making—structuration of the problem, determination of the context, and so on—demand search, acquisition, processing, and application of information. Consequently, evaluation of information involves evaluation of these operations and means used in their performance. This brings us to the following schema of information processes in decision-making.

Schema (13.1): Search → acquisition → processing → application/utilization

Each of the processes in Schema (13.1) has three basic components:

- Input information
- Participating means, which include used means
- Output information

These three components form the following structure:

Schema (13.2): Input information → means used → output information

Taking into account that input information comes from some source while output information goes to some information destination—which can include or coincide with the receptor/receiver—it is reasonable to extend Schemas (13.2) and (13.3).

Schema (13.3): Information source → input information → participating means → output information → information destination

Because participating means including the channel, noise source, message, and signals, we see that Schema (13.3) is a generalization of Shannon's model of communication (Shannon, 1948 [40]) presented in Fig. 13.1.

In addition, all three components of information processes in decision-making have one of the following three modalities, i.e., they can be:

- Tentative or intended component, e.g., tentative or intended information
- Actual or used component, e.g., actual or used information
- Necessary or required component, e.g., necessary or required information

Thus, to have a high-quality process, it is necessary to evaluate all information involved and take actions to improve information quality. As the quality of information depends on features of the means (tools) used for information search, acquisition, processing, and application/utilization, it is also useful to evaluate these means (tools) aiming at their improvement (if possible).

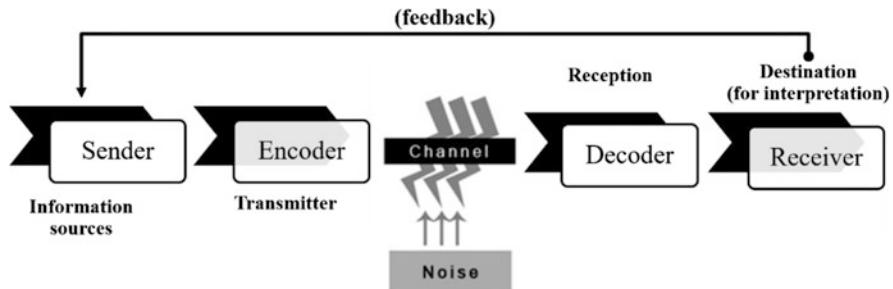


Fig. 13.1 Shannon's model of communication

In addition, evaluation of actions (stage 7 in the SDM model) includes estimation (measurement) of the corresponding properties of actions such as feasibility, acceptability, and desirability. It is also important to understand the risks involved with the possible actions to fully grasp the existing situation and potential risks. Only knowing these and some other parameters of actions, it is possible to select an optimal action (route of operation) with high chances of success.

In a general case, evaluation of information includes estimation of such properties (features) as relevance, reliability, adequacy, completeness, correctness, exactness, importance of information, and some others [1].

To efficiently organize evaluation, it is necessary to discern properties of three types [29, 41]:

- *Quantitative properties* take values in sets of numbers, e.g., classical probability takes values in the interval $[0, 1]$, while information entropy takes values in the set of all positive real numbers (Shannon [40]).
- *Qualitative properties* take values in sets with an order relation, e.g., qualitative probability takes values in partially ordered sets (Burgin, [30]), while many properties represented by linguistic variables, such as age or size, take values in the set of triangular fuzzy numbers (Zadeh [42]).
- *Nominal properties* take values in sets without order relations and their values are treated simply as labels or names.

Note that the considered features of information can be represented by quantitative, qualitative, or nominal properties. For instance, it is possible to represent correctness of information by a quantitative measure, by the linguistic variable that takes values in the ordered set *{completely incorrect, partially incorrect, partially correct, essentially correct, strongly correct, completely correct}* and by the linguistic variable that takes values in the set of two labels *{incorrect, correct}*.

Quantitative properties are preferable for evaluation of actual components of processes including actual information, while qualitative properties more realistically estimate potential, tentative, intended, or projected components of processes including potential, tentative, intended, or projected information [1].

Nominal properties are useful, for example, for identification of the possible causes of the problem under consideration, constraints posed by the existing situation, or issues that require decision-making (cf., e.g., Ryan and Shinnick [6]; Citroen [43]).

However, as Knight and Burn write, despite the sizable body of literature on information quality, relatively few researchers have tackled the difficult task of quantifying conceptual definitions of information quality and suggesting methods for its estimation (Knight and Burn [44]). Nevertheless, information quality has many numerical measures and comprises various measured attributes as it is practically undeniable that no one “magic number” can give out a measurement for all features of information that might be considered significant [1]. Some of these attributes are related to directly measurable numerical quantities, while numerical quantities related to others are only calculated from direct measurements. To attain a wide-ranging information quality representation by numerical attributes and to elaborate efficient numerical measures for these numerical, it is convenient to use the general theory of evaluation and measurement [45]. According to this theory, the process of measurement/evaluation has three main stages:

- *Preparation*
- *Realization*
- *Analysis*

The first step in evaluation/measurement preparation is determination of a specific criterion for evaluation/measurement. Such a criterion determines the mission of evaluation/measurement. Criteria of information quality include such properties as relevance, reliability, exactness, adequacy, completeness, correctness, convenience, user friendliness, and so on. However, such properties are not directly measurable and it is possible only to estimate them. To achieve this goal, it is necessary to utilize matching indicators or indices. With respect to information quality such indicators are called *general information quality measures* or *metrics* [1]. However, a chosen indicator can be too complicated for direct estimation. This causes a stipulation to set up more specific attributes of the evaluated object. To get these attributes, quantifiable tractable questions are formulated. The obtained attributes work as *indices* for the considered criterion. It means that the second stage of evaluation is selection of indices that reflect chosen criteria. In some cases, an index can overlap with the corresponding criterion or a criterion can be one of its indices. However, often it is impossible to obtain exact values for the chosen indices. For instance, it is impossible to perform infinitely many operations or to do measurement with absolute precision. What is feasible to do is only to obtain some estimates of indices. As a result, the third stage of evaluation includes elaboration or location of estimates or *indicators* for selected indices. In the case of information, these indicators have the form of *constructive (procedural) information quality measures*.

As a result, to achieve correct and sufficiently precise evaluation, preparation must include the following operations:

1. Selection of evaluation criteria
2. Assigning characteristics (indices) to each of the selected criteria
3. Representing indices by indicators (estimates)

This exhibits that the complete process of evaluation preparation has the following structure:

Schema (13.4): Criterion → index → indicator

Efficient creation of information quality measures consists of the following three stages:

1. Setting goals specific to needs in terms of purpose, perspective, and environment.
2. Refinement of goals into quantifiable tractable questions.
3. Construction of an information quality measure and data to be collected (as well as the means for their collection) to answer the questions.

Information quality measures can be useful in practice only when there are procedures and algorithms of measurements based on these measures. This implies necessity in additional stages for the measure development.

4. Designing procedures/algorithms for data collection in the measurement.
5. Designing procedures/algorithms for computing measurement values.
6. Designing procedures/algorithms for analysis of measurement results.

An efficient theoretical base for elaboration of information measures is provided by the system of axiological principles from the general theory of information [1].

13.4 Conclusion

We have developed a comprising model of decision-making, which is called the SDM model, and based on this model elaborated a general approach to information evaluation in decision-making. To apply this approach, it is necessary to select criteria for information evaluation in each stage of decision-making described by the SDM model, construct/select indices and indicators developing corresponding measures, and then to elaborate algorithms and procedures of measurement based on these measures. It is possible to use this approach for improving quality of information in decision-making by people and for building better computer decision support systems and expert systems [49].

Acknowledgement The author would like to express gratitude to Éloi Bossé for his useful remarks.

References

1. M. Burgin, *Theory of Information: Fundamentality, Diversity and Unification* (World Scientific, New York/London/Singapore), p. 2010
2. H.A. Simon, *The New Science of Management Decision* (Harper & Row, New York, 1960)
3. J.-C. Pomerol, F. Adam (2004) Practical decision making – From the legacy of Herbert Simon to decision support systems, in *Proceedings of Decision Support in an Uncertain and Complex world: The IFIP TC8/WG8.3 International Conference*, pp. 647–657
4. H. Raiffa, *Decision Analysis: Introductory Lectures on Choices Under Uncertainty* (Addison-Wesley, Reading, Massachusetts, 1968)
5. D. Thakur, Role of Information in Decision-making, Computer Notes. <http://ecomputernotes.com/mis/decision-making/role-of-information>. Accessed 15 Apr 2018
6. G. Ryan, E. Shinnick, The role of information in decision making, in *Encyclopaedia of Decision Making and Decision Support Technologies*, (Ideas Group Reference, Pennsylvania/London, 2008)
7. L. Mann, Becoming a better decision maker. *Aust. Psychol.* **24**(2), 141–155 (1989)
8. L. Mann, R. Harmoni, C. Power, The GOFER course in decision making, in *Teaching Decision Making to Adolescents*, (Lawrence Erlbaum Associates, Hillsdale, 1991), pp. 61–78
9. K.L. Guo, DECIDE: A decision-making model for more effective decision making by health care managers. *Health Care Manag.* **27**(2), 118–127 (2008)
10. P. Brown, Career coach: decision-making. *Pulse* (2007). <http://www.pulsetoday.co.uk/career-coach-decision-making/10967084.article>
11. S. Bezerra, Y. Cherrault, J. Fourcade, G. Veron, A mathematical model for the human decision-making process. *Math. Comput. Model.* **24**(10), 21–26 (1996)
12. M. Burgin, M.L.A. Smith, Unifying model of concurrent processes, in *Proceedings of the 2007 International Conference on Foundations of Computer Science (FCS'07)*, (Press, Las Vegas, 2007), pp. 321–327
13. M. Burgin, M. Smith, Compositions of concurrent processes, in *Communicating Process Architectures*, (IOS Press, Scotland, September, 2006), pp. 281–296
14. J. Pijanowski, The role of learning theory in building effective college ethics curricula. *J. Coll. Char.* **10**(3), 1–13 (2009)
15. T. Postmes, R. Spears, S. Cihangir, Quality of decision making and group norms. *J. Pers. Soc. Psychol.* **80**(6), 918–930 (2001)
16. E.A. Griffin, Interact system model of decision emergence of B. in *A First Look at Communication Theory*, ed. by A. Fisher (McGraw-Hill, New York, 1991), pp. 253–262
17. P.A.M. Dirac, *Spinors in Hilbert Space* (Plenum, New York, 1974)
18. Feynman, R. P. (1987) Negative Probability, in *Quantum Implications: Essays in Honour of David Bohm*, Routledge & Kegan Paul Ltd, London/New York, pp. 235–248
19. C. Ferrie, J. Emerson, Frame representations of quantum mechanics and the necessity of negativity in quasi-probability representations. *J. Phys. A Math. Theor.* **41**, 352001 (2008)
20. F. Kronz, Actual and virtual events in the quantum domain. *Ontology Studies* **9**, 209–220 (2009)
21. D. Duffie, K. Singleton, Modeling term structures of defaultable bonds. *Rev. Financ. Stud.* **12**, 687–720 (1999)
22. P.A. Forsyth, K.R. Vetzal, R. Zvan (2001) *Negative Coefficients in Two Factor Option Pricing Models*, Working Paper (electronic edition: <http://citeseer.ist.psu.edu/435337.html>)
23. Haug, E.G. (2004) Why so Negative to Negative Probabilities, *Wilmott Magazine*, Sep/Oct, pp 34–38
24. M. Burgin, G. Meissner, Negative probabilities in modeling random financial processes. *Integration* **2**(3), 305–322 (2010)
25. M. Burgin, G. Meissner (2012) Negative Probabilities in Financial Modeling, *Wilmott Magazine*, March 2012, pp. 60–65

26. J.A. de Barros, G. Oas, Negative probabilities and counter-factual reasoning in quantum cognition. *PhysicaScripta* **T163**, 014008 (2014)
27. J.A. de Barros, Decision making for inconsistent expert judgments using negative probabilities, in *Quantum Interactions*, (Springer, Berlin/Heidelberg, 2013), pp. 257–269
28. Burgin, M. Extended Probabilities: Mathematical Foundations, Preprint in Physics,math-ph/0912.4767, 2009 (electronic edition: <http://arXiv.org>)
29. M. Burgin, *Theory of Knowledge: Structures and Processes* (World Scientific, New York/London/Singapore), p. 2016
30. Burgin, M. (2016) Probability theory in relational structures, *J. Adv. Res. Appl. Math. Stat.*, v. 1, No. 3&4, pp. 19–29
31. Abramsky, S. and Brandenburger, A. An operational interpretation of negative probabilities and no-signalling models, in *Horizons of the Mind. A Tribute to Prakash Panangaden*, Lecture Notes in Computer Science, vol. 8464, 59–75 (2014)
32. J. Noldus, A short note on extended probability theory, [quant-ph] 29 Sep 2015 (arXiv:1509.09281)
33. L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Syst.* **1**, 3–28 (1978)
34. K.-J. Zimmermann, *Fuzzy Set Theory and its Applications* (Kluwer Academic Publishers, Boston/Dordrecht/London, 2001)
35. M. Oussalah, On the Qualitative/Necessity Possibility Measure, *I. Inform. Sci.* **126**, 205–275 (2000)
36. M. Burgin, N. Debnath, Quality of software that does not exist, in *Proceedings of the ISCA 20th International Conference Computers and their Applications*, (ISCA, New Orleans, Louisiana, 2005), pp. 441–446
37. Martinez, *Negative Math*, in *How Mathematical Rules Can Be Positively Bent*, (Princeton University Press, Princeton, 2006)
38. D.M. Burton, *The History of Mathematics* (The McGraw Hill Co., New York, 1997)
39. M. Burgin, Problems in the system of scientific knowledge. *Proceedings* **1**(3), 273–275 (2017). <https://doi.org/10.3390/IS4SI-2017-03947>
40. C.E. Shannon, The mathematical theory of communication. *Bell Syst. Tech. J.* **27**(1), 379–423 (1948).; No. 3, pp.623–656
41. M. Burgin (1985) Abstract theory of properties, in *Non-classical Logics*, Moscow, pp. 109–118 (in Russian)
42. L.A. Zadeh (1973) *The Concept of a Linguistic Variable and its Application to Approximate Reasoning*, Memorandum ERL-M 411, Berkeley
43. C.L. Citroen, The role of information in strategic decision-making. *Int. J. Inf. Manag.* **31**(6), 493–501 (2011)
44. S.-A. Knight, J. Burn, Developing a framework for assessing information quality on the world wide web. *Inform. Sci. J.* **8**, 159–172 (2005)
45. M. Burgin, L Kavunenko. Measurement and evaluation in science, STEPS Center, Kiev, 1994 (in Russian)
46. M. Burgin, Interpretations of negative probabilities, Preprint in Quantum Physics,quant-ph/1008.1287, 2010, 17 p. (electronic edition: <http://arXiv.org>)
47. M. Burgin, Axiomatizing negative probability. *J. Adv. Res. Appl. Math. Stat.* **1**(1), 1–17 (2016b)
48. K. Parviainen, The Importance of Valid and Real-Time Information in Decision-Making (<https://www.m-brain.com/blog-posts/the-importance-of-valid-and-real-time-information-in-decision-making/>). Accessed 15 Apr 2018
49. R.H. Sprague, H.J. Watson, *Decision Support Systems: Putting Theory into Practice* (Prentice Hall, Englewood Cliffs), p. 1993
50. M.C. Yovits, A. de Korvin, R. Kleyle, L. Medsker, M. Mascarenhas, The Relationship Between Information and Decision Making and the Effect on the Reliability and Failure of Information Systems, in *Information Systems: Failure Analysis*, NATO ASI Series (Series F: Computer and Systems Sciences), eds. by J. A. Wise, A. Debons, vol. 32, (Springer, Berlin/Heidelberg, 1987)

Chapter 14

Evaluating and Improving Data Fusion Accuracy



John R. Talburt, Daniel Pullen, and Melody Penning

Abstract Information fusion is the process of combining different sources of information for use in a particular application. The production of almost every information product incorporates some level of data fusion. Poor implementation of data and information fusion will have an impact on many other key data processes, most particularly data quality management, data governance, and data analytics. In this chapter we focus on a particular type of data fusion process called entity-based data fusion (EBDF) and on the application of EBDF in high-risk applications where accuracy of the fusion must be very high. One of the foremost examples is in healthcare. Fusing information belonging to different patients or failing to bring together all of the information for the same patient can both have dire, even life-threatening, implications.

Keywords Entity-based data fusion · Probabilistic matching · Precision · Recall · *F*-Measure · Data quality management · Quality control · Quality assurance

14.1 Introduction

Data fusion, sometimes called data integration, is simply the process of combining different sources of information for use in a particular application [3]. The production of almost every information product incorporates some level of data fusion. Despite this fact, data fusion is often taken for granted, and its implementation is delegated to low-level programmers or database administrators.

J. R. Talburt (✉) · D. Pullen
University of Arkansas at Little Rock, Little Rock, AR, USA
e-mail: jrtalburt@ualr.edu; dlpullen@ualr.edu

M. Penning
University of Arkansas for Medical Sciences, Little Rock, AR, USA
e-mail: mpenning@uams.edu

Poor implementation of data fusion will have an impact on many other key data processes, most particularly data quality management, data governance, and data analytics.

In this chapter we focus on a particular type of data fusion process called entity-based data fusion (EBDF) and on the application of EBDF in high-risk applications where the accuracy of the fusion must be very high. One of the foremost examples is in healthcare. Fusing information belonging to different patients or failing to bring together all of the information for the same patient can both have dire, even life-threatening, implications. Other examples of high-risk EBDF applications include law enforcement and fraud detection.

Because of the potential for adverse events occurring due to errors in these applications, the supporting data fusion processes need to employ specialized matching techniques and ancillary processes not commonly used in lower-risk applications. Some of the most important are:

- Probabilistic value-level matching
- Validation of blocking alignment
- Production of review indicators
- Clerical review and investigation of indicators
- Robust metadata for process traceability and auditability
- Systematic and periodic measurement of EBDF performance
- A comprehensive data quality management program

Each of these topics will be explored in more detail in this chapter.

14.2 An Overview of Entity-Based Data Fusion

Entity-based data fusion (EBDF) is the process of integrating information about the same real-world entity coming from different sources. From an EBDF perspective, entities are real-world objects with distinct identities. For example, customers of a business, patients of a hospital, students of a school system, products of a manufacturer, or locations of service. The primary issue with EBDF is, in the modern world, we model and describe these entities in an information system. Because the information stored is stored in a computer, it is subject to many processes which can easily transform and copy the information related to these entities. As a result, there are often many disparate records in the information system describing characteristics of the same entity.

EBDF has a broad range of applications in areas such as law enforcement [16], education [17, 18], and healthcare [1, 12, 32]. For example, a hospital patient will have separate records for each hospital visit. In addition, each visit or encounter will generate many additional records such as a record for each laboratory test, each treatment, each drug administered, and so on. Aside from the records of medical treatment, there will also be billing records associated with each visit and

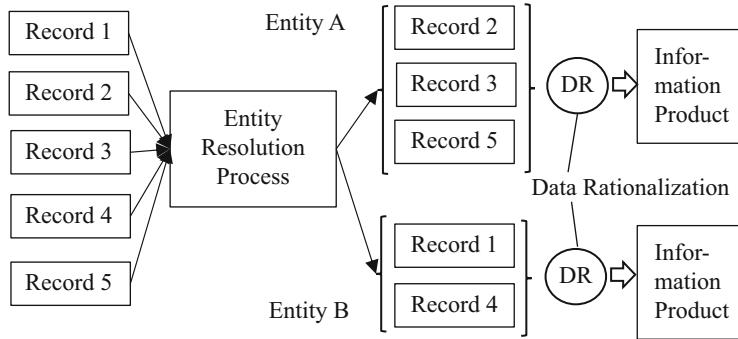


Fig. 14.1 The two steps of the entity-based data fusion process

treatment. Because these records are produced by different systems, they typically have different record layouts and different formatting standards.

EBDF is a two-step process [22] as illustrated in Fig. 14.1. The first step is entity resolution (ER). ER is the process to determine whether two records are referencing the same entity or different entities [24]. Figure 14.1 illustrates five entity records resolving to two entities, Entity A and Entity B by the ER process. In this example, Records 2, 3, and 5 carry information for Entity A, and Records 1 and 4 have information for Entity B. Groups of records referencing the same entity are called “clusters.” All of the records in the same cluster are given the same identifier value called a “link.” For this reason, ER is sometime referred to as “record linking.”

After the records have been clustered according to entity, the second step is the data rationalization step. The purpose of the data rationalization process is to extract and reconcile possibly conflicting or incomplete information in the records in preparation for creating the final information product [8, 9, 31, 33]. For example, in a healthcare application, the incoming records could be patient records describing various diagnoses, treatments, medications, or laboratory tests, and the final product is the patient’s medical chart.

14.2.1 The Probabilistic Nature of Entity Resolution

As shown in Fig. 14.1, the first step in data fusion is the entity resolution (ER) process. ER is sometimes called record linking, data matching, or record de-duplication. ER is the process of determining when two information system references to a real-world entity are referring to the same entity or to different entities [24]. ER represents the “sorting out” process when there are multiple sources of information that are referring to the same set of entities. For example, the same patient may be admitted to a hospital at different times or through different departments such as inpatient and outpatient admissions. ER is the process

of comparing the admission information for each encounter and deciding which admission records are for the same patient and which ones are for different patients.

14.2.2 The Difference Between ER and Matching

Entity resolution is often confused with data matching in both understanding and terminology. By definition the goal of ER is to determine when two references to real-world entities are referencing the same entity or different entities. When two references are for the same entity, they are classified as “equivalent references.” The confusion between matching and equivalence arises because of the similarity assumption that is the primary driver for ER.

The similarity assumption of entity resolution states: The more similar the values of identity attributes between two entity references, the more likely the references are equivalent, and the less similar the values, the less likely they are equivalent [23].

Methods for assessing similarity comprise an entire discipline called data matching [2, 7]. The important point is while matching has a high positive correlation with equivalence, it only represents a likelihood or probability of equivalence or non-equivalence, not a certainty. Not all equivalent references match. For example, patient records for “Mary Doe, 123 Oak” and “Mary Smith, 345 Elm” could be for the same patient, married and moved to a new address. Failing to recognize the equivalence of dissimilar entity references (failure to link) is called a false negative error. Negative because it was a decision not to link, and false because it was a bad decision.

Conversely, not all matching references are equivalent. For example, the patient records “Jim Jones, 67 Pine” and “James Jones, 67 Pine” could be for different patients, a father and his son with similar names and living at the same address. Linking two non-equivalent entity references into the same cluster is called a false positive error. Positive because it was a decision to link, but false because it was a bad decision.

The goal of ER is to link only equivalent references (true positive decision) and to not link non-equivalent references (true negative decision). The accuracy of the ER process is the degree to which this goal has been achieved by reducing or eliminating the false positive and negative errors in the process.

14.3 Identifying and Addressing Sources of Errors in the EBDF Process

Almost all of the errors in the EBDF process are introduced at one of four points in the process; these are:

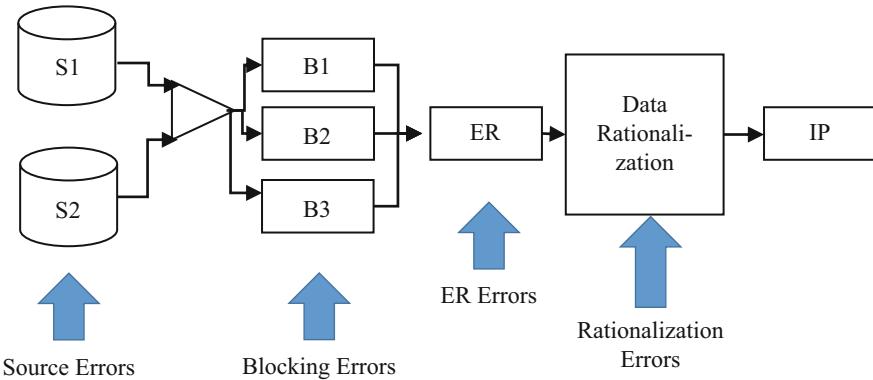


Fig. 14.2 Points in the EBDF process where most errors occur

1. Errors in source information
2. Errors in blocking
3. Errors in entity resolution
4. Errors in data rationalization

These are illustrated in Fig. 14.2 showing the steps in the information product (IP) build process.

14.3.1 Errors in Source Information

No matter how well the data fusion process is performed, it cannot overcome errors of fact in the source information. Certainly some data quality defects such as inconsistent formatting can be addressed and corrected with appropriate tools, but there is no simple solution for detecting and correcting erroneous information. Increasingly EBDF processes involve semi-structured and unstructured sources such as social media that present a new set of challenges [14].

Accuracy is the most difficult dimension of data quality to address. It requires verification by the original source of the information or by comparing the information to a reference source known to be accurate. The former is often impractical because of the time and expense involved, and the latter often does not exist for the domain of interest. Instead, most EBDF processes rely on validation instead of verification.

Validation is using rules to test for outlier conditions in the source information. Validation is when rules are used to test for outlier values or unexpected conditions in the source information.

Validation and accuracy have an asymmetric relationship. While it is true that source information failing validation rules usually indicates the information is inaccurate, the converse is not true. Source information passing validation is not

necessarily accurate. Even if a date-of-birth value is within a valid range and in the correct format, it does not mean it is accurate, i.e., the actual date of birth for the person. To know it is accuracy would require verification by the person or by comparison to an authoritative source such as health records or driver's license information. Even the authoritative source method for accuracy is in itself problematic. It requires an ER process to link the record in the information source to the corresponding record in the authoritative source, and this ER process is itself subject to error.

14.3.2 Errors in Blocking

ER is defined as a pairwise classification process. An ER system is asked to judge if two references to the same entity are equivalent or not. If the ER process decides the references are likely to be equivalent, it links them together; otherwise it does not link them. In theory, this decision would be made for every pair of references in the source information. However, making every possible comparison is impractical even in the new age of Big Data and massively parallel computing platforms.

The reason it is impractical is a matter of mathematics. The number of possible pairwise comparisons for a set of N objects is

$$\text{Pair count} = C(N, 2) = \frac{N \cdot (N - 1)}{2}$$

But even for a small amount of data, say 1000 records, the total number of possible pairs is 499,500, almost half a million. In many applications today, a source size of 200 million (2×10^8) is not to be considered large. But for a source file of this size, the total number of possible pairwise comparisons is approximately 10^{16} or a quadrillion. Even using a highly parallel, high-speed processing system, an ER process for these many comparisons could take several weeks to execute.

Therefore, as a practical matter, ER is always performed on subsets of the source data. The subsets are called "blocks" and the process for creating them is called "blocking" [2]. The ER process only compares entity references within the same block as a way to reduce the number of comparisons to a manageable amount. The smaller the blocks the faster the ER process will execute.

However, ER blocking presents a conundrum. If two equivalent references are placed in different blocks, then the two references will not be compared by the ER process and consequently will not be linked and clustered together. Not linking equivalent references is a false negative error lowering the accuracy of the ER process. Therefore, the goal of blocking is to always place equivalent pairs of entity references in the same block. So here is the conundrum, if we already knew which pairs of references were equivalent, we would not need an ER process in the first place.

The solution is to treat blocking as a “rough match” or “pre-match” to place the references most likely to match in the same block. After the pre-match blocking, the blocks are sent to the ER process for the final refinement of the matching decisions. A blocking strategy will depend heavily on the matching strategy employed by the ER engine.

Suppose one of the matching rules used by a patient ER system requires both references to have the last name. Then the blocking process needs to assure all source references with the same last name are placed in the same block. However, as a practical matter there will be other rules for matching that do not require last names to be the same because a patient may change his or her last name between encounters.

Most systems implement blocking through the generation of match keys. A match key is a string value made by concatenating together pieces of identity attribute values. The match keys need to be in alignment with the match rules. Alignment is fairly straightforward when using “if-then” style rules. It is a matter of translating the match rule requirement into a match key requirement. For example, if a match rule requires an exact match on a student last name and a SOUNDEX match on the first name (first names that sound alike), then a match key formed by concatenating the letters of the last name with the SOUNDEX code of the first name would support this rule. All references with the same last name spelling and the same first name SOUNDEX code would be brought together into the same block for comparison by the ER engine.

If there are other match rules that have a different requirement such as the same first name and the same date of birth, then another match key generator would have to be created to support this rule. This is because some records for the same person could have different last names (married name versus maiden name) and be matched by this rule. However, the first match key would not place such a pair of records into the same block because it requires the last names to be the same. So a different match key is needed to support the second rule, e.g., the concatenation of the first name with the digits of the date of birth. Because most ER systems can match records in many different ways, they must generate several different match keys for each reference.

The match rule in the ER engine is said to be “in alignment” with match key blocking if the following statement is always true: Whenever two references match by a rule, then the two references will share the same value for at least one match key.

Errors caused by misalignment between blocking and the match rule are often difficult to detect. Misalignment almost always creates false negative errors which are difficult to detect. One method for validating alignment is to take a sample of references representative of the different ways of matching. Then perform ER on the sample with the blocking in place, and then execute the same ER process again without the blocking in place. Both results should be the same. If they are not, it shows some references were not linked because the references were placed in different blocks and are never compared even though the references matched.

A complex match rule usually requires several match key generators. This is especially true when using the scoring (probabilistic) rule strategy for matching as

discussed in the next section. As the number of match keys increases, the block sizes will increase, and at some point, defeat the entire purpose of blocking which is to reduce the number of pairwise comparisons that must be made.

Strategies for blocking can be as complex as the matching strategies they support. At this point it suffices to say that blocking is an often overlooked source off error in the EBDF process. For a more complete discussion of blocking strategies, see Christen [2], and for Big Data see Talburt and Zhou [23].

14.3.3 Errors in ER

Because of the similarity assumption, all automated ER systems will always have some level of false positive and false negative errors. Over the years a number of ER matching strategies have been introduced to try and reduce these errors as much as possible.

Most ER systems currently use one of two types of matching strategies. The first matching strategy, often referred to as “deterministic” matching, uses “if-then” logic with Boolean operators “and/or” to make their decisions [4]. The second strategy, often referred to as “probabilistic,” uses the model developed by Fellegi and Sunter [5] a scoring method. Of course, the truth is that all ER techniques are both deterministic and probabilistic, but this terminology is already embedded in the literature. Presently, a third strategy using machine learning is being explored [11].

The earliest ER systems employed Boolean rules because they are easy to construct, explain, and maintain. The major weakness of the Boolean rule strategy is that it operates at the attribute level of the references. In other words, if a Boolean rule is comparing first name values, then a match of “JOHN” as the first name value in one reference to the first name value “JOHN” is a second reference is no different than a match of “BARTHOLOMEW” as a first name value one reference to the first name value of “BARTHOLOMEW” in another record. They are both just first name matches, and both similarities carry the same weight.

If the match on first names is coupled with a match on last name, then it forms an “AND” clause of the form of a Boolean proposition “If the first names match and the last names match, then link the references together as equivalent references.” Boolean rules are still used extensively in ER systems, especially in support of low-risk EBDF applications such as marketing where accuracy is not paramount.

14.3.4 Using Frequency-Based, Probabilistic Matching

The scoring rule strategy was developed to try and improve the accuracy of the ER processes using Boolean rules by drilling down the attribute-value level. The degree of match between two attribute values such as first names would depend on the actual value of the name, i.e., the name itself.

In general, the scoring rule operates as follows. First all of the values of the identity attributes are compared. For each comparison, if the values of the identity attribute agree (match), then an agreement weight is added to a match score. If the values disagree, then a disagreement weight smaller than the agreement weight (possibly a negative value) is added to the match score. After all identity attribute values have been compared, the total match score is compared to a predetermined threshold value. If the total score of the pair is greater than or equal to a preset threshold value, then the references are judged equivalent and linked together; otherwise the references are not linked.

The magnitude of the agreement weight assigned to an attribute value is based on the probability that different entities would share the same value. The more frequently a value is shared by different entities, the smaller its agreement (match) weight should be. Conversely, the less frequently different entities share a value, the higher the agreement weight assigned to the value. In effect, the weights reflect the discrimination power of the attribute value.

Using the same example as before, it would be true for most populations of students in the United States; the first name value of “JOHN” should be assigned a lower agreement weight than the first name value of “BARTHOLOMEW.” The reason is in a typical population of students it is much more likely many students will share the first name “JOHN” while fewer students will share the first name “BARTHOLOMEW.” In a scoring-based approach, the higher the frequency of an attribute value in the target population, the less a match on this value is an indicator of equivalence and the smaller its agreement weight should be.

More precisely, the agreement weight is the logarithmic value of a ratio of two conditional probabilities calculated by the formulas [7]

$$\text{Agreement weight} = \log_2$$

$$\left(\frac{\text{Probability of agreement given equivalent references}}{\text{Probability of agreement given non-equivalent references}} \right)$$

And the disagreement weight is calculated by the formula

$$\text{Disagreement weight} = \log_2$$

$$\left(\frac{1 - (\text{Probability of agreement given equivalent Refs})}{1 - (\text{Probability of agreement given non-equivalent Refs})} \right)$$

The weights can be calculated and assigned either at the attribute level or at the value level, e.g., the weight of agreement on first names in general or the weight of agreement on “JOHN.” However, when assigned at the value level, the scoring rule can significantly increase the accuracy of the ER process because it takes into account the relative discrimination power of different identity attribute values. Consequently, frequency-based probabilistic matching is generally the best choice for high-risk data fusion application such the master patient index (MPI) in healthcare.

For practical reasons, only the most frequent attribute values are assigned an individual agreement and disagreement weight. Typically, a frequency threshold is established for each attribute based on an analysis of data from the target population. Only the values of the attribute occurring with a frequency above this threshold are assigned individual agreement and disagreement weights. Values of the attribute with frequencies below the threshold are given a default, low-frequency agreement and disagreement weight.

Even though the frequency-based scoring rule can attain higher levels of ER accuracy for most applications, this matching strategy does have some drawbacks as well. One weakness is the scoring rules sensitivity to missing values. During the scoring process, if one or both of the values for an identity attribute is missing, then assigning an agreement or disagreement weight does not make sense. For this reason, many systems use a default weight for missing value conditions, usually zero.

Another problem with the scoring rule is blocking. Because the scoring rule works on value-based weights, it can be very difficult to align the scoring rule with match key blocking. When using Boolean rules, the matching requirements are always associated with specific attributes. However, with the scoring rule the requirement is for the total score to be above the match threshold. This can usually be achieved through many different combinations of attribute values and weights. Identifying these combinations and translating them into match key generators can be a challenge.

One final issue with the scoring rule is the issue of versioning and maintenance. With increased accuracy also comes increased sensitivity. The weights and match threshold are calculated on the data as they existed at first implementation. Yet the scoring rule can be sensitive to even modest shifts in the frequency of attribute values. As new sources are added or when existing sources undergo change, the accuracy of the ER may begin to fall if the weights are not recalibrated to produce optimal results.

In addition, it is very difficult to address a particular linking issue by making weight adjustments in the scoring rule. The weights are highly coupled, and making adjustments the weights to solve one particular type of error has the potential to create many other new errors. Even if a change is made to only one weight, the scoring rule should go through the complete regression testing regimen and treated as an ER version change.

14.4 Clerical Review and Assertion as a Supplement to Automation

Because of the probabilistic nature of matching, all automated ER systems will introduce some level of the false positive and false negative linking error into the EBDF process. For critical or high-risk applications, the level of accuracy achieved by the automated ER process alone is often not sufficient. When this happens, the

automated process must be supplemented with an additional manual ER process called “clerical review” or “remediation” [23].

Clerical review has its roots in the Theory of Record Linking described by Fellegi and Sunter [5]. The proof of their theorem relies on the ability to have reviewers (persons) make the correct linking decisions for a certain group of entity reference pairs. Because it is a manual process, clerical review is only used in high-risk EBDF processes where the extra time and expense are justified.

Just as with blocking, clerical review presents a second conundrum. Because not every pair of references can be manually reviewed, the goal of clerical review is to inspect only those pairs of references where the automated process may have made an error. But of course, if the automated process knew it was making an error, it could be programmed not to make the error. Again the solution lies in probability. Even though the automated system may not know it is making an error, it can detect the match conditions which have higher probabilities of an incorrect linking decision. These conditions are called “clerical review indicators.”

Another advantage of the scoring rule is it has a built-in review indicator, the match threshold. Pairs of references scoring very close to the match threshold, either above or below, have a higher likelihood of being a false positive or false negative, respectively. It is a simple matter to set a tolerance level above and below the match threshold as a review indicator. Pairs of reference scoring within the tolerance window around the threshold can be written to an exception file for later clerical review.

It is important to note that effective clerical review is done by people who understand the entity domain, not by matching experts. Presumably the matching experts have already transferred all of their knowledge into the rules of the matching process. The role of the clerical reviewer is not to determine how well the references match, but to determine if the references are equivalent or not. This often requires the reviewer to use other information external to the EBDF process itself. For example, if the entities are patients, it may require accesses other hospital systems, public information system, or even making calls to the patients in question to make a correct linking decision.

Another important aspect of clerical review is the reviewer decisions are not directly implemented by the reviewer. The decisions are first recorded as transactions, and the transactions are then processed by a special configuration of the ER system. This delay also provides for another level of quality assurance. It allows time for a reviewer’s decision to undergo a second review before being processed by the system.

The entire clerical review process should be defined by a set of data governance policies and procedures. The policies and procedures will define the clerical review process, the roles and responsibilities of the reviewers, and the conditions under which reviews may be escalated to even higher levels of review if necessary.

Clerical requires an auxiliary support system to allow a reviewer to see and understand the current linking configuration in the system called out for review by the review indication. These clerical review systems require a robust graphical user interface (GUI) providing a full view of not only the review indication but also the

complete context of the other records in the cluster. The system should allow the reviewer to drill down into the complete context of the references.

Once a final decision has been made by the reviewer and approved, then the next step is to “assert” the decision transaction into the ER system. An assertion is simply the implementation of the reviewer’s decision in the ER system.

Clerical review systems should support two types of assertions, correction assertions and confirmation assertions. When the reviewer determines the indicated references are incorrectly linked, the review must design a correction assertion to remedy the error. For example, if two indicated references were linked incorrectly (false positive), then the reviewer would create a “split assertion” to unlink the references and place them into the correct clusters. On the other hand, if the reviewer determines the indicated references are equivalent, but were not linked by the automated system, then the review would create a “consolidation” assertion to force the system to link the references.

Review indicators only report conditions likely to be in error. The reviewer may decide an indication is not an error, i.e., the automated ER process made the correct decision. If the reviewer decides the automated system was correct, the review should still make a confirmation assertion. A confirmation assertion does not change the linking configuration of the system. Instead, it embeds special metadata into the system to prevent the same clerical review indication from being reported again. Confirmation assertion helps reduce the overall time and effort to conduct clerical review on a regular basis. Once a pair of references indicated for clerical review has been reviewed and determined to be correct, then the system should not select the same pair for clerical review again.

14.4.1 Robust Metadata

Clerical review assertions point to the need for the ER system use robust metadata to support in the EBDF process. For example, a split assertion to repair a false positive link made by the automated system should also create metadata to prevent the same false positive error from occurring in the future. Otherwise, the same automated matching rule will continue to make the same mistake of linking the references. This can start a loop where the automated rules make a false positive link, the pair is indicated for review, the reviewer splits the references, the automated system links them again, and cycle repeats over and over again.

In addition to metadata needed to persist an assertion, there should be metadata to annotate the assertion for future reference and audit. The annotation would identify the reviewer making the assertion, the type of the assertion, the date of the assertion, and possibly notes or other information.

The work of the reviewer will also be helped by metadata inserted by the automated ER process itself, for example, having an identifier for each reference that is unique across the entire system, a way to identify the source of the reference, the date and process run that created a cluster, the date and rule that added each reference to the cluster, and so on.

14.4.2 Error in Data Rationalization

By definition, data fusion is bringing together multiple sources of information. The purpose of the ER process is to assure that the information being brought together is for the same entity. However, even though the information is for the same entity, it does not mean the information will be consistent. The two main issues faced in data rationalization are conflict resolution and value harmonization.

Conflict resolution occurs when two or more sources report different values for the same attribute. Depending on the attribute, the conflict can be a good thing or a bad thing. Conflict resolution is an issue faced daily by large data brokers [8, 9]. As a simple example, consider conflicting dates of birth reported for the same person. Assuming the ER process is correct and the records are actually for the same person, then we know there should only be one correct value for the date of birth. Resolving this conflict requires selecting one value as the correct value. However, for some types of products and application, all of the values are reported.

Of course each EBDF application will have its own unique set of issues, but in general, selection from conflicting values is made by a simple “voting” method, by source reliability, or by a combination of both. For example, if there are ten sources and seven report the same date of birth and the other three report different values, then by the voting scheme, the most reported value would be selected.

On the other hand, it often occurs that some sources are deemed to be more reliable and authoritative. In this case, the selected date of birth would be from the most reliable source even if it is the minority value. It may also occur that in some instances, the most reliable source does not have a value. In that case, the secondary rule may be the voting rule. Conversely, if the voting rule is primary, but there is not a predominate value, then the secondary rule may be to select from the more reporting reliable source.

This example also demonstrates why source validation alone is insufficient. As long as all of the reported dates of birth pass the source validation rule, the record will be passed forward into the ER process. The date-of-birth validation is usually just a simple check of whether the date is within some reasonable range of dates. It is not until the records reach the rationalization step that the error is discovered. Again, it is important to understand that passing source validation does not mean the source values are accurate.

There are more sophisticated approaches to making a selection from conflicting values. For example, the use of genetic algorithms showed improvement over more naïve voting methods [31]. Selection from among conflicting values is where some of the newer machine learning capabilities can be applied. Experimenting with the effectiveness of newer data mining and machine learning techniques to value selection is an area where data fusion research could be advanced.

Conflicting values for an attribute are not always bad. For some attributes they are expected. For example, location information such as address. Between encounters, a patient may have moved and report a different address. In cases where different values are acceptable, the question is not necessarily selecting the correct value, but

selecting a preferred value. For address, the preferred value might be either current address or it might be the billing address. The selection rule logic will be based on the specific requirements of the product.

Another issue for rationalization is value harmonization. Harmonization is when values for the same attribute have the same meaning, but are reported in different ways. Using the healthcare example again, it might be diagnoses given in different coding schemes where one is given according to ICD-10 coding and the other according to SNOMED.

A similar situation occurs when values are coded in “brackets.” Records from one clinic may report “pre-high blood pressures” as 120 to 140, but another may define the same category as “130 to 150.” It is unclear whether the actual value should be assumed between 120 and 150 or if it should be assumed between 130 and 140.

Even though data rationalization is the final step in the EBDF process, depending upon the application, it may not be final step in the production of the information product. Rationalization may simply be a precursor step to further processing.

14.5 Measuring ER Accuracy

Periodic and system measurements are necessary to effectively monitor and improve the accuracy of the ER step in the EBDF process. There are two principal measures of ER accuracy, precision and recall [2]. Precision measures the probability that linking decisions will be correct. The formula for precision is

$$\text{Precision} = \frac{\text{Number of true positive links}}{\text{Total number of links made}}$$

In other words, when the ER system decides to link two references, what percentage of the time are the references actually equivalent. Cautious or conservative match criteria tends to yield higher precision, i.e., only linking when there is strong similarity or other supporting evidence of equivalence. However, high precision often comes at the cost of low recall.

The complementary measure to precision is recall. In high-risk applications, failing to link equivalent references (false negative errors) can be as detrimental as false positive errors. Recall measures the percentage of equivalent references linked by the ER process to the total number of equivalent pairs (i.e., total possible links). The formula for recall is

$$\text{Recall} = \frac{\text{Number of true positive links}}{\text{Total number of equivalent pairs of references}}$$

When both measures are 100% (1.00), then the goal of ER has been reached. If two references are linked, then they are equivalent, and if two references are

equivalent, then they are linked. Often these two measures are combined into a single value called the F -measure. The F -measure is the harmonic mean of the precision and recall given by

$$F\text{-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{(\text{Precision} + \text{Recall})}$$

The difficulty with these performance formulas is they all depend on knowing which references are equivalent and which ones are not. Counting true positive pairs assumes you know which pairs are equivalent, and if one knew all of the equivalent pairs as called for the recall calculation, then there would be no need to have an ER process.

In general, these calculations can only be made for some sample of the references. Recently, there have been some strides made to develop methods for obtaining reliable estimates of these measurements with a manageable level of effort. The first method was developed by Pullen [20] and relies on stratified sampling to estimate precision and recall. The second was developed by Penning [19] based on inferred methods developed in the field information retrieval.

14.5.1 Pullen's Stratified Sampling Method

14.5.1.1 Methodology Overview

Unlike many types of machine learning and classification problems, ER has a unique advantage among the many differences and other disadvantages. The similarity assumption of entity resolution provides an approximation for the truth of ER outcomes. This distinct difference provides a mechanism for which the ER results can be scored and later stratified. The similarity assumption was originally leveraged by the Fellegi-Sunter Model of Record Linkage [5]. In this model, high scoring (i.e., more similar) records are automatically matched by the system, while low scoring (i.e., less similar) records are not automatically matched by the system.

Measuring the quality of ER results requires some type of manual review of matched pairs of records to achieve a high degree of accuracy in the quality estimates. Unfortunately, manually assessing a large amount of record pairs is time consuming and expensive. This method uses techniques from the Fellegi-Sunter model of record linkage, stratified sampling, and manual review to provide an accurate estimation of the precision, recall, and f -measure of a set of ER results with an amount of effort that aligns with the organization's available resources.

The degree of accuracy in the estimated measures increases with the resources and time expended by the organization. Additionally, the techniques outlined in this approach provide insight into thresholds for optimal clerical review processes with a desire to minimize the expense of resources when initially measuring the quality of ER results and during on-going quality improvement and monitoring initiatives for the ER process.

The Fellegi-Sunter scoring rule is defined by a set of attributes with a corresponding agreement and disagreement weight. During the pairwise comparison, the agreement weight is applied when two attribute values agree. In contrast, the disagreement weight is applied when two attribute values disagree. To calculate attribute weights for the scoring rule, a proxy for equivalence is needed. In the Fellegi-Sunter theory, equivalence represents the truth in the real world. Unfortunately, knowing or identifying equivalence is not practical in the real world. If it was known, measuring the precision, recall, and F -measure of ER results would be trivial. For this technique, the linking from the existing ER result is used as a proxy to equivalence for calculating attributes scores and applying the scoring rule for the assessment process. The use of existing linkage for generating attribute weights has been tested and proven in previous work [26]. Furthermore, attribute values are considered to be in agreement if the two values match after changing the values to an upper case and trimming any leading or trailing whitespace characters.

14.5.1.2 Calculating Attribute Weights

For this approach, the algorithm used to calculate the agreement weights and disagreement weights is the Fellegi-Sunter probabilistic model for estimated weights under the assumption of conditional independence of the identity attributes [7]. Let $\{a_1, a_2, \dots, a_n\}$ represent the identity attributes of the references. For each attribute we must calculate two conditional probabilities.

$$m_i = \text{Probability}(\text{values of } a_i \text{ agree} | \text{references are equal})$$

$$u_i = \text{Probability}(\text{values of } a_i \text{ agree} | \text{references are not equal})$$

The agreement weight and disagreement weight for each attribute is calculated from these values as follows:

$$\text{AgreeWeight}_i = \log_2 \left(\frac{m_i}{u_i} \right)$$

$$\text{DisagreeWeight}_i = \log_2 \left(\frac{1 - m_i}{1 - u_i} \right)$$

14.5.1.3 Applying Weights for Positive Outcomes

The pairwise comparisons are performed within all existing clusters in the ER results for analysis of positive outcomes (TP and FP). Due to the dependency on the ER links as a proxy for equivalence, the attribute agreement and disagreement

Table 14.1 Example EIS

Reference ID	First name	Middle name	Last name	Date of birth	System ID Num
1	Joey	M.	Johnson	3/25/1995	111111111
2	Joseph		Johnson	3/25/1995	111111111
3	Joey	Sterling	Johnson	9/13/1997	111111111

Table 14.2 Example attribute agreement and disagreement weights

Weight type	First name	Middle name	Last name	Date of birth	System ID Num
Agreement	5	3	5	10	20
Disagreement	-4	0	-4	-8	-20

Table 14.3 Applied weights for the example pairs

Pair	First name	Middle name	Last name	Date of birth	System ID Num	Total score
(1,2)	-4	0	5	10	20	31
(1,3)	5	0	5	-8	20	22
(2,3)	5	0	5	-8	20	22

weights vary for different ER results of the same dataset. Though, the same set of weights can be reused for different ER results. As an example, consider the references shown in Table 14.1, and their corresponding agreement and disagreement weights shown in Table 14.2.

In Table 14.1, there are three records. This means there are three pairs in total that need to be processed for this cluster. Using the Reference ID field to identify each record, the pairs are (1,2), (1,3), and (2,3).

When the attribute values for two references are the same, the agreement weight is applied. If the values are different, the disagreement weight is applied. For the pair (1,2), the First Name values are different. So, the disagreement weight for the first name, -4, is applied. This is performed for each attribute for the pair.

Table 14.3 shows all of the weights for each attribute in each pair. After calculating all of the attribute weights, the results are summed to provide total score for the pair. The results are shown in Table 14.3.

To understand the usefulness of this score as a stratification method, it is important to understand that the score of two pairs is calculated in a manner that depends on the number of similar or dissimilar attribute values shared between the two references being compared. If they are defined as similar by a specified similarity function, an agreement weight is applied. If they are dissimilar by a specified similarity function, a disagreement weight is applied. This means that a higher score corresponds with the number of attributes that match. A higher score signifies a stronger match. A lower score signifies a weaker match.

Consider Fig. 14.3 for a visual representation of this. This corresponds with the similarity assumption of entity resolution.



Fig. 14.3 Visualizing binned pairwise scores

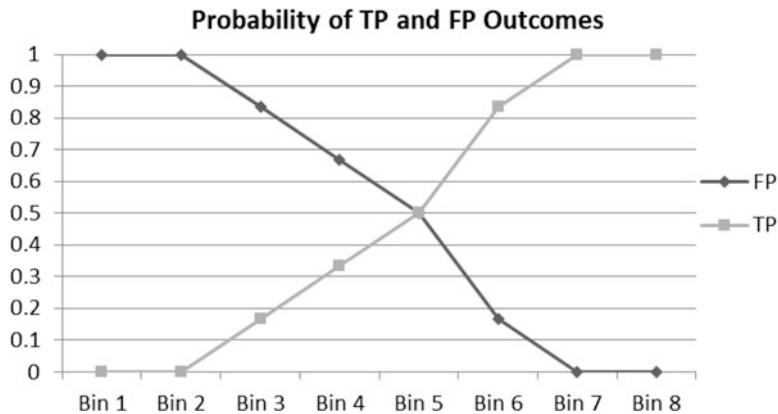


Fig. 14.4 Correlation of score and probability of error

Each of the bins represents a spectrum of values for a score of matching pairs. In the example, each spectrum has a span of 50. The left-most bin represents the pairs that have the least similarity. The right-most bin represents the pairs that have the greatest similarity.

If using equal distance for the bin intervals, the bins represented above will very likely be of unequal population and show a drop off in size as the score reaches the extreme upper or lower ends of the spectrum. By using this information, it is expected that a correlation of the score and the probability of an error should appear as in Fig. 14.4.

14.5.1.4 Applying Weights for Negative Outcomes

The previous example applies to the processing undertaking for positive outcomes (i.e., true positive or false positives). Negative outcomes (i.e., false negatives or true negatives) present a challenge that requires a modified approach. The evaluation is comprised of references across more than one cluster produced by the ER process. In practice, many of these evaluations involve two clusters but may include more.

To find the cross-cluster relationships, a split key is utilized. In this process, the split key acts as a Boolean match rule with a very high recall but relatively low precision. This contrasts with the typical approach of high precision but low recall Boolean rules used by ER systems for pairwise matching. By biasing toward high recall, this allows the split key to identify even the most difficult matches. To reduce the number of potential missed matches that must be manually reviewed, two additional steps are applied: (1) Relationships identified by the split key must disagree with the result produced by the ER process and (2) references that meet the criteria of the split key and the first requirement are grouped together into a value cluster from which the scoring method is applied.

The first step is used to remove potential TP decisions that have no impact on identifying the FN errors and reduce the review set to just potential TN or FN outcomes. The entropy calculated in the second step follows the process described in the false positive indicator entropy approach. The only difference is that this is performed on a value cluster produced by the split key instead of a cluster produced by the ER process.

To understand the application of this process, consider the following scenario with three clusters as shown in Table 14.4.

To identify the missed matches, System ID Number will be used as a simple split key. Using the split key and only including those references that match by the key and disagree with the ER results produces the cluster shown in Table 14.5. Though this example uses a single term key, it is possible to use a multi-term key that even includes fuzziness using a derived hash code comparator such as Soundex, NYSIIS, or regular expression-based transformations.

Table 14.4 Set of three example EIS

Reference ID	First name	Middle name	Last name	Date of birth	System ID Num	Link ID
1	Joey	M.	Johnson	3/25/1995	111111111	A
2	Joseph		Johnson	9/13/1997	111111111	A
3	Joey	Sterling	Johnson	9/13/1997	222222222	A
4	Joseph	Michael	Johnson	3/25/1995	111111111	B
5	Joseph	Michael	Johnson	3/25/1995	111111111	B
6	Joey		Johnson	3/25/1995	111111111	C
7	Joey	M.	Johnson	3/25/1995	333333333	C
8	Joseph	Michael	Johnson	3/25/1995	111111111	C

Table 14.5 Value cluster produced by split key

Reference ID	First name	Middle name	Last name	Date of birth	System ID Num	ER Link ID
1	Joey	M.	Johnson	3/25/1995	111111111	A
2	Joseph		Johnson	9/13/1997	111111111	A
4	Joseph	Michael	Johnson	3/25/1995	111111111	B
5	Joseph	Michael	Johnson	3/25/1995	111111111	B
6	Joey		Johnson	3/25/1995	111111111	C
8	Joseph	Michael	Johnson	3/25/1995	111111111	C

After this value cluster is produced, the scoring method is applied as described previously for positive outcomes.

14.5.1.5 Estimating Precision, Recall, and *F*-Measure

This method uses the output of the manual review to calculate estimates of the total error rate of the ER result. To calculate these estimates, each of the bins are weighted by their portion of the total sample space. The weight is calculated as

$$\text{Bin weight} = \frac{\text{Bin size}}{\text{Total size}}$$

This calculation is performed for all the bins. The manual review process outputs a set of counts of true or false outcomes. These are divided by the sample size for each bin to come up with an estimate of the probability of an error or non-error within that bin. In the case of positive (match) outcomes, the errors are FPs and the non-errors are TPs. For negative (non-match) outcomes, the errors are FNs and the non-errors are TNs. This is calculated as

$$P(\text{error}) = \frac{\text{Nbr of errors}}{\text{Sample size}}$$

$$P(\text{non-error}) = \frac{\text{Nbr of non-errors}}{\text{Sample size}}$$

Finally, the weights are applied to each bin and summed to calculate an estimate for the total ER result across all bins. This is performed for positive outcomes and negative outcomes. This is calculated as

$$P(\text{error}) = \sum_{i=1}^n (\text{Weight}_n \cdot P_n(\text{error}))$$

$$P(\text{non-error}) = \sum_{i=1}^n (\text{Weight}_n \cdot P_n(\text{non-error}))$$

This process can provide estimates for the FP, TP, FN, and TN outcomes. The FP, TP, and FN pair-count estimates can be used for calculating the traditional cluster-based precision, recall, and *F*-measure. In most cases, TN estimates provide no insight into the quality of the data due to the class imbalance of negative and positive outcomes and limitations imposed by using split indicators for aggregating records into value clusters. The split indicators are limited in their recall in a manner similar to the ER results. It is assumed that all pairs of records not generated by the split indicator process have a near-zero probability of being an FN.

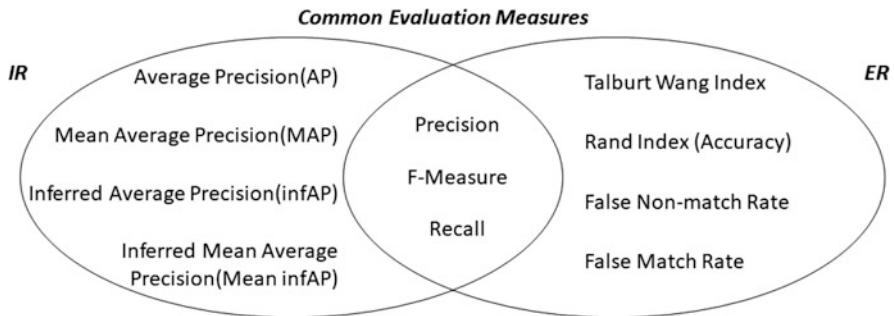


Fig. 14.5 Quality measures

14.5.2 Penning's Method for Inferred Precision

Entity resolution (ER) results are evaluated by classifying and counting record pairs. Specifically, error rates are calculated by counting the number correct and incorrect linking decisions. But if the true links are not known, how can we count them? If we knew the correct links, then ER would be unnecessary. Current methods to address the problem require truth sets, benchmark sets, or pseudo-truth sets. A truth set is made up of known true matches but is only a subset of the population (or ER would be unnecessary) and may not contain all error types [18]. Benchmark sets are the results of previous reliable ER systems but these sets may have undetected errors. Finally, pseudo-truth sets are scored matches where high scores represent high match confidence or true matches and low scores represent low match confidence or true non-matches [28]. These sets have the same issues as benchmark sets. In fact, none of these solutions is ideal [23]. The problem is how to find the error rate without counting. One approach to solve this problem is to borrow techniques from the information retrieval (IR) community.

Information retrieval (IR) is the study of methods of pinpointing information resources that will fit specific information needs. They have faced this same challenge and solved it by using “inferred” measures that are good estimators when datasets are large and the truth is unknown [29]. The problem is in adapting these IR measures to an ER context. In order to borrow these techniques, ER will need to be recast as information retrieval. To begin with an overview of quality measures from both disciplines will be helpful. Figure 14.5 shows the relationship between IR and ER measures.

14.5.2.1 ER Measures

Accuracy

The Rand index is used as an accuracy measure and provides a good general comparison of the portion of correctly assigned results to those that were incorrectly assigned [21]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

where true positives are TP, false positives are FP, true negatives are TN, and false negatives are FN.

F-Measure

F-measure is an approximation of crossover point of precision and recall or the “... is the weighted harmonic mean of precision and recall” [15]. Precision tells us how many retrieved are actually relevant $\text{TP}/(\text{TP} + \text{FP})$, and recall tells us how many relevant were returned $\text{TP}/(\text{TP} + \text{FN})$.

$$F\text{-measure} = \frac{2 (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

Talbur-Wang Index

The Talbur-Wang Index provides a method of ER system comparison that is more efficient since it does not require the TP, FP, TN, and FN counts. Instead it only counts the number of partitions (link groups) and the overlaps between the groups:

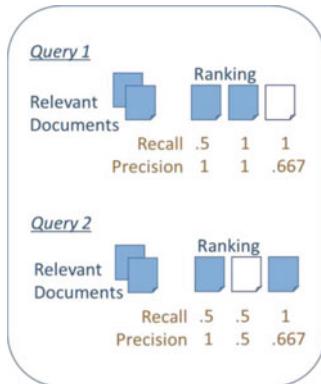
$$TWi(A, B) = \frac{\sqrt{|A| * |B|}}{|V|}$$

where A and B are the counts of linked groups created by the two ER systems and V is the number of overlaps (non-empty intersections) between the groups. The index ranges in value from 0 to 1. The value is 1 if, and only if, the two ER processes link the input references into exactly the same groups [23].

14.5.2.2 IR Measures

Average Precision and Mean Average Precision

Average precision (AP) is the approximation of the area under the precision and recall curve for the top k documents. It can be computed by taking the average of the precision values at relevant documents. Mean average precision (MAP) is a single measurement for an IR system not an individual query. It is the mean of average precision values and is used to provide a single number in order to compare systems. This measure has been developed over decades and has been shown to be a stable discriminator between systems which is used by the IR community extensively [15]. The computation of AP is shown in Fig. 14.6.



$$\text{Query 1} = (1 + 1)/2 = 1$$

$$\text{Query 2} = (1 + .667)/2 = .8335$$

$$\text{MAP} = (1 + .8335)/2 = 0.91675$$

Note that the AP score was higher for query 1 than query 2 – this is due to the ranking of relevant documents

Fig. 14.6 Average precision [19]

Inferred Average Precision and Mean Average Precision

Challenges caused by very large datasets moved the IR community toward development of an inferred measure [25, 29, 30]. In 2005 the inferred version of average precision was developed. Not long after the Text Retrieval Conference (TREC), a well-respected test bed for evaluation of IR systems, run by the US National Institute of Standards and Technology, began using this new measure. AP can be explained as the expected outcome of a random experiment where relevant documents are sampled to provide values for rank levels. For each sampled value, this computation will be recalculated to determine the expected precision values at each of the sampled ranks. The precision values can then be averaged, just as in the descriptive version of AP, to produce an inferred MAP value.

Figure 14.7 shows how this computation is applied. In this example relevant documents are those in positions 1, 3, and 5. The document in position 2 is non-relevant and the document in position 4 is of unknown relevance. Precision and recall are calculated at each ranked position and AP is calculated twice, once assuming the unknown document is relevant and once assuming it is non-relevant resulting in AP values of 0.804 and 0.756. These values demonstrate how the inferred AP value of 0.756 is used as an estimator for AP. Because this example is so small, sampling is done at every relevant k , for a large set sampling is done as needed (Fig. 14.8).

The formula for the estimated precision at rank k is given by

$$E[\text{precision at rank } k] = \frac{1}{k} \cdot 1 + \frac{(k-1)}{k} \left(\frac{|d100|}{k-1} \cdot \frac{|\text{rel}| + \epsilon}{|\text{rel}| + |\text{nonrel}| + 2\epsilon} \right)$$

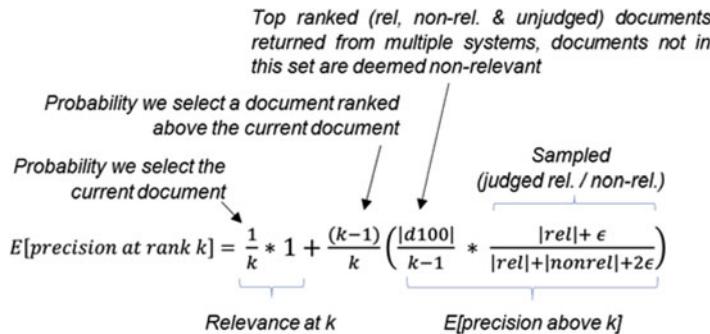


Fig. 14.7 Computation of inferred precision [29]

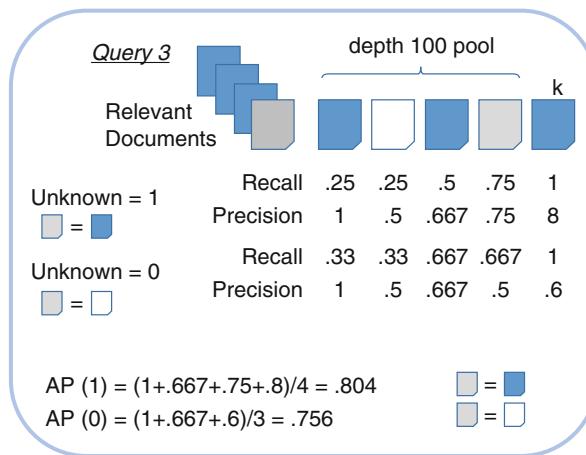


Fig. 14.8 Mean average precision [19]

The estimated precision for ranks 5, 3, and 1 are calculated as

$$E [\text{precision at rank } 5] = \left(\frac{1}{5} \cdot 1 \right) + \left(\frac{(5-1)}{5} \right) \cdot \left(\frac{4}{(5-1)} \cdot \left(\frac{2}{4} \right) \right) = 0.6$$

$$E [\text{precision at rank } 3] = \left(\frac{1}{3} \cdot 1 \right) + \left(\frac{(3-1)}{3} \right) \cdot \left(\frac{2}{(3-1)} \cdot \left(\frac{1}{2} \right) \right) = 0.667$$

$$E [\text{precision at rank } 1] = \left(\frac{1}{1} \cdot 1 \right) + \left(\frac{(3-1)}{3} \right) \cdot \left(\frac{2}{(3-1)} \cdot \left(\frac{1}{2} \right) \right) = 1$$

And finally the inferred average precision is given by

$$\text{inf AP} = \frac{(1 + 0.667 + 0.6)}{3} = 0.756$$

With this background in mind, we can look at how an IR inferred technique can be used for ER error rate estimation.

14.5.2.3 Inferred Average Precision Applied to ER

Evaluating ER quality using inferred AP will require an understanding of ranking as it relates to entity resolution, incomplete truth sets, and low error rates. To recast ER as IR, we will view each linked cluster in the entity resolution results as the equivalent of an IR query results. These sets are comparable in that they both refer to a presumably related set of information—documents in the IR case and single entities in the ER case. Viewing ER in this way ranking will be accomplished by pairwise comparisons of the members of a cluster.

The cluster pairs can be scored based on match quality which will depend on the dataset being processed. There are many possible approaches to scoring match quality and care must be taken to avoid overlap between this scoring system and the ER matching system algorithms. Each pair will receive a score reflecting their match strength and the scores can then be ranked accordingly. Low error rates will be addressed by stratified sampling [20]. Sampling from groups where higher potential for error will performed proportionally to account for low error rates in specific cluster populations.

An example population would be clusters which are split. A split cluster is a single cluster according to one ER system that has been broken into multiple clusters by another. Membership in a split verses a congruent cluster suggests a higher potential for errors. Finally, incomplete pseudo-truth sets, those with unknowns as shown in Fig. 14.7, can be addressed by remembering that the IR query results correspond to the ER cluster. Pairs in a cluster with the highest match score correspond to the judged relevant documents. Remaining pairs below the non-relevant cutoff correspond to the judged non-relevant group. The pairs between these scores correspond to the unjudged group.

- Query (pool) \Rightarrow ER cluster, information on one entity or topic
- Judged relevant \Rightarrow pairs with the highest match score
- Judged non-relevant \Rightarrow pairs falling below a non-relevant cutoff
- Not judged \Rightarrow pairs falling above the cutoff but not having membership in the cluster with the highest match score

14.5.2.4 Implementation

The implementation of inferred average precision for ER is straightforward conceptually and more challenging algorithmically. First, sampling 0.5% of the pairs with random or stratified sampling methods is recommended [28]. Next, rank the cluster pairs based on their similarity score, and choose the scores that will be the break-off points for pseudo-true and pseudo-false as well as unknown [5, 28]. Ranking in this way is not unusual. It can be accomplished using a variety of similarity evaluation methods independently or as part of layered ranking (ranking with multiple methods). All pairs sharing the highest score are “pseudo-true” equivalencies, and those with the lowest score are “pseudo-false” equivalencies [28]. Finally, use these values compute the average precision and the mean average precision. Implementing this algorithm is challenging since data must be maintained on multiple levels simultaneously. Inference is however, a necessity in ER today and provides us with good error estimates.

14.6 An EBDF Data Quality Management Process

Many topics have been discussed, and all are important in understanding the EBDF process. However, the question is how to apply this knowledge in a systematic and comprehensive way to reduce the errors and, consequently, to reduce the risks of an EBDF process. To accomplish this, we propose a structure for an EBDF data quality management process to bring all of these discussion points to bear on the problem.

Product thinking is the key to successful quality management. Product quality management processes, including those for information products, are now well-developed and well-understood. Managing information as product has been a cornerstone of the data quality discipline for the past 20 years based on the research carried out by the MIT Information Quality Program [13, 27]. By taking a product perspective, it has been possible to leverage all of the principles and practices developed for total quality management (TQM) that has revolutionized manufacturing.

The implementation of a data quality management (DQM) process has four components [10]:

1. Data quality planning
2. Data quality control (QC)
3. Data quality assurance (QA)
4. Data quality improvement

14.6.1 Data Quality Planning

The starting point is data quality planning. Data quality planning entails developing and managing data quality requirements, policies, and standards. If you have a high-risk EBDF process, then you must start by identifying nature of the defects in the final product that create the risks and then determine what are the acceptable levels for the rates of these defects. These product requirements become the input for the next component of QC.

14.6.2 Data Quality Control

QC measures the degree to which final products meet data requirements. Basically QC is product inspection. Product inspection should be an ongoing regimen in every organization. When an EBDF process produces too many products for 100% inspection, then QC should be done through sampling using the established techniques of statistical process control [6].

QC is essential for DQM. It helps to prevent releasing defective products to the product users, and it provides a monitor or control to signal when something has gone wrong with the production system. In statistical process control terminology, QC can signal an “out of control” process. While QC is extremely important, it only detects errors, it does not prevent them. The prevention of errors is the role of the QA process.

14.6.3 Data Quality Assurance

QA supports QC by assuring that the component parts of the product meet certain requirements before the final product is assembled, thus helping to prevent making a defective product. While some people tend to use the terms QA and QC interchangeable, they have specific meanings in DQM and quality management in general. QC is inspection of the final product, whereas QA is assuring the component parts of the product fall within expected tolerances. In some sense, QA can be thought of as QC for the product parts, i.e., to what degree are the parts meeting their requirements.

Figure 14.2 provides a QA map for the EBDF process. It shows the four points in the EBDF production process where adherence to tolerances could have the most impact in preventing defects in the final product. The four points are the source information, the blocking process, the ER process, and the rationalization process.

14.6.3.1 QA for Source Information

For source information, QA is typically a collection of data validation rules. These rules test the quality of the values in individual fields for completeness, conformity to required format, timeliness, and other dimensions of data quality relevant to the final product. Other rules may test relationships between related data items, and in relational models, rules to test if the relationships are present and correct.

Validation rules can also be longitudinal. To help detect unexpected changes in source file, each source file is profiled when it is received. Profiling collects the statistics on the file content, such as the frequency of each distinct value for an attribute, the number of missing values for an attribute, and the largest and smallest values of an attribute. When comparing the profile of an incoming file with profiles of previously received files from the same source, large variations in the profile statistics could indicate the layout or content of the file has changed or the wrong file has been sent in error.

14.6.3.2 QA for the Blocking Process

QA for blocking requires is typically done by running the blocking and ER process against benchmark test data to see if the expected results are achieved. QA for blocking tends to be more episodic than ongoing. Events such as modifications to the blocking process or modifications to the matching rules should be triggers for QA for blocking.

As noted earlier, if the blocking process is in complete alignment with the matching rules, then running the ER process on a sample of records with and without blocking should yield the same clustering of the input. The test set used for blocking QA should contain at least one pair of records to exercise each match condition.

14.6.3.3 QA for the ER Process

QA for the ER process is the most extensive. It has three major components, benchmark or regression testing, clerical review, and periodic accuracy measurement. Just as with blocking, whenever there are changes to the upstream systems such as adding a new data source or a change in blocking rules, the ER process should undergo a complete regression testing regimen with benchmark data to assure all expected results are achieved. For high-risk EBDF processes, there should be two additional QA checks for the ER step.

The second QA process is clerical review. For a high-risk EBDF process it is essential for the ER process to produce clerical review indicators. Clerical review indications can be produced during the ER process or produced by a post-ER process examining the cluster produced after the ER process. As noted earlier, scoring-based ER engines can easily product clerical review indications by simply noting when a pair of references produce a score very close the match threshold.

Run-time review indicators are possible for Boolean rule systems as well. The indicators are produced by incorporating a secondary set of “soft” Boolean rules. These review rules are soft in the sense they impose less stringent match conditions, i.e., a near match or partial match. When a pair of references fails to trigger a normal linking rule, but triggers one of the review rules, the pair is output for clerical review. The main issue with soft Boolean rules is they only indicate for false negative pairs and not for false positive. False positive indications for Boolean rules are usually generated by a post-ER process.

There are a number of techniques for generating post-ER review indicators. These are programs that run after the ER process and examine the clusters of linked records created by the ER process. Most post-ER indicators are based measuring the entropy of a cluster or a merged pair of clusters. The entropy of a cluster represented the degree of organization or consistency of the cluster values. If the entropy of a cluster is low, it means that the values of the references in the cluster are organized in the sense most of the values for the same attribute agree with each other. However, if the entropy is high, it signals more disorganization within the cluster, i.e., many different values for the same attribute across the references in the cluster. Clusters with very high entropy could be an indication the ER system has linked together non-equivalent references, i.e., an indication of a false positive error. On the other hand, if two clusters are merged together and the merged cluster has a very low entropy, then it could indicate all of the references should be linked together in the same cluster, i.e., an indication of a false negative error. The entropy-based scoring model is very similar to the Pullen process for estimating precision and recall discussed in the section Measuring ER Accuracy.

The third QA process for ER is the periodic and systematic estimation of the ER system’s precision and recall. These estimates are essential for monitoring and controlling ER accuracy. For large data applications, the Pullen method and the Penning method described in section Measuring ER Accuracy can be used to produce these measures.

14.6.3.4 QA for Data Rationalization

The QA for data rationalization depends heavily on the application context. Because data rationalization is the final step of the EBDF process, it is at or near point at which the final information product is generated. For this reason, the QC process of inspecting the final products can also serve as one of the QA processes for this step. Just as with the preceding points of QA, data rationalization needs its own suite of regression and validation tests to be applied when there are changes to any of the three upstream processes, i.e., sourcing, blocking, and ER.

14.6.4 Data Quality Improvement

While QC and QA detect and monitor data quality problems, the overall goal of a DQM process is to continually improve the quality of the final product. The data quality improvement process has three main parts. The first is error reporting and collection, the second is root cause analysis, and the third is the implementation of process improvements.

Perhaps one the most critical parts of the data quality improvement process is collecting and documenting the errors. All errors across the entire EBDF process should be collected into one central location. Systemic errors can often escape detection when errors are collected and addressed in separate systems.

Error reporting should be organized and comprehensive. There should also be a well-understood classification scheme for the type and location of the error along with supporting metadata. Error metadata could include information about the person reporting the error, when it was reported, the severity/priority of the problem, and any error messages that were produced. If immediate corrective action was taken for the error, then the data governance policy and procedure under which the correction was made should be reported as well. In organization under a data governance program, error correction should only be done by responsible data stewards following approved data governance policies and procedures.

After reporting, the second part of improvement is the periodic and systematic review and analysis of the error reported. The primary goals of an error log review are to uncover the root cause(s) of each error and to classify errors with the same causes into groups.

The third part of improvement is to prioritize each error group by considering the impact of its errors on the EBDF process versus the time and cost necessary to correct and prevent the errors in the future. The prioritization process requires input from both IT and the business to understand the benefits and costs for each group.

Once the prioritization is complete, the final step is to create a project plan for correcting and preventing the errors with the highest priority. Each improvement project should be validated at this conclusion to determine the degree to which the projected benefits from error reduction were attained versus the actual time and cost of the project.

14.7 Conclusion

In this chapter we have discussed the EBDF process and the primary sources of EBDF errors. We have also discussed and suggested some of the best practices for monitoring and reducing these errors in high-risk applications. Finally, we have laid out the fundamental components of the data quality management plan for continual data and process improvement.

However, more progress is needed. Most of the processes discussed rely too much on human intervention. Continued research and case studies are needed to bring to bear new methods and technologies to address these problems through increased automation.

Two areas of particular interest for research are methods and technologies to efficiently perform ER and data fusion in large-scale, distributed processing environments. Another centers on how to take full advantage of machine learning and other AI techniques at the frontiers of ER and data fusion research to improve current QC and QA assurance processes and to develop new processes.

References

1. P. Christen, Febrl- A freely available record linkage system with a graphical user interface, in *Proceedings of the Australian Workshop on Health Data and Knowledge Management (HDKM)*. Conferences in research and practice in information technology (CRPIT), Wollongong, January 2008, vol. 80
2. P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection* (Springer, New York, 2012)
3. A. Doan, A. Halevy, Z. Ives, *Principles of Data Integration* (Morgan Kaufmann, Waltham, 2012)
4. A. Eram, A.G. Mohammed, V. Pillai, J.R. Talburt, Comparing the effectiveness of deterministic matching with probabilistic matching for entity resolution of student enrollment records, in *22nd MIT International Conference on Information Quality (ICIQ-2017)*, Little Rock, 6–7 October 2017, pp. 14:1–14:12
5. I. Fellegi, A. Sunter, A theory for record linkage. *J. Am. Stat. Assoc.* **64**(328), 1183–1210 (1969)
6. C. Fisher, E. Lauria, S. Chengalur-Smith, R. Wang, *Introduction to Information Quality* (MIT Information Quality Program, Cambridge, MA, 2008)
7. T.N. Herzog, F.J. Scheuren, W.E. Winkler, *Data Quality and Record Linkage Techniques* (Springer, New York, 2007)
8. G. Holland, J.R. Talburt, A framework for evaluating information source interactions, in *2008 Conference on Applied Research in Information Technology*, ed. by C. Hu, D. Berleant (University of Central Arkansas, Conway, 2008), pp. 13–19. <http://research.acxiom.com/publications.html>
9. G. Holland, J.R. Talburt, An entity-based integration framework for modeling and evaluating data enhancement products. *J. Comput. Sci. Coll.* **24**(5), 65–73 (2010)
10. ISO 8000-Part 61, *Data Quality Management: Process Reference Model* (ISO copyright office, Geneva, 2016)
11. F. Kobayashi, A. Eram, J. Talburt, Entity resolution using logistic regression as an extension to the rule-based OYSTER system, in *Proceedings: IEEE International Conference on Multimedia Information Processing and Retrieval (IEEE MIPR 2018)*, Miami, 10–12 April 2018 (accepted for publication)
12. E. Lawley, Building a health data hub. March 29, 2010. Nashville Post (online version, downloaded July 24, 2010)
13. Y.W. Lee, L.L. Pipino, J.D. Funk, R.Y. Wang, *Journey to Data Quality* (MIT Press, Cambridge, MA, 2006)
14. D. Mahata, J.R. Talburt, A framework for collecting and managing entity identity information from social media, in *19th MIT International Conference on Information Quality*, Xi'an, 1–3 August, 2014, pp. 216–233

15. C.D. Manning, P. Raghavan, H. Schütze, *An Introduction to Information Retrieval* (Cambridge University Press, Cambridge, England, 2009)
16. E. Nelson, J.R. Talburt, Improving the quality of law enforcement information through entity resolution, in *2008 Conference on Applied Research in Information Technology*, ed. by C. Hu, D. Berleant (University of Central Arkansas, Conway, 2008), pp. 113–118. <http://research.acxiom.com/publications.html>
17. E. Nelson, J.R. Talburt, Entity resolution for longitudinal studies in education using OYSTER, in *Proceedings: 2011 Information and Knowledge Engineering Conference* (IKE 2011), Las Vegas, 18–20 July 2011, pp. 286–290
18. M. Penning, J.R. Talburt, Information quality assessment and improvement of student information in the university environment, in *The 2012 International Conference on Information and Knowledge Engineering* (IKE'12), Las Vegas, 16–29 July 2012, pp. 351–357
19. M. Penning, Inferred error rates for entity resolution, Doctoral Dissertation, University of Arkansas at Little Rock, Published by Proquest, 2016
20. D. Pullen, A system for stratified sampling of entity resolution results to assess and improve accuracy with minimal clerical review, Doctoral dissertation, University of Arkansas at Little Rock, Published by Proquest, 2017
21. W.M. Rand, Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**(336), 846–850 (1971)
22. J.R. Talburt, R. Hashemi, A formal framework for defining entity-based, data source integration, in *2008 International Conference on Information and Knowledge Engineering*, ed. by H. Arabnia, R. Hashemi (CSREA Press, Las Vegas, 2008), pp. 394–398
23. J.R. Talburt, Y. Zhou, *Entity Information Life Cycle for Big Data: Master Data Management and Information Integrations* (Morgan Kaufmann, Waltham, 2015)
24. J.R. Talburt, *Entity Resolution and Information Quality* (Morgan Kaufmann, San Francisco, 2011)
25. E.M. Voorhees, W. Hersh, Overview of the TREC 2012 medical records track, in *The Twenty-First Text Retrieval Conference (TREC 2012) Proceedings, National Institute of Standards and Technology*, 2012
26. P. Wang, D. Pullen, J.R. Talburt, N. Wu, Iterative approach to weight calculation in probabilistic entity resolution, in *2014 International Conference on Information Quality*, Xi'an, 1–3 August 2014
27. R.Y. Wang, A product perspective on total data quality management. *Commun. ACM* **41**(2), 58–65 (1998)
28. W.E. Winkler, *Automatically Estimating Record Linkage False Match Rates* (Census Bureau, Statistical Research Division, Washington, DC, 2007)
29. E. Yilmaz, J.A. Aslam, Estimating average precision with incomplete and imperfect judgments, in *Proceedings of the Fifteenth ACM International Conference on Information and Knowledge Management*, ACM Press, New York, NY, 2006
30. E. Yilmaz, E. Kanoulas, J.A. Aslam, A simple and efficient sampling method for estimating AP and NDCG, in *Proceedings of the Thirty-First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, Singapore, 2008
31. Y. Zhou, A. Kooshesh, J. Talburt, Optimizing the accuracy of entity-based data integration of multiple data sources using genetic programming methods. *Int. J. Bus. Intell. Res.* **3**(1), 72–82 (2012)
32. Y. Zhou, J. Talburt, Y. Su, L. Yin, OYSTER: a tool for entity resolution in health information exchange, in *5th International Conference on the Cooperation and Promotion of Information Resources in Science and Technology* (COINFO'10), Beijing, 27–29 November 2010, pp. 356–362
33. Y. Zhou, J.R. Talburt, Entity identity information management, in *International Conference on Information Quality 2011*, Adelaide, 18–20 November 2011, electronic proceedings at: http://iciq2011.unisa.edu.au/doc/ICIQ2011_Proceeding_Nov.zip

Part II

Aspects of Information Quality in Various Domains of Application

Chapter 15

Decision-Aid Methods Based on Belief Function Theory with Application to Torrent Protection



Simon Carladous, Jean-Marc Tacnet, Jean Dezert,
and Mireille Batton-Hubert

Abstract In mountainous areas, decision-makers must find the best solution to protect elements-at-torrential risk. The decision process involves several criteria and is based on imperfect information. Classical Multi-Criteria Decision-Aiding methods (MCDAs) are restricted to precise criteria evaluation for decision-making under a risky environment and suffer of rank reversal problems. To bridge these gaps, several MCDAs have been recently developed within belief function theory framework. The aims of this chapter are to introduce how these methods can be applied in practice and to introduce their general principles. To show their applicability to the real-life problem, we apply them to the Decision-Making Problem (DMP) comprising the comparison of several protective alternatives against torrential floods and selection of the most efficient one. We finally discuss the method improvements to promote their practical implementation.

Keywords Decision under uncertainty · Imperfect and conflicting information · Multicriteria decision analysis · Belief functions · Risk analysis and mitigation · Torrent protection

S. Carladous (✉)

Département Risques Naturels (DRN), Office National des Forêts (ONF), Grenoble, France
e-mail: simon.carladous@onf.fr

J.-M. Tacnet

Snow Avalanche Engineering and Torrent Control Research Unit (ETNA), Université Grenoble Alpes, Irstea – UR ETGR, St-Martin d'Hères Cedex, France
e-mail: jean-marc.tacnet@irstea.fr

J. Dezert

ONERA, The French Aerospace Lab, Palaiseau Cedex, France
e-mail: jean.dezert@onera.fr

M. Batton-Hubert

Institut Henri Fayol, UMR LIMOS 6158, Ecole Nationale Supérieure des Mines de Saint-Etienne, Saint-Etienne Cedex 2, France
e-mail: mbatton@emse.fr; Mireille.BATTON-HUBERT@emse.fr

15.1 Introduction

Mountainous torrential floods are different from plains' floods because of high flow velocity and high concentration of materials in flowing. Materials come from headwaters, are transported in a channel by debris flows or bedload transport, and finally spread on the alluvial fan. As shown in Fig. 15.1, they put people, buildings, and networks at risk.

Protective systems aim is to reduce damage on elements-at-risk. Therefore, they have specific functions. For instance, the check dam series maintain materials in headwaters whereas sediment traps stop materials before they reach elements-at-risk [1]. In practice, risk managers decide on actions based on several criteria, for example, cost vs. damage reduction. An example of a practical Decision-Making Problem (DMP) is given in Fig. 15.2. The goal is to compare several potential protective actions A_i within a torrential watershed (i) to assign each alternative to a class (or label) as classically done by experts, or (ii) to rank all alternatives according to a preference order, or (iii) to choose the best alternative [2].

Classical Multi-Criteria Decision-Aiding methods (MCDAs) such as Analytic Hierarchy Process (AHP) [3], Technique for Order Preference by Similarity to the Ideal Solution (TOPSIS) [4], and Cost-Benefit Analysis (CBA) [5] help to make decision on such Multi-Criteria Decision-Making (MCDM) problems. While evaluations of criteria in practice are done with different units and scales, imperfect, provided by more or less reliable sources, and made under an epistemically uncertain environment [6], classical MCDAs only consider perfect criteria evaluation, suffer from rank reversal problems, and are limited to decisions under a risky environment.

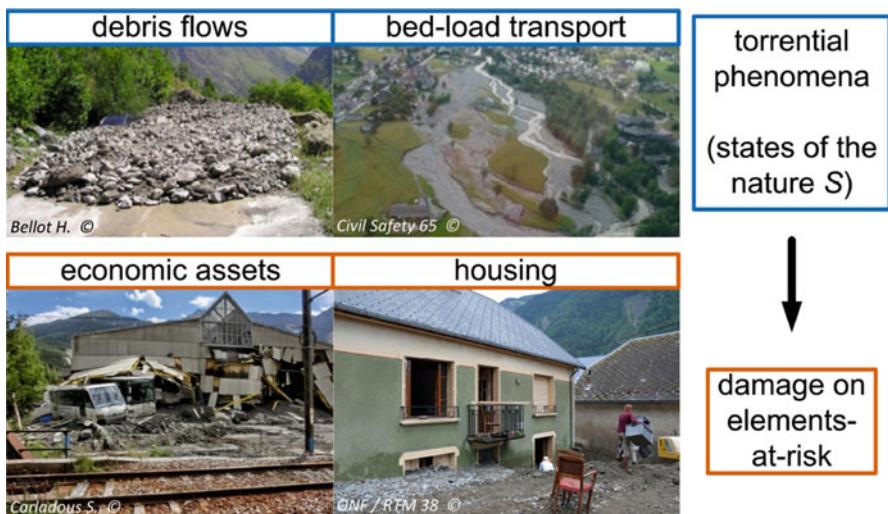


Fig. 15.1 Torrential phenomena and examples of elements-at-risk

To bridge these gaps, several MCDA were developed within belief function theory framework. As for any MCDA, the first step is to specify the DMP, potential alternatives, decision criteria [2], their scoring scale, and their importance weights. To assign each alternative to a qualitative label (i.e., good, bad, very bad, etc.), Evidential Reasoning for Multi-Criteria Decision Analysis (ER-MCDA) approach extends the AHP by taking into account imperfect evaluation of each criterion provided by several sources. Belief function theory is coupled [7] with fuzzy sets [8] and possibility theories [9]. Belief Function-based Technique for Order Preference by Similarity to the Ideal Solution (BF-TOPSIS) methods are more robust to rank reversal phenomena to rank all alternatives than classical MCDA [10]. Cautious Ordered Weighted Averaging with Evidential Reasoning (COWA-ER) [11] and Fuzzy COWA-ER (FCOWA-ER) [12] improve initial OWA [13] to help to make decision under an epistemically uncertain environment.

This chapter shows how these new methods can be combined and applied in practice. Therefore, Sect. 15.2 not only recalls basics of MCDM problems and decision-making under uncertainty but also basics of fuzzy set, possibility, and belief function theories. Section 15.3 introduces general principles of ER-MCDA, BF-TOPSIS, and FCOWA-ER methods. They are then applied in Sect. 15.4 to the same DMP introduced in Fig. 15.2. In Sect. 15.5, we finally discuss needed improvements to encourage their practical implementation.

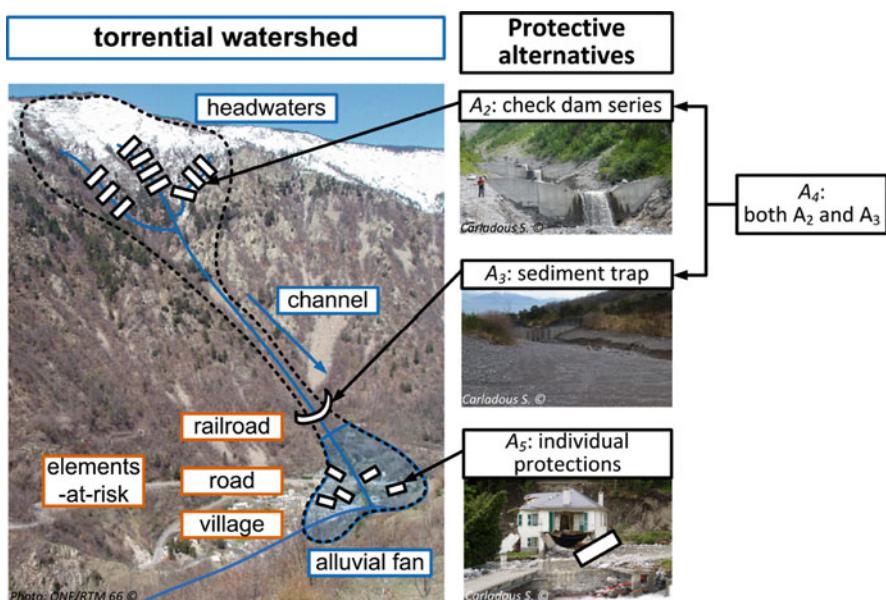


Fig. 15.2 A real-life DMP within a torrential watershed

15.2 Basics of Decision-Making and Imperfect Information

This section provides a formal description of the DMPs, models of representing imperfect information, and decision methods given it.

15.2.1 Formalization of Decision-Making Problems

Any DMP is about comparing M alternatives $A_i \in \mathcal{A}$ and selecting the best one. A decision-maker (DM) faces a MCDM problem if decision depends on several criteria g_j , $j = 1, \dots, N$. A set \mathcal{S} represents the states of the nature. Since the beginning of the twentieth century, it has been proposed to distinguish decision-making under risk from decision-making under uncertainty, given the DM knowledge on \mathcal{S} [14]. This subsection introduces related formalisms to represent the whole MCDM problem under uncertainty.

15.2.1.1 Multi-criteria Decision-Making Problem

A DM assigns an importance weight ω_j to each criterion g_j , $j = 1, \dots, N$. Respecting the condition $\sum_{j=1}^N \omega_j = 1$, the vector $\mathbf{w} = [\omega_1, \dots, \omega_j, \dots, \omega_N]$ represents the DM preference over these criteria. For each g_j , a specific scoring scale X_j is defined. The DM scores each alternative A_i based on each g_j . This score is denoted $x_{ij} \in X_j$. The DM eventually provides the $M \times N$ score matrix $\mathbf{S} = [x_{ij}]$ defined by Eq. (15.1) [15].

$$\mathbf{S} \triangleq A_i \begin{pmatrix} g_1, \omega_1 & \dots & g_j, \omega_j & \dots & g_N, \omega_N \\ x_{11} & \dots & x_{1j} & \dots & x_{1N} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{iN} \\ \vdots & & \vdots & & \vdots \\ x_{M1} & \dots & x_{Mj} & \dots & x_{MN} \end{pmatrix} \quad (15.1)$$

15.2.1.2 Decision-Maker Preferences

The DM has preferences not only over criteria but also to compare alternatives according to each criterion. First, AHP helps to establish the N -vector \mathbf{w} which represents the DM preferences over criteria by comparing criteria pairwisely [3]. Second, DM has a preference ordering between all alternatives $A_i \in \mathcal{A}$, given a scoring scale X_j for each g_j . For instance, considering three alternatives A_1 , A_2 , and A_3 and their scores $x_{1j}, x_{2j}, x_{3j} \in X_j$, the DM preference can be represented by:

- a total pre-order which assumes preference transitivity: if $A_1 \succ A_2$ and $A_2 \succ A_3$, then $A_1 \succ A_3$ [16].
- a partial pre-order which relaxes transitivity assumption [2].

15.2.1.3 From Decision-Making Under Risk to Decision-Making Under Uncertainty

In practice, torrential hazard is generally represented by a finite set of states of the nature $\mathcal{S} = \{S_1, \dots, S_k, \dots, S_K\}$, as recalled in Fig. 15.1. Each S_k is commonly referred to as scenario [17]. Given S_k , C_{ik} is the global payoff expected for each alternative A_i . DM provides the $M \times K$ payoff matrix $\mathbf{C} = [C_{ik}]$ defined by Eq. (15.2) [11].

Given \mathbf{C} , decision-making depends on the DM knowledge on \mathcal{S} [11]:

- Decision-making under certainty: since only one S_k is known and certain to occur, it consists in choosing the best A_{i^*} with $i^* \triangleq \arg \max_i \{C_{ik}\}$.
- Decision-making under risk (or aleatory uncertainty): the true state of the nature is unknown, but one knows all the probabilities $p_k = P(S_k)$. In the context of natural hazards, the expected payoff $E[C_i] = \sum_k p_k \cdot C_{ik}$ is generally computed for each A_i . The best A_{i^*} is with $i^* \triangleq \arg \max_i \{E[C_i]\}$.
- Decision-making under ignorance: one assumes no knowledge about the true state of the nature but that it belongs to \mathcal{S} . Yager's Ordered Weighted Averaging (OWA) approach [13] can be used to make a decision in this context.
- Decision-making under uncertainty: a belief structure characterizes the knowledge on S . In practice, this is the closest representation of torrential hazard knowledge. Its elicitation by subjective probabilities $p_k = P(S_k)$ is usually used. Thus, decision-making is similar to decision-making under risk [18]. A more interesting approach is the OWA proposed by Yager [13] and improved in [11, 12].

$$\mathbf{C} \triangleq \begin{matrix} & S_1 & \dots & S_k & \dots & S_K \\ A_1 & \left(\begin{matrix} C_{11} & \dots & C_{1k} & \dots & C_{1K} \\ \vdots & & \vdots & & \vdots \\ C_{i1} & \dots & C_{ik} & \dots & C_{iK} \\ \vdots & & \vdots & & \vdots \\ C_{M1} & \dots & C_{Mk} & \dots & C_{MK} \end{matrix} \right) \\ A_i & & & & & \\ A_M & & & & & \end{matrix} \quad (15.2)$$

Formalisms of MCDM problems and decision-making under uncertainty are complementary in representing the DMP. A multi-criteria aggregation based on \mathbf{S} can give each vector $\mathbf{C}_k \triangleq [C_{1k}, \dots, C_{ik}, \dots, C_{Mk}]$ of \mathbf{C} . For each A_i and g_j , computing the expected payoff $E[C_i]_j$ from \mathbf{C} can provide $x_{ij} = E[C_i]_j$ in \mathbf{S} [6].

15.2.1.4 Several Types of Imperfect Information

Whatever the DMP, decisions depend on the quality of information used to assess scores x_{ij} ($i = 1, \dots, M$; $j = 1, \dots, N$), payoffs C_{ik} , and states of the nature S_k ($k = 1, \dots, K$). There are various types of information imperfection [19]:

- inconsistency, which is related to conflict between sources such as several experts;
- imprecision referring, for example, to interval of numerical values;
- incompleteness, which represents the lack of information while data exist;
- aleatory uncertainty referring to aleatory events;
- epistemic uncertainty, which is linked to the lack of knowledge.

In practice, probabilities are usually used to represent imperfect information. A first criticism is their limit¹ to represent uncertainty, while other formalisms are available: sets for imprecision, fuzzy sets for vagueness [8], possibility distributions, and imprecise probabilities for both uncertainty and imprecision [9, 27]. A second criticism is the use of subjective probabilities [18] both to decide under ignorance [20] and to represent the DM attitude with few information [21]. Belief function theory allows taking into account all types of imperfect information but also to make decisions under ignorance and epistemic uncertainty [22].

15.2.1.5 What Is the Decision-Making Problem About?

A DMP is about comparing the M alternatives A_i gathered in the set \mathcal{A} . In practice, three different objectives can be given [2]. For instance, the aim is to compare $M = 4$ potential protective actions A_i within a torrential watershed based on their efficiency:

1. to assign each A_i to a predefined qualitative class (or label) of efficiency such as “high,” “medium,” “low,” and “none” [23];
2. to rank all A_i , $i = 1, \dots, M$, totally or partially: for instance, $A_3 \succ A_4 \succ A_1 \succ A_2$ is a total order, while $A_3 \sim A_4 \succ A_1 \sim A_2$ is a partial order;
3. to choose the best alternative $A_{i^*} \in \mathcal{A}$, for instance, A_3 .

¹Indeed, the ignorance of a parameter value x belonging to $[a, b]$ is usually modeled by a uniform probability distribution function (pdf) over $[a, b]$, which yields from the probability calculus to a nonuniform pdf of $1/x$ on $[1/b; 1/a]$. This result is not acceptable from the ignorance modeling standpoint because if one has no specific information on x , we cannot get more information on $1/x$ but that $1/x$ belongs to $[1/b; 1/a]$. Therefore the uniform pdf often used to model ignorance in probability theory is problematic.

15.2.2 Imperfect Information: From Representation to Decision-Making

This subsection details the three main steps to take into account imperfect information: (1) representation, (2) combination and propagation, and (3) decision [24].

15.2.2.1 Representation of Imperfect Information Provided by a Source

- **Fuzzy set theory** was developed to represent linguistic assessment of fuzziness [8]. Given individual elements x of the universe of discourse X , the membership function $\mu_\theta(x) \in [0, 1]$ associates each $x \in X$ to the fuzzy set θ with the grade of membership $\mu_\theta(x)$. As shown in Fig. 15.3a, a simple way to represent a membership function is to use a trapezoidal membership function defined by the quadruplet $\{a, b, c, d\}$ in Eq. (15.3) [25]: $[a, d]$ is the fuzzy set support denoted by $supp_\theta$, while $[b, c]$ is its core $core_\theta$. Given X , $\bar{\theta}$ is the complement fuzzy set of θ defined by Eq. (15.4) (Fig. 15.3a), and a mapping model [19] is a set Θ of n fuzzy sets θ_e , for $e = 1, \dots, n$ (Fig. 15.3b). Given two fuzzy sets θ_1 and θ_2 , the membership function $\mu_{\theta_1 \cup \theta_2}$ defined by Eq. (15.5) represents their union (Fig. 15.3c), while their intersection $\mu_{\theta_1 \cap \theta_2}$ is defined by Eq. (15.6) [8].

$$\begin{cases} 0 & \text{if } x \notin supp_\theta \\ \frac{x-a}{b-a} & \text{if } x \in [a, b] \\ 1 & \text{if } x \in core_\theta \\ \frac{x-d}{c-d} & \text{if } x \in [c, d] \end{cases} \quad (15.3)$$

$$\mu_{\bar{\theta}}(x) \triangleq 1 - \mu_\theta(x), x \in X \quad (15.4)$$

$$\mu_{\theta_1 \cup \theta_2}(x) \triangleq \max_{x \in X}(\mu_{\theta_1}(x), \mu_{\theta_2}(x)) \quad (15.5)$$

$$\mu_{\theta_1 \cap \theta_2}(x) \triangleq \min_{x \in X}(\mu_{\theta_1}(x), \mu_{\theta_2}(x)) \quad (15.6)$$

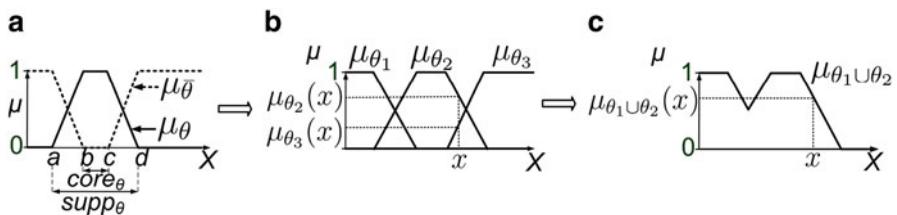


Fig. 15.3 (a) Fuzzy set Θ and its complement $\bar{\Theta}$; (b) mapping model ($n = 3$); (c) union

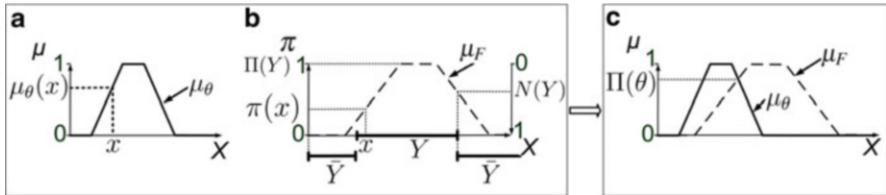


Fig. 15.4 (a) Fuzzy set Θ ; (b) possibility distribution $\pi = \mu_F$; (c) possibility measure of Θ given π

- Zadeh, Dubois, and Prade then developed the *possibility theory* in the fuzzy logic framework [9, 27]. Considering the fuzzy set F of possible values of $x \in X$, the possibility distribution π is given by $\mu_F(x) \triangleq \pi(x) \in [0, 1]$ [26]. Given Y a subset of X and \bar{Y} its complement, the possibility and necessity measures are $\Pi(Y) \triangleq \sup_{x \in Y} \pi(x)$ [9] and $N(Y) \triangleq 1 - \Pi(\bar{Y})$, $\forall Y, \bar{Y} \subseteq X$ [27], as shown in Fig. 15.4b. $\Pi(Y)$ and $N(Y)$ are considered as the upper and lower bounds of the probability $P(Y)$. Given X , the membership function μ_θ , and a possibility distribution μ_F (Fig. 15.4c), the possibility measure of θ denoted by $\Pi(\theta)$ is defined by Eq. (15.7) [9].

$$\Pi(\theta) \triangleq \sup_{x \in X} \mu_{\theta \cap F}(x) \quad (15.7)$$

- In the meantime, Shafer introduced the *belief function theory*, also called Dempster-Shafer Theory (DST) [22]. The Frame of Discernment (FoD) is a finite set $\Theta = \{\theta_1, \dots, \theta_e, \dots, \theta_n\}$, with $n > 1$, which gathers the potential answers of the DMP under concern. In DST, all FoD elements are assumed exhaustive and mutually exclusive. The power set of Θ denoted by 2^Θ is the set of all subsets of Θ , the empty set \emptyset included. The complement of a subset $A \in 2^\Theta$ is denoted \bar{A} . Its cardinality is $|A|$. A source (or body) of evidence is characterized by a basic belief assignment (BBA) $m^\Theta(\cdot)$, which is a mapping $m^\Theta(\cdot) : 2^\Theta \rightarrow [0, 1]$ that satisfies $m^\Theta(\emptyset) = 0$, and $\forall A \neq \emptyset \in 2^\Theta$ the condition $\sum_{A \subseteq \Theta} m^\Theta(A) = 1$. The vacuous BBA models the full ignorance of the source of evidence. If $m^\Theta(A) > 0$, A is a focal element of $m^\Theta(\cdot)$. m_A^Θ denotes the categorical BBA which focuses on $A \neq \emptyset$. More precisely, $m_A^\Theta(A) = 1$ and $m_A^\Theta(Y) = 0$ for any $Y \neq A$. Focal elements of a Bayesian BBA are only singletons on 2^Θ . Given the FoD Θ for final decision, it is possible to represent imperfect evaluation of the score x_{ij} , for each alternative A_i according to each criterion g_j , through the BBA $m_{ij}^\Theta(\cdot)$ providing the $M \times N$ BBA matrix $\mathbf{M}^\Theta = [m_{ij}^\Theta(\cdot)]$. Thus, it must be compared with the $M \times N$ score matrix $\mathbf{S} = [x_{ij}]$, defined by Eq. (15.1). Given $m^\Theta(\cdot)$, belief and plausibility functions are, respectively, defined by:

$$\text{Bel}^{\Theta}(A) \triangleq \sum_{Y \subseteq A | Y \in 2^{\Theta}} m^{\Theta}(Y) \quad (15.8)$$

$$\text{Pl}^{\Theta}(A) \triangleq \sum_{Y \cap A \neq \emptyset | Y \in 2^{\Theta}} m^{\Theta}(Y) \quad (15.9)$$

Considering that the universe of discourse X is the FoD Θ , the plausibility measure $\text{Pl}^{\Theta=X}(A)$ is a possibility measure $\Pi(A)$, $\forall A \subseteq \Theta = X$ [26]. $\text{Bel}^{\Theta}(A)$ and $\text{Pl}^{\Theta}(A)$ are interpreted as lower and upper bounds of the unknown probability $P^{\Theta}(A)$. The interval $\text{BI}^{\Theta}(A) \triangleq [\text{Bel}^{\Theta}(A), \text{Pl}^{\Theta}(A)]$ is its belief interval. Its length $\text{Pl}^{\Theta}(A) - \text{Bel}^{\Theta}(A)$ characterizes the uncertainty, also called ambiguity, on A [28].

Given $m^{\Theta}(\cdot)$, several transformations help to approximate probability function $P^{\Theta}(\cdot) : \text{Smets' pignistic transformation provides } \text{BetP}^{\Theta}(\cdot)$ [29], DSmP transformation gives $\text{DSmP}_{\epsilon}^{\Theta}(\cdot)$ where $\epsilon \geq 0$ is a tuning parameter [30] (Vol. 3), and others.

Shafer's exhaustivity assumption means that the FoD is considered as a "closed world" (*c.w.*). In some practical problems, this assumption is too strict and it is more convenient to consider the original FoD as an "open world" (*o.w.*).

1. In Smets' Transferable Belief Model (TBM) [31], $\Theta^{o.w.} \triangleq \{\theta_1, \dots, \theta_q\}$ and $\emptyset = \bar{\Theta}^{o.w.}$. One has $\sum_{A \in 2^{\Theta}} m(A) = 1$, and one allows $m(\emptyset) \geq 0$.
2. In Yager's approach [32], the open world is closed by an hedge element θ^c , so that $\Theta^{c.w.} \triangleq \Theta^{o.w.} \cup \{\theta^c\}$. The cardinality $|\theta^c|$ is not known.

Shafer's mutual exclusivity assumption can be also too strict. Dezert-Smarandache Theory (DSmT) framework modifies DST to relax this assumption and proposes new techniques to combine the sources of evidence and to make a decision [30].

15.2.2.2 Combining Information Provided by Several Sources of Evidence

First of all, the source reliability and its importance must be clearly distinguished. Reliability is the source objective ability to give the correct solution of the DMP [33]. For each source s_q , $q = 1, \dots, Q$, it is represented by a reliability discounting factor $\alpha_q \in [0, 1]$ [34]. Given the initial BBA $m_q^{\Theta}(\cdot)$ provided by s_q , Shafer's discounting method defined by Eq. (15.10) is generally used to provide the discounted mass $m_{\alpha_q}^{\Theta}(A)$ [22].

$$m_{\alpha_q}^{\Theta}(A) \triangleq \begin{cases} \alpha_q \cdot m_q^{\Theta}(A) & \text{if } A \in 2^{\Theta} \neq \Theta \\ \alpha_q \cdot m_q^{\Theta}(A) + (1 - \alpha_q) & \text{if } A = \Theta \end{cases} \quad (15.10)$$

Importance is the subjective weight granted to the source by DM [33]. In a MCDM problem, each criterion can be considered as a source represented by a BBA $m_j^\Theta(\cdot)$. Each weight ω_j is the importance discounting factor used to provide the discounted BBA² $m_{\omega_j}^\Theta(\cdot)$.

$$m_{\omega_j}^\Theta(A) \triangleq \begin{cases} \omega_j \cdot m_j^\Theta(A) & \text{if } A \in 2^\Theta \neq \emptyset \\ \omega_j \cdot m_j^\Theta(A) + (1 - \omega_j) & \text{if } A = \emptyset \end{cases} \quad (15.11)$$

Once BBAs have been discounted, the combination of distinct sources of evidence is denoted by \oplus to provide the combined BBA $m_\oplus^\Theta(A)$, $A \subset \Theta$. The largely used initial Dempster's rule (*DS*) [22] has been subject to strong debates in fusion community, showing it does not behave well in high conflicting case [35] but also in low conflicting cases [36].

As a consequence, since the 1990s, many alternatives have been proposed to combine belief functions more or less efficiently. The Proportional Conflict Redistribution (PCR) rules have been developed in DSMT [30] (Vol. 3) to palliate disadvantages of the classical Dempster's fusion rule [37]. PCR rule n° 6 (PCR6) defined by Eq. (15.12) for combining two sources of evidence ($K = 2$) $m_1^\Theta(\cdot)$ and $m_2^\Theta(\cdot)$ is also consistent for more than two bodies of evidence ($K > 2$) [38].

$$\begin{aligned} m_{\text{PCR6}}^\Theta(A) &\triangleq \sum_{\substack{X_1, X_2 \in 2^\Theta \\ X_1 \cap X_2 = A}} m_1^\Theta(X_1) \cdot m_2^\Theta(X_2) \\ &+ \sum_{\substack{Y \in 2^\Theta \setminus \{A\} \\ A \cap Y = \emptyset}} \left[\frac{m_1^\Theta(A)^2 \cdot m_2^\Theta(Y)}{m_1^\Theta(A) + m_2^\Theta(Y)} + \frac{m_2^\Theta(A)^2 \cdot m_1^\Theta(Y)}{m_2^\Theta(A) + m_1^\Theta(Y)} \right] \end{aligned} \quad (15.12)$$

Combination by PCR6 fusion rule of the N importance discounted BBAs $m_j^\Theta(\cdot)$ defined by Eq. (15.11) provides a BBA denoted $m_{\text{PCR6}_\emptyset}^\Theta(\cdot)$ with $m_{\text{PCR6}_\emptyset}^\Theta(\emptyset) > 0$. Then, we commit zero to the mass of the empty set, and we normalize this BBA to get a proper normalized BBA $m_{\text{PCR6}}^\Theta(\cdot)$ with $m_{\text{PCR6}}^\Theta(\emptyset) = 0$; see [33] for details.

15.2.2.3 Decision-Making Given a Combined Belief Mass

Given a BBA $m^\Theta(\cdot)$, choosing a singleton $\hat{\theta} \in \Theta$ or a subset $\hat{A} \subseteq \Theta$ is the decision issue. In general, it consists in choosing $\hat{\theta} = \theta_{e^*}, e = 1, \dots, n$ with $e^* \triangleq \arg \max_e C(\theta_e)$, where $C(\theta_e)$ is a decision-making criterion chosen according

²For a technical reason, one allows to commit some mass on the empty set in this discounting. This is not a problem because the final fusion result will be normalized.

to the DM attitude: belief for a pessimistic attitude, plausibility for an optimistic one, one of the probabilistic transformations for an attitude of compromise.

In general, the DM attitude is not well known in DST. Moreover, in some practical cases, taking into account non-singletons $A \subseteq \Theta$ is needed to decide. For these cases, the minimum of any strict distance metric $d(m^\Theta, m_A^\Theta)$ between $m^\Theta(\cdot)$ and the categorical BBA $m_A^\Theta(\cdot)$ can be used in Eq. (15.13) [39]. If only singletons of 2^Θ are accepted, decision is defined by Eq. (15.14).

$$\hat{A} = \arg \min_{A \in 2^\Theta} d_{BI}(m^\Theta, m_A^\Theta) \quad (15.13)$$

$$\hat{\theta} = \arg \min_{\theta_e \in \Theta} d_{BI}(m^\Theta, m_{\theta_e}^\Theta) \quad (15.14)$$

Among the few true distance metrics³ between two BBAs $m_1^\Theta(\cdot)$ and $m_2^\Theta(\cdot)$, the belief interval-based Euclidean $d_{BI}(m_1^\Theta, m_2^\Theta) \in [0, 1]$ is based on Wasserstein's distance [40] and provides reasonable results [41].

The quality indicator $q(\hat{A})$ defined by Eq. (15.15) evaluates how good the decision \hat{A} is with respect to other focal elements: the higher $q(\hat{A})$ is, the more confident DM should be in its decision \hat{A} . If only singletons of 2^Θ are accepted, $q(\hat{\theta})$ is defined by Eq. (15.16).

$$q(\hat{A}) \triangleq 1 - \frac{d_{BI}(m^\Theta, m_A^\Theta)}{\sum_{A \in 2^\Theta \setminus \{\emptyset\}} d_{BI}(m^\Theta, m_A^\Theta)} \quad (15.15)$$

$$q(\hat{\theta}) \triangleq 1 - \frac{d_{BI}(m^\Theta, m_{\theta_e}^\Theta)}{\sum_{e=1}^n d_{BI}(m^\Theta, m_{\theta_e}^\Theta)} \quad (15.16)$$

15.3 Belief Function-Based Decision-Aiding Methods

Classical Decision-Aiding Methods (DAMs) have some limitations: (i) classical MCDAs do not consider imperfect evaluations of criteria, (ii) ranking can be affected by rank reversal problems [42, 43], and (iii) probability framework is limited by an epistemic uncertainty affecting the knowledge on the states of the nature \mathcal{S} [20, 21]. This section introduces new belief function-based DAMs which help to overcome these three limitations using (i) Evidential Reasoning for Multi-Criteria Decision Analysis (ER-MCDA) [44], (ii) Belief Function-based Technique for Order Preference by Similarity to Ideal Solution (BF-TOPSIS) methods [39], and (iii) Fuzzy Cautious Ordered Weighted Averaging with Evidential Reasoning (FCOWA-ER) [12], respectively.

³For any BBAs x, y, z defined on 2^Θ , a true distance metric $d(x, y)$ satisfies the properties of non-negativity ($d(x, y) \geq 0$), non-degeneracy ($d(x, y) = 0 \Leftrightarrow x = y$), symmetry ($d(x, y) = d(y, x)$), and triangle inequality ($d(x, y) + d(y, z) \geq d(x, z)$).

15.3.1 ER-MCDA: Multi-criteria Assignment Given Imperfect Scores

As detailed in Fig. 15.5, ER-MCDA methodology [44] is an extension of the Analytic Hierarchy Process (AHP) [3]. Given the FoD for decision Θ (step 1), it associates fuzzy logic framework and belief function theory to represent imperfect evaluation of the score of each alternative $A_i, i = 1, \dots, M$ based on each criterion $g_j, j = 1, \dots, N$, potentially provided by several sources $s_q, q = 1, \dots, Q$, through a BBA $m_{q,ij}^\Theta(\cdot)$ (step 2). A second improvement is the combination of BBAs taking into account reliability α_q of each source and importance ω_j of each criterion g_j (step 3). It finally helps to assign each A_i to the element $\hat{\theta}(A_i)$ (step 4).

15.3.1.1 ER-MCDA-Step 1: DMP Formalization

The scoring scale X_j is specified for each criterion $g_j, j = 1, \dots, N$. The FoD for decision Θ is also defined: for instance, we consider four qualitative labels (or classes) θ_e of efficiency with $\Theta = \{\theta_1 = \text{high}, \theta_2 = \text{medium}, \theta_3 = \text{low}, \theta_4 = \text{none}\}$.

15.3.1.2 ER-MCDA-Step 2: BBA $m_{q,ij}^\Theta(\cdot)$ Construction

For each $g_j, j = 1, \dots, N$, the mapping model is first provided (Fig. 15.3): n fuzzy sets μ_{j,θ_e} represent a partial pre-order of the DM preference for X_j [44]. Then, for each alternative $A_i, i = 1, \dots, M$, each source (e.g., an expert) s_q provides its imprecise and uncertain evaluation of $x_{ij} \in X_j$ through a possibility distribution $\pi_{q,ij}$ (Fig. 15.4).

Given these elements, the mapping process consists in transforming each $\pi_{q,ij}$ into a BBA $m_{q,ij}^\Theta(\cdot)$ using the g_j mapping model. The initial mapping process [19] was based on a geometric transformation and was restricted to provide only Bayesian BBAs. A new mapping model was recently developed to provide general BBAs [45, 46], for each A_i and each g_j :

1. since fuzzy sets are given for an open world, Yager's hedged element θ^c [32] is used to provide membership functions in an hedged world (*c.w.*), and all membership functions $\mu_{j,X}, X \neq \emptyset \in \Theta$ are built applying Eq. (15.5);
2. all membership functions $\mu_{j,\bar{X}}, \bar{X} \neq \emptyset \in \Theta$ are built applying Eq. (15.4);
3. given the possibility distribution $\pi_{q,ij}$, Eq. (15.7) gives the possibility measures $\Pi_{q,ij}(\bar{X})$ corresponding to the plausibility measure $\text{Pl}_{q,ij}^\Theta(\bar{X})$;
4. the belief function $\text{Bel}_{q,ij}^\Theta(\cdot)$ is directly obtained such as the BBA $m_{q,ij}^\Theta(\cdot)$.

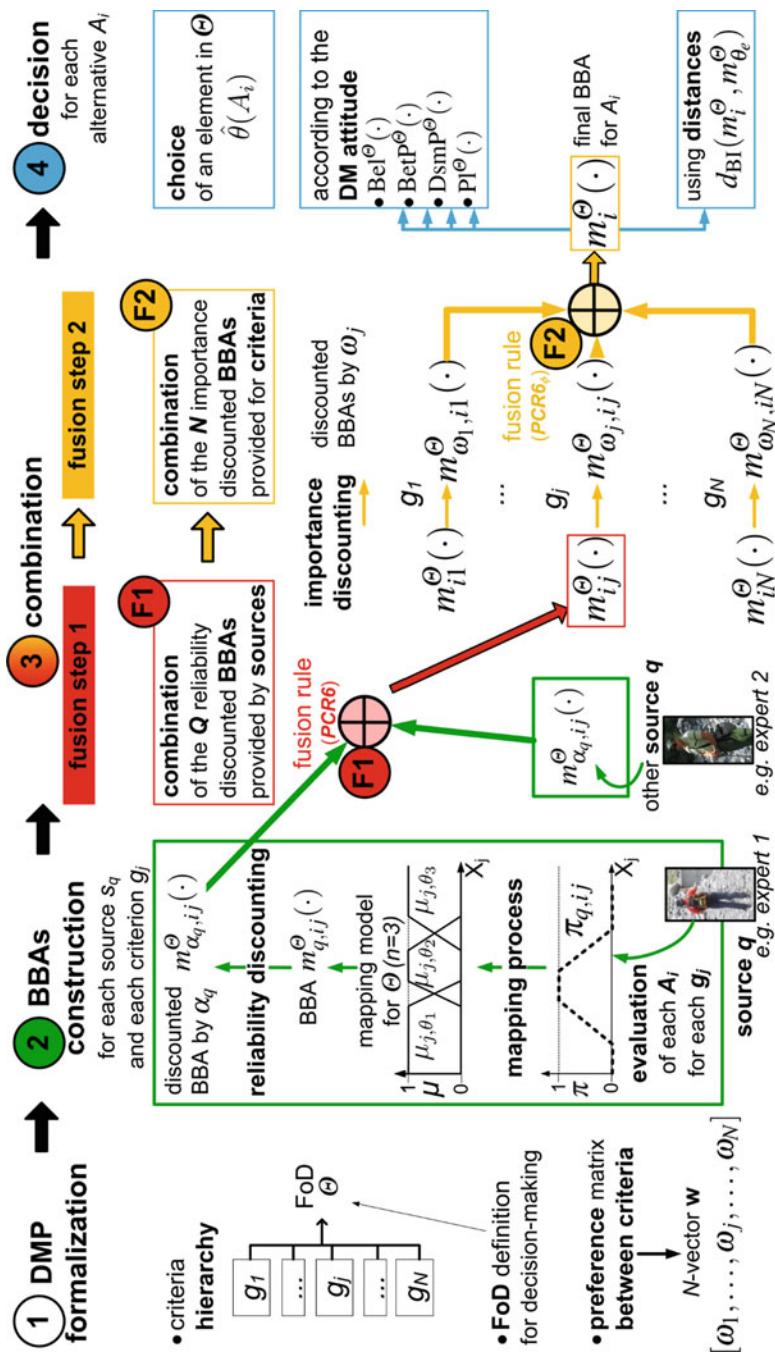


Fig. 15.5 The four steps of ER-MCDA

15.3.1.3 ER-MCDA-Step 3: Combination of Two BBAs

For each A_i and each g_j , fusion 1 combines BBAs provided by Q sources s_q , taking into account reliability factor α_q and using PCR6 fusion rule. For each A_i , fusion 2 combines BBAs provided for N criteria g_j , taking into account importance factor ω_j and using PCR6 $_\emptyset$ fusion rule. Final BBA $m_i^\Theta(\cdot)$ is obtained.

15.3.1.4 ER-MCDA-Step 4: Making Decision for Each A_i

Given Θ and $m_i^\Theta(\cdot)$, DM must assign each A_i to labels by choosing $\hat{\theta}(A_i)$. Therefore, DM can decide according to a pessimistic attitude (max of belief), an optimistic one (max of plausibility), or an attitude of compromise (max of subjective probability).

For the latter, whatever the probability transformation, the cardinality $|\theta^c|$ must be known. Therefore, a strong hypothesis is to consider $|\theta^c| = 1$ which can be theoretically discussed. Using Eq. (15.14) to decide through the minimal of belief interval distance does not involve any hypothesis on it.

15.3.2 BF-TOPSIS: A More Robust Multi-criteria Ranking

Given the score matrix \mathbf{S} defined by Eq. (15.1), the classical MCDA such as the AHP [3], Technique for Order Preference by Similarity to the Ideal Solution (TOPSIS) [4], or Estimator Ranking Vector (ERV) [47] are limited by rank reversal problems [42, 43]. As detailed in Fig. 15.6, the four new BF-TOPSIS methods [10] are inspired by the ERV to avoid a normalization step and by TOPSIS to compare each A_i with an ideal best and an ideal worst solutions. They are based on a common preliminary step and have an increasing computation complexity and robustness to rank reversal problems [6, 10].

15.3.2.1 Preliminary Step: DMP Formalization and BBA $m_{ij}^\mathcal{A}(\cdot)$ Construction

A DMP is about ranking all alternatives A_i and choosing the best one $A_{i^*} \in \mathcal{A}$: the FoD Θ for decision is the set of alternatives \mathcal{A} . Given the score matrix \mathbf{S} (see Fig. 15.6), this common step consists of constructing the $M \times N$ BBA matrix $\mathbf{M}^\mathcal{A} = [m_{ij}^\mathcal{A}(\cdot)]$.

For each A_i and g_j , the positive support $\text{Sup}_j(A_i) \triangleq \sum_{k \in \{1, \dots, M\} | x_{kj} \leq x_{ij}} |x_{ij} - x_{kj}|$ and the negative one $\text{Inf}_j(A_i) \triangleq -\sum_{k \in \{1, \dots, M\} | x_{kj} \geq x_{ij}} |x_{ij} - x_{kj}|$, respectively, measure how much A_i is better and worse than other alternatives according to g_j . Given $A_{\max}^j \triangleq \max_i \text{Sup}_j(A_i)$ and $A_{\min}^j \triangleq \min_i \text{Inf}_j(A_i)$, each $m_{ij}^\mathcal{A}(\cdot)$ is defined by:

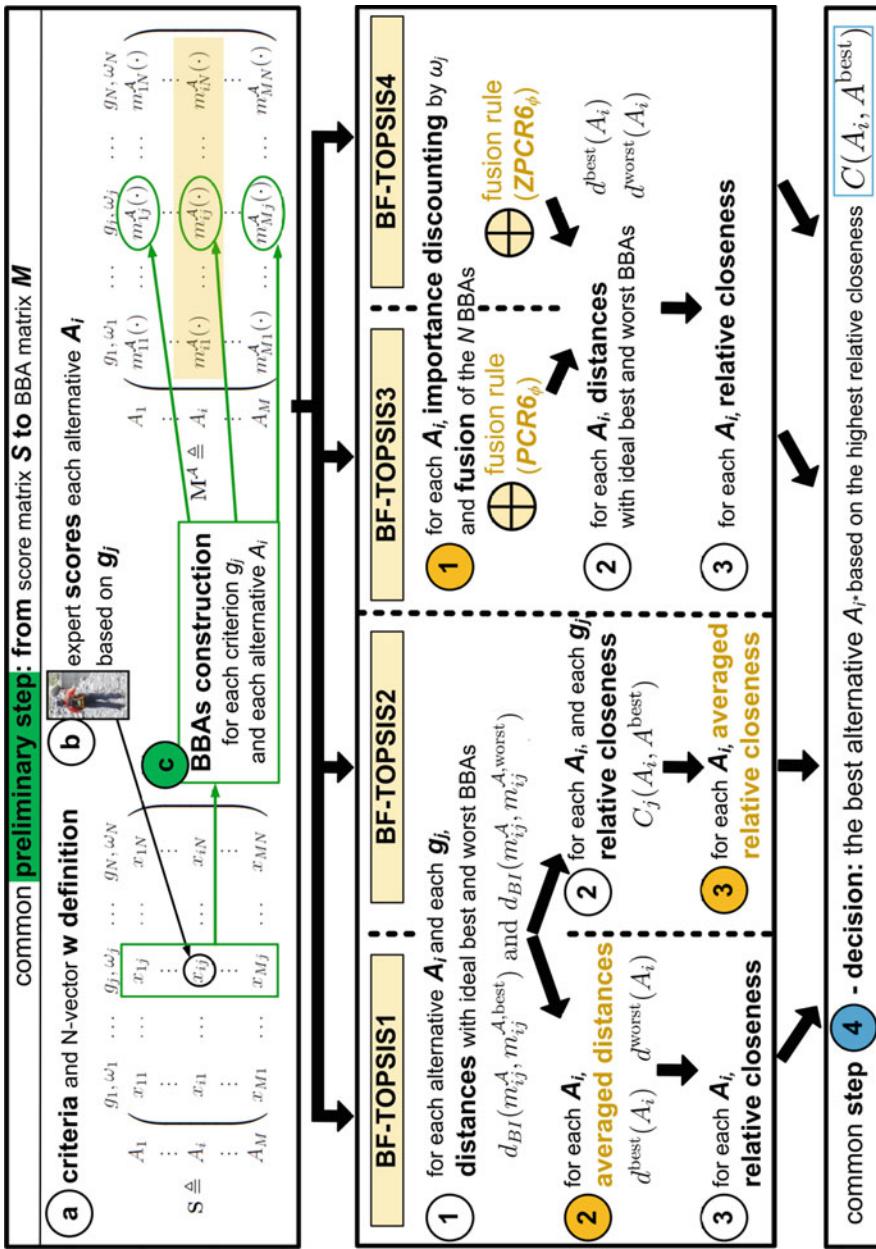


Fig. 15.6 Main steps of the four BF-TOPSIS methods

$$m_{ij}^{\mathcal{A}}(A_i) \triangleq \begin{cases} \frac{\text{Sup}_j(A_i)}{A_{\max}^j} & \text{if } A_{\max}^j \neq 0 \\ 0 & \text{if } A_{\max}^j = 0 \end{cases} \quad (15.17)$$

$$m_{ij}^{\mathcal{A}}(\bar{A}_i) \triangleq \begin{cases} \frac{\text{Inf}_j(A_i)}{A_{\min}^j} & \text{if } A_{\min}^j \neq 0 \\ 0 & \text{if } A_{\min}^j = 0 \end{cases} \quad (15.18)$$

$$m_{ij}^{\mathcal{A}}(A_i \cup \bar{A}_i) \triangleq m_{ij}^{\mathcal{A}}(\Theta) \triangleq 1 - (\text{Bel}_{ij}^{\mathcal{A}}(\bar{A}_i) + \text{Bel}_{ij}^{\mathcal{A}}(A_i)) \quad (15.19)$$

The four BF-TOPSIS methods differ from each other in the way they process the matrix $\mathbf{M}^{\mathcal{A}}$. All of them compute the relative closeness of each alternative A_i with an ideal best solution A^{best} denoted by $C(A_i, A^{\text{best}})$. The preference ordering of all alternatives is built given the following criterion: the higher it is, the better A_i is. An extension of BF-TOPSIS methods for dealing with imprecise scores is proposed in [48].

15.3.2.2 BF-TOPSIS1

- (1) For each A_i and g_j , the ideal best and worst BBAs are defined by $m_{ij}^{\text{best}}(A_i) \triangleq 1$ and $m_{ij}^{\text{worst}}(\bar{A}_i) \triangleq 1$ which are used to compute the distances $d_{BI}(m_{ij}^{\mathcal{A}}, m_{ij}^{\mathcal{A},\text{best}})$ and $d_{BI}(m_{ij}^{\mathcal{A}}, m_{ij}^{\mathcal{A},\text{worst}})$.
- (2) The respective averaged distances $d^{\text{best}}(A_i)$ and $d^{\text{worst}}(A_i)$ are computed by weighting previous distances by importance weights ω_j of criteria g_j .
- (3) For each A_i , the relative closeness is computed by:

$$C(A_i, A^{\text{best}}) \triangleq \frac{d^{\text{worst}}(A_i)}{d^{\text{worst}}(A_i) + d^{\text{best}}(A_i)} \quad (15.20)$$

15.3.2.3 BF-TOPSIS2

- (1) This step is the same as for BF-TOPSIS1. (2) For each A_i and g_j , the relative closeness $C_j(A_i, A^{\text{best}})$ is computed. (3) The averaged relative closeness $C(A_i, A^{\text{best}})$ is computed by weighting $C_j(A_i, A^{\text{best}})$ by importance weights ω_j of criteria g_j .

15.3.2.4 BF-TOPSIS3

- (1) For each A_i , the N BBAs $m_{ij}^{\mathcal{A}}(\cdot)$ are combined through PCR6 fusion rule to give $m_i^{\mathcal{A}}(\cdot)$ taking into account the importance factor ω_j of each criterion g_j [33]. (2) For

each A_i , the ideal best and worst BBAs allow to give $d^{\text{best}}(A_i) = d_{BI}(m_i^{\mathcal{A}}, m_i^{\mathcal{A},\text{best}})$ and $d^{\text{worst}}(A_i) = d_{BI}(m_i^{\mathcal{A}}, m_i^{\mathcal{A},\text{worst}})$. (3) This step is the same as for BF-TOPSIS1.

15.3.2.5 BF-TOPSIS4

This method differs from BF-TOPSIS3 only by the choice of the ZPCR6 fusion rule [49] instead of PCR6 rule of combination.

15.3.3 FCOWA-ER: Choice Under Epistemic Uncertainty

Given the payoff matrix \mathbf{C} defined by Eq. (15.2), the DMP requires choosing the best alternative $A_{i^*} \in \mathcal{A}$. COWA-ER has been proposed [11] for such decision-making given uncertain knowledge on \mathcal{S} . It mixes cautiously the principle of Yager's Ordered Weighted Averaging (OWA) approach based on belief function theory [13] with fusion rules, notably the PCR6 one [30]. As detailed in Fig. 15.7, FCOWA-ER [12] is a modified version of COWA-ER using fuzzy sets which improves performances of COWA-ER and reduces its computational burden.

15.3.3.1 From the OWA Approach...

Under ignorance, Yager uses the OWA operator as a weighted average of ordered values of a variable defined by Eq. (15.21). For each A_i , $i = 1, \dots, M$, it consists in choosing a normalized set of weighting factors $\mathbf{W}_i = [w_{i1}, \dots, w_{ik}, \dots, w_{iK}]$, where $w_{ik} \in [0, 1]$, $\sum_k w_{ik} = 1$, and \mathbf{W}_i depends on the DM attitude: $\mathbf{W}_i = [0, 0, \dots, 0, 1]$ represents the pessimistic attitude, while $\mathbf{W}_i = [1, 0, \dots, 0, 0]$ is used for the optimistic one. The OWA value V_i is computed for the collection of payoffs $C_{i1}, C_{i2}, \dots, C_{iK}$, with b_{ik} as the k^{th} largest element in it. The best A_{i^*} is chosen with $i^* \triangleq \arg \max_i \{V_i\}$.

Under epistemic uncertainty, considering the states of the nature \mathcal{S} as the FoD, Yager represents the DM belief structure by a BBA $m^{\mathcal{S}}(\cdot) : 2^{\mathcal{S}} \rightarrow [0, 1]$, which is characterized by the s focal elements $X_r \in 2^{\mathcal{S}}$. For each alternative A_i , restricting the states of the nature to $S_k \in X_r$, one has $\mathbf{M}_{ir} \triangleq \{C_{ik} | S_k \in X_r\}$, $r = 1, \dots, s$. For each A_i , each X_r , and some DM attitude chosen a priori, the OWA value $V_{ir} = \text{OWA}(\mathbf{M}_{ir})$ is computed. The derivation of a generalized expected value C_i of payoff is defined by Eq. (15.22). The best A_{i^*} is thus chosen with $i^* \triangleq \arg \max_i \{C_i\}$.

$$V_i \triangleq \text{OWA}(C_{i1}, C_{i2}, \dots, C_{iK}) = \sum_k w_{ik} \cdot b_{ik} \quad (15.21)$$

$$C_i = \sum_{r=1}^s m^{\mathcal{S}}(X_r) V_{ir} \quad (15.22)$$

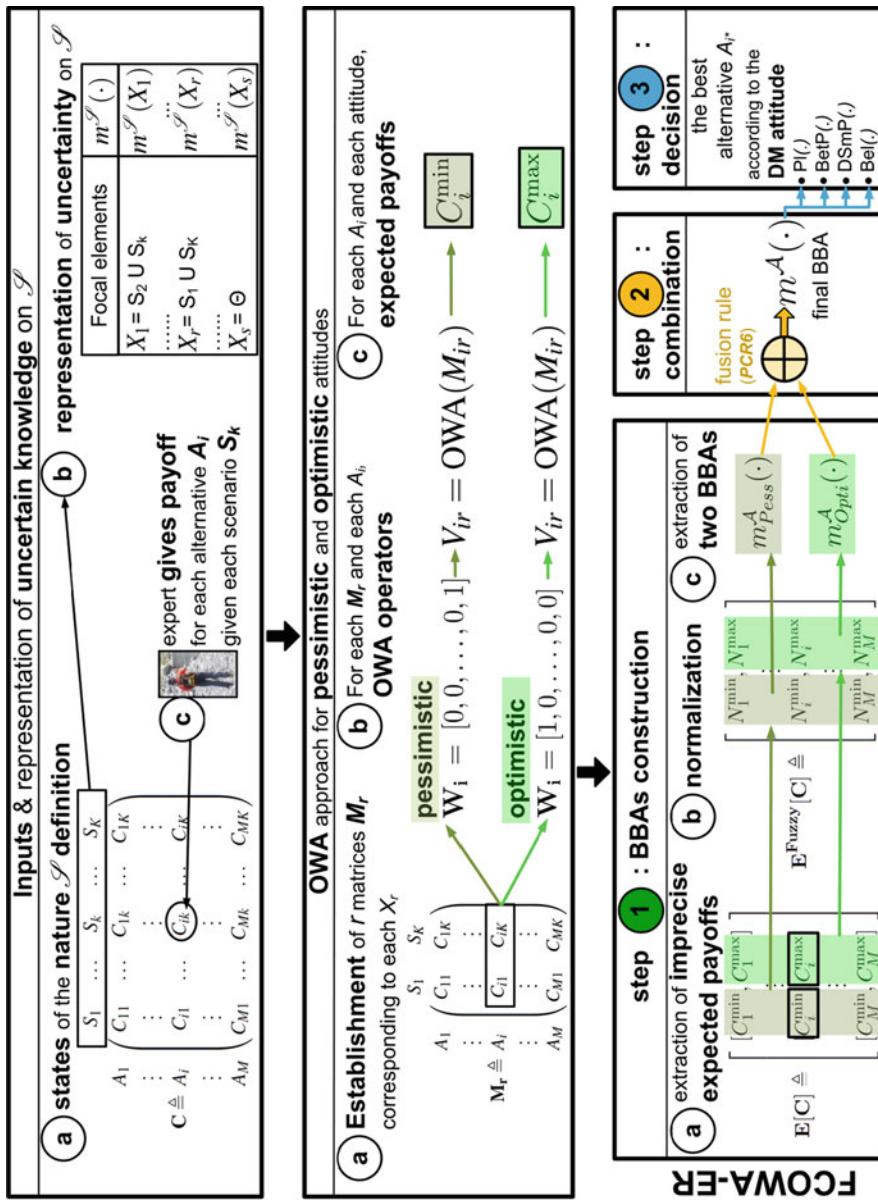


Fig. 15.7 Steps of FCOWA-ER

15.3.3.2 ...to the COWA-ER and FCOWA-ER Approaches

For each alternative A_i , COWA-ER method exploits only the results of the two extreme attitudes (pessimistic and optimistic OWA) jointly [11]. Decision-making under uncertainty is thus based on the M imprecise valuations (or intervals) of expected payoffs gathered in the M -vector $\mathbf{E}[\mathbf{C}]$ given by Eq. (15.23).

FCOWA-ER method [12] has been then developed to go beyond two COWA-ER limitations. (1) The BBAs obtained by using α -cuts are consonant support (nested in order) without any correlation between information sources. (2) The computational time for making the combination does not depend on the number M of alternatives.

15.3.3.3 FCOWA-ER-Step 1: Construction of BBAs

Each column in $\mathbf{E}[\mathbf{C}]$ is, respectively, normalized to obtain the column-wise normalized expected payoff $\mathbf{E}^{\text{Fuzzy}}[\mathbf{C}]$ given by Eq. (15.24).

$$\mathbf{E}[\mathbf{C}] \triangleq \begin{bmatrix} [C_1^{\min}, C_1^{\max}] \\ [C_2^{\min}, C_2^{\max}] \\ \vdots \\ [C_M^{\min}, C_M^{\max}] \end{bmatrix} \quad (15.23)$$

$$\mathbf{E}^{\text{Fuzzy}}[\mathbf{C}] \triangleq \begin{bmatrix} N_1^{\min}, N_1^{\max} \\ N_2^{\min}, N_2^{\max} \\ \vdots \\ N_M^{\min}, N_M^{\max} \end{bmatrix} \quad (15.24)$$

The vectors $\mu_1 = [N_1^{\min}, \dots, N_M^{\min}]$ and $\mu_2 = [N_1^{\max}, \dots, N_M^{\max}]$ can be seen as two fuzzy membership functions (FMFs) $\mu : A_i \in \mathcal{A} \rightarrow [0, 1]$. Given the FoD $\mathcal{A} = \{A_1, A_2, \dots, A_M\}$, they are, respectively, converted into two BBAs $m_{\text{Pess}}^{\mathcal{A}}(\cdot)$ and $m_{\text{Opti}}^{\mathcal{A}}(\cdot)$ using the α -cut approach [50], considering M as the number of α -cuts.

15.3.3.4 FCOWA-ER-Steps 2 and 3: Combination of the Two BBAs and Decision

The two BBAs $m_{\text{Pess}}^{\mathcal{A}}(\cdot)$ and $m_{\text{Opti}}^{\mathcal{A}}(\cdot)$ are combined with the PCR6 fusion rule. The decision is about choosing A_{i^*} according to the DM attitude or the minimal distance.

15.4 Application to Efficiency of Torrential Protective Actions

The example previously introduced in Fig. 15.2 is used to show how new belief function-based MCDAs can help DMs to decide on real DMPs. The DMP is first formalized. ER-MCDA, BF-TOPSIS, and FCOWA-ER are then successively applied without detailing computation steps: we only provide inputs and main results.

15.4.1 Formalization of the Decision-Making Problem

The problem is about comparing under uncertainty several torrential protective actions based on their efficiency by involving several criteria.

15.4.1.1 Multi-criteria Decision-Making Problem

Alternatives and decision criteria, with their scoring scales and importance weights, must be specified to provide the structure of the score matrix \mathbf{S} defined by Eq. (15.1).

The set \mathcal{A} gathers $M = 5$ protective alternatives A_i (Fig. 15.2):

- A_1 : doing nothing;
- A_2 : building check dam series in headwaters;
- A_3 : building a sediment trap on the alluvial fan apex;
- $A_4 = A_2 \cup A_3$: building both check dam series and a sediment trap;
- A_5 : building individual protections for each element at-risk.

On the one hand, these actions aim at reducing potential damage on elements-at-risk. Several types of damage can occur such as housing destruction or environmental damage due to destruction of dangerous sites, etc. Their assessment in monetary value can strongly be debated as, for example, for human casualties [6]. On the other hand, each alternative involves high investment and maintenance cost.

A DM thus considers $N = 5$ criteria g_j with specific scoring scale X_j [6] to compare alternatives according to their efficiency. He wants to minimize g_1 and g_2 (decreasing preference) and to maximize g_3 , g_4 , and g_5 (increasing preference), with:

- g_1 : investment cost in ϵ ($x_{i1} \in X_1 = \mathbb{R}^+$);
- g_2 : annual maintenance cost in ϵ ($x_{i2} \in X_2 = \mathbb{R}^+$);
- g_3 : annual risk reduction of damaged houses surface in m^2 ($x_{i3} \in X_3 = \mathbb{R}^+$);
- g_4 : annual risk reduction in *human casualties* ($x_{i4} \in X_4 = \mathbb{R}^+$);
- g_5 : annual risk reduction in *number* of dangerous sites ($x_{i5} \in X_5 = \mathbb{R}^+$).

In practice, for each A_i and g_j , annual risk reduction $\Delta R_j(A_i)$ is computed as reduction of potential damage expected value: $\Delta R_j(A_i) = R_j(0) - R_j(A_i)$, where $R_j(0)$ is the baseline risk (without A_i) and $R_j(A_i)$ is the residual risk with A_i [6].

Given this set of criteria, DM uses the AHP process [3] to define the 5-vector of their importance weights: $\mathbf{w} = [0.08, 0.04, 0.10, 0.46, 0.32]$.

15.4.1.2 Decision-Making Under Uncertainty

Damage assessment depends on torrential hazards: without flood, there is no damage; during a big but rare flood, damage are higher; damage are much higher with a big debris flows. In practice, several scenarios of torrential hazards are thus taken into account to assess annual risk reduction criteria. They must be specified to provide the structure of the payoff matrix \mathbf{C} defined by Eq. (15.2).

A DM considers a set of states of the nature \mathcal{S} with $K = 7$ scenarios as follows:

- liquid floods without bedload transport : S_1 with $Q_l < Q_{l1}$ ⁴; S_2 with $Q_l \geq Q_{l1}$;
- floods with bedload transport : S_3 with $V_s < V_{s1}$ ⁵; S_4 with $V_s \geq V_{s1}$;
- debris flow : S_5 with $V_l < V_{l1}$ ⁶; S_6 with $V_{l1} \leq V_l < V_{l2}$; S_7 with $V_l \geq V_{l2}$

15.4.2 ER-MCDA to Assign an Efficiency Label to Each Alternative

In practice, a first DMP is about assigning each alternative to a qualitative efficiency label. Therefore, ER-MCDA methodology can be used.

15.4.2.1 ER-MCDA Inputs

FoD for decision gathers $n = 4$ exhaustive and mutually exclusive efficiency labels, with $\Theta = \{\theta_1 = \text{no}, \theta_2 = \text{low}, \theta_3 = \text{medium}, \theta_4 = \text{high}\}$. The mapping model of each criterion is provided in Fig. 15.8.

Two sources (experts) s_q , $q = 1, 2$, are assumed totally reliable ($\alpha_1 = \alpha_2 = 1$). For each A_i and g_j , each one provides the possibility distributions $\pi_{q,ij}$ in Table 15.1.

⁴ Q_l = liquid flow.

⁵ V_s = solid volume.

⁶ V_l = debris flow volume.

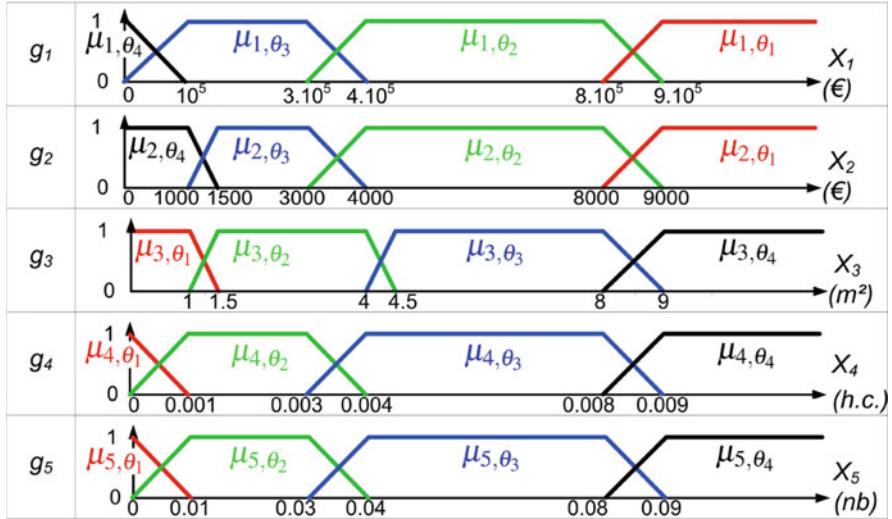


Fig. 15.8 ER-MCDA-input: $N = 5$ mapping models

Table 15.1 ER-MCDA-input: imprecise evaluations $\pi_{q,ij}$ ($q = 1, 2$; $i = 1, \dots, 5$; $j = 1, \dots, 5$)

A_i	s_q	g_1^a	g_2^a	g_3^b	g_4^b	g_5^b
A_1	s_1	0,0,0,0	0,0,0,0	0,0	0,0	0,0
A_1	s_2	0,0,0,0	0,0,0,0	0,0	0,0	0,0
A_2	s_1	$(2, 3, 3, 5).10^5$	$(3, 5, 7, 8).10^3$	1.1, 4.3	$(2.1, 8.9).10^{-3}$	$(3.4, 6.4).10^{-2}$
A_2	s_2	$(1, 2, 2, 3).10^5$	$(2, 3, 4, 5).10^3$	2, 5.2	$(1.2, 10.3).10^{-3}$	$(4.3, 8.7).10^{-2}$
A_3	s_1	$(2, 4, 4, 6).10^5$	$(1, 1.5, 2, 3).10^3$	1.2, 4.8	$(2.3, 11.8).10^{-3}$	$(5.1, 8.5).10^{-2}$
A_3	s_2	$(4, 5, 6, 8).10^5$	$(0.5, 1, 1, 2).10^3$	0.1, 5.5	$(3.1, 13.1).10^{-3}$	$(3.7, 8.5).10^{-2}$
A_4	s_1	$(4, 7, 7, 11).10^5$	$(4, 7.5, 9, 11).10^3$	3.3, 8.45	$(3.5, 16.8).10^{-3}$	$(6.2, 9.4).10^{-2}$
A_4	s_2	$(5, 7, 8, 11).10^5$	$(2.5, 4, 5, 7).10^3$	3.1, 8.9	$(3.4, 16.1).10^{-3}$	$(4.3, 9.2).10^{-2}$
A_5	s_1	$(9, 10, 12, 14).10^5$	0, 0, 0, 0	4.2, 9.2	$(3.7, 8.4).10^{-3}$	$(3.1, 8.4).10^{-2}$
A_5	s_2	$(8, 9, 10, 11).10^5$	0, 0, 0, 0	4.65, 8.25	$(2.1, 9.3).10^{-3}$	$(1.5, 9.2).10^{-2}$

^aCriteria assessed by {a,b,c,d} as shown in Fig. 15.3

^bCriteria assessed by {a,d} ($a=b$ and $c=d$)

15.4.2.2 ER-MCDA Results

For each criterion g_j , the new mapping process is used in ER-MCDA-Step 2 to map each $\pi_{q,ij}$ of Table 15.1 into the g_j mapping model of Fig. 15.8. It provides BBAs $m_{1,ij}^\Theta(\cdot)$ for s_1 and $m_{2,ij}^\Theta(\cdot)$ for s_2 . Applying ER-MCDA-Step 3 on those BBAs, $M = 5$ BBAs $m_i^\Theta(\cdot)$ are finally obtained in Table 15.2.

To avoid assumption $|\theta_c| = 1$, ER-MCDA-Step 4 is based on computing distances $d_{BI}(m_i^\Theta, m_X^\Theta)$, from each column of Table 15.2. For each A_i , the chosen focal element $\hat{X}(A_i) \in 2^\Theta$, the corresponding value of minimal distance d_{BI}^{\min} , and the decision quality $q(\hat{X}(A_i))$ are given in Table 15.3.

Table 15.2 ER-MCDA-Step 3 results: BBAs $m_i^\Theta(\cdot)$ obtained for each alternative A_i

Focal elements X	$m_1^\Theta(\cdot)$	$m_2^\Theta(\cdot)$	$m_3^\Theta(\cdot)$	$m_4^\Theta(\cdot)$	$m_5^\Theta(\cdot)$
θ^c	0.00049	0.00075	0.00067	0.00079	0.04062
$\theta_1 \cup \theta^c$	0.96997	0	0	0	0.03157
$\theta_2 \cup \theta^c$	0	0.26680	0.02393	0.02293	0
$\theta_1 \cup \theta_2 \cup \theta^c$	0	0	0	0.00830	0
$\theta_3 \cup \theta^c$	0	0.36178	0.97036	0.71702	0.86412
$\theta_2 \cup \theta_3 \cup \theta^c$	0	0.36113	0.00416	0.02852	0
$\theta_4 \cup \theta^c$	0.02954	0	0.00018	0.17687	0.03597
$\theta_3 \cup \theta_4 \cup \theta^c$	0	0.00954	0.00070	0.04557	0.02772

Table 15.3 ER-MCDA-Step 4 results: decision based on $d_{BI}(m_i^\Theta, m_X^\Theta)$

Alternative A_i	d_{BI}^{\min}	$\hat{X}(A_i)$	$q(\hat{X}(A_i))$
A_1	0.0172	$\theta_1 \cup \theta^c$	0.9990
A_2	0.1855	$\theta_2 \cup \theta_3 \cup \theta^c$	0.9878
A_3	0.0154	$\theta_3 \cup \theta^c$	0.9991
A_4	0.1319	$\theta_3 \cup \theta^c$	0.9919
A_5	0.0567	$\theta_3 \cup \theta^c$	0.9968

Solution A_1 of doing nothing is mainly no efficient. Alternative A_2 is lowly or mediumly efficient, while the three other ones are mediumly efficient. It corresponds to a partial preference order: $A_3 \sim A_4 \sim A_5 \succeq A_2 > A_1$.

As shown in Table 15.3, quality indicators of decisions are similar for A_3 and A_5 and better than for A_4 . For A_2 , global mass is more distributed with $m_2^\Theta(\theta_3 \cup \theta^c) = 0.36$, $m_2^\Theta(\theta_2 \cup \theta_3 \cup \theta^c) = 0.36$, and $m_2^\Theta(\theta_2 \cup \theta^c) = 0.27$: decision quality is less good.

15.4.3 BF-TOPSIS to Rank Alternatives

Another DMP is about ranking all potential solutions given previous criteria. BF-TOPSIS methods are compared to help to solve it.

15.4.3.1 BF-TOPSIS Inputs

A precise score of each A_i based on each g_j must be provided to obtain the score matrix \mathbf{S} , consistent with possibility distributions $\pi_{q,ij}$ given in Table 15.1. To take into account decreasing preference, initial scores of g_1 and g_2 are multiplied by -1 providing the score matrix \mathbf{S}^{pref} defined by Eq. (15.25).

Table 15.4 BF-TOPSIS results: relative closeness $C(A_i, A^{\text{best}})$

Alternative A_i	BF-TOPSIS1	BF-TOPSIS2	BF-TOPSIS3	BF-TOPSIS4
A_1	0.12	0.12	0.03	0.03
A_2	0.49	0.49	0.68	0.68
A_3	0.66	0.66	0.85	0.85
A_4	0.69	0.69	0.88	0.88
A_5	0.92	0.92	0.97	0.97

$$\mathbf{S}^{\text{pref}} \triangleq \begin{pmatrix} & g_1 & g_2 & g_3 & g_4 & g_5 \\ A_1 & 0 & 0 & 0 & 0 & 0 \\ A_2 & -300000 & -6000 & 5 & 0.007 & 0.02 \\ A_3 & -300000 & -1500 & 5 & 0.008 & 0.04 \\ A_4 & -600000 & -7500 & 7 & 0.008 & 0.05 \\ A_5 & -1000000 & 0 & 7 & 0.008 & 0.1 \end{pmatrix} \quad (15.25)$$

15.4.3.2 BF-TOPSIS Results

Given \mathbf{S}^{pref} , applying the four BF-TOPSIS methods provides relative closeness $C(A_i, A^{\text{best}})$ of each alternative A_i with the ideal best solution A^{best} in Table 15.4.

In this case, whatever the BF-TOPSIS method used, preference ranking of all alternatives according to descending order of $C(A_i, A^{\text{best}})$ is $A_5 \succ A_4 \succ A_3 \succ A_2 \succ A_1$. In [6], for this application case, authors not only give computation details but also a comparison with classical MCDA methods such as CBA and AHP. They show that CBA is very sensitive to criteria choice for monetary valuation such as the human life (g_4) and that BF-TOPSIS is more robust to rank reversal problems than AHP.

15.4.4 FCOWA-ER to Choose the Best Alternative Under Uncertainty

The final practical DMP is about choosing the best solution to implement, considering the knowledge on the states of the nature.

15.4.4.1 FCOWA-ER Inputs

For each $A_i, i = 1, \dots, 5$ and each scenario $S_k, k = 1, \dots, 7$, C_{ik} is the efficiency level in the 5×7 matrix \mathbf{C} defined by Eq. (15.26). It can be extracted after implementing ER-MCDA to solve a previous MCDM problem, given a specific S_k . A quantitative transformation of labels $\theta_e, e = 1, \dots, n$ into $[1; 10]$ (the higher

Table 15.5 FCOWA-ER result: credibility, BetP, DSmp $_{\epsilon=10^{-6}}$, and plausibility of A_i efficiency

A_i	$\text{Bel}^{\mathcal{A}}(A_i)$	$\text{BetP}^{\mathcal{A}}(A_i)$	$\text{DSmp}^{\mathcal{A}}(A_i)_{\epsilon=10^{-6}}$	$\text{Pl}^{\mathcal{A}}(A_i)$
A_1	0.000000	0.027908	0.000004	0.139530
A_2	0.000000	0.060282	0.000008	0.269030
A_3	0.000000	0.132010	0.000013	0.470130
A_4	0.000000	0.180390	0.000015	0.566890
A_5	0.404960	0.599420	0.999960	1.000000

is score, the higher is payoff) is proposed. Results of ER-MCDA in Tables 15.2 and 15.3 help give payoffs for S_4 .

$$\mathbf{C} = \begin{bmatrix} 9 & 3 & 2 & 2 & 1 & 1 & 1 \\ 10 & 8 & 4 & 3 & 2 & 1 & 1 \\ 10 & 7 & 6 & 6 & 4 & 4 & 1 \\ 10 & 6 & 8 & 7 & 3 & 2 & 1 \\ 10 & 8 & 6 & 6 & 6 & 5 & 1 \end{bmatrix} \quad (15.26)$$

Expert represents uncertainty on the states of the nature \mathcal{S} through a BBA $m^{\mathcal{S}}(\cdot)$. $s = 4$ focal elements $X_r \in 2^{\mathcal{S}}$ are considered:

$$m^{\mathcal{S}}(X_1) = m^{\mathcal{S}}(S_1 \cup S_3 \cup S_5) = 0.4, m^{\mathcal{S}}(X_2) = m^{\mathcal{S}}(S_2 \cup S_4 \cup S_5 \cup S_6) = 0.25, \\ m^{\mathcal{S}}(X_3) = m^{\mathcal{S}}(S_7) = 0.1, m^{\mathcal{S}}(X_4) = m^{\mathcal{S}}(\mathcal{S}) = 0.25.$$

X_1 , X_2 , and X_3 are partial ignorances, and X_4 is the full ignorance.

15.4.4.2 FCOWA-ER Results

Given payoff matrix \mathbf{C} and BBA $m^{\mathcal{S}}(\cdot)$, applying FCOWA-ER steps, provides the two BBAs $m_{\text{Pess}}^{\mathcal{A}}(\cdot)$ and $m_{\text{Opti}}^{\mathcal{A}}(\cdot)$ for the FoD \mathcal{A} (step 1) which are combined to give the final BBA $m_{\text{PCR6}}^{\mathcal{A}}(\cdot)$ through PCR6 fusion rule (step 2). Table 15.5 shows values of $\text{Bel}^{\mathcal{A}}(\cdot)$, $\text{BetP}^{\mathcal{A}}(\cdot)$, $\text{DSmp}_{\epsilon=10^{-6}}^{\mathcal{A}}(\cdot)$ and $\text{Pl}^{\mathcal{A}}(\cdot)$ based on $m_{\text{PCR6}}^{\mathcal{A}}(\cdot)$ (step 3).

Whatever the decision rule, the best action A_{i*} is always A_5 . The total preference ranking is deducted: $A_5 \succ A_4 \succ A_3 \succ A_2 \succ A_1$.

15.5 Conclusions and Perspectives

In practice, torrential risk managers must decide on the best action to reduce damage on elements-at-risk. Therefore, the comparison of efficiency of potential alternatives is generally used. Each one can be assessed through qualitative labels that require a partial ranking of alternatives, but it generally cannot help to choose the best action that requires a total preference ranking.

The decisions are based on several criteria, such as costs and different types of damage reduction, imperfectly assessed under an epistemically uncertain environment corresponding to torrential hazards. Confronted with such difficulties, decisions are generally based on expert knowledge which directly takes into account imperfect information.

This chapter shows how recent developments of MCDAs based on belief function theory actually can help decision-makers in their decision process. A practical example is proposed and new methods, showing their possible combination for a global decision-making was applied.

Whatever the method is used, the first step of any DMP is to define alternatives to compare, criteria to take into account, their importance weights which represent the DM preference between them, and the set of states of the nature. Given the scores of each alternative based on each criterion, the final step is about aggregating this multi-criteria and multi-scenario evaluation to help decision.

The three methods applied in this chapter (ER-MCDA, BF-TOPSIS, FCOWA-ER) are combined at a combination step, while the PCR6 fusion rule is preferred. Nevertheless, analyzing the effect of this choice on the results of each method should be done.

To solve MCDM problems, ER-MCDA helps to take into account imperfect evaluation of criteria potentially provided by several sources. FoD of decision is first specified through qualitative or quantitative labels. Each expert is considered as source who gives a possibility distribution (imprecise scoring) and a mapping model based on fuzzy sets. ER-MCDA makes it possible to choose a label for each alternative providing the quality of this decision.

The four BF-TOPSIS methods help a total preference ranking of all alternatives with a better robustness to rank reversal problems than classical MCDAs. It is based on a precise score matrix representing the MCDM problem. Using intermediary results of ER-MCDA as an intermediary decision step helps to take into account imprecise scoring in BF-TOPSIS.

FCOWA-ER is different from the two previous methods because it proposes a method to solve DMP under uncertainty. It improves the OWA method used when the knowledge of the states of the nature is uncertain. It was first developed to decide given a precise scoring of each payoff. As for BF-TOPSIS, it is possible to apply ER-MCDA for each scenario and to propose a quantitative transformation if qualitative labels are used. It thus helps to take into account initial imperfect scoring.

From an operational point of view, this chapter shows how theoretical methods can help a better formalization of the decision-making process. Indeed, expert and DM elicitation is always needed to express, given the DMP under concern, the criteria to take into account, their importance, and the preferences for their evaluation. Moreover, methods have been applied to DMPs related to protection efficiency, but they are generic and can be applied for any other DMP.

Acknowledgements This study was partially funded by the French Agricultural and Forest Ministry (MAA) and the French Environment Ministry (MTES).

References

1. S. Carladous, G. Piton, A. Recking, F. Liébault, D. Richard, J.-M. Tacnet, D. Kuss, F. Philippe, Y. Queffélan, O. Marco, Towards a better understanding of the today French torrents management policy through a historical perspective, in *3rd International Conference FLOODrisk*, Lyon (2016)
2. B. Roy, *Méthodologie Multicritère d'Aide à la Décision* (Economica Collection Gestion, Paris, 1985)
3. T.L. Saaty, *Multicriteria Decision Making – The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation* (McGraw-Hill, Pittsburgh, 1980)
4. Y.J. Lai, T.Y. Liu, C.L. Hwang, TOPSIS for MODM. *Eur. J. Oper. Res.* **76**, 486–500 (1994)
5. N. Hanley, C.L. Spash, *Cost-Benefit Analysis and the Environment* (Edward Elgar, Cheltenham, 1993)
6. S. Carladous, J.-M. Tacnet, J. Dezert, D. Han, M. Batton-Hubert, Evaluation of efficiency of torrential protective structures with new BF-TOPSIS methods, in *19th International Conference on Information Fusion*, Heidelberg (2016)
7. J.-M. Tacnet, M. Batton-Hubert, J. Dezert, A two-step fusion process for multi-criteria decision applied to natural hazards in mountains, in *International Workshop on the Theory of Belief Functions*, Brest (2010)
8. L.A. Zadeh, Fuzzy sets. *Inf. Control* **8**, 338–353 (1965)
9. L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Syst.* **1**, 3–28 (1978)
10. J. Dezert, D. Han, H. Yin, A new belief function based approach for multi-criteria decision-making support, in *19th International Conference on Information Fusion*, Heidelberg (2016)
11. J.-M. Tacnet, J. Dezert, Cautious OWA and evidential reasoning for decision making under uncertainty, in *14th International Conference on Information Fusion*, Chicago (2011), pp. 2074–2081
12. D. Han, J. Dezert, J.-M. Tacnet, C. Han, A fuzzy-cautious OWA approach with evidential reasoning, in *15th International Conference on Information Fusion*, Singapore (2012), pp. 278–285
13. R. Yager, Decision making under Dempster-Shafer uncertainties. *Stud. Fuzziness Soft Comput.* **219**, 619–632 (2008)
14. F.H. Knight, *Risk, Uncertainty, and Profit* (Houghton Mifflin Company, 1921)
15. A. Schärfig, *Décider sur plusieurs critères. Panorama de l'aide à la décision multicritère* (Presses Polytechniques et Universitaires Romandes, Lausanne, 1985)
16. J. von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, 1944)
17. B. Mazzorana, J. Hübl, S. Fuchs, Improving risk assessment by defining consistent and reliable system scenarios. *Nat. Hazards Earth Syst. Sci.* **9**, 145–159 (2009)
18. L.J. Savage, *Foundations of Statistics* (Wiley, New York, 1954)
19. J.-M. Tacnet, Prise en compte de l'incertitude dans l'expertise des risques naturels en montagne par analyse multicritères et fusion d'information. Ph.D. Thesis, ENSMSE: Saint-Étienne (2009)
20. K.J. Arrow, L. Hurwicz, An optimality criterion for decision-making under ignorance, in *Uncertainty and Expectations in Economics*, ed. by C.F. Carter, J.L. Ford, G.L.S. Shackle (Basil Blackwell, Oxford, 1972), pp. 1–11
21. M. Allais, The so-called Allais paradox and rational decisions under uncertainty, in *Expected Utility Hypotheses and the Allais Paradox. Contemporary Discussions of the Decisions Under Uncertainty with Allais' Rejoinder*, ed. by M. Allais, O. Hagen (Springer, New-York, 1979)
22. G. Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, 1976)
23. S. Carladous, M.-P. Michaud, J.-M. Tacnet, Q. Delvienne, A survey on protection work databases used at the Alpine space level: analysis of contents and state of the art related to protection work effectiveness assessment in START-it-uP (WP4). European START-it-uP project, Grenoble (2014)

24. A. Martin, *Cours sur La fusion d'informations* (ENSIETA, Rennes, 2005)
25. D. Dubois, H. Prade, Fuzzy sets, probability and measurement. *Eur. J. Oper. Res.* **40**, 135–154 (1989)
26. D. Dubois, H. Prade, Properties of measures of information in evidence and possibility theories. *Fuzzy Sets Syst.* **24**, 161–182 (1987)
27. D. Dubois, H. Prade, *Possibility Theory – An Approach to Computerized Processing of Uncertainty* (Springer, Paris, 1988)
28. R.P. Srivastava, Decision making under ambiguity: a belief-function perspective. *Arch. Control Sci.* **6(XLII)**, 5–27 (1997)
29. P. Smets, Decision making in the TBM: the necessity of the pignistic transformation. *Int. J. Approx. Reason.* **38**, 133–147 (2005)
30. F. Smarandache, J. Dezert, *Advances and Applications of DSMT for Information Fusion*, vols. 1–4 (American Research Press, 2004–2015). <http://www.onera.fr/fr/staff/jean-dezert>
31. P. Smets, The transferable belief model. *Artif. Intell.* **66**, 191–234 (1994)
32. R.R. Yager, Hedging in the combination of evidence. *J. Inf. Optim. Sci.* **4**, 73–81 (1983)
33. F. Smarandache, J. Dezert, J.-M. Tacnet, *Fusion of Sources of Evidence with Different Importances and Reliabilities* (Workshop on the Theory of Belief Functions, Brest, 2010)
34. A. Martin, A.-L. Jousselme, C. Osswald, Conflict measure for the discounting operation on belief functions, in *11th International Conference on Information Fusion*, Cologne (2008), pp. 1003–1010
35. L.A. Zadeh, *On the Validity of Dempster's Rule of Combination. Memo M79/24* (University of California, Berkeley, 1979)
36. A. Tchamova, J. Dezert, On the behavior of Dempster's rule of combination and the foundations of Dempster-Shafer theory, in *6th International Conference of Intelligent Systems*, Sofia (2012), pp. 108–113
37. J. Dezert, A. Tchamova, On the validity of Dempster's fusion rule and its interpretation as a generalization of Bayesian fusion rule. *Int. J. Intell. Syst.* **29**, 223–252 (2014)
38. A. Martin, C. Osswald, A new generalization of the proportional conflict redistribution rule stable in terms of decision, in *Advances and Applications of DSMT for Information Fusion – Collected Works*, ed. by F. Smarandache, J. Dezert, vol. 2 (American Research Press, Rehoboth, 2006)
39. J. Dezert, D. Han, J.-M. Tacnet, S. Carladous, Decision-making with belief interval distance, in *4th International Conference on Belief Functions*, Prague (2016)
40. A. Irpino, R. Verde, Dynamic clustering of interval data using a Wasserstein-based distance. *Pattern Recognit. Lett.* **29**, 1648–1658 (2008)
41. D. Han, J. Dezert, Y. Yang, New distance measures of evidence based on belief intervals, in *3rd International Conference on Belief Functions*, Oxford (2014), pp. 432–441
42. Y.-M. Wang, Y. Luo, On rank reversal in decision analysis. *Math. Comput. Model.* **49**, 1221–1229 (2009)
43. Y.B. Shin, S.G. Lee, D. Chun, D. Chung, A critical review of popular multi-criteria decision making methodologies. *Issues Inf. Syst.* **14**, 358–365 (2013)
44. J.-M. Tacnet, J. Dezert, M. Batton-Hubert, AHP and uncertainty theories for decision making using the ER-MCDA methodology, in *International Symposium on AHP*, Sorrento (2011)
45. S. Carladous, J.-M. Tacnet, J. Dezert, G. Dupouy, M. Batton-Hubert, A new ER-MCDA mapping for decision-making based on imperfect information, in *4th International Conference on Belief Functions*, Prague (2016)
46. S. Carladous, Approche intégrée d'aide à la décision basée sur la propagation de l'imperfection de l'informaiton – application à l'efficacité des mesures de protection torrentielle. Ph.D. Thesis, ENSMSE: Saint-Étienne (2016)

47. H. Yin, J. Lan, X.-R. Li, Measures for ranking estimation performance based on single or multiple performance metrics, in *16th International Conference on Information Fusion*, Istanbul (2013)
48. J. Dezert, D. Han, J.-M. Tacnet, Multi-criteria decision-making with imprecise scores and BF-TOPSIS, in *20th International Conference on Information Fusion*, Xi'an (2017)
49. F. Smarandache, J. Dezert, Modified PCR rules of combination with degrees of intersections, in *8th International Conference on Information Fusion*, Washington, DC (2015), pp. 2100–2107
50. M.C. Florea, A.-L. Jousselme, D. Grenier, E. Bossé, A critical review of popular multi-criteria decision making methodologies. Approximation techniques for the transformation of fuzzy sets into random sets. *Fuzzy Sets Syst.* **159**, 270–288 (2008)

Chapter 16

An Epistemological Model for a Data Analysis Process in Support of Verification and Validation



Alicia Ruvinsky, LaKenya Walker, Warith Abdullah, Maria Seale, William G. Bond, Leslie Leonard, Janet Wedgwood, Michael Krein, and Timothy Siedlecki

Abstract The verification and validation (V&V) of the data analysis process is critical for establishing the objective correctness of an analytic workflow. Yet, problems, mechanisms, and shortfalls for verifying and validating data analysis processes have not been investigated, understood, or well defined by the data analysis community. The processes of verification and validation evaluate the correctness of a logical mechanism, either computational or cognitive. Verification establishes whether the object of the evaluation performs as it was designed to perform. (“Does it do the thing right?”) Validation establishes whether the object of the evaluation performs accurately with respect to the real world. (“Does it do the right thing?”) Computational mechanisms producing numerical or statistical results are used by human analysts to gain an understanding about the real world from which the data came. The results of the computational mechanisms motivate cognitive associations that further drive the data analysis process. The combination of computational and cognitive analytical methods into a workflow defines the data analysis process. People do not typically consider the V&V of the data analysis process. The V&V of the cognitive assumptions, reasons, and/or mechanisms that connect analytical elements must also be considered and evaluated for correctness. Data Analysis Process Verification and Validation (DAP-V&V) defines a framework and processes that may be applied to identify, structure, and associate logical

A. Ruvinsky (✉) · L. Walker · W. Abdullah · M. Seale · W. G. Bond · L. Leonard
US Army Engineer Research and Development Center – Information Technology Laboratory,
Vicksburg, MS, USA
e-mail: alicia.i.ruvinsky@erdc.dren.mil; lakenya.k.walker@erdc.dren.mil;
warith.i.abdullah@erdc.dren.mil; maria.a.seale@erdc.dren.mil; william.g.bond@erdc.dren.mil;
leslie.c.leonard@erdc.dren.mil

J. Wedgwood · M. Krein · T. Siedlecki
Lockheed Martin Advanced Technology Laboratory, Cherry Hill, NJ, USA
e-mail: janet.e.wedgwood@lmco.com; michael.krein@lmco.com; timothy.siedlecki@lmco.com

elements. DAP-V&V is a way of establishing correctness of individual steps along an analytical workflow and ensuring integrity of conceptual associations that are composed into an aggregate analysis.

Keywords Verification · Validation · Data analysis · Data model · Analysis model · Epistemology

16.1 Introduction

Warnings and anecdotes about risks and pitfalls that are common to the process of data analysis are prevalent in state-of-the-science reports by the data science online community. According to Overton [14], “The challenge is finding the analytical approach that will get you safely to a prediction.” Data analysts from a variety of disciplines describe common hazards and warnings of analytical pitfalls. For example, Vasileva [20] lists several issues that may lead to erroneous analysis, including letting bias influence analysis and incorrectly stating the hypothesis. Due to the compositional nature of the data analysis process (DAP) and the inconsistent, impressionable, and rash tendency of human nature, data scientists are at high risk of either solving the wrong problem (faulty validation) or solving the right problem incorrectly (faulty verification).

The real world is teeming with examples of data scientists wasting creative effort by throwing good ideas into solutions that work brilliantly for problems that they weren’t actually tasked to solve. Consider an example from academia in which a team of talented professors and graduate students sit down at a conference table to discuss a simulation project. The simulation will provide situational awareness for first-responders and managers of a large facility in making decisions about facility evacuation during an emergency.

The project has produced several undeniably great ideas that bore fruit, and demonstrations show simulated people evacuating the simulated venue. Interesting research opportunities are apparent, and goals are being met. Perhaps most satisfying is seeing the simulated lives saved on screen. Discussions turn to proposed mechanisms and metrics for validation and verification of the simulation. Now that the project is showing potential, the research team wants to check with potential users and other experts to ascertain the correctness of the solution to their problem. There is even excitement within the team to see how the experts respond to this research.

Yet, the project leader suddenly informs everyone that the product the team worked hard to create must be scrapped. Despite everything working as planned, it just doesn’t solve the users’ problem. Unfortunately, all of this creative work occurred before anyone thought to ask the first-responders what questions the simulation should answer or what actions those answers might elicit. For example, no one thought to consult a psychologist for a list of behaviors real people might exhibit under the range of conditions imposed by an evacuation event. No one asked

the facility managers which doors could be left opened or closed, or what sorts of emergencies have occurred at other, similar venues. In short, no one asked the stakeholders or experts which questions to ask; what reasonable assumptions could be made; or what constitutes correct answers. When the team sat down to frame their solution, neither the problem nor the purpose was properly understood. The ideas that solved the team's problems were innovative but misguided because they were based on unverified and/or invalid assumptions, problem specifications, and purpose.

There are many kinds of fatal yet avoidable flaws described by the data analysis community such as:

- *Inconsistent data representation:* Situations in which the definition of variables leveraged by an algorithm are not consistent with the purpose of the data analysis. This disconnection between the purpose of the analysis and the input to the algorithm often occurs at the transition from abstract concepts defined in the analytic purpose to more specific as real-world metrics or mechanisms relevant to the operationalized analysis. The challenge here is encapsulated by Kaplan's paradox of conceptualization which states, "The proper concepts are needed to formulate a good theory, but we need good theory to arrive at proper concepts" [10]. Essentially, the less we know about a phenomenon, the less likely we are to define it (and hence measure it) correctly. Adcock and Collier describe a systematization process to facilitate the identification and specification of concepts and indicators being investigated by an analytic process [2]. Examples of decompositions of definitions from abstract purpose to specific real-world measures can be found in Fearon and Laitin [6].
- *Data analysis with no relevant context:* Applying an algorithm to data is not data analysis. Data analysis requires reasoning for choosing an appropriate algorithm to apply to a relevant data set in order to produce information that addresses a particular question or purpose [11].
- *Phantom populations or sample bias:* To draw appropriate conclusions about a population, data analysis must be conducted on a representative sample of that population. Sometimes data analysts do not realize that their data contains biases that misrepresent their target population. An example presented in Harford [7] describes a mobile app called Boston Bump that claimed to identify potholes on streets in Boston based on accelerometer data from smartphones. What they did not consider is that the data would be generated by a biased sample of the inhabitants and commuters that owned and operated smartphones. This population was not representative of the breadth of commuters on Boston's city streets.
- *Misdirected or unfalsifiable models:* Misdirected models result when researchers create models based on biased or mistaken theories. These cases may result in self-fulfilling analysis of the data generated by the biased model, thereby perpetuating the misdirection and characterizing unfalsifiable science. For example, Google Flu Trends' (GFT) claim to fame was its ability to quickly, accurately, and cheaply predict flu spread based on theory-free analysis of correlation between

search terms made by people and whether they had flu symptoms.¹ GFT became more notorious 4 years after the publication of its success in nature. That year, GFT was significantly wrong in its forecasting. The pit GFT fell into was that, as a theory-free analysis, it could not effectively falsify any of its underlying associations. If the correlations being leveraged is, “searching for ‘flu symptoms’ means you have flu symptoms,” the algorithm simply used the correlation to make predictions without evaluating the falsifiability of the correlation’s claim. This resulted in a misdirected model that was responsive, fast, and inexpensive but also wrong [7, 12].

Data analysis processes involve evaluating objects and observations by collecting and analyzing data from the real world. In turn, analysts use these mechanisms to understand reality better. We must systematically and consistently question and evaluate the constructs we use to define and design the data analysis processes by which we interpret and understand the world. Otherwise, the answers our mechanisms produce and our subsequent understanding of the world are at risk of being meaningless or wrong; the work and inspiration that went into their creation wasted.

The following chapter tackles validation and verification (V&V) of data analysis processes (DAPs). DAPs are often less explicit or well defined than software engineering, modeling, simulations, or other processes routinely subjected to V&V, because the analysis process is not a deliverable product as is software, models, and simulations. Despite the complication in delineating the steps to a DAP, their validation and verification is no less important. This chapter demonstrates how to delineate, decompose, and document the steps and components of a DAP, yielding a more concrete path to reliable, repeatable, and defendable solutions.

This chapter presents a model for DAPs to support V&V and considers what such a model entails, as well as the pitfalls of not having V&V for DAPs. Section 16.2 provides definition and background to support our discussion of the data analysis process and mechanisms of V&V. Section 16.2 also identifies and describes shortfalls that DAPs are susceptible to without an explicit V&V process integrated into the analytic workflow. Section 16.3 presents a model for capturing and documenting foundational information about a DAP, as well as the logical constructions built upon this foundation. This is followed by the presentation of a V&V process for DAP that leverages the foundational and constructed definitions to establish specific dimensions of verification and validation of the DAP. Section 16.4 then presents two use cases to demonstrate how V&V for DAP has been leveraged to identify and mitigate V&V issues in data analysis processes. The chapter is concluded with a brief summary and vision of the future.

¹It is important to note that Google Flu Trends is no longer active having been terminated in 2015 [17].

16.2 What Is Data Analysis and DAP?

In most cases where some kind of data is used for gainful insight, lack of data is not the issue. In fact, there is often so much data to ingest that it becomes difficult to procure any perception from the data. Consequently, there exists a need for a method to interpret the data efficiently; hence, there is data analysis. A community-built definition for data analysis is “a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making” [22]. The process of data analysis consists of mechanisms that make assumptions, identify or create logical associations, and generalize or otherwise trade complexity or uncertainty for clarity and the potential for generalizability.

When approaching any data analysis problem, the analytic options can seem overwhelming. At a high level, the options are categorized into four distinct areas. A robust analytic environment includes descriptive (what happened?), diagnostic (why did it happen?), predictive (what will happen?), and prescriptive (how can we make it happen?) analytics. In order to have a holistic view of data, all options co-exist and complement each other. Descriptive analytics are used to summarize and describe data. Creditors use descriptive analytics to assess credit risk by analyzing a person’s past financial behavior. This provides a good indication of their current and future financial performance, but it is not a predictive analysis in that there is no predicted debt or some other forecasted value. Diagnostic analytics help discover or determine why something happened through the use of tools and techniques that incorporate search, filter, and compare functionality, as well as integrating other data sets to integrate other contexts into an analysis. For example, education may be relevant to better understanding a consumer’s spending habits. Predictive analytics moves beyond what happened and why to discover insight about future events. Common uses include fraud detection, marketing optimization, risk reduction, resource management, etc. Prescriptive analytics provide methods for determining what decisions should be made or steps taken to produce an intended outcome. It pushes the analytic process from hindsight and insight to foresight analysis [3].

The sequence of interrelated steps utilized in data analysis can be defined as the data analysis process (DAP). The DAP is often embedded in an analyst’s mental model of what the data represents in the world and the mechanisms that create or impact the data. Traditionally, the DAP begins with a defined and documented problem. The problem definition and corresponding documentation focuses the entire analysis process on obtaining results. In some cases, enough information is not available at the start of the analysis process. Therefore, defining the problem and planning each step will aid in developing guidelines to follow throughout the entire project. Once a plan is developed, data is prepared by obtaining, cleaning, normalizing, and transforming data into an optimized data set.

16.2.1 What Is Verification and Validation?

Verification and validation (V&V) are well-known and well-used terms within many scientific fields. As with any well-used term, definitions abound with varying specificity.² A popular and useful formulation for identifying and distinguishing verification and validation is the quintessential verification question of “Did we build the thing right?” and the quintessential validation question of “Did we build the right thing?”³ By asking whether we “built the thing right,” we measure the fidelity of the model, asking whether it conforms to the specifications that purport to describe its operation. By asking whether we “built the right thing,” we measure the reliability of the model, asking whether it will reliably and accurately accomplish the work in the world that is promised.

V&V are vital techniques for controlling risk. By vetting a DAP along these dimensions, costly errors resulting from the deployment of flawed steps along the process can be avoided. Traditionally, a formal approach to V&V is based on three fundamental assumptions [13]:

1. The work done in the world by the vast majority of DAPs will require the manipulation of things affected by the physical laws of nature, rather than the manipulation of human beings affected by the less rigorous laws of psychology or other social sciences.
2. The science used by the designers of these DAPs is itself valid and unproblematic.
3. The DAPs to be subjected to V&V are computational in nature and operationalized as computer software.

From the perspective of decision-makers, V&V certainly adds value to mission accomplishment. The problem is that V&V is an expensive, difficult, and perpetual process. Our task is to formulate a protocol that adapts the traditional approach to V&V to the distinctive challenges presented by DAPs; i.e. DAPs are primarily designed and intended for understanding and/or manipulation by human beings and human groups. As guided by DOD’s injunction in VV&A Recommended Practices Guide [21]: The Key Concepts (2011), the specific techniques and procedures used to accomplish V&V “must be tailored to match the nature of the problem.”⁴

For this work, our working definition for each term based on VV&A Recommended Practices Guide [21]: The Key Concepts (2011) are as follows:

²Attempts to nail down V&V in the “soft” sciences over the decades has resulted in various assertions of types of validity that does little to clarify the space and contributed greatly to confounding the terminology regarding the study of validation in these spaces [2, 5, 13].

³“Key Concepts of VV and A” Sept. 15, 2006; official DOD pp. 7–8; <http://vva.mscos.mil/Key/key-prd.pdf>.

⁴Ibid. p. 6.

- *Verification:* The process of determining that a model implementation and its associated data accurately represent the developer's conceptual description and specifications.
- *Validation:* The process of determining the degree to which a model and its associated data provide an accurate representation of the real world from the perspective of the intended uses of the model.

By delineating these terms, we not only nail down the V&V aim for this work with respect to DAP, but we are also able to use these definitions to expose why V&V of DAP is so critical and challenging. The analysis presented here is aimed at capturing the mental model such that it may be documented, verified, and validated.

16.2.2 Shortfalls of DAP Without V&V

The DAP leverages aspects of the scientific method. Since the scientific method does not support validation and verification, neither does the current DAP. For example, it does not require assumptions or boundaries of the hypothesis to be captured. Though good science does implicitly capture assumptions and boundaries, it does not explicitly depend on them; therefore, the process does not break in their absence or even create opportunities for exposure of lack of assumption or boundary specification. Consequently, the current DAP does not guarantee results that are verified and valid. Due to its lack of V&V, the current DAP succumbs to (1) a lack of accountability of inconsistent assumptions, (2) a lack of falsifiability, and (3) increased misrepresentation of the data.

16.2.2.1 Inconsistent Assumptions

Inconsistent assumptions in the DAP refer to speculations about the data or analytics on the data that are not steady throughout the life cycle of the process. Such fallibility emerges when assumptions go undocumented and unchecked. For example, if at one stage of the DAP, the assumption is made that the associated data is measured in some units, and then, at some other stage in the DAP, this assumption is unaccounted for and the data is assumed to be measured in another unit and analyzed as such. A prime example of this type of V&V oversight is NASA's Mars Climate Orbiter in which distinct software elements of the orbiter assumed the use of different metric systems causing an accumulation of error in the spacecraft's trajectory. The ultimate and costly result was the loss of the orbiter [19]. Conflicting assumptions can propagate through the DAP, resulting in unforeseen inaccuracies.

16.2.2.2 Falsifiability Management

“A *falsifiable* claim is one for which there is some observation (or set of observations) we could make that would show us that the claim is false” [18]. Falsifiability was historically associated with the medical field. For example,

Benjamin Rush, a famous American surgeon and patriot, would justify treatment of bloodletting with the logic that: If a patient died, “The disease was too far gone for the treatment to work” if the patient recovered, then, “the treatment worked!” (Cedar’s Digest). This is not a sufficient method for determining when an experiment fails. If there is no way to determine failure, there is no way to minimize or eliminate failure. Without V&V, the current DAP does not have an expedient for identifying what qualifies as a failed experiment. Hence, the current DAP does not support a means to verify and validate evidence of failure, thus it does not support falsifiability management. Lack of falsifiability poses an issue in the DAP when an occurrence in the data or analytics is not desirable, but because there is no falsifiability, there is no account for this failure.

16.2.2.3 Misrepresentation of Data

Misrepresentation of data is concerned with using data that is not sufficient for the problem domain. This happens a lot with re-use of data. Data collected for one purpose is used for another purpose. This poses a problem in the current DAP without V&V because encodings of data can be dated or obsolete. Without V&V, such inaccuracies will not be identified.

16.3 Modeling, Verifying, and Validating the DAP

Traditionally, V&V is an empirical or operational level process which assesses how well a model performs when compared to (1) the model developer’s specification (i.e., verification), and (2) the real world (i.e., validation). A DAP is a workflow of logical and associated analyses; it is not solely the specific analytical steps that compose the process. The specific steps may be traditionally V&V’ed as they are evidential in nature, but the association of one step to the next is logical in nature and hence an aspect of the brain’s knowledge generation. We need another apparatus for modeling the DAP to support V&V of this knowledge.

The proposed model for verification and validation (V&V) of the data analysis process (DAP) is initially based on the Kantian construction of Transcendental Deduction [8] in which the philosopher Immanuel Kant claims that we can only know reality as our active minds structure, organize, and form our conditioned experiences of reality [4] and delves into an epistemological analysis of how people create knowledge in the world. The epistemology of knowledge is what we aim

to capture in this model of the DAP to enable the verification and validation of that knowledge and ultimately generate confidence and trust in the particular documented DAP.

16.3.1 An Epistemological Hierarchy Model for Decomposing the DAP

“Thoughts without content are empty; intuitions (perceptions) without concepts are blind.” – Immanuel Kant

The DAP is a workflow of logical and associated analyses applied to data to generate an understanding of the latent or indirect connections that exist within the data. Similar to human perceptions of the world, the DAP proposes to produce knowledge about the world by collecting observations of the world and reasoning on these observations. These processes cannot take place in a vacuum but must be firmly situated on an ontological foundation that provides meaning to the analytical result. One method of achieving this ontological foundation is through development of an epistemological decomposition of the data and the analytics of the DAP.

Kant begins an epistemological decomposing of data and analysis by declaring that without concepts our perceptions of the world are without meaning. Our measures of perceptions (i.e., data) require valid association to concepts in or about the world that the measures are describing. Similarly, analytical thought requires valid association to worldly observation. Without validity from concept to data and data to reason, what is produced from the DAP cannot be claimed as knowledge. In this way begins our epistemological decomposition from concepts to data and from data to analytical thought. (See Fig. 16.1.)

It is useful to note that the model presented here builds significantly on an epistemological model developed by Ruvinsky, Wedgwood, and Welsh [16] for V&V of scientific inquiry into social science models.⁵ In the Ruvinsky, Wedgwood, and Welsh model, the aim is to decompose a social science model into the pieces of knowledge that it leverages in its representation of the world in order to evaluate whether the composition of those pieces, as defined by the model, is correct and consistent with the human experience of the real world. According to Ruvinsky, Wedgwood, & Welsh:

The Epistemological Hierarchy (EH) is a framework for parsing distinct yet related levels of analysis from extremely abstract (i.e., theory) to extremely concrete (i.e., data).

The motivation for such a hierarchical decomposition of a model was based on the

⁵Though the purpose for defining an epistemological hierarchy (EH) model of knowledge elements was for evaluating the verification and validation of Human, Social, Cultural, Behavioral (HSCB) models, the mechanism is applicable to any kind of inquiry-based modeling. The prerequisite for an EH decomposition of a model is a Kantian composition of knowledge elements defined as observable concepts and reasoned understanding over those concepts [8].

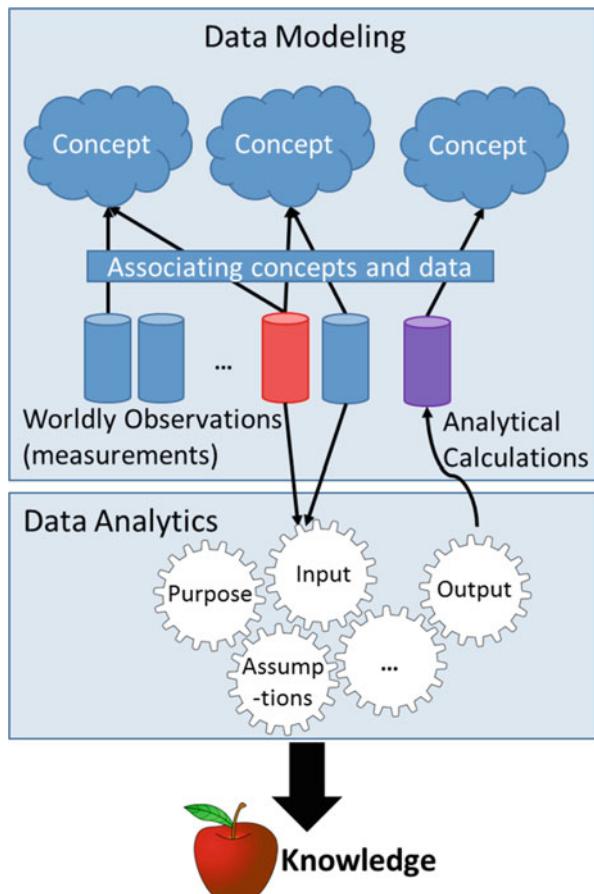


Fig. 16.1 Data Analysis Process consists of identifying relevant concepts, defining their measurement, and reasoning over the measurement data with analytic mechanisms

need to interrogate the correctness (verification) and appropriateness (validation) of the informational aspects of a model beyond solely the data it produces. MESA (Model Evaluation, Selection and Application) is able to assess the V&V of a model at each stage of model development, from the initial theoretical conceptualization that inspires the creation of a model to the specific operationalizations that define the real-world measures of relationships to be tested. In order to analyze a model in this way, one must decompose a model in terms of the knowledge it leverages and produces at each stage of development. In other words, one must break a model down into its epistemological elements ranging from the social ontology to the raw data.

In defining an epistemological decomposition of the DAP, we will leverage Ruvinsky, Wedgwood, & Welsh's Epistemological Hierarchy (EH) decomposition of a model for "better capturing and understanding information that a model uses to represent the world." The EH model-based decomposition of a DAP must include

a representation and understanding of the data itself, followed by a decomposition of the analysis that will leverage the data. The Epistemological Hierarchy Model for DAP (EHM-DAP) will be made up of two EH models: a data model and an analysis model. In Sect. 16.3.1.2, we will present a decomposition of data that supports measurement validity between the data and its associated concepts. This will be followed in Sect. 16.3.1.3 by the re-appropriation of Ruvinsky, Wedgwood, and Welsh's epistemological model of scientific inquiry [16] from scientific inquiry to the less stringent analytic process of data analysis. The extension to Ruvinsky et al. [16] presented here is a description of how the data model and analysis model work together in the DAP thereby enabling V&V of the DAP (Sect. 16.3.1.4).

It is important to note that this is not a model for *doing* a DAP; it is a model for V&V of the DAP. Though it is consistent with how data analysis can be done, it is not designed to optimize the design or development of the DAP. The recommended use is to apply this model alongside the actual data analysis design/development process such that as data is explored, the data model may be instantiated, verified, and validated; as appropriate analysis is identified, the analytic model may be instantiated, verified, and validated; and as both of these models evolve, the changes may be captured, verified, and validated. Examples of instantiating or evolving the data and analytic models are provided in Sect. 16.4.

16.3.1.1 The DAP Model: Data + Analysis

Data analysis aims to figure out what story the data is able to tell. Though there are disparate applications of data analysis throughout scientific, business, economic, political, and various other domains, there is general consensus as to what data analysis is. A representative working community definition is as follows.

The process of evaluating data using analytical and logical reasoning to examine each component of the data provided. This form of analysis is just one of the many steps that must be completed when conducting a research experiment. Data from various sources is gathered, reviewed, and then analyzed to form some sort of finding or conclusion.⁶

The DAP is essentially an iterative scientific inquiry. An individual analysis within a DAP is a single scientific inquiry whose result informs the purpose of the DAP and evolves the analysis forward into a refined analysis, a refined data set, or both. To capture this process, the DAP model merges the epistemological models for data and analysis into a single model and incorporates a workflow that captures the process of data analysis.

The DAP model will capture data, analysis, and the flow that integrates data and analysis together to generate new analytical knowledge.

⁶<http://www.businessdictionary.com/definition/data-analysis.html>

In Sect. 16.3.1.3, we present how the data model captures the conceptual construction of the data set being investigated. In Sect. 16.3.1.3, we present how the analysis model captures the analytical construction from theory to social model to hypothesis to application. The epistemological model of data analysis will integrate the models of data and analysis. This model and the analytic process that leverages this model will be further described in Sect. 16.3.1.4 once the individual data and analytic models are described.

16.3.1.2 DAP Composite Model: Data Model

As Kant stated, “... intuitions (perceptions) without concepts are blind.” Kant’s claim is that perceiving, measuring, and collecting data about the world must be associated with concepts to create knowledge about the world. Data is worthless if we have no awareness of what the data is and what it describes. The purpose of an epistemological data model is to associate data with concepts and abstract or refine those concepts as guided by the data analysis purpose. The data model is capturing knowledge about the world and building associations between knowledge elements resulting in conceptual representations of the world. For example, a data set collected for the weight of apples is associated with the concept of “apple”; apples are associated with the concept of food supply; food supply is associated with cargo. This concept of cargo may then be used by an analytical model for calculating a distribution plan. Since apples are a specific version of the cargo concept, the distribution plan may be instantiated with the apples cargo. The association of apple to food supply to cargo is an example of a data model capturing a representation of the world as guided by the analytical purpose, specifically distributing apples. (The relationship between the data model and the analytical model is presented in Sect. 16.3.1.1.)

The data model does not require a purpose for analysis. We may simply be interested or curious about a data set and want to learn more about it without having an analytic purpose in mind. It is, though, helpful to be aware of the analytic purpose that motivated the generation of the data set to provide better, more appropriate curation, pre-processing, and understanding of the data and even, more importantly, reuse of an existing data set. Since there is always some amount of uncertainty attached to data, awareness of the data’s context enables reliability. For example, when counting homeless people in a city, how does the data collector define “homelessness”? Is someone considered homeless if they are unemployed yet living with charitable family? [15] According to Yau, “[Context] can completely change your perspective on a data set, and it can help you decide what the numbers represent and how to interpret them” [23].

The Epistemological Data Model Hierarchy is one of the two composite pieces of the DAP model and is represented in Fig. 16.2.

The initial decomposition of the data is driven by a need to know what the data is and is based on any available information or resource that may be used to better understand what the data was designed to represent in the world [2]. For

Epistemological Data Model Hierarchy	Definition of hierarchical level	Example: Data set containing tweets
Abstract Concept	An abstract concept is a concept that has a broad constellation of meanings and understandings that may be associated with it.	Society
Domain Concept	A domain concept is a narrowing of the selection of concepts and relevant meanings and understandings of the abstract concept. A domain concept further specifies or determines the particular meaning of the abstract concept.	Economic State Social State Political State
Specific Concept	The specific concept is a refinement of the domain concept such that the specific concept is the lowest level of abstraction characterizing the things in the world that will be measured. Below specific concept, there is no more abstraction; there is now precision.	Social State Vulnerabilities Social State Manifestation
Systematized Data Concept	The systematized data concept is an operationalization of the concept as a measurable aspect of the real world.	News Reports Social Media Data
Model Data	The model data are the actual data in the data set. This data may be raw measurements of the real world, or they may be calculated or aggregated data.	Facebook posts tweets

Fig. 16.2 Epistemological Data Model Hierarchy (EDMH) decomposition, definition, and example

example, if social media data is gathered for an investigation into indicators of social unrest, then the “social media data” systematized concept will be positioned as an instantiation of the “social state manifestation” specific concept (see Fig. 16.2). If the social media data was gathered for investigation of crowd sourced weather reports, then the “Social Media Data” systematized concept will be positioned as an instantiation of “Weather Events” specific concept. It might be possible to conceive of the data in another manner, but the conceptualization of data beyond the parameters for which it was collected will be motivated and driven by an analytic purpose within the DAP model described in Sect. 16.3.1.3.

The dark blue boxes in Fig. 16.2 represent an associated thread of concepts from an abstract concept to model data. The light blue boxes are other related concepts for which there may also be instantiations. For example, this thread instantiates “social state” as a domain concept to “society”, but societies also have domain concepts of economic and political states, among many others.

The data being analyzed may have been generated to support the purpose of the analyst’s current DAP (see Sect. 16.3.1.3 for more information regarding DAP purpose), or the data may have been generated for a different purpose altogether and is being reused for the analyst’s current DAP. In some cases, if the desired data is not available (e.g., contains personal information, is classified, etc.), the data may serve as a proxy for the desired data. In this situation, the EDMH process can be

helpful in illuminating nuances related to the differences between the desired data and the proxy data and what impact that might have on any analyses performed with the proxy data. In any such case, awareness of the purpose that drove the data's generation is relevant to better understanding the data's meaning, structure, and significance [15, 23].

Once the data sets have been explored, the data analyst may begin to leverage the data model via the analytic model described in Sect. 16.3.1.2. Guided by the data analysis purpose, the analytic model may result in refinements, extensions, or augmentations of the data model. This intertwined process of data conceptualization and theory formulation will be further described in Sect. 16.3.1.3.

In Sect. 16.3.2, we will show how the modeling of data supports DAP validation. The opportunity for validation will present itself in the model of the DAP as the analyst begins to leverage the observed concepts captured by the data model to produce new knowledge by way of reasoned understanding. The V&V process conducted on the instantiated DAP model will evaluate the appropriateness of the application of the analysis to the data. Before DAP modeling may be verified and validated, though, the reasoning will need to be identified and documented in the analysis model presented next.

16.3.1.3 DAP Composite Model: Analysis Model

As Kant stated, “Thoughts without content are empty...” The reasoned understanding as arising from our brains requires content as an input. This content amounts to knowledge. Kant defines knowledge from observed data and reasoned understanding. As such, the analysis model is tasked with capturing the reasoned understanding applied throughout the DAP to observed knowledge (i.e., data) as well as reasoned knowledge (i.e., supported conclusions).

Every DAP must have a goal or goals in mind in order to focus the direction of analysis. These goals are captured as purpose. The DAP analysis model purpose is the analytic reasoning that motivates and drives the DAP. (See Fig. 16.3.) If the aim of an analytic process is simply to explore the data, then that in itself is a purpose that guides the kind of analytic choices that the analyst makes. The purpose is critical because it guides the analyst regarding how to instantiate a conceptual

	Definition	Example
Data Analysis Process Purpose	The data analysis process purpose is the desired aim or goals that motivates and directs the data analysis process.	What factors contribute to a political candidate winning an election?

Fig. 16.3 DAP purpose definition and example

Epistemological Analysis Model Hierarchy	Definition of hierarchical level	Example: What is the current sentiment of a population
Theory	Abstract statements about reality describing relationships between or among concepts	Society exhibits a social state at any given moment in time
Social Model	A representation of real world system behavior based on theories and concepts	Social state consists of a sentiment characteristic
Hypothesis	Conjectures within a theory regarding the relationship of two concepts to be explored in the Social Model	Social anguish is part of the sentiment characteristic of a social state
Application	A description of how the Social Model will be refined and the concepts operationalized	Social media analytics can expose social anguish messaging in social media data
Implementation	The equations, parameter settings, and coding rules to enable execution of the application model and manipulation of the raw data into model parameters	Sentiment analysis on twitter data to expose sentiment vector

Fig. 16.4 Epistemological Analysis Model Hierarchy (EAMH) decomposition, definition, and example

construction around the data as represented by the data model (described in Sect. 16.3.1.2). Consider, for example, the apple concept defined in Sect. 16.3.1.2. In that example, apples were defined as an instance of a food supply in order to conceive of it as possible cargo. Apples may also be considered in other ways such as fruit, a source of nutrients, etc. How the concept of apples is instantiated within the data model of the DAP model will depend on the purpose of the analysis.

To capture the epistemological elements that researchers utilized in a model development, Ruvinsky, Wedgwood and Welsh designed an epistemological hierarchy of levels of abstraction of social science research design shown in Fig. 16.4. The Epistemological Analysis Model Hierarchy is the second of two composite pieces of the DAP model.

16.3.1.4 Modeling DAP Workflow Leveraging EH Models of Data and Analysis

The DAP model integrates and extends the individual data and analysis models presented in earlier sections. In particular, the DAP model extends the analysis model by adding conceptual levels to the analytic model hierarchy that are instantiated as abstract concepts in the data model hierarchy. The analytic model is also extended to include a data level that specifies the actual data sets to be leveraged by the analysis. Along these lines, the data model is extended to include not just model data defined as raw input data but also calculated model data which are data sets that result from analytics executed on raw data sets. A visual representation of the integrated and extended data and analytic models into the DAP model is shown in Fig. 16.5.

The DAP model ensures representation of the full spectrum of the DAP. The instantiation of the DAP model for a specific DAP will not likely be performed from top to bottom (or vice versa) in one analytical thread. There are sublevels of the DAP that correspond to different data analysis tasks. These sublevels may be instantiated

Epistemological Analysis Model Hierarchy	Epistemological Data Model Hierarchy
Ontology: The set of background entities and beliefs about the world that pertain to and characterize a basic structure of reality	Abstract Concept
Paradigm: The intuition under which the specified ontology elements are experienced. According to Kant, the paradigms that form human knowledge are the Faculty of Sense (e.g., time and space) and the Faculty of Understanding (e.g., causation, inertia, etc.)	
Theory	Domain Concept
Social Model	
Hypothesis	Specific Concept
Application	Systematized Data Concept
Implementation	Calculated Model Data: The calculated model data is calculated data from either model data or other calculated model data.
Data: Actual data produced by selection of data bases and methods to access data sets from specific databases	Model Data

Fig. 16.5 Epistemological DAP model: an integration and extension of the epistemological analysis model and the epistemological data model

individually, yet inform each other to ensure consistency of analysis. For example, when executing data cleaning, the analyst will focus on the implementation layer and data layer of the analytic hierarchy model and the calculated model data and model data of the data hierarchy model. This analysis will generate the identification and cleaning of data consistent with the requirements of the implementation layer. At another time in the analysis, the same analyst is interested in investigating hypothesized relationships between variables in the data set in which the analyst will focus on the higher levels of the DAP model and capturing the conceptual logic of the relationships being investigated.

The DAP model is designed to be flexible and robust so that an analyst can leverage sub-elements or components of the model as needed. Though DAP modeling is not restricted to a specific process, we will provide an example that goes through multiple steps of a typical DAP. We will begin with a data set to be investigated and modeled. As shown in the example in Fig. 16.6, given a data set of Twitter Data, the data set is cleaned and oriented into a conceptual framework that will inform and guide what and how this data is used for analysis. The dark blue boxes in Fig. 16.6 represent the instantiation of the specific Twitter data set from raw model data to abstract concepts. The light blue boxes represent similar data/concept elements within the same data ecosystem as the instantiation of the Twitter data model. For example, social media data is similar in context to news reports. Both of these data elements may be used to represent social discourse, though in this instance, we are dealing with Twitter data which is an instance of social media data which serves as a means of investigating and measuring social discourse.

Epistemological Analysis Model Hierarchy	Example Data Analysis Process	Epistemological Data Model Hierarchy
PURPOSE:		
Ontology	Economy Government Society	Abstract Concept
Paradigm	Faculty of Sensibility Faculty of Understanding Time/Space Causal	
Theory	Economic State Social State	Domain Concept
Social Model	Political State	
Hypothesis	Social Action Social Discourse	Specific Concept
Application	News Reports Social Media Data	Systematized Data Concept
Implementation	Cleaned: text opinion data Cleaned: word count aggregations	Calculated Model Data
Data	Facebook Data Twitter Data	Model Data

Fig. 16.6 Example of an instantiated epistemological data model

From the data set, an analytic purpose will be identified, and an analysis approach modeled. As detailed in Fig. 16.7, the data elements in blue with red border are data products of the analytic model. For example, the “Sentiment” domain concept was brought into the DAP model as a result of the “Social state consists of a sentiment characteristic.” Also, the “Cleaned: sent. Vector” is a calculated model data set consisting of the sentiment vectors that result from analyzing the Twitter data set. It is important to note that the V&V of the implementation of the sentiment analysis code is not what is being conducted here. The V&V of sentiment analysis code is being leveraged by this DAP instance, meaning that the analysis was conducted elsewhere previously and is being leveraged by this DAP instance. Here, we are assuming that the implementation used is verified and valid. Here, we are asking if applying sentiment analysis to the Twitter data for the purpose of capturing social anguish can be verified and validated.

The process will then assume the execution of the analysis on the data and will generate new data that will either support or not support the hypothesis. The result is then made available and integrated into the DAP model and will drive the next step in the analytic process described in Fig. 16.8. Here, we see that the data products generated in Fig. 16.6 drove the generation of new analysis products. For example, the new anguish data set generated from social media data needs to be corroborated to show support for the original application claim that “Sentiment analysis can expose social anguish messaging in social media data.” These new

Epistemological Analysis Model Hierarchy	Example Data Analysis Process		Epistemological Data Model Hierarchy
PURPOSE: How does the population of Country_X feel about government response to Natural_Disaster_Y?			
Ontology	Society		Abstract Concept
Paradigm	Time/Space		
Theory	Society exhibits a social state at any given moment in time	Social State	Domain Concept
Social Model	Social state consists of a sentiment characteristic	Sentiment	Domain Concept
Hypothesis	Social anguish is part of the sentiment characteristic of a social state	Social Discourse Social anguish	Specific Concept
Application	Sentiment analysis can expose social anguish messaging in social media data	Social Media Data	Systematized Data Concept
Implementation	Sentiment analysis on twitter data to expose sentiment vector Extract social anguish sentiment measures from sentiment vector	Cleaned: sent. vector Cleaned: anguish sent. data	Calculated Model Data
Data	Twitter Data		Model Data

Fig. 16.7 Example of instantiated analytic purpose and epistemological analysis model

analytic components to the DAP model instance then drive new data components, and the process continues until all claims and purpose are ultimately, satisfactorily addressed. With respect to the example presented here, the integration of the analytic result and the next analysis it drives will be the final step shown in this example, though the process itself would continue until all claims are resolved and the analytic purpose is met satisfactorily.

The interplay between analysis and data is an important aspect of data analysis that this model strives to capture. According to Abraham Kaplan, “Proper concepts are needed to formulate a good theory, but we need a good theory to arrive at the proper concepts . . . The paradox is resolved by a process of approximation: the better our concepts, the better the theory we can formulate with them, and in turn, the better the concepts available for the next, improved theory” [9]. The DAP model presented here attempts to capture this iterative process to facilitate the improvement of concepts and theory.

Though the analyst may clean the data and model relationships in the data non-contiguously, the analyst must ensure consistency in logic between the modeling

Epistemological Analysis Model Hierarchy	Example Data Analysis Process		Epistemological Data Model Hierarchy
PURPOSE: How does the population of Country_X feel about government response to Natural_Disaster_Y?			
Ontology	Society		Abstract Concept
Paradigm	Time/Space		
Theory	Society exhibits a social state at any given moment in time	Social State	Domain Concept
Social Model	Social state consists of a sentiment characteristic	Sentiment	
Hypothesis	Social anguish is part of the Does social anguish data from social media corroborate with other indicators of social anguish?	Social Discourse Social anguish	Specific Concept
Application	Sentiment analysis can expose Social anguish data from social media corroborates with social anguish indicators such as ...?	Other Social State Content Provider	Systematized Data Concept
Implementation	Sentiment analysis on twitter data to expose sentiment vector Extract social anguish sentiment measures from sentiment vector	Cleaned: sent. vector Cleaned: anguish sent. data	Calculated Model Data
Data	Twitter Data		Model Data

Fig. 16.8 Interaction between analytic development and data development of the DAP

and data cleaning such that the clean data captures appropriate representations of the measured relationship. The DAP model supports the verification and validation of the DAP such that any issues with consistency of logic are exposed for detection. The process of using the DAP model to support V&V is described in detail in Sect. 16.3.2.

Once the analysis is executed on the data and a result is attained, this result will generate new data that will either support or not support the hypothesis. Support is good but not the end. Analysis continues as the analyst must build more support for a claim by showing that other representations of the relevant concepts also demonstrate the same support. Lack of support is also not the end. The analytic construct may be further explored by other approaches such as affiliating the hypothesis with different specific concepts or exploring different relationships with the same data set.

16.3.1.5 User Guide to Instantiating a DAP Model

The DAP model described above is intended to be flexible and robust to support any approach to data analytics. To show how the model may be used to capture the DAP for verification and validation of that process, we present the questionnaire developed by Ruvinsky, Wedgwood and Welsh to facilitate the capture of the information specified by each data and analytic component of the model. Their questionnaire entitled the Epistemological Decomposition (or “e-decomp”) is presented in Table 16.1 [16].

16.3.2 *Evaluating the V&V of the DAP Decomposition*

V&V assesses how correctly a DAP is capturing and producing knowledge about a phenomenon of interest. To assess the V&V of the DAP models, one needs to define a strategy that overcomes the shortfalls described in Sect. 16.2.2. At the heart of the strategy presented here is a move to go beyond viewing verification and validation as solely empirical testing. Our approach is to provide mechanisms and techniques for evaluating knowledge by considering other aspects of knowledge that contribute to the verification and validation of knowledge. Our approach begins with the decomposition of a DAP in terms of its data model and its analysis model with respect to its epistemological elements ranging from ontology to purpose to raw data as described in Sect. 16.3.1. Beyond organizing and documenting the informational elements of a DAP, the epistemological hierarchy of the DAP provides a structure upon which verification and validation techniques may be based. To be clear, the epistemological hierarchy is not intended as a means of verifying or validating models but rather as a structure for organizing our definition and knowledge of a DAP so that verification and validation may be performed upon the structure. The epistemological content generated by the decomposition is used by the V&V tools and techniques presented here.

V&V of a DAP may be considered from two distinct perspectives: focal and contextual. Focal V&V considers the verification and validation of a model with respect to its ability to explain the target phenomena for which the model was intentionally built. Contextual V&V assesses the flexibility of a model with respect to contextual vulnerabilities or dependency. The perspective we address here is that of focal V&V, evaluating the DAP model based on the domain space for which the model was intended to perform.

We also define operational V&V techniques as those traditional V&V techniques assessing the empirical output from an algorithm or model (DoD 2006), such as cross-validation testing, analysis of variance assessment, etc. Operational V&V techniques complement the focal V&V techniques described here because operational V&V evaluates analytic performance, while focal V&V will explore beyond the empirical output analyzed by operational techniques and into a theoretical and epistemological scrutiny of the DAP model. In particular, operational V&V

Table 16.1 Questionnaire supporting the epistemological decomposition of the data and analysis models presented in Ruvinsky et al. [16]

EH tiers	Sublevels	Question
Model purpose		What is the question that the analysis is trying to answer?
Conceptual levels	Social ontology	What is the world made up of relevant to the analysis?
	Paradigm	What is the paradigm (i.e., grand scheme or worldview) that underlies the analysis? From which particular abstract concepts is this paradigm made? What is the broad domain addressed by the paradigm?
Theoretical levels	Theory	What are the theories/arguments that the model is designed to investigate? For each theory, what are the domain concepts associated with the relationship being explored? To which abstract concept is each domain concept associated? How are these domain concepts arranged into theoretical relationships? What assumptions (i.e., theoretical relationships not explicitly investigated by this model) is the theory making about the world? In other words, what theoretical claims is the theory leveraging? These may be theories proposed by other researchers and presented in other works cited by this model's researcher in supporting documentation. What is the phenomenon that this theory is trying to explain?
		What are the specific concepts associated with the relationships being explored? To which domain concept is each specific concept associated? How are these specific concepts arranged into described theoretical relationships?
		What assumptions about the context (i.e., aspects of the social model not explicitly investigated by this model) is the social model making about the world?
		What is the descriptive context in which the social model has instantiated the theory?
		What is the hypothesis (are the hypotheses) being investigated? What are the specific concepts associated with the relationships being explored? Decompose the variables associated with each hypothesis into independent, dependent, or conditional variables of the hypothesis. What is the scope of inquiry represented by these hypotheses?
	Application	What is the particular empirical context space into which the social model is being applied? (e.g., the country to which the model is being applied; the geographic region; the demographic; etc.) What are the boundary conditions associated with this empirical context space?
		What is the methodological context into which the social model is being applied? (e.g., What is the general methodology that is being applied to the model? What unit of analysis?)
		What are the boundary conditions associated with this methodological context space?

(continued)

Table 16.1 (continued)

EH tiers	Sublevels	Question
		What are the principles behind this methodology? What are the systematized data concepts associated with the specific concepts being explored in the hypothesis? To which specific concept is each systematized data concept associated? How are the relationships between systematized data concepts, specified in the hypothesis, applied in this experimentation context?
	Implementation	What specific methodological protocols are being used for this model? What are the formal expressions of the systematized data concepts that are input, output, and conditional variables? What are the formal expressions of the relationships between those input, output, and conditional variables? What are the boundary conditions associated with this choice of method?
	Data	What data sources does the model use for each variable? For each variable, what is the coding required (if any) and what data sources are required? What is the unit of measure for the input, output, and conditional variables? What are the observed relationships between input, output, and conditional variables? What is the description of the data (i.e., data type, time frame, definition, and data source) represented by the measured input data?

techniques are applied to evaluate data analysis algorithms that produce information. The DAP leverages or assembles data analytic algorithms and their operational V&V assessments to create an analytic process that transforms information into knowledge about the question or purpose being investigated. The purpose of the focal V&V approach presented here is to show how V&V of a DAP beyond empirical evaluation (with respect to the domain for which the process was defined) may be established.

16.3.2.1 The Epistemological Hierarchy as a V&V Checklist

The process of generating the epistemological decomposition of a DAP as it is being developed may provide early evidence and insight into the state of V&V of a data analysis process as it unfolds. Early evidence of DAP's issues with V&V allows for failing early so that problems are exposed during development of a DAP and may hence be corrected easily versus exposing issues after development resulting in wasted effort in developing a flawed analysis and cost in backtracking the analysis in order to correct the issue. Using the epistemological decomposition as a kind of checklist during the development of a DAP is useful for assessing whether the

analysis has the appropriate construction to support V&V and enables immediate and less costly awareness and correction of data analysis issues. The epistemological decomposition as a checklist is also useful when evaluating an existing DAP because it facilitates an assessment of whether it is feasible and cost-effective to continue with the V&V of a pre-defined DAP.

Using the epistemological decomposition of the DAP as a means of evaluating the preparedness of the process for a V&V assessment is called V&V Early CHeckpoint via Epistemological Critique (E-CHEC). If V&V failure is detected at any point along the epistemological decomposition of a DAP, the analysis is preemptively halted from any further V&V analysis because any epistemological structure based on an un-validated construct must be assumed to be itself un-validated. This not only helps in managing the verification and validation of a DAP, but it also facilitates improving the process by homing in on the pieces of the DAP model's epistemological structure that have failed verification or validation. In this way, an analyst can focus their efforts on improving the process model, and V&V of that model does not need to start from scratch each time, only needing to assess the portion of the epistemological hierarchy beginning from the point of failure from the previous V&V assessment (assuming the higher – more abstract – epistemological elements were not altered).

16.3.2.2 V&V Methodology

The approach to V&V of a DAP via the epistemological decomposition of the process is to assess modularly the V&V of each epistemological level of the process model. (See Sects. 16.3.1.1 and 16.3.1.2 for a description of the Epistemological Hierarchy models for data and analytics, respectively.) Consequently, the state of V&V of a model will be an aggregation of the V&V assessment of each epistemological level.

To address the verification of a specific level of the hierarchy, one must consider the surrounding levels. In other words, assessing the verification of one level of the hierarchy requires asking questions about that level with respect to its relationship to the levels around it. Since verification is an assessment of the implementation of a claim (i.e., “did I build the thing right?”), the analysis investigates the fidelity of the implementation’s abstraction to the claim, or conversely, the fidelity of the claim’s specification to the implementation. In other words, the line of questioning of the current level becomes, “Did I make the right abstraction of the lower level entity, or specification of the higher level entity based on (1) what has been defined at the lower/higher level, and (2) the analytic purpose?” For example, if the domain concept specified for the context model level of the DAP model is “Social State,” and the analytic purpose is to expose social sentiment, then the specific concept of “social discourse” is an adequate specification for the hypothesis level because it is consistent with the analytic intent of the analyst. Considering another example, if the hypothesis is “Social anguish is part of the sentiment characteristic of a social state,” and the analytic purpose is to expose social sentiment, then the context

model definition of “Social state consists of sentiment characteristics” is an adequate abstraction for the hypothesis level because it is logical and consistent with the modeler’s intent.

Validation is assessed by considering the modeled DAP hierarchy with respect to external constructs. In other words, the line of questioning of the current level becomes, “Did I make abstractions or specifications that are consistent with (1) evidence-based constructs prominently leveraged by the community, and (2) the analytic purpose?” For example, a domain of “social state” is valid for an analytic purpose of understanding sentiment of a population because (1) social state is a well-identified concept and (2) social state consists of a sentiment dimension which is relevant to the analytic purpose.

16.3.2.3 V&V Dimensions

As Sect. 16.3.2.2 describes, the V&V analysis we present is performed on each level of the epistemological hierarchy of the DAP model. To delve into the theoretical and epistemological scrutiny of a DAP model, we begin by defining aspect of V&V relevant to an epistemological analysis. In our approach, we considered verification and validation individually and identified dimensions for each. The dimensions of interest that we defined are listed in Table 16.2.

Table 16.2 Dimensions of verification and validation

V&V process	Dimension	Definition
Verification	Consistency of concepts	An assessment of how well concepts are specified and operationalized between levels of the hierarchy with respect to consistency
	Consistency of Relationships	An assessment of how well relationships are specified and operationalized between levels of the hierarchy with respect to consistency
	Consistency of assumptions	An assessment of how well assumptions are specified and operationalized between levels of the hierarchy with respect to consistency
Validation	Utility	An assessment of the usefulness of the epistemological element with respect to model purpose
	External consistency	An assessment of the consistency of epistemological elements of a model with respect to other epistemological elements external to the model
	Prominence	An assessment of breadth of use by the user and modeling communities
	Accuracy	An assessment of the accuracy of epistemological elements with respect to empirical findings derived from the specific element in question

16.3.2.4 V&V Questionnaire and Measures

For each dimension of V&V that we identified, we constructed questions to extract a measure for the dimension in the form of a rubric-based assessment. These questions are compiled into a V&V questionnaire designed as a tool for evaluating the adequacy of the DAP model epistemology. The assessment of the V&V of a DAP model has been decomposed into questions to be asked of each levels of knowledge in the epistemological hierarchy. The V&V questions are shown in Table 16.3. These questions make up the V&V questionnaire.

To provide a common metric by which to measure each V&V dimension per sublevel of the epistemological hierarchy, we constructed a rubric consisting of “A”, “B”, “C”, or “D” scores for each question corresponding to values of 4, 3, 2, or 1, respectively, such that a higher valued score indicates a better score in assessing valid or verified epistemological element. Since the flavor of each score is very similar among all of the questions, below we present a small subset of the focal V&V questions along with their rubrics. In particular, this sample will present the hypothesis level questions and their rubrics. Each question is indexed as either “VE” if it is a verification question or “VA” if it is a validation question.

(VE) 1. *Concept Consistency:* Are the specific concepts that compose the hypothesis included among those determined to compose the context model?

- (a) The concepts are clearly and completely consistent between sublevels. The evaluator can follow the evolution of concepts across levels of the epistemological hierarchy.
- (b) The concepts are clearly but not completely consistent between sublevels. The evaluator can follow the evolution of concepts across levels of the epistemological hierarchy.
- (c) The concepts are not clearly consistent between sublevels OR there are concepts missing. The evaluator cannot follow the evolution of the concepts across relevant levels of the epistemological hierarchy.
- (d) The concepts are not clearly consistent between sublevels AND there are concepts missing. The evaluator cannot follow the evolution of the concepts across relevant levels of the epistemological hierarchy.

(VE) 2. *Relationship Consistency:* Is the relationship between independent and dependent variables as proposed by the hypothesis within the set of described and contextualized relationships posited by the social model?

- (a) The relationships are described clearly and with reasonable completeness. The evaluator is able to follow the evolution of the relationships across levels of the EH.
- (b) The relationships are described clearly but not completely. The evaluator notes important gaps in the evolution of the relationships.
- (c) The relationships are not described clearly, are not described completely, OR are described incorrectly. The evaluator cannot follow the evolution of the relationships across relevant levels of the EH.

Table 16.3 Questions from Focal V&V Questionnaire (verification questions in blue; validation questions in green)

EH level	Sub-level	Focal V&V dimension	Question
Conceptual levels	Ontology	Utility	Are the ontological claims about the existence of categories of events and phenomena relevant to the model purpose?
		External consistency	Can the ontology co-exist with other ontologies without contradicting them?
		Prominence	Are the ontological beliefs widely subscribed to?
		Accuracy	Has subscription to those ontological beliefs elsewhere ultimately yielded empirically valid findings?
	Paradigm	Concept consistency	Are the abstract concepts that compose the paradigm accounted for in the categories established by the social ontology?
		Utility	Is the domain addressed by the paradigm appropriate to the model purpose?
		External consistency	Can the paradigm co-exist with other paradigms without contradicting them?
		Prominence	Is the paradigm widely subscribed to?
Theoretical levels	Theory	Accuracy	Has subscription to the paradigm elsewhere ultimately yielded empirically valid findings?
		Concept consistency	Do the domain concepts that compose the theory represent faithfully the abstract concepts determined to compose the paradigm?
		Relationship consistency	Is the set of relationships posited by the theory consistent with the set of axioms at the core of the paradigm?
		Assumption consistency	Do the assumptions stipulated by the theory reflect the expectations and understandings established by the paradigm?
		Utility	Is the phenomenon explained by the theory relevant to the model purpose?
		External consistency	Can the theory co-exist with other theories without contradicting them?
		Prominence	Is the theory widely used?
		Accuracy	Has usage of the theory elsewhere ultimately provided empirically valid findings?
	Model	Concept consistency	Do the specific concepts that comprise the social model represent faithfully the domain concepts determined to comprise the theory?
		Relationship consistency	Is the described and contextualized set of relationships posited by the social model consistent with the abstract set of relationships at the core of the theory?
		Assumption consistency	Do the assumptions stipulated by the social model represent the assumptions of the theory?

(continued)

Table 16.3 (continued)

		Utility	Is the descriptive context of the social model associated with the model purpose?
		External consistency	Can the social model co-exist with other social models without contradicting them?
		Prominence	Is the social model widely used?
		Accuracy	Has usage of the social model elsewhere ultimately provided empirically valid findings?
Hypothesis		Concept consistency	Are the specific concepts that compose the hypothesis included among those determined to compose the social model?
		Relationship consistency	Is the relationship between independent and dependent variable proposed by the hypothesis within the set of described and contextualized relationships posited by the social model?
		Assumption consistency	Is the scope of the hypothesis consistent with the assumptions of the social model?
		Utility	Does understanding the dependent variable explained by the hypothesis further the model purpose?
		External consistency	Can the hypothesis co-exist with other hypotheses without contradicting them?
		Prominence	Has this hypothetical claim been widely made?
		Accuracy	Has this hypothetical claim elsewhere ultimately produced empirically valid findings?
Operational levels		Concept consistency	Are the systematized data concepts that compose the application faithful representations of the specific concepts that compose the hypothesis? Are the principles of the methodology appropriate for evaluating the specific concepts that compose the hypothesis?
		Relationship consistency	Are the relationships between specified systematized data concepts presented as input, conditional, and output variables consistent with the relationships between independent and dependent variables proposed by the hypothesis? Are the principles of the methodology appropriate for evaluating the relationship between independent and dependent variables proposed by the hypothesis?
		Assumption consistency	Do the boundary conditions set by the empirical domain of the application meet the scope of inquiry stipulated by the hypothesis? Do the boundary conditions set by the methodological domain of the application meet the scope of inquiry stipulated by the hypothesis?
		Utility	Is the particular empirical domain designated by the application appropriate to the model purpose?

(continued)

Table 16.3 (continued)

		Is the particular methodological domain designated by the application appropriate to the model purpose?
	External consistency	Can the application co-exist with other applications without contradicting them?
	Prominence	Has this application been widely used?
	Accuracy	Has this application been used elsewhere and ultimately produced empirically valid findings?
Implementation	Concept consistency	Do the formal expressions that compose the implementation accurately transcribe the systematized data concepts that compose the application? Are the methodological protocols as expressed in the formal expressions of the implementation consistent with the methodology posited in the application?
	Relationship consistency	Are the formal expressions of relationships between input, conditional and output systematized data concepts consistent with the relationships between systematized data concepts described in the application?
	Assumption consistency	Do the boundary conditions of the empirical domain set in the implementation respect the boundary conditions of the empirical domain set in the application? Do the boundary conditions of the methodological domain set in the implementation respect the boundary conditions of the methodological domain set in the application?
	Utility	Does this formal expression of the elements of the application inform the model purpose?
	External consistency	Can the implementation co-exist with other implementations without contradicting them?
	Prominence	Has this implementation been widely used?
	Accuracy	Has this implementation, when used elsewhere, ultimately produced empirically valid findings?
	Concept consistency	Is the unit of measurement for input data consistent with the formal expression of the input systematized data concept? Is the unit of measurement for output data consistent with the formal expression of the output systematized data concept?
	Relationship consistency	Are the calculated relationships between input and output data consistent with the formal expressions of relationships stipulated in the implementation?
	Assumption consistency	Does the description of the data produced by the measures of data match the boundary conditions of the empirical domain set in the implementation? Does the description of the data produced by the measures of data match the boundary conditions of the methodological domain set in the implementation?

(continued)

Table 16.3 (continued)

	Utility	Does the expression of raw observations inform the model purpose?
	External consistency	Can the data co-exist with other data within the same empirical and methodological domain without contradicting them?
	Prominence	Have these data been widely used?
	Accuracy	Have these data been empirically validated?

- (d) The relationships are not described clearly, are not described completely, AND are described incorrectly. The evaluator cannot follow the evolution of the relationships across relevant levels of the EH.
- (VE) 3. *Assumption Consistency*: Is the scope of the hypothesis consistent with the assumptions of the social model?
- (a) The scope is described clearly and with reasonable completeness. The evaluator can follow the evolution of assumptions across levels of the EH.
 - (b) The scope is clearly described, but not completely such that important gaps have been noted. The evaluator can follow the evolution of assumptions across relevant levels of the EH.
 - (c) The scope is not clearly described, is not completely described, OR is described incorrectly. The evaluator cannot follow the evolution of the assumptions across relevant levels of the EH.
 - (d) The scope is not clearly described, is not completely described, AND is described incorrectly. The evaluator cannot follow the evolution of the assumptions across relevant levels of the EH.
- (VA) 4. *Utility*: Does understanding the dependent variable explained by the hypothesis further the model purpose?
- (a) The utility of the artifact with respect to the model purpose is obvious. The artifact's motivations and boundaries are clearly described and are in line with the stated model purpose.
 - (b) The utility of the artifact with respect to the model purpose is evident. There is little elaboration on the artifact's motivations or boundaries such that consistency with the model purpose is not clear.
 - (c) The utility of the artifact with respect to the model purpose is discernible with effort. The evaluator has to make significant assumptions as to the artifact's motivations and boundaries to establish consistency with the model purpose.
 - (d) The utility of the artifact with respect to the model purpose is not discernible. Even with effort on the evaluator's part, the artifact does not show consistency with the model purpose.

(VA) 5. *External Consistency*: Can the hypothesis co-exist with other hypotheses without contradicting them?

- (a) The artifact can coexist with other artifacts at this level from other model ecologies in the same school of thought. The consistencies are clearly acknowledged and addressed.
- (b) The artifact can coexist with other artifacts at this level from other model ecologies in the same school of thought. The consistencies are not clearly acknowledged or addressed.
- (c) The artifact cannot coexist with other artifacts at this level from other model ecologies in the same school of thought. The inconsistencies are clearly acknowledged or addressed.
- (d) The artifact cannot coexist with other artifacts at this level from other model ecologies in the same school of thought. The inconsistencies are not clearly acknowledged or addressed.

(VA) 6. *Prominence*: Has this hypothetical claim been widely made?

- (a) The component is well-known. An evaluator who is familiar with the research landscape would be familiar with this component.
- (b) The component is somewhat known. An evaluator who is familiar with the research landscape would have passing familiarity with this component.
- (c) The component is relatively unknown. An evaluator who is familiar with the research landscape might have passing familiarity with this component.
- (d) The component is completely alien. An evaluator who is familiar with the research landscape is very unlikely to be familiar with this component.

(VA) 7. *Accuracy*: Has this hypothetical claim elsewhere ultimately produced empirically valid findings?

- (a) The evaluator is aware of one or more instances where this model artifact at this level of the hierarchy has been used in other analyses. In addition, as far as the evaluator knows, all of these analyses have ultimately provided empirically valid findings.
- (b) The evaluator is aware of one or more instances where this model artifact at this level of the hierarchy has been used in other analyses. In addition, as far as the evaluator knows, at least one of these analyses has ultimately provided empirically valid findings.
- (c) The evaluator is aware of one or more instances where this model artifact at this level of the hierarchy has been used in other analyses. In addition, as far as the evaluator knows, these analyses have shown to be inconclusive or invalid.
- (d) The evaluator is not aware of any other instances of the usage of this model artifact.

16.3.2.5 V&V Inter-coder Reliability

With the potential of various data analysts and analysis consumers providing information about the V&V of a DAP, differences in evaluations among reviews is inevitable. To address this, we consider the application of inter-coder reliability (ICR) experimentation to show how coded reviews can be used to assess the V&V of a DAP model. Our observations with ICR experiments of the V&V Questionnaire as applied to the domain of computational social science showed that though there were differences in responses among reviewers, there was generally consensus emerging among the reviewers for each question. Our conclusion to the experiment was that as the number of individuals that review a model increases, the convergence of a verified and valid model will converge. With a small set of reviews there may be a lot of noise, but a model that is valid and/or verified will converge to high value scores. On the contrary, if a model contains issues with its validity or verifiability, this too will converge in the form of predominately low scores for the focal V&V assessment. Models for which convergence does not arise represent models based on principles that are under contention in the research community.

Our vision of this evaluation process is that as V&V scores of data analysis models begin to converge, high-likelihood issues in the model will be exposed and refinements will occur. Iterations of this process will result in the convergence of high-use models to high-value scores. In other words, the most relevant and/or useful data analytic processes will become increasingly correct and appropriate.

16.4 Use Cases

This chapter presents a model for capturing the epistemological decomposition of a data analysis process (DAP) for the purposes of supporting an inquiry of the decomposed DAP for measures of validation and verification with respect to the verification dimensions of consistency of concepts, relationships, and assumptions, as well as the validation dimensions of utility, external consistency, prominence, and accuracy. The following two use cases describe distinct, real-world examples of data analysis processes leveraging aspects of an epistemological decomposition and ultimately V&V of the data analysis process. The proposed V&V for DAP analysis is intended to be flexible and robust in scale such that the DAP decomposition and V&V analysis may expose vulnerabilities and dependencies at varying scope and scale based on the analyst's intent. These two examples effectively use the proposed tools for V&V of DAP analysis at somewhat varying levels of specificity.

16.4.1 Aircraft Prognostics Use Case

The Aircraft Prognostics Use Case is an example that demonstrates a DAP process in a non-social science domain. What appears at first to be a fairly straightforward process of using aircraft sensor data to predict aircraft faults is actually fraught with opportunities to use the data in ways that are inappropriate for individual supervised and unsupervised learning algorithms. Although the epistemological decomposition presented in this example does not consider the upper layers of the ontology and the paradigm, there are concepts that clearly need to be understood and agreed upon by all participants involved in analyzing and using the data. As data scientists use this data to develop prognostics and attempt to compare or combine prognostics, the assumptions made throughout the DAP resulting in these prognostics must be verified and validated; otherwise, the comparisons/combinations of these prognostics will not only likely be incorrect but also potentially dangerous (e.g., parts that, by other measures, should be replaced, are left in service).

This use case begins decomposing the Aircraft Prognostics DAP at the theory level of the analysis model, along with the associated domain concepts of the data model. For this epistemological decomposition, the top levels of the hierarchy, the paradigm and ontology, were not expanded. These were deemed to be irrelevant to the current analysis and as such were not expanded.

16.4.1.1 Purpose

The purpose of the Aircraft Prognostics Use Case is to answer the question: *Can the health of a specific aircraft component, particularly the Line Replaceable Unit (LRU), be classified through observation of aircraft sensor data and/or field-collected repair and maintenance (R&M) data?*⁷ The goal of such a classification is to identify leading indicators that predict component failure. If component failure can be predicted, it is possible to decrease the amount of time an aircraft is unavailable due to unplanned maintenance. A clearly defined purpose enables a clearer look at the big picture of the data.

This study started the model and data analysis process at the theory level, along with the domain concepts. These elements are represented at the top of the hierarchy in Fig. 16.9. For this analysis, the top levels of the hierarchy, the paradigm and ontology were not expanded as they were deemed to be out of scope of the current analysis.

⁷The R&M data consists of names of particular LRUs diagnosed as “faulty” and dates they were removed from the aircraft.

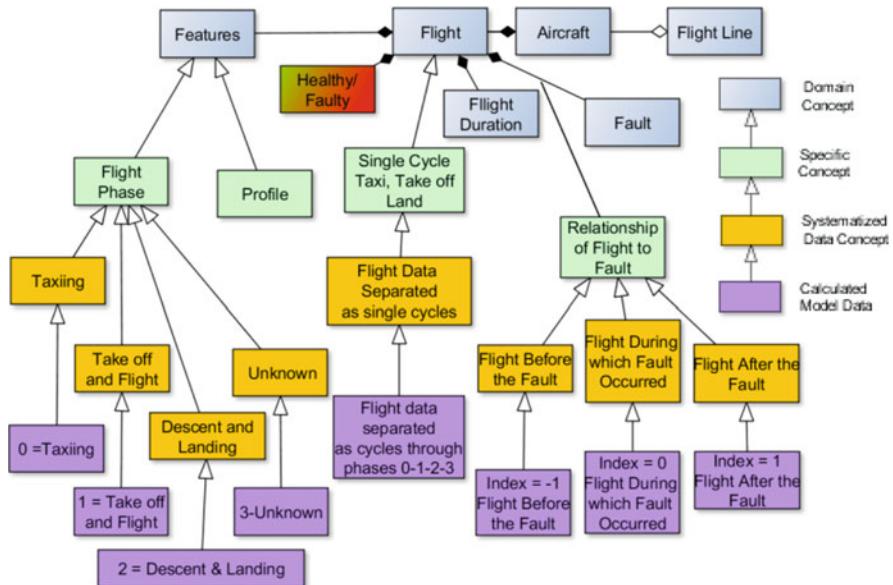


Fig. 16.9 An epistemological hierarchy of the aircraft prognostics use case beginning with domain concepts and operationalizing down to calculated model data (see legend for color definition)

16.4.1.2 Theory

An aircraft exhibits a relative state of health that changes over time. In general, the state of health of various components degrades at different rates over time as the aircraft is operated. The state of health can be improved by replacing degraded or faulty components. Leading indicators of failure may reduce unplanned downtime of the aircraft, increasing aircraft availability and possibly reducing costs.

By design, some aircraft may have specific sensors for specific LRUs to indicate when they have failed. In prognostics, the analytical interest lies in anomalous behavior in the sensor reading for a specific LRU or readings from other sensors prior to the specific LRU's failure, as opposed to sensor readings after the component has failed. The longer the lead time between the detection of the indicator and the occurrence of the failure and the more precise the indicator, the more valuable the indicator is. Once an indication of impending failure occurs, decision makers can determine when to replace the component. Prognostic information such as these indicators used in a decision support system may reduce unplanned down time of the aircraft.

This analysis would be trivial if all LRUs were instrumented with sensors that detected their degradation. However, this is not physically possible and was not the case in the aircraft that gave rise to this Use Case.

16.4.1.3 Social Model

The social model of the aircraft is the representation of the real-world system behaviors based on theories. The social model can be expressed as follows:

An aircraft exhibits a health state that consists of characteristics of data from its sensors over some period of time and the history of its Repair and Maintenance Data. These data may be indicative of both the operation of the aircraft and the response of the aircraft as it is operated.

The second statement of the social model is an important qualification. If the data only provided information about how the aircraft was being operated (altitude, speed, pitch, roll, etc.), then there could be no expectation that analysis of the data would provide any health indication of the aircraft systems. On the other hand, if the data are only providing response information (vibration, component temperature, air flow), it is possible that health indicators could be found, but they will be noisy due to different operational modes of the aircraft impacting the response information. Consequently, with both response and operational sensors, we have the best of both worlds.

These observations led to the refinement of the domain concept (see Fig. 16.9) to include specific concepts related not only to the sensor data features that had already been considered but also to features related to how the aircraft was being operated, including flight phase and mission.⁸

16.4.1.4 Hypothesis

The importance of a correct hypothesis cannot be overstressed. When formulating the hypothesis, it is often useful to revisit the purpose. The hypothesis needs to directly address the purpose. The relationship and the direction of the relationship being measured by the hypothesis are critical to achieving the purpose of the study. The hypothesis for this use case can be expressed as follows:

As an LRU degrades, the degradation will be detectable through analysis of the sensor data.⁹ Furthermore, we will be able to develop thresholds related to the degradation to classify flights as “healthy” or “faulty.”

Although the health of the aircraft is likely on a continuum, the hypothesis brings into focus the desire to identify healthy and faulty data in a binary sense. If unable to successfully differentiate the data from healthy and faulty flights, in general, it

⁸The Mission is defined as specific characteristics of how the aircraft is being flown. In this case, the Mission was induced from the data. Eventually, the Data Analytic Process to produce the Mission will require its own pair of hierarchies in order to be properly characterized, verified and validated.

⁹Indicators may also be found in the R&M data. For instance, sequences or co-occurrences of LRU removals may be used to predict faults. Analysis for this DAP is not included in this use case.

is clear that a prognostic cannot be developed. At this stage, it was recognized that the relationship of the flights to the faults was important for machine learning algorithms.

16.4.1.5 Application

At this point in the study, a team of four data scientists set out on four separate paths to develop prognostic indicators. The thought was that even if the predictors that were developed were somewhat weak, they could be combined using an ensemble of methods resulting in a usable predictor that would provide prognostic benefits in an operational setting. However, if these predictors were going to be compared or combined, there had to be agreement as to what data should be used for the analysis. Such an agreement turned out to be non-trivial and was addressed at the application level of analysis.

The application level is used to instantiate the hypothesis. An LRU failure can be detected by applying both supervised and unsupervised learning techniques to the sensor data. For each of the four methods that were explored at this step, each data scientist identified which supervised or unsupervised method of anomaly detection he or she planned to use.

On the data model side, the *Systematized Data Concepts* that correspond to the application analysis model are shown in orange in Fig. 16.9. At this point, the descriptions of the data are still text descriptions, but they are very specific concerning which data is to be used. This is an important step, as it will signify which approaches can be reasonably compared. If the approaches are combined using an ensemble method, differences in the selected data should be reflected in differences in the ability of each model to predict under different circumstances.

To understand the development of the systematized data concepts, it is helpful to think about aircraft behavior. Aircraft don't just fly. They also undergo engine tests on the ground, during which time the engine is turned on and off, but the aircraft doesn't fly. Flight phase arose from our social model. Flight phase was more specifically defined to be taxiing, take off and flight, descent and landing, and unknown.

Similarly, aircraft fly different missions as defined by a series of "Flight Phases". For example, they might fly *touch and goes* (short loops around the airport to practice landing and taking off), or they might fly at tens of thousands of feet for many hours.

For each fault, we identified the flight before the fault, the flight during which the fault most likely occurred and the flight after the fault. For dates where there were multiple flights, it was not clear exactly when the fault occurred. In these cases, it was assumed that the fault occurred on the last flight of the day. From an ontological standpoint, these nodes are connected to the nodes that are connected to the "Relationship of Flight to Fault," reflecting a specific combination of fault and flight. The combination of fault and flight is used for training machine learning algorithms.

16.4.1.6 Implementation

The implementation model is the point at which the decision is made regarding the specific algorithms to build and the data to use. At this point, the final tweaks on the data cleaning are performed, and the algorithms are coded.

For example, to label the data for machine learning, we defined the flight before the day on which the fault occurred to be Index = -1, the flight on the day of the fault to be Index = 0, and the flight after the fault was fixed to be “healthy.” The exact number of flights before and after that is considered healthy by the different data scientists vary but can be clearly specified by each investigator through updates to the data model. “Clean” data was broadly defined as data that was at least five flights away from a specific fault for a specific aircraft. “Super Clean” data was defined as data that was at least five flights away from all faults for a specific aircraft. These refinements aren’t shown on the ontology in the interest of readability.

Figure 16.9 shows a portion of the ontology derived after multiple iterations through the refinement process of “Domain Concept → Specific Concept → Systematized Data Concept → Calculated Model Data”.

16.4.1.7 Verification

Verification of a model (i.e., “Was it built, or ‘operationalized’, correctly?”) is achieved by looking down the hierarchy, while validation (Was the right thing built, i.e., did it accomplish the purpose?) is achieved by looking up the hierarchy. Verification is described here in the step between the four different flight phases and the labeling of the implementation data. We had been given code that was used to determine whether an aircraft was taxiing, taking off, landing, etc. Making this determination turned out to be very important. When we first defined the flight phases and extracted the data according to the flight phases from the files that were taken from the aircraft, we were surprised to see multiple flights in one file (multiple sequences of Flight Phases). If we had not identified flight phases as important to the study, we would have been inadvertently performing machine learning algorithms on mixtures of individual and multiple flights. We therefore concluded that our operationalization of flight phases was verified.

On the other hand, since we were handed the code for the flight phase labeling, we had not verified that the code worked correctly and subsequently discovered that it had some deficiencies. In particular, we found that it was overzealous in separating flights and misinterpreted short durations of weight on wheel bounces (part of the algorithm) as meaning a new flight was starting. We ended up with hundreds of unrealistically short flights, indicating that we weren’t properly separating out flights. Once this was fixed, and we could verify that it was separating out flights correctly, we looked up the hierarchy to make sure that our flight phase enumeration was useful toward assisting in developing machine learning algorithms. For example, did we need to figure out what state(s) “Unknown” really was? Over time, we have gained confidence in this enumeration and to date have not identified any reason to change it.

16.4.2 Climate Data Analytics Use Case

The physical system of earth is complex. Despite advances in technology, physical scientists (i.e., climate, ocean, hydrology, etc.) continue to face uncertainties resolving components of the earth system within large-scale numerical models. Nevertheless, progress has been made, and forecast products have improved significantly from micro-scale to planetary-scale predictions. A function of such improvement lies within data acquisition and data integrity, upon which the models build an abstraction to predict and forecast events. All models have this dependency regardless of domain; however, the environment for which physical data is being retrieved is rapidly changing and negatively impacting forecasting accuracy of physical models.

Historical meteorology data sets, which generally range from 40 to 80 years (statistically significant values over time), are utilized in models to build components of the earth system. When the model is executed, 4–6 years (called “spin up time”) is allowed to build the world, and predictions can then be made for the length of time required by the experiment. Scientists make their first assumption that the data sets are valid enough to satisfy a stochastic, non-deterministic representation of the physical system. In other words, the scalar values and their temporal and geospatial attributes are sufficient to define an initially stable abstraction of the world from which to generate predictions. This assumption is challenged, given that *climate change* is rapidly convoluting earth’s thermodynamic and chemo-energetic environment. Within the past few years, models across the globe have struggled to predict tropical cyclone (or hurricane) intensity and movement. In October 2016, Hurricane Matthew rapidly intensified into a category 5 storm, despite the high windshear environment (15–20 knots) and despite sea surface temperatures above average during the previous several years *without* a category 5 storm. Neither NOAA’s Hurricane Weather Research Forecast (HWRF) model, the European Centre for Medium-Range Weather Forecast (ECMWF) model, nor the Geophysical Fluid Dynamic Laboratory (GFDL) model captured this behavior. In 2017, a similar rapid intensification event occurred with Hurricane Harvey – even with the storms western, and northwestern quadrants over Texas, it intensified from category 3 to 4, defeating all models. Soon after, Hurricane Irma defied the textbooks and shocked everyone by sustaining category 5 strength over 2 continuous days in the Atlantic Ocean. Again, no models predicted this behavior. These storms are just a few examples (Hurricane Patricia which occurred in 2015, and the western Pacific surge in Super Typhoons since 2012 [i.e., Haiyan, Nepartek] for example) of the on-going issues the scientific community is facing.

Models have a dependency on data, and large-scale models *statistically require* large-scale data sets spanning multiple decades to ensure sufficient information to build an abstraction of the physical world. Data from 3 to 7 years is incredibly small by comparison and would produce erroneous results. However, significant features within the climate system, such as increased atmospheric thermal profile, dramatically increased atmospheric carbon dioxide concentrations (and associated

thermal and chemical feedback), increased average water vapor concentration over time (due to warming on average), and global increases in temperatures never measured before by humankind *compared* to the previous 40–80 years have impacted the earth system in ways models *do not know how to represent*. Yet, in order to ensure our predictions are feasible, we must capture this behavior.

Applying an additional assumption that “tweaking” a parameter by which you see today versus what you *do not* see in the past and running predictions from that is simply *bad science*. This practice gives rise to even more questions. How long should you “tweak” that parameter? On what regions of the planet? Will those changes correlate with subsequent parameters (e.g., increased air temperature forces oceans to absorb heat, causing oceans to thermally expand, changing oceanic circulations, etc.). More bad science arises with the assumption that increased data resolution (either temporally or geospatially) will dramatically remove levels of uncertainty. This actually leads to *more* uncertainties between scales, let alone physics (e.g., between quantum and relativity) and limiting the *feasibility* of producing a *reliable* prediction at all.

A more systematic approach to resolve a solution would be to implement a model to identify pitfalls in our assumptions of our data to fit our models, such as the epistemological model (EM) (Fig. 16.5). Let’s be clear: *verifying* physical data is redundant. Wind is wind, humidity is humidity; the instruments to collect physical data are designed and calibrated to gather specific physical parameters. Validation is the real issue. The EM is flexible, meaning regardless of the domain or stage of the DAP, the researcher can apply certain categories of the EM to suit his or her purpose.

As shown in the instantiation presented in Fig. 16.10, the aggregated physical data is cleaned and organized into specific applications. Dark blue boxes represent a larger-scale and longer-term modeling purpose, while light blue represents a

Epistemological Analysis Model Hierarchy		Example Data Analysis Process		Epistemological Data Model Hierarchy
PURPOSE				
Application	Long-term predictions	Short-term predictions	Systemized Data Concept	
Implementation	Large-scale models	Medium-scale models	Calculated Model Data	
Data	Multi-Decadal Data	Sub-decadal Data	Model Data	

Fig. 16.10 An instantiation of the EM for decomposing the climate data analytics problem space based on temporal (e.g., long- vs. short-term predictions) and geospatial (large vs. medium scale) dimensions

medium-scale and shorter-term modeling purpose. Indeed, all physical data can be utilized for any range of forecasting; however, the scale of the data set can either benefit or harm the analyses for the specific model for which the data is to be applied. In this way, the EM allows the researcher to outline the purpose of his or her data, for both the analysis model and data model. And though the data have been categorized to suit certain modelling purposes, all data can functionally represent the physical world, just on different time scales. This provides insight: although our multi-decadal data sets do not capture recent features from climate change, that data is still needed. Also, the models by which multi-data sets are implemented are *not obsolete*. Sub-decadal data sets are equally representative for physical systems, providing higher resolution to medium-scale models under feasible conditions compared to large-scale models. The EM helps us infer that in order to capture sub-decadal features, we need to implement a method that allows us to utilize as *much data as possible* to gauge uncertainties and determine data that is or isn't useful.

To address this issue, Abdullah, Reddy, Butler, and Walters [1] implement a Bayesian Belief Network (or Bayesian Neural Network, BNN) as a surrogate model to ensemble models. The principle goal is to allow the BNN to “check” whether the model-generated data to predict a hurricane is taking into consideration features in the ocean-atmosphere interface (OAI) to produce the storm. The BNN can feasibly provide probabilistic calculations to inform the researcher, among the generated data *what data is useful and what is not*. In other words, it's a *validation-optimization model*.

Through a structured analysis model, like the EM, it's conceivable to deduce where problems in data verification and validation arise. More importantly, the EM can provide direction behind initial assumptions and garner different methodologies to solve the researcher's problem.

16.5 Conclusions

All knowledge generation processes must systematically and consistently question and evaluate the constructs used of interpreting and understanding the world. Otherwise, the knowledge produced based on these unverified and un-validated underlying constructs is at risk of being meaningless or wrong. A Data Analysis Process (DAP) does not explicitly support validation and verification. The process of data analysis does not necessarily break in the absence criteria such as valid constructs or consistent assumption. As a consequence, a DAP alone does not guarantee results that are verified and valid and is susceptible to vulnerabilities such as (1) a lack of accountability of inconsistent assumptions, (2) a lack of falsifiability, and (3) increased misrepresentation of the data.

V&V for a data analysis process (DAP) assesses how correctly a DAP is capturing and producing knowledge about a phenomenon of interest. In order to assess the V&V of a DAP, one needs to define a strategy that supports the

evaluation of a DAP with respect to measuring how well the DAP (1) does what it is meant to do by its designer and (2) is consistent with reality. This chapter presented a model for a data analysis process that supports verification and validation. This model aims to capture an epistemology of data being leveraged or created by a DAP and the analysis being composed by a DAP. At the heart of the strategy presented here is a move to go beyond viewing verification and validation as solely empirical testing. Our approach is to provide mechanisms and techniques for evaluating the data constructs and analytic composition of a DAP by considering their epistemological underpinnings. Exposing and documenting these epistemologies enable the verification and validation of a DAP and ultimately generate confidence and trust in a DAP and its products.

The approach to V&V of a DAP via its epistemological decomposition is to assess each epistemological level of the decomposed DAP model as an individual module. The V&V evaluation of a model will result from an aggregation of the V&V assessment of each epistemological level. The dimensions of verification and validation presented here in support of modular evaluation of a data analysis process are consistency of concepts, relationships, and assumptions, as well as utility, external consistency, prominence, and accuracy. The dimensions are useful and relevant to understanding the provenance of an analytic process with respect to its origins as well as its applications.

The evaluation of V&V for a DAP is based on user-defined scores to questions about the epistemological decomposition of a data analysis process. As a particular DAP becomes more prominent or useful in the community, more V&V evaluations will be made of the DAP. These evaluations will either (1) provide awareness of flaws and feedback on issues to support the improvement of the DAP or (2) reinforce validity and verification of the DAP thereby encouraging further use and extension of the DAP.

References

1. W. Abdullah, R. Reddy, C. Butler, W. Walters, Utilizing Bayesian belief networks to model the ocean-atmosphere interface. *J Miss Acad. Sci* **63**(1), 121–122 (2018)
2. R. Adcock, D. Collier, Measurement validity: A shared standard for qualitative and quantitative research. *Am. Polit. Sci. Rev.* **95**(03), 529–546 (2001). <https://doi.org/10.2307/3118231>
3. A. Bekker, *4 types of Data Analytics to Improve Decision-Making* (Science Soft, 2017), [Online]. Available: <https://www.scnsoft.com/blog/4-types-of-data-analytics>. Accessed 3 Dec 2018
4. F.C. Copleston, *A History of Philosophy* (Image Books, Garden City, 1964)
5. M. Cronbach, P. Meehl, Construct validity in psychological tests. *Psychol. Bull.* **52**(4), 281–302 (1955)
6. J.D. Fearon, D.D. Laitin, *Ordinary Language and External Validity: Specifying Concepts in the Study of Ethnicity**. LiCEP Meetings (2000), Retrieved from <https://web.stanford.edu/group/fearon-research/cgi-bin/wordpress/wp-content/uploads/2013/10/Ordinary-Language-and-External-Validity-Specifying-Concepts-in-the-Study-of-Ethnicity.pdf>

7. T. Harford, Big data: A big mistake? *Significance* **11**(5), 14–19 (2014). <https://doi.org/10.1111/j.1740-9713.2014.00778.x>
8. I. Kant, *Critique of Pure Reason*, 1. paperback ed., 15. print (Cambridge Univ. Press, Cambridge [u.a.], 2009)
9. A. Kaplan, *The Conduct of Inquiry* (Chandler, San Francisco, 1964)
10. A. Kaplan, *The conduct of inquiry* (Transaction Publishers, 1973). Retrieved from https://books.google.com/books?id=ks8wuZHSKs8C&pg=PA53&lpg=PA53&dq=Abraham+Kaplan%27s+paradox&source=bl&ots=bHV9ptpV3g&sig=8_k3iRGHtuBuIovAcZSGqLwTTYo&hl=en&sa=X&ved=0ahUKEwjzvrDS777YAhVDRN8KHaxIBA4Q6AEISTAI#v=onepage&q=Abraham+Kaplan's+paradox&f=false
11. C. Kufs, The five pursuits you meet in statistics. (Stats With Cats Blog, 2010), Retrieved May 10, 2018, from <https://statswithcats.wordpress.com/2010/08/22/the-five-pursuits-you-meet-in-statistics/>
12. D. Lazer, R. Kennedy, What we can learn from the epic failure of google flu trends. (WIRED, 2015), Retrieved May 10, 2018, from <https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/>
13. I.S. Lustick, M.R. Tubin, Verification as a form of validation: Deepening theory to broaden application of DOD protocols to the social sciences, in *Proceedings of the 4th International Conference on Applied Human Factors and Ergonomics*, (San Francisco, 2012). Retrieved from <http://lustickconsulting.com/data/Verification as a Form of Validation - Lustick, Tubin.pdf>
14. J. Overton, *Going Pro in Data Science* (O'Reilly Media, Inc, 2012). Retrieved from https://www.oreilly.com/data/free/files/going-pro-in-data-science.pdf?mkt_tok=eyJpIjoiWW1GbU1XSmhNRGMwTkRVdylsInQiOijGNIRsSFZnZExYXC9wR0ZOZWZOaWZ1ZHFUzjBFM1RhblFJSHM4VmpibW5udVwvY2FLRVVKVFDsQzlCNnV6ZEQ3Nkl3VEg3c09idhZWU5YNEVCTllySjM0eCtNRGJnQnpsR1Q0QTFaU
15. J.A. Paulos, *Metric Mania* (The New York Times, 2010)
16. A. Ruvinsky, J. Wedgwood, J. Welsh, Establishing bounds of responsible operational use of social science models via innovations in verification and validation, in *2nd International Conference on Cross-Cultural Decision Making*, 2012
17. F. Sailer, Google Flu Trends is dead – long live Google Trends? (UCL Research Department of Primary Care and Population Health Blog, 2018), Retrieved August 23, 2018, from <http://blogs.ucl.ac.uk/pcph-blog/2018/01/23/google-flu-trends-is-dead-long-live-google-trends/>
18. J.D. Stemwedel, Basic concepts: Falsifiable claims. – Adventures in ethics and science. Retrieved August 24, 2018, from <http://scienceblogs.com/ethicsandscience/2007/01/31/basic-concepts-falsifiable/>
19. A.G. Stephenson, D.R. Mulville, F.H. Bauer, G.A. Dukeman, P. Norvig, L.S. LaPiana, R. Sackheim, *Mars Climate Orbiter Mishap Investigation Board Phase I Report* (1999), Retrieved from http://sunnyday.mit.edu/accidents/MCO_report.pdf
20. K. Vasileva, Common mistakes in data analysis – The Data Nudge – Medium. (2017), Retrieved May 10, 2018, from <https://medium.com/the-data-nudge/common-mistakes-in-data-analysis-951e366084b9>
21. VV&A Recommended Practices Guide, (2011). <https://vva.msco.mil/Key/key-pr.pdf>
22. Wikipedia_contributors, Data analysis. (2018), Retrieved October 5, 2018, from https://en.wikipedia.org/w/index.php?title=Data_analysis&oldid=838877371
23. N. Yau, *Why Context is as Important as the Data Itself* (2010)

Chapter 17

Data and Information Quality in Remote Sensing



John Puentes, Laurent Lecornu, and Basel Solaiman

Abstract Remote sensing datasets are characterized by multiple types of imperfections that alter extracted information and taken decisions to a variable degree depending on data acquisition conditions, processing, and final product requirements. Therefore, regardless of the sensors, type of data, extracted information, and complementary algorithms, the quality assessment question is a pervading and particularly complex one. This chapter summarizes relevant quality assessment approaches that have been proposed for data acquisition, information extraction, and data and information fusion, of the remote sensing acquisition-decision process. The case of quality evaluation for geographic information systems, which make use of remote sensing products, is also described. Aspects of a comprehensive quality model for remote sensing and problems that remain to be addressed offer a perspective of possible evolutions in the field.

Keywords Remote sensing · Acquisition-decision process · Geographic information systems · Data-information fusion quality

17.1 Introduction

Earth monitoring for analyzing its past and current state as well as longitudinal evolution is a complex endeavor. Such monitoring utilizes technologies to measure remotely physical variables of a given area, without being in direct contact with the observed elements. Known as remote sensing, these technologies rely basically on the principle of sensing specific types of radiations emitted and/or reflected by the

J. Puentes (✉) · L. Lecornu

IMT Atlantique, Lab-STICC, Technopole Brest Iroise – CS 83818, Brest, France
e-mail: John.Puentes@imt-atlantique.fr; Laurent.Lecornu@imt-atlantique.fr

B. Solaiman

IMT Atlantique, Technopole Brest Iroise – CS 83818, Brest, France
e-mail: Basel.Solaiman@imt-atlantique.fr

studied zones in order to detect, identify, classify, and map the objects and surfaces of interest. Sensors range from radars, radiometers, sounders, and spectrometers in satellites to optical spaceborne or airborne cameras, sonars on ships, Light Imaging, Detection, and Ranging (LIDAR) on moving vehicles or aircraft, and more recently sensors on demand on autonomous and remotely operated devices like aerial, terrestrial, and underwater drones. A large and increasing variety of applications is therefore accessible for numerous distance ranges, spatial-temporal resolutions, and radiations. Some examples of remote sensing applications associated with their respective sensor are¹:

- Radar – measures of aerosols, rain, snow, soil moisture, altitude, sea surface temperature, weather prediction, earth deformation, and images of land and sea extent
- Radiometer – measures of sea ice extent and thickness, land and vegetation cover mapping, sea level change, greenhouse gases, near-surface wind speed, water resource monitoring, and wildfire detection
- Sounder – simultaneous 3D measures and profiles of air temperature, pressure, and moisture, atmospheric composition, ozone mapping, and profiles of gases affecting ozone chemistry
- Spectrometer – measures of solar irradiance, solar energy, solar wind condition, atmospheric carbon dioxide from space, images of clouds, vegetation classification, mineral mapping, and ocean color
- Optical sensing – measures of soil types, plants' reflectance, global ocean properties, change detection, definition of surface and terrain models, identification of geographical hazards, urbanization, forestry, and farming follow-up
- Sonar – measures of bathymetry; characterization of bottom type and composition; seafloor mapping and imaging; detection of underwater sounds, mines, and shipwrecks, and localization of fish
- LIDAR – measures of carbon cycle, vertical profiles of clouds and aerosols, structure of clouds, spatial and temporal structure of forests, characterization of plankton properties, creation of urban elevation maps, and monitoring of fire areas

Studies of these geophysical phenomena through time produce very large, periodic or irregular heterogeneous data streams [1]. As a consequence of the constant increase in sensors resolution and performance, an accelerated progression of remote sensing data production has been taking place. Sensor data streams have reached several GB per second [2], while physical models search to integrate part of that data making use of enhanced spatial and temporal grids [3]. The fusion of observations from multiple sensors applying improved dynamic models is leading thus to the so-called big data assimilation [4]. Regardless of the sensors, kind of data, information extracted from it, and processing algorithms, the quality assessment question is a pervading one. In spite of the trend toward huge data volumes, data

¹<https://earthdata.nasa.gov/user-resources/remote-sensors>.

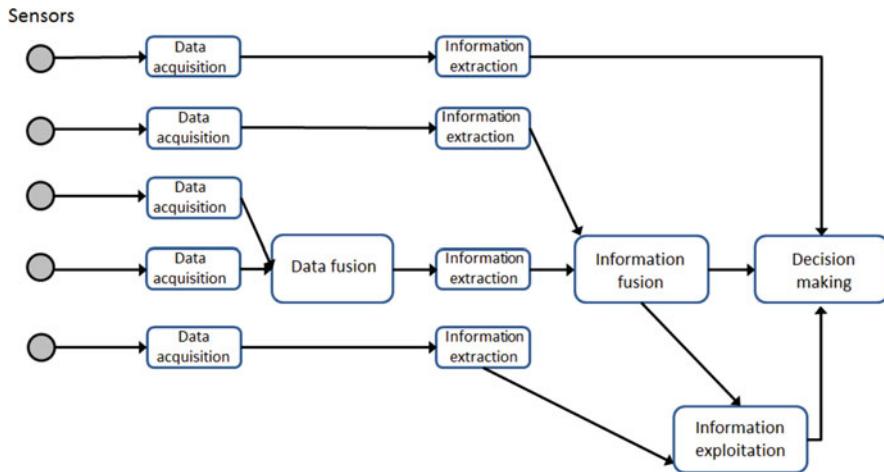


Fig. 17.1 Quality assessment in a remote sensing process from data acquisition to decision-making. All the stages – data acquisition, data fusion, information extraction, information fusion, information exploitation, and decision-making – require quality evaluation

gaps limit considerably the possibilities of information extraction [5, 6]. Moreover, sensors may produce noisy, incomplete, approximate, and untrustworthy data, which can also be redundant and/or contradictory, depending on sensor characteristics, operational conditions, and uncontrollable external factors. Information extracted from poor- or low-quality data will be as a result also of unacceptable quality but, worst of all, might induce wrong costly interpretations and decisions.

Similar to any fusion system, in remote sensing systems, quality assessment can be considered as a problem to be solved at each one of the different phases that are part of any acquisition-decision process (Fig. 17.1). Globally, it is essential to evaluate the quality of:

- Data acquired by one or multiple sensors and, if required, fused data
- Extracted and fused information
- Exploitation of extracted information
- Making decisions

Consequently, quality assessment in remote sensing systems is a complex problem, mainly because of four reasons:

- Data quality evaluation must be adapted to each kind of sensors.
- Awareness of information quality is highly contrasted across use cases.
- Process and system architectures differ from one application to another.
- Whenever the context of a given process changes, it is necessary to reformulate the quality analysis strategy.

Quality assessment in remote sensing is necessary to cope with the variable features of heterogeneous increasing data volumes, acquired by multiple types of

sensors, operating under different conditions and contexts for diverse applications. Conceiving quality analysis and evaluation methods for remote sensing requires therefore fundamental understanding of quality components and their interaction through the acquisition-decision workflow. Namely, besides individual quality measures at different workflow stages, it is also required to understand the impact of data quality on information and decision quality, as well as embedded fusion processes and application of extracted information. This chapter examines those elements by providing an overview of initiatives and required efforts in the domain. Examples of elaborated quality approaches for remote sensing are examined in Sect. 17.2. The particular case of quality assessment for geographic information systems, which make use of remote sensing products and complete the acquisition-decision-making process, is described in Sect. 17.3. Essential aspects of a perspective comprehensive quality model for remote sensing are described in Sect. 17.4. Conclusions are summarized in Sect. 17.5.

17.2 Quality Assessment Approaches

Perfect remote sensing acquisition conditions are very rare, making quite common the generation of datasets of extracted information with multiple imperfections. Additionally, when data acquisition modalities are fused to improve information extraction, imperfections are likely to have a considerable impact on final results. However, despite multiple initiatives, the absence of reliable generalized quality evaluation approaches impedes to assess systematically remote sensing data and information quality [7]. Since an automatic quality control is a complex task with variable types of outcomes, part of quality evaluation has been carried out manually by experts who provide arbitrary global qualitative estimations. On the other hand, several initiatives have been searched to solve the quality evaluation problem for particular use case characteristics, application requirements, or user preferences. This section recapitulates some of the main ideas about quality assessment approaches in remote sensing concerning data from sensors, extracted information, and fusion.

17.2.1 Data Quality

In an ideal process, the validity of remotely collected data should be verified consistently, i.e., by comparing measured and reference values. Given the covered surfaces, required technical infrastructure, and associated costs, exhaustive earth monitoring ground truth necessary to make preliminary validations of data quality is rare [1, 8, 9]. Yet, one domain, in which reference values can be frequently obtained, is sea surface temperature (SST) measurement [10]. For instance, direct measurements can be obtained by means of hull contact sensors or buckets on ships, as well as fixed or drifting buoys. Despite the sensors' appropriate performance,

these measures depend strongly on correct calibration, geographic deployment zone, and functioning conditions that make it necessary to apply various quality tests to verify infrared SST satellite values. In situ buoy data are checked for self-consistency and cross-consistency with other data to calibrate SST satellite observations [11–13]. Verifications include among others: sensors drift, geolocation, time stamps, spikes, reference background, and cross-platform validation. Matchup databases result from the combination of in situ and satellite SST observations, permitting to validate satellite SST data [14] or to verify the pertinence of in situ measurements [15]. Furthermore, cloud contamination and atmospheric path length corrections were adjusted applying background field checks, i.e., compared with respect to the known variances of local and regional errors, assuming that these are normally distributed. Quality tests were also applied to detect aerosol contamination (smoke, spray, or dust particles that attenuate signals). Regional aerosol transport models [16] provided related knowledge depending on instrument wavelengths, to identify different levels of aerosol contamination. The reduction of erroneous cloud detection was examined as a way to improve satellite SST measurements accuracy by means of a Bayesian approach [17]. Cloud detection estimations were compared to compiled quality flags of buoy measurements [18], to improve the quality of satellite SST data. Nevertheless, cold water makes it very complex to identify clouds in SST images [19]. Finally, the accuracy of satellite SST values was examined under cloud coverage and compared to compiled in situ buoy measurement [20].

SST values are then corrected or rejected according to a transfer model that considers the vertical distribution of temperature and the aerosol height and estimated dispersion. Errors of acquired data were compared to background error limits to decide if SST observations were consistent depending on seasonality [10]. Internal error verifications of measurements were also carried out with respect to a complete observations dataset [21]. It is important to note that, since data quality within a forecasting system is analyzed in a sequential manner from acquisition to prediction, unnoticed errors at the initial and intermediate levels are seldom detected later at the decision level.

Beyond these particular examples of sensors' data quality assessment, basic raw data quality controls in oceanography are commonly of five types [22]: prescreening, plausibility, internal consistency, external consistency, and mutual consistency. These controls are part of larger quality control procedures that estimate data coherence across a system. At a practical level, four levels of data quality control have been implemented from the sensor to the prediction levels [10]. With the objective of minimizing wrong decisions, the four control levels search to: remove obvious errors and anomalies; decide to accept, reject, or manually process remaining data; carry out structured tests to detect unreliable data; and characterize data integrity with respect to obtained analyses and forecasts. In other disciplines, sensor accuracy was verified according to fixed control points on the scanned surface using variable acquisition parameters [23]; given different frequency bands and noisy acquisition conditions [24]; according to accuracy and precision [1]; or depending on multiple resolutions – spatial, radiometric, spectral, and temporal – along with focal length, point spread function, and modulation transfer function [8].

17.2.2 *Information Quality*

Depending on the use case, quality assessment of extracted information can prevail over data quality evaluation. Since remote sensing datasets are much likely to contain various types of imperfections, extracted information is consistently altered to a variable degree, depending on data acquisition constraints and final product requirements. Considering the difficulties of endorsing detailed sensor data quality verification, the quality of extracted information is frequently evaluated instead. Although quantitative information quality analysis should be applied to reproduce results, subjective evaluation is also a somewhat common practice, alone or combined with other approaches. To compare raw and processed images, it can take the form of numeric marks as 1 (excellent) to 5 (very poor) or descriptive labels. Visual quality segmentation evaluated by an expert as poor, medium, or good was compared to associated geometric information – area, perimeter, and shape – to determine the impact of segmentation quality on the classification of remote sensing images [25]. Two subjective quality estimation levels were defined to configure product settings and to give a global quality evaluation [26]. Product settings were configured depending on the quality labels: very low, marginal, questionable, average, intermediate, acceptable, high, or perfect. And global quality was estimated to be as follows: not produced, acceptable, or good. Visualization for quality evaluation of processed images of terrain models was carried out to qualitatively identify the type of dataset that permits to recognize indicators of landslide activity, according to distances between points and density of points [27].

Another reported information of quality evaluation technique for remote sensing is based on per-pixel analysis, which focuses on the separate evaluation of information conveyed by each acquired value to be used with other estimations. For instance, to assess the quality of object detection and classification, geometric features like detected objects location, size, and spatial extent, represented in terms of verified groups of pixels, were used to validate classification correctness and completeness [28]. Global quality estimation was obtained combining these and user measurements to be compared to those of reference data with proper quality. A frequent issue in remote sensing images is the variable quality due to cloud coverage, snow, shadows, and any other unwanted surface representation that hides searched information. To address this issue, some approaches represent data gaps for detecting invalid pixels and possible interpolation solutions. Data gaps were studied determining the feasibility of interpolations by applying spatial and temporal variations per region and season to characterize different quality settings [26]. Interpolation viability depended on the number of invalid pixels, gap length, and corresponding data quality setting. Qualitative labels – passed, suspect, and failed – were employed besides quantitative estimations, to define the pixel quality of images having different scan geometry and land area with variable cloud contamination [29]. This approach permitted to determine the retrieval quality of land surface albedo. Per-pixel quality estimations were also applied to analyze cloud denseness and shadows on images, in order to decide when it was possible to use suboptimal images [30].

Sometimes when two or more sensors at different times or view angles are used to scan a region, image registration is necessary as a preprocessing step to align or resample the images before segmentation, classification, or fusion can be applied. One of the examined quality-related problems has been to determine if a given quality indicator can permit to identify which is likely to be the most appropriate registration transformation. To this end, using ground control points and statistical and geometrical indicators, quality assessment of image registration suggested the appropriate transformation for the registration of image to image spaceborne images, image to map airborne images, and image to map spaceborne images [31]. Quality evaluation was defined as tests to detect possible deviations from the null hypothesis but letting the user select a transformation depending on the results of interpretation. Since registration may be incorrect because of radiometric anomalies, data gaps, the presence of mosaics in a dataset, and cloud coverage, to reduce such risk a quality monitoring of monthly image composites was defined [32].

Image segmentation is a significant product of remote sensing campaigns that also needs to be examined by quality analysis. The main principle of quality evaluation in this case is to use known pre-defined geometric objects as reference to validate the relative pertinence of segmentation. However, multiple geometric features are necessary, since a single one cannot account for most of quality aspects [33]. In applications like mapping, geomorphological representations of landforms comprising shape features (length, width, and elongation ratio), orientation, and position accuracy were compared to evaluate the quality of segmented remotely sensed images [34]. An equivalent approach but without utilizing the ground truth was designed to examine the quality of reconstructed 3D roofs from dense airborne laser scanner [35]. The main goal was to generate a graphic representation of quality levels symbolized with colors over reconstructed structures for evaluation by users. It included accuracy, completeness, and correctness, geometric quality elements to quantify the orientation and boundaries of planes, the characteristics of isolated contours, and the coherence of closed representations and distances.

In multiple applications, segmentation is followed by classification – pixel-based (supervised and unsupervised) or object-based – to define land cover and land use patterns, mainly used by resource management, landscape planning, and longitudinal environmental follow-up. Besides the dependency on reference information (in the supervised and object-based cases), classification also relies strongly on the pertinence of segmentation, making necessary to implicitly evaluate both. Nonetheless, it is important to note that a product quality does not only depend on high classification accuracy [36]. An early approach defined the quality of image classification in land cover maps as a relation between the accuracy, the classification purpose, and the cost of wrong classification [37]. This work excluded user subjectivity in the evaluation. Quantitative, area, perimeter, and shape, and qualitative, visual qualification labels, criteria were defined to estimate the impact of segmentation quality on classification quality by using the comparison of reference surfaces as the reference [25]. Also, the quality of forest classification was related to the quality of segmentation and was defined by the differences of overlap, positions, and distances between segmented and reference objects [38]. Other approaches

studied additional methods of information quality evaluation, for example, a method measuring the effect of different signal-to-noise ratios and their corresponding impact on visual perception [39, 40], the agreement of displacement directions, magnitudes and gradients with user visual perception [41], and the pertinence of context in anomaly detection [42].

17.2.3 *Fusion and Quality*

Multivariate data and information are present in a large extent in remote sensing processes, mainly because some sensors alone cannot provide all the necessary information and also to take advantage of the possibilities offered by combined sensing schemes. Complementary scene representations are generated by making use of two or more sensors, at the same or different times, or the same sensor repeatedly to improve the results by enlarging combined parameters like spectral and spatial resolution. Fusion is then required to obtain enhanced data and information for segmentation, classification, and decision support. Several works have examined fusion quality as an associated question rather than a central problem. For example, it was studied to know if the quality of spectral and spatial fused images could be evaluated by the blur on different bands [43] and to estimate the variability of fusion results measuring quality by standard statistics indicators without separating signal and noise [44]. The theoretical principles of quality based on accuracy [45] and how common quality factors like variable noise, averaging, changeable signal-to-noise ratio, and blurring can be modeled [46] were also analyzed.

Otherwise, fused image pertinence has been evaluated depending on the availability of reference images. When reference images were available, fused image quality was measured by comparative metrics as the root mean square error, relative global error, mean bias, percentage fit error, signal-to-noise ratio, peak signal-to-noise ratio, correlation coefficient, mutual information, and structural similarity index measure. In the absence of reference images, such metrics as standard deviation, entropy, cross entropy, spatial frequency, and fusion index were used. Several so-called protocols have been proposed to evaluate the quality of algorithms that use one single-band (panchromatic) image of high spatial resolution to increase the spatial resolution of a higher-spectral-resolution multispectral image. These protocols evaluate the quality of fused images depending on the coherence and similarity of synthesized images compared to the corresponding expected images acquired by a sensor [47]; correlation between high-frequency information extracted from the single-band and fused images [48]; combination of spatial and spectral distortion indices assuming that spectral similarity relationships are invariant to scale changes [49]; and separate measurement of spectral and spatial quality using matched low-pass and high-pass complements of modulation transfer function filters, respectively [50].

The evaluated appropriateness of studied models was constrained to particular cases, without a wide general application scope. This trend was confirmed by

the lack of adapted tools for automatic quality assessment to compare results of airborne satellite multisensor image fusion [51] and evaluation of integrated spatial and spectral quality in the context of high panchromatic and low-spatial-resolution images [52]. Moreover, the heterogeneity of other proposed quality approaches to evaluate fusion results – like the combination of qualitative quality and basic statistical measures [53–55], purely visual qualitative criteria [56], or quantitative metric based on the human vision system [57] – is an indication of the multiple known problems to infer how quality should be assessed in different fusion scenarios. The restricted scope of proposed quality metrics suggests that quality evaluation of fused images is rather a complex open problem as recently summarized in [58, 59].

17.3 Quality Assessment for Geographic Information Systems

Remote sensing products as land cover (physical properties of land) and land use (human activities on the mapped surface, among others) are commonly integrated with the geographic information systems (GIS). Collected data and extracted information about large areas can be further exploited dynamically within a GIS system to store, process, and analyze spatial information through time, making use of specialized tools (Fig. 17.2). Some well-known applications are visualization, cartography, geographic space modeling, identification of natural and artificial objects, follow-up of land and water use, management of natural resources, disaster monitoring, evaluation of human activity impact on the environment, epidemiology, infrastructure design, etc. More specifically, infection risk profiling and propagation models of various parasitic infections were defined according to socioeconomic status, access to water, and education level [60]. Patterns of ecological and environ-

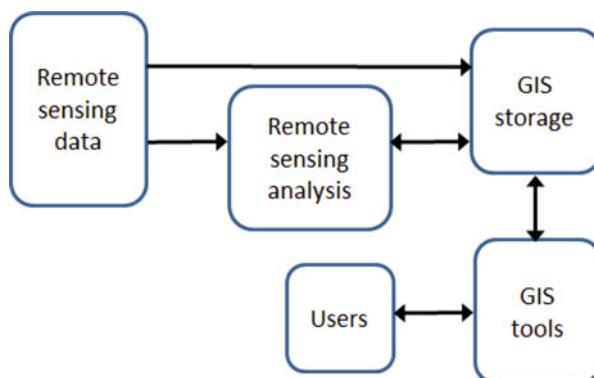


Fig. 17.2 Remote sensing and the GIS framework

mental factors related to parasitic infections were observed and detected too [61]. The integration of remote sensing products and GIS has been also applied to flood management during its development [62], the estimation of groundwater potential [63], and the evolution of land use depending on urban growth [64]. Currently, applications considering monitoring of crops and forests in precision agriculture [65] as well as spatial variability analysis in soil erosion [66] by conceptual models have been studied.

Early awareness of GIS quality underlined the importance of spatial data fitness to a given purpose, allowing the user to make informed decisions [67]. The underlying quality evaluation road map highlighted the need to trace how original raw data were modified to obtain information, explaining applied quantitative measures, along with the validation of completeness, consistency, accuracy, and time coherence. Multiple efforts have been deployed since then, placing spatial data quality as a core sub-discipline of geographic information sciences [68]. As a result, spatial data quality has become a focal point of specific international standards, based on research contributions, according to the notion of fitness for use, although there is a notable lack of shared terminology among different groups of interest that conceive, develop, maintain, and use GIS.

Even if the quality of GIS inputs from remote sensing products may have been previously evaluated, specific quality assessment for GIS intends to guarantee that efficient use will be given to expensive data and information. A first standardization attempt for data and quality information was introduced by the International Organization for Standardization (ISO) to address quality management and assurance [69]. Since then, other agencies have defined additional abstract standards relating to quality and metadata, e.g., CEN/TC 287, ISO/TC 211, and OGC, to define experimental modules for different needs, design extensible and sharable concepts, and conceive web services, respectively [70]. Some of these standards have been revisited to take into account sensing technology and GIS tools evolution. ISO extended the geographic metadata standards ISO 19101 [71] and ISO 19115 [72]. It also defined a reference geographic imagery processing model, ISO 19101-2:2008 [73], to include sensor data as image information and imagery knowledge. A schema required for describing imagery and gridded data was specified in ISO 19115-2:2009 [74], to combine information about spatial representation, data quality, and acquisition. Alternatively, ISO/TC 211 [75] searches to structure standards for georeferenced data and information, including imagery, metadata, and data quality models, for the development of specific applications.

An extensive attempt to standardize GIS data quality elements is ISO 19157:2013 [76]. It examines spatial quality of a product in comparison with its specification, according to six elements, which describe how a dataset meets product or user criteria. The elements are:

- Completeness: presence or absence of features, their attributes, and relationships
- Thematic accuracy: correctness of nonquantitative attributes and of features classifications, accuracy of quantitative attributes, and their relationships

- Logical consistency: degree of adherence to logical rules of data structure, attribution, and relationships
- Temporal quality: accuracy of the temporal attributes and temporal relationships of features
- Positional accuracy: accuracy of assigned position indicators.
- Usability: conformance to a set of user requirements, i.e., suitability of a dataset quality for a particular application (less structured than the other elements)

Three complementary quality elements – confidence, “representativity,” and homogeneity – provide quantitative and qualitative accounts on the quality evaluation against stated criteria. Whenever detailed product specification is not available or the given specification lacks quantitative measures and descriptors, qualitative evaluation applies.

Besides the literature describing the standardization efforts, an exhaustive theoretical and operational description of concepts, problems, and existing solutions for the production and utilization of properly qualified geographic data is presented in [77]. There are also academic works that have investigated how to make use of existing quality description and evaluation resources. An approach was proposed to relate quality information to user operations, making the user aware of documented quality issues [78]. This procedure required incorporating information about data quality as part of GIS metadata, which is not a general practice [79]. Such metadata should be comprehensive and properly presented to increase trust in information, which requires the development of adapted tools to facilitate user understanding. Fitness for use of spatial data was examined in the context of ecological assessment and monitoring [80]. It was based on two concepts: a set of user-defined values considered as specific quality indicators, i.e., expected quality, and a set of indicators or critical factors that reveal if a given dataset is suitable for a specified application context. With the advent of open and collaborative platforms, geographical data – mainly map, image, and text – are also generated by untrained volunteer individuals, who annotate data referring it to specific geographic locations. Lack of volunteers’ knowledge, spatial coverage heterogeneity, and inexistence of standards in this domain make necessary to study also the quality of volunteered geographic information [81].

17.4 Toward a Comprehensive Quality Model

Quality measurement, analysis, and interpretation in remote sensing differ across sensing devices, data and information fusion, standards, and applications. Consequently, quality evaluation has been defined in numerous manners for different systems and applications. Even if reported works share the same global objective of assessing quality to improve decision-making, they appear paradoxically as separate efforts without a common comprehensive model. In domains in which quality assessment has been an early interest, concerted models emerged. Namely, in information management systems [82], the main ideas of fitness for use and

evaluation according to dimensions emerged [83], along with some specific measures [84]. Furthermore, works on web information systems investigated the interest of quality in data mining [85] and databases [86]. However, very few works have studied specific methodological aspects of quality assessment of remote sensing components. Examples of these studies are the interestingness and data quality dimensions for sensor streams [87]; incorporation of quality information sources in decision workflows for information fusion [88]; and quality evaluation of information fusion systems [89]. Accordingly, the need for comprehensive remote sensing quality assessment models represents one important challenge in geospatial research [7].

Existing quality assessment models from other disciplines cannot be directly applied to remote sensing because of substantial differences in measurement variables, times, test points, acquisition-decision workflows, and overall objectives. As a possible alternative, since remote sensing systems are constructed by assembling sensors, smart components, control and communication devices, and data and information processing algorithms, along with human interaction to carry out intelligent monitoring, it can be considered as an instance of cyber-physical systems [90–92]. Such concept could permit to use a reference quality evaluation framework while refraining from very specific isolated qualitative or statistical descriptions of quality. In addition, cyber-physical systems – designed also for acquisition-decision-making workflows – share with remote sensing systems equivalent requirements for quality assessment models [93]. Nevertheless, it is important to note that most works on data and information quality for remote sensing systems do not refer specifically to the cyber-physical system concept despite using it implicitly. In the following subsections, we provide the main common definitions applied to data and information quality assessment, adapted from the framework of cyber-physical systems, as described in [94].

17.4.1 Data Quality

Data are defined as streams of bits with no comprehensible sense, including multidimensional acquired signals, system commands, operators' inputs, and decision-maker requests, among others. Data usually resulted from modeling and representation. Modeling results in <entity, attribute, value> triplets that allow for defining the observation realm, as the measured values of attributes corresponding to an entity. Representation involves displaying or recording these triplets. The data abstraction level – necessary data detail from simple to elaborated spatial, physical, and temporal details – will then be determined by the context of use. Data quality is thus mainly understood as the characterization of five key imperfections:

- Erroneous: Data are erroneous when values are different from the true data.
- Incomplete: Data are not fully supplied as expected because of missing values.
- Imprecise: Data inaccuracy does not permit to identify true values but possible approximations.

- Uncertain: Data cannot be specified with absolute confidence.
- Unavailable: The system cannot obtain some sets of values because of its limitations.

Given that erroneous data affect the integrity of a system, it should be discarded when detected. This action avoids extracting information from wrong data. Data altered by the last four imperfections could still supply partially valid information, depending on available observations through time, as well as changes in associated operational conditions. For instance, unavailable data can be estimated when necessary operational conditions are known.

17.4.2 *Information Quality*

Data processing permits to extract information, i.e., data with a semantic sense in a particular context combining the functional context of the given sub-system and the system it is integrated to. Due to the fact that any information can be a part of different tasks, quality evaluation should be task-independent. For this reason, three categories of dimensions are necessary to estimate information quality, namely, intrinsic, contextual, and extrinsic. The *intrinsic* category is the group of 13 quality dimensions defined for an isolated sub-system (Table 17.1). Eight quality dimensions that describe a sub-system as part of a complete system are included in the *contextual* category (Table 17.2). Evaluations of streams quality when multiple sub-systems are interconnected are comprised in the ten dimensions of the *extrinsic* category (Table 17.3).

Table 17.1 Intrinsic information quality dimensions

Name	Description
Source precision	The extent to which every information under constant source acquisition conditions remains the same
Accuracy	The extent to which extracted information is close to the true information
Objectivity	The extent to which information is unbiased, unprejudiced, and impartial
Reputation	The extent to which information is highly regarded in terms of its source or content
Obsolescence	The extent to which information is valid through time
Freshness	The extent to which information is new
Acquisition cost	The cost to acquire the information
Readability	The extent to which data used to obtain information are noiseless and intelligible
Resolution	The extent to which data used to obtain information are sampled
Integrity	The extent to which information is complete and accurate and the provider sub-system is fully available
Consistency	The extent to which information is accessible in the same format consistency
Uniqueness	The extent to which information is not repeated

Table 17.2 Contextual information quality dimensions

Name	Description
Real precision	The extent to which every information under constant sub-system use conditions remains unchanged
Clarity	The extent to which information is comprehensible along with other information
Trust	The extent to which information is trustworthy
Value added	The extent to which information is useful and provides advantages from its use
Timeliness	The extent to which information is expected by the system at a certain time
Completeness	The extent to which information is known in a complete context
Concision	The extent to which information is compactly represented
Volume	The extent to which the volume of information is appropriate for the task at hand
Believability	The extent to which information is regarded as true and credible

Table 17.3 Extrinsic information quality dimensions

Name	Description
Accessibility	The extent to which information is available and easily and quickly retrievable
Security	The extent to which access to information is restricted appropriately to maintain its confidentiality
Ease of use	The extent to which information is easy to use and apply to different tasks
Manipulation	The extent to which an information unit is easy to manipulate
Interpretability	The extent to which information is in appropriate languages, symbols, and units and the definitions are clear
Compatibility	The extent to which information is comprehensible for different sub-systems
Format	The extent to which information respects a specific format
Understandability	The extent to which information is easily comprehended
Redundancy	The extent to which other sub-systems provide the same information
Coherence	The extent to which information is consistent with respect to other information

These nonexhaustive tables of information quality dimensions are proposed as a suitable basic structure to be completed by further developments, along with the main objective of building a shared quality evaluation model. Both quantitative and qualitative aspects of information quality assessment are included. Such aspects are commonly required for use cases that cannot be only evaluated quantitatively or qualitatively. The flexible nature of this basic structure permits to include required dimensions of new use cases in a reliable manner for remote sensing. This approach requires nevertheless categorizing groups of measures, depending on sensors, extracted information, and decisional contexts [95]. In this way, adapted quality aspects can be evaluated at different stages of the acquisition-decision process, taking into account the singularity of processes and necessary measurements.

Information quality evaluation according to three categories of dimensions intends to take the abstraction of *sub-system* as the basic reference functional module to structure the measurements. Therefore, the essential assessment concerns

intrinsic information quality dimensions of any system module. This implies that the availability of features characterizing, for example, the location of the sub-system or to which system it is transmitting a stream, is required to extend the evaluation at the contextual and extrinsic levels, respectively. On the other hand, the aforementioned theoretical model is rarely applied exhaustively, because all the necessary elements are not always necessary or available. Additionally, real operational conditions may change dynamically, requiring different partial quality evaluations, instead of static evaluations.

17.5 Conclusions

Remote sensing is characterized by multiple types of systems with varied scale and complexity. As these systems continue to be developed, a considerable issue is the analysis of huge data volumes for decision support requiring taking into account specific operational contexts. Since numerous factors can alter collected data and extracted information, decision support can be lightly or severely impacted in an unmanageable manner. Methodologies and models for quality estimation emerge therefore as a possibility to determine whether collected data and extracted information are relevant for supporting decisions. It appears however that quality evaluation in remote sensing lacks a strong theoretical ground shared by stakeholders.

Several reasons explain this consideration. Acquired data have multiple imperfections associated with operational conditions and sensor characteristics, which make necessary to construct separate quality indicators. For instance, image acquisition is impacted by sensing technology, sensor calibration, spatial and temporal resolution, sensor inherent distortion, and restricted coverage, among others. Also, ground truth is very limited, leading to building data quality analysis on numerous assembled suppositions. Despite the latter, impact of data acquisition quality on registration, segmentation, classification, and fusion, quality assessment is not yet understood as a critical component of the processing chain but rather as a by-product that permits to show how result pertinence conform to the searched objective. Moreover, the amount of quality metrics constantly broadens with the emergence of new sensing technologies, applications, and type of studied problems. Knowing that quality estimations may be inappropriate despite their quantitative nature, complementary qualitative evaluations are commonly left to user interpretation.

On the other hand, quality assessment is considered to be critical when information extracted from remote sensing is integrated to find relations and patterns in the GIS domain. Multiple initiatives have intended to formalize how to assess the quality in cartography and derivate products of information obtained from remote sensing. There is consensus about the notion of map fitness for specific applications, particularly on the quality of all information processing results, which conditions strongly the practical value of products. A rich variety of quality indicators has been conceptualized to evaluate and interpret quality assessments from product input to final product use. Nevertheless, the availability of so many quality indicators

generates at least three new problems: identify an adapted quality metric, correctly integrate it into the concerned application, and provide a suitable visualization of data and information quality indicators for the user. The existing state of GIS-oriented quality systems development is confirmed by the omission of practical reports in the literature about progress on quality assessment implementation.

Given that methodological dispersion of proposed approaches for data and information quality assessment in remote sensing are based mainly on unstructured ad hoc measurements, a preliminary comprehensive quality model was proposed. Taking into account the fact that quality cannot be reduced to tailored measurements, the model outlines some examples of significant quality evaluation aspects, to which existing and future metrics could be associated at particular situations. Furthermore, depending on the data, information, and application types, grouping principles of quality elements could be defined as quality evaluation profiles. Having an adapted comprehensive reference model could potentially orient the flawed growth of quality approaches, providing eventually the basis for a common understanding of operational quality assessment.

Most of raw data and information quality uncertainty will remain unless large open verified ground truth datasets are available. Such datasets must be used not just to compare experimental results but to assure that research is reproducible. Evidently, the collection and management of datasets require a considerable collective effort of all concerned parties. Once it has been obtained, quality assessment makes sense as a whole at the end of the acquisition-decision-making process. Although separate data and information quality measurements may be interesting for experts, a product user does not necessarily have the knowledge or the time to go through those measures to make an evaluation. Therefore, an issue that needs to be addressed is cumulative quality for a decision-maker. In spite of its importance, it appears marginally in the literature given the focus on individual metrics. Supposing that those cumulated metrics exist, two related questions arise: how to define grouping principles to characterize operational quality profiles and how to make automatic quality features selection to form groups according to the data, information, and application types as well as user needs. These two questions imply that automatic quality assessment should be associated with the decision-making process and its possible subjectivity for avoiding additional complexity. Finally, it is necessary to develop quality assessment approaches capable of following up the evolution of earth conditions on large-scale monitored surfaces, with variable temporal resolution in terms of days, weeks, months, years, or even decades. These approaches are necessary to cope with the variable characteristics of voluminous periodical or seasonal datasets, from which merely reduced subsets are currently being exploited. Only a shared understanding of evaluation principles and interests can prompt the development of reliable, flexible, dynamic, and human-friendly quality indicators, within an accepted reference model.

References

1. J.B. Campbell, R.H. Wynne, *Introduction to Remote Sensing* (Guilford Press, New York, 2011)
2. Y. Ma, H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, W. Jie, Remote sensing big data computing: Challenges and opportunities. *Futur. Gener. Comput. Syst.* **51**, 47–60 (2015)
3. T. Miyoshi, K. Kondo, K. Terasaki, Big ensemble data assimilation in numerical weather prediction. *Computer* **11**, 15–21 (2015)
4. T. Miyoshi, M. Kunii, J. Ruiz, G.Y. Lien, S. Satoh, T. Ushio, et al., Big data assimilation – revolutionizing severe weather prediction. *Bull. Am. Meteorol. Soc.* **97**(8), 1347–1354 (2016)
5. E. Osuteye, C. Johnson, D. Brown, The data gap: An analysis of data availability on disaster losses in sub-Saharan African cities. *Int. J. Disaster Risk Reduction* **26**, 24–33 (2017)
6. C. Senf, R. Seidl, P. Hostert, Remote sensing of forest insect disturbances: Current state and future directions. *Int. J. Appl. Earth Obs. Geoinf.* **60**, 49–60 (2017)
7. S. Li, S. Dragicevic, F.A. Castro, M. Sester, S. Winter, A. Coltekin, et al., Geospatial big data handling theory and methods: A review and research challenges. *ISPRS J. Photogramm. Remote Sens.* **115**, 119–133 (2016)
8. L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, *Remote Sensing Image Fusion* (CRC Press, Boca Raton, 2015)
9. L. Gómez-Chova, D. Tuia, G. Moser, G. Camps-Valls, Multimodal classification of remote sensing images: A review and future directions. *Proc. IEEE* **103**(9), 1560–1584 (2015)
10. J.A. Cummings, Ocean data quality control, in *Operational Oceanography in the 21st Century*, (Springer, Dordrecht, 2011), pp. 91–121
11. K.A. Kilpatrick, G. Podestá, S. Walsh, E. Williams, V. Halliwell, et al., A decade of sea surface temperature from MODIS. *Remote Sens. Environ.* **165**, 27–41 (2012)
12. H. Uehara, A.A. Kruts, Y.N. Volkov, T. Nakamura, T. Ono, H. Mitsudra, A new climatology of the Okhotsk sea derived from the FERHRI database. *J. Oceanogr.* **68**(6), 869–886 (2012)
13. F. Xu, A. Ignatov, In situ SST quality monitor (i quam). *J. Atmos. Ocean. Technol.* **31**(1), 164–180 (2014)
14. C. Donlon, I. Robinson, K.S. Casey, J. Vazquez-Cuervo, E. Armstrong, O. Arino, et al., The global ocean data assimilation experiment high-resolution sea surface temperature pilot project. *Bull. Am. Meteorol. Soc.* **88**(8), 1197–1214 (2007)
15. S. Guinehut, C. Coatanoan, A.L. Dhomps, P.Y. Le Traon, G. Larnicol, On the use of satellite altimeter data in Argo quality control. *J. Atmos. Ocean. Technol.* **26**(2), 395–402 (2009)
16. A.S. Bogdanoff, D.L. Westphal, J.R. Campbell, J.A. Cummings, E.J. Hyer, J.S. Reid, C.A. Clayson, Sensitivity of infrared sea surface temperature retrievals to the vertical distribution of airborne dust aerosol. *Remote Sens. Environ.* **159**, 1–13 (2015)
17. C.J. Merchant, A.R. Harris, E. Maturi, S. MacCallum, Probabilistic physically based cloud screening of satellite infrared imagery for operational sea surface temperature retrieval. *Q. J. R. Meteorol. Soc.* **131**(611), 2735–2755 (2005)
18. B.B. Barnes, C. Hu, A hybrid cloud detection algorithm to improve MODIS sea surface temperature data quality and coverage over the eastern gulf of Mexico. *IEEE Trans. Geosci. Remote Sens.* **51**(6), 3273–3285 (2013)
19. M. Bouali, A. Ignatov, Adaptive reduction of striping for improved sea surface temperature imagery from Suomi national polar-orbiting partnership (S-NPP) visible infrared imaging radiometer suite (viirs). *J. Atmos. Ocean. Technol.* **31**(1), 150–163 (2014)
20. P.K. Koner, A. Harris, E. Maturi, Hybrid cloud and error masking to improve the quality of deterministic satellite sea surface temperature retrieval and data coverage. *Remote Sens. Environ.* **174**, 266–278 (2016)
21. J.A. Cummings, Operational multivariate ocean data assimilation. *Q. J. R. Meteorol. Soc.* **131**(613), 3583–3604 (2005)
22. JCOMM Data Management Coordination. Final report of the third session of the JCOMM data management coordination group (jcomm dmcg-iii), Tech. Rep. 56, Intergovernmental Oceanographic Commission of UNESCO and World Meteorological Organization (2008)

23. E. Ahokas, H. Kaartinen, J. Hyppä, A quality assessment of airborne laser scanner data. *Int. Arch. Photogramm. Remote Sens.* **34**(part3), W13 (2003)
24. Z. Wan, Y. Zhang, Q. Zhang, Z.L. Li, Quality assessment and validation of the MODIS global land surface temperature. *Int. J. Remote Sens.* **25**(1), 261–274 (2004)
25. M. Neubert, H. Herold, G. Meinel, Evaluation of remote sensing image segmentation quality—further results and concepts. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **36**(4/C42) (2006). http://www.isprs.org/proceedings/XXXVI/4-C42/Papers/10_Adaption%20and%20further%20development%20II_OBIA2006_Neubert_Herold_Meinel.pdf
26. R.R. Colditz, C. Conrad, T. Wehrmann, M. Schmidt, S. Dech, TiSeG: A flexible software tool for time-series generation of MODIS data utilizing the quality assessment science data set. *IEEE Trans. Geosci. Remote Sens.* **46**(10), 3296–3308 (2008)
27. K.A. Razak, M.W. Straatsma, C.J. Van Westen, J.P. Malet, S.M. De Jong, Airborne laser scanning of forested landslides characterization: Terrain model quality and visualization. *Geomorphology* **126**(1–2), 186–200 (2011)
28. Q. Zhan, M. Molenaar, K. Tempfli, W. Shi, Quality assessment for geo-spatial objects derived from remotely sensed data. *Int. J. Remote Sens.* **26**(14), 2953–2974 (2005)
29. Y. Shuai, C.B. Schaaf, A.H. Strahler, J. Liu, Z. Jiao, Quality assessment of BRDF/albedo retrievals in MODIS operational system. *Geophys. Res. Lett.* **35**(L05407), 5 p (2008)
30. V.E. Brando, J.M. Anstee, M. Wettle, A.G. Dekker, S.R. Phinn, C. Roelfsema, A physics based retrieval and quality assessment of bathymetry from suboptimal hyperspectral data. *Remote Sens. Environ.* **113**(4), 755–770 (2009)
31. H.J. Buitenh, B. Van Putten, Quality assessment of remote sensing image registration-analysis and testing of control point residuals. *ISPRS J. Photogramm. Remote Sens.* **52**(2), 57–73 (1997)
32. T.R. Loveland, B.C. Reed, J.F. Brown, D.O. Ohlen, Z. Zhu, L.W.M.J. Yang, J.W. Merchant, Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. *Int. J. Remote Sens.* **21**(6–7), 1303–1330 (2000)
33. U. Weidner, Contribution to the assessment of segmentation quality for remote sensing applications. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **37**(B7), 479–484 (2008)
34. M.J. Smith, J. Rose, S. Booth, Geomorphological mapping of glacial landforms from remotely sensed data: An evaluation of the principal data sources and an assessment of their quality. *Geomorphology* **76**(1–2), 148–165 (2006)
35. S.O. Elberink, G. Vosselman, Quality analysis on 3D building models reconstructed from airborne laser scanning data. *ISPRS J. Photogramm. Remote Sens.* **66**(2), 157–165 (2011)
36. G.M. Foody, Harshness in image classification accuracy assessment. *Int. J. Remote Sens.* **29**(11), 3137–3158 (2008)
37. P.C. Smits, S.G. Dellepiane, R.A. Schowengerdt, Quality assessment of image classification algorithms for land-cover mapping: A review and a proposal for a cost-based approach. *Int. J. Remote Sens.* **20**(8), 1461–1486 (1999)
38. Y. Ke, L.J. Quackenbush, J. Im, Synergistic use of QuickBird multispectral imagery and LIDAR data for object-based forest species classification. *Remote Sens. Environ.* **114**(6), 1141–1154 (2010)
39. R.D. Fiete, T.A. Tantalo, Comparison of SNR image quality metrics for remote sensing systems. *Opt. Eng.* **40**(4), 574–586 (2001)
40. R.D. Fiete, T.A. Tantalo, J.R. Calus, J.A. Mooney, Image quality of sparse aperture designs for remote sensing. *Opt. Eng.* **41**(8), 1957–1970 (2002)
41. D. Scherler, S. Leprince, M.R. Strecker, Glacier-surface velocities in alpine terrain from optical satellite imagery – Accuracy improvement and quality assessment. *Remote Sens. Environ.* **112**(10), 3806–3819 (2008)
42. Q. Liu, R. Klucik, C. Chen, G. Grant, D. Gallaher, Q. Lv, L. Shang, Unsupervised detection of contextual anomaly in remotely sensed data. *Remote Sens. Environ.* **202**, 75–87 (2017)
43. J. Li, Spatial quality evaluation of fusion of different resolution images. *Int. Arch. Photogramm. Remote Sens.* **33**(B2; PART 2), 339–346 (2000)

44. W. Shi, C. Zhu, Y. Tian, J. Nichol, Wavelet-based image fusion and quality assessment. *Int. J. Appl. Earth Obs. Geoinf.* **6**(3–4), 241–251 (2005)
45. R.G. Congalton, K. Green, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices* (CRC Press, Boca Raton, 2008)
46. Y. Chen, Z.Y. Xue, R.S. Blum, Theoretical analysis of an information-based quality measure for image fusion. *Inf. Fusion* **9**, 161–175 (2008)
47. L. Wald, T. Ranchin, M. Mangolini, Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogramm. Eng. Remote. Sens.* **63**(6), 691–699 (1997)
48. J. Zhou, D.L. Civco, J.A. Silander, A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *Int. J. Remote Sens.* **19**(4), 743–757 (1998)
49. L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, M. Selva, Multispectral and panchromatic data fusion assessment without reference. *Photogramm. Eng. Remote Sens.* **74**(2), 193–200 (2008)
50. M.M. Khan, L. Alparone, J. Chanussot, Pansharpening quality assessment using the modulation transfer functions of instruments. *IEEE Trans. Geosci. Remote Sens.* **47**(11), 3880–3891 (2009)
51. J. Dong, D. Zhuang, Y. Huang, J. Fu, Advances in multi-sensor data fusion: Algorithms and applications. *Sensors* **9**(10), 7771–7784 (2009)
52. M. Ehlers, S. Klonus, P. Johan Åstrand, P. Rosso, Multi-sensor image fusion for pansharpening in remote sensing. *Int. J. Image Data Fusion* **1**(1), 25–45 (2010)
53. W. Wang, F. Chang, A multi-focus image fusion method based on Laplacian pyramid. *J. Comput.* **6**(12), 2559–2566 (2011)
54. X.X. Zhu, R. Bamler, A sparse image fusion algorithm with application to pan-sharpening. *IEEE Trans. Geosci. Remote Sens.* **51**(5), 2827–2836 (2013)
55. Y. Jiang, M. Wang, Image fusion with morphological component analysis. *Information Fusion* **18**, 107–118 (2014)
56. Y. Zhang, R.K. Mishra, From UNB PanSharp to Fuze Go – The success behind the pan-sharpening algorithm. *Int. J. Image Data Fusion* **5**(1), 39–53 (2014)
57. J. Liu, J. Huang, S. Liu, H. Li, Q. Zhou, J. Liu, Human visual system consistent quality assessment for remote sensing image fusion. *ISPRS J. Photogramm. Remote Sens.* **105**, 79–90 (2015)
58. P. Jagalingam, A.V. Hegde, A review of quality metrics for fused image. *Aquatic Procedia* **4**, 133–142 (2015)
59. R.C. Frohn, R.D. Lopez, *Remote Sensing for Landscape Ecology. New Metric Indicators: Monitoring, Modeling, and Assessment of Ecosystems* (CRC Press, Boca Raton, 2017)
60. C. Simoonga, J. Utzinger, S. Brooker, P. Vounatsou, C.C. Appleton, A.S. Stensgaard, et al., Remote sensing, geographical information system and spatial analysis for schistosomiasis epidemiology and ecology in Africa. *Parasitology* **136**(13), 1683–1693 (2009)
61. G.S. Bhunia, M.R. Dikhit, S. Kesari, G.C. Sahoo, P. Das, Role of remote sensing, geographical information system (GIS) and bioinformatics in kala-azar epidemiology. *J. Biomed. Res.* **25**(6), 373–384 (2011)
62. E. Opolot, Application of remote sensing and geographical information systems in flood management: A review. *Res. J. Appl. Sci. Eng. Technol.* **6**(10), 1884–1894 (2013)
63. D. Oikonomidis, S. Dimogianni, N. Kazakis, K. Voudouris, A GIS/remote sensing-based methodology for groundwater potentiality assessment in Tirnavos area, Greece. *J. Hydrol.* **525**, 197–208 (2015)
64. I.R. Hegazy, M.R. Kalooop, Monitoring urban growth and land use change detection with GIS and remote sensing techniques in Daqahlia governorate Egypt. *Int. J. Sustainable Built Environ.* **4**(1), 117–124 (2015)
65. N. Baghdadi, C. Mallet, M. Zribi (eds.), *QGIS and Applications in Agriculture and Forest* (Wiley, Hoboken, 2018)
66. R.J. Patil, *Spatial Techniques for Soil Erosion Estimation: Remote Sensing and GIS Approach* (Springer, Cham, Switzerland, 2018)

67. N.R. Chrisman, The role of quality information in the long-term functioning of a geographic information system. *Cartographica Int. J. Geographic Inf. Geovisualization* **21**(2–3), 79–88 (1984). Part 2 Issues and problems relating to cartographic data use, exchange and transfer
68. R. Devillers, A. Stein, Y. Bédard, N. Chrisman, P. Fisher, W. Shi, Thirty years of research on spatial data quality: Achievements, failures, and opportunities. *Trans. GIS* **14**(4), 387–400 (2010)
69. ISO 8402:1994, Quality management and quality assurance – Vocabulary, <https://www.iso.org/standard/20115.html>
70. S. Servigne, N. Lesage, T. Libourel, Quality components, standards, and metadata, in *Fundamentals of Spatial Data Quality*, (2006), pp. 179–210
71. ISO 19101-1:2014, Geographic information – Reference model – Part 1: Fundamentals, <https://www.iso.org/standard/59164.html>
72. ISO 19115-1:2014, Geographic information – Metadata – Part 1: Fundamentals, <https://www.iso.org/standard/53798.html>
73. ISO 19101-2:2018, Geographic information – Reference model – Part 2: Imagery, <https://www.iso.org/standard/69325.html>
74. ISO 19115:2009, Geographic information – Metadata – Part 2: Extensions for imagery and gridded data, <https://www.iso.org/standard/39229.html>
75. ISO/TC 211 Geographic information/Geomatics, <https://www.iso.org/committee/54904.html>
76. ISO 19157:2013(en), Geographic information – Data quality, <https://www.iso.org/obp/ui/#iso:std:iso:19157:ed-1:v1:en>
77. R. Devillers, R. Jeansoulin, *Fundamentals of Spatial Data Quality* (Wiley, Hoboken, 2010)
78. A. Zargar, R. Devillers, An operation-based communication of spatial data quality. *IEEE Int. Conf. Adv. Geogr. Inf. Syst. Web Serv.*, 140–145 (2009)
79. P. Díaz, J. Masó, E. Sevillano, M. Ninyerola, A. Zabala, I. Serral, et al., Analysis of quality metadata in the GEOSS clearinghouse. *Int. J. Spatial Data Infrastructures Res.* **7**, 352–377 (2012)
80. I. Pócas, J. Gonçalves, B. Marcos, J. Alonso, P. Castro, J.P. Honrado, Evaluating the fitness for use of spatial data sets to promote quality in ecological assessment and monitoring. *Int. J. Geogr. Inf. Sci.* **28**(11), 2356–2371 (2014)
81. H. Senaratne, A. Mobasher, A.L. Ali, C. Capineri, M. Haklay, A review of volunteered geographic information quality assessment methods. *Int. J. Geogr. Inf. Sci.* **31**(1), 139–167 (2017)
82. D.P. Ballou, H.L. Pazer, Modeling data and process quality in multi-input, multi-output information systems. *Manag. Sci.* **31**(2), 150–162 (1985)
83. R.Y. Wang, D.M. Strong, Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.* **12**(4), 5–33 (1996)
84. S.E. Madnick, R.Y. Wang, Y.W. Lee, H. Zhu, Overview and framework for data and information quality research. *J. Data Inf. Qual.* **1**(1), article 2, 22 p (2009)
85. Z. Chen, *Data Mining and Uncertain Reasoning: An Integrated Approach* (Wiley, New York, 2001)
86. Naumann F, From databases to information systems-information quality makes the difference. 6th International Conference on Information Quality. (2001) pp. 244–260
87. A. Klein, W. Lehner, Representing data quality in sensor data streaming environments. *J. Data Inf. Qual.* **1**(2), 10 (2009)
88. Rogova GL, Bosse E, Information quality in information fusion. 13th IEEE Conference on Information Quality in Information Fusion, (2010) pp. 1–8
89. I.G. Todoran, L. Lecornu, A. Khenchaf, J.M. Le Caillec, Fusion systems evaluation, in *Multisensor Data Fusion: From Algorithms and Architectural Design to Applications*, (CRC Press, Boca Raton, USA, 2017), pp. 147–156
90. V. Gunes, S. Peter, T. Givargis, F. Vahid, A survey on concepts, applications, and challenges in cyber-physical systems. *KSII Trans. Int. Inf. Syst.* **8**(12), 4242–4268 (2014)

91. N. Chen, C. Xiao, F. Pu, X. Wang, C. Wang, Z. Wang, et al., Cyber-physical geographical information service-enabled control of diverse in-situ sensors. *Sensors* **15**(2), 2565–2592 (2015)
92. G. Mois, T. Sanislav, S.C. Folea, A cyber-physical system for environmental monitoring. *IEEE Trans. Instrum. Meas.* **65**(6), 1463–1471 (2016)
93. K. Sha, S. Zeadally, Data quality challenges in cyber-physical systems. *J. Data Inf. Qual.* **6**(2–3), 8 (2015)
94. P. Merino Laso, D. Brosset, J. Puentes, Monitoring approach of cyber-physical systems by quality measures. *Lect. Notes Inst. Comput. Sci. Soc. Informatics Telecommun. Eng.* **205**, 105–117 (2016)
95. P. Merino, Laso, D. Brosset, J. Puentes, Analysis of quality measurements to categorize anomalies in sensor systems. *IEEE Comput. Conf.*, 1330–1338 (2017). <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8252077>

Chapter 18

Reliability-Aware and Robust Multi-sensor Fusion Toward Ego-Lane Estimation Using Artificial Neural Networks



Tran Tuan Nguyen, Jan-Ole Perschewski, Fabian Engel, Jonas Kruesemann, Jonas Sitzmann, Jens Spehr, Sebastian Zug, and Rudolf Kruse

Abstract In the field of road estimation, incorporating multiple sensors is essential to achieve a robust performance. However, the reliability of each sensor changes due to environmental conditions. Thus, we propose a reliability-aware fusion concept, which takes into account the sensor reliabilities. By that, the reliabilities are estimated explicitly or implicitly by classification algorithms, which are trained with extracted information from the sensors and their past performance compared to ground truth data. During the fusion, these estimated reliabilities are then exploited to avoid the impact of unreliable sensors. In order to prove our concept, we apply our fusion approach to a redundant sensor setup for intelligent vehicles containing three-camera systems, several lidars, and radar sensors. Since artificial neural networks (ANN) have produced great results for many applications, we explore two ways of incorporating them into our fusion concept. On the one hand, we use ANN as classifiers to explicitly estimate the sensors' reliabilities. On the other hand, we utilize ANN to directly predict the ego-lane from sensor information, where the reliabilities are implicitly learned. By the evaluation with real-world recording data, the direct ANN approach leads to satisfactory road estimation.

Keywords Information fusion · Reliability · Neural networks · Ego-lane estimation · Intelligent vehicles

T.T. Nguyen (✉) · J.-O. Perschewski · F. Engel · J. Kruesemann · J. Sitzmann · J. Spehr
Volkswagen Aktiengesellschaft, Wolfsburg, Germany
e-mail: Tran.Tuan.Nguyen@volkswagen.de; Jan-Ole.Perschewski@volkswagen.de;
Fabian.Engel@volkswagen.de; Jonas.Kruesemann@volkswagen.de;
Jonas.Sitzmann@volkswagen.de; Jens.Spehr@volkswagen.de

S. Zug · R. Kruse
Faculty of Computer Science, Otto-von-Guericke University Magdeburg, Magdeburg, Germany
e-mail: Sebastian.Zug@ovgu.de; Rudolf.Kruse@ovgu.de

18.1 Introduction

Advanced driver assistance systems (ADAS) and automated driving heavily rely on environment perception and especially on road estimation. By that, current research explores various algorithms toward road detection by using multiple sensors, such as camera, radar, lidar, etc. Thereby, one of the biggest challenges is the huge variety of environmental conditions that influence sensor performances. This leads to sensor failures in several scenarios. For instance, a camera-based detection system can provide sufficient results under many weather conditions. However, this system can fail in case of heavy rain, snow, etc. In contrast, radar sensors can detect the surrounding objects despite these conditions since their technology is not affected by rain or snow as cameras. For that reason, it is necessary to combine the data of distinct sensors so that the system can constantly produce sufficient results.

In our previous works, we introduce a multi-source fusion framework for robust ego-lane detection [1–3]. Thereby, we take into account that the sensor reliabilities depend on environmental conditions and can change over time. The reliabilities are estimated by applying different classification algorithms, which are offline trained by using the extracted information from sensors' detections. Consequently, the fusion process based on *Dempster-Shafer theory* incorporates these reliabilities to combine the information of the sources.

In this work, we exploit the possibility of estimating the ego-lane directly by using neural networks. By that, the reliabilities are internally learned by the networks and encoded as weights of the neurons. This differs from the approaches of Nguyen et al. in [3, 4], where the reliability of each source is estimated by training a separate classifier. Furthermore, we integrate new environment information to take advantage of the redundant sensor system, such as detections from a surround view camera system, free space information, etc. To achieve higher accuracy of the classification, we utilize the mutual information of the features to select the features with the greatest influence on the classification. Finally, we evaluate our presented approaches by using a new database of real-world data recordings.

This work is organized as follows: Sect. 18.2 explains three categories of perception approaches toward automated driving and gives an overview of various works. In Sect. 18.3, we introduce our concept of incorporating reliabilities into ego-lane estimation by using different classifiers. Following, Sect. 18.4 applies neural network to explicitly learn the sensors' reliabilities. Afterward, Sect. 18.5 explains our approach of using neural networks to directly estimate the ego-lane. Lastly, Sect. 18.6 presents the experimental results obtained for the feature selection, the reliability estimation, and the final ego-lane estimation.

18.2 Related Work

The approaches in the field of automated driving can be divided into three categories [5], illustrated in Fig. 18.1. The first category consists of *behavior reflex approaches*, which use purely *data-driven techniques*, also called as *AI techniques*, to map sensor data to driving decisions directly. The second category with *direct perception approaches* apply AI algorithms to estimate a selected set of features representing the relevant information of the current environment. Afterward, a simple controller uses these features to realize driving functions. Representing the third category, *mediated perception approaches* build an environment model by processing the sensor data using both *model-based methods* and *AI techniques*, respectively. Based on the generated environment model, AI methods are utilized to derive the driving actions of the vehicle. Following, all categories will be discussed in detail.

18.2.1 Behavior Reflex Approaches

In the early stages of automated driving, Pomerleau et al. propose a behavior reflex approach using an *artificial neural network* (ANN) to estimate the steering angle for an intelligent vehicle [6]. Thereby, the network consisting of only three layers is trained by using a low-resolution 30×32 pixel camera image. Thus, the input layer of the ANN contains 960 neurons. Following, the input layer is fully connected to the hidden layer consisting of five neurons, which in turn is connected to the output layer of 30 neurons. Each neuron in the last layer represents a steering angle that is used to calculate the steering of the vehicle. To provide a stable behavior, the final steering angle is determined by calculating the center of masses of the activations around the highest activated neuron.

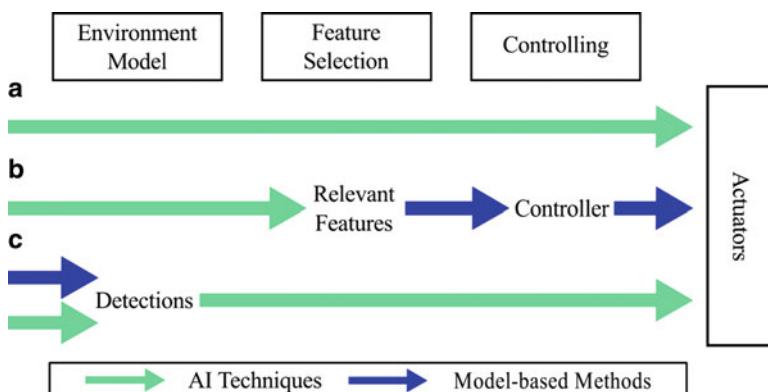


Fig. 18.1 Perception models: (a) Behavior reflex, (b) Direct perception, and (c) Mediated perception approaches

A more sophisticated approach using ANNs is presented by Bojarski et al. [7]. Their system uses *convolutional neural network* (CNN) as recent advances of ANNs. For that, they use the images of three cameras to determine the steering wheel angle. In order to train the multilayer CNN, backpropagation is performed with the mean squared error of the estimated angle to the angle chosen by a human driver. Additionally, they rotate and shift the images to avoid an overfitting to the training data. In their evaluation, they reach an autonomy level of 98%, which is defined as follows:

$$\text{autonomy Level} = \left(1 - \frac{\#\text{interventions} \cdot 6 \text{ seconds}}{\text{elapsedTime}} \right) \cdot 100 \quad (18.1)$$

A similar approach using CNN to determine the steering angle is presented by Chen et al. [8]. The resulting network is able to perform steering with a mean error of 2.42°. However, the authors explain that evaluating the camera images frame by frame is not appropriate since the repetition of the small error in every frame can result in leaving the lane. Thus, they conclude that it is necessary to incorporate temporal information into the network to improve the results in a continuing scenario.

Codevilla et al. [9] propose a more practice-oriented approach by incorporating commands into the learning process. Therefore, they use a camera system which determines the steering angle and acceleration using a CNN. Furthermore, they compare two architectures for their networks. On the one hand, the command input architecture combines the image processing results, the measurements of the environment, and the command by feeding the outputs into fully connected layers, which determine the action. On the other hand, the branched architecture combines the image processing results and environment measurements and forwards the outputs into fully connected layers depending on the command. Impressively, the branched version drove an off-the-shelf 0.20 scale truck nearly perfectly on walkways in a residential area.

The problems of using behavior reflex approaches are that it is hardly possible to install a fail-safe. This can result in accidents in unknown environments and endanger other traffic participants.

18.2.2 Direct Perception Approaches

In [5], Chen et al. introduce a direct perception approach for autonomous driving by choosing a set of 13 features to represent the current environment. These features contain information about the angle between the vehicle and the road, distances to lane markings, and preceding vehicles on other lanes. Using these features, the authors construct a controller, which minimizes the distance to the lane center line and keeps a safe distance to other traffic participants. In order to determine the features, they use two different approaches: a handcrafted GIST system [10] and

CNN. As a result, CNN outperforms the GIST system regarding every parameter. Using the superior CNN, they develop a system that can perform well in both virtual and real environments. Although this approach seems to achieve good results, two problems can occur. First, the controller depends strongly on correct inputs, which cannot be ensured in the current state. Secondly, if this approach needs to be scaled to fully autonomous driving, the selected features will become as complex as in mediated perception approaches. Therefore, the simple controllers will not be sufficient and should be replaced by mediated perception approaches.

Similar to [5], Al-Qizwini et al. provide a different direct perception approach called GIAD [11]. Therefore, they compare the top three CNN architectures, namely, GoogLeNet [12], VGGNet [13], and Clarifai [14]. These CNNs are used to learn five affordance parameters, which are used by the controller to drive the intelligent vehicle. During the training of the CNNs on images provided by TORCS, GoogLeNet outperforms VGGNet and Clarifai. Hence, they use GoogLeNet as the best network to evaluate the automated driving capability in a simulated environment by measuring the mean and deviation to the lane center. Their algorithm performs well and achieves a mean deviation on the evaluation tracks of at most 0.2 m. Although this approach seems to be promising, it suffers from the lack of complexity in comparison to real-world scenarios because of using simulation results. By way of example, they cannot simulate all mistakes that other traffic participants could make to react accordingly.

18.2.3 Mediated Perception Approaches

Mediated perception approaches are characterized by modeling a complex environment representation when combining information from several sensors. Thereby, the biggest challenge is how to handle inconsistency and conflict between the information coming from different sources. Thus, several works investigate the sensor reliability by using different methods, e.g., classifiers [3, 15, 16] and failure models [17]. At the decision layer, these reliabilities can be exploited to fuse only reliable sources.

Frigui et al. present a context-dependent multisensor fusion framework [18]. By that, they use a clustering algorithm to cluster the extracted features. Each cluster represents a certain context and contains data that shows similar characteristics of the environment. Afterward, a reliability of each source is manually defined for each context. This approach can be problematic when the number of features rises, and the clustering algorithms will suffer from the curse of dimensionality. In this case, the number of clusters would rise exponentially.

In [15], Hartmann et al. fuse multiple sensors to create a road model, which is then verified with a digital map. Therefore, they train an ANN using a large database containing sensor data and the associated map geometry. The goal is to assess whether the estimated road model is incorrect and does not match with the digital map. This can be the case when the predicted road course changes due to

construction works or errors of the detection algorithms. As a result, the trained ANN outputs a reliability value representing the probability for an error between the estimated road and the digital map. This approach can detect contradictions, but it cannot decide which source is faulty [19]. Hence, this method could be improved by identifying the incorrect source [20].

Realpe et al. introduce a fault-tolerant object estimation framework [21]. First, objects are separately estimated by using data from each single sensor. For each sensor, the discrepancy of its estimated objects to the reference in the offline evaluation phase is used as weight for the final fusion. This concept is promising, but the reliability estimation could be further increased by using additional context information, such as the road type, where the vehicle is driving on.

Romero et al. present an environment-aware fusion approach for lane estimation [22]. By that, they compare the estimated lane from each sensor with the ground truth. Based on the comparison result, they assign a reliability value to each sensor for the current GPS position. When the vehicle is located at a certain position, the stored reliabilities are used to perform a weighted fusion. However, this approach is not generalizable to new areas since it uses GPS position to predict reliabilities and requires the vehicle to have been there before. Instead of utilizing GPS coordinates, additional features extracted from sensor detections could be used to make the estimations location-independent [19].

The discussed works in this chapter contain interesting approaches, but they still have potential for improvement or are quite work-intensive. Hence, the following chapter will explain our fusion concept.

18.3 Overall Concept

Our fusion concept is an extension of our previous work in [3, 23]. As illustrated in Fig. 18.2, it consists of multiple levels such as in the JDL model [24]. At *Level 0*, the raw sensor data is preprocessed on the basis of physical signal level. At *Level 1*, multiple detection modules iteratively utilize the preprocessed data to estimate and predict the states of different object types. This includes tasks such as object detection, tracking, association, etc. The low-level fusion, e.g., object association of different sensors [25, 26], is taking place here. In our work, the used sensors are delivered with their internal processing modules and provide different results such as lane markings, dynamic objects, etc.

Starting from *Level 2*, we present two different fusion concepts, where reliable sources should be preferred over unreliable sources. In the first approach represented in Sect. 18.4, we utilize artificial neural networks (ANNs) to estimate the reliability of different ego-lane models by using the scenario features, which are extracted from the sensor and contextual information. Afterward, the fusion based on Dempster-Shafer theory utilizes these estimated reliabilities to identify and neglect the unreliable sources. In the second approach, we utilize ANNs to directly estimate the

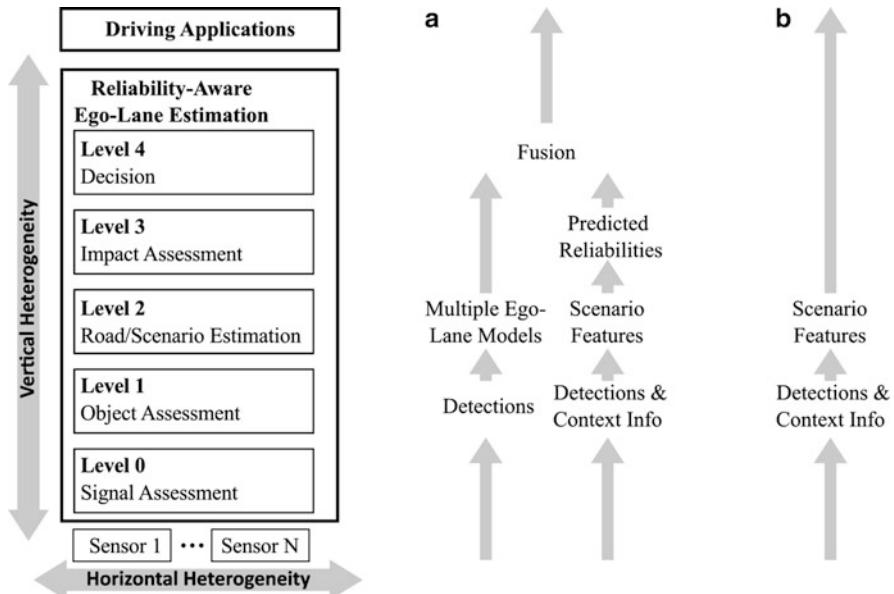


Fig. 18.2 Overview of our two different fusion concepts. While (a) estimates the reliabilities of the separately estimated ego-lane models and incorporates them into the fusion, (b) estimates the ego-lane directly using sensor detections

ego-lane (Sect. 18.5). By that, the network should internally learn the reliabilities of the sources for an optimal estimation. Both concepts are detailed in their respective sections. Since the scenario features are used by both approaches, we will explain them in the following.

18.3.1 Sensor Setup

As shown in Fig. 18.3, we use a setup of three-camera systems in order to detect lane markings. Thereby, each camera system separately provides estimations for the next right lane marking (*RM*) and the next left lane marking (*LM*). In this work, a prefix of “*second*” or “*third*” denotes the affiliation to that particular camera system. If no prefix is given, the estimation belongs to the first camera system. Furthermore, the prototype vehicle also is also equipped with several radar and lidar sensors for a 360° object detection, which is not be further explained here.

By way of example, Fig. 18.4 shows four scenarios with the detected lane markings and objects. The highway in Fig. 18.4a demonstrates an ideal scenario, where all lane markings can be perceived clearly. Thereby, the two front-facing camera systems can detect markings up to 100 m, while the third camera system has a shorter detection range of about 20 m. In this scenario, the vehicle can use

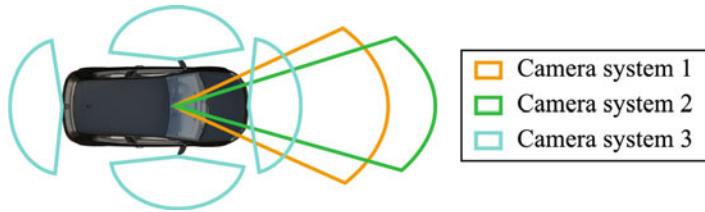


Fig. 18.3 The prototype vehicle with three-camera systems: Two are front-facing and differ slightly in field of view; the third consists of four fish-eye cameras for a surround view. The positions of other sensors such as lidars, radars, and ultrasonic sensors are not shown here

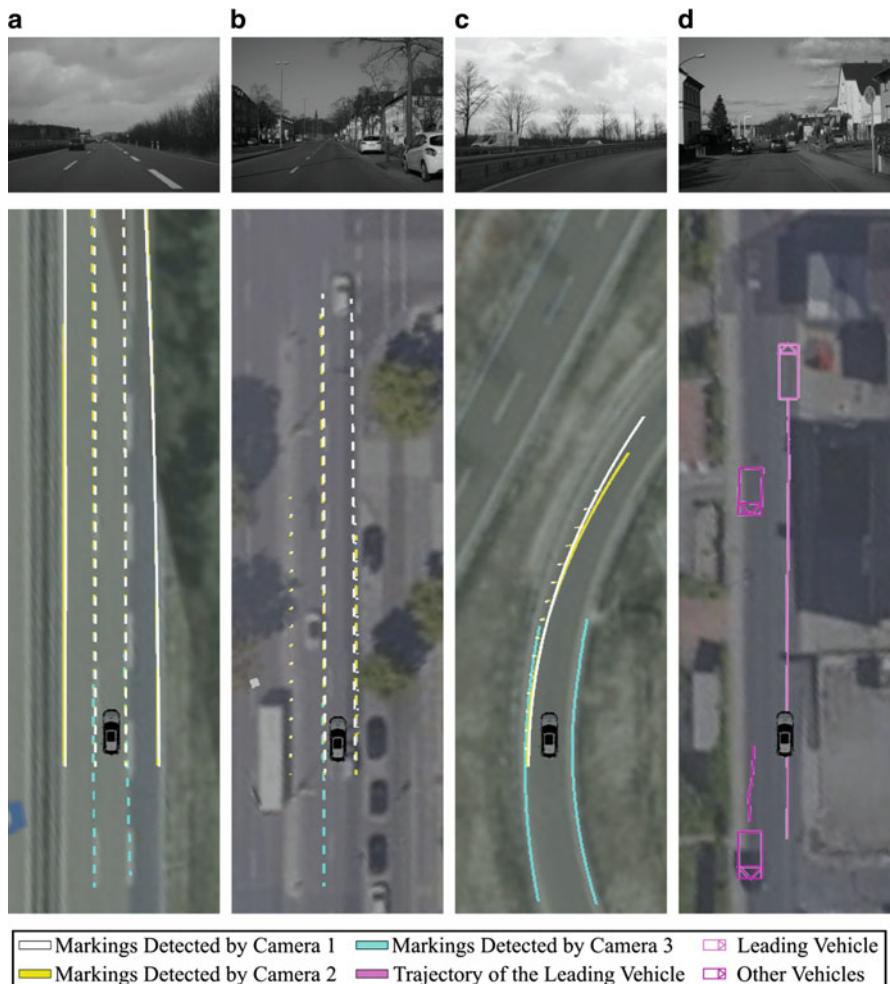


Fig. 18.4 First row: images from the first camera. Second row: visualization of detection results of all three cameras and object estimations on *Google Maps*

any marking from the first two cameras or a combination of them to estimate the current ego-lane. As opposed to this, Fig. 18.4b depicts an urban scenario, where the detection ranges of all cameras are smaller than in the highway scenario. Moreover, markings are not existing on the right side so that only Camera 1 and Camera 2 can identify the curbstone as lane boundary. In contrast, the left lane marking is perceived clearly by all cameras. Therefore, the vehicle should orientate to the left lane marking. Especially in the on-ramp scenario in Fig. 18.4c, the third camera system outperforms the rest by detecting markings on both sides up to 20 m away. Here, the first two camera systems cannot recognize the right marking due to their narrow field of views (Fig. 18.4). In order to handle this scenario, the vehicle should utilize the detected markings of Camera 3. Last, Fig. 18.4d depicts another urban scenario with no markings on both sides. Unfortunately, none of the cameras can detect the curbstone stably. Only the leading vehicle can be detected so that its trajectory should be used to generate an ego-lane hypothesis.

18.3.2 Scenario Features

This section explains in detail the composition of the scenario features, which we extract from sensor and context information. In this work, all lane markings as well as the trajectory of the leading vehicle (*ACC object*) are modeled by an approximation of the clothoid model [27]:

$$y(x) \approx \phi_0 \cdot x + \frac{C_0}{2}x^2 + \frac{C_1}{6}x^3 \quad (18.2)$$

$$= a_1 \cdot x + a_2 x^2 + a_3 x^3 \quad (18.3)$$

A subset of the used scenario features is generated from these clothoid parameters, which can be seen in Table 18.1. Additionally, this table contains a likelihood ξ , representing a measure of uncertainty about the existence of an object. Moreover, Table 18.1 also contains the estimated lane width $Lane_w$, the feature *free*, that expresses the amount of free space along the clothoid evaluated with an occupancy grid built by using lidar data. Furthermore, we introduce several consensus features

Table 18.1 Sensor-related and consensus features of all markings and the trajectory of the leading vehicle:
 $h \in \{LM, RM, SLM, SRM, TLM, TRM, ACC\}$

Feature	Description	Feature	Description
$h x_0$	Start on the x-axis	$h y_0$	Lateral offset
$h l$	Clothoid length	$h \phi$	Clothoid angle
$h c_0$	Curvature	$h c_1$	Curvature change
$h \xi$	Existence likelihood	$Lane_w$	Lane width
$CS_h l$	Deviation to l	$CS_h \phi$	Deviation to ϕ
$CS_h c_0$	Deviation to c_0	$CS_h c_1$	Deviation to c_1

Table 18.2 Motion parameters of object $o \in \{Ego, ACC\}$

Feature	Description	Feature	Description
o_x	x-position	o_{v_x}	Longitudinal velocity
o_y	y-position	o_{v_y}	Lateral velocity
o_{v_ϕ}	Yaw rate	o_{vLM}	Velocity to LM
o_{turn}	Tri-state	o_{vRM}	Velocity to RM

measuring the deviations to respective average values. Last, the type of all lane markings is also utilized, e.g., solid, dashed, and curbstone.

For moving objects like the ego-vehicle and the leading vehicle, we extract various motion parameters, as seen in Table 18.2.

Furthermore, we utilize external contextual features extracted from a navigation map. These include *roadType* (e.g., highway, rural, urban, connection), *linkType* (e.g., ramp, roundabout), *laneClass* (e.g., normal, split, merge, intersection), and *cityLimitStatus* (e.g., inside, outside). Additional features are the mean width $\mu_{EgoLaneWidth}$ and the standard deviation $\sigma_{EgoLaneWidth}$ of the ego-lane.

Instead of using these features directly to train ANNs as in [3], we normalize these features and encode them to reach a higher classification performance, which is described in the following section.

18.3.3 Preprocessing Features

If a sensor provides data directly to ANN, the input can suffer from artificial semantic through different ranges and meanings of the data. For example, comparing the *roadType* that are denoted by natural numbers, the distances between two categories are varying even though the semantics are not different. Thereby, the difference between a highway and an urban scenario is equal to the difference between a highway and a rural scenario. Therefore, the distance between these categories should not differ. For that reason, we apply *one-hot encoding* to the categorical input data. By that, a one-hot encoding transforms a categorical feature with n categories into a vector of n entries, where each entry is set to one if the index corresponds to the respective category and to zero otherwise as

$$\text{one-hot} : \{0, 1, \dots, n-1\} \rightarrow \{0, 1\}^n, \text{ one-hot}(k)_i = \begin{cases} 1 & i = k \\ 0 & \text{else} \end{cases} \quad (18.4)$$

Another challenge is the huge variety of ranges in the data set. For instance, the length l of the lane markings can reach up to 100 m, while the angle ϕ varies between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$. Hence, l has a bigger influence on the results until the network learns to reduce its influence by adapting the weights. Therefore, the convergence of the network is slower than the case where alldata is in similar ranges. For that

reason, we apply the following *min-max-scaling* to each feature, so that all values are in the interval $[-1, 1]$ with:

$$\text{scale}(x) = \frac{2 \cdot (x - \min_x)}{\max_x - \min_x} - 1 \quad (18.5)$$

18.4 Reliability Estimation

This section presents the application of ANNs as reliability estimators to the reliability-aware road fusion framework of Nguyen et al. [3, 23]. For this purpose, this section starts by explaining the fusion framework and the model-based ego-lane generation in greater detail. Following, we select for each ego-lane model the most important features, which are obtained by applying the feature selection method mutual information (MI). Afterward, we present the structure and training process of ANNs based on chosen features and introduce different fusion strategies.

18.4.1 Concept

Sections 18.2 and 18.3 clarify the relevance of fusing multiple sources for road estimation. By that, a proper incorporation of reliabilities can leverage the fusion's performance [19]. Thus, we present a multisensor fusion framework, which continuously estimates the sensor reliabilities and uses them to perform the fusion. Adapted from [23], Fig. 18.5 shows different layers of the framework, whereby the contributions of this work are highlighted in green.

At *Layer 0*, different sensor inputs are processed. This preprocessed data is then passed to *Layer 1*, where different types of information are estimated, e.g., lane markings, free space information, vehicles, etc.

At *Level 2*, several hypotheses for the current ego-lane are generated using a model-based approach from Toepfer et al. [1]. Additionally, here we also generate the *scenario features*, which are extracted from sensor detections and contextual information. By way of example, the parameters describing the lane markings are selected, such as the length, the curvature, etc. Moreover, we extend the feature set from [28] with the *consensus features*, which describe the similarity among the lane markings and the driven trajectory of the leading vehicle.

In the offline phase of the *Level 3*, the estimated ego-lane hypotheses are compared with the ground truth, which is represented by the driven trajectory of human drivers. If the deviation from the ground truth exceeds a predefined threshold, the hypothesis will be considered as unreliable and vice versa. Together with the corresponding features, they are stored in a database to train different classifiers. By that, one classifier is trained to predict the reliability of each ego-lane model. During the online phase, each estimated ego-lane is assigned with a predicted reliability from the corresponding classifier.

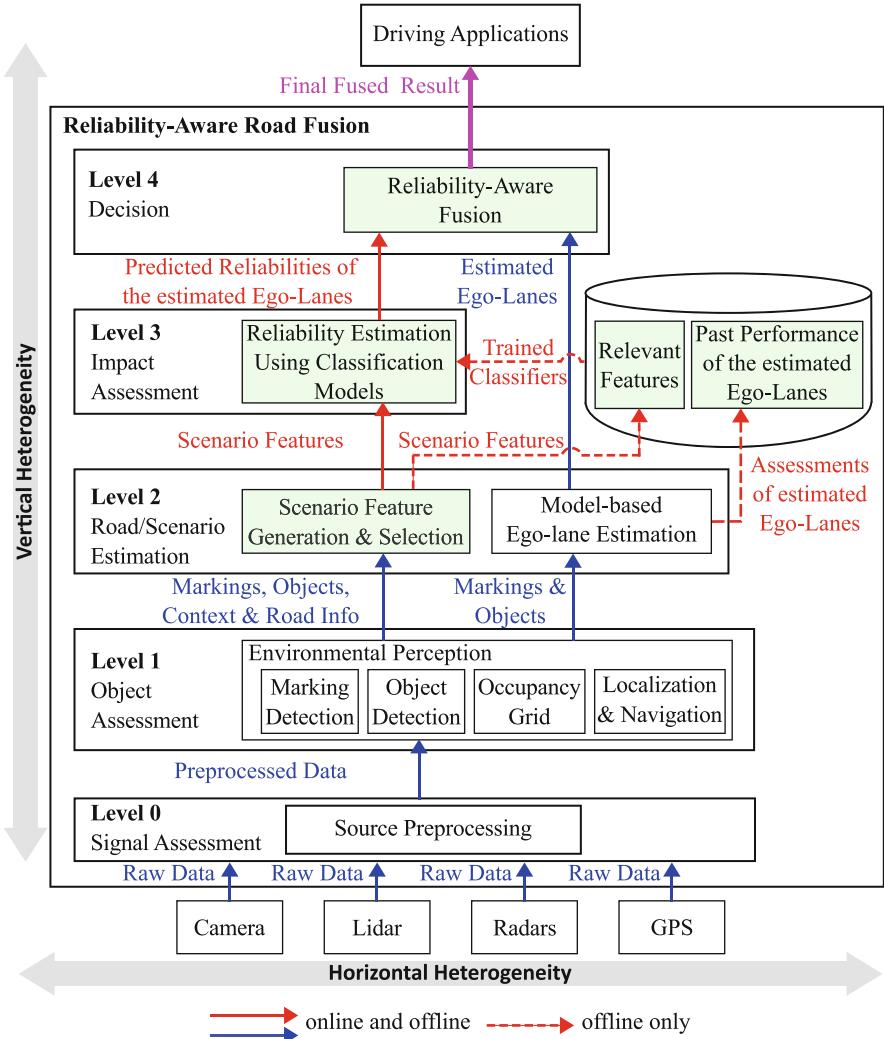


Fig. 18.5 Reliability estimation and reliability-aware fusion as an additional supervision system within the road estimation task [23] (Blue: Data for road detection; Red: Reliability information)

As the last layer, *Level 4* fuses different models depending on the predicted reliabilities. Following, the final ego-lane estimation is then used to perform driving functions.

In this work, we apply *mutual information (MI)* to detect nonlinear relations between the scenario features and the reliability values [29]. Additionally, ANNs are employed as reliability estimators since they perform well in many other tasks and could increase the reliability estimation result [5, 30, 31].

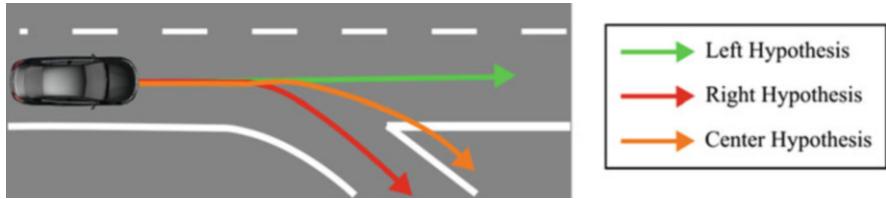


Fig. 18.6 Three estimated ego-lane hypotheses for each camera system [2]

18.4.2 Hypotheses

This section introduces different types of ego-lanes, which are created from lane markings and leading vehicles. Thereby, the detection of lane markings is performed independently for each camera system. In general, the results of each system are used to generate three model-based ego-lane hypotheses (Fig. 18.6). By that, the *left hypothesis (LH)* and the *right hypothesis (RH)* use only the left and right lane markings, respectively. The *center hypothesis (CH)* utilizes the detected lane markings on both sides. By applying this process to the three-camera systems, we can receive up to nine ego-lane estimations. Additionally, the *vehicle hypothesis (VH)* represents the trajectory of the leading vehicle as shown in Fig. 18.4d. This leads to the following set H of hypotheses, where the prefixes “F,” “S,” and “T” indicate the first, second, and third camera system, respectively:

$$H = \{FLH, FRH, FCH, SLH, SRH, SCH, TLH, TRH, TCH, VH\}$$

18.4.3 Feature Selection

Since information from multiple sources is incorporated, the generated feature vector consists of hundreds of elements. Training classifiers with all these features would be computationally expensive, and the results can worsen due to the curse of dimensionality [32]. Moreover, not all features directly affect the reliabilities. Therefore, we perform a feature selection so that only the most relevant features are used to train the classifiers.

For this work, we apply the method *mutual information (MI)*, which is a measure of the dependency between two variables [29]. It is used to determine the information about a variable through another variable. For this purpose, MI is not using the covariance like the linear correlation coefficient but the distance between two probability distributions. Hence, MI can describe nonlinear relationships between two variables. Assuming an independent, identical distribution of a set of N bivariate measurements $\{t_i = (x_i, y_i) \mid i = 1, \dots, N\}$ of the features $X = \{x_1, \dots, x_N\}$ and $Y = \{y_1, \dots, y_N\}$, the mutual information of X and Y is defined as follows:

$$I(X, Y) = \int \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy \quad (18.6)$$

where $p(x, y)$ is the joint probability density and $p(x)$ and $p(y)$ are the marginal probability densities of X and Y , respectively.

Since the densities are not always known, an approach of approximating MI is applied. Therefore, the values of X and Y are sorted into containers of finite sizes, which is described in the following:

$$I_{\text{cont}}(X, Y) = \sum_{ij} p(i, j) \log \frac{p(i, j)}{p(i)p(j)} \quad (18.7)$$

where $p(i, j) = \int_i \int_j p(x, y) dx dy$, $p(i) = \int_i p(x) dx$, and $p(j) = \int_j p(y) dy$. By that, \int_i denotes the integral over container i and \int_j denotes the integral over container j .

The number of entries of each container is counted and

$$p(i) \approx n_x(i)/N \quad (18.8)$$

$$p(i) \approx n_y(i)/N \quad (18.9)$$

$$p(i, j) \approx n(i, j)/N \quad (18.10)$$

are approximated, where $n_x(i)$ and $n_y(i)$ represent the number of entries of the respective container i of X and Y , and $n(i, j)$ denotes the number of overlapping entries. When the number of containers is increased toward to infinity and the size of the containers is aiming toward zero, I_{cont} converges to I .

18.4.4 Training Process

During the offline training phase, the database is divided into training and testing datasets (Fig. 18.7). Afterward, the data is resampled to balance the number of negative and positive samples. As a result, the resampled datasets contain the same amount of samples for both classes to avoid bias during the training [33]. Following, a feature vector \mathbf{X}_h with four different categories for each sample of h is generated as

$$\mathbf{X}_h = [s_h, \tau, \gamma_{\text{int}}, \gamma_{\text{ext}}] \quad (18.11)$$

where s_h describes the sensor information, τ represents the consensus features, and γ_{int} and γ_{ext} denote the internal information (e.g., odometry data) and environment information (e.g., the road type), respectively [23]. After creating \mathbf{X}_h , an error metric is applied to the ego-lane hypothesis h to determine the label L_h . By that, L_h will

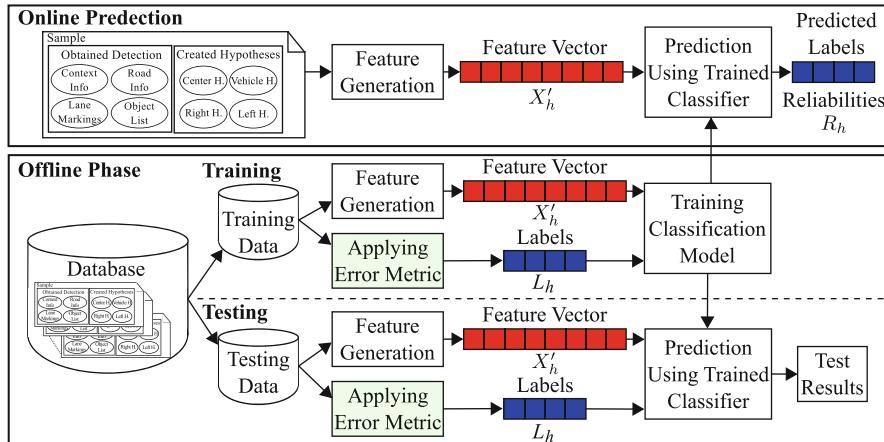


Fig. 18.7 Overview of the application of the classifier [23]

be considered as *reliable* if the deviation of h from the reference is smaller than a predefined threshold. We will explain the used metric in Sect. 18.6.1.

To evaluate the trained networks, the testing dataset is used. Thereby, the created feature vectors are passed directly to the networks and their predictions are compared with the actual test target. The evaluation process and the results are explained in greater detail in Sect. 18.6.

18.4.5 Artificial Neural Networks for Reliability Estimation

In order to estimate the reliability of each hypothesis $h \in H$, we train an ANN ANN_h for each h separately. Thereby, the output of ANN_h represents the estimated reliability R_h . The structure of each individual ANN is shown in Fig. 18.8.

After applying MI to the feature vector \mathbf{X}_h , the 25 most relevant features \mathbf{X}'_h are used as the input for the training. Thereby, these features are then preprocessed by the normalization and one-hot encoding described in Sect. 18.3.3. As a consequence, the processed feature vector can have i elements with $i \geq 25$ due to the one-hot encoding. Since the networks are fully connected, each neuron's input function receives the output from all neurons of the preceding layer.

The next five layers consist only of *rectified linear units (ReLU)*, i.e., they employ $f(x) = \max(x, 0)$ as their activation function. These layers only differ by the number of neurons. Starting with 25 neurons in the first layer, the number is reduced by five for every succeeding layer. The last layer has only one neuron and a sigmoid activation function to produce an output between zero and one, which represents the final reliability value of the corresponding ego-lane model.

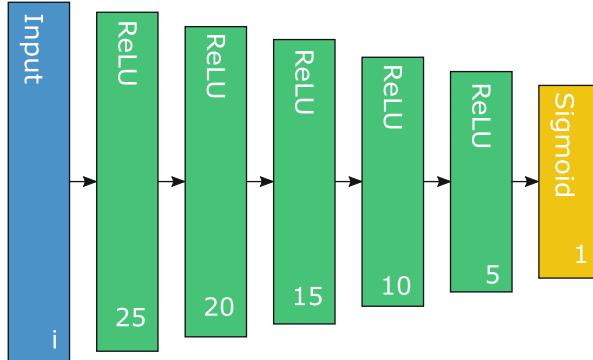


Fig. 18.8 Structure of ANNs toward reliability estimation

During the training, the label vector L_h is compared with the estimation produced by the network to update the weights of the neurons. Since the basic backpropagation algorithm often suffers from contrary training examples and requires a high number of iterations until convergences [32], we apply *Stochastic Gradient Descent (SGD)*, an advanced backpropagation algorithm, to update the weights [34]. Instead of minimizing the total error as the basic backpropagation, SGD minimizes the empirical risk over the training data $D = \{(x_i, y_i) | i = 1, \dots, n\}$ as

$$E(f) = \frac{1}{n} \sum_{i=1}^n l(f(x_i), y_i) \quad (18.12)$$

where l denotes the loss function describing the loss of the prediction $f(x_i)$ regarding the target y_i . In this work, we use the squared Euclidean loss function, which is defined as

$$E_{L2}(f) = \frac{1}{2n} \sum_{i=1}^n \|f(x_i) - y_i\|_2^2 \quad (18.13)$$

For convenience, the loss is divided by two for an easier derivative of the squared Euclidean loss. For an optimal gradient in the learning phase, the gradient has to be calculated in every iteration, which produces a heavy computational effort. Hence, SGD estimates the gradient by using a batch $B \subset D$, which is significantly smaller than D with $|B| \ll |D|$.

$$E_{L2}(f) = \frac{1}{2|B|} \sum_{i=1}^{|B|} \|f(x_i) - y_i\|_2^2 \quad (18.14)$$

By that, Eq. 18.14 describes the risk that is optimized in each iteration. During the weight adaption, the learning rate needs to be decreased to achieve convergence. Although better results can be achieved, the gradient descent can still get stuck in a local minimum of the empirical risk [35]. Therefore, a momentum term is used in the weight adaption, which helps the network to converge faster and leave local minima [36]:

$$\Delta w_{t+1} = \mu \Delta w_t - \alpha \nabla E_{L2}(f) \quad (18.15)$$

where w_t and w_{t+1} are the weights, μ is the momentum, Δw_t is the weight change in step t , and α is the base learning rate. This technique can increase the performance of ANNs as described in [37].

To train the networks, we set the base learning rate $\alpha = 0.1$. Every 100,000 iterations, the learning rate α is multiplied with a factor $\gamma = 0.8$ to support the converging of the networks. In total, each network is trained with 1,000,000 iterations using a batch size of $|B| = 4$. The momentum of the weight change is chosen as $\mu = 0.1$.

18.4.6 Incorporating Reliabilities into Fusion

During the online prediction phase, a feature vector is generated for each ego-lane hypothesis. The trained ANNs take these vectors as input and predict the reliability values, which are used to combine the ego-lanes. Thereby, the quality of the fusion is restricted by the quality of the source [16]. However, different fusion strategies create results with varying quality. Therefore, this section presents several basic strategies and a more complex strategy based on Dempster-Shafer theory [3, 20].

18.4.6.1 Basic Strategies

Following, we introduce several basic fusion strategies:

Baseline (BE) The standard road estimation approach from [1] serves as a baseline strategy.

Average fusion (AVG) By this strategy, every estimation model is equally involved in the fusion. This is one of the easiest approaches, but AVG will not produce the best results because inferior models can impair the fused result.

Weight-based fusion (WBF) As an extension of AVG, the reliability R_h of every model h can be utilized as weight for the fusion. Using R_h allows to disregard unreliable models and focus on the combination of the remaining reliable models.

Winner-take-all (WTA) WTA selects solely the ego-lane model with the greatest R_h , and all other hypotheses are discarded.

Minimum (MIN) The ego-lane model with the smallest R_h , i.e., the most unreliable one, is chosen for the fusion. This strategy is needed to prove that the unreliable sources can be identified by the classifiers and assigned with lower reliabilities.

Random (RAN) As an additional baseline, *RAN* chooses a hypothesis arbitrarily.

18.4.6.2 Dempster-Shafer Theory (DST)

The theory of belief functions was developed by Dempster and Shafer in [38]. Its application is the combination of several unreliable sources to a total result which often occurs in reality. As introduced by Nguyen et al. [3], the reliability of each ego-lane hypothesis h can be modeled as a *frame of discernment* $\Theta_h = \{\rho_h, \bar{\rho}_h\}$, which consists of two statements *Reliable* ρ_h and *Unreliable* $\bar{\rho}_h$. The following steps are taken under the assumption that the reliabilities of the ego-lane models are independent [19]. Since DST also models a belief function for the situation, where both states of Θ_h can occur, adding ρ_h and $\bar{\rho}_h$ does not have to result in one as compared to the classical probability theory. This is difficult to represent using the Bayesian probabilistic model. As a consequence, the power set Φ_h for a hypothesis h is defined as:

$$\Phi_h = 2^{\Theta_h} = \{\emptyset, \{\rho_h\}, \{\bar{\rho}_h\}, \{\rho_h, \bar{\rho}_h\}\} \quad (18.16)$$

where $\{\rho_h, \bar{\rho}_h\}$ describes the occurrence of both possibilities. The mass function for the reliability of the model h at time t is defined as follows:

$$\begin{aligned} \sum_{\theta \in \Phi} m^t(\theta) &= 1 \text{ with } m^t(\emptyset) = 0, \quad m^t(\{\rho_h\}) = R_h^t \cdot PR_h \\ m^t(\{\bar{\rho}_h\}) &= (1 - R_h^t) \cdot PR_h \quad m^t(\{\rho_h, \bar{\rho}_h\}) = 1 - PR_h \end{aligned} \quad (18.17)$$

where PR_h represents the precision of the neural network ANN_h , which estimates the reliability R_h of h . By that, PR_h is determined by evaluating the classifier ANN_h offline using test data. Assuming two different times t and t' are both independent, the fusion of m^t and m^{t+1} is defined as:

$$m_F(z) = m^t \otimes m^{t+1}(z) = \frac{\sum_{x,y \subseteq \Phi, x \cap y = z} m^t(x) \cdot m^{t+1}(y)}{1 - \sum_{x,y \subseteq \Phi, x \cap y = \emptyset} m^t(x) \cdot m^{t+1}(y)} \quad (18.18)$$

Every hypothesis' reliability consists of two parts. The *belief* b_F and the *plausibility* pl_F . The first describes the belief in the correctness of the hypothesis and the second the plausibility of the hypothesis:

$$b_F(\{\rho_h\}) = \sum_{X \subseteq \{\rho_h\}, X \neq \emptyset} m_F(X) = m_F(\{\rho_h\}) \quad (18.19)$$

$$pl_F(\{\rho_h\}) = \sum_{\rho_h \in X} m_F(X) = m_F(\{\rho_h\}) + m_F(\{\rho_h, \bar{\rho}_h\}) \quad (18.20)$$

To compare the estimated R_h of each hypothesis, the average of *belief* and *plausibility* is used, like in [23, 39]:

$$p_F(\{\rho_h\}) = \frac{b_F(\{\rho_h\}) + pl_F(\{\rho_h\})}{2} \quad (18.21)$$

Using $p_F(\{\rho_h\})$ as the weight for the respective hypothesis and a predefined threshold ϵ_R , only the most reliable hypotheses are allowed to take part in the fusion.

Instead of an explicit reliability estimation, the next section will describe another fusion approach, which estimates the ego-lane directly by using sensors detections.

18.5 Ego-Lane Estimation Using Artificial Neural Networks

18.5.1 Concept

An alternative approach for ego-lane estimation can be performed with artificial neural networks, whose architecture is shown in Fig. 18.9. Hereby, *Level 0*, *Level 1*, and *Level 2* are analogous to the reliability estimation process in Sect. 18.4. By using the generated scenario features from *Level 2*, we apply ANNs as *regressors* to estimate the clothoid parameters of the ego-lane at *Level 3* and *Level 4*. Thereby, we create the training data by taking the human-driven path as a reference, which Sect. 18.5.2 will explain in detail. Moreover, we will present the network structure in Sect. 18.5.3 and the training procedure in Sect. 18.5.4.

18.5.2 Ground Truth Acquisition

An important task is creating the reference data, which is used as targets to train ANNs. For that reason, we use real-world data recordings provided by the test vehicle to determine the necessary coefficients. During a local simulation of the recordings, the positions and the orientations of the vehicle are saved in a database by reconstructing the human-driven path (Fig. 18.10). For time t , the reference is created by an approximation of a clothoid using the points p_0, \dots, p_n . By that, p_0 represents the current vehicle position at time t and p_1, \dots, p_n the vehicle positions at time $t+1, \dots, t+n$. Therefore, the consecutive points p_1, \dots, p_n are rotated and translated in the coordinate system of p_0 . As a result, the reference ego-lane is represented by an approximation of a clothoid [27]:

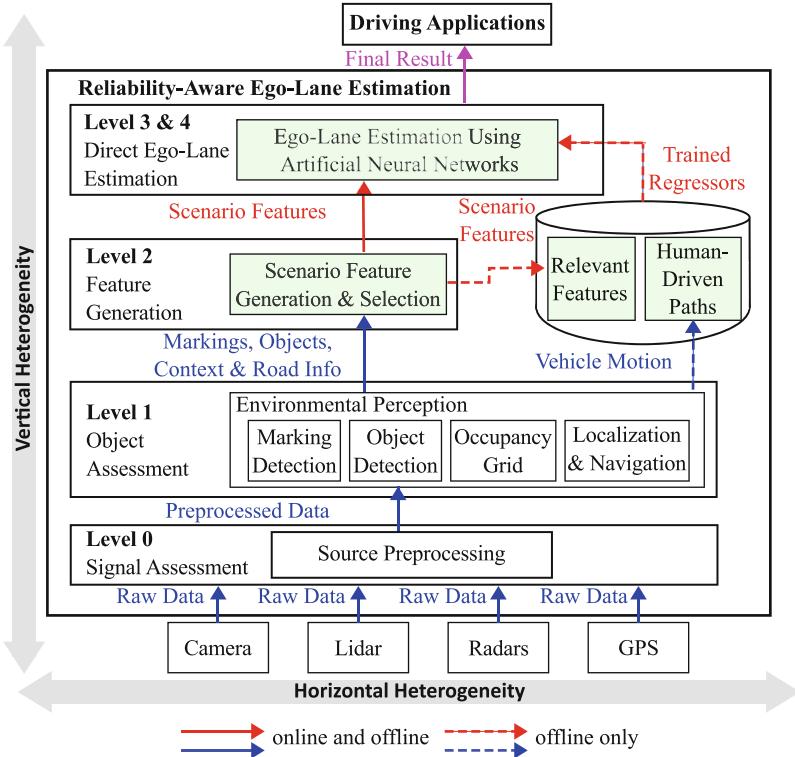


Fig. 18.9 Direct ego-lane estimation using artificial neural networks (Blue: sensor information; Red: reliability information)

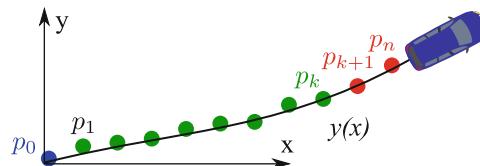


Fig. 18.10 The ground truth at p_0 is acquired by using linear polynomial regression of p_0, \dots, p_k . The points p_{k+1}, \dots, p_n are available, but they are left out due to exceeding the maximal distance or angle

$$y(x) \approx \phi_0 \cdot x + \frac{C_0}{2}x^2 + \frac{C_1}{6}x^3 \quad (18.22)$$

$$= a_1 \cdot x + a_2 x^2 + a_3 x^3 \quad (18.23)$$

We determine a_1 , a_2 and a_3 by applying linear polynomial regression. Therefore, we construct the following linear system using the consecutive points.

$$\mathbf{y} = \begin{bmatrix} y_{p_1} \\ \vdots \\ y_{p_n} \end{bmatrix} = \begin{bmatrix} x_{p_1} & x_{p_1}^2 & x_{p_1}^3 \\ \vdots & \vdots & \vdots \\ x_{p_n} & x_{p_n}^2 & x_{p_n}^3 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \mathbf{X} \cdot \mathbf{a} \quad (18.24)$$

Next, this system is solved by using Moore-Penrose inverse regression [40] since \mathbf{X} is most of the time not invertible. For that reason, the parameters can be calculated by using

$$\mathbf{a} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (18.25)$$

Consequently, the coefficients are $\phi_0 = a_1$, $C_0 = 2a_2$, and $C_1 = 6a_3$. Basically, this process could be applied to all consecutive points p_1, \dots, p_n in the recording, but this is neither representative for an estimation nor applicable due to the computational effort. Hence, we reduce the number of points by choosing a subset of only k consecutive points as

$$\left\{ p_k \mid \sum_{i=1}^k \text{distance}(p_{i-1}, p_i) < 50 \wedge |\text{direction}(p_k)| < 15^\circ \right\} \quad (18.26)$$

First, only the first k points that are less than 50 m away from the start point p_0 are selected. Secondly, the orientation of these points has to be smaller than 15° to achieve a sufficient approximation by the polynomial.

Since the manually-driven path is used to calculate the targets to train ANNs, we have to remove samples/situations where the driver leaves the current ego-lane. For example, such samples can be obtained by intersections, lane change, overtaking maneuvers, etc. Additionally, the samples that do not contain any information about the road course are also removed since ANNs cannot produce any useful estimation in such scenarios.

18.5.3 Structure

An important decision is the choice of a structure for ANNs. We also decide to use one network for each parameter to preserve expressiveness. Each ANN has seven layers consisting of a decreasing amount of neurons as displayed in Fig. 18.11. The first layer contains 80, the second 60, the third 40, the fourth 20, the fifth 10, the sixth 5, and the last 1 neuron. All layers except the last layer consist of *rectified linear units (ReLU)*. The last layer has the identity function as activation function to enable an output of arbitrary real numbers. We choose this structure since the layers using ReLU can deal with not linearly separable data. Additionally, the layers decrease in the number of neurons to generalize the scenario features in small steps.

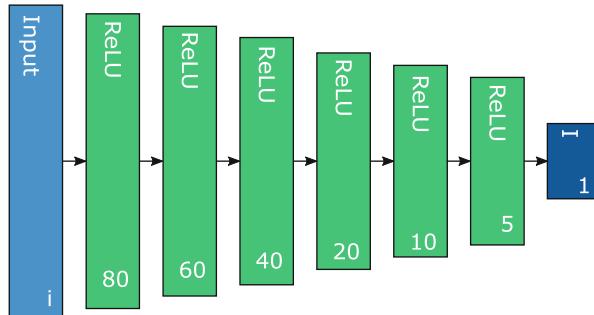


Fig. 18.11 Structure of the ANN to estimate each single parameter of the clothoid model. The text describes the activation function in the layer, where ReLU denotes a layer of rectified linear units and I is the identity function

18.5.4 Training

Analogously to the reliability estimation, we use the stochastic gradient descent from Sect. 18.3 with an Euclidean loss for 100,000 iterations. Using this technique, a learning rate of $\alpha = 0.0001$ that decreases to 0.9 times itself every 10,000 iterations is chosen. Additionally, we chose a momentum weight of $\mu = 0.0000001$ that is multiplied by 0.01 after the same amount of iterations. Moreover, we use a batch size of $|B| = 25$.

During this process, we scale the learning targets by multiplying them by 10000. Hereby, the real appearing value range becomes bigger, so that the impact of the gradient is bigger and leads to faster convergence. Furthermore, the training data set is resampled regarding the *roadType*, so that the trained networks can perform well in each category.

18.6 Experimental Results

In this section, we use real-world data recordings to evaluate our introduced fusion concepts. Figure 18.12 shows the routes, where the prototype vehicle drove in Wolfsburg and its surroundings. Thereby, we planned our routes in order to archive a balanced distribution of highway, ramp, rural and urban scenarios.

First, we will present the evaluation concept. Following, the impact of the feature selection with *mutual information* is analyzed. Afterward, the reliability estimation and the final performance of both fusion concepts are presented.



Fig. 18.12 Driven roads for recording training and testing data

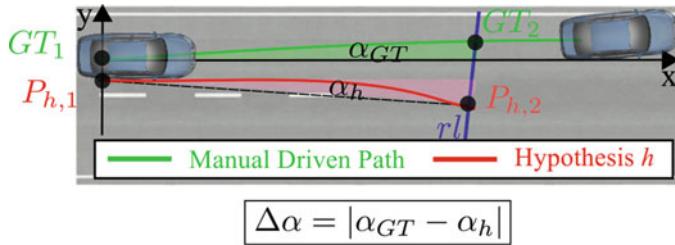


Fig. 18.13 Metric to measure reliability of the estimated ego-lanes [28]

18.6.1 Concept

In the following, we use the angle metric presented by Nguyen et al. [28] to assess the reliability of the estimated ego-lanes. Instead of using highly-precise DGPS and digital map as the authors in [1, 15, 26], this metric incorporates the human-driven path as reference, which can be reconstructed with standard and cheap motion sensors. As shown in Fig. 18.13, this metric measures the angle deviation $\Delta\alpha$ between the estimated lane and the manually driven path for a run length rl starting from the position of the ego-vehicle at time t . The motivation for this metric is because human drivers cannot drive perfectly on the lane centerline during recording data. This leads always to small lateral offsets between the estimation and the reference, even when the estimation could be detected perfectly [2]. By using the angle deviation $\Delta\alpha$, only the parallelism between the hypotheses and the driven path is taken into account.

As Fig. 18.13 shows, the angle α_h of the hypothesis h is calculated using the position $P_{h,2} = (x_{h,2}, y_{h,2})$ at the run length rl and its start position $P_{h,1} = (x_{h,1}, y_{h,1})$. The ground truth is reconstructed by using the human-driven path,

where $GT_1 = (0, 0)$ represents the ego-vehicle's position at time t and $GT_2 = (x_{GT,2}, y_{GT,2})$ denotes the position at time t' with $t' \gg t$. In other words, GT_2 represents the position of the vehicle after driving rl meters. As a result, the angle difference can be calculated as

$$\Delta\alpha = \left| \arctan\left(\frac{y_{h,2} - y_{h,1}}{x_{h,2} - x_{h,1}}\right) - \arctan\left(\frac{y_{GT,2}}{x_{GT,2}}\right) \right| \quad (18.27)$$

By that, we consider an ego-lane estimation as *reliable* if its angle deviation is smaller than 2° for $rl = 30$ m. For the sake of completeness, we also use the lateral offset $\Delta d = |y_{GT,2} - y_{h,2}|$ as another criterium when evaluating the hypotheses to be comparable with related works.

18.6.2 Result of Feature Selection

This section discusses the results of feature selection with *mutual information* (*MI*) by showing the 10 highest ranked features for each hypothesis in Figs. 18.14

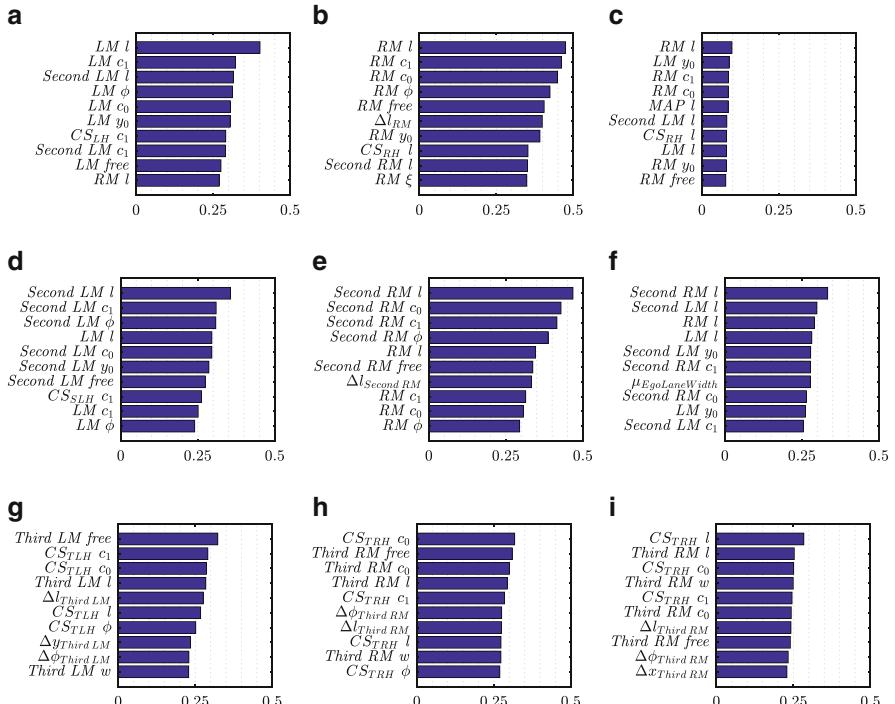
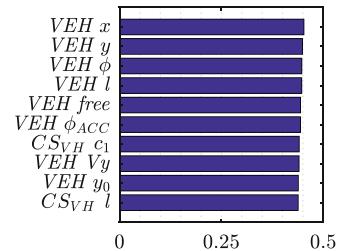


Fig. 18.14 The 10 highest ranked features for ego-lane models, which are generated by using lane markings. (a) LH. (b) RH. (c) CH. (d) SLH. (e) SRH. (f) SCH. (g) TLH. (h) TRH. (i) TCH

Fig. 18.15 The 10 highest ranked features for VH



and 18.15. Thereby, LM denotes the left ego-lane marking and RM the right ego-lane marking respectively. The prefix *Second* and *Third* represents the camera system, where the lane marking is coming from. Moreover, VEH denotes features of the ACC object.

As shown in Fig. 18.14, the length l , the curvature c_0 , the curvature change c_1 , and the yaw angle ϕ of the lane markings are very important for the ego-lane models, which are created by involving lane markings. Besides, one notices the distinct difference between the hypotheses, which only use the right or the left lane marking. For LH , SLH , and TLH , the most important features come from the corresponding left ego-lane markings and some of the consensus features. For RH , SRH , and TRH , only features concerning the right lane markings and features belonging to these hypotheses are ranked as important. The only exception can be found for LH .

Figure 18.14a–f shows that the features of the lane markings from the first camera and the second camera are sometimes mixed for the hypotheses LH , RH , CH , SLH , SRH , and SCH . The reason is because of the similar characteristics and the installation positions of both cameras. In contrast, only the lane markings received from the third camera and their belonging features are important for TLH , TRH , and TCH due to the different field of view.

Figure 18.15 shows that almost all features acquired from the leading vehicle are very relevant for VH . It is also interesting and correct that none of the marking information can be found here.

In summary, the main impact on the reliabilities of the hypotheses comes from the according detection source. Hence, a reliability estimator can be trained by using only the data of the corresponding detections. Furthermore, the observation that the first and second camera features are correlated indicates a strong redundancy between the cameras. For the evaluation of the classifiers, the neural networks from Sect. 18.4.5 are trained using the 25 highest ranked features.

18.6.3 Result of Reliability Estimation

To measure the classifier's performance, we use the $F_{0.8}$ -Score which is defined as

$$F_\beta = \frac{(1 + \beta^2) \cdot PR \cdot RC}{(\beta^2 \cdot PR) + RC} \quad (18.28)$$

where $PR = TP/(TP + FP)$ is the precision and $RC = TP/(TP + FN)$ is the recall. Moreover, TP denotes the number of true positive samples, FP the number of false positive samples and FN the number of false negative samples. The motivation of using a $F_{0.8}$ -Score is that we want to increase the impact of the precision on the result and penalize false positives more than false negatives, since automated driving is a safety-critical application. The higher the $F_{0.8}$ -Score, the better a classifier.

Figure 18.16 shows the classification results of ANN when predicting the reliability for the ten hypotheses. Since we perform a down-sampling on the evaluation data, there are the same numbers of reliable and unreliable samples. This is indicated by the *maximum availability*, i.e., the amount of positive samples over all samples, which is equal to 0.5 in most cases. Only for *FLH*, *FRH*, *SRH* and *SCH* from highway scenarios no down-sampling is needed, since all samples are positive. However, ANN estimates some hypotheses, such as *FRH*, *FCH*, *SLH* and *TCH*, to be reliable for all samples. This leads to a low $F_{0.8}$ -Score of around 0.7.

For highway scenarios, the hypotheses *FLH*, *FRH*, *SRH*, and *SCH* have the best performance of around 100% (Fig. 18.16b). Following this, the performance for

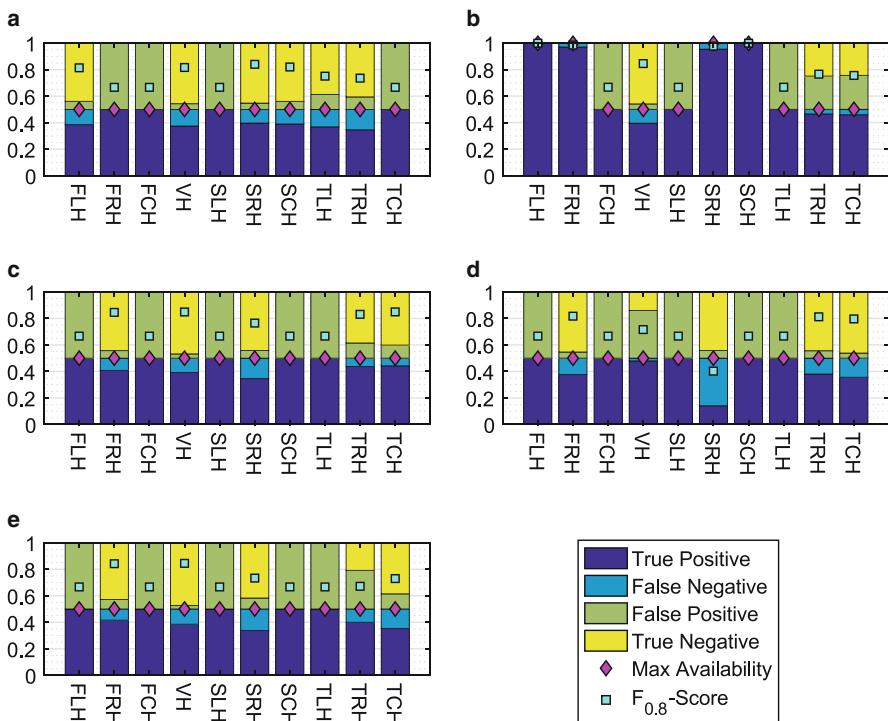


Fig. 18.16 Classification performance of ANN when predicting our ten hypotheses in different scenarios. **(a)** Overall. **(b)** Highway. **(c)** Rural. **(d)** Urban. **(e)** Connections

VH is about 80%. Moreover, ANN also performs well for *VH* in other scenarios. For rural scenarios in Fig. 18.16c, the classification performances for the ego-lane models based on the right lane markings are around 85%, which is better than the results of models based on the left lane markings. The center hypotheses *FCH* and *SCH*, which incorporate both lane markings, cannot improve their estimation result using the right markings and have the same bad classification performance as the left models *FLH* and *SLH*. Only *TCH* can make more use of both markings and is therefore ranked better than *FCH* and *SCH*. In urban scenarios, the performances of all classifiers decrease due to the variety of situations, where markings are sometimes not existing (Fig. 18.16d). In connection scenarios, all classifiers perform worse since the front-facing cameras cannot detect markings well here due to the narrow fields of view. It can be seen that *VH* has the highest performance in connection scenarios.

By using ANN as reliability estimators, the next section will evaluate the results of different fusion strategies from Sect. 18.4.6 and compare them with the direct ego-lane estimation approach from Sect. 18.5.

18.6.4 Result of Ego-Lane Estimation

The final estimation results are compared using two metrics: the commonly used lateral offset and the angle deviation from [28]. Both metrics are applied to the hypotheses at different run lengths to investigate the estimation quality both in close distance and in far distance to the vehicle.

As a general observation from Fig. 18.17, over 75% of the samples of each fusion strategy reach an angle difference of $\Delta\alpha < 2^\circ$ and a lateral offset of $\Delta d < 1$ m. In the following, ANN denotes the fusion concept, where ANNs are used to directly estimate the parameters of the ego-lane. By comparing ANN with different fusion approaches, ANN turns out to perform well regarding short distances (Fig. 18.17a, b). However, both metrics agree that the error of ANN increases significantly as the distance grows. Thus, all fusion approaches outperform ANN after a run length of 28 m (Fig. 18.17g, h). The reasons for these results are the two design decisions when using ANNs for the direct ego-lane estimation process in Sect. 18.5. First, the usage of the polynomial for the ground truth acquisition induces an error, which is especially great in strong curves because the assumption of an angle below 15° does not hold. Second, the representation as a polynomial has the disadvantage of highly amplifying small mistakes in the estimation. For instance, if the ideal parameters are denoted by ϕ_0 , C_0 and C_1 and the estimated parameters are denoted by $\tilde{\phi}_0$, \tilde{C}_0 and \tilde{C}_1 , each estimated parameter can be written as

$$\tilde{\phi}_0 = \phi_0 + \epsilon_{\phi_0} \quad (18.29)$$

$$\tilde{C}_0 = C_0 + \epsilon_{C_0} \quad (18.30)$$

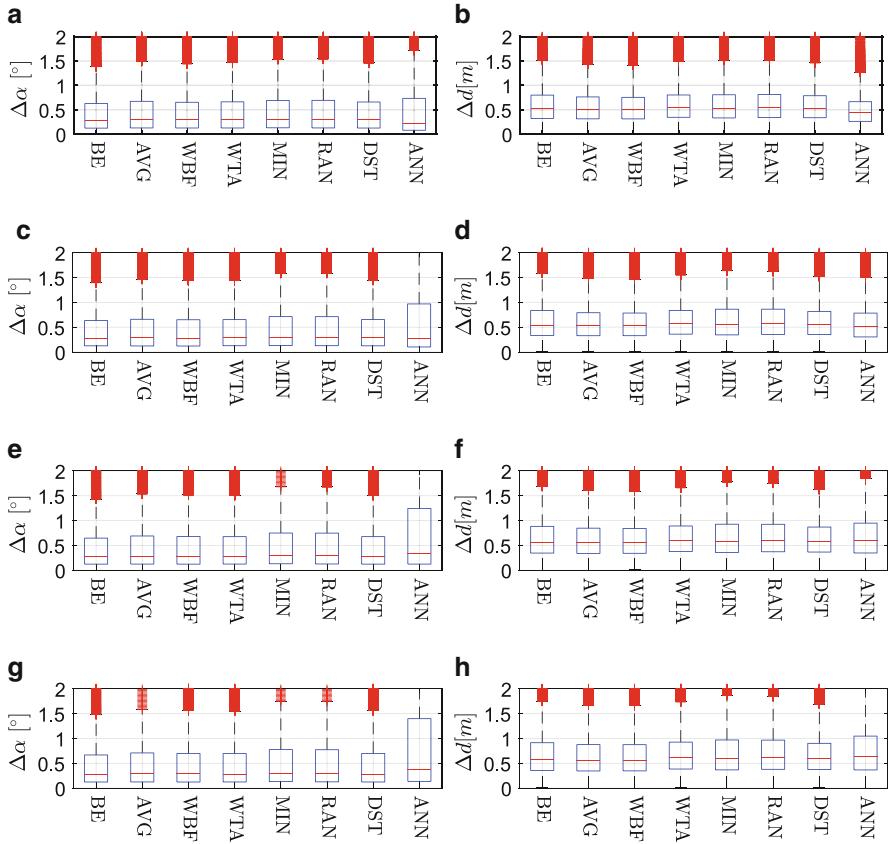


Fig. 18.17 Performance of different fusion strategies and ANN measured by the angle deviation $\Delta\alpha$ and the lateral offset Δd to the ground truth at various distances. We excluded all samples with no reliable hypothesis. **(a)** Angle deviation at 16 m. **(b)** Lateral offset at 16 m. **(c)** Angle deviation at 22 m. **(d)** Lateral offset at 22 m. **(e)** Angle deviation at 28 m. **(f)** Lateral offset at 28 m. **(g)** Angle deviation at 31 m. **(h)** Angle deviation at 31 m

$$\tilde{C}_1 = C_1 + \epsilon_{C_1} \quad (18.31)$$

where ϵ_p denotes the error in the estimation of parameter p . Next, the impact of the estimation error can be determined as the absolute error

$$e_{\text{abs}} = \left\| \phi_0 \cdot x + \frac{C_0}{2}x^2 + \frac{C_1}{6}x^3 - \left((\phi_0 + \epsilon_{\phi_0}) \cdot x + \frac{C_0 + \epsilon_{C_0}}{2}x^2 + \frac{C_1 + \epsilon_{C_1}}{6}x^3 \right) \right\| \quad (18.32)$$

$$= \left\| \epsilon_{\phi_0} \cdot x + \frac{\epsilon_{C_0}}{2}x^2 + \frac{\epsilon_{C_1}}{6}x^3 \right\| \quad (18.33)$$

If the error at a distance of 31 m is considered and $\epsilon_{\phi_0} = 0$ and $\epsilon_{C_0} = 0$, the absolute error e_{abs} is $\left\| \left(4965 + \frac{1}{6} \right) \epsilon_{C_1} \right\|$. Hence, an error in C_1 greater than 0.00021 ($\approx 0.012^\circ$) leads to a lateral offset of more than one meter. Analogously, the error in C_0 has a significant impact. For that reason, small errors in estimations can lead to poor performance of ANN with increasing run lengths.

To evaluate the final performance of the fusion strategies and ANN, we use the *availability* (AV), which is given by the proportion of samples with a correct ego-lane estimation over all samples [23]. By that, a strategy is considered as available only when the following conditions are fulfilled. First, the strategy provides an estimate for the given sample. Secondly, the angle deviation $\Delta\alpha$ of the provided estimate must not exceed 2° .

For highways and rural roads, Fig. 18.18b, c show that the performances of all strategies are near 100% because of good road conditions in these scenarios. As expected, the performances of all strategies are lower in urban areas (Fig. 18.18d) due to the variety of situations. In on- and off-ramp scenarios, all strategies have their lowest availability (Fig. 18.18e). Compared to BE from [1], our fusion can enable an increase of up to 5 percentage points regarding the availability.

Furthermore, ANN has the lowest availability for all scenarios with the exception of connections. The overall low availability is expected looking at the performance regarding $\Delta\alpha$. Unfortunately, ANN performs even worse considering that AV is mostly smaller than MIN, which selects the hypothesis with the lowest reliability. In contrast, ANN achieves the best performance in connection scenarios. This is due to the weaker dependence on lane markings, which are hard to detect in curves. Hence, ANN can comprehend the lack of lane marking detection. Moreover, the results could be improved by using a different representation that suffers less from

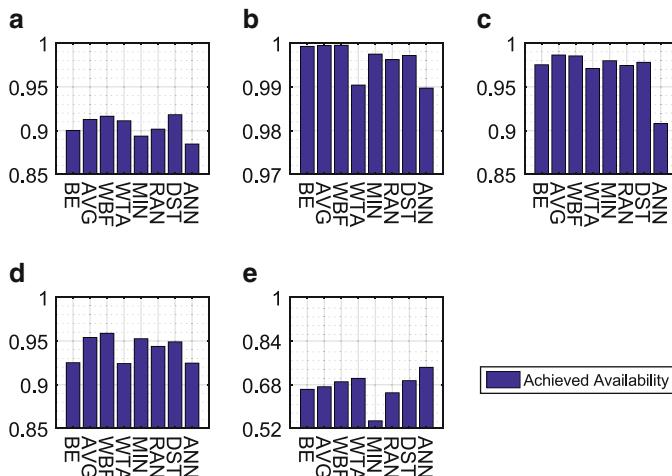


Fig. 18.18 Comparison of the achieved availability of different ego-lane estimation models in different scenarios. (a) Overall. (b) Highways. (c) Rural. (d) Urban. (e) Connection

errors in the prediction and the ground truth acquisition. As a consequence, the results could be used to improve the performance especially in curve scenarios.

18.7 Conclusion

In this work, we present two fusion concepts for ego-lane estimation by using multiple sensors and neural networks. The first approach estimates for each source a reliability value, which indicates whether the source is correct for the current situation or not. Based on the predicted reliabilities, the fusion will prefer reliable sources over unreliable sources, such as by giving greater weights to reliable hypotheses or by excluding unreliable sources from the fusion. Instead of explicitly estimating reliabilities, the second approach uses neural networks to directly estimate the ego-lane. Thereby, the reliabilities are internally learned and encoded as weights of the neurons. Compared to a standard road estimation approach from [1], our approach can increase the availability by up to 5 percentage points.

In future work, we want to improve both fusion concepts by changing the net structure and utilize different structures for different hypotheses and parameters respectively. Additionally, a further improvement of the feature selection needs to be done by comparing the performance of the same classifier using different features. The direct ego-lane estimation performs slightly worse than the results of other fusion strategies regarding the angle deviation and availability. However, the performance in connection scenarios is better than all other fusion approaches. For that reason, a possible use for ANNs would be to incorporate the estimation into the fusion framework and improve the performance in connection scenarios. When training ANNs, we found that the representation of the targets as an approximation of a clothoid is not appropriate due to the large amplification of errors in the estimation. Hence, a scalar field could be used instead, where the values above and below a threshold represent the lane. Furthermore, we plan to improve both neural network approaches by incorporating temporal information and using recurrent neural networks. This can lead to more sufficient estimations in all scenarios.

References

1. D. Töpfer, J. Spehr, J. Effertz, C. Stiller, Efficient scene understanding for intelligent vehicles using a part-based road representation, in *IEEE Conference on Intelligent Transportation Systems* (2013), pp. 65–70. <https://doi.org/10.1109/ITSC.2013.6728212>
2. T.T. Nguyen, J. Spehr, M. Uhlemann, S. Zug, R. Kruse, Learning of lane information reliability for intelligent vehicles, in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems* (2016), pp. 142–147. <https://doi.org/10.1109/MFI.2016.7849480>
3. T.T. Nguyen, J. Spehr, J. Xiong, M. Baum, S. Zug, R. Kruse, Online reliability assessment and reliability-aware fusion for ego-lane detection using influence diagram and Bayes filter, in *IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems* (2017), pp. 7–14

4. T.T. Nguyen, J. Spehr, S. Zug, R. Kruse, Multi-source fusion for robust road detection using online estimated reliabilities. *IEEE Trans. Indus. Inf.* 1 (2018). <https://doi.org/10.1109/TII.2018.2865582>
5. C. Chen, A. Seff, A. Kornhauser, J. Xiao, DeepDriving: learning affordance for direct perception in autonomous driving, in *IEEE International Conference on Computer Vision* (2015), pp. 2722–2730
6. D.A. Pomerleau, Efficient training of artificial neural networks for autonomous navigation. *Neural Comput.* 3(1), 88–97 (1991). <https://doi.org/10.1162/neco.1991.3.1.88>
7. M. Bojarski, D.D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L.D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, K. Zieba, End to end learning for self-driving cars (2016). CoRR abs/1604.07316
8. Z. Chen, X. Huang, End-to-end learning for lane keeping of self-driving cars, in *2017 IEEE Intelligent Vehicles Symposium (IV)* (2017), pp. 1856–1860. <https://doi.org/10.1109/IVS.2017.7995975>
9. F. Codevilla, M. Müller, A. Dosovitskiy, A. López, V. Koltun, End-to-end driving via conditional imitation learning (2017). CoRR abs/1710.02410
10. A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42(3), 145–175 (2001). <https://doi.org/10.1023/A:1011139631724>
11. M. Al-Qizwini, I. Barjasteh, H. Al-Qassab, H. Radha, Deep learning algorithm for autonomous driving using googlenet, in *2017 IEEE Intelligent Vehicles Symposium (IV)* (2017), pp. 89–96. <https://doi.org/10.1109/IVS.2017.7995703>
12. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions (2014). ArXiv e-prints
13. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition (2014). CoRR abs/1409.1556
14. M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks (2013). CoRR abs/1311.2901
15. O. Hartmann, M. Gabb, R. Schweiger, K. Dietmayer, Towards autonomous self-assessment of digital maps, in *Proceedings of the IEEE Intelligent Vehicles Symposium* (2014), pp. 89–95. <https://doi.org/10.1109/IVS.2014.6856564>
16. G.L. Rogova, V. Nimier, Reliability in information fusion: literature survey, in *7th International Conference On Information Fusion* (2004), pp. 1158–1165
17. T. Brade, S. Zug, J. Kaiser, Validity-based failure algebra for distributed sensor systems, in *IEEE International Symposium on Reliable Distributed Systems* (2013), pp. 143–152. <https://doi.org/10.1109/SRDS.2013.23>
18. H. Frigui, L. Zhang, P. Gader, Context-dependent multi-sensor fusion for landmine detection, in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium* (2008), pp. II–371–II–374. <https://doi.org/10.1109/IGARSS.2008.4779005>
19. T.T. Nguyen, J. Spehr, J.-O. Perschewski, F. Engel, S. Zug, R. Kruse, Zuverlässigkeitssbasierte Fusion von Fahrstreifeninformationen für Fahrerassistenzfunktionen, in *Proceedings 27. Workshop Computational Intelligence*, ed. by F. Hoffmann, E. Hüllermeier, R. Mikut (KIT Scientific Publishing, Karlsruhe, 2017), pp. 33–49
20. T.T. Nguyen, J. Spehr, J. Sitzmann, M. Baum, S. Zug, R. Kruse: Improving ego-lane detection by incorporating source reliability, in *Multisensor Fusion and Integration in the Wake of Big Data, Deep Learning and Cyber Physical System*, ed. by S. Lee, H. Ko, S. Oh. Lecture Notes in Electrical Engineering, vol. 501 (Springer International Publishing, Cham, 2018)
21. M. Realpe, B.X. Vintimilla, L. Vlacie, A fault tolerant perception system for autonomous vehicles, in *Proceedings of the 35th Chinese Control Conference* (2016), pp. 6531–6536. <https://doi.org/10.1109/ChiCC.2016.7554385>
22. A. Rechy Romero, P.V. Koerich Borges, A. Elfes, A. Pfrunder, Environment-aware sensor fusion for obstacle detection, in *Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems* (2016), pp. 114–121. <https://doi.org/10.1109/MFI.2016.7849476>

23. T.T. Nguyen, J. Spehr, D. Vock, M. Baum, S. Zug, R. Kruse, A general reliability-aware fusion concept using DST and supervised learning with its applications in multi-source road estimation, in *2018 IEEE Intelligent Vehicles Symposium (IV)* (2018), pp. 597–604
24. F.E. White, A model for data fusion, in *Proceedings of the First National Symposium on Sensor Fusion* (1988)
25. C. Gackstatter, P. Heinemann, S. Thomas, B. Rosenhahn, G. Klinker: Fusion of clothoid segments for a more accurate and updated prediction of the road geometry, in *13th International IEEE Intelligent Transportation Systems Conference* (2010), pp. 1691–1696. <https://doi.org/10.1109/ITSC.2010.5625270>
26. T.T. Nguyen, J. Spehr, H. Lin, D. Lipinski, Fused raised pavement marker detection using 2D-Lidar and mono camera, in *IEEE International Conference on Intelligent Transportation Systems* (2015), pp. 2346–2351
27. E.D. Dickmanns, B.D. Myslivetz, Recursive 3-D road and relative ego-state recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **14**(2), 199–213 (1992). <https://doi.org/10.1109/34.121789>
28. T.T. Nguyen, J. Spehr, J. Xiong, M. Baum, S. Zug, R. Kruse, A survey of performance measures to evaluate ego-lane estimation and a novel sensor-independent measure along with its applications, in *IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems* (2017), pp. 239–246
29. A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information. *Phys. Rev. E Stat. Nonlinear Soft Matt. Phys.* **69**(6 Pt 2), 066138 (2004). <https://doi.org/10.1103/PhysRevE.69.066138>
30. J. Pradeep, E. Srinivasan, S. Himavathi, Diagonal based feature extraction for handwritten character recognition system using neural network, in *2011 3rd International Conference on Electronics Computer Technology (ICECT)* (2011), pp. 364–368. <https://doi.org/10.1109/ICECTECH.2011.5941921>
31. A. Graves, A.R. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2013), pp. 6645–6649. <https://doi.org/10.1109/ICASSP.2013.6638947>
32. R. Kruse, C. Borgelt, C. Braune, S. Mostaghim, M. Steinbrecher, *Computational Intelligence: A Methodological Introduction*. Texts in Computer Science, 2nd edn./2016 edn. (Springer, London, 2016)
33. C.M. Bishop, *Pattern Recognition and Machine Learning*. Information Science and Statistics (Springer, New York, 2006)
34. L. Bottou, Stochastic gradient descent tricks, in *Neural Networks: Tricks of the Trade*, ed. by G. Montavon, G.B. Orr, K.R. Müller, 2nd edn. (Springer, Berlin/Heidelberg, 2012), pp. 421–436
35. L. Bottou, Stochastic gradient learning in neural networks. *Proc. Neuro-Nîmes* **91**(8), 687–696 (1991)
36. N. Qian, On the momentum term in gradient descent learning algorithms. *Neural Netw.* **12**(1), 145–151 (1999). [https://doi.org/10.1016/S0893-6080\(98\)00116-6](https://doi.org/10.1016/S0893-6080(98)00116-6), <http://www.sciencedirect.com/science/article/pii/S0893608098001166>
37. I. Sutskever, J. Martens, G. Dahl, G. Hinton, On the importance of initialization and momentum in deep learning, in *Proceedings of the 30th International Conference on Machine Learning – Volume 28*, ICML’13 (2013), pp. III–1139–III–1147. <http://dl.acm.org/citation.cfm?id=3042817.3043064>
38. G. Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, 1976)
39. M. Aeberhard, S. Paul, N. Kaempchen, T. Bertram, Object existence probability fusion using Dempster-Shafer theory in a high-level sensor data fusion architecture, in *Proceedings of IEEE Intelligent Vehicles Symposium* (2011), pp. 770–775. <https://doi.org/10.1109/IVS.2011.5940430>
40. A.E. Albert, A.L. Albert, *Regression and the Moore-Penrose Pseudoinverse*. Mathematics in Science and Engineering: A Series of Monographs and Textbooks (Academic, New York, 1972)

Chapter 19

Analytics and Quality in Medical Encoding Systems



John Puentes, Laurent Lecornu, Clara Le Guillou, and Jean-Michel Cauvin

Abstract Medical practice support intends to provide important complementary information for diagnosis by preprocessing voluminous data available on separate, distributed, commonly noninteroperable applications of complex existing medical information systems. Such technology is being investigated to support medical encoding, which manually identifies groups of patients with equivalent diagnosis to determine healthcare expenses, billing, and reimbursement. Medical encoding is expensive, takes considerable time, and depends on multiple scattered and heterogeneous data sources. This chapter summarizes some relevant approaches and findings that illustrate how the considerations of information quality and analytics technologies may enable to improve medical practice. Essential components of a conceived medical encoding support system are described, followed by the associated data analysis, information fusion, and information quality measurement. Results show that it is possible to process, generate, and qualify pertinent medical encoding information in this manner, meeting physicians' requirements, making use of data available in existing systems and clinical workflows.

Keywords Computerized decision aid · Medical encoding support · Information analysis · Information fusion · Information quality

J. Puentes (✉) · L. Lecornu

IMT Atlantique, Lab-STICC, Technopole Brest Iroise – CS 83818, Brest, France
e-mail: John.Puentes@imt-atlantique.fr; Laurent.Lecornu@imt-atlantique.fr

C. Le Guillou · J.-M. Cauvin

CHRU Brest, Medical Information Department, DIM - Hôpital de La Cavale Blanche, Boulevard Tanguy Prigent, Brest, France
e-mail: clara.leguillou@chu-brest.fr; jean-michel.cauvin@chu-brest.fr

19.1 Components of a Medical Practice Support System

Medical practice support systems have emerged out of the unsettled definition and development of complex hospital information systems, rather inconsistent with field reality. Historically, medical staffs have been constrained to use functionally limited data and information processing applications implemented according to a widespread proprietary development strategy. Consequently, medical practice support tools search to enable diagnosis under the best viable conditions, making accessible related medical multimedia data and information. These tools include calculations based on extracted values, allowing the physician to interpret all those elements before taking a decision or performing an action. Multiple detailed information made available in this manner is expected to reduce manual data analysis and handling, within an environment of isolated applications. Therefore, medical practice support tools could enhance medical staff possibilities of action, by giving access to preprocessed data and information, instead of manually searching to come up with possible answers, when it becomes necessary to analyze huge data sets.

How can the integration of information and analytics technologies for medical practice support, enable existing intricate information systems to become truly adapted to users' needs and field realities? Which is the interest and implication of processing data, information, and quality, for medical practice support tools? This chapter intends to provide some clues to answer these questions by means of approaches conceived to cope with the complexity of medical information exploitation. We describe contributions proposed to improve medical encoding and to verify its quality with respect to a consensus of expert medical coders. Presented works focus on support for the interpretation and use for decision-making of voluminous medical encoding information. The developed approaches meet specific physicians' requirements in daily medical practice conditions, using open architectures to enable further applications evolution. In the rest of this section, the main elements of a medical practice support system are described, taking as an example a medical encoding support system, before presenting in the next two sections relevant processing, fusion, and quality evaluation approaches.

19.1.1 *Medical Encoding Support*

The primary objective of medical encoding is to identify groups of patients with equivalent diagnosis, in order to determine the corresponding healthcare expenses, billing, and reimbursement. Medical activities are thus being increasingly evaluated at various levels of health organizations by means of encoded information. Besides patient management, encoding relevance also affects epidemiologic, safety, research, and health policies decisions [1]. Medical encoding assigns codes to define care events that occurred during an inpatient stay. Codes represent [2] main and secondary diagnoses, complications, and comorbidities, as well as primary and

secondary procedures. Presently, medical encoding is carried out in two different manual manners: by expert coders and by physicians. Expert coders produce lists of codes that are considered to be exhaustive, without having additional patient information, i.e., collected during an examination or follow-up. Physicians code essential aspects of the care episodes having some knowledge of the specific patient history. Physicians generate only a subset of the code list produced by expert coders, because of their focus on current diseases, restricted awareness of encoding guidelines, and reduced time to assign codes. In both cases, medical encoding is expensive since human coders have to examine hundreds of candidate codes in encoding references, along with scrutinizing the patient record, to define the most appropriate code list. However, the pertinence of resulting code sets depends strongly on the variable coders' expertise, fluctuating between two types of erroneous results [1]: under- or overencoding.

Furthermore, medical encoding is constrained by the intrinsic complexity of having access to information required to accomplish it. Whereas data and information sources are substantially scattered, software applications have a reduced or nonexistent interoperability. As a consequence, most of the information cannot be used unless a concentrator collects and indexes it [3]. Figure 19.1 illustrates the main information streams required for medical encoding, identified in a French hospital information system (HIS). In this scheme, clinical activities are planned and followed as movements in the permanent patient ID database, while results and interpretations are stored in the hospital patient record, associated with the corresponding bills. Moreover, the PMSI (*Programme de Médicalisation des Systèmes d'Information* in French) and decision modules play a key role, summarizing related quantitative and standardized medical productivity information.

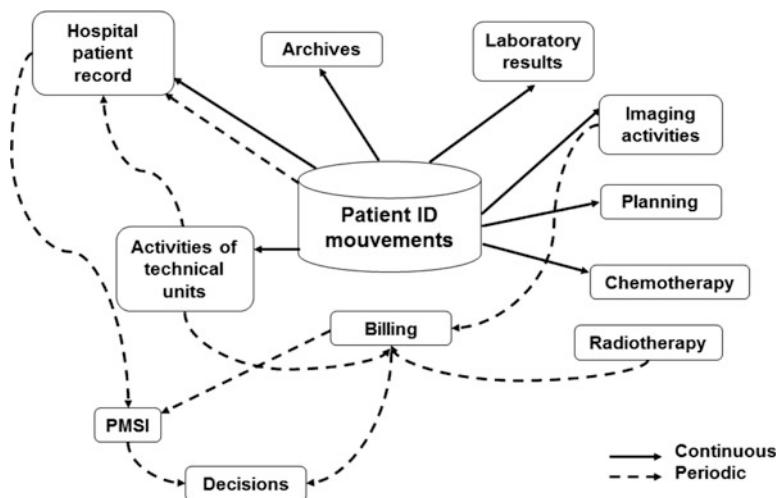


Fig. 19.1 Continuous and periodic hospital information streams necessary for medical encoding. (Adapted from [3])

Given the complexity of medical encoding, as well as its growing importance, computer-assisted technology emerges as an alternative way to analyze available patient data and information. A medical encoding support system intends to automatically generate a list of most pertinent medical codes, from which human coders can select in a more efficient manner the ones corresponding to a specific inpatient stay. In the next subsections, we present a schematic description of such a system based on published works, conceived, developed, and validated at the university hospital CHRU Brest, France.

19.1.2 Architecture

To generate medical code lists, the encoding support system processes the outputs of six HIS sources: laboratory results, rules to link laboratory results and diagnoses, previous discharge summaries, actuarial curves of codes chronicity, current discharge summary, and knowledge about relations between procedural and administrative parts. Depending on these information sources, three information processing tools –analysis of laboratory results [3], previous codes analysis [4], and probability analysis [5] – are used to generate semantic labels, estimation of pertinent codes, or probabilities of codes (Fig. 19.2), respectively. These partial results are aggregated to generate a unique list of ranked codes by considering their complementary pertinence. The generated encoding support list is then examined by the physician to select the most appropriate codes. Although additional HIS sources such as detailed clinical reports were also available, tested information extraction approaches were not adapted for encoding support.

Ground truth to verify the pertinence of encoding support quality was provided by consensus of three expert coders who coded manually the same information processed by the encoding system. Whereas the encoding system generated automatically the corresponding lists of plausible codes in few seconds, several weeks of analysis were necessary for the expert coders.

19.1.3 Inputs and Standards

Examined documents processed by the encoding system consist of laboratory results, medical unit discharge summaries, administrative documents, and encoding references. Standardized information about hospitalizations is collected in the medical unit discharge summary (Fig. 19.3). Each medical unit that provided healthcare during a patient stay, reports patient demographic data (age, sex), admission and discharge dates, main and secondary diagnoses, a severity index, and associated diagnosis codes, complemented by diagnostic and therapeutic procedure codes.

The CHRU Brest uses among others a general encoding reference based on the International Statistical Classification of Diseases and Related Health Problems –

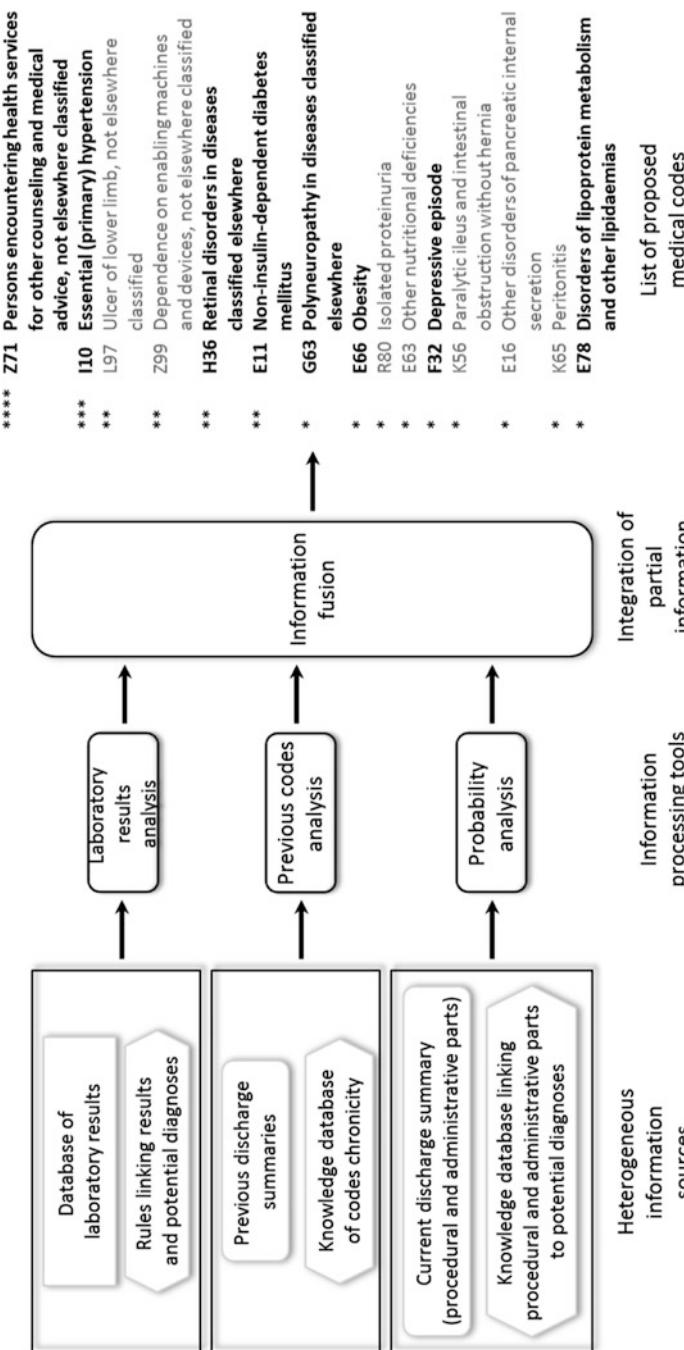


Fig. 19.2 Main components of the described medical encoding support system and example of a generated ranked code list of an inpatient stay at the endocrinology clinical department. Symbols “***” indicate the evaluated degree of pertinence and codes in bold were chosen by the physician [6]

Primary keys	
Birth date	
Sex	
Medical unit	
Admission date and modalities	Discharge date and modalities
Primary Diagnosis	Secondary Diagnosis
Severity Simplified Index	
List: Associated Diagnosis Codes, 1...N	
List: Diagnostic and therapeutic procedure codes (1,...N)	

Fig. 19.3 Elements of the medical unit discharge summary [3]

10th revision (ICD-10), compiled by the World Health Organization [7]. ICD-10 is used in numerous countries for registering morbidity and mortality causes and to facilitate the organization of healthcare services. It contains nearly 17,000 entries for symptoms, diseases, traumatisms, and other reasons to use health services, classified in 22 chapters. Each diagnosis code is composed of a letter, followed by 2–4 digits. Another encoding reference, the Common Classification of Medical Procedures (CCMP) [8] is a French nomenclature that describes medical procedures, using a code composed of four letters and three digits. Each letter denotes a part of the technical procedure context. CCMP and ICD-10 form a hierarchical encoding reference, consistent with medical knowledge, permitting to handle information at different aggregate levels.

19.1.4 Outputs

Three intermediate outputs are obtained after processing the initial heterogeneous input information (Fig. 19.2): semantic labels, scores of codes chronicity, and probabilistic values. Using rules to link laboratory results and potential diagnoses, the laboratory results analysis proposes codes of equivalent importance by using a semantic label, for example, rare or often, previously defined by an expert. Since some codes reappear when a chronic disease needs several hospital stays, analysis of previous codes determines the set of relevant recurrent diagnosis codes as a function of the time passed since their last occurrence. That enables to identify for any given patient the most pertinent previously utilized codes, which can be potentially applicable to the current stay. Finally, making use of a knowledge base formed by probabilistic predictions of diagnosis codes depending on patient age, sex, hospital stay length, diagnoses, and medical acts, the impact of these parameters is determined for each individual case.

These three intermediate outputs – semantic labels, scores of codes chronicity, and probabilistic values – cannot be separately interpreted by the physician, making necessary to generate a unique list of potential codes in order to select the most appropriate codes. A ranked and contextualized list of suggested diagnoses codes is generated by considering the events that took place during the patient hospital stay. To this end, a fusion algorithm aggregates the three partial code lists produced by the previously described methods by taking into account heterogeneity and complying with the respective indicators of code relevance.

19.2 Analytics and Fusion

As described previously, three independent information processing approaches handle documents stored in the HIS to produce medical codes by linking laboratory results and potential diagnoses by means of semantic labels [3]; evaluating previously assigned codes to identify frequency patterns [4]; and analyzing probabilistic relations between diagnoses codes and procedural parts of discharge summaries [5]. Partial heterogeneous information extracted by each method is subsequently integrated to generate one list of codes per inpatient stay [9]. This section describes the basic principles of the three information extraction approaches and the information fusion.

19.2.1 *Laboratory Results Analysis*

Data corresponding to laboratory examinations and associated patient condition are analyzed to define a subset of related diagnosis codes. The analysis relies on the interaction of three modules [3]. First is a set of rules describing the patient condition, the characteristics of relevant laboratory results, and proposing a set of diagnosis codes for that specific data collection. Second is an alerts module designed to notify a physician that the system has found a group of facts agreeing with a rule applied to a specific patient case and enabling therefore to provide relevant coding. An alert with a pertinence degree value is activated when a rule is satisfied. The third is an interfaces management module designed to access the different data sources, handling data heterogeneity and variability through time. Rules represent if-then statements with one or several conditions and a conclusion that are achieved when the conditions are satisfied. Conditions include a context to define the associated data source.

19.2.2 Previous Codes Analysis

The recurrence of diseases elements is analyzed to identify if previously existing codes show unambiguous temporal occurrence patterns. It is assumed that a chronic disease, likely to be described in some or all previous hospital stays, can be proposed as a pertinent coding option for the current stay, depending on certain conditions [4]. Adapted actuarial survival models of codes are estimated on a reference database of discharge summaries that represent millions of hospital stays during several years. Actuarial survival models of codes are then used to estimate the recurrence of diagnosis codes linked to chronic diseases. Calculated estimations constitute a knowledge base of recurrence rates for each diagnosis code, depending on the elapsed time since it was previously used. Additional knowledge represented by codes assigned to the patient during the 2 years preceding the current hospital stay is applied to construct the proposed list of related codes. Besides the ordered codes and associated labels, results include precision and recall rates that define the code importance.

Hence, coding of a new stay relies on diagnoses of all previous stays during 2 years. For that reason, all previous codes and associated delays (time between the end of the last stay when the code appeared and the beginning of the current stay) are included. Only codes having a minimal delay are selected to assign a reappearance rate. If reappearance rates are higher than an experimentally determined threshold, a ranked list of codes is proposed according to those results.

19.2.3 Probabilistic Analysis

Making use of a knowledge base extracted from a large database of anonymous discharge summaries, codes are also predicted according to the probabilistic relations between some fundamental healthcare variables. Collected probabilistic predictions suggest diagnosis codes depending on patient age, sex, hospital stay length, diagnoses, and medical acts. To reduce the number of combinations, probabilities of diagnoses are grouped according to the impact of age, stay length, diagnoses, and medical acts [5].

The probability of diagnosis calculated using the preceding elements considers four information sources: the set of age, sex, and stay duration; the medical or functional unit that provided healthcare; the medical procedures already encoded; and already coded diagnoses. Conditional probabilities of each information source with respect to the other three are calculated separately and combined linearly, applying four weights that depend on the performance of each individual information source (evaluated experimentally). The relevance of resulting proposed codes is indicated by the corresponding final probability.

19.2.4 Information Fusion

Lists of codes generated by the analysis of laboratory results, previous codes, and probabilities are heterogeneous and represent three separate partial lists of suggested codes and relevance values. A suitable information fusion method is, therefore, necessary to obtain the code list generated by the medical encoding support system [9]. A fusion by voting [10] counts whenever a code was suggested by the previous information extraction methods. Since extracted information is likely to be complementary, many codes obtain the same quantity of votes, implying that information sources degree of credibility must be taken into account. Empirically, when credibility is applied to improve codes ranking, it assigns the best prediction scores to patterns of previously existing codes, followed by probabilities, and laboratory results.

Another suitable fusion approach is to order proposed diagnosis codes according to the different relevance values. Nevertheless, relevance values are also heterogeneous, and, therefore, a common base for the analysis is necessary. Code accuracy is estimated to cope with this difficulty. It is defined as the ratio of the number of times each code is chosen by the physician to the number of times the code is proposed by the information extraction method. To reduce the uncertainty of probability estimations, recall rates, and particularly semantic labels, the intrinsic accuracy of each code is represented by a possibility value [11], enabling to combine them. The accuracy of each information extraction method is examined as a transformed probabilistic relevance (necessity/possibility pairs), to define an experimental conversion table. Results are smoothed to achieve a monotonically decreasing trend. This permits to cope with the problem of finding a monotonic conversion function to transform relevance values, given that their range is divided into accuracy intervals. Values are merged using a conventional fusion operator, e.g., the *max* operator.

To evaluate the performance of these two fusion methods, a sample of 1000 discharge summaries representing hospital stays of more than 24 h, having at least two diagnosis codes and one medical procedure code, was used. Reference codes were those produced by physicians for the same patients. Information retrieval estimators – precision and recall – were applied to measure fusion results. Precision is defined as the number of pertinent identified codes divided by the total number of retrieved codes, and recall is defined as the number of pertinent identified codes divided by the total number of known existing pertinent codes in the generated list. Figure 19.4 shows the obtained precision and recall rates curves, depending on the number of generated pertinent codes. Both information fusion methods propose relatively the same codes in the first 15 ranks, confirming for the tested data set their compared stability. However, the rank of a relevant diagnosis code may be increased, if it is found by only one extraction method and has a low credibility factor.

While the highest precision is obtained by the probabilistic analysis, recall results indicate that the two fusion approaches outperform the three partial information extraction methods. Note that there are some particular differences related to

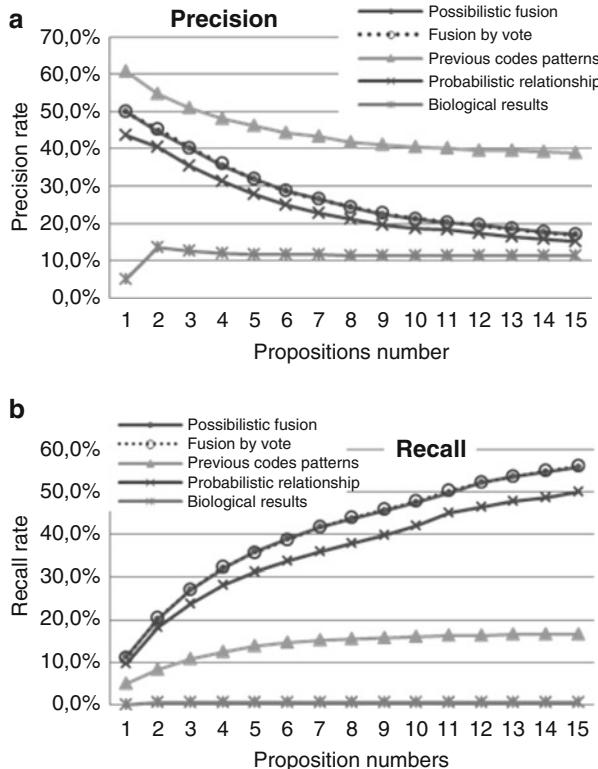


Fig. 19.4 Evaluation of the precision (a) and recall (b) rates [9]

ranking disparities not revealed by these curves. To understand those differences, a coding example is analyzed. Table 19.1 presents the coding of a hospital stay made by a physician, and Table 19.2 presents the ordered lists of codes proposed by each information fusion method. Both fusion methods propose the same 14 codes and list one expected but not proposed (n.p.) code. Results of Table 19.2 show that, as a consequence, the valid codes order is different: four codes, N18, Z49, Z94, and D64, are placed at positions 1, 3, 5, and 6 by the possibilistic fusion and positions 1, 3, 10, and 5 by the vote fusion, respectively. These variations generate differences not observed in Fig. 19.4, which have nevertheless an impact when the physician uses the diagnosis coding support system because the order of proposed codes is different.

Codes obtained from biological laboratory examinations have limited influence on the final results, mainly because a condensed set of examinations not widely representative of most known cases was used. It can also be observed that although previous code analysis identifies a relatively smaller amount of proper codes, most of these codes are relevant. In general, diagnosis coding results generated by both

Table 19.1 Example of stay coding done by a physician [9]

Discharge summary ICD10 codes and label	
D64	Other anemias
N18	Chronic kidney disease (CKD)
T86	Failure and rejection of transplanted organs and tissues
Z00	General examination and investigation of persons without complaint and reported diagnosis
Z49	Encounter for care involving renal dialysis
Z94	Transplanted organ and tissue status

Table 19.2 Proposed coding (for the stay of Table 19.1) by the two fusion methods [9]

Proposed ICD10 codes and label	Possibilistic		Vote	
	Rank	Value	Rank	Sources
D47 Other neoplasms of uncertain behavior of lymphoid, hematopoietic, and related tissue	13	0.130	8	1
D64 Other anemias	6	0.632	5	1;2
D69 Purpura and other hemorrhagic conditions	8	0.437	6	1;3
E11 Type 2 diabetes mellitus	4	0.867	4	1;2
E79 Disorders of purine and pyrimidine metabolism	11	0.25	9	1
I10 Essential (primary) hypertension	12	0.25	12	2
I15 Secondary hypertension	n.p.		15	2
I25 Chronic ischemic heart disease	14	0.12	13	2
K74 Fibrosis and cirrhosis of liver	9	0.388	7	1
N17 Acute renal failure	2	1.87	2	1;2;3
N18 Chronic kidney disease (CKD)	1	1.87	1	1;2;3
N99 Intraoperative and postprocedural complications and disorders of genitourinary system, not elsewhere classified	7	0.5	n.p.	2;3
Y43 Primarily systemic agents adverse effect	15	0.1	14	2
Z49 Encounter for care involving renal dialysis	3	1.042	3	1;2
Z51 Encounter for other aftercare	10	0.25	11	2
Z94 Transplanted organ and tissue status	5	0.68	10	2

Sources: 1, Previous codes patterns; 2, Probabilistic relationships; 3, Biological results; n.p. not proposed

information fusion methods are equivalent, albeit some particular differences that make either one of the methods closer than the other to the expected code list.

19.3 Quality Evaluation

Since exploited information sources are heterogeneous, the encoding support output is affected in various manners: inappropriate code ranking, missing codes, dispersion of correct codes among incorrect ones, and noise induced by long sequences

of incorrect codes. Expert coders and physicians are able to rule out autonomously most of the information heterogeneity consequences, applying learned heuristics and/or having access to additional patient documents. Encoding support systems cannot solve those problems in an equivalent manner, making necessary to measure the information quality of each generated code list. This section describes the approaches conceived to follow up the quality of the encoding support process and to estimate the quality of automatically generated code lists.

19.3.1 Process Quality

Taking into account the complexity of a medical encoding support system and the complementarity of its multiple interacting parts, it is necessary to go beyond the simple unrelated estimation of separate data or information quality. To qualify up to what point coding support was correctly accomplished, a methodology to evaluate information quality was defined and validated [12]. It offers the possibility to comply with the encoding support system evolution and to explain how changes in coding support modules have an impact on the quality of coding results.

To characterize the pertinent quality dimensions, an adapted set of quality criteria is determined for every basic component of the encoding support system. Each component is then modeled by a quality transfer function (analog to a transfer function in signal processing) that represents how the component processing task affects information quality at a local level. The information quality of a global process results from the aggregation of the calculated local quality transfer functions. Since local quality transfer functions depend only on their respective inputs and outputs, quality propagation through the system can be evaluated for a static system configuration, as for any other configuration, regardless of data and information heterogeneity.

The quality transfer function Q_f provides the component's output quality as a relation of the input information (I_{in}) and its quality (Q_{in}) as:

$$Q_{out} = Q_f(I_{in}, Q_{in}) \quad (19.1)$$

Since the notion of quality is multidimensional, the function Q_f is also multidimensional. Input quality is evaluated using three criteria: accuracy (ratio of the correct codes to the total number of codes), completeness (ratio of registered codes to all the codes that could have been registered), and currency (percentage of extracted codes that are up to date). The quality of components' output is evaluated by correctness (proportion of correct codes), completeness, and reliability (codes' confidence degree). Figure 19.5 illustrates an example of the application of the method to a particular data set on which the inputs of each processing module are qualified in order to obtain the respective output quality.

The proposed method has the potential to identify faulty system components from the quality point of view by permitting the verification of the local quality

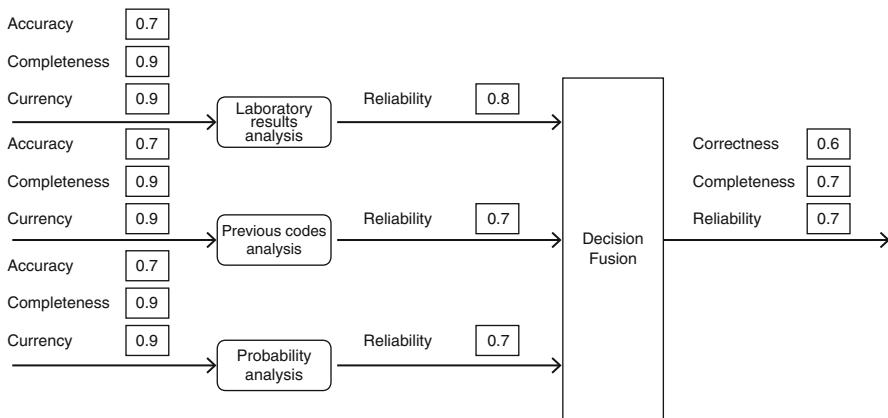


Fig. 19.5 Example of process quality evaluation. (Adapted from [12])

according to the application conditions. A considerable advantage is also that changes or updates of processing system components do not imply a complete additional analysis. Only the quality transfer function of the pertinent component needs to be determined before calculating the whole system quality, abstracted as the propagation of quality values from one component to another. Having an estimation of the global process quality implies also the feasibility of comparing system architectures.

19.3.2 Quality of Medical Coding Lists

A significant problem with medical encoding support systems is how to measure the proposed code lists appropriateness in terms of fitness for use, i.e., quality, according to the distribution of correct and incorrect codes along the list, the amount of expected correct codes, and the variable list length. That information quality measurement should assert the practical value of any code list in a manner suitable to the adapted encoding practices of hospitals. Automatically generated lists of codes represent nevertheless information with variable quality. Based on how expert coders and physicians make use of computer-assisted medical encoding, a quality measure was defined and validated to evaluate codes accuracy (A), dispersion (D), and noise (N) in the whole generated list, independently of its content and length [6].

Expert coders and physicians first review the available inpatient stay documents. Next, they study the corresponding code list generated by the encoding support system by knowing from the documentation review how many diagnoses and procedures should be coded. Since the proposed code list is formed by correct and incorrect codes, a strategy is applied to make use of it. The whole list is implicitly divided into three observation windows of variable lengths. Required

correct codes should ideally be found in the first window, but if some or all expected correct codes are not in that window, the second window is inspected. The third window is partially or fully examined when correct codes are obviously still missing after the second window inspection. These three observation windows are called compactness (*CW*), dispersion (*DW*), and noise (*NW*) windows, respectively.

The number of correct codes found in the *CW* window defines the first component of the encoding information quality measurement (IQ_1). This accuracy value (A_{CW}) is calculated as the ratio of proper codes to the total amount of necessary correct codes. In an equivalent manner, appropriate codes are identified in the *DW* to define the A_{DW} accuracy value with respect to the total amount of necessary correct codes. Contrary to IQ_1 , the second component (IQ_2) of the encoding information quality measurement is altered by dispersion and noise according to usability. Also, information quality IQ_1 varies in the interval [0,1] and information quality IQ_2 is within [-1,1], due to the combined influence of noise and dispersion. Usability implies that the user should be able to simultaneously quantify the global quality and identify each component IQ_1 and IQ_2 . Both conditions are satisfied when drawing associated quality values in a 2D representation space. As a result, the information quality of a code list is fully represented by a single point within a normalized triangular space (Fig. 19.6), partitioned by iso-quality lines (oblique downward lines from left to right), on which the best global quality is at point (0, 1) and the worst at point (-1, 0). Moreover, the model is suitable to examine and compare, by using a unique scheme, the information quality of hundreds of code lists, showing their practical value for encoding.

A significant finding of this quality model is the definition of iso-quality lines, which represent all the points having the same global quality value, with different quality components contributions. It depicts clearly the fact that a given quality value can result from countless configurations of the quality factors. Besides, unlike 2D accuracy diagrams, the quality variation between two lists with the same correct codes, but organized differently, can be detected in a simple manner along the iso-quality lines.

19.4 Conclusion

The exponentially growing availability of numerous and varied technologies is both demanding and stimulating for the future of medical practice support. This chapter presented a medical encoding support system from the perspective of data analysis, information fusion, and information quality, which is embedded in the process of synthetizing lists of relevant codes to facilitate the work of human coders, making use of data and information acquired and produced by a complex distributed medical information system. Results suggest that the integration of information and analytics technologies has a clear potential to make evolve existing systems, by creating suitable functionalities adapted to users' need and professional context realities. Such progress does not require to modify existing systems but to be able to collect

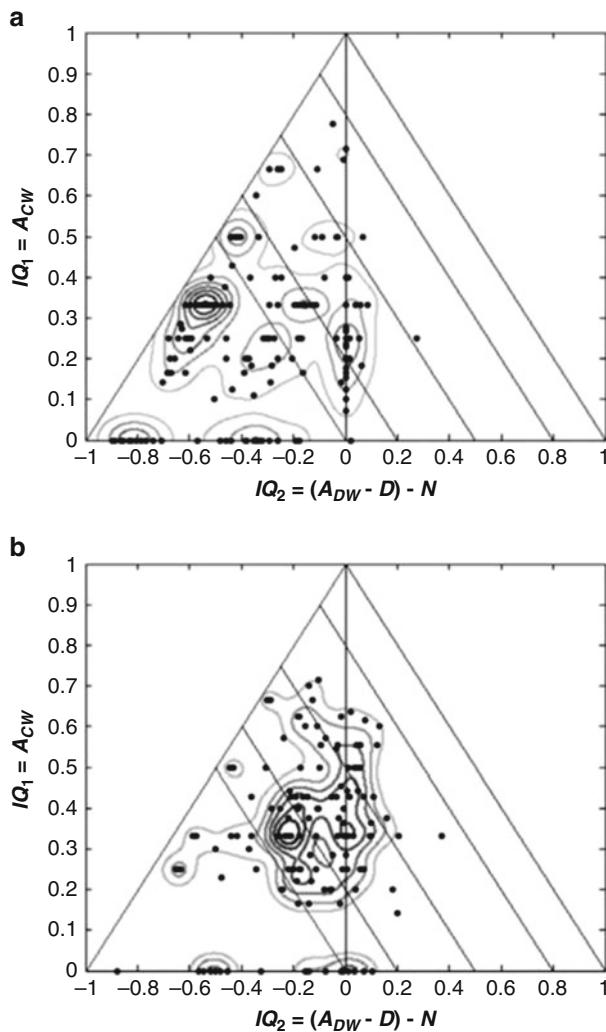


Fig. 19.6 Information quality points of code lists generated by the encoding support system and corresponding density-level curves to define clusters of codes for the (a) traumatology, 198 lists, and (b) obstetrics—141 lists clinic departments of the same hospital. (Adapted from [6])

relevant data and information, making accessible to the physician the necessary information to support medical practice.

A current medical mindset transition from a resource-based to a results-based culture entails a growing interest in medical practice decision support processes of significant added benefit compatible with existing clinical workflows. That is the case of the described system, which provides a considerable aid to simplify a complex process by using the same elements as medical staffs. Otherwise, despite the fact that physicians can be relatively tolerant to wrong information, quality

evaluation is mandatory because of the heterogeneous and imperfect intrinsic nature of medical data and information. Although quality evaluation of information fusion systems is at its beginnings, the two proposed approaches to estimate the quality of the whole process and generated code lists illustrate the difficulty of this question and confirm the need to carry out such measurements, beyond the performance evaluation of individual information processing algorithms.

We emphasize the advantages of information and analytics technology combination to blueprint medical practice support tools adapted to user needs, developed by using flexible open technologies, and deployed according to physicians' requirements. Addressing open development approaches, clinical workflow, and specific field realities will enable the adapted conception of systems for highly specialized medical tasks and services, expecting to resolve the chronic mismatch between technology features and end-user acceptance.

References

1. P. Cheng, A. Gilchrist, K.M. Robinson, L. Paul, The risk and consequences of clinical miscoding due to inadequate medical documentation: a case study of the impact on health services funding. *Health Inf. Manag.* **38**(1), 35–46 (2009)
2. C.J. Buck, *Step-by-Step Medical Coding* (Elsevier Health Sciences, St. Louis, 2016)
3. L. Lecornu, C. Le Guillou, G. Thillary, et al., C2i: a tool to gather medical indexed information. Paper presented at the 9th IEEE international conference on information technology and applications in biomedicine, Larnaca, 5–7 November 2009
4. L. Lecornu, C. Le Guillou, F. Le Saux, et al., ANTEROCOD: actuarial survival curves applied to medical coding support for chronic diseases. Paper presented at the 32nd IEEE International Conference of the Engineering in Medicine and Biology Society, Buenos Aires, 29 August–4 September 2010
5. L. Lecornu, G. Thillary, C. Le Guillou, et al., REFEROCOD: a probabilistic method to medical coding support. Paper presented at the 31st IEEE international conference of the engineering in medicine and biology society, Minneapolis, 3–6 September 2009
6. J. Puentes, J. Montagner, L. Lecornu, J.M. Cauvin, Information quality measurement of medical encoding support based on usability. *Comput. Methods Prog. Biomed.* **112**(3), 329–342 (2013)
7. World Health Organization, International classification of diseases: ICD-10, vol. I–XXII, 2016, <http://apps.who.int/classifications/icd10/browse/2016/en>
8. M. Maravic, C. Le Bihan, P. Landais, La classification commune des actes médicaux (CCAM) : de la description à la tarification. *Rev. Rhum.* **70**(9), 785–789 (2003)
9. L. Lecornu, C. Le Guillou, F. Le Saux, et al., Information fusion for diagnosis coding support. Paper presented at the 33rd IEEE International Conference of the Engineering in Medicine and Biology Society, Boston, 30 August–03 September 2011
10. I. Bloch, *Fusion d'informations en traitement du signal et des images* (Lavoisier, Hermès Science, Paris, 2003)
11. D. Dubois, H. Prade, Théorie de possibilité, théorie des probabilités et logiques multiples évaluées: une clarification. *Annales des Mathématiques et de l'Intelligence Artificielle* **32**, 35–66 (2001)
12. I.G. Todoran, L. Lecornu, A. Khenchaf, J.-M. Le Caillec, A methodology to evaluate important dimensions of information quality in systems. *ACM JDIQ.* (2015). <https://doi.org/10.1145/2744205>

Chapter 20

Information Quality: The Nexus of Actionable Intelligence



Marco Antonio Solano

Abstract Information quality (IQ) plays a critical role in the ability of a high-level information fusion (HLIF) system of systems (SoS) to achieve actionable intelligence (AI). Whereas the need for information quality management in traditional information systems has been understood for some time, and its issues are fairly well mitigated, the challenges pertaining to the relatively new field of high-level information fusion remain significant. Principal and unique among these challenges are the multitude of issues which arise from the inherent complexity in high-level information fusion system of systems and which permeate throughout the various interdependent phases of its life cycle. Actively managing information quality in HLIF is essential in ensuring that they do not adversely impact decision-making and the ability to determine the best course of action (COA). Accordingly, in an effort to advance this critical facet of high-level information fusion, this chapter proposes an end-to-end framework that enables (a) the development of an information quality meta-model (IQMM), (b) the characterization of information quality elements, (c) the assessment of impacts of information quality elements and their corresponding mitigation, and (d) the integration of these aforementioned objectives within the HILF processes and life cycle.

Keywords Information quality · Quality meta-model · Decision support · Actionable intelligence · Information fusion · Quality framework

20.1 Introduction

Information quality (IQ) originated more as a secondary consideration within information systems (IS); that is, it was understood that there is an inherent relevance, yet it rarely surfaced as a key concern until the results were not what

M. A. Solano (✉)

Raytheon IIS (Intelligence, Information and Services), Dulles, VA, USA
e-mail: marco_a_solano@raytheon.com

was expected, or not within the norm. All too often, information quality was either assumed or at best reacted upon, and, as such, not usually designed in.

Furthermore, there are tools available as “plug-ins” for traditional information systems, such as an extract-transform-load (ETL) process in a database-centric project, which provide capabilities that address several issues related to data quality, e.g., missing values, duplicate values, inconsistent entries, and outliers. More recently, master data management (MDM) solutions help the configuration management of the enterprise’s critical data holdings. The access to generic tools for information quality furthers the notion that there are sufficient after-the-fact recourses available to deal with the ubiquitous information quality issues reactively. Whereas these tools may be appropriate for certain aspects of information quality, such as data marts and data warehouses, they represent a rather small sliver of the comprehensive solution space and, as such and by themselves, may not be relied upon to provide a holistic IQ approach for the significantly more complex information quality inherent to HLIF system of systems (SoS). It is important to notice that information quality measures must be executed throughout the entire information exploitation chain to achieve viable actionable intelligence.

Figure 20.1 depicts this gap between traditional data quality (DQ) and information quality (IQ), where IQ is a requisite for achieving actionable intelligence. Note that IQ overlaps DQ instead of just picking up from where DQ left off. This overlap denotes that IQ needs to be implemented for actionable intelligence from the outset, and therefore, it may not be necessarily sufficient to execute mitigation in tandem, but rather a distinct approach is needed.

The information fusion (IF) community, both academia and industry, have conducted many strides in developing and evolving IF concepts, accommodating the growing complexity, and creating pragmatic implementations. Even more importantly, despite often diverse priorities and expertise that underlie the domains of low-

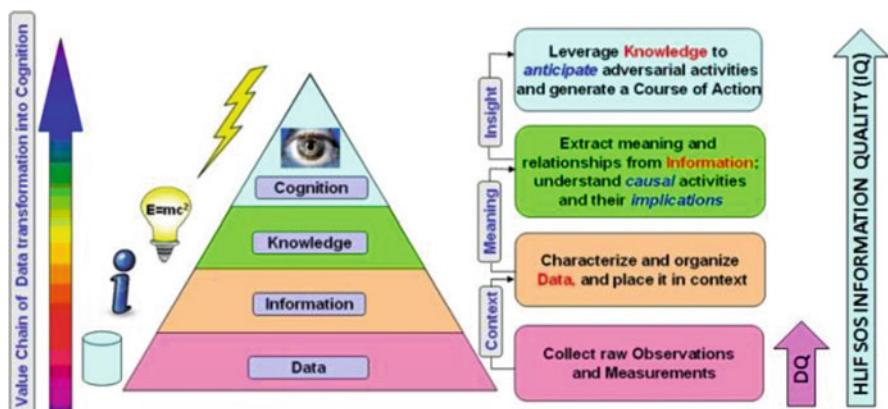


Fig. 20.1 Bridging the gap between data quality (DQ) and information quality (IQ)

and high-level fusion, the IF community has been able to propose, develop, and maintain cohesive integrated frameworks to support end-to-end HLIF systems [21].

Given the current level of maturity achieved in the domain of information fusion, it is now befitting, if not somewhat overdue, that a rigorous examination of information quality be undertaken. The objective is to formally define frameworks that incorporate and integrate the ability to mitigate information quality issues from the outset and throughout the HLIF life cycle.

Accordingly, in an effort to advance this critical facet of high-level information fusion, this chapter proposes an end-to-end framework that enables (a) the development of an information quality meta-model, (b) the characterization of information quality elements, (c) the assessment of impacts of information quality elements and their corresponding mitigation, and (d) the integration of these aforementioned objectives within the HILF processes and life cycle.

20.2 Information Quality and Information Fusion

As a result of the advances in the information fusion community, the view of the original Joint Directors of Laboratories (JDL) Fusion model has evolved to include and integrate additional considerations. A holistic JDL view currently includes functionality for both low- and high-level fusion comprising Level 0 (raw data processing, feature refinement), Level 1 (object refinement), Level 2 (situation assessment), Level 3 (threat assessment), and Level 4 (process refinement) [22], as well as advanced topics for user refinement (proposed Level 5) and mission management (proposed Level 6) [3].

This succinct recapitulation of the fundamental JDL model is essential in understanding the critical role that information quality plays within the high-level information fusion (HLIF) domain.

20.2.1 Actionable Intelligence

By and large, actionable intelligence (AI) is the desired outcome of HLIF. Actionable intelligence enables a decision-maker to formulate a course of action (COA), i.e., it can be acted upon, usually with the intent of gaining an advantage, and usually predicated on a scenario where all the desired information may not be readily available, and the available information is imputed with various degrees of uncertainty. The consequences of engaging in a course of action based on “wrong” or “bad” information may vary significantly, from an inconvenient yet innocuous temporary corporate embarrassment to the dire long-lasting repercussions of inciting an international conflict.

This latter consequence is often referred to within the intelligence community (IC) as an “intelligence failure”; it connotes a catastrophic series of events resulting

from a decision which was based on flawed intelligence. “Intelligence errors are factual inaccuracies in the analysis resulting from poor or missing data. Intelligence failure is the systemic organizational surprise resulting from incorrect, missing, discarded or inadequate hypothesis” [17].

Whereas arguably information quality may influence some of these aforementioned dynamics, e.g., the role a presentation style of the information results may play in influencing the decision-maker [30], it is evident that information quality is indeed a key contributing factor to the overall integrity of the resultant actionable intelligence of a high-level information fusion system.

20.2.2 The Impetus for Information Quality in Information Fusion

When juxtaposed, it becomes evident that HLIF is several orders of magnitude more complex than traditional information systems, e.g., general business applications, data marts, data warehouse, and to some degree, at the higher end of the information systems spectrum, even the corporate-level decision support systems (DSS) and business intelligence (BI) applications. Ostensibly, the latter exhibits primitive parallels with information fusion systems. Accordingly, while existing information quality management techniques and tools from information systems apply to some facets of HLIF, by and large there are no corresponding mature solutions for most of the intricate and numerous interdependent components that form an end-to-end HILF system.

Moreover, a central property of HLIF is that the complexity of the end-to-end system is greater than the sum of the complexity of all its interdependent yet individual parts. This gives rise to questions such as how can the information quality of the overall HLIF system be defined?”, let alone measured from the individual heterogeneous constituent information quality elements.

From a cost-benefit perspective, the motivation for mitigating information quality shortcomings in HILF systems becomes self-reinforcing when the gravity of the potential impacts of failing to do so is taken into account. Unmitigated IQ issues represent an inherent and unmanageable risk to achieving the objectives of an HLIF SoS.

As information fusion is positioned for a breakthrough, given the advances in auspicious technologies such as mobile and cloud computing, data analytics, sensor management, and information security, it is opportune to undertake the systematic resolution of the many vexing issues induced by information quality shortcomings.

Table 20.1 provides a set of information quality dimensions to help frame the initial problem space and identify and characterize high-level system requirements.

It stands to reason that the initial dimensions assessment may exhibit a degree of subjectivity of its own. Notwithstanding, its primary purpose is to enable the comprehensive and holistic view of how information quality impacts the various

Table 20.1 Information quality dimensions for qualifying initial IQ concerns/requirements

Information Dimension/(Characteristic Abbreviation) <i>Characteristic Name:</i> Description	
Information Quality Impact	
(P)	<i>Primary:</i> Information is the root cause of the quality issue at hand
(S)	<i>Secondary:</i> Information itself is a secondary contributor, e.g., algorithm, process
Information Quality Scope	
(T)	<i>Tactical:</i> Pertaining a particular fusion level, e.g., L0, L1, L2, L3
(S)	<i>Strategic:</i> Applicable throughout the fusion levels or life cycle
(B)	<i>Both:</i> Impacting both tactical and strategic aspects of HLIF
Information Quality Composition	
(M)	<i>Monolithic:</i> Information with homogeneous morphology
(H)	<i>Heterogeneous:</i> Information comprising multiple phenomenologies
Information Quality Evaluation	
(O)	<i>Objective:</i> Information can be quantified through objective measures
(S)	<i>Subjective:</i> Information may only be assessed through subjective characterization
(B)	<i>Both:</i> Encompasses both objective and subjective measures and characterization

facets of HLIF. In turn, this categorization will allow for a systematic analysis and development of an integrated approach to address IQ concerns.

Table 20.2 recapitulates representative HLIF information quality concerns collated from existing research [4, 5, 7, 28–31] and also proposes additional candidates. This list is by no means exhaustive, but the breadth of its scope does convey the diverse expertise required to properly address them. Additionally, a systems engineering methodology is adopted whereby the topics are presented as HLIF system requirements. These requirements have been qualified along four dimensions to provide a richer context with which to analyze them and their impact on the overall HLIF system and to help elicit additional information quality-related concerns. These dimensions are as follows:

This representative list of requirements underscores the critical role that information quality plays in the overall HLIF system, for example, how it drives *accurate*, *precise*, and *reliable* actionable intelligence.

20.2.3 *The Need for an Integrated and Holistic Approach*

Whereas a traditional information system may architecturally be described as a system, with all its design, development, and implementation rigor and processes appertaining, HLIF is more aptly described as a system-of-systems.

The implication being that the inherent complexity of a SoS is orders of magnitude greater, and as such it is imperative to undertake a systematic and integrated approach to managing IQ concerns throughout the HLIF life cycle.

Accordingly, it is suitable that the disciplines systems engineering (SE) and systems architecture (SA) be leveraged as a central approach with which to frame the

Table 20.2 Selected key information quality concerns and requirements for HLIF SoS

Information Quality Concerns/Requirements		Impact	Scope	Composition	Evaluation
The following topics should be read as IQ requirements, i.e.: The HLIF system shall have the ability to:					
1 ...identify information quality elements	S	B	H		O
2 ...derive the total HLIF System Quality Measure	S	S	H	S	
3 ...integrate IQ in the HLIF life cycle	P	B	H	O	
4 ...parametrize the impact of IQ on actionable intelligence	P	S	H	S	
5 ...determine a confidence for subjective IQ elements	P	B	H	S	
6 ...aggregate individual IQ assessments of diverse data sources	P	S	H	S	
7 ...determine quality of service for HLIF components	S	B	H	O	
8 ...evaluate the impact of individual IQ elements	P	B	M	S	
9 ...determine an algorithm sensitivity based on its IQ elements	S	T	H	O	
10 ...evaluate the effects of information granularity (e.g., resolution)	P	B	M	B	
11 ...evaluate the processing performance impacts of IQ elements	S	B	H	O	
12 ...determine the cost-benefit trade-offs of IQ	S	S	H	S	
13 ...assess interoperability among information providers and consumers	S	T	H	O	
14 ...provide information security (confidentiality, integrity, availability)	S	S	H	O	
15 ...manage information pedigree throughout the HLIF life cycle	S	S	H	O	
16 ...mitigate IQ-induced degradation	P	B	H	O	
17 ...determine the root cause of IQ-related HLIF system impacts	S	S	H	O	
18 ...establish an extensible IQ metadata framework for HLIF	S	S	H	S	
19 ...assign weights to intra- and inter-IQ element vectors	P	T	B	S	
20 ...determine and calculate IQ measures of effectiveness	P	S	B	S	

problem space of IQ in HLIF SoS and formulate its corresponding solution space. Both SE and SA disciplines provide the proven systematic approach and ability to draw from and integrate the multidisciplinary expertise required to properly manage the intricate aspects of information quality in high-level information fusion system-of-systems.

In this vein, the following HLIF IQ tenets are proposed. These two architecture principles (APs) combined provide simple yet powerful complementary perspectives for managing information quality concerns:

- Information quality must be fit for purpose.
- Information quality is enabled by managing information as an asset.

Fit for purpose [10] incorporates IQ considerations with respect to achieving the objective of an HLIF, i.e., actionable intelligence, by enabling a top-down analysis. That is, fit for purpose is based on the decision-maker needs. Another key concept associated with fit for purpose is to establish a threshold that represents what is good enough for a particular mission; i.e., it ties with a customer's measures of effectiveness (MOE), which flow down to the system of systems measures of performance (MOP), key performance parameters (KPPs), and technical performance measures (TPMs) [26].

The information as an asset architecture principle [13] incorporates IQ considerations as a key architectural building block with which to synthesize a robust bottoms-up design for HLIF system of systems. Information as an asset is essential given the core value of information as a resource and its role in decision-making. Furthermore, this is a key principle upon which information management roles, e.g., information custodian and information owner, should be considered and used to dovetail with information governance.

These high-level and low-level perspectives are integrated with the systematic and iterative methods of (a) characterizing the IQ problem space, (b) assessing the impact of IQ on achieving optimal actionable intelligence, (c) providing a corresponding solution space comprising tools to mitigate the impacts, and (d) integrating these steps within the HLIF life cycle.

In contrast with traditional information systems, where a reactive IQ approach is often good enough, HLIF SoS, due to its highly iterative nature and cascade effects of propagating IQ issues in tandem with processing and algorithm execution, means that no amount of ex post facto IQ mitigation can properly address the IQ deficiencies introduced in upstream processes.

As mentioned by Hall et al. [17], description of fusion sensors and downstream processing alone cannot correct for upstream sensor data errors. More importantly, Hall et al. quantify that the combination of marginal sensor performance will not yield an improved result. The implication is that IQ will only be effective if proactively addressed and accordingly cannot be bolted on to a system, but rather should be designed in.

Figure 20.2 depicts the proposed information quality approach to optimize actionable intelligence in high-level information fusion system of systems. It leverages formal systems engineering and architecture concepts and is requirements driven.

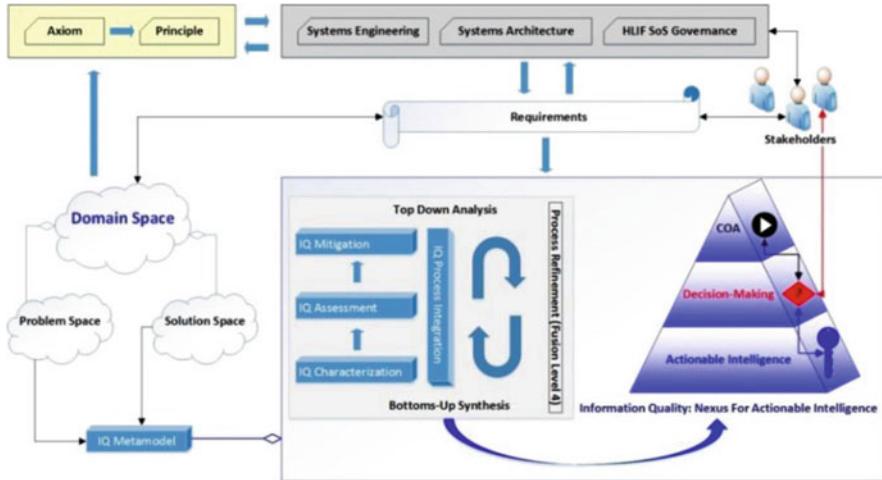


Fig. 20.2 Strategic approach for an information quality framework in HLIF SoS

Decision-makers use actionable intelligence to determine a course of action (COA). Multidisciplinary eclectic teams participate in systems engineering, architecture, and governance functions.

20.3 Concepts and Primitives: Outlining an IQ Meta-model

Information quality is a subdomain of high-level information fusion; as such, it requires that a common ground be established whereby its domain scope, which is composed of the corresponding problem space and solution space, may be fully qualified.

The problem space is defined as the comprehensive set of elements, e.g., topics, items, concerns, and requirements, that exist in a domain and which fully identify, qualify, and quantify the problems to be solved and, accordingly, for which a solution must be formulated.

The solution space is defined a set of enablers comprising the architecture, processes, tools, and solutions, which provide the requisite remedies to address and manage the problem space.

Consequently, as a prerequisite to embarking on a holistic systematic approach to resolving IQ challenges in HLIF, the terms, definitions and concepts, as well as the implication of these should be developed and defined first. An added advantage of this undertaking is that it also elicits perspectives that help frame the context of IQ within the various complex facets of an HLIF SoS that would otherwise not be apparent [18].

Ultimately, the intent of this effort is to aid in developing an HLIF IQ meta-model (IQMM) that will help built a comprehensive framework to analyze, create rules, set up constraints, and generate models applicable to the IQ problem and solution spaces.

20.3.1 Key Terms and Definitions

A succinct exploration of literature from outset of the first information management systems to present-day complex decision support systems will yield the genesis of many information- and information quality-related concepts and terms.

Although many of these definitions may be useful and extend to the domain of information fusion, it is advantageous that a specific vernacular be developed with which to clearly articulate the problem and solution space particular to the HLIF domain.

Figure 20.3 places the terms and concepts in context of the high-level information fusion system-of-systems domain dynamics. A simplified input-process-output (IPO) view is used to depict how information is exploited throughout the iterative Level 0 to Level 6 information fusion cycles – highlighted by the light blue flow and feedback connections. Externalities may interact with the HLIF as well as with the key stakeholder: the decision-marker highlighted by the gray dashed connectors. Information quality permeates throughout the HLIF life cycle, shown in the solid dark-blue connectors; the dashed dark blue line indicates the critical role IQ plays in the overall mission success.

The terms and definitions corresponding to the headings presented in the following tables (Tables 20.3, 20.4, and 20.5) are mapped to the components

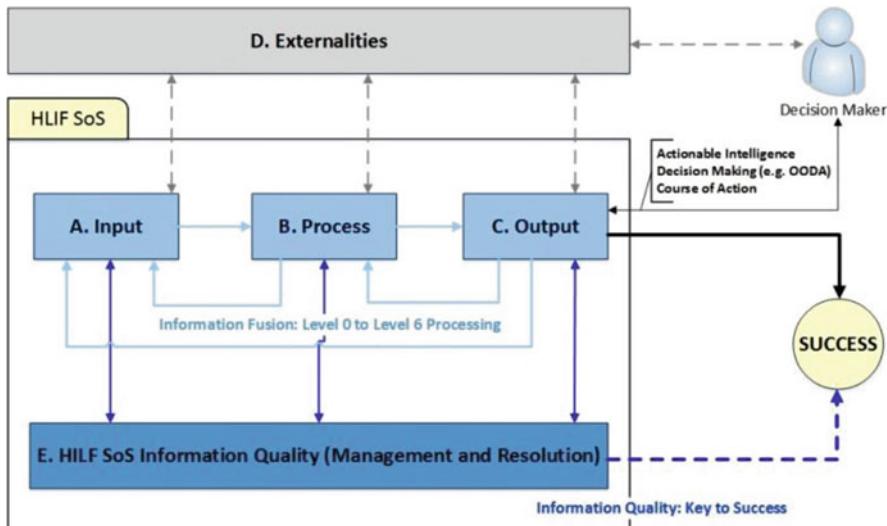


Fig. 20.3 Genesis of an IQ meta-model: placing terms in context of the HLIF domain

Table 20.3 HLIF information quality subdomain-related terms and concepts (Part I)

A	Input: Information Characterization
A.1 Sources	
No.	(Context Qualifier) Term < Alternative Term > [example]: Definition
1	(Sensor/Input) Phenomenology : A sensor's (or more generically, an information collection mechanism) inherent capability to collect and characterize information across a specific phenomenology. Analogous to how humans perceive their surroundings, e.g., touch, sight, sound, taste, smell. Examples of sensors collection modalities include radar, electro-optical, biometrics, multispectral, and infrared. Information collected via specialized sensors is by and large the purview of low-level fusion. However, the end-to-end fusion process that is part of HLIF SoS includes and extends these data concerns
2	Collection discipline : These are specialized collections types with methods and means honed to gather information along a highly focused area of concern; usually around which there is an underlying community of interest (COI) subject matter experts (SMEs) with specific domain skills and knowledge. Examples are open source intelligence (OSINT), imagery intelligence (IMINT), signals intelligence (SIGINT), geospatial intelligence (GEOINT), communications intelligence (COMINT), and moving objects intelligence (MOVINT). A sensor and its underlying phenomenology are usually designed with the objective of collecting information along one of these specialized collection disciplines
3	(Application) Domain : Specialized project, business, or organizational focus area for an information fusion SoS, e.g., defense, medical, transportation, marine biology, space exploration
4	Transdimensional information category : A high-level coarse categorization of information into (a) geospatial, (b) temporal, and (c) semantic (everything else not spatial or temporal) vectors. A multidimensional, i.e., 4 (space + time) + n (semantic features), framework which places any object whether real or abstract, simple or complex (multipart compound) in a space-time reference with all its semantic attributes fully characterized. The utility of this model is that it enables the abstraction of "objects" in a real-world space-time model to analyze its spatial-temporal dynamics and semantic changes within a spatial-temporal framework
A.2 Capture/Store/Representation	
No.	(Context Qualifier) Term < Alternative Term >: Definition
5	Information set : <Table> collection of objects. An information set comprises the collection of information elements of like-attributes and properties that are cohesively managed (e.g., stored, updated) or processed through an algorithm. An example is an information set representing a company's employees, containing the names (attributes) with the corresponding properties (e.g., string) in a specific format (e.g., relational)
6	(Information) Object : <Row/Tuple> [e.g., person]: lowest level of granularity, i.e., an information items, entry. An object comprises properties (i.e., attributes/features that characterizes it). An object is analogous to a "row" of a table in an RDBMS
7	(Object) Feature : <Column/Attribute> [e.g., person:age] describes the attributes or characteristics of the information. Used for qualifying an information object and its properties (e.g., name, string, valid values, range) equivalent to a column in a database table
8	(Feature) Property < Type > [e.g., person:age:number]: Information object features have properties which define their characteristics, e.g., string, number, integer, float. This concept is key in IQ as it pertains to IQ issues which arise from storing (including precision), algorithm efficiency, and especially interoperability

(continued)

Table 20.3 (continued)

A	Input: Information Characterization
9	(Property) Format [e.g., person:age:number:Integer (0–140)]: An extension to property which characterizes the representation, internal or external of the data, e.g., property: numeric, format: integer32, float; property: string, format: alphanumeric (A/N) with all lower caps, may use regular expressions, e.g., [a-z][0–9], meaning all lowercase characters and digits zero through nine with no special characters allowed. Useful for data cleansing operations to ensure that data complies with established formats and used for data transformations, e.g., feature: birthday, property: date, format DD/MM/YY to YYYY-MM-DD
10	(Information Set) Morphology : Related to the form, format, structure of data; e.g., the same data may be represented as text, binary, relational, XML, RDF etc. The same element (e.g., “birthday” represented as a string or as a specialized “date” property). This also includes encoding such as used to represent text in ASCII or EBCDIC (an older encoding, yet still in use, which may cause conversion information quality issues)
11	(Information Set) Type/Format : Akin to encoding, an information set type is associated with its underlying storage, encoding, and access representation format. Information set types include (a) audio (e.g., *.wav, *.mp3, *.aiff), (b) imagery (e.g., *.nift, *.jpeg, *.tiff, *.bmp), (c) video (e.g., *.avi, *.flv, *.wmv, *.mp4, *.mov), (d) structured and (e) unstructured information (e.g., *.ppt, *.doc, *.pdf), (f) relational, and (g) geospatial. For example, audio may include .wav, mpeg. It is common to use these information types as is (e.g., listen to a voice track or process them for specific words or to automatically generate a transcript). Likewise, imagery may be exploited to generate objects via for example, feature extraction algorithms such as buildings and cars and motion imagery, that is, video, to generate tracks from moving objects. Additionally, central to information quality, is that these types of data (i.e., their formats) are often further categorized by their compression characteristics. This may have a significant impact on algorithm processing and the ability to generate accurate and precise features as well as directly impact on information quality of the output of an algorithm
12	(Information Set) Characteristic : Related to information set type, but a high-level categorization for otherwise nondescript data set characteristics, for example, (a) Raw : usually a proprietary output format of a level 0/1 collection from a sensor or device; e.g., DSLR cameras may capture images in a raw format, as a staging process, or prior to converting the data to a standard format for further downstream processing or exploitation. (b) Unstructured : Information that has no inherent or explicit morphology, e.g., text data. Various information elements may be included in unstructured data, which is then required to be mined or processed to extract objects and features. However, some algorithms allow the native exploitation of unstructured data such as support vector machines for automatic categorization of documents, e.g., PDF, MS Word. (c) Homogeneous : A collection of commonly formatted information sets. By and large, information in data warehouses and data marts or in traditional database management systems may represent various “types”/“entities” of information, but are represented fairly homogeneously, e.g., stored in an RDBMS, which allows it to be managed in a common way. (d) Heterogeneous : A collection of diverse information sets, both morphologically and endemically (e.g., categorized by different phenomenologies (radar, electro-optical) or representation types (audio, video, text, binary, geospatial). A characteristic of heterogeneous information is that they are usually the domain of low-level fusion, i.e., object refinement, feature refinement, and require post-processing before they may be able to be comingled or fused with other types of data. (e) Information dimensionality : In machine learning and statistics, dimensionality reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. It can be divided into feature selection and feature extraction. An information set with numerous features/attributes may be characterized as a high-dimensional information set; certain algorithms may be adversely impacted by high dimensionality, and consequently IQ may suffer

Table 20.4 HLIF information quality subdomain-related terms and concepts (Part II)

B.	Process: information manipulation and actions
No.	(Context Qualifier) Term < Alternative Term > [example]: Definition
13	Information at rest: Information that is in a quiescent state, either stored as a source of reference or staged for subsequent processing. Information at rest may have specific related IQ requirements
14	Information in motion: Information that is actively being processed as in by an algorithm, or manipulated as in an ETL process, or transferred either internally between and among processes or fusion phases, or externally with other departments, or organizations, either being ingested or being disseminated. Information actively being read as input from various sources or written as output to various sources
15	(Information) Mutation: Changes in data characteristics as it undergoes transformation related to data preparation, data algorithms, data evolution, during either processing, data in motion, or exploitation. A key facet of IQ that is often overlooked and requires special IQ strategies to ensure that information is not degraded throughout the fusion life cycle
16	(Information) Processing: Any action undertaking where information is input and which results in either (a) the same information being output in different formats, morphologies, i.e., transformed, or (b) new information being generated, e.g., based on an algorithm report, or (c) where information is being ingested or transferred, whereby no change in either format or morphology is occurring, but rather data is “moved” from one point to another
17	(Information) Exploitation: The ability to infer, compute, or generate “additional” information or characterize information by examination or algorithmic processing, either simple or complex. This is the de facto purview of an information fusion system, as information is exploited throughout the different and iterative fusion levels via, e.g., enrichment, combination, fusion, recombination, and algorithmic processing
18	(Information) Pedigree < lineage>: Metadata about the information’s source, genesis, or origin. It is considered a subset of metadata that is specifically maintained for the purpose of documenting actions and impacts to the information as it moved (information in motion, information mutation) throughout the processing chain, e.g., algorithms, data profiling, transformation, migration, dissemination. This information is a key enabler to improving IQ as it allows tracing to a point or process in time or life cycle, where the IQ issue was likely introduced, thereby identifying the root cause and enabling IQ mitigation

depicted in the previous figure, namely input, process, output, externalities, and IQ management and resolution, respectively.

This extensive list of terms and definitions serves as (i) a draft proposal to be reviewed by the HLIF IQ community of interest in order to establish a common vernacular and a reference IQMM and (ii) to provide the requisite context to understand the concepts explored throughout this writing.

Table 20.6 captures the impact of externalities on an HLIF SoS; externalities represent a key, yet often under-considered dimension of information quality. Stakeholders often set measures of effectiveness based on strategic and tactical externalities.

A course of action should be set based on the dynamics of externalities, and in turn, the implication of the externalities should directly set information quality parameters for risk, relevance, urgency, and cost-benefit trade-offs.

Table 20.5 HLIF information quality subdomain-related terms and concepts (Part III)

C.	Output: strategic results and activities
No.	(Context Qualifier) Term < Alternative Term >: Definition
19	Actionable intelligence: Strategic and tactical information that is generated from various information collections and processed in such a way that enables a decision-maker to undertake an action that will lead to an advantageous outcome from the decision-maker's perspective
20	Course of action: A prescribed set of actions (steps) predicated on actionable intelligence that enables achieving the objectives as set forth by the decision-maker
21	Decision-making framework: A model presuming a rational approach, vis-à-vis intuitive approach, whereby the internal and external forces and dynamics, including the cognitive processes of the stakeholders, especially those innate to the decision-maker or the collective ones of the decision-making team, play a role in and influence the selection of a particular course of action from among several alternative possibilities. Decision-making frameworks in HLIF include models such as observe-orient-decide-act (OODA)

Table 20.6 HLIF information quality subdomain-related terms and concepts (Part IV)

D.	Externalities: external forces, dependencies, and interdependencies
No.	(Context Qualifier) Term < Alternative Term >: Definition
22	Compliance: Regulatory, policies, which may affect various facets of information management, e.g., dissemination, security, standards
23	Macroeconomics: Large-scale economic factors, usually at the national or global level, e.g., interest rates, trade, productivity. Whereas these considerations do not have a direct impact on the design or operation of an HLIF, macroeconomic factors may influence the stakeholders, and, accordingly, may alter requirements for, e.g., MOEs and MOPs, and, accordingly, may alter the thresholds or risk and cost-benefit analysis of actionable intelligence and constrict a course of action
24	Natural and environmental: Considerations for resiliency, disaster recovery, fail-over. A strategic architectural consideration may drive the selection of particular deployment regions, as in, e.g., cloud data center selections
25	Interorganizational: Considerations for interface control documents (ICD), access control, e.g., identity and access management architecture, interoperability, information lineage, and provenance, including establishing authoritative sources and information jurisdiction
26	Geopolitical: National and transnational factors, including dynamics which may result in limited or denied access to information and dynamics which may result in conflict and directly drive the urgency of actionable intelligence and create a higher risk and consequence for a given course of action

Table 20.7 lies at the heart of describing the IQ-specific problem and corresponding solution spaces. Outlining a granular meta-model will provide the greatest flexibility in analyzing and developing mitigation measures that align with the HLIF objectives.

The dichotomy of defining a general IQ framework and a domain-specific one lies at the heart of striking a balance between reaching a level that provides the broadest utility for many domains and the level that provides the deepest depth of

Table 20.7 HLIF information quality subdomain-related terms and concepts (Part V)

E.	Information Quality: Analysis, Management, and Remediation
No.	(Context Qualifier) Term < Alternative Term> : Definition
27	Information quality: Inherent or otherwise externalities-driven characteristics associated with information that is collected, stored, migrated, created, processed, and disseminated or produced and consumed in any fashion throughout the HLIF SoS life cycle that will have a primary or secondary impact on actionable intelligence, the decision-making framework, and the ability to generate a corresponding optimal course of action
28	Information quality concern: An information quality-related requirement. A quality-related issue in the problem space, which requires a corresponding mitigation or resolution in the solution space. Quality concerns are considered to be at the strategic level, and they may be grouped in dimensions to help categorization. An initial set of IQ concerns is presented in Table 20.2
29	Information quality dimension: Strategic concerns that help categorized IQ concerns as they relate to HLIF characteristics. Dimensions tie or groups IQ elements or IQ vectors to specific facets of a HLIF SoS, e.g., the aforementioned dimensions of impact, scope, composition, and evaluation referenced in Table 20.1. A specific view, contextual perspective, topic, or concern related to HLIF SoS IQ from which information quality elements or vectors may be elicited. Dimensions may be general or domain specific and may overlap. Identifying information dimensions is a useful way of eliciting IQ-related requirements. IQ dimensions may comprise IQ vectors, which in turn comprise IQ elements, but the association between IQ dimensions and IQ vectors is not as coupled, i.e., hierarchical, as between IQ vectors and IQ elements
30	Information quality vector: A related set of IQ elements; a logically cohesive grouping of IQ elements. IQ vectors may have commonalities among the element properties, their impact to the HLIF, and may be mitigated with similar approaches
31	Information quality element: A specific characteristic of an information quality concern that may impact actionable intelligence or a course of action. The lowest level (granular, atomic) characteristic of an information quality concern that is to be managed. IQ elements are addressed by specific IQ resolutions. Hierarchically they are defined as IQ dimension::IQ vector::IQ element
32	Information quality taxonomy: The hierarchical organization (tree-like structure) of IQ elements, vectors, and dimensions. The scheme within which IQ elements may be systematically classified (e.g., element properties) and categorized (e.g., IQ vectors)
33	Information quality ontology: A set of concepts, categories, and properties that qualify the relationships among IQ elements, IQ vectors, and IQ dimensions. It helps ascribe key dynamic interdependencies between and among constituent IQ elements that may not be described using solely a hierarchy in order to determine strategic and tactical mitigation approaches
34	Quality attributes: Quality attributes are the overall factors that affect runtime behavior, system design, and user experience. They represent areas of concern that have the potential for application-wide impact across layers and tiers. Some of these attributes are related to the overall system design, while others are specific to runtime, design time, or user-centric issues. The extent to which the application possesses a desired combination of quality attributes such as usability, performance, reliability, performance, scalability, availability, and security indicates the success of the design and the overall quality of the software application. They are usually architecturally significant requirements that demand the architects' attention

(continued)

Table 20.7 (continued)

E.	Information Quality: Analysis, Management, and Remediation
35	Quality impact assessment: The qualification or quantification of an impact that a particular IQ vector or element may have on the overall IQ objectives or targets/goals. Mitigation and resolution of impacts directly correlate to mission success and fit for purpose as set by requirements
36	Quality objectives: Strategic in nature, based on MOEs or equivalent
37	Quality targets < Goals>: Metrics, MOPs, TPMs, SLAs, or equivalents
38	Quality process: Any of a series of individual or interdependent IQ-related process, running in parallel or in tandem, such as analyzing, monitoring, reporting, and controlling quality of processes in the life cycle of an HLIF
39	Information quality governance: A body with the authority to establish policies and enforcing them with regard to planning, monitoring, and managing information assets
40	Information contamination: Propagation of information quality degradation from one element or vector to another due to intermingling, aggregation, processing, and feedback loops
41	Quality mitigation: A process to control quality via correction techniques. Comprises quality remediation and quality resolution. Intended to be used as a general term, i.e., not pertaining to a particular level of granularity of an IQ concern or issue. Quality mitigation may apply loosely to IQ dimensions, vectors, or elements
42	Quality remediation: A strategic IQ mitigation technique, as it may be applied specifically to an IQ to vector. A remediation is descriptive in nature. The full context is that IQ vectors have corresponding quality remediation approaches
43	Quality resolution: A tactical and very specific solution ascribed to an IQ element, i.e., for a specific scenario (including application). A specific quality resolution for one scenario of the same information may not be applicable or transferable to another scenario. Quality resolution is prescriptive in nature. The full context is that IQ elements have corresponding quality resolutions

knowledge to be directly applicable to a specific domain, e.g., defense. Therefore, this work strives to develop a multi-domain framework with defense domain-specific examples. Additionally, the intent is to generate an HLIF IQ vernacular, and only leverage the quality assurance/quality control (QA/QC) lexicon from traditional quality management where directly pertinent and useful, and avoid overloading these existing terms whenever and wherever possible.

20.3.2 Strategic Implications

Following is an expansion of a selected set of the aforementioned terms and concepts and their corresponding implication for information quality in HLIF SoS. These definitions enable a methodical analysis of the pertinent IQ considerations and facilitate their categorization within the problem and solution space and further the IQMM.

20.3.2.1 Actionable Intelligence

The key goal of actionable intelligence (AI) is to ensure that the decision-maker has access to the best actionable intelligence. This implies in part that AI should be *accurate* and *precise*. It is clear that these objectives can be traced to information quality concerns, such as the granularity or resolution of information sets and the predictive accuracy of forensic and predictive algorithms used. However, there are more subtleties involved, such as that information needs to be *relevant* and especially *timely*.

Information relevancy may drive several facets of IQ in HLIF such as the need to manage the collection of information (quantity and type) and to understand which information is actually contributing to the accuracy and precision of forensic and predictive algorithms.

Actionable intelligence is perishable, meaning it expires. Once AI is made available, it is only valid for a period of time, after which its value, or ability to act upon, degrades until it is no longer viable. This is especially true in scenarios where decision-makers are evaluating a course of action (COA) against an opponent or adversary. This is due to the fact that an adversary is actively engaging in countermeasures which may include compromising the integrity of the information or subverting the COA. This drives a significant number of IQ issues that need to be taken into consideration and correspondingly mitigated. Externalities, in general, influence the context of AI and the evaluation of COAs. It is recommended that actionable intelligence, and the corresponding course of action, take into consideration the dynamic nature brought about by an adversarial scenario [9, 37].

20.3.2.2 The Broader Context of Decision-Making

To complement the dynamics of AI and COA framed in the context of decision-making, two additional concepts are included. First, to represent the agility of the decision-making process, Boyd's observe-orient-decide-act (OODA) is leveraged. This decision model captures the agility requirements of the decision-making process which can then be mapped and integrated with the HLIF JDL-based model. It is considered advantageous to have the capability to execute the OODA loop more expeditiously than one's adversary.

This is often referenced as closing the OODA loop and can be a key to determining IQ-related measures of effectiveness (operational measures of success, i.e., achieving the mission operational objectives), measures of performance (measures which characterize physical or functional attributes that need to be met to achieve MOEs), key performance parameters (critical system capabilities that must be met for a system to achieve its operational goals), and technical performance measures (measures used to assess a design or component capability, enables the architecture components to be evaluated for compliance with requirements, and enables the ability to assess technical performance risk). Performance requirements

are usually divided into objective and threshold, representing desired performance and minimum required performance, respectively.

A second resource is a threat assessment quantification model based on capability, intent, and opportunity (COI) of an opponent [34]. The COI model helps define the information collection requirements and corresponding IQ concerns for its three information vectors, namely, capability (e.g., weapons, skills), intent (e.g., political climate), and opportunity (e.g., profiles for potential location, time, target, and victims). These two models in conjunction with the AI and COA considerations provide a rich source for identifying and even quantifying IQ needs.

Accordingly, in an HLIF SoS, information quality is directly or indirectly correlated to the extent it influences and partakes in the decision-making process and its results, and as such, several decision-making frameworks that apply to information fusion should be leveraged, e.g., Blasch et al. cognitive OODA [5].

20.3.3 Genesis of an IQ Meta-model: An Initial Proposal

Figure 20.4 depicts a proposed information quality meta-model (IQMM). This IQMM only covers a selected portion of the concepts and terms presented hitherto. The objective is to use this model as an enabler to document and help analyze the domain space. As the model is enriched throughout additional revisions and contributions from the information fusion community, it can also serve as a platform to exchange ideas and critiques that will help mature the IQ research. A meta-model promotes the organized understanding of a domain space, by capturing its primitives (fundamental concepts and terms, i.e., building blocks) and framing them with rules, constraints, and relationships. The development of an information quality meta-model enables the systematic analysis of the problem space and the corresponding synthesis of the solution space and is the cornerstone to developing a comprehensive ontology.

20.4 Identifying, Characterizing, and Organizing IQ

Identifying the key IQ elements in the domain of interest is arguably the first step to methodically address information quality shortcomings in an HLIF SoS. The broad nature of the applicability of information fusion to various domains, e.g., defense, medical, and transportation, will invariably result in the overloading information quality terms.

By *overloading*, it is meant that a quality element may be ambiguous due to its meaning being contingent on its usage context. For example, the quality concern of “integrity” may mean “consistency” or “completeness” in one context, yet in the context of information security, it means that information should not be altered, either by error or malice.

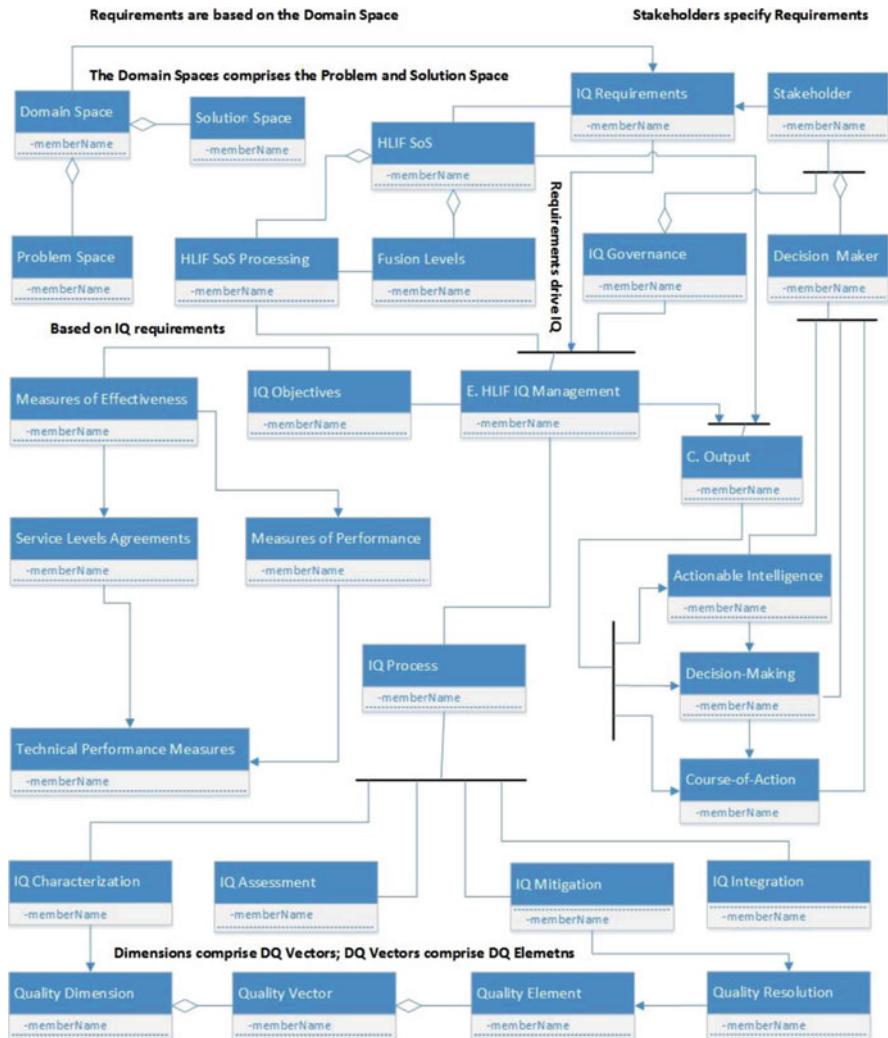


Fig. 20.4 Initial HLIF SoS information quality meta-model

Consequently, it is unlikely that there exists a single and absolute ontology that perfectly describes the problem and solution space for all domains. Nor should a single IQ ontology be necessarily desirable, due to its inherent monolithic implementation and respective lack of flexibility, which would make it a suboptimal solution, especially for an HLIF SoS which may span various domains. The use of *Namespaces* is recommended to characterize IQ elements within a specific domain.

Table 20.8 Characterizing information quality elements

Characterization	Property	Description
Objective	Judgment and personal bias do not significantly alter the characterization.	
	Measurable	(O M): Objectively and numerically measured or calculated, e.g., speed, length, time, distance
	Classifiable	(O C): Objectively categorized, e.g., color, blue, red
Subjective	Judgment and personal bias may significantly alter the characterization	
	Quantifiable	(S Q): Has an inherent quantifiable aspect but cannot be directly measured or calculated, e.g., small, large, medium; fast, slow. However, values may be subjectively attributed to the classifications and operations on them may still preserve the initial logic, e.g., adding two or more small objects
	Classifiable	(S C): Does not have an inherent quantifiable aspect and cannot be consistently categorized. Must use subjective descriptors to categorize, e.g., trustworthiness, competency, viability, completeness, secure. Values may be subjectively attributed and operations among these may lack traditional logic, making it rather ambiguous to infer anything for the resultant operation, e.g., what does it mean to aggregate something or someone with medium trustworthiness to something or someone semi-capable?

20.4.1 *Information Quality Elements: Characterization*

Attributes characterize information quality elements (IQEs) comparable to how information features::properties characterize information objects; e.g., age::number and name::string characterize data object “person.” A key concern in IQ is the ability to recombine (among disparate types), aggregate (among levels or layers of granularity), and propagate (throughout information exploitation processes) information quality measures.

Table 20.8 outlines proposed ways for systematically characterizing IQEs with the objective of facilitating IQ impact analysis.

The utility of this approach is to provide a framework for developing *common denominators* with which to perform meaningful numerical operation IQEs. By meaningful it is meant that the numerical operation will yield a logical, repeatable, and useful result which can be acted upon.

20.4.2 *Eliciting IQEs: Using Dimensions – Selected Examples*

To aid in the process of identifying IQ elements, the concepts presented in Sect. 20.3 Concepts and Primitives: Outlining an IQ Meta-Model are revisited and expounded to identify dimensions. An IQ dimension affords a unique perspective or vantage point of the HLIF SoS that helps in eliciting quality concerns and organizing them into information quality vectors and decomposing these into information quality

elements. This process entails both a top-down and bottoms-up approach and should be iterated until the desired level of granularity and fidelity is achieved.

The objective is to leverage and be able to dovetail with frameworks for information quality ontologies such as Rogova and Bossé' Ontology of Quality of Information Context [30], Chatschik et al. Quality of Information [2], Costa et al. URREF ontology [7], and Blasch et al. measures of effectiveness [4].

Using dimensions related to IQ functional requirements, and leveraging the previous concept of namespacing, will facilitate the process of combining taxonomies and ontologies [19, 24], thereby enriching the IQ body of work.

20.4.2.1 Quality Attributes: A System-of-Systems Perspective

Quality attributes (QAs) do not directly relate to information quality in the traditional sense, i.e., directly related to the underlying information, but rather are considered nonfunctional requirements of a system of systems.

Since QAs are usually not direct requirements levied upon an HLIF SoS, they are often overlooked for information quality impact.

The “-ilities” as they are also known is an extensive list, but the main reason for emphasizing them is that they are key to the information fusion SoS architecture. It is incumbent on the architects and stakeholders to identify the tie between these QAs and the corresponding IQ elements they drive. Table 20.9 provides an example of correlating IQs to QAs.

The following is a more comprehensive albeit not exhaustive list of quality attributes: accountability, accuracy, adaptability, administrability, affordability, agility, auditability, autonomy, compatibility, composability, configurability, correctness, credibility, customizability, discoverability, durability, extensibility, fault-tolerant, fidelity, flexibility, integrity, interoperability, maintainability, modifiability, portability, recoverability, repeatability, scalability, survivability, standards-compliance, simplicity, traceability, understandability, upgradability, vulnerability, usability, sustainability, etc.

Many of these QAs may in turn drive or be driven by MOEs, MOPs, and TPMs. The principal takeaway is that IQ permeates the entire HILF life cycle and

Table 20.9 QAs as drivers for identifying IQ elements

No.	Quality Attribute	Information Quality Element
1	Relevance	Contribution factors to algorithms, e.g., MDL, domain space
2	Reliability	Measures of objective and subjective data reliability
3	Accessibility	Constraints for access control, e.g., privileges, RBAC, ABAC
4	Availability	Information in motion, backups, redundancy
5	Usability	Common information formats, morphology
6	Interoperability	Interorganizational information exchanges
7	Scalability	Metadata for information dimensionality and volume impact

Table 20.10 Dimension: quality attributes

Dimension	IQ Vector	IQ Element	Attribute Characterization: Description
QAs	Availability ⁽²⁾	<i>Fail-over, resiliency, clustering concerns</i>	
		Configuration	O C ⁽¹⁾ :active-active/active passive
		Performance	O M: 99% (3.65 days downtime per year)
		PIT Recovery	O M: point in time recovery; 1 hour
	Accessibility	<i>Security-related concerns</i>	
		Confidentiality	O C: encryption, access control
		Integrity	O C: hashing
	Interoperability	<i>Information exchanges concerns</i>	
		Morphology	O C: structure
		Format	O C: storage

accordingly should not be a mere afterthought. Table 20.10 lists sample IQ vectors and elements for selected quality attributes.

Reference Table 20.18 Subjective Information Input Source Assessment: *Reliability* for descriptions of (O|M) and (O|C) (superscript 1).

Note that *availability* (superscript 2) is being used as both an IQ vector and an IQ element; however, namespacing may be used to provide the proper context for the usage of the terms, i.e., QAs::Availability vs. QAs::Accessibility::Availability. In the first instance, *availability* is presented in the context of an IQ vector, representing an IQ concern related to performance, such as determined by meantime between failures (MTBF) and meantime between repairs (MTBR). In the second case, *Availability* is presented in the context of an IQ element such as it pertains to a “security” IQ vector, e.g., a potential IQ threat of denial of service.

Therefore namespacing enables tuning the characterization of IQ concerns. Which one to choose depends on the particular scenario of the problem and solution space being considered, as well as the mitigation strategy; both “types” of *Availability* will impact information quality, especially with respect to timeliness of actionable intelligence, but they will likely require distinct resolutions.

20.4.2.2 The Static Nature of Information in HILF SoS

The static nature of information comprises the descriptive metadata for a particular information set. This description should be focused solely on the innate nature of the information itself, and not on its representation, such as morphology or format. Aspects of the information or information sets at rest need to be understood and qualified, e.g., is the information encrypted; will the information be used in its natural state; and how often will the information be refreshed.

Static nature or information at rest may also include interim staging areas between fusion processes or staging ingest and dissemination platforms. Table 20.11 provides sample vector and elements for the information at rest IQ dimension.

Table 20.11 Dimension: information at rest metadata (innate IQ properties)

Dimension	IQ Vector	IQ Element	Attribute Characterization: Description
Info-at-Rest	Temporal	<i>Info sets/objects with temporal attributes, i.e., time</i>	
		Resolution	O M: what is the resolution of temporal attributes? This may be important for interval analysis or sequence analysis such as Allen's Temporal Intervals [1]
	Geospatial	<i>Info sets/objects with geospatial attributes, i.e. coordinates</i>	
		Resolution	O M: e.g. for area, is the resolution square inch, meter, kilometer?
	Semantic	<i>Info sets/objects with semantic attributes, e.g., color, names</i>	
		Granularity	O C / S C: at what level of granularity has the information for the object been collected, e.g., city, county, state, country?
		Collection discipline	O C: e.g., HUMINT, OSINT
		Collection date	O M: may be relevant in establishing timeline information or setting policies to age data sets, e.g., archive, purge
	Digital Audio	Raw data sets (i.e., still need to undergo object/feature extraction)	
		Bit depth	O M: determines the dynamic range, difference between loudest and softest sound; analogous to intensity
		Sampling frequency	O M: determines the lowest and highest pitch that can be stored

There is no end to the quantity of metadata that can be used to describe the relatively static properties of objects or their corresponding information sets. There is also no de jure method for identifying and attributing a hierarchy of quality elements.

As before, the recommendation is that the identification of IQEs and their allocation into IQVs be driven by requirements and bottoms-up and top-down evaluation of how they impact actionable intelligence.

20.4.2.3 The Dynamic Nature of Information in HILF SoS

The dynamic nature of information is arguably one of the most overlooked concerns in information quality. The complexity of information quality quickly multiplies when considerations such as how specific information quality properties may impact the predictive capability of algorithms; e.g., some algorithms can handle multidimensionality very well, while others do not; some algorithms may handle sparsity very well, e.g., blanks or empty fields, while others do not.

The dynamics of information flows should also be considered such as (a) the cascading effects, e.g., what happens when information mutates as it passes sequentially from one algorithm to the next or loops back to some previous step or

phase in the fusion process, and (b) flows into a fusion process, which inputs may be the aggregate of multiple previous processing results, each of them with different degrees of quality. This quality divergence among multiple inputs raising concerns such as (a) what is the resultant information quality, or should the information even be mixed or amalgamated?; (b) what are the consequences of using heterogeneous information or mixing information with various degrees of granularity; i.e., some algorithms are more susceptible than others when it comes to rounding errors or aggregating information with diverse resolution.

Other considerations of information in motion include volumetrics, i.e., the amount of information to be processed, which represents a sizing and performance concern, and other storage and database requirements, such as sizing of data and caches and buffers, as well as backup and retention policies, as it goes through the pipeline of tasking, collection, processing, exploitation, and dissemination (TCPED) [35]. Orchestration also plays a critical role in achieving quality and reliability in an end-to-end TCPED information fusion enterprise [35]. Accordingly, metadata regarding these information-in-motion dynamics need to be included as IQEs.

Note that interoperability was also a vector identified in the quality attributes dimension; this is not incongruent to itself, but it reflects that (a) there may be concerns regarding interoperability, which may impact facets of various quality dimensions, i.e., IQ vectors, and as before, how to best organize them depends on how they impact actionable intelligence or the fusion process as well as to how to mitigate them. It is beneficial to get as granular as possible in order to better quantify and mitigate the issues.

Table 20.12 introduces information-in-motion IQ concerns. Superscript (1): Reference additional definition of terms in Table 20.3 HLIF information quality subdomain-related terms and concepts (Part I), Section A.2 Capture/Store/Representation.

Table 20.12 Dimension: information in motion

Dimension	IQ Vector	IQ Element	Attribute Characterization: Description
Info in Motion	Pedigree	<i>Considerations for provenance and processing history</i>	
		History	O C: Lineage information, transactions, etc
		Source	O M: Original source, point of contact
		Versions	O C: Version history where used
	Interoperability	<i>Considerations for characterizing information sets</i>	
		Type ⁽¹⁾	O C: Applicable types
		Morphology ⁽¹⁾	O C: Applicable morphology
		Characteristic ⁽¹⁾	O C: Applicable characteristics
		Standards	O C: Applicable standards
	Volumetrics	<i>Considerations for performance and management impacts</i>	
		Size	O M: Megabytes, gigabytes
		Throughput	O M: Bytes per second, pixels per hour

Note that whereas information set characterization also applies to information at rest, this metadata or IQ elements need to be captured for each step that undergoes information mutation.

20.4.2.4 Real-Time and Streaming Information Considerations in HILF SoS

At this point, it is opportune to address another facet of information in motion and that is real-time information or streaming information acquisition and processing. Examples include moving target indicators (MOVINT) or detection and track generation and other streaming information, e.g., audio and video. Systems such as supervisory control and data acquisition (SCADA) from industrial control systems (ICS) may have diverse proprietary configurations, and managing trends in real time is key to achieving information quality; subsequently real-time information acquisitions and input from SCADA are also subject to information quality considerations [38].

It is understood that certain formats or protocols may encompass compression (superscript 1) specification as well. Notwithstanding, it is useful to consider compression explicitly or at least compression options within different protocols or format standards, as it is analogous to “resolution” and consequently may determine how much the information may be exploited (note that the exception may be lossless compression).

Regarding support data (superscript 2), it is often the case that ancillary information is streamed or ingested in parallel. Sometimes ingest is made with complementary information and associated exploitation support data (ESD) is made available before or after the streaming event. There are two types of exploitation support data: (a) telemetry or acquisition data may be streamed that indicates the health and status, as well as control capabilities of the primary information acquisition, and (b) ESD, which is associated with the ability to exploit data, usually in downstream processes. The sensitivity to these quality elements are contingent on the real-time ingest requirements, processing timelines, and, of course, the critical lead time needed to derive actionable intelligence (Table 20.13).

20.4.2.5 Externalities Considerations in HILF SoS

Previously high-level information fusion “systems” have been more aptly defined as system of systems (SoS); this consideration is critical when considering externalities, as systems in the end-to-end workflow from fusion level 0/1 to level 6 comprise an intricate, interconnected, and interdependent maze of not only separate systems but fundamentally separate organizations. This means that interorganizational challenges including the chain of custody, information ownership, deconfliction policies (e.g., which organization is considered the “master” when synchronizing

Table 20.13 Dimension: real-time and streaming data

Dimension	IQ Vector	IQ Element	Attribute Characterization: Description
Real Time	Ingest	<i>Considerations for real-time acquisition and processing</i>	
		Data rates	O M: e.g., bits per second, frames per second
		Compression ⁽¹⁾	O M: contingent on data type
		Storage	O M/S: e.g., buffering, storage
	Support Data	<i>Ancillary data needed for exploitation, e.g., telemetry, esd⁽²⁾</i>	
		Delay	O M: delta in (s, min)
		Utility	O/S C: attributes used for processing aid
	Interoperability	<i>Considerations for real-time interoperability and integration</i>	
		Protocols	O C: encryption, access control
		Format	O C: hashing
		Structure	O C: description

Table 20.14 Dimension: externalities

Dimension	IQ vector	IQ element	Attribute characterization: description
Externalities	Data transfer	<i>Enterprise data transfers</i>	
		Network rate	O M: bits per second
		Protocol	O C: e.g., TCP, UDP, JAS ⁽¹⁾
		Replication	O C: deconfliction rules, master-master ⁽²⁾
		Buffer/throttle	O M: e.g., amount of buffer and throttle
	Policies ⁽³⁾	<i>Regulatory, statutory, and interorganizational policies</i>	
		Standards	O C: format, transfer, APIs
		Privacy/integrity	O C: hashing, PKI, certifications ⁽⁴⁾
		Retention rules	O M: archive type ⁽⁵⁾ and duration rules

repositories across several sources), as well as regulatory policies, have to be managed at both the national and sometimes transnational level, e.g., European Union.

Furthermore, the complexities can be compounded quickly when considering that many organizations are moving to the Cloud, which means at the very least this implies a third-party infrastructure as a system (IaaS), and likely platform as a system (PaaS), and in many cases software as a system (SaaS) as well. This paradigm shift has numerous implications for information quality in HLIF SoS, as it adds yet another party, for which information security, performance (both processing and network), information integrity, and availability need to be considered. Here is where systems engineering and system requirements for information quality come into play; performance requirements need to be decomposed and flowed down, and the right service-level agreements need to be negotiated. On the pro side of third party IaaS argument is of course that it offers a solid infrastructure that usually includes fail-over and disaster recover (DR) capabilities and, as such, greatly reduces the burden of designing and implementing these from scratch (Table 20.14).

The Joint Architecture Study (JAS) (superscript 1) protocol [12, 20] uses the lower-level SpaceWire data link layer to provide reliable packet delivery services to higher-level host application processes. It is an example protocol implementation based on information quality for both performance and reliability, critical by itself when having to ensure intrasystem performance but more so when transferring data among external segments.

Often, certain nodes or segments in an HLIF SoS which span (by definition) multiple organizations are designated and implemented as information sinks (specialized nodes that receive information from other nodes and combine them and serve as a centralized master repository). This of course brings about serious considerations for information quality with respect to synchronizing timestamps and reconciling differences; this becomes the purview of information replication, where rules must be set up in a master-master or master-slave information replication environment. These rules include setting up an authoritative hierarchy, such as when two records with different values are ingested they may be reconciled, e.g., by latest or first timestamp or by location, i.e., the one considered authoritative.

Policies (superscript 3) affect many areas, especially security, retention, and data exchanges. The key difference, vis-à-vis implementing the same functionality without being driven by policies is for example, the additional effort required when implementing security controls such as auditing and logging, and security accreditation to ensure compliance.

Although closely related to security, requirements such as public key infrastructure (PKI) (superscript 4) may drive additional quality-related requirements such as setting revocation lists and setting up certificate authorities. Clearly this last example is within the purview of security engineering; however, in system of systems, security requirements usually flow down to the components implementing the functionality, and impacts to quality and performance are properly allocated. Accordingly, this aspect is more of an indirect impact to information quality, but nevertheless one that cannot be ignored.

Archiving rules (superscript 5) play a critical role in complying with policies for retention. Additionally, the selection of archiving mechanism has a direct impact on availability and performance.

20.4.2.6 Fusion Processing Metadata: Algorithm-Information pairing

There may also be both (a) initial concerns, to be addressed at the outset of a design, and (b) operational concerns, to be managed as the system matures. For example, after an HLIF SoS begins to use information sets, some may be marked or tagged as more effective than others for use with certain algorithm or processes.

A salient characteristic of information fusion is not only the information quality concerns regarding “input” information, but when it comes down to performance and predictive accuracy, information also plays a role in both performance and predictive capability of an algorithm; e.g., some algorithms do well with high dimensionality (e.g., number of independent variables), while others do not; some

Table 20.15 Dimension: processing metadata

Dimension	IQ vector	IQ element	Attribute characterization: description
Algorithm Profiles			
Performance	<i>Processing performance based on data type</i>		
	Algorithm data	O C: matrix of processing ⁽¹⁾	
Accuracy	<i>security related concerns</i>		
	Algorithm data	O M: matrix of results, e.g., ROC Type I/II ⁽²⁾	

algorithms do well with sparsely populated information, e.g., “blanks” for a field (i.e., missing values), while others do not.

Additionally, information fusion is seldom the application of a single process or algorithms but rather an exhaustive interdependent and not always sequential chain of processes and algorithms throughout the L0 to L3 fusion levels. Gaining insight into information quality impacts within the steps of the end-to-end flow enables an optimal result and significantly contributes to fusion level 4, i.e., process refinement.

Accordingly, a key concept for improving fusion quality is the process of classifying algorithms themselves and cataloging the results for discovery, that is to create an algorithm marketplace, where various users can inquire as to which algorithms perform better with certain data and to pair information to processes and algorithms, based on profile and characteristics which will yield optimal results (Table 20.15).

A matrix is an efficient way of mapping algorithms to data types and qualifying both processing performance and algorithm accuracy. Additional information such as the volume of data processed is useful in order to normalize the entries, e.g., minutes per 100 entries or minutes per megabyte.

The accuracy matrix should include metadata on the control set used to create the predictive model. The accuracy itself should include both Type I (false positives) and Type II (false negatives) errors. Finally, algorithms and processes should be categorized, e.g., forensic modeling, data cleansing, and predictive modeling, and within these allow for even more granular breakdown, e.g., predictive modeling (parametric, nonparametric, Naive Bayes, k-nearest neighbor, majority classifier, support vector machines, neural networks), to further allow for fine tuning of fusion level 4 (process refinement).

20.4.3 Organizing IQEs: IQ Vectors, Taxonomies, and Ontologies

It stands to reason that the previous sections are only selective and nominal examples of how to elicit IQEs by leveraging the organized dimensions of concern driven by the HILF SoS requirements.

A comprehensive example is beyond the scope of this writing and even impractical, as it surely would have to be vetted by an engineering or architecture board overseeing the particular design of the system. This implies that whereas there may be a large commonality of information quality concerns across systems, the particulars of HLIF SoS may drive very unique IQ requirements.

Akin to the information quality meta-model presented in Sect. 20.3.3, it is recommended to organize information quality dimensions, vectors, and elements using primarily taxonomies and ontologies. To manage the complexity of information quality commensurate with an enterprise system, it is recommended that an ontology tool be used as part of a formal process. The advantage of organizing IQs concerns is to facilitate the discovery of patterns and identify areas of overlap or gaps. Additionally, this effort will set the stage for the next step which is to identify impacts correlated with these IQ concerns.

20.5 Assessing and Mitigating the Impact of IQ

HLIF SoS is by definition multidimensional, multilayered, and multiprocess, with both parallel and sequential paths. Information quality assessment and corresponding mitigation follow the same complex pattern as a high-level information fusion SoS itself. No single cohesive or holistic definition exists or, for that matter, should exist, as doing so may preclude optimization, meaning that the complexity is such that a general approach would likely be inefficient and ineffective.

Having laid the foundation of IQ meta-model and IQ characterization in the previous section, it now stands to reason that the next step is to decompose and classify IQ considerations based on the IQ elements, vectors, and dimensions. There is no single metric that can capture this complexity, and there is no single approach to aggregate IQ metrics. Although the consequences of poor information quality issues throughout the system will manifest themselves at the output, i.e., actionable intelligence; this does not imply that the root cause is a single enterprise issue.

The more pertinent approach is to frame information quality assessment in the proper context, e.g., what is the process and what corresponding resolution should be applied. Also, “good enough” is often more actionable than chasing an open-ended information quality concern.

Another example of misapplying IQ at the enterprise level is that of information sparseness; it may seem desirable to have as much information and all fields and attributes properly populated. However, some algorithms actually perform fairly well with sparse information, and others will not necessarily provide a proportional return toward a downstream result. So framing the question “what is the information quality profile required for a particular algorithm or process?” is a more relevant tactical question that may serve the strategic HLIF SoS IQ concerns as well.

IQ assessment needs to be framed in the context of its IQ dimensions and the point of diminishing returns needs to be ascertained. Excessive quality controls without a commensurate return represent an opportunity cost, i.e., an extra step in

the end-to-end process that is made at the expense of not accomplishing something else that may yield better results. In the final analysis, the impact assessment and mitigation trade-space need to be juxtaposed with time and cost considerations and the expected improvements on IQ, all of which should be congruent with the objectives of the mission as set by the decision-maker. The cost of IQ mitigation should be commensurate with its returns.

20.5.1 Recapitulation of Traditional IQ Mitigation Techniques

There are significant advances in understanding quality and uncertainty for low-level fusion; for many sensor types, the science is fairly mature with regard to the mathematical rigor, e.g., random and systematic (bias error), the ability to compute error ellipse, and other quantifiable analysis regarding issues that impact IQ in sensors. Other facets related to quality can be measured with high confidence, e.g., the spatial and temporal resolution of information collections; and certainly these should be included in categorizing the accuracy and precision of data that is used in HLIF as well.

Additionally, several facets of traditional data warehousing and data marts quality assessment for business intelligence (BI) should be leveraged such as data profiling, which is usually done via extract-transform-load (ETL) tools. A significant amount of information quality issues that impact BI also apply HLIF SoS, and as such, these tools and techniques should be applied where applicable.

Table 20.16 lists a recapitulation of these better known issues, their *impact* as related to HLIF SoS and their corresponding *mitigation* techniques, such as data cleansing. While these issues and resolution are common in BI applications, their impacts to HLIF are even more profound, due to the continual information exploitation processes which then may cascade quality errors downstream and which, furthermore, may prevent the ability to perform data enrichment, whereby additional relevant information is identified and associated with other information elements.

Granularity (superscript 1) is an often overlooked information quality concern. Granular data is a key enabler to profiling data for distribution, e.g., numerical, date ranges, and geospatial granular hierarchy, e.g., city, county, state, and country. This is used in data marts, and business analytics, but essential for information fusion, which enables analytics with fine-grain slices, which often yields improved results over information sets where various granularities are comingled.

20.5.2 Measuring Quality: There Is No Enterprise Silver Bullet

The case for information quality has been supported by the fusion community, yet a key question still lingers with regard to how to ascribe an HLIF SoS overall value to

Table 20.16 Information profiling: basic assessment and mitigation for IQ issues

IQ Element	Scenario/Impact	Mitigation
Duplicates	Multiple entries for the “same” entity, produces inconsistent results and generates poor models	Identification of duplicate entries, creation of primary keys
Spelling Variants	Variants for information representations includes spelling errors. Prevents consistent analysis and generates poor models	Soundex rules (phonetic coding) and mappings to consolidate variants into one consistent value
Blanks/Nulls/Zero	Missing values, sparse data, attributes (columns) for entries (rows) may be lacking values. Prevents accurate model generation. Impacts algorithm processing for classifiers	Generation of defaults based on profiling rules or mapping to specific values based on other populated fields. Other rules include using average or mode
Distributions	Min, max, ranges. Invalid information elements or outliers. This will skew algorithms	Distributions are critical for numeric fields. It helps identify outliers and ensure that the collected information represents the population statistics; accordingly, the use of classical statistics is paramount to ensure statistical valid data sets Additionally, especially if the information fusion node is a large data hub collector, it is critical to validate the data against known numeric ranges (coordinates, age, speed) or known values (geographic locations)
Reformatting	The worst-case scenario is when the same field has inconsistent formatting, e.g., a field string may contain dates in various formats, e.g., mm//dd/yy or yyyy/dd//mm, or a combination with timestamp, which requires in many cases time zone adjustment	Use of mapping or transformation scripts. A common technique is the use of XML style sheet language transformations (XSLTs), which cannot only automate conversions, but also provide transformations between XML document types
Granularity ⁽¹⁾	This is less evident than other quality issues; it concerns collecting data and bundling them into one superset. This may cause suboptimal predictors. The classic example is to combine bimodal data or data at different resolutions, e.g. by city, county, state, country, all in one data set	Use of classical statistics can easily reveal bimodal data, which appears as distinct peaks in the probability density function. For categorical information, especially temporal and geospatial, it is best to generate data cubes

information quality. Throughout this writing, information quality has been framed within the context of achieving actionable intelligence. However, IQ does not stand in a vacuum, so the recommended question that should be addressed is: “What is the appropriate IQ for the system given a particular AI objective?” Meaning IQ that may be good enough for one scenario may not be necessarily appropriate for

another. And perhaps this should be the approach, as it is from this perspective that we can mitigate many vexing IQ shortcoming and answer the persistent question of how to assign values to IQ.

HLIF IQ should be assessed in context of achieving an AI objective. Accordingly, metrics should include the context with which to assess the specific impact as well as the overall impact. IQ elements such as timeliness, and relevance, and resolution may qualify information but not necessarily represent an information quality issue unless specific impact may be attributed. Accordingly, a key IQ characterization is that of relevance – with respect to achieving AI and COA objectives.

Yet relevance may not be assessed until after processing, such as doing predictive fit, covariance, and minimum description length (MDL); that is, certain type of quality impacts such as those related to algorithm processing may only be assessed *a posteriori*. This information should then be used in feedback loop to improve the overall fusion IQ as part of fusion level 4, i.e., process refinement.

Accordingly, de Villiers et al. implement the URREF ontology and introduce a key concept of data criterion with weights based on IQ elements such as problem relevance and credibility [8]. This may serve to identify key performance parameters (KPPs) be monitored and managed throughout the various fusion levels.

Ultimately, there may be general terms to IQ based on how useful the actionable intelligence is and how effective the course of action was with respect to a certain objective. In this vein, “satisfaction” information should be collected; satisfaction information serves as future feedback to qualify a particular scenario, at a particular point in time and compare results given by various processes and information. Additionally, the value in collecting satisfaction information is to qualify the overall success of the mission and find key attributes that may be traced to contributing IQ concerns.

20.5.3 Assessing an HLIF SoS Information Quality Maturity

It is critical to consistently assess IQ throughout the fusion processes. Uncertainty introduced into the system by using subjective IQ measures results in propagating errors throughout the information exploitation value chain. To the degree that IQ processes are integrated within the HLIF life cycle itself, it may lead to amplification of uncertainty through feedback loops; e.g., IQ issues present at the outset are then reintroduced in subsequent estimations. A consistent set of quantifiable IQ measures correlating information issues to their corresponding results will enable quality to be systematically monitored and managed.

It is proposed that the overall IQ maturity be assessed similarly to systems maturity; i.e., IQ implementations, just like systems, have a life cycle, and as such, they mature from the conceptual phases to operational systems.

One way of considering the reliability of technology devices of system is to leverage the technology readiness level (TRL). A variant, called the system readiness level (SRL) [32], is more apt to qualify a maturity system of systems such

Table 20.17 System readiness levels/quality readiness level

Level	S: SRL description/Q: QRL description
SRL/QRL 1	S: System concept: system functionality qualitatively understood Q: IQ requirements identified
SRL/QRL 2	S: System technologies: technology and implementation understood Q: IQ dimensions, vectors, and elements identified
SRL/QRL 3	S: System proof of concept: experimental evidence Q: IQ requirements decomposed and allocated to subsystems and components
SRL/QRL 4	S: Component verification: components built and tested in laboratory Q: IQ assessment implemented at select fusion levels
SRL/QRL 5	S: Component validation: components tested in relevant environment Q: IQ proof of concept validated against selected components
SRL/QRL 6	S: Prototype demonstration: prototype in relevant environment Q: IQ demonstrated for selected subsystems against relevant scenario
SRL/QRL 7	S: Operational demonstration: integrated prototype operational Q: Quality mitigation applied and integrated at various levels
SRL/QRL 8	S: Actual system demonstration: representative system demonstrated Q: Quality mitigation applied at various levels by not integrated
SRL/QRL 9	S: Operational system: production system in operational environment Q: Fully integrated quality process in lockstep with fusion level 4

as an HLIF, as it takes into consideration its components or systems as well and the interfaces among them, and the corresponding required integration. This provides a valuable framework to apply to a subjective system which may not lend itself to be aptly quantifiable. Additionally, as previously mentioned, information quality is a function of how the system is fit for purpose with respect to a particular objective, and as such, IQ is contingent on the system readiness or maturity level. Table 20.17 proposes a quality readiness level (QRL) tiered approach that is congruent and commensurate with its corresponding system readiness level.

20.5.4 Nonquantifiable IQ: A Model for Subjective Assessment

Adjudicating subjective parameters is an extremely delicate effort; until there are enough scenarios where the inputs may be evaluated for fit for purpose with regards to the AI objectives, allocating subjective parameters may not provide pragmatic value; i.e., it may not be possible to validate the values. Furthermore, the more cascading subjective evaluations are made, i.e., the more subjective IQ measures feed into other processes, the more the meaning of a quality assessment may be diluted.

This section walks through an example of how to attribute subjective criteria and generate a more complex model that relies on qualitative constructs more so than quantitative measurements.

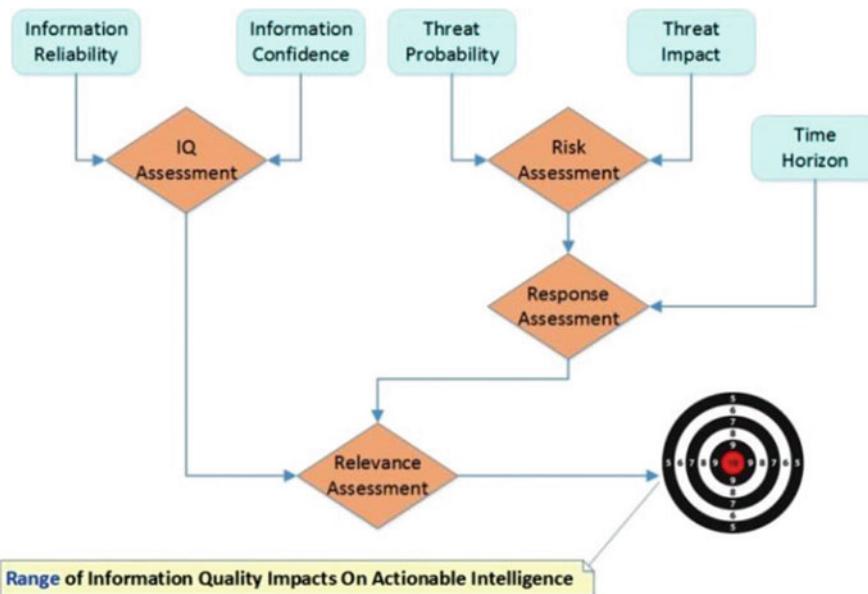


Fig. 20.5 Information quality impact on time-critical AI evaluation

Figure 20.5 depicts a likely workflow of HLIF SoS, whereby the impact of information quality on time-critical actionable intelligence is assessed. This workflow depicts the complexity of aggregating subjective assessments and highlights how the meaning is diluted when cascading multiple decision nodes. Note that IQ impact and assessment is usually a range. This range may serve to categorize mitigation strategies.

The utility of this assessment approach is that it ties several aspects of information fusion with regard to actionable intelligence and course of action. The threat probability and impact relate to fusion level 3 and, placed into the context of a time horizon, represent a decision-maker's context for how to respond, i.e., the OODA loop.

Table 20.18 provides a tiered approach for subjectively assessing an information source's reliability. This assessment approach is applicable to information generated from both human sources and information generated by technologies of diverse maturities; the corresponding SRL assignments are shown. Weight and scale are additional parameters that help fine-tune the model.

Weight represents a measure of proportion with respect to the other criteria, i.e., trustworthy to not trustworthy, whereas scale is used to tune that particular criteria as appropriate for a given scenario.

The net effect would be the same as using a single factor denoting proportion, but the advantage is that it provides flexibility in combining measures either by multiplication or addition and assessing the end results with respect to the tuning

Table 20.18 Subjective information input source assessment: reliability

Criteria	Description	Weight	Scale
Trustworthy	HUMINT: No doubt about the source's authenticity, trustworthiness, or competency. History of complete reliability. SRL: 9	3	5
Dependable	Minor doubts. History of mostly valid information. SRL: 7, 8	2	4
Neutral	This may be assigned to a new source that has undergone interim vetting only; should not be confused with "undetermined." SRL: 6	2	3
Not Dependable	Significant doubts. Provided valid information in the past SRL: 3, 4, 5	2	2
Not Trustworthy	Lacks authenticity, trustworthiness, and competency. History of invalid information. SRL: 1, 2	1	1
Undetermined	Insufficient information to evaluate reliability. May or may not be reliable	X	TBD

Table 20.19 Subjective alternate corroborating source assessment: confidence

Criteria	Description	Weight	Scale
Probable	Greater than 95% confidence. Confirmed: logical, consistent with other relevant information, confirmed by independent sources	1	5
Likely	Greater than 75% but less than 95%. Probably true: logical, consistent with other relevant information, not confirmed	1	4
Neutral	Greater than 25% but less than 75%. Possibly true: reasonably logical, agrees with some relevant information, not confirmed	1	3
Not Likely	Greater than 5% but less than 25%. Doubtfully true: not logical but possible, no other information on the subject, not confirmed	1	2
Not Probable	Less than 5%. Improbable: not logical, contradicted by other relevant information	1	1
Undetermined		X	TBD

parameters and adjusting them if necessary. Note that the criteria are meant to be ordered; it follows then that any combination of weight and scale must also preserve the order, albeit the relationships between the set of ordered criteria need not be linear.

Table 20.19 describes adjudication guidelines for confidence, which includes the ability to corroborate with other sources or validate with other evaluation frameworks. Probability definitions have been based on statistical significance, but may be adjusted as well as weight and scale to tune the model based on feedback, i.e., fusion level 4.

Figure 20.6 depicts the resulting matrix of reliability and confidence and provides a recommendation for classifying the results in five tiers, ranging from *Best* to *Bad*, with corresponding actionable directives. This example highlights that all facets of categorizing subjective data are very broad in nature, but the takeaway is that if

CONFIDENCE	Description	RELIABILITY						INFORMATION QUALITY ASSESSMENT					
		Unknown	Not Trustworthy	Not Dependable	Neutral	Dependable	Trustworthy	Recommended Action:					
		Weight	X	1	2	3	4	5	1. Best: Use Information Set as is.	2. Good: Use, but look at opportunity to improve. Flag issues.	3. Average: Proceed with caution - Do not use as is. Remediate.	4. Below Average: Use only under extreme factors. Archive.	5. Bad: Avoid or Discard the Information Set. Quarantine.
Scale	TBD	1	2	2	2	2	3		Classification Distribution: Equal Intervals (Max / No. Categories)				
Probable	5	1	TBD	5	20	30	40	75	Classification	Ceiling	Range	Description	Total
Likely	4	1	TBD	4	16	24	32	60	1	75	=> 60 <= 75	BEST	2
Neutral	3	1	TBD	3	12	18	24	45	2	60	=> 45 < 60	GOOD	1
Not Likely	2	1	TBD	2	8	12	16	30	3	45	=> 30 < 45	AVERAGE	4
Not Probable	1	1	TBD	1	4	6	8	15	4	30	=> 15 < 30	BELOW AVERAGE	7
Unknown	X	TBD	N/A	TBD	TBD	TBD	TBD	TBD	5	15	=> 1 < 15	BAD	11
													25

Fig. 20.6 IQ assessment based on reliability and confidence

PROBABILITY	Description	THREAT						IMPACT ASSESSMENT (RISK)					
		Unknown	Low	Medium Low	Moderate	Medium High	High	Recommended Action:					
		Weight	X	1	2	3	4	5	1. Catastrophic: cannot recover; long-lasting damage; Top Priority	2. Severe: significant consequences; Heightened precautions	3. Moderate: can recover in case of failure; Actively Manage	4. Minor: no lasting consequences; Monitor only and reevaluate	5. Negligible: Insignificant impact; Acknowledge only in Risk Register
Scale	TBD	1	1	1	1	1	3		Classification Distribution = Custom (Focus on High-Values)				
Probable	5	2	TBD	10	20	30	40	150	Classification	Ceiling	Range	Description	Total
Likely	4	1	TBD	4	8	12	16	60	1	75	=> 60 <= 150	CATASTROPIC	2
Neutral	3	1	TBD	3	6	9	12	45	2	60	=> 30 < 60	SEVERE	4
Not Likely	2	1	TBD	2	4	6	8	30	3	45	=> 15 < 30	MODERATE	5
Not Probable	1	1	TBD	1	2	3	4	15	4	30	=> 5 < 15	MINOR	6
Unknown	X	TBD	N/A	TBD	TBD	TBD	TBD	TBD	5	15	=> 1 < 5	NEGLIGIBLE	8
													25

Fig. 20.7 Impact assessment based on threat and probability

the criteria can be bounded, then pragmatic actions to mitigate information quality may be implemented. Currently, there is no clear definition for undetermined, and the recommendation is that information, for which neither confidence nor reliability can be ascertained, be quarantined and thereby excluded from processing until its pedigree can be duly assessed.

Figure 20.7 qualifies the impact of information quality based on risk. Risk assessment is based on threat impact and probability, which is a common approach in business management for mitigating various types of risks, such as cost, schedule, and technical risk.

In the case of information fusion, the level of risk for a particular scenario for which actionable intelligence is sought is what needs to be assessed and managed. A higher risk scenario commands more diligence on IQ. Note that the classification of the resulting matrix, i.e., impact assessment results, has been modified to focus on high values. This strategy helps to more effectively assign resources for the critical areas, i.e., *catastrophic* and *severe* impacts.

Figure 20.8 depicts the resulting assessment matrix from combining risk level with a time horizon. This assessment can be used to properly qualify the timeliness of information that plays a role in determining the corresponding AI and course of

	Description	RISK ASSESSMENT						Classification Distribution = Custom (Focus on High-Values)	Total
		Unknown	Negligible	Minor	Moderate	Severe	Catastrophic		
		Weight	X	1	2	3	4	5	
TIME-HORIZON	Immediate	5	4	TBD	20	40	60	80	300
	Impending	4	3	TBD	12	24	36	48	180
	Near Term	3	2	TBD	6	12	18	24	90
	Foreseeable	2	1	TBD	2	4	6	8	30
	Long Term	1	1	TBD	1	2	3	4	15
	Unknown	X	TBD	N/A	TBD	TBD	TBD	TBD	

Fig. 20.8 Response assessment: IQ time-critical evaluation of risk assessment

	Description	RESPONSE ASSESSMENT						Classification Distribution = Equal Intervals (Max / No. Categories)	Total
		Unknown	Monitor	Analyze	Prepare	Respond	Critical		
		Weight	X	1	2	3	4	5	
IQ ASSESSMENT	Scale	TBD	1	1	1	1	1	1	
	Best	1	1	TBD	1	2	3	4	5
	Good	2	1	TBD	2	4	6	8	10
	Average	3	1	TBD	3	6	9	12	15
	Below Average	4	1	TBD	4	8	12	16	20
	Bad	5	1	TBD	5	10	15	20	25
	Unknown	X	TBD	N/A	TBD	TBD	TBD	TBD	

Fig. 20.9 Time-critical evaluation matrix based on IQ assessment

action, within the right, i.e., impact context. That is, timeliness is more important to scenarios with higher risk; this ties to models such as Boyd's OODA loop.

Again, the high values are the focus of action, making it easier to identify critical items. However, as is the case for all of these resulting matrix classifications, the number of tiers and tier criteria, i.e., ranges, needs to be adjusted to fit the particular scenario and objectives.

Finally, Fig. 20.9 combines the result from Fig. 20.6 with Fig. 20.8. The result is a matrix that allows an action based on information quality for input data that correlates to the impact assessment.

The classification is based on equal intervals and was designed to consider that the impact of information quality increases as it becomes more time-critical, but its overall impact is directly correlated to quality; e.g., bad data may be easier to mitigate at the outset, the monitoring phase, or upstream fusion processes; but it becomes completely inadequate when it must be acted on immediately. That is, a critical scenario requires the best quality information.

The preceding models have the deficiency of only being able to scale linearly, reflecting a pure information quality cascading effect between tiers of information quality and higher phases of the HLIF process, and thus would not take into consideration potential feedback loop mitigation.

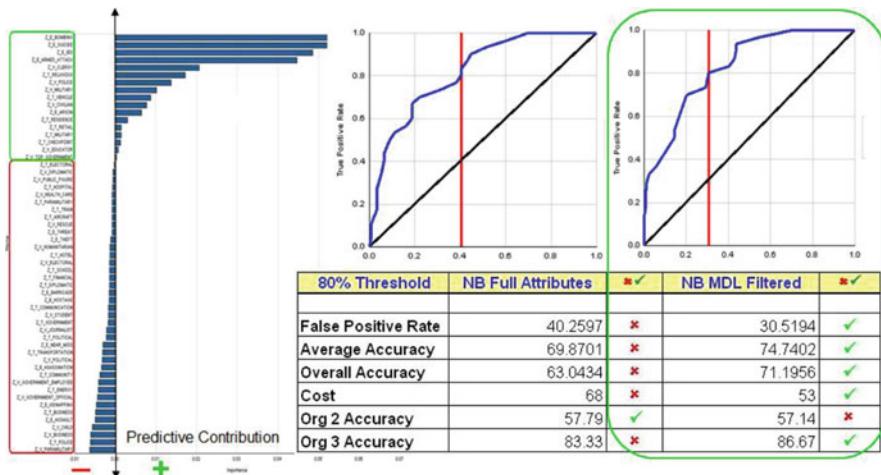


Fig. 20.10 Minimum description length as an IQ selector

For example, it may be desirable to quantify information quality in terms of what HILF life cycle or phase it is encountered, e.g., level 0 to level 6. It may be useful to conceptualize the impact of information quality with respect to time-critical decision-making in the classic important/urgent (*i/u*), important/not urgent (*i/ū*), not important/urgent (*ī/u*), and not important/not urgent (*ī/ū*) quadrants. This approach enables both decision-making and assessment of the underlying information quality for time constraint actions, i.e., information whose use will expire with respect to taking a corresponding course of action.

Figure 20.10 captures this strategy, whereby the aforementioned I/U quadrants are implemented as dual parameters for each entry, and each cell can be individually tuned for each phase, i.e., monitoring, analyze, prepare, respond, critical, as well as for the individual information quality weights.

Notice, for instance, that in the previous assessment, the monitor phase was qualified as low impact irrespective of the information quality, i.e., Weight = “1” and Scale = “1”. This negates the strong possibility of the negative cascading effects that information quality may have in the subsequent phases of the data exploitation value chain, meaning bad data, early one will have an accumulating negative effect in downstream processes.

The interval arranges have been selected to yield an symmetric distribution (5,6,3,6,5) IQ assessments (same assessment classifications as those used in the previous figure) around the neutral categories, i.e., “average information quality” and the “prepare” response phase.

Using this approach, it becomes evident that bad data has a negative cumulative effect by the time it reaches the “prepare” stage. This is congruent with the fact that downstream information quality mitigation cannot always compensate for upstream information quality issues.

20.5.5 Managing IQ with Dimensionality Reduction

When it comes to analysis of information quality, it is useful to split the fusion domain space in the traditional way, as described in the JDL model and expounded by Hall, LLinas, and Waltz [16, 21, 40–42], i.e., fusion levels 0 to 6, and made up of forensic (data mining) and predictive models.

Arguably, an underdeveloped topic in information quality is that of relevance for a particular process or algorithm. One characteristic of relevance is that of how much does a particular data element or data set contributes, if not to the overall result of actionable intelligence, to the individual downstream (downstream from data ingest) algorithms.

20.5.5.1 Minimum Description Length

Process refinement is a JDL model fusion level (level 4), whose main objective is the progressive optimization of the fusion process. Optimization can be stated in terms of improvements in efficiency, e.g., improving processing time, minimizing false positives, and minimizing cost of collecting information, and effectiveness, e.g., collecting only information that contributes to information exploitation and improving accuracy. Process refinement is neither sequential nor between any specific fusion levels; rather, it should be planned as part of the workflow and executed throughout the end-to-end exploitation chain, by implementing feedback loops among processes.

High dimensionality [11, 39] refers to an abstract solution space of a dependent variable as a point in n-dimensional space, whose dimension is determined by the number of independent feature variables. Having an extensive input space may be only “more information” and not necessarily have an effect in the predictive quality of a dependent target attribute – this extraneous information represent noise, which adversely impacts the processing efficiency and effectiveness of algorithms and, consequently, quality.

Furthermore, attributes that do not contribute to the information value exploitation chain represent a cost burden to the collection and preparation process and unnecessarily tie-up resources, e.g., personnel and equipment. This can be exacerbated when it is important to maintain security classification, as the cost of managing information across security classification domains can be significant.

Minimum description length (MDL) [14, 15, 25] can be used to reduce the number of entity attributes by ranking them according to their predictive contribution (a positive value represents predictive contribution, a negative value represents noise). Figure 20.10 Depicts results of running an MDL algorithm; importance is ranked from highest to lowest, with positive values denoting contribution to the predictive process and negative values representing noise. A similar effort can be done with numerical multivariate regression, where the p -value is used as the probability of obtaining a test statistic.

The value of the MDL process is twofold: (a) it is used to reduce the processing cost by requiring less independent variables, which minimizes collection cost and processing time; and (b) it may yield insight into patterns of what type of independent variables are more important to predicting the target value.

Applying MDL algorithms to achieve a reduction in attributes must be evaluated against the resulting predictive profile. While generally reducing attributes to be collected may significantly reduce the collection and processing cost, the predictive capability of the newer model (with less attributed) may impact predictive accuracy and false positives. Additionally all predictive models may change over time given that the dynamics of the variables are in constant flux; an adversary's course of action may change to adapt to a decision-maker's response; so maintaining historical data on the input attributes and corresponding predictive metrics helps in tuning information quality.

In this particular case, as shown by the receiver operating characteristic (ROC) curves in Fig. 20.10, the right ROC shows that accuracy has been improved and false alarms have been reduced with only a subset of information, highlighted by the green rectangle; therefore the information quality element (IQE) for algorithm optimization has even improved at a reduced cost; i.e., less information needs to be collected.

20.5.6 *The IQ Impact-Mitigation Interdependency*

A previous section presented information quality assessment and mitigation as interdependent, instead of two separate sections. The reason for this is that there is no end-to-end holistic framework for how information quality impacts AI at the system level. Intuitively it follows that the aggregate of IQ concerns progressively degrades the IQ of the overall HLIF SoS objectives.

But that does not imply that IQ can be mitigated with a one-shot inoculation. Rather, IQ must be addressed parsimoniously, at the first process or algorithm that is impacted by the lack of information quality or the first IQ point of impact.

To facilitate efficiency, IQ meta-models and a catalog of mitigation techniques for IQ vectors, and elements, should be part of the IQ integration process. IQ concerns need to be mitigated at their first point of impact.

20.6 Integrating IQ Management in the HLIF SoS

Integrating information quality processes in a high-level information fusion system of systems is as complex as the information fusion process itself. Due to feedback loops, the information exploitation chain is not a waterfall process, and accordingly information quality mitigation strategies cannot always be implemented *a priori* or in a predictive sequence.

Decomposing measures of performance (MOPs) into key performance parameters (KPPs) [27] must be flowed down and allocated to system components and may prove challenging. Performance requirements at the component level are not necessarily invariant due to the dynamic nature of interfusion-level processes; i.e., timelines and processing may be contingent on the volume and size of information and other characteristics such as high dimensionality or sparse data. The key to managing SLAs is to keep a processing profile updated and develop heuristics, including predictive load processing to provision additional resources, e.g., CPUs, memory, and storage. This may be accomplished in an elastic environment such as a Cloud.

Managing SLAs requires a more granular approach to accommodate variability driving by scenarios, which may include processing prioritization tied to matching criteria on information ingest. For example, if the information attribute matches x and attribute b matches y, then 90% of the processing chain needs to be completed in z time units. Processing thresholds may be set up based on CPU utilization watermarks that may trigger provisioning of additional compute resources.

Therefore identifying metrics for high-risk concerns requires a dynamic framework that includes the ability to map IQ dimensions to IQ KPIs to granular components and on demand ingest and processing to ensure high confidence in supporting the decision-making cycles [23].

This strategic top-down and tactical bottoms-up integrated analysis and synthesis requires the proper governance to ensure a cohesive enterprise IQ implementation.

20.6.1 Information Governance: The Requisite Oversight for IQ

It follows that defining and integrating critical processes in lockstep with the HLIF SoS from the outset, i.e. requirements phase, are essential in achieving IQ objectives; and this should be the purview of information governance (IG) throughout the entire systems engineering life cycle. There is no information governance without executive-level buy-in and a corresponding dedicated responsible and empowered individual. Information governance is the authoritative enabler that plans, monitors, and enforces controls over all facets of information assets management. In addition to ensuring the mission-driven IQ processes and coordinating with the chief engineer and chief architect of the HLIF, the responsible individual (RI) must also stay abreast of all appertaining external regulations and serve as a liaison between the organization and all other stakeholders with which information exchanges are performed.

The Enterprise Engineering Review Board (EERB), which is ultimately responsible for all technical design, trades, implementation decisions, should comprise the IG RI representing the various interests, both mission and technical of the HLIF SoS; this IG RI is responsible for setting the foundation regarding all IQ matters,

including the adoption of architecture principles such as information as an asset, ensuring IQ requirements are properly flowed down.

Information governance ensures a cohesive approach, analyzing concerns top-down and synthesizing solutions bottoms-up, aligned requirements, and ensures nobody adopts an “it’s the other’s responsibility” attitude. IG RI proactively manages technical and administrative concerns with the corresponding internal and external boards and is a stakeholder in trades that need to balance the IQ drivers with the enterprise needs and requirements. IG is also concerned with information configuration management (ICM) as a key to version control and overall data integrity.

An enterprise HLIF implemented with a system-of-systems approach requires a dedicated team of individuals with the roles and responsibilities at a granular level that interact with the overall technical information fusion subject-matter experts (SMEs) and are ultimately accountable to the IG RI through the established chain of command and governance policies.

This wider IQ team comprises roles and responsibilities such as information owners, information stewards, information administrators, information managers, information architects, information technicians, and information experts (analysts) for specific domains. This embedded approach guarantees that IQ is proactively managed from the outset and throughout the information fusion life cycle, and becomes an integral part of fusion level 4: process refinement.

20.6.2 The HLIF SoS Life Cycle IQ Management

Quality assurance (process oriented), quality control (product oriented), and total quality management (principles and management oriented) are all powerful concepts that by and large should be adopted within the information fusion community. However, just as information fusion has successfully leveraged many disciplines and technologies and yet requires a dedicated framework to realize its objective of actionable intelligence, proven IQ management techniques should also be leveraged but must be tailored to align with HILF SoS IQ dimensions.

Table 20.20 describes the phases and examples for corresponding selected steps. The comprehensive set of processes and steps should include considerations for support activities, such as configuration management, regulations, and implementing principles such as information as an asset.

20.7 Next Steps

Information quality is lagging information fusion; the focus of information fusion systems is advancing low- and high-level fusion architectures. It is understandable that many facets of IQ may not pragmatically develop until after its HLIF counter-

Table 20.20 Guidance on establishing an integrated HLIF SoS IQ life cycle process

Process/Step	Description
P1: Quality Planning	Identifies stakeholders, establishes overall plan, organized from strategic to tactical, i.e., per fusion level or process. Qualifies goals aligned with mission MOEs and establishes high-level metrics MOPs and SLAs that may be decomposed and assigned to various fusion components
S: Mission Objectives	Aligns with mission objectives, sets MOEs; check, measure, adjust, and iterate per AI each cycle. Manage trends proactively
S: IQ Vectors, Elements	Leverage requirements and IQ dimensions to identify and characterize IQ vectors and elements
S: IQ Impact/Resolution/QA	Qualify and quantify the impact and their corresponding resolutions (at the IQ element level) and identify mitigation strategies, how to deploy these to the various IF processes
P2: IQ Execution/QC	Apply (deploy) and execute the IQ resolutions, set-up triggers for executing other information quality control algorithms
P3: IQ Monitor	Compile results and corresponding metrics, analyze, and include in feedback loops. Root cause analysis and adjustment and confirm validity against planning targets and iterate cycle
P4: Governance	This is considered a phase that binds all others together with guidance, decisions, and authority. Information quality governance should be the key to achieving continuous improvement in that it is the phases that enables it and provides oversight of all steps within all other phases

parts are mature, especially given that IQ solutions may not always be developed *a priori*. Notwithstanding, it is worthwhile to prepare and explore how upcoming technologies and advances may impact information quality in HLIF SoS.

The following sections provide some proposals into considerations that may be useful to explore in keeping up with upcoming technologies.

20.7.1 *Information Quality Maturity Model for HLIF SoS*

The concept of an IQ maturity model was introduced earlier in Sect. 20.5. There are several frameworks that tie IQ maturity with CMMI. The information quality management maturity model (IQM3) by Caballero et al. [6] views information as a key corporate asset, which by the way is aligned with our architecture principle: data as an asset. This model is used to develop a framework which integrates with a corresponding management process.

Ryu et al. [33] focus on information value and information services and also develop an information quality management maturity model based on CMMI.

The next step is to develop an integrated and detailed CMMI-based model that expounds the basic concept in Fig. 20.11 and add the framework complexity presented in this chapter.

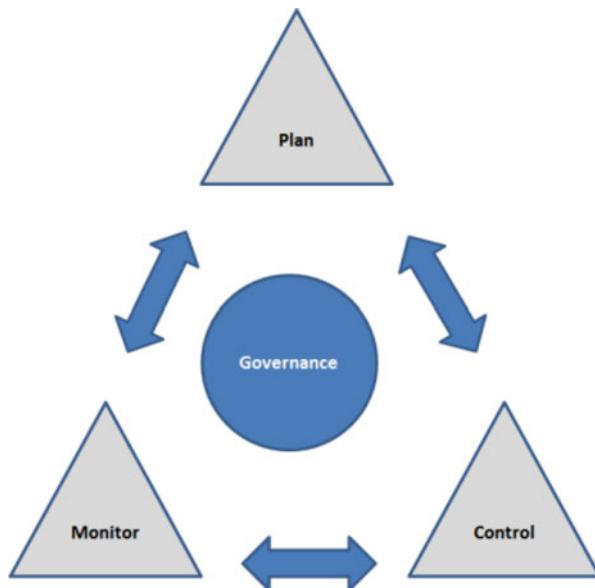


Fig. 20.11 The integrated IQ HLIF process (primitive view)

20.7.2 *IQMM Ontology Development*

Developing an IQ HLIF SoS meta-model ontology is perhaps the most critical endeavor in progressing IQ and also one of its most challenging. This is by no small measure; efforts include consulting and leveraging various expertise and coordinating the requisite systems engineering and architecture resources to realize commonality and developing the intricate relationships of a complex IQ meta-model that comprises all fusion levels.

20.7.3 *Information Quality Resource Catalog*

As discussed in the previous section, leveraging a catalog to provide among other things a marketplace resource and maintain pedigree and information regarding optimized IQ mitigation strategies and resolution would prove invaluable.

It is recommended that the first step is to develop a mapping, starting with categorizing algorithms and qualifying the corresponding data input profiles, i.e., what data is best suited for using a particular algorithm.

20.7.4 The Cost of Information and Information Quality

While proof of concepts are key in providing momentum to a research topic, when it comes to operational systems, they are invariably constraint by resources, i.e., cost, schedule, available technologies. It is imperative to integrate the cost trade-off of IQ in a comprehensive end-to-end framework. ROC curves for cost-benefit analysis of using particular algorithms with respect to their predictive capabilities, i.e., false alarms vs. accuracy.

However the need for a more comprehensive cost-benefit assessment is just about a foregone conclusion, as a system that is not fiscally efficient will ultimately be deprecated.

20.7.5 Dynamic HLIF: Timeliness and Closing the OODA Loop

IQ considerations for real-time acquisition were discussed in previous sections as well as a model on how to include the dimension of urgency to IQ. However, a more comprehensive question is lacking, which is: “how to truly assess the dynamic state of IQ, especially with respect to an adversary who may be implementing counter measures; e.g., misinformation?” So the perspective that needs to be assessed from both an AI and a course of action is how to evaluate, especially including timelines the process within an OODA framework.

Or to propose, what IQ elements related to timeliness need to be addressed to maintain an advantage over an adversary. And, as such, this needs to include a more tight integration with actionable intelligence, decision-making frameworks, and evaluation of effectiveness of any course of action steps taken.

20.7.6 Information Quality in the Cloud

As systems or more importantly system of systems are deployed to the Cloud, more cases are worth examining. The complexity of mixed ownership at the different Cloud tiers, e.g., IaaS, PaaS, and SaaS, coupled by a third party infrastructure with its own technology deployment cycle may cause significant disruption to the main organizations life cycle.

It is proposed that the IQ considerations for HLIF SoS deployed in Clouds be given extra scrutiny and that robust frameworks be developed for such environments.

20.8 Conclusion

Information quality is as complex as high-level information fusion itself, and given its critical role in achieving actionable intelligence, this subdomain must continue to mature not as a secondary fragmented consideration, but rather as a built-in and in lockstep with the advances in HLIF SoS.

The objective of this writing is to instill awareness by covering a broad number of topics to impress upon the community the wide and deep range of repercussions that information quality has on the very objective of information fusion systems, i.e., generating actionable intelligence, enabling the decision-maker, and establishing a course of action.

Information quality is a strategic consideration, and it must follow high-level goals and be congruent and cost-effective with respect to its outcome. Notwithstanding, information quality is neither about an enterprise silver bullet nor a unified and holistic metric which describes the totality of the HLIF SoS IQ. Accordingly IQ is driven by strategic measures, e.g., MOE, MOPs, and KPPs, but must be implemented tactically.

Information quality must undertake the same system of systems engineering approach as HLIF, and to achieve this integration, a cohesive enterprise yet flexible framework must be developed to include an IQ meta-model, the definition of quality concerns based on quality dimensions, quality vectors, and quality elements, quality assessment and resolution, and the requisite IQ governance to manage an integrated IQ process.

In conclusion, it can now be asserted that IQ is fundamental in achieving actionable intelligence. It is engrained and permeates all faces and levels of high-level information fusion, and as such it becomes imperative to proactively mitigate and address IQ issues from the outset.

References

1. J.F. Allen, Maintaining knowledge about temporal intervals, in *Readings in Qualitative Reasoning About Physical Systems* (Elsevier, 1990), pp. 361–372
2. C. Bisdikian et al., Building principles for a quality of information specification for sensor information, in *Information FUSION, 2009. FUSION'09. 12th International Conference on*, IEEE, 2009
3. E. Blasch, S. Plano, DFIG Level 5 (User Refinement) issues supporting situational assessment reasoning. in *Information Fusion, 2005 7th International Conference on*, 2005, p. 1
4. E. Blasch, P. Valin, E. Bosse, Measures of effectiveness for high-level fusion, in *Information FUSION (FUSION), 2010 13th Conference on*, IEEE, 2010
5. E.P. Blasch et al., User information fusion decision making analysis with the C-OODA model, in *Information FUSION (FUSION), 2011 Proceedings of the 14th International Conference on*, IEEE, 2011
6. I. Caballero et al., IQM3: Information quality management maturity model. *J. UCS* **14**(22), 3658–3685 (2008)

7. P.C. Costa et al., Towards unbiased evaluation of uncertainty reasoning: The URREF ontology, in *Information FUSION (FUSION), 2012 15th International Conference on*, IEEE, 2012
8. J. de Villiers et al., Evaluation metrics for the practical application of URREF ontology: An illustration on data criteria, in *Information Fusion (Fusion), 2017 20th International Conference on*, IEEE, 2017
9. DoD, MIL-STD 2525B: Common War-Fighting Symbology, 2005
10. A. DoD, *DoD Architecture Framework Version 2.0 (DoDAF V2. 0)* (Department of Defense, Washington, DC, 2009)
11. D.L. Donoho, High-dimensional data analysis: the curses and blessings of dimensionality, Manuscript, 2000
12. J.W. Enderle et al., *Joint architecture standard (JAS) reliable data delivery protocol (RDDP) specification* (Sandia National Laboratories, 2011) seas.upenn.edu
13. Group, T.O., TOGAF Version 9.1. 2011: Van Haren Publishing
14. P. Grunwald, Introducing the minimum description length principle, 2005 <https://arxiv.org/abs/math/0406077>
15. P. Grunwald, A tutorial introduction to the minimum description length principle. Arxiv preprint math.ST/0406077, 2004
16. D.L. Hall, E. Waltz, *Multisensor Data Fusion, Handbook Multisensor Data Fusion: Theory and Practice*, 2nd edn. Chapter 21. (CRC, 2008)
17. R. Johnston, *Analytic culture in the US intelligence community: an ethnographic study* (Csentral Intelligence Agency. Center For Study Of Intelligence Washington, DC, 2005)
18. B. Khaleghi et al., Multisensor data fusion: A review of the state-of-the-art. *Information Fusion* **14**(1), 28–44 (2013)
19. M. Klein, Combining and relating ontologies: an analysis of problems and solutions, in *IJCAI-2001 Workshop on Ontologies and Information Sharing*, USA, 2001
20. D.S. Lee, M. Heller, M. Napier, *Joint Architecture Standard-Quick Start Guide*, (Sandia National Laboratories (SNL-NM), Albuquerque, 2015)
21. J. Llinas et al., Revisiting the JDL data fusion model II. *Seventh International Conference on Information Fusion*, 2004, pp. 1218–1230
22. J. Llinas, D.L. Hall, An introduction to multisensor data fusion. *Proc. IEEE* **85**(1), 6–23 (1997)
23. V. Masayna, Koronios A., Gao J., A framework for the development of the business case for the introduction of data quality program linked to corporate KPIs & governance, in *Cooperation and Promotion of Information Resources in Science and Technology, 2009. COINFO'09, 4th International Conference on*, IEEE, 2009
24. N.F. Noy, Semantic integration: A survey of ontology-based approaches. *ACM SIGMOD Rec.* **33**(4), 65–70 (2004)
25. M. Pickett, T. Oates, *The Cruncher: Automatic Concept Formation Using Minimum Description Length. Abstraction, Reformulation And Approximation: 6th International Symposium, SARA 2005*, Airth Castle, Scotland, UK, 26–29 July 2005: PRO, 2005
26. G. Roedler, C. Jones, Technical measurement guide, in *International Council on Systems Engineering and Practical Software and Systems Measurement*, 2005
27. G.J. Roedler, C. Jones, Technical measurement. *A collaborative project of PSM, INCOSE, and industry* (Army Armament Research Development and Engineering Center, Picatinny Arsenal, 2005)
28. G. Rogova, E. Bosse, *Information quality effects on information fusion* (DRDC Valcartier, Tech Rept, 2008)
29. G. Rogova, et al. Context-based information quality for sequential decision making, in *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2013 IEEE international multi-disciplinary conference on*, 2013, IEEE
30. G.L. Rogova, E. Bosse, Information quality in information fusion, in *Information FUSION (FUSION), 2010 13th Conference on*, 2010 IEEE
31. G.L. Rogova, V. Nimier, Reliability in information fusion: literature survey, in *Proceedings of the 7th International Conference on Information Fusion*, 2004

32. S. Ross, *Application of System and Integration Readiness Levels to Department of Defense Research and Development* (Deputy Assistant Secretary of the Air Force for Science, Technology and Engineering Washington, DC, 2016)
33. K.S. Ryu, J.S. Park, J.H. Park, A data quality management maturity model. *ETRI J.* **28**(2), 191–204 (2006)
34. M.A. Solano, S. Ekwaro-osire, M.M. Tanik, High-level fusion for intelligence applications using recombinant cognition synthesis. *Inf. Fusion* **13**, 79–98 (2010)
35. M. Spitzer, E. Kappes, D. Böker, Enhanced intelligence through optimized TCPED concepts for airborne ISR, in *Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications IX* (International Society for Optics and Photonics, 2012)
36. C.P. Team, *CMMI® for Development, Version 1.3, Improving processes for developing better products and services. No. CMU/SEI-2010-TR-033* (Software Engineering Institute, 2010)
37. D. Thibault, *Commented APP-6A-Military Symbols for Land Based Systems* (Defence R&D Canada, 2005)
38. A. Tyukov et al., A concept of web-based energy data quality assurance and control system, in *Proceedings of the 14th International Conference on Information Integration and Web-Based Applications & Services*, 2012, ACM
39. M. Verleysen, Learning High-Dimensional Data, in *Limitations and Future Trends in Neural Computation*, 2003
40. E. Waltz, D.L. Hall, Requirements derivation for data fusion systems, Chapter 15, in *Handbook of Multisensor Data Fusion*, 2001
41. E. Waltz, J. Llinas, *Multisensor Data Fusion* (Artech House Boston, 1990)
42. E.L. Waltz, Information understanding: integrating data fusion and data mining processes. in *Circuits and Systems, 1998. ISCAS'98. Proceedings of the 1998 IEEE International Symposium on*, 1998, pp. 6

Chapter 21

Ranking Algorithms: Application for Patent Citation Network



Hayley Beltz, Timothy Rutledge, Raoul R. Wadhwa, Péter Bruck,
Jan Tobochnik, Anikó Fülöp, György Fenyvesi, and Péter Érdi

Abstract How do technologies evolve in time? One way of answering this is by studying the US patent citation network. We begin this exploration by looking at macroscopic temporal behavior of classes of patents. Next, we quantify the influence of a patent by examining two major methods of ranking of nodes in networks: the celebrated “PageRank” and one of its extensions, reinforcement learning. A short history and a detailed explanation of the algorithms are given. We also discuss the influence of the damping factor when using PageRank on the patent network specifically in the context of rank reversal. These algorithms can be used to give more insight into the dynamics of the patent citation network. Finally, we provide a case study which combines the use of clustering algorithms with ranking algorithms to show the emergence of the opioid crisis. There is a great deal of data contained within the patent citation network. Our work enhances the usefulness of this data, which represents one of the important information quality characteristics. We do this by focusing on the structure and dynamics of the patent network, which allows us to determine the importance of individual patents without using any information about the patent except the citations to and from the patent.

H. Beltz · T. Rutledge · R. R. Wadhwa · P. Érdi (✉)

Center for Complex Systems, Kalamazoo College, Kalamazoo, MI, USA

e-mail: hbeltz@umich.edu; Hayley.Beltz14@kzoo.edu; Timothy.Rutledge15@kzoo.edu;
Raoul.Wadhwa.2017@alumni.kzoo.edu; perdi@kzoo.edu

P. Bruck

ProcessExpert Ltd, Budapest, Hungary

J. Tobochnik

Department of Physics, Kalamazoo College, Kalamazoo, MI, USA

e-mail: jant@kzoo.edu

A. Fülöp

Wigner Research Centre for Physics, Hungarian Academy of Sciences, Budapest, Hungary

e-mail: fulop.aniko@wigner.mta.hu

G. Fenyvesi

Poliphon Ltd., Budapest, Hungary

Keywords Ranking · Algorithms · PageRank · Reinforcement learning · Patents · Clustering

21.1 Introduction

Ranking of nodes in a network with a diverse number of connections (degree) is an extensively studied field. In the theory of social networks, centrality measures were constructed to rank nodes of networks based on their (not unique) topological importance. Another family of measures is related to the spectral properties of the adjacency matrix [18], which takes into account the importance of the influence of a neighbor. Importance can be defined recursively. Brin and Page [4] introduced a matching recursive centrality measure called PageRank. The relevance of this algorithm to citation networks was discussed in [15]. By adopting a citation-based recursive ranking method for patents, the evolution of new fields of technologies can be traced.

Our driving question revolves around learning about the behavior of a network from two different levels. We examined the behavior of *classes* of patents using the IPC classification system and also the “attractiveness” of the individual patents. That is, we explored more quantitative methods of determining how influential or important particular patents are. One key characteristic of information quality is quantifying the importance of specific pieces of data. Thus, this work allows one to focus on those patents that are most important, thus enhancing the quality of the information we obtain from the patent network. This is where ranking methods come into play. In this chapter, we discuss some versions of the PageRank algorithm and the reinforcement learning algorithm [11]. Their applicability to ranking patents for the USPTO citation network is demonstrated.

21.2 Patent Citation Network Analysis

New information builds off of previous knowledge, and patents are no exception. When a patent is filed, it will cite other related patents. The applicant for the patent and the corresponding patent clerk are both responsible for selecting the set of cited patents. The patent citation network can be modeled as a directed acyclic graph where patents act as nodes and citations are edges between nodes. Analysis of the resulting graph permits insights into the advances and the current state of innovation and technology. We have previously studied the growth of the patent citation network at both the “microscopic” level of individual patents [8, 9, 12, 22, 23] and the “mesoscopic” [13] levels. Microscopic level studies measured the “attractiveness” of a patent as a function of its age and the number of citations already obtained. At the mesoscopic level, the analysis has been extended to subclasses, and it was demonstrated that it is possible to detect and predict

an emerging new technological trend through the application of an appropriate clustering algorithm. By adopting a citation-based recursive ranking method for patents, the evolution of new fields of technology can be traced [5]. Specifically, Bruck et al. [13] demonstrated that laser/inkjet printer technology emerged from the combination and development of two previously existing technologies: sequential printing and static image production. The dynamics of the citations coming from the different “precursor” classes illuminates the mechanism of the emergence of new fields and allows for the possibility to make predictions about future technological development.

To unlock the mysteries of this network, we apply ranking algorithms as well as examine the behavior of specific classes of patents. First, we include a brief discussion of the IPC classification system. We then briefly mention previous work done on the patent network as a whole. Finally, we consider the large-scale behavior of particular classes in the patent network to show it is similar to a social network.

21.2.1 USPTO Database

In 2015, Google, through collaboration with the United States Patent and Trademark Office (USPTO), made all patents published from 1976–2015 available for bulk download. Each patent was characterized by the following information: the patent number, publication date, an International Patent Classification (IPC) number, cited patent numbers, and one or more other classifications. This data is now hosted by Reed Tech (<https://patents.reedtech.com/>). This data allowed us to construct a network of every US patent published between 1976 and 2015 and perform analysis. Patents published prior to 1976 were not available in digital text format and thus were not included, and we make no claims on the behavior of the network before that time. Before examining this analysis, a quick word is needed about the IPC system.

The World Intellectual Property Organization (WIPO) uses the IPC system to organize patents [25]. The IPC, as the name suggests, contains patents from outside the USA. Although we only examine US patents, we argue that the IPC is the obvious choice for our analysis. The USPTO has several other classification systems, but these were not appropriate for use as there is no internal USPTO classification system such that each patent from 1976 to 2015 had a valid classification. Using the IPC system allowed us to utilize every patent in our database. The IPC and the classes it contains are updated every year, allowing it to act as a template for studying temporal mechanisms. For these reasons, IPC is the clear choice for our studies.

The IPC is a hierarchical structure with eight distinct main branches, denoted by the letters A through H. Each main class is then filled with a series of subclasses, as the subject matter becomes increasingly specific. For example, a patent in the class A61K 9/20 can be understood as:

Human Necessities (A)

Medical or Veterinary Science; Hygiene (61)

Preparations for medical, dental or toilet purposes (K)

Medicinal preparations characterized by special physical form (9)

Pills, lozenges or tablets (20)

Each patent is classified to the lowest “branch” or subclass available, i.e., there are no patents whose entire classification is only A61K.

A final note is that the IPC updates its classification system every year and it is possible for a patent to have its classification changed. These updates and changes to the network are a route we use to explore the trends and behavior of the US patent network.

21.2.1.1 Information Quality of the Data

How do the patent citation data we use satisfy the criteria of high quality of information? They are certainly *accurate* and *credible*, and we don't believe there are significant missing data from the databases. We have some difficulties with *timeliness*, since the approval of a submitted patent needs time. The methods adapted here are based on citations between patents, and any content analysis by using text-mining is neglected. The fact that we are able to learn so much just from citation data shows how this data is both *relevant* and *useful*.

21.2.2 Results

21.2.2.1 Temporal Behavior of Patent Classes

Historically, there have been discussions of the mechanisms of the temporal evolution of social groups [17]. These mechanisms have been classified as growth, decay, birth, death, merge, and split. The first four mechanisms implement changes within one community (cluster). Of the remaining two, the merge mechanism consists of a combination of two communities, while the split mechanism consists of one community breaking up into two smaller communities. The goal of this section is to demonstrate the existence of such mechanisms in the patent universe.

21.2.2.2 Methods

For the patent network, we have chosen man-made classes as our communities. Furthermore, we will be examining subclasses of varying levels in B (performing operations; transporting), D (textiles; paper), and G (physics) to give examples of these six mechanisms of evolving communities. The first four mechanisms – growth,

decay, split, and birth – were originally presented in Beltz et al. [6]. We present these as well as newly found instances of merge and death. To determine the behavior of a particular class or subclass, the first step was to create a list of every patent which resided in that class and the year that patent was published. Next, the number of patents added each year was recorded and examined.

21.2.2.3 Growth Mechanism

Growth occurs when the number of members in a cluster increase over a specified time period. Since patents are only removed from a class in rare cases, almost all patent classes are technically growing by this definition. This would make for an unexciting analysis, so we have decided to make a more meaningful definition of growth in regards to patent classes. We have defined growth in the patent network to be a class in which the *change* in the number of patents being added to a subclass is positive – akin to a positive acceleration. We have identified this mechanism in the relatively large subclass G06F 17/30. This subclass, sitting inside the larger physics class (G), deals with “information retrieval; database structures therefor” in digital computing. An example of a patent in this class would be one involving data mining such as patent 2554323 which is titled “Estimating Computational Resources for Running Data-Mining Services.” The earliest patent in this subclass is from patent number 5414626 from 1995 called “Apparatus and method for capturing, storing, retrieving, and displaying the identification and location of motor vehicle emission control systems.” The growth of this class is shown in Fig. 21.1.

21.2.2.4 Decay Mechanism

Decay of a cluster typically occurs when the number of members of that cluster decreases over time. Since patents are not usually removed from the patent network, we will not see a decay in the typical sense. Thus, to discover decay, we again examined the change in the number of patents added each year. A patent class that grows at a slower rate in consecutive time intervals would fit our description of decay. This definition of decay is simply the opposite of our definition of growth. We have found an instance of decay in class D03D, which includes patents dealing with “woven fabrics; methods of weaving; [or] looms,” part of the larger class of TEXTILES (D). The decay of this class is illustrated in Fig. 21.2.

21.2.2.5 Split Mechanism

The split mechanism occurs when a single cluster splits into at least two smaller ones. The split mechanism is similar to birth, see the section below. In the patent universe, we have defined splitting to be when a new class is created with older patents already inside it. A subset of patent classes fitting this characterization can

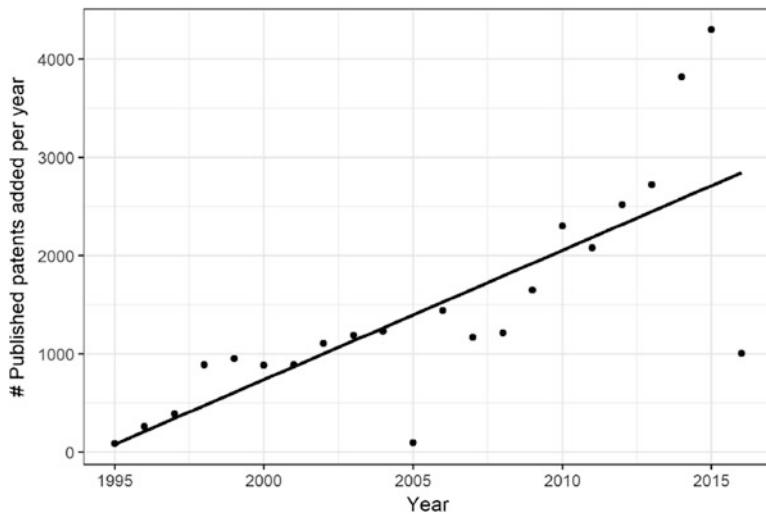


Fig. 21.1 The increasing growth of the G06F 17/30 class is exemplified by the best-fit slope of 131.1 annually published patents, on average. The r^2 value of 0.66 supports the hypothesis of rapid growth within G06F 17/30. We use a linear trend line and r^2 value not to argue that the behavior is linear, but that the general behavior of this class is growth. That is, the slope of the graph (the acceleration) is clearly positive, not negative

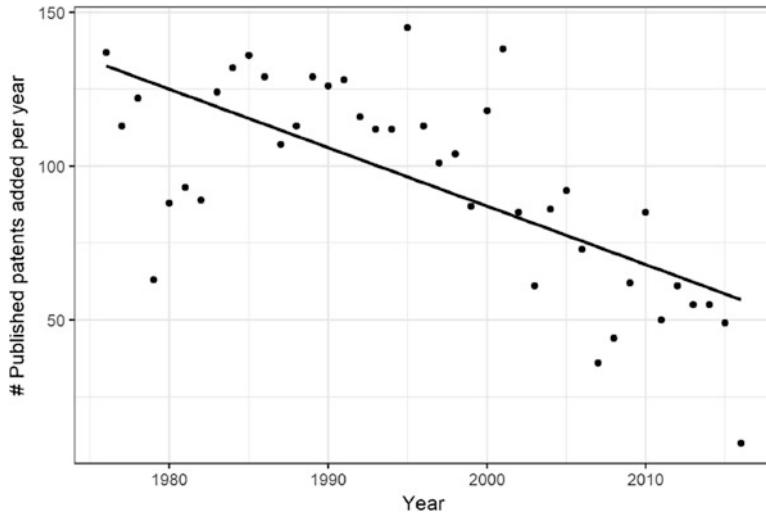


Fig. 21.2 The decay of the D03D class is illustrated by its temporal activity above, and the best-fit slope of -2.8 annually published patents, on average. The r^2 value of 0.72 supports the hypothesis of decay within D03D. Analogous to Fig. 21.1, no argument is being made for linear behavior. Rather, we use the linear best fit line and r^2 value to show the trend of decay or negative acceleration

be easily found by examining the year in which the patents within a given class were approved and comparing these to the date of class creation. If any patent within a class is older than the class itself, then this is an example of a split in the patent universe.

To find this mechanism in the patent universe, we looked at the classes containing microtechnology and nanotechnology. Before nanotechnology was known as its own distinct field, it was a part of microtechnology. Eventually, the class was split when a qualitative distinction between microtechnology and nanotechnology was recognized. Patents dealing with microtechnology remained unchanged, while those dealing with nanotechnology were reclassified into a distinct group. In the IPC, patents relating to microtechnology fall under the class B81, and those relating to nanotechnology fall under the class B82. To ensure that the split here is valid, we must find a single patent older than the class itself. The critical publication is Patent No. 6322713, published in 2001. This patent, entitled “Nanoscale conductive connectors and method for making same” was published about 5 years before the nanotechnology class was created. Thus, it is clear that a split into two, smaller communities occurred in this corner of the patent universe.

21.2.2.6 Birth Mechanism

The birth mechanism is rather self-explanatory: it occurs when a distinct new cluster comes into existence. Birth occurs frequently in the patent universe. Thousands of patents are added each year, and there must be classes to adequately describe them; if the established classes are not sufficient, new ones must be created. When looking for a birth mechanism, it was critical to ensure that we were examining an example of birth and not of a split. The important distinction between the two is that if a class contained patents which were older than the class itself, it should be classified as an instance of the split mechanism. If all patents in a class were published when or after the class itself was established, then the class must have been created for a patent that did not fit in any of the existing classes. This would be the canonical case of the birth mechanism. We were able to identify a class that fit our criteria for a birth mechanism, G01S 19, a class inside the physics branch. This class contains patents that deal with “satellite radio beacon positioning systems that determine position, velocity, or attitude using signals transmitted by such systems.” This class was created in 2010 and contains patents published from 2010 onward. The earliest patent we have data on, Patent No. 7701390, was published on April 20, 2010. From this information, we can conclude that this class was an example of the birth mechanism.

21.2.2.7 Merge and Death Mechanisms

A merge occurs when two distinct communities combine into one larger community. There are two types of events in the patent universe that fall under the merge

		B61C	LOCOMOTIVES; MOTOR RAILCARS (vehicles in general B60; frames or bogies B61F; special railroad equipment for locomotives B61J, B61K)
Note(s)			
This subclass covers:			
			<ul style="list-style-type: none"> • general design features or items of locomotives and motor railcars not otherwise provided for; • non-electric features of electric locomotives.
Locomotives or motor railcars in general or characterised by the type of motive power plant used			
	B61C 1/00	Steam locomotives or railcars (characterised by power transmissions B61C 9/00; engines F01; boilers F22B) [2006.01]	
	B61C 3/00	Electric locomotives or railcars (characterised by power transmission B61C 9/00; electrical features B60L, H02) [2006.01]	

Fig. 21.3 An example of the more subtle type of merging in the patent community. This screenshot of the definition of class B61C 1 shows the relationship between this class and three other classes, two of which are under a completely different main class

umbrella. The first is very literal and is closely linked to our example of death. It can be thought of as such: At time 0, there exist two classes, A and B. At some time in the future, every patent that was inside class B is now in class A. A is now significantly larger and is the merging of two classes. Also, class B no longer exists. We consider this a death in the patent universe.

This event is uncommon in the patent universe and is only able to be discovered by looking through changes in the classification names on the IPC website. Each year, members of the WIPO make edits to the IPC, sometimes performing this form of “death merge.” One example of this was with the classes F25C (“producing, working, or handling ice”) and A23G 9/00 (“frozen sweets, including ice cream, and their production”). These classes have since merged into F25C, and A23G 9/00 has experienced death.

The second example of merge is more subtle. Figure 21.3 shows the breakdown of the classes inside B61C dealing with locomotives and motor railcars. This class contains subclasses that distinguish between electric and steam locomotives. Inside the steam subclass, we note that this subclass relationship is intrinsically linked to power transmissions (B61C 9), engines (F01), and boilers (F22B). We consider this a merger since some of the patents inside these three subclasses have properties that have merged with B61C 1. That is, every patent in B61C 1 must have properties described in B61C 9, F01, or F22B. Note that the reverse is not true. For example, patents in F01 do not have to have characteristics of B61C1. It appears that this merger happened at late as 2006. Since the WIPO does not release older versions of its classifications, we cannot put an exact date on this merger, just an upper bound.

21.3 PageRank: A Brief Summary

PageRank is a recursive algorithm that takes into consideration the effect of the neighbors for the iterative computation of the probability of reaching a node through a random walk having “jump probability” d . This parameter is called the “damping factor,” because it reflects how significant is the effect of the neighbors to the node in question. PageRank’s significance comes from its simplicity; with being

dependent purely on network structure and only requiring one parameter, it allows for the possibility of many modifications. In Eq. (21.1) we have the PageRank value $\text{prob}(j)$ for a single node j . $O(i)$ is the outdegree of patent i , $B(j)$ is the list of patents that cite patent j , and d is the damping factor parameter, which has a value between 0 and 1. Equivalently, the equation for the vector of all the PageRank values, π , is given by Eq. (21.2). T is the transition matrix, and $\vec{\mathbf{1}}$ is a column vector of all ones. Note that this PageRank value is similar to the probability of a random walk to be at node j .

$$\text{prob}_{\text{new}}(j) = \left(d \times \sum_{i \in B(j)} \frac{\text{prob}_{\text{old}}(i)}{O(i)} \right) + \frac{1-d}{n} \quad (21.1)$$

$$\pi_{\text{new}} = d T \pi_{\text{old}} + \frac{1-d}{n} \vec{\mathbf{1}} \quad (21.2)$$

21.3.1 History of PageRank

The PageRank algorithm was first defined by Sergey Brin and Lawrence Page. In the first publication, the sum of the PageRank for all the nodes equals the number of nodes. However, in their second paper, they made a small change to the normalization so that the summation equals unity. Google had the edge over other search engines with this powerful and simple algorithm, and it was the main factor in the immediate success of Google. Other ways to rank the web were simply inefficient and returned too many “junk results” [4, 16].

Now there are countless studies, extensions, and modifications to the original PageRank algorithm. PageRank’s simplicity of only using the information of the network makes it extremely flexible and allows for it to be applied to any type of problem that has a network. There are applications for PageRank in many different fields, ranging from neuroscience to sports [14]. PageRank in practice is a little bit different than just the simple Eq. (21.1). The most common one and normally implied adjustment is to handle the problem of “dangling nodes,” which are nodes with outdegrees equal to 0. This is thought to be a problem for many reasons [26], but perhaps the most simple is that PageRank is based off of the study of Markov Chains, and one of the requirements for convergence is a stochastic matrix. However, if we were to use our algorithm (21.1), we see that if we have a dangling node with no outgoing links, then the column corresponding to that page would only sum to $\frac{dN}{N} = d$ rather than 1. One common adjustment is to force these columns to sum to one, but this modification changes the network so that any dangling node cites all n nodes in the network, changing our (21.1), (21.2) and (21.3). D_n is the set of all dangling nodes in the original network. Note that a network with no dangling nodes is unchanged.

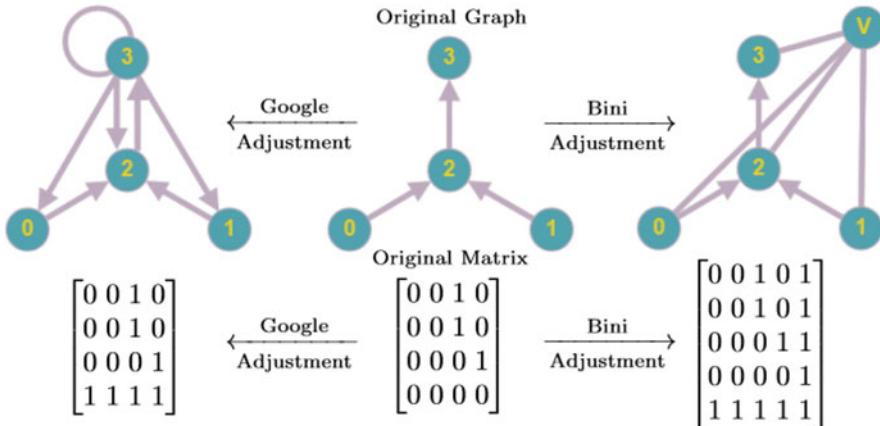


Fig. 21.4 In the center are the original graph and its transition matrix, and then to the left, the Google Adjustment is applied connecting the dangling node to every other node, while on the right, the Bini Adjustment is applied with each node being connected to and from the virtual node, V

$$\text{prob}_{\text{new}}(j) = d \left(\sum_{i \in B(j)} \frac{\text{prob}_{\text{old}}(i)}{O(i)} + \sum_{i \in D_n} \frac{\text{prob}_{\text{old}}(i)}{n} \right) + \frac{1-d}{n} \quad (21.3)$$

This is the most common adjustment for PageRank and is often called the Google Adjustment. It is not the only adjustment that can handle the issue of dangling nodes. Another proposal is called the Bini Adjustment which introduces a new, virtual, node that cites every node and is cited by every node. These adjustments are shown in Fig. 21.4 [10].

21.3.2 The Algorithm

There are two methods of solving for the PageRank of all nodes, the first being to explicitly solve the system of equations, and the other is the more practical computation power method. The first step for either method is applying the adjustment, and for this example, we will focus on the Google Adjustment, Eq. (21.3).

We can now explicitly calculate the PageRank for each node in the network by solving the system of equations generated from Eq. (21.3).

$$\text{prob}(0) = \frac{1}{2d^2 + 3d + 4} \xrightarrow{d=0.85} \approx 0.125$$

$$\text{prob}(1) = \frac{1}{2d^2 + 3d + 4} \xrightarrow{d=0.85} \approx 0.125$$

$$\text{prob}(2) = \frac{2d + 1}{2d^2 + 3d + 4} \xrightarrow{d=0.85} \approx 0.338$$

$$\text{prob}(3) = \frac{1 + d + 2d^2}{2d^2 + 3d + 4} \xrightarrow{d=0.85} \approx 0.412$$

Computational Method The more practical way to calculate PageRank for large graphs uses the Google Matrix, G , which is a reorganized form of Eq. (21.2). Here G is defined in terms of the transition matrix T and the random teleportation component $\frac{1-d}{n}$. J is a matrix of the same size as T , with every element $J_{ij} = 1$.

$$\pi_{\text{new}} = G\pi_{\text{old}} \quad (21.4)$$

$$G = d T + \frac{(1-d)}{n} J \quad (21.5)$$

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \xrightarrow[\text{Adjustment}]{\text{Google}} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \xrightarrow[\text{Matrix}]{\text{Google}} \begin{bmatrix} \frac{d-d^2}{n} & \frac{d-d^2}{n} & \frac{nd+d-d^2}{n} & \frac{d-d^2}{n} \\ \frac{d-d^2}{n} & \frac{d-d^2}{n} & \frac{nd+d-d^2}{n} & \frac{d-d^2}{n} \\ \frac{d-d^2}{n} & \frac{d-d^2}{n} & \frac{d-d^2}{n} & \frac{nd+d-d^2}{n} \\ \frac{nd+d-d^2}{n} & \frac{nd+d-d^2}{n} & \frac{nd+d-d^2}{n} & \frac{nd+d-d^2}{n} \end{bmatrix} \\ = G$$

This matrix G encapsulates the whole PageRank process into a single stochastic matrix. Let our starting state be the vector s_0 . With this matrix, we can carry out one step of the algorithm with one matrix multiplication, that is, $T s_0 = s_1$, or the state after one step. We repeated this process, going from $T s_n = s_{n+1}$ to $\|s_{n+1} - s_n\| = \epsilon$, where ϵ is the convergence factor. As Sergey Brin and Larry Page demonstrated in their paper, this convergence happens very quickly. For more information on PageRank, look at the original paper by Page/Brin [10, 16].

21.4 A Closer Look at the Damping Factor

PageRank is a simple yet powerful algorithm, but the correct interpretation of its only parameter, the damping factor, is still not fully understood. The choice of damping factor can often be underemphasized, generally with the argument to use $d = 0.85$, because that is what Google did. The issue can be more complicated requiring more in-depth consideration for different types of networks. In some carefully constructed networks, small perturbations in the damping factor such as from 0.850 to 0.851 can drastically change the rankings, [3] while analysis of the Stanford Web network can show a damping factor of 0.65 results in a more stable ranking than 0.85 [21].

21.4.1 Rank Reversal

The PageRank value of each and every node is a function of the damping factor. However, how d effects a node's PageRank is different for each node. Figure 21.5 illustrates that the change of the damping factor can lead to the reversal of their ranks; this is called rank reversal. A more concrete understanding of the damping factor and its role in PageRank could lead to a better understanding of the effects of components such as bottlenecks and dangling nodes or how to build better algorithms that are more resilient to manipulation. The more rank reversals a network has, the more possible rankings can be generated from the PageRank algorithm. A rather straightforward sounding way to handle rank reversals would be simply not to use any damping factor, and that is what is done in the method called TotalRank [2]. To rank network nodes in TotalRank, there is no need to explicitly specify the value of d , because the PageRank values corresponding to all possible values of d will be averaged through integration. TotalRank is arguably not a good practical replacement for PageRank, as it takes a significant increase in computation time that is not adequately justified by an objective improvement over PageRank. However, it did remove the ambiguity of PageRank as there is only one possible ranking for any given network.

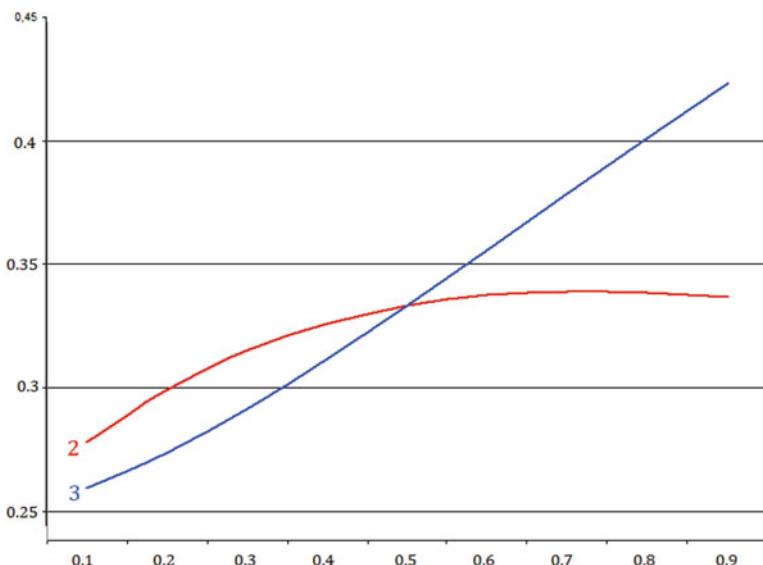


Fig. 21.5 Plot of the PageRank for nodes 2 and 3 in our sample graph. At $d = 0.5$, the PageRank Value of node 2 is the same as node 3

21.4.2 Causes of Rank Reversal

Figure 21.6 illustrates that in a large network (e.g., the USPTO network), some nodes are receiving high ranks at low values of d and high ranks at high values of d , which is primarily influenced by the ratio of the direct and the indirect citations of a given node. These rank reversals are caused by the increasing influence from distant nodes as d increases; certain identifiable network structures may also contribute such as leaf nodes and dangling nodes. Dangling nodes are nodes with no outward links; leaf nodes are nodes with no inward links. In our example graph, Fig. 21.4, we can see that we have two leaf nodes, nodes 0 and 1, and one dangling node, node 3. Leaf nodes play their biggest role at low values of d , since with no incoming links, their PageRank values only decrease as d increases. Dangling nodes have their biggest role at high values of d , because a random walker would tend to move to these dangling nodes in the limit d goes to unity. Dangling nodes aren't as large of an issue since the Google and Bini Adjustment remove dangling nodes from networks by adding more edges. There are also other larger structures such as bottlenecks and rank sinks which may contain more than one node. A bottleneck is a connection between two mostly disjoint components of a network; an example of this would be node 2 in the example graph as it separates nodes 0 and 1 from node 3. Rank sinks are a set of nodes that have very little outbound connections, therefore “trapping” the random walker. Like dangling nodes, rank sinks also have the greatest effect on a network in the limit as d goes to unity [21].

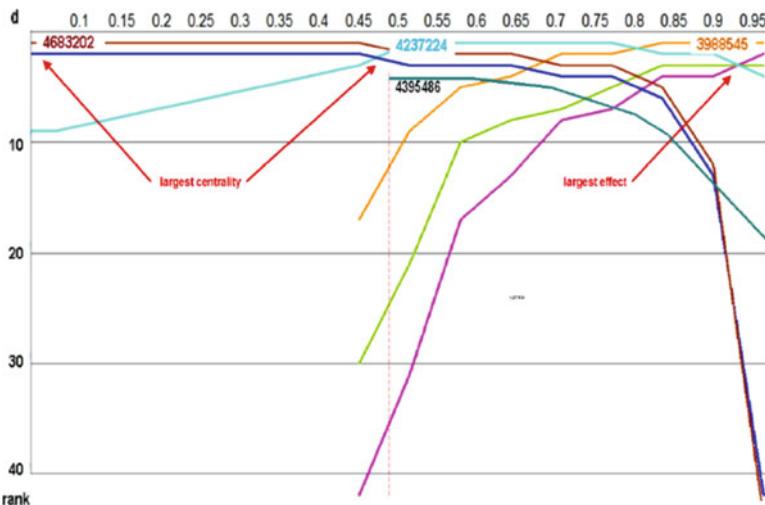


Fig. 21.6 Rank changes of the most important USPTO nodes as a function of d

21.5 Reinforcement Learning

Although PageRank is perhaps the most famous of ranking algorithms, it is not applicable for all situations. While the spirit of the algorithm can be extended to patent citation networks, it needs modifications for application in a directed acyclic graph (DAG). Because of the unidirectional structure of the patent network, applying PageRank will cause significantly more “sinks” of rank, therefore overinflating the importance of the older patents. A “sink” can be thought of as a patent that, due to the nature of the PageRank algorithm and the structure of the network, sees a large amount of PageRank score flowing into it at each iteration of the algorithm. Because of this, we chose to employ an altered version of a reinforcement learning (RL) ranking algorithm developed by Derhami et al. [11]. RL is a machine learning concept that, in its agent-based form, aims to teach the agent in question how to act in the relevant environment by assigning either a reward or punishment for all potential actions [24].

21.5.1 The Algorithm

The following equation shows the calculation of RL Rank:

$$R_{t+1}(i) = \sum_{j \in B(i)} \left(\frac{\text{prob}(j)}{O(j)} (r_{ji} + \gamma R_t(j)) \right) \quad (21.6)$$

where γ “is a discount factor that determines the present value of the future rewards that can be achieved over time,” $O(j)$ is the outdegree of patent j , and r_{ji} is the value of a reward granted by a patent j to the patent i that it cites [11]. The function $\text{prob}(j)$ comes from the equation for PageRank, given by Eq. (21.2). As a reminder this expression is the probability of a random walker on the network being at a patent j with damping factor d . Above, $B(i)$ is the list of patents that cite patent i . Thus, the entire RL ranking algorithm can be described in the following pseudocode:

```

 $\delta \rightarrow 0$ 
while ( $\delta > \epsilon$ )
For every page  $i \in V$ 
 $\text{prob}_{\text{new}}(i) = \left( d \times \sum_{j \in B(i)} \text{prob}(j)/O(j) + (1 - d)/n \right)$ 
End for
 $\delta \leftarrow ||\text{prob}_{\text{new}} - \text{prob}||$ 
 $\text{prob} \leftarrow \text{prob}_{\text{new}}$ 
End while
 $\delta \leftarrow 0$ 
```

```

while ( $\delta > \epsilon$ )
For every page  $p \in V$ 
 $r_{ji} = 1/O(j)$ 
 $R_{\text{new}}(i) = \sum_{j \in B(i)} \left( \frac{\text{prob}(j)}{O(j)} \times (r_{ji} + \gamma R(j)) \right)$ 
End for  $\delta \leftarrow ||R_{\text{new}} - R||$ 
 $R \leftarrow R_{\text{new}}$ 
End while

```

The RL ranking algorithm (above) exhibits rapid convergence, allowing for its use in the relatively large patent citation network [11]. It is important to note that the original algorithm preserves the score initially assigned to nodes with zero indegree. To meaningfully apply it to a citation network, this original algorithm had to be altered since nodes with zero indegree, i.e., those that have not been cited to date, should inherently possess a low score. We modify the algorithm provided in [11] by allowing score flow to occur in all nodes. This implicitly grants a score of zero to nodes with zero indegree once convergence has been reached ($\epsilon = R_t - R_{t-1} < 10^{-9}$ between consecutive iterations for each individual patent).

21.5.2 Comparison with PageRank

With two different methods of ranking, the question of which to use in a given situation arises.

Before answering that question, one might wonder if RL ranking is sufficiently different from PageRank. It turns out that the similarity between the two relies solely on the choice of γ and d . Figure 21.7 illustrates that the predictability of RL scores from PageRank scores at a given moment in time is independent of γ , the discount factor. A more important role is played by the damping factor, d , which nonlinearly decreases the predictive power of PageRank scores as they tend toward unity. We can also compare the scores under the same γ and damping factor, as shown in Fig. 21.8, which plots the PageRank score and the RL score inside a single subclass. We see some slight correlation between scores but nothing too significant. From this, we can see that RL rank is related to PageRank slightly but still differs significantly. In this figure, which plots the PageRank score and the RL score inside a single subclass, we see some slight correlation between scores but nothing too significant. From this, we can see that RL rank is related to PageRank slightly but still differs significantly.

Now the question of when to use PageRank and when to use RL rank can be addressed. One of the key differences between the two ranking methods is whether or not connection from nodes should be treated equally. In PageRank, the importance of a node citation is not considered. The PageRank from being cited by an influential node is the same as that from one that is unheard of. RL

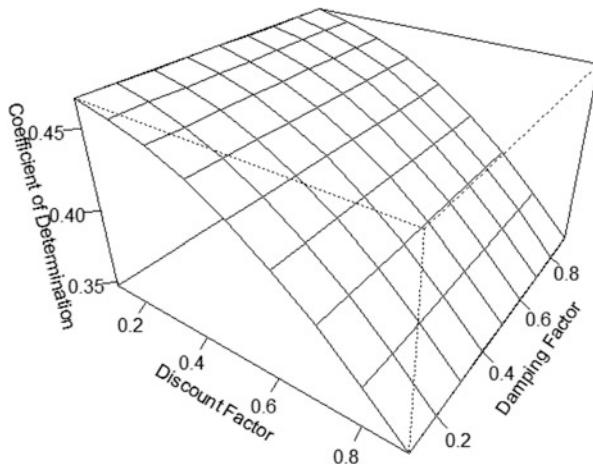


Fig. 21.7 A perspective plot of the predictability of a patent's RL rank score based on its PageRank score. Only patents within A61K or connected to patents in A61K are included. Predictability is operationalized by the *coefficient of determination* (R^2) on the z-axis. The patent citation network is based on its state at the end of 2015. The damping factor, d , and the discount factor, γ , are the independent variables. This figure is built using the procedure in [19]

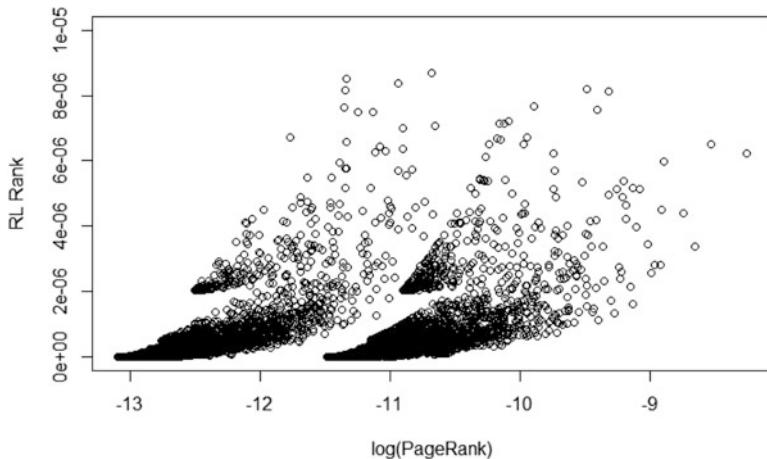


Fig. 21.8 A direct comparison between the PageRank score and the RL rank score of a particular subclass. We can see some slight trends in correlation but nothing incredibly strong

ranking addresses this and allows for the distribution of rank based on the number of outgoing citations. For example, a node that cites 100 nodes will deliver a smaller amount of rank to each child node than a parent node which cites only 5 children. From this, it seems that the optimal ranking system to use depends on the context of your network. Do your nodes differ vastly in importance, and do you want your ranking to address that? If so, RL rank would provide a ranking score that addresses this concern. If you want your rank to only be determined by the connectivity of your graph, then PageRank is the clear choice.

21.5.3 Case Study

We end by examining a use of RL ranking with patents. For this, we examined a subclass A61K9 and conducted a RL ranking of all of the patents inside it. We then examined the behavior and connectivity of these “most influential patents.” That is, we looked at the central patent and its first neighbors (patents that it cites or patents that cite the central patent) at eight different time frames. At each time step, we clustered all of these patents using a multilevel community detection algorithm in the software igraph [1, 7]. A more in-depth description of this process can be found in [6]. The most interesting case we examined was that of patent 3845770, one that describes an osmotic device which enables a variety of drug compounds to be released in a controlled and continuous way for a prolonged period of time. We show the behavior of this patent in Fig. 21.9. This figure, broken up into eight distinct time periods, shows the evolution of this patent. In the first time period, this section of the network contains just two clusters: a bigger (red) and a smaller (blue)

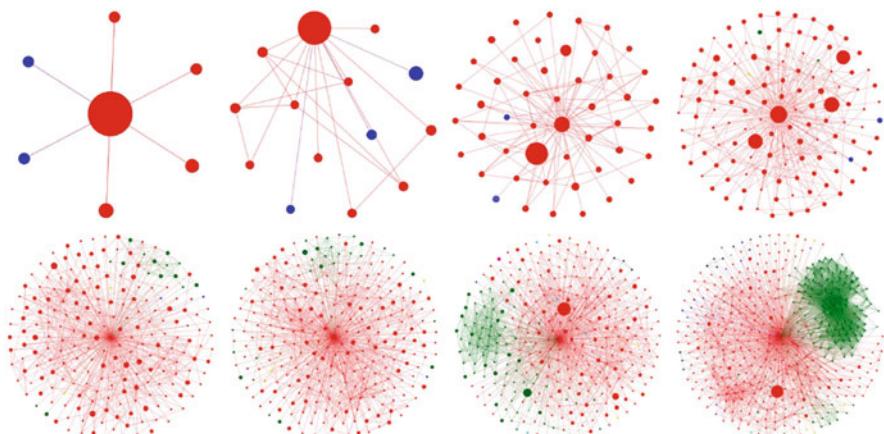


Fig. 21.9 Colors encode clusters. Node size is proportional to the normalized ranking score. 1976–80, 1981–1985, 1986–90, 1991–95, 1996–2000, 2001–05, 2006–10, 2011–15

cluster. The patent of interest belongs to the bigger red cluster. Every node from this group is related to a different kind of drug delivery system or drug dosage form (such as pill, tablet, capsule, or liquid form). The blue cluster also contains patents with similar ideas such as the drug delivery mechanism into the body. In the time period 1990–1995, a new cluster appears with green color. Every patent from this cluster is related to different kinds of dosage forms and delivery systems for opioids. Since 1990, opioid drugs have become very popular in the treatment of chronic pain and cancer [20]. We can see the emergence of the opioid epidemic by the creation of the green cluster and its rapid growth in size. We also see other cluster dynamics at play. The blue cluster seems to stagnate and then essentially disappear. In essence, by using reinforcement learning ranking to identify influential patents, one can see the emergence of historical trends in the patent universe.

21.6 Summary

The patent citation network can be viewed as a time-evolving complex system, and the relationship between the topological structure and the dynamics of the network has been analyzed. Patent data are organized into a hierarchy, where there are patent classes, subclasses, etc.

Inventions often can be described as combinations of already existing technologies, and one of our goals was to identify possible mechanisms of technological evolution reflected by changes in the patent universe. Growth, decay, split, birth, and merge mechanisms were detected.

Next, algorithms for ranking of nodes were discussed and used specifically for patent citation networks. Now it is well-known that the result of the ranking given by the PageRank algorithm depends on the numerical value of the damping factor, and rank reversal happens. The application of the PageRank algorithm to the USPTO database shows that rank stability occurs for smaller values of the damping factor, and massive rank reversal happens for higher values.

The reinforcement learning ranking algorithm proved to be useful for situations when nodes show a high diversity in their importance.

Finally we showed real-world applications by combining reinforcement learning ranking with cluster analysis. We were able to identify influential patents. Knowing which patents are influential is one characteristic of information quality, and thus we are able to obtain useful information from the structure and dynamics of the patent citation network.

Acknowledgements PE thanks the Henry Luce Foundation for support of Complex Systems Studies as Henry R Luce Professor. JT thanks the Herbert H. and Grace A. Dow Foundation for support as the Dow Distinguished Professor of the Natural Sciences.

References

1. V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**(10), P10008 (2008)
2. P. Boldi, Totalrank: ranking without damping, in *14th International World Wide Web Conference (WWW2005)* (ACM Press, New York, 2005), pp. 898–899
3. M. Bressan, E. Piserico, Choose the damping, choose the ranking? *J. Discre. Algorithms* **8**(2), 199–213 (2010)
4. S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* **30**(1–7), 107–117 (1998)
5. P. Bruck, I. Réthy, J. Szente, J. Tobochnik, P. Érdi, Recognition of emerging technology trends: class-selective study of citations in the U.S. patent citation network. *Scientometrics* **107**(3), 1465–1475 (2016)
6. Y. Choe (ed.), in *From Ranking and Clustering of Evolving Networks to Patent Citation Analysis*, Anchorage (IEEE Computational Intelligence Society, 2017)
7. G. Csardi, T. Nepusz, The igraph software package for complex network research. *InterJournal Comp. Syst.* **1695**, 1–9 (2006)
8. G. Csárdi, K.J. Strandburg, L. Zalánayi, J. Tobochnik, P. Érdi, Modeling innovation by a kinetic description of the patent citation system. *Phys. A Stat. Mech. Appl.* **374**(2), 783–793 (2007)
9. G. Csárdi, K.J. Strandburg, J. Tobochnik, P. Érdi, The inverse problem of evolving networks – with application to social nets, in *Handbook of Large-Scale Random Networks*, ed. by B. Bollobás, R. Kozma, D. Miklós (Springer, Berlin/Heidelberg, 2008), chap. 10, pp. 409–443
10. D.A. Bini, G.M. del Corso, F. Romani, Evaluating scientific products by means of citation-based models: A first analysis and validation. *Electron. Trans. Numer. Anal.* **33**, 1–16 (2008)
11. V. Derhami, E. Khodadadian, M. Ghasemzadeh, A.M.Z. Bidoki, Applying reinforcement learning for web pages ranking algorithms. *Appl. Soft Comput.* **13**(4), 1686–1692 (2013)
12. P. Érdi, *Complexity Explained*. Springer Complexity (Springer, New York, 2008)
13. P. Érdi, K. Makovi, Z. Somogyvári, K. Strandburg, J. Tobochnik, P. Volf, L. Zalánayi, Prediction of emerging technologies based on analysis of the US patent citation network. *Scientometrics* **95**(1), 225–242 (2013)
14. D.F. Gleich, PageRank beyond the Web. *SIAM Rev.* **57**(3), 321–363 (2015). <https://pubs.siam.org/doi/abs/10.1137/140976649>
15. S. Maslov, S. Redner, Promise and pitfalls of extending google's pagerank algorithm to citation networks. *J. Neurosci.* **29**, 1103–1105
16. L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: bringing order to the web. *Stanford InfoLab* (1999)
17. G. Palla, P. Pollner, A.-L. Barabási, T. Vicsek, Social group dynamics in networks, in *Adaptive Networks: Theory, Models and Applications*, ed. by T. Gross, H. Sayama (Springer, Berlin/Heidelberg, 2009), chap. 2, pp. 11–38
18. N. Perra, S. Fortunato, Spectral centrality measures in complex network. *Phys. Rev.* **78**, 036107 (2008)
19. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2017)
20. A. Rosenblum, L. Marsch, H. Joseph, R. Porteno, Opioids and the treatment of chronic pain: controversies, current status, and future directions. *Exp. Clin. Psychopharmacol.* **16**(5), 405–416 (2008)
21. S.-W. Son, C. Christensen., P. Grassberger, M. Paczuski, Pagerank and rank-reversal dependence on the damping factor. *Phys. Rev. E* **86**, 066104 (2012)

22. K.J. Strandburg, G. Csárdi, J. Tobochnik, P. Érdi, L. Zálynyi, Law and the science of networks: an overview and an application to the ‘patent explosion’. *Berkeley Technol. Law J.* **21**, 1293 (2007)
23. K.J. Strandburg, G. Csárdi, J. Tobochnik, P. Érdi, Patent citation networks revisited: signs of a twenty-first century change. *North Carolina Law Rev.* **87**(5), 1657–1698 (2009)
24. R.S. Sutton, A.G. Barto, *Reinforcement Learning: An Introduction* (The MIT Press, London, 1998)
25. WIPO, *International Patent Classification(IPC)* (2018)
26. I.C.F. Ipsen, T.M. Selee, PageRank computation, with special attention to dangling nodes. *SIAM J. Matrix Anal. Appl.* **29**(4), 1281–1296 (2008). <https://pubs.siam.org/doi/abs/10.1137/060664331>

Chapter 22

Conflict Measures and Importance Weighting for Information Fusion Applied to Industry 4.0



Uwe Mönks, Volker Lohweg, and Helene Dörksen

Abstract Information sources such as sensors, databases, and human experts serve as sources in order to realise condition monitoring and predictive maintenance in Industry 4.0 scenarios. Complex technical systems create a large amount of data which cannot be analysed manually. Thus, information fusion mechanisms gain increasing importance. Besides the management of large amounts of data, further challenges towards the fusion algorithms arise from epistemic uncertainties (incomplete knowledge) and—mostly overseen—conflicts in the input signals. This contribution describes the multilayered information fusion system MACRO (multilayer attribute-based conflict-reducing observation) employing the BalTLCS (balanced two-layer conflict solving) fusion algorithm to reduce the impact of conflicts on the fusion result by a quality measure which is denoted by *importance*. Furthermore, we show that the numerical stability in heavy conflicts is a key factor in real-world applications. Different examples end this contribution.

Keywords Information fusion · Industry 4.0 · Conflict measures · Importance weighting · Machine learning · MACRO system

22.1 Introduction

Information fusion is an essential methodology in state-of-the-art industrial equipment which makes use of the emerging field of cyber-physical systems (CPS). In the context of the upcoming Fourth Industrial Revolution (Industry 4.0), multiple

U. Mönks (✉)
coverno GmbH, Lemgo, Germany
e-mail: uwe.moens@coverno.de

V. Lohweg · H. Dörksen
inT – Institute Industrial IT, Ostwestfalen-Lippe University of Applied Sciences,
Lemgo, Germany
e-mail: volker.lohweg@hs-owl.de; helene.doerksen@hs-owl.de

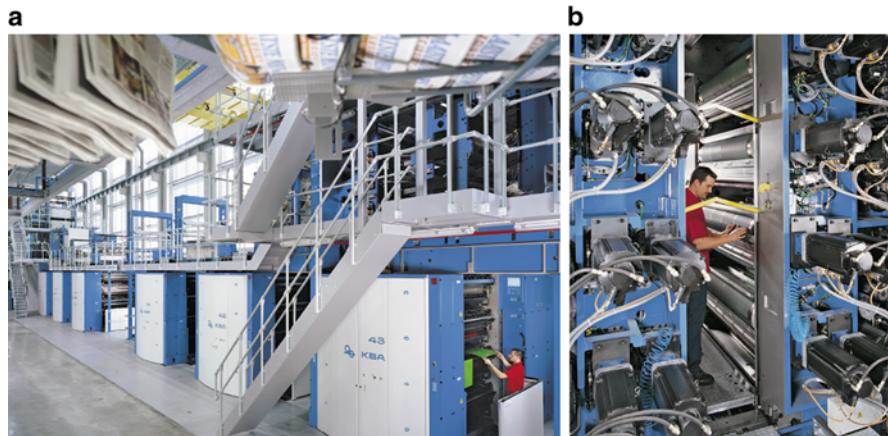


Fig. 22.1 (a) KBA Cortina newspaper offset printing press overview (b) KBA Cortina newspaper offset printing press detail showing its several attached electric drives. (With kind permission of KBA—Koenig & Bauer AG, Würzburg [21])

CPS are necessary for condition monitoring and predictive maintenance in machine and process industries. CPS are physical processing systems equipped with sensory devices which interconnect over communication networks for distributed cognitive information processing applications. Whereas the data amount is usually large because of quasi-unlimited sources (sensors, databases, experts, etc.), the computational resources are generally limited. One example is industrial printing processes, like the newspaper printing process depicted in Fig. 22.1.

Today's state-of-the-art printing systems are driven by hundreds of actuators in the application, along with a number of sensors in the same order of magnitude. On the one hand, these are electric drives moving cylinders, positioning units, or magnetic valves controlling the print colour's application onto the substrate, for example. On the other hand, a vast variety of sensory units are applied for acquisition of different types of data. These may be several different basic physical measures such as pressure or temperature, but also specific process parameters like the quality of inks, as well as many others.

Since the signal sources are distributed over the entire application, all the data must usually be communicated over an appropriate network to all CPS. Conservative approximations show that the bandwidth of standard Fast Ethernet is occupied by already 20 network participants [7, 36]. Nevertheless, these figures point out that centralised systems are not able to handle all occurring data due to restrictions of current fieldbus systems needed to communicate the data. The situation additionally deteriorates when instead of one single process a complete plant consisting of a number of production machines is to be monitored. Inconsistencies and conflict must occur naturally. Such systems don't scale in the end.

As illustrated above, different sensors are increasingly applied in industrial processes to measure and control complete processes, machines, and logistics. One

way to handle the resulting large amount of data created by thousands of different information sources is to employ information fusion systems. Information fusion systems, e.g. condition monitoring and predictive maintenance, combine different sources of information, like databases, sensors, or human experts to generate the current state of a complex system. The result of such information fusion processes is regarded as a *health indicator* of complex systems. Therefore, information fusion approaches are applied to, e.g. automatically inform about a reduction in production quality or detect possibly alarming situations in technical systems.

Besides the management of large amounts of data, further challenges towards the fusion algorithms arise from epistemic uncertainties (incomplete knowledge) in the input signals as well as conflicts between them. In many cases the information captured from the environment may be imprecise, incomplete (scarce), or inconsistent. Furthermore, signal sources may not be reliable [1].

Considering the importance of sensors in information fusion systems in industrial processes in general, defective sensors have several negative consequences. The machine condition is not detected correctly; control processes will not run adequately; it may lead to machine failure, e.g. when wear and tear of a machine is not detected sufficiently in advance—just to name a few critical effects.

A prominent factor which can generate ineffectiveness or even contradictory results is the “conflict between data sources”. That is given whenever the information of at least one source disagrees with the remaining available information. The possible causes of conflict can be numerous. Source deterioration or faults occur especially in real-world problems. Manipulation of the sources (or their information) is also conceivable, especially in security-critical settings. Conflict is formally a form of conscious ignorance. It is, namely, the cause of inconsistency or distorted information [1].

Such information inconsistencies lead to results, which do not represent the actual situation if the conflict has not been recognised and addressed during information processing.

Conflict has been identified as one of the most challenging topics in information fusion (IFU) [20]. Measures of conflict are known in literature. One prominent example is Shannon’s entropy measure [41] which is also applied as a conflict measure [1]. Therefore, it is necessary to extend known fusion concepts insofar that they are able to measure and to handle imprecision and reliability [33]. These aspects must be considered during information processing to obtain reliable results, which are in accordance with the real world. Only then, the obtained information can be regarded to be of high quality. The analysis of the scientific state of the art shows that current solutions fulfil the said requirements at most only partly.

Examples on such insufficiencies based on inconsistencies or conflicts might not only have influences on machine behaviour but have also lethal consequences. On May 9, 2015, Airbus suffered from a crash of one of its A400M military aircrafts during a test flight shortly after take-off: four crew members died, and two were severely injured [22]. Preliminary results of the case’s investigation led to the conclusion that the engines received conflicting commands from the aircraft’s control unit. This resulted in the crash either due to a limitation of the engines’ thrust level

or a complete engine shutdown [14]. Another critical case with luckily no victims was Lufthansa flight LH 1829 on November 5, 2014 [45]. This list of incidents is continued by the Air France crash on the way from Rio de Janeiro to Paris in May 2009 with 228 victims [44] or the crash of an AirAsia Airbus A320 close to the Indonesian coastline in December 2014 (162 people died) [46]—the importance of conflict handling (also of course in other areas) should be clear at this point.

To a certain extent, conflict handling is independent from the model applied to represent information. Whereas probability theory [2, 16, 19], fuzzy set theory [49], and possibility theory [8, 50] need to incorporate further processing steps for conflict handling, Dempster-Shafer theory of evidence [6, 40] is inherently designed to handle conflicts. Conflict in a fusion process represents inherent uncertainty. Therefore, information of the applied sensors and consequently their information itself contained in the result of the fusion process are not 100 % reliable. Thus, an importance measure is directly connected to a conflict measure.

This chapter proposes to apply the multilayered information fusion system multilayer attribute-based conflict-reducing observation (MACRO) [33, 34, 37] employing the balanced two-layer conflict solving (BalTLCS) [26, 33, 37] fusion algorithm to reduce the impact of conflicts on the fusion result. The performance of the contribution is shown by its evaluation in the scope of a machine condition monitoring application under laboratory conditions. Here, the MACRO system yields promising results compared to other state-of-the-art fusion mechanisms.

22.2 Sensor Conflict and Importance in Information Fusion

In this section we describe the nature and psychological effects of *conflict* and *importance* as well as the state-of-the-art of conflict measures and information fusion in this context.

22.2.1 *Conflict and Importance*

Conflict occurs whenever information does not bear evidence for only one opinion/proposition but also for another. This might either be due to actual failure in the observed process or system or caused by one or more defective sensors. The latter case is the most severe one since wrong decisions might be derived if sensors were considered reliable, although they are not.

The basic idea of conflict and importance relies on the fact that non-conflicting information is psychologically co-noted with reliability in the meaning of *importance*. Therefore, in a high-conflict case, the result should be treated as less important for a condition or situation because high conflict indicates low reliability.

22.2.2 State of the Art

A number of publications work on the improvement or substitution of conflict measured and processed in the combination rule of Dempster-Shafer theory of evidence (DST) [6, 40]. Martin et al. propose a conflict measure based on the distance between belief functions. This measure additionally serves to determine a posteriori the reliability of the recently processed data [29, 30]. Smarandache et al. put this new approach into context and benchmarked it against other conflict measures (which they call “contradiction measures”) [42]. A measure based on vector distances between the data to be fused is introduced in [25]. Minor and Johnson do not consider conflict as uncertainty originating from data source reliability issues, but from uncertainty in the frame of discernment. The sources’ reliabilities must be questioned in this case as they are then applied to observe an inappropriate situation (the augmented frame of discernment) [31]. Another conflict measure is developed in [3–5]. It is based on the internal conflict between belief functions, which increases when decreasing belief is assigned to the evaluated propositions. To date, this concept is developed axiomatically [5].

A combination rule along with a conflict measure, based on the average of the individual beliefs, is introduced in [9]. This paper concludes that the arithmetic mean is typically the best combination rule but admits at the same time that the choice of the correct combination rule is context-sensitive. The implicative importance weighted ordered weighted averaging (IIWOWA) operator [23] is an extended version of ordered weighted averaging (OWA) [48], which allows for weighting each element with respect to its importance in the current problem. It is the normalised version of the importance weighted ordered weighted averaging (IWOWA) to achieve value equivalence instead of order equivalence to Yager’s weighted arithmetic mean (WAM) operator [23].

22.3 Models and Measures

Conflict is understood as epistemic uncertainty: if more information is available, it is possible to reduce or resolve the inherent conflict completely. It has a substantial influence on the fusion result. The influence changes with respect to the total number of sources. Its behaviour in such cases must be known in advance. Thus, analytical and numerical evaluations are carried out. DST-based fusion models are capable to handle epistemic uncertainties. Therefore, an approach using DST’s well-defined and researched Dempster’s rule of combination (DRC) [40] as a basis, but tackling its problems, is elaborated and defined in this contribution. This fusion algorithm denoted by BalTLCS applies psychological concepts derived from human group decision processes to stabilise the fusion result. The fusion algorithm is the core fusion operator applied in the MACRO system [33], which is introduced in the following section.

22.3.1 Multilayer Attribute-Based Conflict-Reducing Observation

The architecture of the MACRO fusion system [33, 37] is designed to resemble the actual structure of the monitored system, which is partitioned into several subsystems on the one hand. This kind of architecture is found in contemporary system design of several application fields. On the other hand, MACRO's architecture facilitates the implementation of multimodal systems in the sense of [39].

The purpose of the MACRO fusion system is to determine and assess the state of a complete system by monitoring its subsystems and properties. The following terminology applies:

Definition 1 (MACRO terminology [33]) The terminology in the context of the MACRO system is defined as follows:

Sensors: A set of sensors $\mathcal{S} = \{S_s\}, s \in \mathbb{N}_N$ acquires the signals of the monitored system (physical device) or its environment. Its *physical effects* determine the sensors' signals, which are output as raw data d_s . The term "physical" hereby encloses all effects, which the monitored physical device influences or is exposed to, hence also biological and chemical effects.

Signal conditioning: A number of signal conditioning blocks $SC_j, j \in \mathbb{N}_F$ extract each one feature f_j from the raw sensor data d_s . It may also include signal preprocessing procedures.

Attribute layer: The *attribute layer* consists of a number of attributes, each containing an *attribute fusion* algorithm.

Attribute: An attribute $a \in \mathcal{A}$ represents a characteristic (physical quantity, functionality, component, etc.) of the monitored system that is represented by at least two features. The attributes depend both on the monitored system and the application MACRO is utilised in and are defined by expert's knowledge. Given the hierarchy of the monitored system, four types of attributes are defined in the following taxonomy:

Module attribute: An attribute a is a *module attribute* iff it represents a single module or component that is part of the monitored system.

Physical attribute: An attribute a is a *physical attribute* iff it characterises a single elementary (physical, biological, chemical) phenomenon of a specific module.

Functional attribute: An attribute a is a *functional attribute* iff it characterises functionality of the monitored entity with respect to a specific module.

Quality attribute: An attribute a is a *quality attribute* iff it assesses the output (e.g. fabricated product) of the monitored system.

An attribute's output indicates to which degree its inputs represent the system's normal condition and is denoted by *attribute health*.

System layer: All attribute healths are fused on *system layer* by the *system fusion* algorithm. It determines and assesses the current system state denoted by *system*

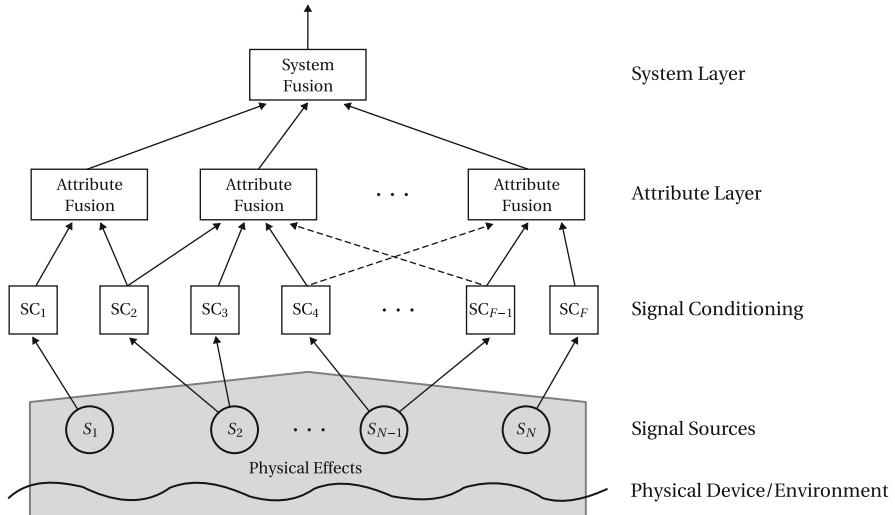


Fig. 22.2 Architecture of the *multilayer attribute-based conflict-reducing observation* system MACRO [33, 34]

health. The system health is MACRO's final output and indicates to which degree the input attribute healths represent the system's normal condition.

MACRO's architecture is depicted in Fig. 22.2.

The MACRO architecture offers the following general key properties:

- The fusion architecture describes the information flow from the bottom to the top and is as such independent from both the choice of signal conditioning approaches as well as from any fusion technique.
- The system health is determined based on a number of attribute healths. An anomaly will be first noticed by a decrease in the system health. The subsequent evaluation of the attribute healths assists in narrowing down the location of the anomaly.
- Due to the resemblance of the physical system in the structure of MACRO, a transparent monitoring system is created. This property assists in interpreting results during MACRO's runtime by the system operator.
- The architecture supports the possibility to implement a distributed information fusion system in real-world applications. Instead of transmitting the raw data to the system layer, the aggregated and thus compressed attribute healths are transmitted.

Its building blocks are instantiated based on [33] in the following. On the lowest layer of the MACRO architecture, N signal sources (*sensors*) S_s capture the physical effects, which a system is exposed to and influences, respectively. Each sensor delivers a signal d_s . These signals are heterogeneous in type and dimension. Consequently, the various signal data are incomparable and must be transferred into

the same space, before information fusion can take place. As information fusion on symbol level is generally too rigid [15, 28], information fusion inside the MACRO architecture is carried out on feature level.

In order to extract the features f from the signals, they are processed in the blocks labelled SC . The processing steps are denoted by *signal conditioning* in the MACRO architecture (cf. Fig. 22.2). Analogue signals are digitised by sampling before features are extracted from them. The content of each of the F signal conditioning blocks is always application-dependent and cannot be defined generally. Whereas one feature is delivered per signal conditioning block, this feature f may be input of an arbitrary number of attributes. The same applies to the sensor's data: the data d_s originates from one sensor S_s and may be the input of a number of signal conditioning blocks SC . Hence, more than one feature may be derived from d_s .

A fuzzy set theory [49] approach has been chosen for modelling the acquired data in a common unitless space between 0 and 1. It is capable to model uncertainty in the data, which is coming from, e.g. sensor noise, and allow variations in the system's behaviour due to environmental changes (e.g. in temperature, humidity), which do not affect the fulfilment of the system's task. The Modified-Fuzzy-Pattern-Classifier (MFPC) [27] models the information by a unimodal potential function applied as fuzzy membership function $\mu_s : \mathbb{R} \rightarrow [0, 1]$. This information model has proven its performance scientifically as well as in real-world applications (e.g. [27, 35, 38]). It employs an automatic learning procedure to determine the membership function based on measurement data during operation of the system to be monitored in its normal condition N^C . Details on the learning procedure are presented in [33, 37].

On the attribute layer, the membership functions representing the observed system's normal condition $N\mu_s$ are fused to determine the attribute health. Uncertainties, to which all the sensor signals and the features are prone, are treated by modelling the signals as fuzzy membership functions. Conflicts between them remain unsolved and are handled by the BalTLCs fusion operation on attribute layer. Groups of features f are constructed, of which each group represents the same component or property of the monitored system. Such a group is denoted by *attribute* a . The features originate from different signal sources S , so that sensor defects affect an attribute only to some degree during fusion. This also decreases an attribute's uncertainty.

An attribute's importance $I_a \in [0, 1]$ represents the weight of an attribute in the fusion on system layer: the higher an attribute's importance, the more the attribute influences the system fusion result. The importance of a MACRO attribute is determined continuously based on the conflict between the attribute's inputs during fusion on attribute layer. Hence, this information is to be incorporated on system layer during determination of the system health. Note that manual determination of the importance is also possible, e.g. a priori (by an expert), and set statically. A dynamic approach is nevertheless more beneficial as dynamic changes of the monitored system (change of the system's operation point, varying sensor reliabilities, etc.) are considered.

Whereas the OWA operator [48] is suitable for integrating optimism in the fusion process by its andness [33], it is not prepared to consider attribute importances. Importances are integrated to the OWA operator by the implicative importance weighted ordered weighted averaging (IIWOWA) operator introduced in [23]. It supports the integration of both parameters andness and importance and is proposed to determine the *system health*.

The entire approach supports that faulty sensors, which are in contradiction with the other fault-free sensors, do not significantly affect the overall fusion result. This is achieved first by the BalTLCS fusion itself, which inherently detects and handles conflicts between inputs. In addition, the amount of conflict determined by BalTLCS is forwarded to the subsequent IIWOWA fusion operation on system layer. Here, attributes containing a considerable amount of conflict are devalued because their conflict is interpreted as uncertainty connected with the attribute. Consequently, attributes containing no or only a small amount of conflict are regarded as important and contribute more to the system health than the unimportant attributes, which are full of conflict. Hence, the confidence and thus the quality of the overall result are increased compared to fusion approaches not incorporating such mechanisms.

The BalTLCS fusion operator is the integral part allowing for quality-aware information fusion based on importance measures and is introduced in the next section.

22.3.2 Numerical Stability's Influence on Information Fusion Quality

The BalTLCS fusion operator is on the one hand based on Two-Layer Conflict Solving (TLCS) to exploit its positive properties elaborated in [33]. BalTLCS offers the following properties:

- adoption of effective human group decision-making principles,
- determination of conflicts between inputs,
- solution of the conflicts, such that their effect on the fusion result is decreased,
- creation of intuitive fusion results, also in high-conflict cases.

On the other hand, the deficiencies of TLCS identified in [33, 37] (with specific focus on its counter-intuitive fusion results in high-conflict cases) are mitigated. This consequently facilitates high-quality fusion results. In order to illustrate the mentioned findings on TLCS, its numerical stability in the range of the conflict's limits is evaluated in the next section. These will show the potential for improvements, which BalTLCS incorporates.

22.3.2.1 Two-Layer Conflict Solving

The TLCS fusion approach was analysed in detail in [33]. It was shown that a number of amendments need to be incorporated in its original definition from [24] to avoid undefined numerical situations. Thus, TLCS is applied according to [33] in this contribution:

Definition 2 (Adapted two-layer conflict solving fusion [33]) For $n \geq 2$ sensors S_s and $o \geq 2$ propositions A_i , the TLCS fusion operation determines the fused basic belief assignment (BBA) to proposition A by

$$m(A) = \frac{\sum_{s=1}^n m_s(A) + (\text{Bc}(n) + \text{acc}(K_{\text{cm}})) \cdot \text{CMDST}(A)}{n + \text{Bc}(n) + \text{acc}(K_{\text{cm}})}, \quad (22.1)$$

where $\text{Bc}(n) = \binom{n}{2}$ is the binomial coefficient, the *accord* $\text{acc}(K_{\text{cm}}) = |\log(K_{\text{cm}})|$, the *adapted conflicting factor*

$$K_{\text{cm}} = \frac{1}{\text{Bc}(n) - k_{\text{cm}} + \varepsilon}, \quad \varepsilon \in \mathbb{R} \text{ with } 0 < \varepsilon \ll 1, \quad (22.2)$$

the conflicting coefficient

$$k_{\text{cm}} = \sum_{s=1}^{n-1} \sum_{t=s+1}^n \sum_{i=1}^o m_s(A_i) \cdot (1 - m_t(A_i)), \quad (22.3)$$

and the Conflict-Modified-DST

$$\text{CMDST}(A) = \frac{\sum_{s=1}^{n-1} \sum_{t=s+1}^n m_s(A) \cdot m_t(A)}{\sum_{s=1}^{n-1} \sum_{t=s+1}^n \sum_{i=1}^o m_s(A_i) \cdot m_t(A_i)}. \quad (22.4)$$

The analysis in [33] that TLCS creates counter-intuitive fusion results due to numerical instabilities. In order to illustrate the cause of the numerical instability in case of maximum conflict, Conflict-Modified-DST (CMDST) is investigated in the scope of a relaxed maximum conflict case.

Definition 3 (Relaxed maximum conflict [33]) Let n denote the number of sensors S_s , o the number of propositions A_i with $o = n$, and λ the conflict relaxation parameter with $\lambda \ll 1$. The case of *relaxed maximum conflict* is defined such that each sensor S_s assigns a BBA of $m_s(A_i) = 1 - \lambda$ to an arbitrary proposition A_i and to another proposition $A_j \neq A_i$ a BBA of $m_s(A_j) = \lambda$. All other propositions are assigned zero BBAs.

Table 22.1 Example BBAs $m_s(A_i)$ for three sensors S_s and three propositions A_i in the case of *relaxed maximum conflict* [33]

	A_1	A_2	A_3
S_1	$1 - \lambda$	λ	0
S_2	λ	0	$1 - \lambda$
S_3	0	$1 - \lambda$	λ

The other sensors assign BBAs in the same way such that $\sum_{i=1}^o m_s(A_i) = 1$ for all s and $\sum_{s=1}^n m_s(A_i) = 1$ for all i . Hence, for each proposition A_i sensor S_s assigns BBA $m_s(A_i) = 1 - \lambda$, and another sensor $S_t \neq S_s$ assigns $m_t(A_i) = \lambda$, whereas all remaining BBAs are 0.

A valid example of relaxed maximum conflict for $n = o = 3$ is provided in Table 22.1.

The CMDST fusion result for arbitrary n in the case of relaxed maximum conflict is obtained as given in [33]:

$$\text{CMDST}(A_i) = \frac{1}{n} \cdot \frac{1 - \lambda}{1 - \lambda + \frac{\varepsilon}{n\lambda}}. \quad (22.5)$$

Analytic evaluation of $\lim_{\lambda \rightarrow 0} \text{CMDST}(A_i)$ yields

$$\begin{aligned} \varepsilon > 0 : & \quad \text{CMDST}(A_i) = 0 \\ \lim_{\lambda \rightarrow 0} \frac{1}{n} \cdot \frac{1 - \lambda}{1 - \lambda + \frac{\varepsilon}{n\lambda}} &= 0 \quad \lim_{\lambda \rightarrow 0} \frac{1}{n} \cdot \frac{1 - \lambda}{1 - \lambda} = \frac{1}{n} \end{aligned} \quad (22.6)$$

The variable $0 < \varepsilon \ll 1$ was introduced in Eq. (22.2) in order to numerically avoid undefined states of K_{cm} in high-conflicting cases ($k_{\text{cm}} \rightarrow \text{Bc}(n)$) and is thus an artificial addition. In CMDST however, it determines and falsifies the output, also analytically (cf. Eq. (22.6)).

The error term $\Delta = \frac{\varepsilon}{n\lambda}$ in Eq. (22.5) shows that ε and λ are proportionally dependent on each other. The reason for the discovered numerical instabilities in case of $\lim_{\lambda \rightarrow 0} \text{CMDST}(A_i)$ for all $\varepsilon > 0$ follows from [33]

$$\lim_{\lambda \rightarrow 0} \frac{\varepsilon}{n\lambda} = \infty,$$

regardless of the number of sensors n . As a consequence, TLCS yields counter-intuitive fusion results in high-conflicting cases due to numerical instabilities.

In order to bound this effect and facilitate fusion quality, the error term Δ is limited, as derived in the following. The value of its constituent ε is typically predefined in implementations by the accuracy of the software or the processing unit; n is the number of sensors applied in the fusion process. Hence, Δ depends on λ , for which the lower bound λ is approximated by

Proposition 4 (Lower bound of the conflict relaxation parameter [33]) *In order to limit the error term Δ , the lower bound of the conflict relaxation parameter λ is determined by*

$$\lambda \geq \frac{\varepsilon}{n\Delta}. \quad (22.7)$$

The following example provides approximations of λ for MATLAB implementations.

Example 1 (Lower bounds of the conflict relaxation parameter in MATLAB implementations [33]) MATLAB implements floating-point numbers according to the IEEE standard 754 [17, 18, 43]. Its accuracy¹ is $\varepsilon = 2^{-52} \approx 2.2204 \cdot 10^{-16}$. Then the following lower bounds for λ result in case of arbitrarily chosen Δ :

$$\begin{aligned} \Delta \leq 10^{-10} : & \quad \Delta \leq 10^{-5} : \\ \lambda \gtrsim \frac{1}{n} \cdot 2.2204 \cdot 10^{-6}, & \quad \lambda \gtrsim \frac{1}{n} \cdot 2.2204 \cdot 10^{-11}. \end{aligned}$$

The trade-off between the error and allowable maximum conflict in order to guarantee numerical stability is as follows: the less error Δ is allowed, the larger is the smallest allowable value of λ . This consequently means that maximum conflict in the input data must be more relaxed the smaller Δ is allowed to be.

From a practical point of view, this evaluation facilitates a check on the input data when relaxed maximum conflict in the data is detected. If $\lambda \geq \frac{\varepsilon}{n\Delta}$ is not satisfied, a critical situation for numerical stability is detected, in which the results cannot be trusted. Such situation is avoidable by exception handling: the fusion result is annotated as being unreliable and thus not further processed automatically. At the same time, a notification is triggered for the on-duty supervisor who then can investigate the scene manually.

The number of sensors n influences on the one hand the lower bound of λ and on the other hand determines the *condition* of $\text{CMDST}(A_i)$. The absolute condition relates the difference in the output of $\text{CMDST}(A_i)$ to the infinitesimal difference of its input, hence

$$\text{cond}(\text{CMDST}(A_i)) = \left| \frac{\delta \text{CMDST}(A_i)}{\delta \lambda} \right|.$$

A problem is denoted *well-conditioned* if its condition is $\text{cond} \approx 1$, i.e. an infinitesimal change of the input results in an infinitesimal change of the output. This is interpreted as numerical stability of the problem. It is numerically evaluated

¹The presented information is derived from MATLAB 2016a (9.0.0.341360) 64-bit for Microsoft Windows. The declarations are also valid for earlier versions of MATLAB according to [32].

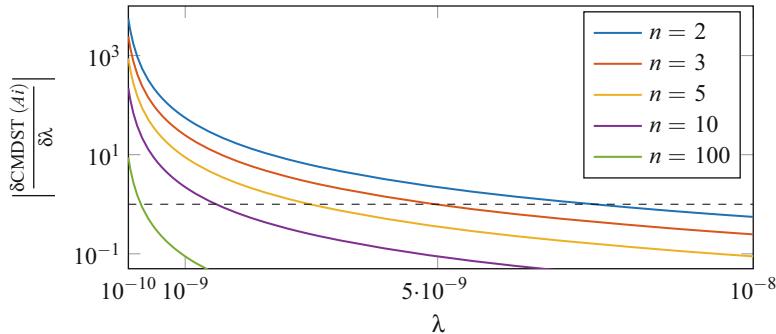


Fig. 22.3 Absolute condition of $\text{CMDST}(A_i)$ in the relaxed maximum conflict case determined numerically for $\varepsilon = 2^{-52}$, $\delta\lambda = 10^{-12}$, and $n \in \{2, 3, 5, 10, 100\}$. The dashed line represents $\text{cond}(\text{CMDST}(A_i)) \approx 1$, denoting the boundary below which a problem is denoted well-conditioned [33]

for CMDST with respect to λ in the relaxed maximum conflict case with $\varepsilon = 2^{-52}$, $\delta\lambda = 10^{-12}$, and $n \in \{2, 3, 5, 10, 100\}$. The results are visualised in Fig. 22.3.

The CMDST combination rule is well-conditioned for roughly $\lambda > 10^{-8}$ regardless of n . In this context it must be noted that increasing n facilitates numerical stability for smaller values of λ and hence precise processing of $\text{CMDST}(A_i)$. At the same time, it must be ensured that λ does not fall below the boundary determined by Eq. (22.7) in order to satisfy the desired limit of the error term Δ . The next section will show that BalTLCS is instead numerically stable and not demanding such adjustments in its parameters.

22.3.2.2 Balanced Two-Layer Conflict Solving

The Two-Layer Conflict Solving (TLCS) approach introduced by Li and Lohweg is a promising candidate for attribute layer fusion. However, its analysis shows that adaptations are necessary especially when real-world applications are considered [33, 37]. These findings are applied in the design of the BalTLCS fusion algorithm. It determines intermediate fusion results with respect to non-conflicting and conflicting BBAs, which are subsequently combined in an additive way:

Definition 5 (Balanced two-layer conflict solving Fusion [33, 37]) For $n \geq 2$ sensors S_s and $o \geq 2$ propositions A_i , the BalTLCS fusion operation determines the fused BBA to proposition A by

$$m(A) = m_{nc}(A) + m_c(A). \quad (22.8)$$

where the *non-conflicting part* is determined as

$$m_{nc}(A) = \frac{1}{Bc(n)} \sum_{s=1}^{n-1} \sum_{t=s+1}^n m_s(A) \cdot m_t(A), \quad (22.9)$$

and the *conflicting part* is determined as the arithmetic mean of all input BBAs weighed by *normalised conflicting coefficient c*

$$m_c(A) = c \cdot \frac{1}{n} \sum_{s=1}^n m_s(A), \quad (22.10)$$

with c modelling the degree of conflict between individual beliefs as

$$c = \frac{1}{Bc(n)} \sum_{s=1}^{n-1} \sum_{t=s+1}^n \sum_{i=1}^o m_s(A_i) \cdot (1 - m_t(A_i)) = 1 - \sum_{i=1}^o m_{nc}(A_i). \quad (22.11)$$

Whereas the non-conflicting part is determined by pairwise aggregation, the conflicting part considers all sensors at the same time. Hence, BalTLCS follows the same concept, which is applied in TLCS: decision-making in the whole group employs the intermediate result, which has been found in “bilateral discussions” and the original BBAs of all sensors.

The BBA assigned to the frame of discernment, which represents the amount of ignorance, is determined by

$$m(\Theta) = 1 - \sum_{A_i \subset \Theta} m(A_i).$$

Considering the limits of conflict, the following properties of BalTLCS are derived:

No conflict: $c = 0 \Rightarrow m_c(A) = 0 \Rightarrow m(A) = m_{nc}(A)$.

Maximum conflict: $c = 1 \Rightarrow m_{nc}(A) = 0, m_c(A) = \frac{1}{n} \Rightarrow m(A) = \frac{1}{n}$.

Hence, if no conflict occurs, the non-conflicting part m_{nc} determines the overall fusion result. If the conflict is maximal, then all information sources have to be taken into account, which is achieved by m_c determining the arithmetic mean of all sensory hypotheses. A balance between conflicting and non-conflicting beliefs is established by the additive connection utilising the conflicting coefficient c as a control parameter.

This balancing is illustrated by the following numerical examples on the conflict's limits given in Table 22.2.

The evaluation yields the expected results. In the case of no conflict, all aggregated belief is assigned to proposition A_1 , which is the only proposition that the sensors assign beliefs to. The conflicting parts m_c are all 0 due to the same

Table 22.2 Example BBAs $m_s(A_i)$ for three sensors S_s and three propositions A_i in the case of (a) *no conflict*, and (b) *maximum conflict* along with the results obtained by the BalTLCS fusion approach [33, 37]

(a) No conflict					No conflict	Maximum conflict
	A_1	A_2	A_3	c	0.000	1.000
S_1	1	0	0	$m_{nc}(A_1)$	1.000	0.000
S_2	1	0	0	$m_{nc}(A_2)$	0.000	0.000
S_3	1	0	0	$m_{nc}(A_3)$	0.000	0.000

(b) Maximum conflict					No conflict	Maximum conflict
	A_1	A_2	A_3	$m_c(A_1)$	0.000	0.333
S_1	1	0	0	$m_c(A_2)$	0.000	0.333
S_2	0	0	1	$m_c(A_3)$	0.000	0.333
S_3	0	1	0	$m(A_1)$	1.000	0.333
				$m(A_2)$	0.000	0.333
				$m(A_3)$	0.000	0.333

reason; thus, the fusion result is determined only by the non-conflicting parts. In the case of maximum conflict, no BBAs are assigned to the non-conflicting parts m_{nc} . Instead, all belief is assigned to the conflicting parts m_c , to which the beliefs of all sensors in the respective propositions are equally assigned. Consequently, each fusion result is determined only by its conflicting part.

The critical case with respect to numerical stability for TLCS and its constituent component CMDST is the relaxed maximum conflict case (cf. Example 3). The BalTLCS approach yields in this case

$$m(A_i) = \frac{1}{n}.$$

Considering its condition with respect to λ

$$\text{cond}(m(A_i)) \left| \frac{\delta m(A_i)}{\delta \lambda} \right| = 0,$$

numerical instabilities will not affect the result: regardless of the change of λ , the result will be the same. In contrast to TLCS, no artificial parameter needs to be introduced in the constituent parts of BalTLCS to numerically avoid undefined situations. The operations employed in BalTLCS on the BBAs to be fused are bounded [33, 37]. Hence, no numerical instabilities are expected.

The numerical examples presented above validate the stability of BalTLCS also in case of no and maximum conflict. In contrast to TLCS, BalTLCS yields the expected fusion results. Conflict is also determined and considered in the fusion process. The importance measure introduced in the following section is based on this conflict measure.

22.3.3 Conflict as a Measure of Importance

Conflict in a fusion process represents inherent uncertainty. Therefore, the information of the applied sensors and consequently the information contained in the result of the attribute's fusion are not 100% reliable. Thus, the *importance measure* I_a of the attribute a is defined as follows:

Definition 6 (Importance measure [33, 37]) Let I_a be the information weight in a fusion process, which estimates the impact of a conflict regarding the aggregation of sensor information in attribute a . Let ${}_a^N\mu$ be the fused result of a balanced two-layer conflict solving (BalTLCS) process regarding proposition ${}^N\mathcal{C}$ with the conflicting coefficient $c_a \in [0, 1]$. Then $I_a : c_a \rightarrow [0, 1]$ is the corresponding information weight of the fusion result ${}_a^N\mu$, which is dependent on the attribute's conflicting coefficient c_a . The information weight is denoted by *importance measure*. It is determined by

$$I_a = 1 - c_a. \quad (22.12)$$

Proof In case of low conflict ($c_a \rightarrow 0$), the importance must be high and vice versa. Hence, the importance moves in the opposite direction of the conflicting coefficient. Therefore, the sum of conflicting coefficient and importance must be constant, hence $c_a + I_a = \sup(c_a) = 1$ due to $c_a \in [0, 1]$ and $I_a \in [0, 1]$. It follows

$$I_a = 1 - c_a.$$

□

The conflicting coefficient encodes information about the uncertainty involved in the fusion process: the smaller c_a , the lower is the uncertainty [33, 37]. This principle is exploited in the concept of *importance*. The importance is the complement of the conflicting coefficient. This expresses that the fusion result is more important the less conflict has been determined during fusion and vice versa. In addition, sensor defects directly influence the conflict between sensor inputs. Hence, the conflict (hence the attribute's negated importance in the context of MACRO, delivered at no additional cost) seems to be an appropriate indicator for a possible sensor defect.

22.4 Applications

Industrial applications are in transition towards modular and flexible architectures that are capable of self-configuration and self-optimisation. This is due to the demand of mass customisation and the increasing complexity of modern industrial systems. Sensors, actuators, but also other sources like databases serve as data

sources for the realisation of condition monitoring in industrial applications or for the acquisition of characteristic parameters, such as production speed or reject rate. The data originates from sources which are spatially distributed over the shop floor. Modern industrial plants are equipped with an increasing number of sensors generating a large amount of data. The task of processing these amounts becomes increasingly complex. Computation takes longer and necessary communication may exceed available bandwidth. Furthermore, machine operators are unable to properly process and draw correct conclusions from the generated information [36]. Complex machineries make it increasingly difficult for a system designer to comprehend the overall industrial plant. In such complex systems, the acquisition of sensor signals must be designed very carefully and tailored optimally towards the specific application. This challenge aggravates with the progressing introduction of modular and flexible systems and devices.

The conversion to modular systems, like in *intelligent technical systems* or *cyber-physical systems* for Industry 4.0, is related to challenges in all disciplines. Consequently, diverse tasks like information processing, extensive networking, or system monitoring using sensor and information fusion systems need to be reconsidered. In modern industrial plants, the idea of flexible systems and devices will be realised, especially at runtime. Up until now, flexibility was often only predesigned, which demands a designer to consider all possible situations beforehand. This is about to change in modern applications. Distributed sensor and information fusion for system monitoring, which must reflect the increasing flexibility of fusion systems, are in the focus in such applications. With respect to the human system designer and operator, an IFU system has to be transparent, understandable, and traceable. These properties allow erroneous situations to be properly detected and resolved adequately. Consequently, the following requirements for a design methodology are derived [33]:

- The application's requirements have to be fulfilled by a proper selection of sensors with respect to the measured quantity, the measurement range, and resolution.
- The system designer should only be assisted in the design process and must remain as final decision instance. Solutions for the design should at most be suggested such that the system designer can choose the most appropriate one.
- Each design of an IFU system depends on the specific application. Nevertheless, partial solutions are reusable and should therefore be considered before identifying a completely new IFU system design. Consequently, repositories for storage of problem formulations and solutions have to be available that hold information in a defined manner to identify similarities.
- Attributes of the MACRO system or their input signals include descriptive information to automatically generate, update, and destroy the attributes. Hence, available autoconfiguration mechanisms have to be extended by a fusion system design methodology in order to be able to process self-descriptive data that originates from intelligent sensors.

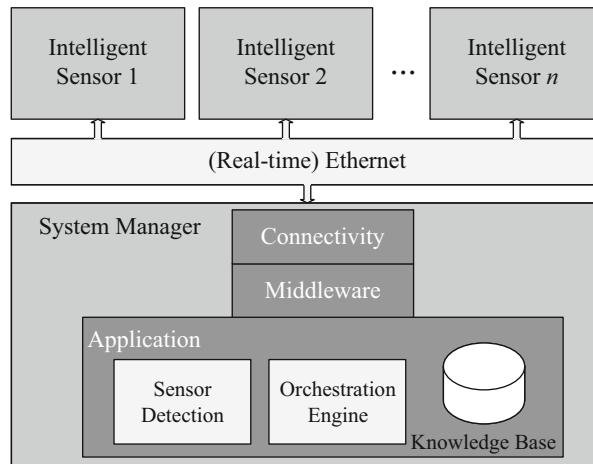


Fig. 22.4 Architecture of the self-configuring fusion system [11]

One previously published approach relies on a network of self-descriptive intelligent sensor nodes including self-description information for the automatic design and update of sensor and information fusion systems [12, 13]. It automatically designs an IFU system but gives the system designer the final authority over the actual implementation. The methodology relies on a rule-based decision system, which evaluates semantic descriptions delivered by the involved sensors. The architecture of the IFU system is depicted in Fig. 22.4. It forms a sensor network consisting of n intelligent sensors that are capable to communicate among themselves and with the *system manager*. The system manager implements functionalities for automated system design and self-configuration.

In this case the system manager is a central processing unit to detect available sensors and process their self-descriptive information in order to propose an IFU system for the specific application. It also monitors the IFU system for changes and adjusts it accordingly. Fritze et al. summarise distributed systems in the form of *multi-agent systems* and their advantages regarding self-organisation and self-adaptation in [13]. Because of better scaling and the avoidance of a single point of failure, such systems fit also into the concept of automated fusion system design. Nonetheless, industrial applications require real-time communication channels for process data exchange to be able to react to changes in process real time. Thus, it must also be considered in IFU, especially in distributed IFU systems. There is currently no real-time communication standard that is capable to fulfil the requirements for decentralised data exchange. Therefore, a central processing unit is indispensable when real-time communication is required. Consequently, the concept presented in the following additionally incorporates a central processing unit in form of the system manager.

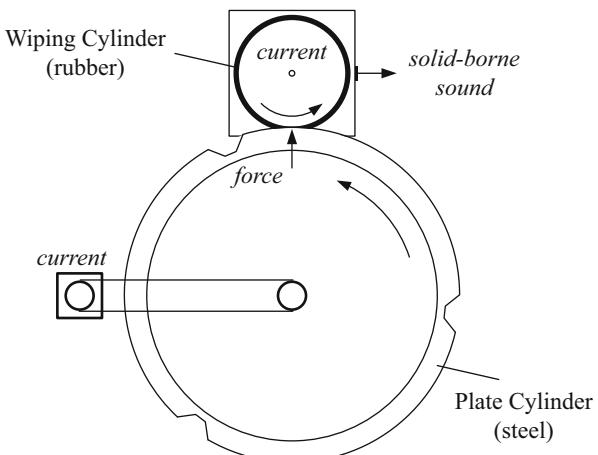
Trust in the processed information is crucial at this point. This applies to both the self-descriptive information and the sensor signals, which are fused in the IFU system. Importance measures, such as the one introduced in Sect. 22.3.3, help to integrate and maintain trustworthiness. Generally speaking, the specific importance of a piece of information is considered by weighting the information. Smaller importance measures denote smaller importance but are also applied to express the level of confidence assigned to the piece of information.

The importance measure defined in Sect. 22.3.3 evolves from the MACRO information fusion system. There, it is applied to devalue attributes incorporating high levels of conflict. This consequently leads to higher quality of the entire fusion system result. Mönks et al. showed this, e.g. in a laboratory printing unit supervision experiment [33, 37]. The utilised demonstrator contains models of two cylinders applied in the printing unit, which are turned by electric drives. The pressure between the *wiping cylinder* having a rubber surface and the steel-surfaced *plate cylinder* is freely adjustable. Four analogue sensors (force, solid-borne sound, electric current of each drive) continuously acquire data during operation to monitor the process. The demonstrator setup is schematically shown in Fig. 22.5.

The signals of the sensors are processed by MACRO as well as TLCS and the quasi-standard naïve Bayes and Support Vector Machine (SVM) fusion operators [33]. In order to illustrate the effect of the importance measure incorporated in MACRO, this contribution concentrates on naïve Bayes and MACRO fusion. The findings are similar for the remaining fusion operators and are found in [33].

During the experiment, the solid-borne sound sensor is manipulated to enforce a conflict between the sensor signals. The manipulation is induced by manual and continuously increasing low-pass filtering of the signal between 03:45 (min:s) and 04:33 (min:s) of the experiment. The filter status is then kept until 06:36 (min:s), when the low-pass filter is disabled. In addition a real defect of the demonstrator

Fig. 22.5 Structural design of the printing unit simulator along with the applied sensors (printed in italic) [47]



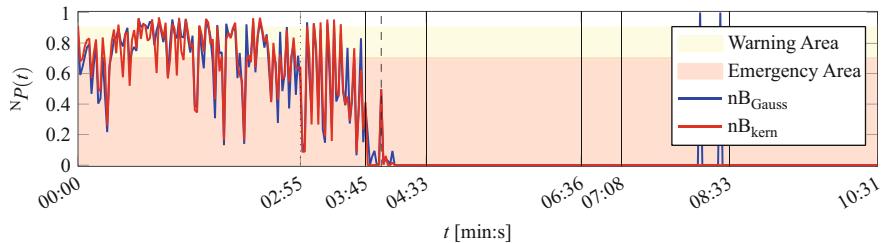


Fig. 22.6 System health evaluation over time during manipulated operation of the printing unit demonstrator by one-class naïve Bayes applying Gaussian (nB_{Gauss}) and kernel density estimated (nB_{kern}) priors [33]

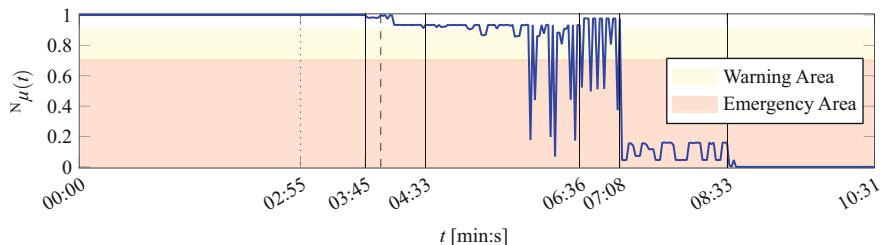


Fig. 22.7 Evaluation of the system health N_μ over time obtain by MACRO fusion during manipulated operation of the printing unit demonstrator [33]

occurs from 07:08 (min:s) on. The fusion results yielding the *system health* of the printing unit demonstrator are depicted in Fig. 22.6 for naïve Bayes and in Fig. 22.7 for MACRO fusion.

The “sensor defect” from 03:45 (min:s) on leads the naïve Bayes fusion result in the wrong direction: it classifies the demonstrator’s system health being in the emergency area from that moment on, although no defect occurred at the demonstrator itself. MACRO’s importance weighting on system layer results in correct classification of the demonstrator status: the conflict between the solid-borne sound sensor’s signal and the other sensors leads to a decreased importance measure for the respective attributes. Their attribute healths are therefore incorporated to a smaller extent in the system health determination, leading to a compensation of the sensor defect. For further details on the experiment illustrating the beneficial influence of the importance measure introduced in Sect. 22.3.3, the reader is referred to [33, 37].

Besides direct incorporation of the importance measure in the fusion process, it is also applicable to monitor sensors for defects. Ehlenbröker et al. exploit MACRO’s multilayered structure to determine sensor reliabilities [10]. Their approach necessarily needs groups of sensors, which are delivered at no additional cost by MACRO’s attributes. Consistencies between sensor signals are determined among the sensors of each attribute. The more consistent the signals, the higher the reliability of the considered sensor. If the continuously monitored reliability falls

below a given threshold, the sensor is considered to be defect. The defective sensor can then be repaired or replaced to maintain information fusion quality. Detailed information is found in [10]. As increasing conflict within an attribute will lead to decreased consistency between sensor signals, the importance seems also to be appropriate to serve as a measure for monitoring sensor reliability.

22.5 Conclusions

The handling of conflicts between information sources is crucial for the reliability of the result of an information fusion application. This contribution focuses on conflict and importance represented in the multilayer attribute-based conflict-reducing observation (MACRO) information fusion system. The contribution describes the attribute layer fusion algorithm balanced two-layer conflict solving (BalTLCS), which is capable to determine conflicts between fusion inputs and decrease their effect on the fusion result. Furthermore, the numerical stability in the context of heavy conflicts and the related importance is described. This fact is crucial in application implementations. Different examples, which show the applicability of the described conflict/importance measure, are given.

References

1. B.M. Ayyub, G.J. Klir, *Uncertainty Modeling and Analysis in Engineering and the Sciences*. (Chapman & Hall/CRC, Boca Raton, 2006)
2. C.M. Bishop, *Pattern Recognition and Machine Learning*. Information Science and Statistics, 8th edn. (Springer, New York, 2009)
3. M. Daniel, Belief functions: a revision of plausibility conflict and pignistic conflict, in *Scalable Uncertainty Management*, ed. by D. Hutchison, T. Kanade, J. Kittler, J.M. Kleinberg, F. Mattern, J.C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M.Y. Vardi, G. Weikum, W. Liu, V.S. Subrahmanian, J. Wijnen. Lecture Notes in Computer Science, vol. 8078 (Springer, Berlin/Heidelberg, 2013), pp. 190–203
4. M. Daniel, Properties of plausibility conflict of belief functions, in *Artificial Intelligence and Soft Computing*, ed. by D. Hutchison, T. Kanade, J. Kittler, J.M. Kleinberg, F. Mattern, J.C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M.Y. Vardi, G. Weikum, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh, J.M. Zurada. Lecture Notes in Computer Science, vol. 7894 (Springer, Berlin/Heidelberg, 2013), pp. 235–246
5. M. Daniel, Conflict between belief functions: a new measure based on their non-conflicting parts, in *Belief Functions: Theory and Applications*, ed. by F. Cuzzolin. Lecture Notes in Computer Science, vol. 8764 (Springer, Cham, 2014), pp. 321–330
6. A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Stat. **38**, 325–339 (1967)
7. H. Dörksen, U. Mönks, V. Lohweg, Fast classification in industrial big data environments, in *19th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA 2014)* (2014), pp. 1–7

8. D. Dubois, H. Prade, *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Softcover reprint of the original 1st edn., 1988 edn. (Plenum Press, New York/London, 1988)
9. L. Dymova, P. Sevastjanov, K. Tkacz, T. Cheherava, A new measure of conflict and hybrid combination rules in the evidence theory, in *Artificial Intelligence and Soft Computing*, ed. by D. Hutchison, T. Kanade, J. Kittler, J.M. Kleinberg, A. Kobsa, F. Mattern, J.C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, D. Terzopoulos, D. Tygar, G. Weikum, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh, J.M. Zurada. Lecture Notes in Computer Science, vol. 8468 (Springer, Cham, 2014), pp. 411–422
10. J.F. Ehlenbröker, U. Mönks, V. Lohweg, Sensor defect detection in multisensor information fusion. *J. Sens. Sens. Syst.* **5**(2), 337–353 (2016)
11. A. Fritze, U. Mönks, V. Lohweg, A concept for self-configuration of adaptive sensor and information fusion systems, in *21st International Conference on Emerging Technologies & Factory Automation (ETFA 2016)*, Berlin (2016)
12. A. Fritze, U. Mönks, V. Lohweg, A support system for sensor and information fusion system design. *Procedia Technol.* **2016**(26), 580–587 (2016)
13. A. Fritze, U. Mönks, C.A. Holst, V. Lohweg, An approach to automated fusion system design and adaptation. *Sensors* **17**(3), 601 (2017)
14. M. Gebauer, Luftwaffe zweifelt an Absturzursache. *DER SPIEGEL* **2015**(22), 64 (2015)
15. D.L. Hall, J. Llinas (eds.), *Handbook of Multisensor Data Fusion*. The Electrical Engineering and Applied Signal Processing Series (CRC Press, Boca Raton, 2001)
16. J.Y. Halpern, *Reasoning about Uncertainty* (The MIT Press, Cambridge, 2005)
17. IEEE Computer Society, *IEEE Standard for Binary Floating-Point Arithmetic: IEEE Std 754™-1985* (IEEE, New York, 1985)
18. IEEE Computer Society, *IEEE Standard for Floating-Point Arithmetic: IEEE Std 754™-2008* (IEEE, Piscataway, 2008)
19. E.T. Jaynes, *Probability Theory: The Logic of Science*, 1. publ., repr. edn. (Cambridge University Press, Cambridge, 2003)
20. B. Khaleghi, A. Khamis, F.O. Karay, S.N. Razavi, Multisensor data fusion: a review of the state-of-the-art. *Inform. Fusion* **14**(1), 28–44 (2011)
21. Koenig & Bauer AG, Press images (2014). <http://www.kba.com/en/downloads-glossary/downloads/press-images/>. Accessed 22 Apr 2016
22. J. Kuri, Absturz des Airbus A400M: Doch Softwarefehler in der Triebwerksteuerung (2015). <http://heise.de/-2678691>. Accessed 25 July 2017
23. H.L. Larsen, Importance weighted OWA aggregation of multicriteria queries, in *18th International Conference of the North American Fuzzy Information Processing Society (NAFIPS 1999)* (1999), pp. 740–744. <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=781792>
24. R. Li, V. Lohweg, A novel data fusion approach using two-layer conflict solving, in *International Workshop on Cognitive Information Processing (CIP 2008)* (IEEE, 2008), pp. 132–136. <http://www.eurasip.org/Proceedings/Ext/CIP2008/papers/1569094849.pdf>
25. J.W. Li, Z.T. Hu, L. Zhou, Representation method of evidence conflict based on vector measure, in *Control Conference (CCC), 2014 33rd Chinese* (2014), pp. 7445–7449. <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6896238>
26. V. Lohweg, U. Mönks, Sensor fusion by two-layer conflict solving, in *2nd International Workshop on Cognitive Information Processing (CIP 2010)* (IEEE, 2010), pp. 370–375. <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5604094>
27. V. Lohweg, C. Diederichs, D. Müller, Algorithms for hardware-based pattern recognition. *EURASIP J. Appl. Signal Process.* **2004**(12), 1912–1920 (2004)
28. R.C. Luo, M.G. Kay, Data fusion and sensor integration: state-of-the-art 1990s, in *Data Fusion in Robotics and Machine Intelligence*, ed. by M.A. Abidi, R.C. Gonzalez (Academic, Boston, 1992), pp. 7–136
29. A. Martin, About conflict in the theory of belief functions, in *Belief Functions: Theory and Applications*, ed. by T. Denoeux, M.H. Masson. Advances in Intelligent and Soft Computing, vol. 164 (Springer, Berlin/Heidelberg, 2012), pp. 161–168

30. A. Martin, A.L. Jousselme, C. Osswald, Conflict measure for the discounting operation on belief functions, in *2008 11th International Conference on Information Fusion* (2008), pp. 1–8
31. C. Minor, K. Johnson, Reliable sources and uncertain decisions in multisensor systems, in *SPIE Sensing Technology + Applications*, SPIE Proceedings (SPIE, 2015), p. 949803
32. C.B. Moler, *Numerical Computing with Matlab* (Society for Industrial and Applied Mathematics, Philadelphia, 2004)
33. U. Mönks, *Information Fusion Under Consideration of Conflicting Input Signals* (Springer, Berlin/Heidelberg, 2017)
34. U. Mönks, V. Lohweg, Fast evidence-based information fusion, in *4th International Workshop on Cognitive Information Processing (CIP 2014)* (IEEE, 2014), pp. 1–6
35. U. Mönks, D. Petker, V. Lohweg, Fuzzy-pattern-classifier training with small data sets, in *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, ed. by E. Hüllermeier, R. Kruse, F. Hoffmann. Communications in Computer and Information Science, vol. 80 (Springer, Berlin/Heidelberg, 2010), pp. 426–435
36. U. Mönks, H. Trsek, L. Dürkop, V. Geneiß, V. Lohweg, Towards distributed intelligent sensor and information fusion. *Mechatronics* **34**, 63–71 (2015)
37. U. Mönks, H. Dörksen, V. Lohweg, M. Hübner, Information fusion of conflicting input data. *Sensors* **16**(11), 1798 (2016)
38. M. Niederhäuser, V. Lohweg, Application-based approach for automatic texture defect recognition on synthetic surfaces, in *IEEE International Conference on Emerging Technologies and Factory Automation, 2008. ETFA 2008* (2008), pp. 229–232. <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4638397>
39. A. Ross, A.K. Jain, Multimodal human recognition systems, in *Multi-sensor Image Fusion and Its Applications*, ed. by Z. Liu, R. Blum. Signal Processing and Communications, vol. 26 (CRC Press, Boca Raton, 2005), pp. 289–301
40. G. Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, 1976)
41. C.E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(4), 623–656 (1948)
42. F. Smarandache, D. Han, A. Martin, Comparative study of contradiction measures in the theory of belief functions, in *2012 15th International Conference on Information Fusion (FUSION) (2012)*, pp. 271–277. <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6289814>
43. The MathWorks, Inc., Floating-Point Numbers. http://www.mathworks.com/help/matlab/matlab_prog/floating-point-numbers.html. Accessed 23 July 2016
44. G. Traufetter, Gehirnschlag im Cockpit. *DER SPIEGEL* **2010**(8), 120–123 (2010)
45. G. Traufetter, Auf Absturz programmiert. *DER SPIEGEL* **2015**(13), 120–121 (2015)
46. G. Traufetter, Steigflug ins Verderben. *DER SPIEGEL* **2015**(2), 116 (2015)
47. K. Voith, S. Glock, U. Mönks, V. Lohweg, T. Türke, Multi-sensory machine diagnosis on security printing machines with two-layer conflict solving, in *SENSOR+TEST Conference 2011* (AMA Service GmbH, Wunstorf, 2011), pp. 686–691
48. R.R. Yager, On ordered weighted averaging aggregation operators in multicriteria decision-making. *IEEE Trans. Syst. Man Cybern.* **18**(1), 183–190 (1988)
49. L.A. Zadeh, Fuzzy sets. *Inform. Control* **8**(3), 338–353 (1965)
50. L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Syst.* **1**, 3–28 (1978)

Chapter 23

Quantify: An Information Fusion Model Based on Syntactic and Semantic Analysis and Quality Assessments to Enhance Situation Awareness



Leonardo Castro Botega, Allan Cesar Moreira de Oliveira, Valdir Amancio Pereira Junior, Jordan Ferreira Saran, Lucas Zanco Ladeira, Gustavo Marttos Cáceres Pereira, and Seiji Isotani

Abstract Situation awareness is a concept especially important in the area of criminal data analysis and refers to the level of consciousness that an individual or team has about a situation, in this case a criminal event. Being unaware of crime situations can cause decision-makers to fail, affecting resource allocation for crime mitigation and jeopardizing human safety and their patrimony. Data and information fusion present opportunities to enrich the knowledge about crime situations by integrating heterogeneous and synergistic data from different sources. However, the problem is complicated by poor quality of information, especially when humans are the main sources of data. Motivated by the challenges in analyzing complex crime data and by the limitations of the state of the art on critical situation assessment approaches, this chapter presents Quantify, a new information fusion model. Its main contribution is the use of the information quality management throughout syntactic and semantic fusion routines to parameterize and to guide the work of humans and systems. To validate the new features of the model, a case study with

L. C. Botega (✉) · V. A. Pereira Junior · G. M. Cáceres Pereira
Graduate School in Information Science, São Paulo State University (UNESP), Marília, Brazil
e-mail: botega@univem.edu.br

A. C. M. de Oliveira · J. F. Saran
Computer Science and Information Systems, University Centre Eurípides of Marília (UNIVEM),
Marília, Brazil
e-mail: allan_oliveira@univem.edu.br

L. Z. Ladeira
Institute of Computing, State University of Campinas (UNICAMP), Campinas, Brazil

S. Isotani
Institute of Mathematics and Computer Science, University of São Paulo (USP), São Carlos,
Brazil
e-mail: sisotani@icmc.usp.br

real crime data was conducted. Crime reports were submitted to the modules of the model and had situations depicted and represented by an Emergency Situation Assessment System. Results highlighted the limitations of using only lexical and syntactical variations to support data and information fusion and the demand and benefits provided by quality and semantic means to assess crime situations.

Keywords Criminal data and information fusion · Criminal information quality management · Crime situation awareness

23.1 Introduction

Situation awareness (SAW) is an important cognitive process of decision-makers in several critical areas. It concerns the perception of the presence and nature of the entities of interest in the environment, the understanding of their meaning and the importance of their individual and collective actions, and the projection of their status in the near future.

In the field of emergency management, SAW is a crucial factor for the success of operations involving humans. A limited SAW can compromise operators' understanding of what is happening and lead to a poor decision-making, which can result in disastrous consequences for people, property, or the environment.

Human operators that are aware of an emergency can not only characterize entities, events, and their relationships but also reveal trends and the existence of threats and infer the increase or decrease of imminent risks.

Although SAW cannot guarantee better decision quality, its improvement can help operators to maintain a superior knowledge of current events and situations. Operators of emergency services can be routinely subjected to information overload, especially because of the inherent need to perform multiple tasks. Supporting situation awareness of the operators is a challenge fundamental to the effectiveness of their activities.

Supporting SAW is even more challenging when data are provided by humans, as is the case when a report of a crime is offered by a victim or a witness. Typically, such data can be incomplete, outdated, inconsistent, and sometimes even irrelevant to the associated event. In addition, that data can also be influenced by human factors such as stress, fear, and cultural particularities. The presence of low-quality data and information influences the computational methods that use this human report as input to infer useful information to support operators in developing SAW.

To overcome this problem, information fusion (IF) processes have been designed to guide the development of systems. They comprise acquisition, inference, evaluation, and representation phases of high-level situational information. These systems typically use multiple heterogeneous data sources and computational intelligence to support environmental changes and help operators to develop SAW [11, 13, 14, 17, 18].

However, determining how a semiautomated situation analysis activity can be structured to better amplify operator's SAW is still a challenging issue for the fusion community, especially in the field of emergency management. In addition, there is still difficulty in dealing with quality problems inherent to human produced data [5, 6, 23, 25, 28].

Lack of knowledge about the quality of data and information propagated by fusion processes (such as information assessment process, inference process, information recovery process, or simply process) can also lead operators to uncertainties and errors in SAW, thereby degrading decision-making. The visual representation of situational information should consider data and information quality indexes to better support human observations and interpretation of a situation.

In this context, data fusion systems dedicated to supporting SAW can be augmented by information quality management processes, benefiting both the automated data assessment routines and the human understanding on crime situations. These assessment processes can be oriented and parameterized by quality indexed information. Quality indexes can also help operators by increasing their confidence in situational information and stimulate a proactive interaction with the system.

Furthermore, automated systems are great at processing a large amount of data but may fail on determining connections and meaning of data. Hence, quality assessment can support automated or machine-human processes to better complement each other, sharing objectives and contributing to the construction of situational knowledge. Thus, SAW can be better and more quickly acquired, maintained, and even reacquired [4, 6, 19]. Consequently, quality-aware information fusion systems (IFS) must present capabilities and processes to reveal, process, represent, and mitigate information limitations.

Literature presents data and information fusion models that explicitly describe the role of the human operator in semiautomatic approaches, typically originating from the JDL (Joint Directors of Laboratories), DFIG (Data Fusion Information Group), and User-Centered Information Fusion models. In these models humans are solely consumers of information or are active participants in managing and transforming information [3, 6]. These models are limited on presenting solutions for the problems associated with the human's deeper involvement in the process of transforming information to enrich SAW in IFS. More recent approaches present opportunities for human interaction throughout each level of fusion [1, 15, 19].

However, there are not many records of human-system collaboration supported by information quality in scenarios where time is a critical factor. In addition, known approaches are limited to providing refinements in a reactive fashion to the final product of the process [6].

The goal of this project is to present the Quality-aware Human-driven Information Fusion Model (Quantify) that aims to contribute to the improvement of the SAW of human operators. This model can be used by an emergency situational assessment system, dealing with scenarios that are complex, changeable, and dynamic, in which information is constantly evaluated and parametrized by several variables.

In addition, another objective of this chapter is to present details on how Quantify employs a combination of syntactic and semantic methods for a hierarchical and

multicriteria integration of information. This chapter also shows how this model deals with fusion of semantic information using information quality criteria. These advances aim to increase the inference capacity of complex information and to contribute to the SAW of analysts.

The pillars of the Quantify model are the processes for continuous assessment of data and information quality for the orientation of the IF process and the means for the semantic analysis of qualified situational information.

Specifically, through the Quantify model, this chapter aims to demonstrate:

- How the information quality management (inference, representation, and mitigation) can be beneficial in global and local contexts, at either low or high fusion levels, and contribute to SAW;
- How to ensure the propagation and usefulness of qualified information produced by humans, up to the highest levels of abstraction, useful for situation assessment (SA);
- How integrated syntactical and semantic analysis can contribute to empowering the inference capabilities of the fusion process;
- How to deal the information as linked knowledge about crime situations that gets enriched over time, produced by humans and machines, and its connections with the other steps of the fusion process.

To demonstrate in practice the use of Quantify, this chapter will end with a study case of real-time crime assessment.

23.2 The Incorporation of Data and Information Quality in the Fusion Process

Mapping complex entities, such as humans and their interactions with the real world, is a challenging process. The dynamic and complex nature of interactions between people, objects, and places demands the use of comprehensives computational techniques to reveal their states over time.

The process that fusion systems use to understand human interactions starts with the search and determination of which entities are present in a real scenario. Next, the states of the entities are determined, formed by their physical characteristics, position, orientation, and other data relevant to the domain. Finally, fusion systems establish the possible relationships among entities, relating each entity context and state to one another. These relationships may help humans and systems to understand situations.

Achieving SAW is a challenging process that SA systems seek to support. With the advent of SAW models, especially the Endsley model [12, 13], new IF models, architectures, and processes emerged aiming to support the development of SA systems.

In this context, the knowledge of the quality of information by the IFS emerges as a complementary resource to the process of inference of situations. Also, IFS act as technology to support the existing models and processes for the acquisition and maintenance of SAW.

The benefits of knowing the quality of information in an IF process at each of the possible levels of inference of the JDL model are the reliability of data sources and the effectiveness of data preparation algorithms (Level 0), the completeness and accuracy of the identification of objects (Level 1), the integrity of a relation between objects (Level 2), the assertiveness of a projection (Level 3), and the graphical representation of information (Level 5). Knowledge of information quality can not only support decisions at each level but also influence and parameterize internal inference routines.

In addition to contributing to the operationalization of the internal mechanisms of data fusion levels, knowledge about the quality of data and information can also contribute to the relationships between the levels of fusion, i.e., to help determine and direct information by virtue of desired outputs and inputs for each level. Quality indexes operate as a lever to determine the usefulness of information throughout the process from one level to another. This routine further contributes to fill the gap between low-level and high-level IF inferences [2, 4].

Among the challenges of incorporating data and information quality into a fusion process, we can highlight the role of information quality and the management of information dynamics for defining information quality. Regarding the first challenge, it is known that computerized processes of an IFS (e.g., mining, integration, and correlation) infer new information on a distributed, asynchronous, and dynamic fashion. To collaborate with each other, these processes must have a mechanism that qualifies each new data or information produced with a quality indicator (quality metadata). Thus, the parameterization of the process gains a new variable (different from attributes or objects) that must be considered every time a new fusion process is performed. This routine contributes to quality information reaching the upper levels of the process.

In addition, by helping to parameterize automation, the second challenge is to properly represent information quality and stimulate interactions of operators on specialized user interfaces and visualizations dedicated to SAW. To visually stimulate the perception of operators in the search for patterns and relationships, it is necessary to use cues or suggestions that qualify the information. These signals help to justify human behavior and explain why information is accepted or not and, consequently, can help guide operators to improve the quality of information through a continuous refinement process.

23.3 Quality-Aware Human-Driven Information Fusion Model

To overcome some of the information quality assessment challenges, this paper proposes a fusion model called Quantify (Quality-aware Human-driven Information Fusion Model) (Fig. 23.1) [9]. The major differentials of this model are the combination of syntactic and semantic approaches to assess data and the use of information quality management throughout the fusion process.

Quantify consists of six internal processes: data acquisition, data and information quality assessment, object assessment, situation assessment, information representation, and user interfaces (UI).

This model has the goal to orient the development of IF systems, dedicated to supporting the assessment of situations that occur in complex real-time scenarios, especially when it is hard to acquire reliable information. These complex scenarios comprise highly complex entities that interact and relate to each other to form situations, which evolve in time and space [3, 6].

Among the main features of the model are mechanisms designed to:

- Manage data and information quality (infer, represent, and mitigate) in local and global contexts of the IF process at low and high levels of abstraction;
- Support operators in improving their perception and understanding of the situation and in orienting and refining information;
- Parameterize automated processes of the IF routine using qualified situational information in syntactic and semantic fusion processes.

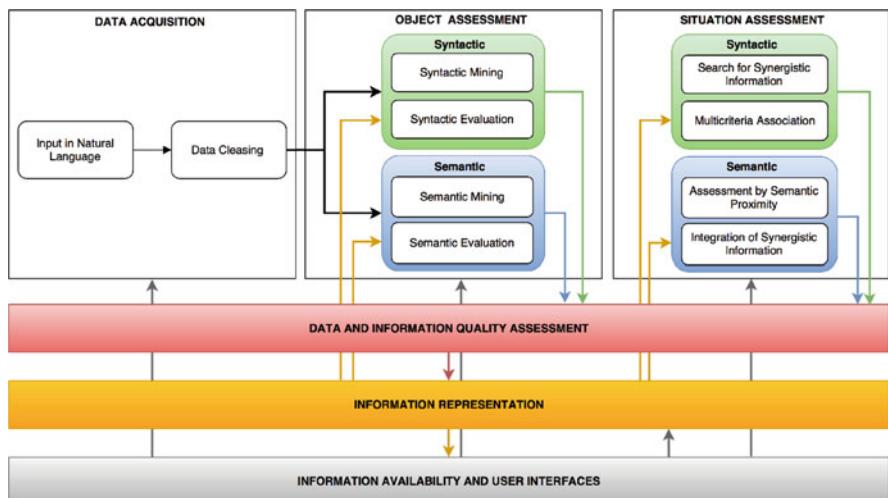


Fig. 23.1 Quantify model with syntactic and semantic information fusion processes in details

This model also comprises:

- A complete process of SA over complex real-time scenarios, with internal processes of acquisition, processing, representation, and refinement of situations that have human-provided data produced by heterogeneous sources;
- A cyclic, iterative, and interactive operation, which allows operators to accompany changes in the situation;
- A set of mechanisms to manage data and information quality to assess each new information inferred, to enrich situational representation, to parameterize processes, and to orient operator in the refinement task;
- A semantic fusion approach that uses semantic models (ontologies) with associated data quality to improve the finding of synergistic information in human provided data;

In the next sections, we will describe the Quantify model, as well as its internal processes in detail.

23.3.1 Data Acquisition

In complex scenarios, there are multiple types of data available, such as audios, text messages from social networks, records from historical databases, camera images, and information from diverse subsystems. Each application has sources and input data that may be used to perform the assessment of a situation.

Therefore, the internal process of human-provided data acquisition is responsible for collecting information generated by humans and making it available for the use of other internal mechanisms of Quantify. The result of this process is the identification and classification of objects, attributes, and preliminary situations, according to an application domain. To achieve this objective, this process is structured in three stages, namely, (1) *obtain sentences*, (2) *grammatical analysis of sentences*, and (3) *search and identification of relevant objects*. To perform the stage *obtain sentences*, natural language processing (NLP) techniques are used to transcribe the audio and to format it in a string structure. This step can be accomplished with a tool like the one provided by Google [10, 26].

In the emergency management domain, data from social networks, like Twitter, can also be used through its public API. Posts that report a situation are searched based on the objects previously identified by NLP. Once data has been captured, transcribed and stored in a structured way, it can then be sent for a sentence analysis, which is performed to identify patterns and logical sequences of characters and words [21, 27].

At the *grammatical analysis of sentences* step, the input text must be analyzed in real time by a grammar checking tool, such as CoGrOO [16, 26]. Thus, it makes it possible to add labels such as nouns, number, object, or any other classification. It is also possible to connect the sentences obtained in the input text.

Each object found is classified, along with its attributes, by using keywords. These keywords have already been defined through the analysis of several sentences and also as a product of systems' requirements [20, 23].

The *search and identification of relevant objects* processes elements defined as important in fulfilling the requirements. During the process of defining these requirements, meaningful words are defined in an account (report), generating lists of words classified in different categories, such as tagCor (tag for color) and tagTypePhysical (tag for physical type) [16, 24].

In this way whenever a word from any of these categories is found, new analyses are performed of the following words of the input text data, seeking additional meanings such as status, situation, and even quality of objects, people, or situations. While analyzing the result of the classification of a word, it is possible to infer what type of information it represents, such as addresses, names, etc. To determine the possibility of a next word, several block words are analyzed and compared to a glossary.

At the end, the identified objects and the first situations are encapsulated in an object model (e.g., JavaScript Object Notation – JSON) and submitted to the next internal process to assess the information according to quality dimensions and metrics.

23.3.2 Data and Information Quality Assessment

The data and information quality assessment internal process aims at qualifying the situational information by quantifying quality dimensions for the guidance and parameterization of the fusion process as a whole, so that other processes can use the qualified information [8].

The quality assessment is applied to raw data and also to the situational information after it has been formed and represented as linked relations (situations formed by pieces of information).

The dimensions assumed to perform the assessment of human provided crime data are:

- Timeliness (considering how fresh the data is);
- Completeness (the percentage of attributes and objects a situation has);
- Temporal completeness (how complete is time-referring data);
- Consistency (the alignment of new processed data with situational information);
- Relevance (the new data is useful to the current situational data);
- Syntactic precision (the data are within an acceptable threshold of syntactic variation);
- (Un)certainty (trust of the system in the information)

This process also relies on the “Methodology for Data and Information Quality Assessment in the Context of Emergency Situational Awareness” (IQESA), described in Botega et al. to assess and evaluate data and information quality [8].

The IQESA methodology is performed in three steps: elicitation of information quality requirements, in which quality requirements are determined for a specific domain by interviews and a goal-driven task analysis (GDTA) [9, 14, 24]; definition and use of functions and metrics to assess quality dimensions, in which metrics are defined and applied to infer quality indexes; and finally the representation of situational information, in which data and the associated quality are represented in the form of semantic models (ontologies).

23.3.3 Object and Situation Assessment: Information Fusion Using Information Quality

23.3.3.1 Syntactic Information Fusion

The process of syntactic information fusion can be abstracted in two main stages, each with its internal mechanisms that play specific roles in improving the representativeness of information. The stages are *search for synergistic information* and the *multicriteria association*.

After acquisition and quality assessment, a new object is produced. This resulting object corresponds to what we know as L1 data fusion results, comprised by objects and attributes found with quality indexes attributed to each object [22, 23].

The decoded object and its attributes feed the creation of a preliminary ontology (Fig. 23.2), which represents an initial situation with its current classes and object instances (semantic data associated with ontology classes). This situation has people, objects, and places, each with their respective attributes, with indications of common activities between them (relationship properties), defined at acquisition time.

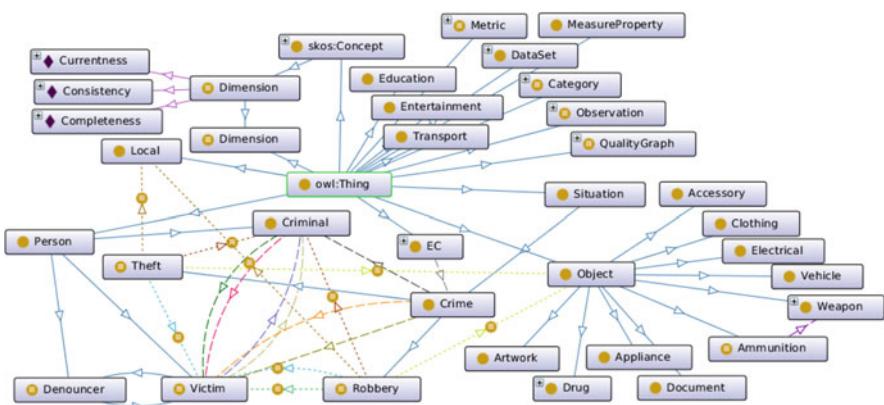


Fig. 23.2 Ontology of a crime emergency situation

The internal process of “*information representation*” is responsible for that task and directly connected to all other processes that use its output and provide information to it. They provide/use information asynchronously, distributed and on a dynamic fashion. The resulting data is also transported to/from the situational knowledge frequently.

This instantiated ontology composes the input for the IF phase. Among the input parameters are the objects identified in the previous phases; the type of data source, which properties must be present; and even a quality threshold of information.

Once the fusion process is started, *search for synergistic information* is performed between classes that are already present in the current ontology and that can hold information of objects, attributes, properties, and quality indexes that have some kind of correspondence.

After a search in information already represented in the ontology, a new search is made for more information that has not yet been considered in the process, from the same source or from other data sources, and which have already been submitted to the “*data acquisition*”. This routine is designed to obtain new information about the associated objects at any given time, validating and giving greater consistency to the already defined information. The entry to this process is either an isolated information (Level 1 JDL) or an relationship between objects (Level 2 JDL), and the result of this step may be a new object or a new situation.

This process can be implemented by data mining techniques, for example, by the Apriori algorithm [7, 9, 16], which infers the frequency of the presence of certain information when analyzed in relation to the rest of the available input data (from current or other sources). This inference is made by considering an information support formula (covariance).

The results of this process are new objects, attributes, quality indexes, and properties that may complement the current information held by the ontology.

The next stage, the *multicriteria association*, computes the synergistic information using predefined criteria for quality indexes and semantic properties. The results are insertions of new information into the ontology, satisfying the similarity found in the context of the original information and satisfying the multicriteria process [9, 24].

Two criteria are currently suggested, one is the input from the operator during execution time and the other based on the information and knowledge obtained through the analysis of requirements by developers, before the operation of the system, and that automatically affects the algorithms of this process. As a result of the automated part, all the initial information that is submitted to the fusion process is analyzed for synergy (have their syntactic or semantic similarities checked). The results are the discovery of new attributes, properties, and even new objects, in a combined and hierarchical way, resulting in new situational information. This result can be resubmitted to the previous synergistic data search process, increasing the process’s ability to find new information and further consolidate the information already found.

Hence, information is increasingly specialized, enriching the current situation with more details and qualified data. The syntactic fusion is cyclical and is

performed until the result of the multicriteria association does not reach the requirements previously defined, i.e., until the quality levels are sufficient for the decision-maker.

The resulting information is submitted to the information quality evaluation process, now punctuating the new information found, while also focusing on the indexes of the current situation. After this process, the information reevaluated will be reinstated in the ontology.

In the automated fusion process performed after the acquisition and assessment of information quality, the greatest possible number of associations is made between the objects, their attributes, properties, and quality indexes. For such, it is considered the existence of two or more data sets available from the same source or from different sources.

This process, called the primary fusion, employs primarily criteria for automated fusion. These criteria can include either a minimum level of quality or priority of object properties, which is useful to define what should be processed and shown to the operator first. These priorities are defined by the information requirements gathered through questionnaires filled by several specialists of various functions and career time.

In the case of the on-demand fusion by a human operator, the algorithm is activated once again, but the integration options are selected entirely by the operator through the user interface, rather than considering all possible combinations of objects, attributes, and properties identified in step of acquisition. This process of association, now manual, though based on objects and attributes, is strongly supported by indexes of quality and hypotheses employing information related to previously classified objects that were obtained in past cycles or different data sources.

Since this process is performed by the operator utilizing a user interface, the criterion for the data fusion process (e.g., quality indexes or an object characteristic or even a physical property) can be chosen and changed by the same operator, as well as the removal of predefined criteria by the requirement analysis. This capability provides for the flexibility of the structure to receive and process different criteria for a given situation, as well as allows operators to interact with the system based on their experiences and knowledge.

23.3.3.2 Semantic Information Fusion

Based on the identified objects and attributes inferred by the implementation of the model described in the previous section, a preliminary ontology is instantiated. For example, in the crime domain, this ontology classes represent victims, criminals, stolen things, information quality, and location, each with their respective attributes and relations. The ontology also reveals the existence of semantic properties of the information (meanings), as shown in the example of Fig. 23.2 [24].

This part of Quantify is also responsible for generating input to the information fusion process by considering the semantic aspects, that is, the meaning of certain

information according to a context. This technique greatly enhances the power of SA, because instead of analyzing the structure of a word, it seeks to analyze its semantic connections (whether there are common contexts that would have some possibility of building a situation based on such linked information).

Let's consider two situations transcribed in different ways. Even if the meaning of their objects is the same, they may not have been considered synergistic in the process of syntactic fusion, for example, considering the situation "man flew" and, in another situation, "a guy ran." They are completely different information from the syntactic point of view, but in a semantic context that considers meaning, this sentence has points of similarity.

The semantic fusion process is performed by an algorithm based on data mining techniques, using the same Apriori technique employed on the syntactic analysis. The result of this semantic search procedure is the same as the ones of syntactic process but grouped into collections with a degree of similarity based on their meaning. In this process, there is a possibility to find terms that do not have associated meaning. Hence, they cannot be assigned to an existing ontology class. In the next process, these collections will be integrated into a new situation in order to contribute to improve quality scores and the real meaning of this information.

Each new information is submitted to the quality assessment layer where quality scores are assigned to it. Further, using expected threshold values, it is decided whether the routine continues or whether the new information should be disregarded.

Finally, information generated in both semantic and syntactic processes are compared based on the quality indexes. Better ranked information is then sent one last time to the quality assessment layer, and if the updated values are not sufficient for the criteria chosen by the expert, the complete process of SA can be redone. After that, they are displayed to the experts to ensure that all processing possibilities have been done.

This option of forwarding the information to the syntactic and semantic assessment layers can be automated or triggered by the human operator, which can start a fusion event via the user interface and demand that each situation to be processed two or more times.

The result selected after several evaluation cycles is sent to the information representation layer, which will be instantiated and considered as the most current situation, as explained in the next section.

23.3.4 SAW-Driven User Interfaces

The user interface (UI) aims at specifying a sequential activity routine to fully manage the situational information, generated, propagated, and maintained by the Quantify model [9, 20].

To this purpose, it includes a user interface in the Quantify modeling dynamics, not only to represent situational knowledge but also to promote the two-way

relationship with the other stages of the process. In addition, it also seeks to specify the process management routines as a whole to promote the necessary refinements to the SAW process.

As previously discussed, situational knowledge was constructed by the “*data acquisition*” phase, qualified by the “*data and information quality assessment*” phase under the requirements pertinent to the domain of complex environments, and enriched by “*object and situation assessment*” and is now visually represented in this phase.

Additionally, the UI also receives from the previous process the metadata that qualifies the situations, enables the use of information visualization techniques, and graphically represents this accumulated and qualified knowledge.

The development of the SAW-oriented interface must follow the design principles of Endsley [12], which are organize information according to objectives, present level 2 of SAW directly, support global SAW and information filtering, support local and global trust verification, represent historical events to follow information evolution, and support uncertainty and quality management.

23.4 Case Study

23.4.1 *Situational Awareness and the Problem of Crime Analysis in Brazil*

Critical SAW-oriented systems, such as risk management and risk assessment systems, require specialized intelligence to provide operators a dynamic understanding of what is going on or what has happened in an environment. In Brazil, criminal record databases, based on unstructured human provided data, have problems related to the quality of information, mainly the reliability of registered addresses as places of crime. These problems are due to the imprecision of the information obtained from the victims, and the lack of prioritization of this data by collectors (civilians or military police), which focuses more on the description of the event than on location data.

In addition, most electronic system recording events allow the completion of the record even without the address of the fact or only with a reference point (e.g., a restaurant, a store, or a public place). This aspect is particularly important for the decision-making process, since the criminal mapping as a data analysis tool for defining public policies has become popular in Brazil. However, by ignoring the poor quality of location information and the absence of data processing routines, criminal information systems are often based on georeferenced maps and do not reflect the actual information about the crime incidents.

In a complex decision-making environment, commanders need a clear, concise, and accurate assessment of the situation and whether there is any risk to people's lives, patrimony, or environment [28]. To support the production of information

useful for supporting and maintaining SAW for solution of problems dealing with criminal behavior and their diverse environmental contexts, the techniques of filtering, mining, and data integration, elements present in IF processes, are critical.

An appropriate SAW-oriented fusion synergistically integrates information into the current situational picture, performs an analysis of the input information, and commits to providing information according to the needs and expectations of the expert.

Supported by semantic fusion and the quality of information, opportunities for improving the parameters (or criteria) for data fusion have become imminent, enhancing the possibilities for effective contributions to the process of SA [2, 5, 6, 17, 28].

23.4.2 Risk Analysis Through Syntactic and Semantic Perspectives

The case study is based on the situation assessment supporting situation awareness of a real crime situation, more specifically a crime of robbery, reported to the police emergency response service. The reported information was submitted to each step of Quantify and its algorithms. Results are analyzed based on the acquisition, mining, fusion, and representation of relevant information useful for decision-making. The IF consists of two processes of data assessment: syntactic and semantic. Complementing them, there are other processes crucial for implementing IF, such as object identification, information representation, and quality assessment.

SA starts with the identification of entities and objects present in the reports, which is based on NLP techniques, using a rich vocabulary of the language, focused on the emergency management domain, more specifically the analysis of real-time crime data. After the identification of entities and objects, the information, communicated through JSON objects, starts being instantiated in small ontologies (Figs. 23.6, 23.7 and 23.8) that will later compose an entire situation. The information representation module is used here to support the IF processes during execution, by maintaining the most current version of situational information.

Each result is also submitted to the module of quality assessment, which at this time only evaluates the completeness of the identified entities and objects. Data will return to this module each time new inferences or any kind of information transformation is made.

Hence, the execution flow follows the cycle: object identification, representation, quality assessment, and back to the representation module. The following reports were submitted to Quantify, and the results are presented and discussed below, highlighting objects of interest.

Crime Report 1 The *victim* stopped the *vehicle* at a semaphore. Then, he was surprised by *two individuals* in a *black motorcycle*. The *motorcycle passenger* hit the glass of the car *with a gun*. *Threatened the victim and subtracted his vehicle*.

The criminals escaped with the destination location ignored. *Victim* does not have conditions to describe the authors in detail.

Crime Report 2 *Two guys on a dark motorcycle stole a car at the semaphore at the intersection between Mooca and Taquari Streets, pointed a revolver, and took the woman's car. One of them was in a red coat.*

Crime Report 3 *Two men robbed a gray car in the street from Mooca to the side of Santander Bank. The men were on a black motorcycle, one of them was in a blue jacket and jeans. They left in the direction of Hospital Villa-Lobos.*

23.4.3 Syntactic Analysis

The syntactic fusion process starts by searching for synergistic information in data sources and previously processed situations, looking for a similarity in their syntax. For instance, the contents of the reports are analyzed by making Boolean comparisons if one word is equal to another, considering some word variations such as tension, gender, and radical.

After the synergistic search among the reports, those that present a sufficient level of synergy are grouped in data sets. Analyzing the three reports used in this case, we can note only some terms that satisfy this syntactic fusion condition, such as “two,” “black,” “Mooca,” and “motorcycle.” However, these terms, even if they are covariant, do not express any rich or explicit meaning enough to carry out a fusion of information between the reports.

Finally, these terms go through the multicriteria association, which considers not only synergy but also quality index assigned to data and inferred information. This activity associates new collected information to the current situation, so that the final product is a single situation, as complete and detailed as possible.

The result is sent again to the quality assessment layer to be updated. The inferred situations that did not meet a predefined quality level are temporarily stored to be compared with the results of the semantic fusion, which occurs in parallel. Higher-quality information is permanently aggregated to the current situation.

The result of the syntactic fusion process for the three reports is shown in Fig. 23.3. The main downfalls of the syntactic fusion are the lack of properties among the identified terms (activities that imply a relation between objects, e.g., “wallet belongs to victim” or “criminal ran to subway”) and the capacity of recognizing multiple similar objects of the same class with the same meaning (e.g., victim's car and criminal's car).

Using only this process could lead to major failures in the final result, directly affecting quality indexes such as completeness and consistency. These failures in turn may affect the process of building SAW.

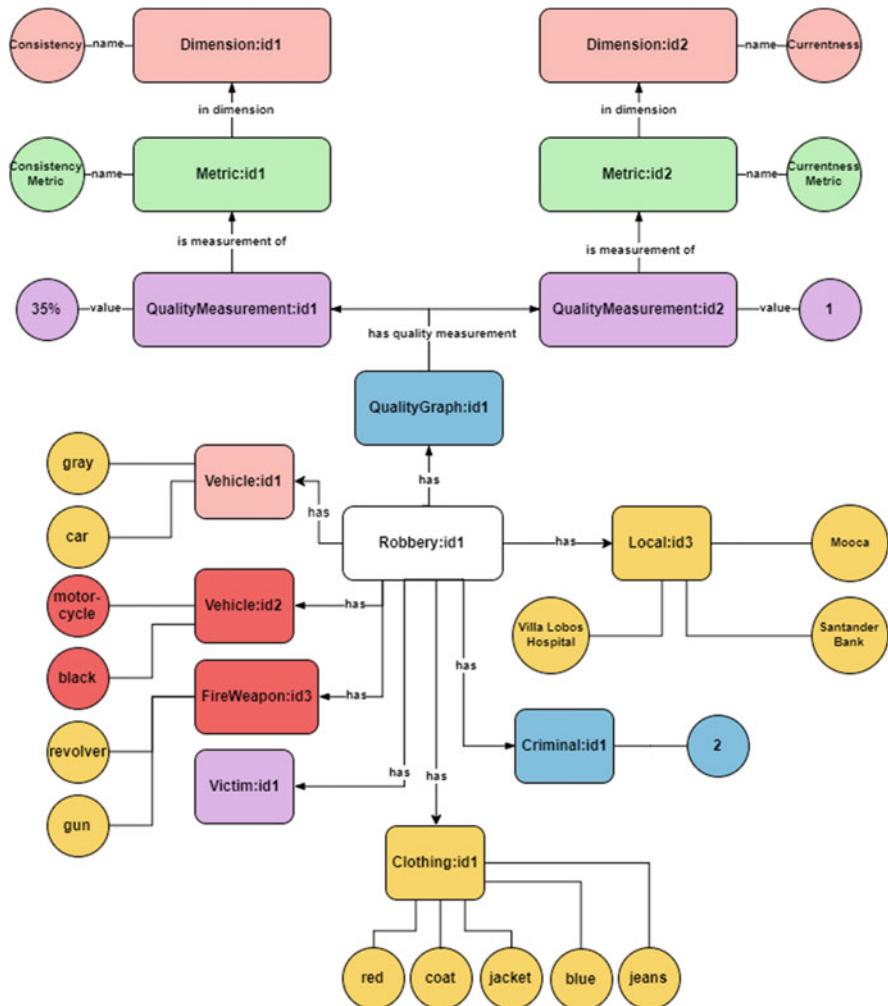


Fig. 23.3 Situation ontology instantiated with the result of the syntactic analysis of the reports

23.4.4 Semantic Analysis

At the beginning of the semantic process, information is not analyzed directly, i.e., by the way the terms are written, as is in the syntactic analysis. At first, each report is structured in ontological instances, based on an ontology developed for this domain. Each of the reports described above will be a different instance, depending on the situation it represents.

This process of semantic instantiation starts from the identification of elements that are stored in the information representation layer. Then, for each report,

SPARQL queries are made in the domain ontology for each previously identified element. These queries seek to identify which class or set of classes of the ontology best represents the elements and the probable properties that characterize them, even if they are not explicit in the report. The result of this process is an instance of a situation, which does not necessarily represent a crime.

At this point, we have several instances belonging to the same events without relationships between them. The next step is to perform SPARQL queries, now considering a local context, i.e., inside each class, to identify common properties of instances of elements, to make it possible to infer new objects, properties, and attributes, in order to define a situation for each report. These instances are shown in Figs. 23.4, 23.5, and 23.6.

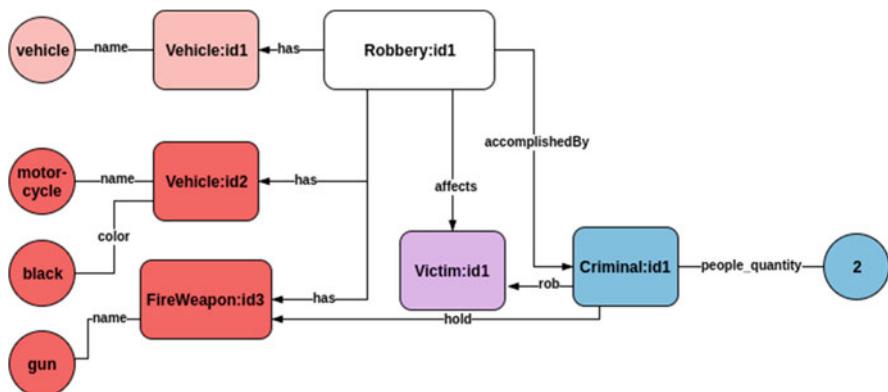


Fig. 23.4 Situation ontology instantiated with the information from crime report 1

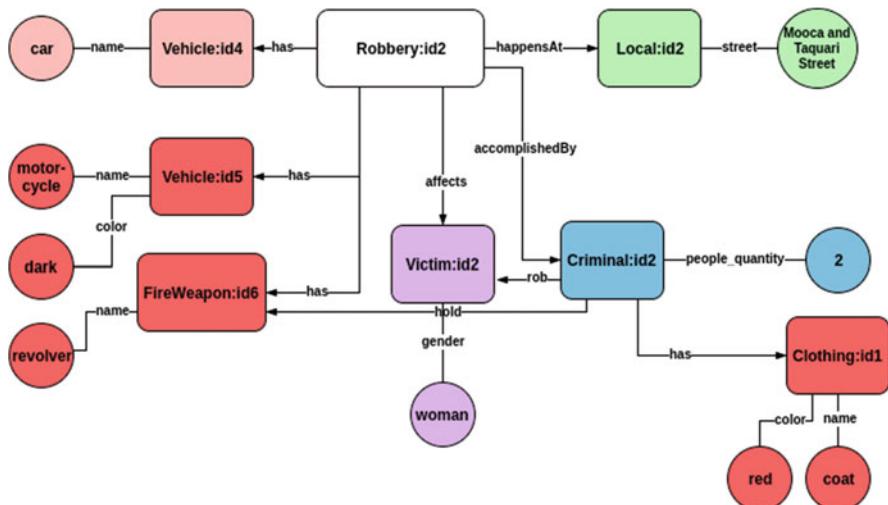


Fig. 23.5 Situation ontology instantiated with the information from crime report 2

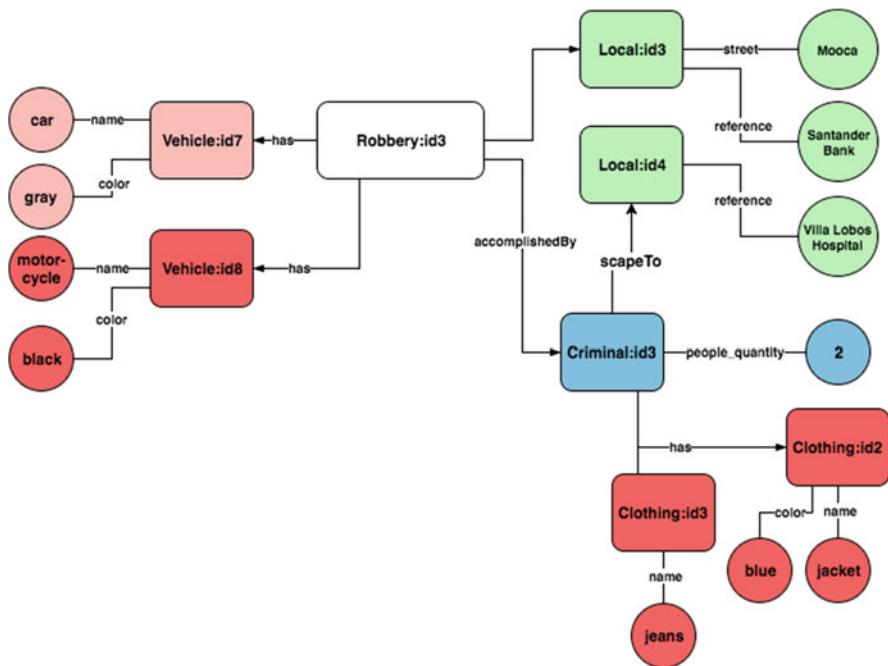


Fig. 23.6 Situation ontology instantiated with the information from crime report 3

In these figures, the rectangles represent the instances, and the circles represent the attributes. The arrows and links represent the connections and their properties between classes and attributes, defined according to a vocabulary that was also developed for the domain and shown in the following figures.

The instances are assigned according to the classes they represent to facilitate the understanding of the situation. Pink represents objects related to the victim; red represents objects related to the criminal; green are site-related; purple is the victim; blue are criminals; and white represents the instance of the situation.

The semantic processing starts with the analysis of the elements found in Crime Report 1, being *victim*, *vehicle*, *two individuals*, *black motorcycle*, *motorcycle passenger*, *with a gun*, *threatened*, *subtracted*, and *his vehicle*.

Based on these terms, queries are made to the situational ontology, which will return which class they fit into, or if they are just properties. In this case, a set of classes is returned with victim, criminal, and object and some properties, such as threatened and subtracted, which may represent a theft.

With SPARQL queries to a rich ontology, it is also possible to make associations between distinct terms. For instance, Report 1 does not have the term “criminal,” and yet a criminal class was identified, because in the ontology, one of the instantiated terms that characterizes a criminal and is present in Report 1 is “individuals.”

In another example, we perceived a “hold” property of a criminal for a firearm, indicating that the criminals have a gun, even though there is no term “held” in the report. This is because the term “with” is on the report, and through the ontology, it is possible to establish a covariance between them. At the end of the semantic analysis, the entities are correlated according to properties, dependencies, or relationships found in the reports and predicted in the vocabulary. The result (Fig. 23.4) for Crime Report 1 is also persisted in the *information representation* layer.

This routine of instantiating the ontology with inferred objects, attributes, and properties regarding a single situation occurs for all reports submitted to the semantic process. Figures 23.5 and 23.6 shows the results of this process for Reports 2 and 3. As it is possible to observe in Figs. 23.1, 23.2, and 23.3, the three reports seem to refer to the same situation, and fragments of it were described by different people present in the situation.

This relationship between the situations found in the reports is easily understood by a human, who deduces the situations and interprets them into a situational knowledge (what is known about a situation). However, in a more complex scenario, humans become error-prone and cannot absorb all the characteristics of a situation.

At this point semantic fusion combines new reports with instances of RDFs (previously processed information stored in the information representation layer). The goal is to build a computational model very close to a human mental model, through ontologies and vocabularies. This process uses the information in the information representation layer, after the process of semantic identification.

Semantic fusion is very similar to the semantic identification, but at a higher level of significance, once its inferences are made using instances from all reports. Also, semantic fusion considers all the properties presented in each internal instance of the elements.

At the end of the semantic fusion, we will have a new set of possible situations varying in the organization and presence of the elements and properties. Each of these possible situations is saved in the representation layer and has its quality assessed.

Then the multicriteria assessment is performed, mainly evaluating the improvements in the quality indexes. The situation that presents the best quality indexes and that satisfies criteria elicited in the requirements, like the presence of some specific element, will be elected as the final situation resulting from the fusion. Again, this situation is persisted by the ontology in the *representation* layer and later presented at the interface.

The semantic fusion of the analyzed reports allows to obtain an information with greater added value by the junction of the terms found. In this case study with the three reports, the fused information allows the identification of the clothes of both criminals, the characteristics of the stolen object, the weapon used by the criminals and vehicle of escape, as well as the place where the crime occurred. The examples show that the crime reports alone do not present these situations explicitly.

The semantic fusion result is shown in Fig. 23.7, which uses the same colors as before, with the difference that the yellow color represents the information that was fused from the three reports. In addition to the presence of more instances and

attributes in the fused situation, it is also possible to note new properties, which were not explicit when considering each report separately.

Another point to be noted is that the quality of the information can be assessed in the ontology to allow the verification of what information has a better quality and thus to decide what should be used in the fusion or not. For this assessment we used the Data Quality Vocabulary (DQV) ontology that is built on top of another quality ontology Dataset Quality Vocabulary (DAQ). DQV allows the creation of instances of categories, dimensions, and metrics to map quality measurements.

For each instance of the ontology, an instance of quality graph is created, making the connection between all quality measures applied to that instantiated information. Figure 23.7 shows the result of the semantic fusion with quality assessment, showing the dimensions of consistency and currentness for the theft situation assessed.

To graphically represent the results of this study, a system called Emergency Situation Assessment System (ESAS) (Fig. 23.8) was developed, guided by the Quantify model. This system has the capabilities of dealing with human-generated input and inferring what is/was going on by processing the natural language, which is useful to real-time (emergency) or risk analysis (historical data).

Figure 23.8 also shows the UI of ESAS, which has in the top right corner the “Event Table,” where human-generated input is reproduced, highlighting the transformations over the raw data, e.g., the discovery of a new relevant entity such as a criminal and its characteristics. In the bottom right corner of ESAS, there is the “Map of Reports.” The display shows the data sources placed on their origin location (where the reports came from). The raw data can be extracted from sources by user interaction with the placed pins. This map can also be populated by data from social network (e.g., Twitter posts).

In the left side, there is the “Situation Graph.” This display contains a hierarchical structure that represents the current situation picture, i.e., what is going on, with the central node being the situation itself, the next level the classes that composes the situation, and the leaves the instances of each classes that specifies the event. The color of the nodes represents what the information quality is, ranging from solid red when the quality is low to solid green when the quality is high.

23.5 Conclusions

This chapter presented a new IF model named Quantify and highlighted how it deals with situational information by syntactic and semantic processes. In general, Quantify aims to improve SAW for human operators that assess situations to make decision in complex scenarios, such as in risk and emergency management.

This work also demonstrated the way of incorporating new objects and a situation assessment cycle by utilizing the semantic fusion routines into the IF process. This approach can be inserted in situation evaluation routines, so that decision-makers may reason about information quality dimensions and seek better

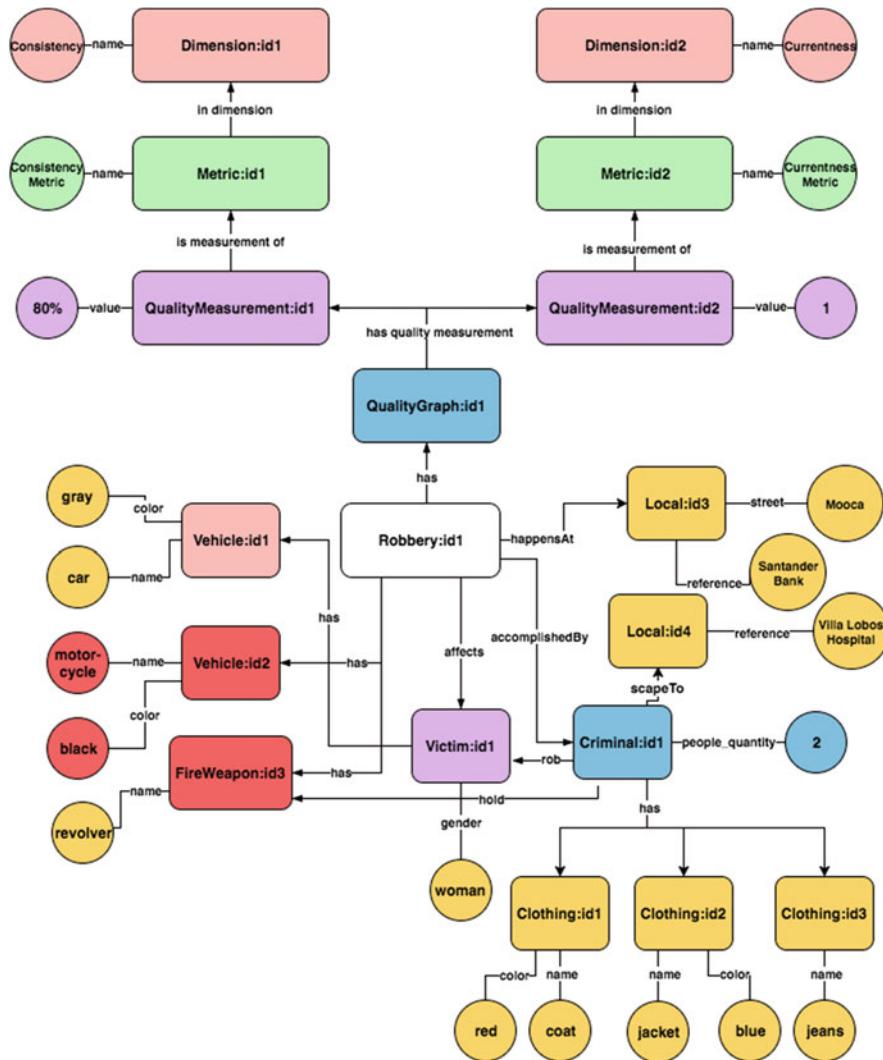


Fig. 23.7 Situation ontology instantiated after the semantic fusion, with the information from all reports associated with a quality index

quality. Moreover, this work also presented methods for fusing information by using multiple hierarchical and representation criteria of knowledge.

The use of quality indexes may contribute in the future to the fusion of data from physical sensors with human-generated data. The Quantify and its IF process, with the associated methods, were validated by the results of the acquisition of useful information for supporting SAW, according to the requirements set by domain

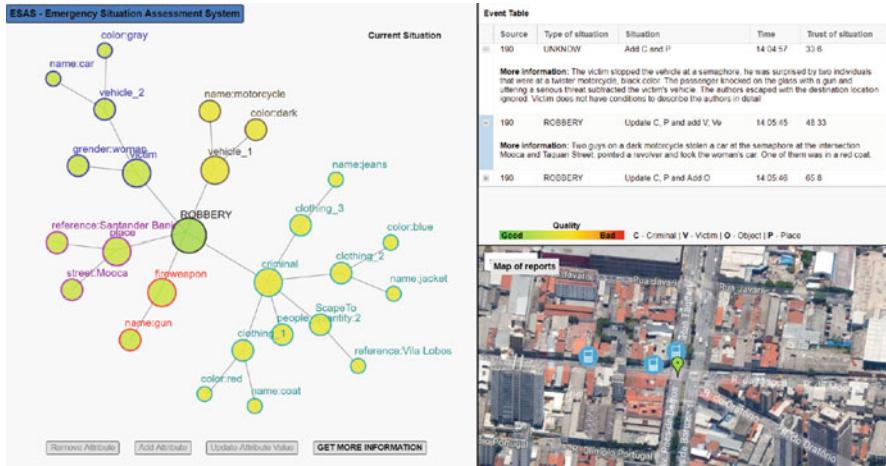


Fig. 23.8 Emergency Situation Assessment System (ESAS)

experts. This work also showed that data and information quality can act as a method for integrating heterogeneous data.

The information required to develop SAW was successfully built incrementally using syntactic and semantic inputs. The use of multicriteria information fusion enabled the assessment of situations by generating various possibilities for integration of synergistic information for the analysis of a specialist.

The continuous assessment of data and its quality proved that, at each evolution in the situation, improved and updated information were available for fusion and graphic representation, even if recently acquired and inferred. These assessment routines also proved the capabilities of Quantify in processing human feedbacks and supporting their interactions with the automation. Hence, with the establishment of new connections on situational information, by semantics and quality assessments, the authors state that the awareness of decision-makers on critical situations can be improved and their uncertainty mitigated.

References

1. E. Blasch, Level 5: user refinement issues supporting information fusion management, in *2006 9th International Conference on Information Fusion*, Florence, vol. 5, July 2006, pp. 1–8
2. E. Blasch, High level information fusion (HLIF): survey of models, issues, and grand challenges, in *IEEE A&E Systems Magazine* (2012), pp. 4–20
3. E. Blasch, S. Plano, DFIG level 5 user refinement issues supporting situational assessment reasoning, in *2005 7th International Conference on Information Fusion*, Philadelphia, vol. 5 (2005), pp. xxxv–xlivi

4. E. Blasch, P. Valin, A.L. Jousselme, D. Lambert, É. Bossé, 2 Top ten trends in high-level information fusion, in *Proceedings of the 15th International Conference on Information Fusion (FUSION)*, Singapore (2012), pp. 2323–2330
5. E. Blasch, J. Schubert, K.B. Laskey, G.W. Ng, R. Nagi, P.C.G. Costa, D. Stampouli, J. Schubert, P. Valin, Issues of uncertainty analysis in high-level information fusion, in *Proceedings of the 15th International Conference on Information Fusion (FUSION)*, Singapore (2012), p. 13
6. E. Blasch, A. Steinberg, S. Das, J. Llinas, C. Chong, O. Kessler, E. Waltz, F. White, Revisiting the JDL model for information exploitation, in *16th International Conference on Information Fusion*, Istanbul (2013), pp. 129–136
7. L.C. Botega, C. Berti, R.B. Araújo, V.P.A. Neris, A model to promote interaction between humans and data fusion intelligence to enhance situational awareness, in *Human-Computer Interaction. Theories, Methods, and Tools. HCI 2014*, ed. by M. Kurosu. Lecture Notes in Computer Science, vol. **8510** (Springer Cham, 2014), https://doi.org/10.1007/978-3-319-07233-3_37
8. L.C. Botega, J.O. Souza, F.R. Jorge, C.S. Coneglian, M.R. de Campos, V.P.A. Neris, R.B. Araújo, Methodology for data and information quality assessment in the context of emergency situational awareness. *Univ. Access Inf. Soc.* **16**(4), 889–902 (2016)
9. L.C. Botega, V.A.P. Junior, A.C.M. Oliveira, R.B. Araujo, J.F. Saran, L.A. Villas, Quality-aware human-driven information fusion model, in *International Conference on Information Fusion*, Xian (IEEE Computer Society, 2017), pp. 1–10
10. C. Chelba, D. Bikell, M. Shugrina, P. Nguyen, S. Kumar, Large scale language modeling in automatic speech recognition (2012). *CoRR*, abs/1210.8
11. M.R. Endsley, The challenge of the information age, in *Proceedings of the Second International Workshop on Symbiosis of Humans, Artifacts and Environment*, Kyoto (2001)
12. M.R. Endsley, What is situation awareness? in *Designing for Situation Awareness: An Approach to User-Centered Design*, 2nd edn. (CRC Press, Boca Raton, 2011), pp. 13–30
13. M.R. Endsley, Final reflections: situation awareness models and measures. *J. Cogn. Eng. Decis. Mak.* **9**, 101–111 (2015)
14. M.R. Endsley, D.G. Jones, *Designing for Situation Awareness: An Approach to User-Centered Design*, 2nd edn. (Taylor & Francis, London/New York, 2012)
15. D.L. Hall, J. Llinas, M.D. McNeese, Modeling and mapping of human source data. Technical report, College of Information Sciences and Technology, The Pennsylvania State University (2011)
16. V.A.P. Junior, M.F. Sanches, L.C. Botega, C.S. Coneglian, Towards semantic fusion using information quality awareness to support emergency situation assessment, in *Advances in Intelligent Systems and Computing*, vol. 444 (Springer, Cham, 2016), pp. 145–155
17. M. Kokar, M.R. Endsley, Situation awareness and cognitive modeling. *IEEE Intell. Syst.* **27**(3), 91–96 (2012)
18. J. Llinas, C. Bowman, G. Rogova, A. Steinberg, Revisiting the JDL data fusion model II, in *7th International Conference on Information Fusion*, Stockholm (2004)
19. M. Nilsson, J.V. Laere, T. Susi, T. Ziemke, Information fusion in practice: a distributed cognition perspective on the active role of users. *Inform. Fusion* **13**(1), 60–78 (2012)
20. N. Oliveira, F.R. Jorge, J.O. Souza, V.A.P. Junior, L.C. Botega, Development of a user interface for the enrichment of situational awareness in emergency management systems, in *Advances in Safety Management and Human Factors. Advances in Intelligent Systems and Computing*, vol. **491**, ed. by P. Arezes (Springer, Cham, 2016), https://doi.org/10.1007/978-3-319-41929-9_17
21. A.C.M. Oliveira, L.C. Botega, J.F. Saran, J.N. Silva, J.O.S.F. Melo, M.F.D. Tavares, V.P.A. Neris, Crowdsourcing, data and information fusion and situation awareness for emergency management of forest fires: the project DF100Fogo (FDWithoutFire). *Comput. Environ. Urban. Syst.* (2017). <https://doi.org/10.1016/j.compenvurbsys.2017.08.006>
22. V.A.Pereira Junior, M.F. Sanches, L.C. Botega, C.S. Coneglian, N. Oliveira, R.B. Araújo, Using semantics to improve information fusion and increase situational awareness, in *Advances in Intelligent Systems and Computing*, vol. 491 (Springer, Cham, 2016), pp. 101–113

23. V.A. Pereira Junior, M.F. Sanches, J.F. Saran, C.S. Coneglian, L.C. Botega, R.B. Araujo, Towards semantic fusion using information quality and the assessment of objects and situations to improve emergency situation awareness, 2016 Eleventh International Conference on Digital Information Management (ICDIM), Porto, pp. 260–265, doi: [10.1109/ICDIM.2016.7829794](https://doi.org/10.1109/ICDIM.2016.7829794)
24. V.A. Pereira Junior, J.F. Saran, L.Z. Ladeira, J.H. Martins, V. Pagotti, A.M. Souza, L.A. Villas, L.C. Botega, Beyond syntactic data fusion in the context of criminal data analysis, in 2017 12th Iberian Conference on Information Systems and Technologies (CISTI) (IEEE, Piscataway, 2017), pp. 1–6
25. J.J. Salerno, Information fusion: a high-level architecture overview, in *Fifth International Conference on Information Fusion*, Maryland, vol. 1 (2002), pp. 680–686
26. M.F. Sanches, V.A.P. Junior, J.O. de Souza, C.S. Coneglian, F.R. Jorge, N.P. Oliveira, L.C. Botega, Objects assessment approach using natural language processing and data quality to support emergency situation assessment, in *18th International Conference on Human-Computer Interaction*, Toronto (Springer, Berlin/Heidelberg, 2016), pp. 238–244
27. J.F. Saran, V. Mendes, V.A.P. Junior, C.G. Santos, G. Nascimento, M.F.D. Tavares, A.C.M. Oliveira, L.C. Botega, Data and information fusion in the context of emergency management: the DF100Fogo project, in *2017 12th Iberian Conference on Information Systems and Technologies (CISTI)* (IEEE, Piscataway, 2017), pp. 1–6
28. N.A. Stanton, J. Piggott, Situational awareness and safety. *Saf. Sci.* **44**, 0–17 (2001)

Chapter 24

Adaptive Fusion



Vincent Nimier and Kaouthar Benameur

Abstract This chapter describes a methodology for adaptively incorporating reliability of information, provided by multiple sensors, to improve the quality of the data fusion result. The adaptivity is achieved by dynamically utilizing auxiliary information, comprising the measure of performance of each sensor or contextual information when it is available. A comparative study of the results obtained by using either source of auxiliary information for adaptivity is presented in this paper.

Keywords Contextual information · Indigenous information · Exogenous sensing conditions · Supervision information · Supervised data fusion · Adaptive data fusion

24.1 Introduction

The latest developments in perception systems support the need for the joint use of multiple sensors. Indeed, the expected benefits are promising: a greater ability to analyze complex situations and an increased robustness to the environment. Applications of perception systems range from the industrial environment with assembly tasks, mobile robotics, to military domain mainly in the field of C4ISR. Fusion of heterogeneous information is still under investigation and can be considered as a very active research field.

In the fields of detection and estimation, the proposed algorithms very often rely on a probabilistic modeling. In many of these algorithms, the probability distributions of multiple sources are assumed known, independent, constant, and context independent. However, they are often dependent on the context characteristics defining external conditions and vary according to them. This is particularly true for multisensor systems. Indeed, if we consider, for example, a system composed of

V. Nimier (✉) · K. Benameur (✉)
DTIS/ONERA, Palaiseau, France
e-mail: nimier@onera.fr; benameur@onera.fr

a radar and a camera operating in the visible domain, then the information on the use time of the system, day or night, or on the expected altitude of the target, high or low, can affect directly the performance of the system. The use of contextual information in data fusion has received a renewed interest [1–3] and [4]. For example, in [2], the authors list a wide variety of contextual information and present their use at the different levels of JDL model. In [5], the problem of the utilization of contextual information in Level 1 fusion is considered, and five different categories of contextual information are defined: domain knowledge, environment to hardware processing, known entity distribution, traffic behavior history, and road information for traffic tracking.

In this chapter we describe a sensor fusion algorithm driven by *supervision information* represented either by contextual variables when they are available or/and by the current performance of each sensor. Although, it is true that environmental conditions for a sensor can lead to bad performance, the opposite is not true, and bad performance for a sensor does not imply bad environmental conditions. So the two sources of supervision information are not independents, but in the following, we do not consider correlation between both sources.

Consideration of the supervision information allows for improving the quality of the fusion algorithm results. In this paper, we will formally define the context as a supervision information, which is considered as the first-class category of domain knowledge. We will then present the use of the contextual information in the estimation process followed by the integration of the proposed concept in a multisensor fusion methodology for air target tracking.

24.2 Introduction to Supervision Information

The supervision process described in this chapter is performed by employing auxiliary information. This information is defined as supervision information and is used for quantifying reliability of the sensors. The supervision information introduced in this chapter is measured either by endogenous information characterizing the performance of the sensor themselves or by exogenous variables characterizing the sensing conditions. Let us consider a fuzzy event [5] representing a fusion result of n sensors. Let z be the supervision information for a set of sensors participating in fusion, and $\mu_i^j(z_i^j)$ is a fuzzy membership function characterizing reliability of sensor i given the characteristic j of the supervision information. The membership function $\mu_i^j()$ of a fuzzy event takes value in the interval [0 1], where the value 1 means that the sensor is totally reliable, and therefore, its measurement must be a part of the fusion process, while the value 0 means that the sensor is unreliable, and its information must be discarded. An example of exogenous auxiliary information can be the weather which has an influence on the quality of an image provided by an outdoor camera or the SNR of the signal of a radar. An example of endogenous auxiliary information can be the measure of the distance between the sensor measurement and the expected measurement value based on a model.

In the reminder of the chapter, the auxiliary exogenous information will be called contextual information. The nature and origin of this information are very diverse. It can be obtained by utilizing measurements provided by annexes sensors such as those measuring rainfall, pressure, or outdoor temperature, etc. Similarly an operator, through a man/machine interface, can also give valuable information on operational conditions. It is important to notice that contextual information may not be totally reliable or even relevant since, for example, a sensor measuring rainfall cannot be completely reliable itself. Similarly, an operator can be tired or overcharged. Context quality can be considered as a higher level quality. Additional information provided by some sensor signal processing will be designated as endogenous information. This information can be obtained, for example, by measuring SNR, the correlation peak width, and other parameters or indicators that, in some cases, can assess the measurement quality. In this paper an operating status of each sensor is defined by both exogenous and endogenous information, and, the quality of its measurements is considered in the framework of fuzzy logic. This information is evaluated by an expert by assessing the performance and limitations of each sensor by means of membership functions defined on supervision information variable. Therefore, in a real situation, and knowing values for each supervision information variable, the membership functions establish valid measurements from a selected set of sensors. Thus, it is possible to define the association of sensor measurements that are best suited to a particular context and to take only into account in the estimation process the measures resulting from this association.

24.3 Supervision Space

The supervision information is defined by supervision variables. Let S be a system composed of n sensors, and let z define a supervision information variable on a supervision space $Z \subseteq R^n \times P$ with $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$ where p is the maximal number of supervision variables that may be used by a sensor. For a sensor s_i , his set of supervision variable is a vector $z_i = \{z_i^1, \dots, z_i^q\}$ with $q \leq p$. As described in the previous section, z_i^j can be the value of a contextual variable such as a measurement of the fog density, of the rainfall, of the SNR, or a measurement of an endogenous variable. The “validity” of a sensor defines the fact that the sensor provides measurements in accordance to the model used in the estimation process. Let a membership function $\mu_i^j(z_i^j)$ of z_i^j define the validity domain of the sensor s_i (Fig. 24.1). In the following, we denote s_i the sensor i and in bold \mathbf{s}_i its inclusive (to be defined later) validity domain.

From Fig. 24.1, we can see that a membership function $\mu_i^j()$ for $z_i^j \in [0, \kappa_1]$ is equal to 1. Therefore, sensor s_i is valid, i.e., the measurements it provides correspond to the model being used in the estimation process. For $z_i^j \in [\kappa_2, \infty]$, the sensor is not valid, and the provided measurements do not correspond to the used model. For $z_i^j \in [\kappa_1, \kappa_2]$, we use a linear interpolation. If the sensor has more

Fig. 24.1 Membership function of fuzzy set

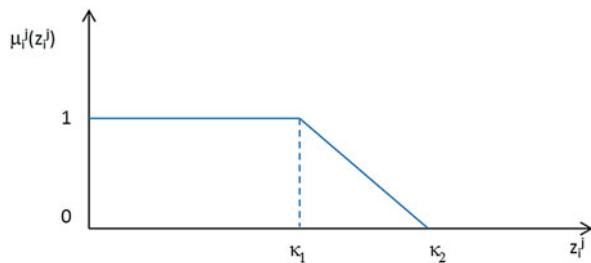
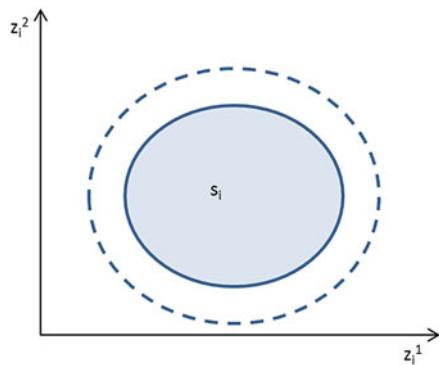


Fig. 24.2 Validity of sensor s_i in the space formed by two supervision variables



than one supervision variable, then its validity can be defined by an intersection of domains obtained for each supervision variable (Fig. 24.2).

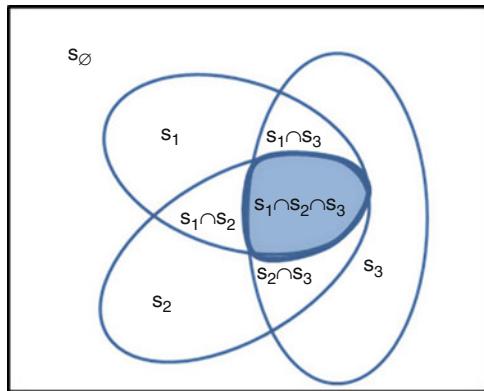
In Fig. 24.2, the blue zone defines the validity domain of the sensor s_i in the case of two supervision variables. It is the domain, where the membership function $\mu_i(z_i^1, z_i^2) = 1$. The dotted line represents the border where the membership function $\mu_i(z_i^1, z_i^2) = 1$ changes from positive value to zero.

This domain is generally defined by a human operator in accordance with his expertise. It can be also defined through learning or experimental tests. The objective is to specify, once the supervision variable is found, the range of this variable corresponding to optimal sensor performances.

Figure 24.3 shows the validity domains s_1 , s_2 , and s_3 of three sensors with only two supervision variables. For simplicity, and unlike in Fig. 24.2, we limit our drawing to the domain where the membership function is equal to 1. We can note that a new domain appears in the figure. This new domain, defined by s_\emptyset , is the domain where no sensor is valid.

For a classical fusion algorithm, which does not consider supervision variable, the implicit assumption is that all the sensors are valid. This implies that sensors are actively measuring at the intersection of their validity domains, shown by a colored area in Fig. 24.3.

Fig. 24.3 Validity domains of three sensors



24.3.1 Inclusive and Exclusive Domains

We now introduce the notion of inclusive and exclusive validity domains. We call a sensor validity domain *inclusive* when we consider a validity domain of a single sensor regardless of validity domains of other sensors participating in fusion (see Fig. 24.2). The exclusive validity domain is the domain where only the considered sensor is valid but other sensors not.

For a system S with n sensors, let us define $\mathbf{2}^S = \{s^\emptyset, s^1, s^2, \dots, s^{[1,2]}, \dots, s^{[1,2,\dots,n]}\}$ a power set, i.e., the set of exclusive validity domains of all possible combination of sensors. The general formula between inclusive s_j domain and exclusive s^j domain is given by:

$$s^J = \bigcap_{j \in J} s_j \cap \bar{s}_i \quad (24.1)$$

where $J \subseteq \{1, \dots, n\}$ and \bar{s}_i defines the complement of the sensor s_i validity domain. We can note that $s^\emptyset = \bar{s}_1 \cap \bar{s}_2 \cap \dots \cap \bar{s}_n$ represents the absence of any valid sensor. The examples of inclusive and exclusive validity domains for sensor 1, with two supervision variables, are given in Figs. 24.4 and 24.5, respectively.

24.3.2 Probability of a Sensor or a Subset of Sensors Validity

In this section we explain how to calculate the probability of a sensor association that will be used to represent the sensor adaptation to the situation under consideration. We define the probability of a fuzzy event σ_i = “sensor s_i is valid” by utilizing the sensor validity domain defined by the membership function of a fuzzy set. We define the set of events $\Omega = \{\sigma_1, \sigma_2, \dots, \sigma_n\}$ that are not exclusively independent. From this set, we can define a set of events that are exclusive Ω . It follows, that two cases corresponding to two types of probabilities will be defined. The first one will be called *inclusive probability* of the group in reference to the inclusive

Fig. 24.4 The inclusive domain of sensor 1

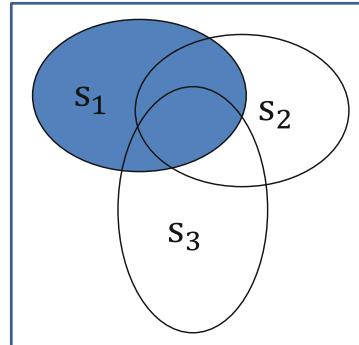
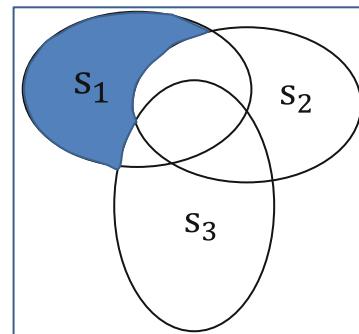


Fig. 24.5 The exclusive domain of sensor 1



domain of validity. The second one will be called exclusive probability of the group in reference to the exclusive validity domain.

Knowing this membership function and the probability density of the measurement vector, we can define the probability of a fuzzy event σ_i = “the sensor is valid” knowing that the supervision information variable z is measured at z_m by:

$$P(\sigma|z_m) = \int \mu_i(z) p(z|z_m) dz \quad (24.2)$$

where z_m is the value of the supervision variable. The definition of the probability of a fuzzy event is given by Zadeh in [6]. When the uncertainty relating to the measurement of a supervision variable is insignificant, we can replace probability density $p(z|z_m)$ by a Dirac $\delta(z - z_m)$ such that the validity probability of the sensor is given by the value of the membership function at the measured supervision variable:

$$P(\sigma_i|z_m) = \mu_i(z_m) \quad (24.3)$$

When the supervision variables are independent, and taking the operator \min as a conjunctive operator, the previous equation can be written as follows:

$$P(\sigma_i|z_m) = \int \min\left(\mu_i^1(z_1), \dots, \mu_i^p(z_p)\right) p(z|z_m) dz_1 \dots dz_p \quad (24.4)$$

24.3.3 Inclusive Validity Probability of a Group of Sensors

In the case of a group of sensors, their inclusive validity probability is equal to the probability of the associated fuzzy event.

For $J \subseteq \{1, \dots, n\}$, the membership function of the event $\bigcap_{i \in J} \sigma_i$ is given by:

$$\mu_J(z) = \text{Min}_{i \in J} (\mu_i(z)) \quad (24.5a)$$

$$P \left(\bigcap_{i \in J} \sigma_i | z_m \right) = \int \mu_J(z) P(z | z_m) dz \quad (24.5b)$$

24.3.4 Exclusive Validity Probability of a Group of Sensors

As we have defined the inclusive s_i and exclusive s^i validity domains of sensor s_i , we can now define the inclusive event σ_i = “the sensor s_i is valid” without any information on sensor j with $j \neq i$ and $j \in \{1, \dots, n\}$ and on the exclusive event σ^i = “only the sensor s_i is valid and the other sensors are not valid.”

Once the inclusive validity probability is defined by (Eq. 24.5), we can determine the exclusive validity probability. It is given by the formula [7]:

$$\beta_J = P(\sigma^J) = \sum_{\{I \subseteq N \cup \phi / J \subseteq I\}} (-1)^{|I-J|} P \left(\bigcap_{i \in I} \sigma_i \right) \quad (24.6)$$

Where $\beta_\emptyset = P(\sigma^\emptyset) = P(\bigcap_{i \in N} \overline{\sigma_i})$ with $N = \{1, \dots, n\}$ and $\overline{\sigma_i}$ defines the event that sensor i is not valid. I and J are subsets of the set N such that $J \subseteq I$ and $I \subseteq N \cup \phi$. The notation $|I-J|$ defines the cardinality of the subset $I-J$. For the sake of simplicity, we replace $P(\sigma^J | z_m)$ with $P(\sigma^J)$. There are as many probability expressions as there are elements in N that means 2^n . The probability β_J verifies the following condition:

$$\sum_{J \subseteq N \cup \emptyset} \beta_J = 1 \quad (24.7)$$

24.4 Supervised Estimation

The probabilities defined above are used to evaluate the validity of all possible groups of sensors. When several associations of sensors are valid at the same time (many possible groups of sensors), the data fusion result is the weighted mean of the

estimates based on measurements provided by the different groups of sensors with weightings equal to the probabilities of associations.

We distinguish two estimation types: a static one, in which the estimates at the present time do not depend on the previous ones, and a dynamic one, where the present estimates depend on previous estimates.

For static estimation the algorithm outlined is presented in [8]. There are two levels of processing. The high level is the context analysis. Based on membership function defined by human experts, the parameters z_j obtained, for example, from an external sensor (humidity, temperature, etc.) or obtained from an internal sensor parameter or from signal processing (SNR, pic correlation width, etc.) are analyzed to define the context. If no contextual exogenous supervision information variables are available, the supervision process can be achieved by considering the defined distance between the actual measurement of a sensor and the predicted one provided by the fusion process. The objective is to penalize the sensor with a measurement such that the distance between the measurement and the prediction is too much important. As a result of this analysis, the coefficients β_J are transmitted to the low level processing to weight different possible associations of sensors.

The next subsection will address the dynamic problem, namely, the problem of tracking a target with a system composed of n sensors.

24.5 Dynamic Estimation

24.5.1 Basic Idea

The basic idea for introducing supervision in a data fusion algorithm is to use the law of total probability. Indeed the estimator at the time k is:

$$\hat{x}_{k/k} = E(x_k | Y_k) \quad (24.8)$$

where Y_k is the vector representing all the measurements provided by all sensors until time k . If we consider all events σ^J with $J \subseteq \{\emptyset\} \cup N$, we can obtain the estimator like:

$$\hat{x}_{k/k} = E(x_k | Y_k) = \int x_k p(x_k | Y_k) dx_k \quad (24.9)$$

If we consider:

$$p(x_k | Y_k) = \sum_{J \subseteq \{\emptyset\} \cup N} p(x_k | Y_k, \sigma^J) P(\sigma^J) \quad (24.10)$$

The combination of Eq. (24.10) in Eq. (24.9) we obtain

$$\begin{aligned}
\hat{x}_{k/k} &= \int x_k \sum_{J \subseteq \{\emptyset\} \cup N} p(x_k / Y_k, \sigma^J) P(\sigma^J) dx_k \\
&= \sum_{J \subseteq \{\emptyset\} \cup N} P(\sigma^J) \int x_k p(x_k / Y_k, \sigma^J) dx_k \\
&= \sum_{J \subseteq \{\emptyset\} \cup N} P(\sigma^J) E(x_k / Y_k, \sigma^J)
\end{aligned} \tag{24.11}$$

The event σ^J means that only the measurements provided by the group of sensors J are valid, so the expectation $E(x_k / Y_k, \sigma^J)$ can be replaced by $E(x_k / Y_k^J, \sigma^J)$. The estimation problem is to calculate

$$\hat{x}_{k/k} = \sum_{J \subseteq \{\emptyset\} \cup N} P(\sigma^J) E(x_k / Y_k^J, \sigma^J) \tag{24.12}$$

24.5.2 Problem Formulation

Let x_k be the state vector at the time k . The dynamic model is supposed to be linear and invariant with equation:

$$x_k = Fx_{k-1} + v_k, \tag{24.13}$$

where F is the state transition matrix. The noise v_k is a Gaussian stochastic process with zero mean and covariance matrix $E(v_i v_j^T) = Q\delta_{ij}$, where δ is the Kronecker symbol.

For the observation model, we consider n observation equations, where each equation corresponds to one sensor. Even if it is rarely the case, by sake of simplicity, we assume that the observation model is linear that lead us to use classical linear Kalman filter. If the observation models are not linear, then EKF (extended Kalman filter) or UKF (unscented Kalman filter) or other variants may be used.

The equation is as follows:

$$y_k^i = H_i x_k + b_k^i, \tag{24.14}$$

where $i \in N = \{1, \dots, n\}$ n is the number of sensors. The matrices $H_1 \dots H_n$ are n observation matrices, one for each sensor. The observation noise, b_k^i , is Gaussian with zero mean and covariance matrix $E(b_k^i b_l^j) = R_i \delta_{kl} \delta_{ij}$. The set of measurements provided by the sensor i up to the time k is denoted as $Y_k^i = \{y_l^i\}_{l=1}^{l=k}$, and the set of all measurements from all the sensors is denoted as $Y_k = \{Y_k^i\}_{i \in N}$. Moreover for a subset $J \subseteq N \cup \emptyset$, we define $Y_k^J = \{Y_k^i\}_{i \in J}$ the set of measurements up to k provided by the association of a group of sensors identified by J .

24.5.3 Equations of the Filter

(a) *Update equation for a group of associated sensors*

For a group of associated sensors [9, 10] whose indices are elements of J , the estimate is given by:

$$\hat{x}_{k/k}^J = E(x_k | Y_k^J) \quad (24.15)$$

The optimal estimate at time k is provided by the relation:

$$\hat{x}_{k/k}^J = \hat{x}_{k/k-1}^J + \sum_{i \in J} K_j^i(k) (y_k^i - H_i \hat{x}_{k/k-1}^J) \quad (24.16)$$

where K_j^i is the Kalman gain associated with sensor i belonging to the set of sensors J . Given that Kalman equations for a multisensory system are easier expressed in the information form, in the following, we consider the information presentation. The Kalman gain is defined by:

$$K_j^i(k) = P_J(k/k) H_i^T R_i^{-1} \quad (24.17)$$

The covariance matrix $P_J(k|k)$ is given by the formula:

$$P_J^{-1}(k/k) = P_J^{-1}(k/k-1) + \sum_{i \in J} H_i^T R_i^{-1} H_i \quad (24.18)$$

(b) *Update equation for an association of sensors with reliability defined by supervision variables*

Now, we derive the expression for a global estimate by considering all possible associations of sensors J with $J \subseteq N \cup \{\emptyset\}$ given by:

$$\hat{x}_{k/k} = E(x_k | Y_k) \quad (24.19)$$

To take into account the supervision variables, we develop the following formula:

$$\hat{x}_{k/k} = x_k^\emptyset \beta_\emptyset(k) + \sum_{J \subseteq N} \beta_J(k) E(x_k | Y_K^J) \quad (24.20)$$

The global estimate is composed of elementary estimates provided by the association of sensors. When none of the sensors is valid, we consider the prediction $\hat{x}_{k/k-1}$ for X_k^\emptyset . In (24.20) the elementary estimates $E(x_k | Y_K^J)$ can be substituted by

their values given in (24.13) where each elementary prediction $\hat{x}_{k/k-1}^J$ of the set of sensors J is replaced by the global prediction $\hat{x}_{k/k-1}$. The relation (24.20) can then be written as follows:

$$\hat{x}_{k/k} = \hat{x}_{k/k-1} + \sum_{J \subseteq N} \sum_{i \in J} \beta_J(k) K_J^i(k) (y_k^i - H_i \hat{x}_{k/k-1}) \quad (24.21)$$

This last relation can be simplified by considering the following notation:

$$K_i(k) = \sum_{\{J | i \in J\}} \beta_J(k) K_J^i(k) \quad (24.22)$$

Where $\{J | i \in J\}$ defines the set of all groups of sensors containing sensor i . Equation (24.21) can then be formulated as follows:

$$\hat{x}_{k/k} = \hat{x}_{k/k-1} + \sum_{i \in N} K_i(k) (y_k^i - H_i \hat{x}_{k/k-1}) \quad (24.23)$$

The covariance matrix is given by the relation:

$$P(k/k) = \sum_{J \subseteq N \cup \{\emptyset\}} \beta_J(k) \left[P_J(k/k) + (\hat{x}_{k/k} - \hat{x}_{k/k}^J) (\hat{x}_{k/k} - \hat{x}_{k/k}^J)^T \right] \quad (24.24)$$

This is similar to the covariance matrix expression of a Gaussian mixture. The matrix $P_J(k/k)$ is the covariance matrix of the elementary estimate $\hat{x}_{k/k}^J$ obtained following the association of sensors J .

Finally with the updated state and covariance, the prediction equations are classically given by:

$$\hat{x}_{k+1/k} = F \hat{x}_{k/k} \text{ and } P(k+1/k) = F P(k/k) F^T + Q \quad (24.25)$$

24.6 Algorithm

Although the previous equations can be directly implemented with any programming language, in data fusion, we often prefer the sequential form of the filter. Firstly, this form is easier to implement, and secondly, it is well appropriate for asynchronous sensors. In the next section, the sequential version of the algorithm is presented in the following two forms: with and without taking into accounts the supervision information.

For simplicity, we consider a system composed of three sensors. The generalization of the algorithm to a system composed of n sensors is straightforward.

24.6.1 Fusion Algorithm

We consider here the sequential form of the Kalman filter for a synchronous system. The state equations for a system S composed of three sensors are the following:

$$S \begin{cases} x_k = Fx_{k-1} + v_k & \text{Dynamics} \\ y_k^1 = H_1 x_k + b_k^1 & \text{Sensor 1} \\ y_k^2 = H_2 x_k + b_k^2 & \text{Sensor 2} \\ y_k^3 = H_3 x_k + b_k^3 & \text{Sensor 3} \end{cases} \quad (24.26)$$

The recursive form [11] for the above system allows conducting the processing in four consecutive steps. The first one, corresponding to s_1 , is the prediction step. The second one is the estimation of an intermediate state x_k^1 . The third step, corresponding to s_2 , allows the estimation of a second intermediate state x_k^2 . The last step, corresponding to s_3 , allows the estimation of the global state x_k .

Step 1

From the estimated state given by the $k-1$ iteration, the prediction is given by the classical Eq. (24.23) above.

Step 2

The prediction is updated based on the formula:

$$\hat{x}_{k/k}^1 = \hat{x}_{k/k-1} + K_1(k) \left(y_k^1 - H_1 \hat{x}_{k/k-1} \right) \quad (24.27)$$

and

$$P_1(k/k) = (I - K_1(k)H_1) P(k/k-1) \quad (24.28)$$

with:

$$K_1(k) = P(k/k-1) H_1^T \left(H_1 P(k/k-1) H_1^T + R_1 \right)^{-1} \quad (24.29)$$

Step 3

Considering s_2 , the prediction is simply obtained by substituting the predicted state $\hat{x}_{k/k}$ by $\hat{x}_{k/k}^1$ and $P(k/k)$ by $P_1(k/k)$ and updating the state and the covariance matrix. The Kalman gain is now:

$$K_2(k) = P_1(k/k) H_2^T \left(H_2 P_1(k/k) H_2^T + R_2 \right)^{-1} \quad (24.30)$$

Step 4

This step is the same as the previous one but now we substitute $\hat{x}_{k/k}^1$ by $\hat{x}_{k/k}^2$ and $P_1(k/k)$ by $P_2(k/k)$. The Kalman gain is now:

$$K_3(k) = P_2(k/k) H_3^T \left(H_3 P_2(k/k) H_3^T + R_3 \right)^{-1} \quad (24.31)$$

Remark

In each update step (2, 3, 4), the noise covariance matrix R_i of the associated sensor i is taken into account in the Kalman gain. It is this covariance matrix that weights contribution of the sensor measurement with respect to the prediction in the calculus of the estimate. The larger the covariance of a sensor noise (the trace of the matrix is big), the less the measurement of this sensor is taken into account in the fusion process. It is a first approach to adapt the fusion algorithm to the deterioration in the sensor performance. The approach is to calculate the covariance matrix of the noise regularly for each sensor and to take into account the new covariance matrix in computing the Kalman gain. However, this method is not sufficient because it is possible to calculate a good covariance matrix with a sensor that is providing false information.

24.6.2 Supervised Fusion Algorithm

The proposed algorithm is based on the previously presented algorithm with the main difference that we must calculate $2^n - 1$ estimates for a system with n sensors instead of n as it was the case in the previous algorithm. Each estimate corresponds to an association of sensors. The number of the combinations should not be an obstacle because in general a multisensor system is composed of a limited number of sensors (two to three sensors).

We will describe here five processing steps. The first step is still devoted to the prediction. The following three steps are for computing estimates for different association of sensors. The fifth step is to supervise the different associations of sensors by computing a weighted mean of the estimates.

Step 1

Like the preceding algorithm, this step is used to predict the estimate and the covariance matrix. From the estimate computed at the preceding iteration $\hat{x}_{k-1/k-1}$ and $P(k-1/k-1)$, this step computes the prediction of $\hat{x}_{k/k-1}$ and $P(k/k-1)$ according to Eq. (24.23).

Step 2

In this step the predicted state and covariance matrix are updated by the measurements y_k^i provided by the three sensors in order to obtain three estimated states $\hat{x}_{k/k}^i$ and $P_i(k/k)$ with $i \in \{1, 2, 3\}$. Each state and covariance matrix will be used in the Step 3 for the second phase of the estimation.

Step 3

In the considered case, there are three couples of sensor association: {1, 2}, {1, 3}, and {2, 3}. For the couple {1, 2}, for example, we note the estimate $\hat{x}_{k/k}^{12}$ and its covariance matrix $P_{12}(k/k)$. To compute this estimate and this covariance matrix, we simply take the estimate of $\hat{x}_{k/k}^1$ and $P_1(k/k)$ and update them with the measurement y_k^2 and the noise covariance R_2 provided by sensor 2. For the couple {1, 3}, we take the estimate $\hat{x}_{k/k}^1$ and the covariance matrix $P_1(k/k)$, and we update them by considering the measurement and covariance matrix of sensor 3. For the couple {2, 3}, we use the same approach as the previous one: the state $\hat{x}_{k/k}^2$ and the covariance matrix $P_2(k/k)$ are updated by the measurement and noise covariance R_3 provided by sensor 3.

Step 4

At this level, the global fusion of the three sensors corresponding to the subset {1, 2, 3} is made. We can take one of the three states computed in the preceding step and update it by the measurement from the sensor that is not already contributing to the update of the selected state. For example, if we take the state $\hat{x}_{k/k}^{12}$ and its covariance matrix $P_{12}(k/k)$, we update them by the measurement y_k^3 and the covariance matrix R^3 . In order to obtain $\hat{x}_{k/k}^{123}$ and $P_{123}(k/k)$, the order of the indices does not matter. In fact $\hat{x}_{k/k}^{123} = \hat{x}_{k/k}^{231} = \hat{x}_{k/k}^{312}$.

Step 5

This fifth step introduces the supervision information in the computation of the estimate. At the time of the estimation, each supervision variable has a value which allows computing the coefficients $\beta_j(k)$ obtained from the membership function of the fuzzy set. These coefficients are used to compute the global estimate $\hat{x}_{k/k}$ and the global covariance $P(k/k)$ based on the Eqs. (24.21, and 24.22).

24.7 Simulations

24.7.1 Simulation Conditions

The purpose of the following simulations is to fuse the measurements of three sensors and evaluate the feasibility of the introduced adaptive method utilizing supervision variables. The first sensor providing angle measurements y_1 is a search radar with poor resolution, but this sensor is never jammed or deceived. The second sensor with angle measurements y_2 is a tracking radar which is jammed in the time interval [75, 125]. The third sensor is an I.R camera providing angle measurements y_3 . On, below, we can see some decoys in the time interval [20s 120s] (Fig. 24.6).

The measurement y_1 has a Gaussian noise with standard deviation $\sigma_1 = 6^\circ$. The measurements y_2 and y_3 have equal Gaussian noises with standard deviation

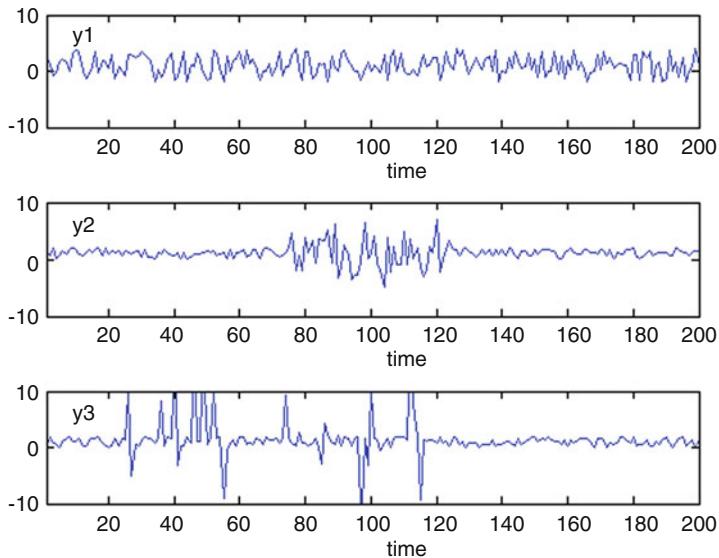


Fig. 24.6 Sensor measurements (x axis time in s and y axis angle in degrees)

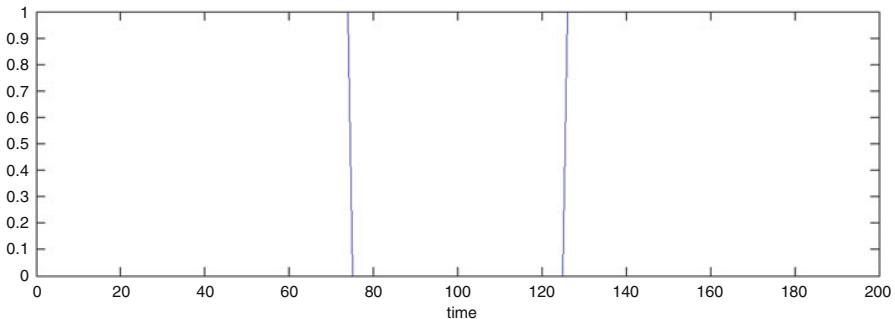


Fig. 24.7 Time evolution of $\mu_2(z_2)$, X -axis is the time in second, and Y -axis is a logical value without dimension

$\sigma_2 = \sigma_3 = 2^\circ$. Between times 75 s and 125 s, we add a noise with a standard deviation $\sigma_2 = 10^\circ$ to the measurement y_2 . To simulate a jamming between times 20s and 120 s, we add randomly, in time, a peak with a Gaussian amplitude and a standard deviation $\sigma_2 = 20^\circ$ to simulate a decoy.

On the tracking radar, a parameter z_2 can indicate the instant when the radar is jammed. For example, z_2 can be the SNR ratio and when this ratio is under a certain value z_m then $\mu_2(z_m) = 0$. The membership function of the fuzzy set on this parameter takes the values given in Fig. 24.7 above.

For the detection of the decoy, the parameter z_3 is used, and the fuzzy set function on this parameter takes the values given in Fig. 24.8 below. The parameter z_3 can be

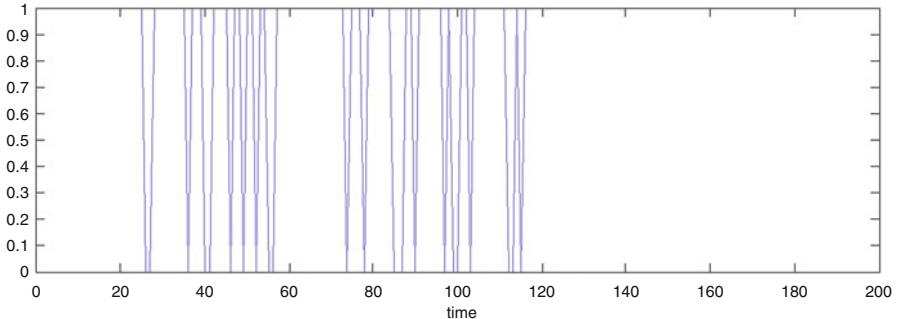


Fig. 24.8 Time evolution of $\mu_3(z_3)$, X -axis is the time in second, and Y -axis is a logical value without dimension

a parameter score given by image processing and $\mu_3(z_3) = 0$ when the score is under a certain limit z_l . We do not describe the contextual variables or their membership functions because they are very application dependent.

24.7.2 Results with Contextual Information

The results of the simulation are given in Fig. 24.9 and Fig. 24.10.

In Fig. 24.9, the upper plot shows the result of the tracking estimate x_f without taking into account the context, while the lower plot shows the results of the tracking estimate x_{fc} when the context is taken into account.

We can see that the estimate x_{fc} is less noisy than x_f . This difference can be seen in another way in Fig. 24.10.

In this figure, we can see that before time $t = 20$ the errors of x_f and x_{fc} are the same. It is normal because only the probability of the association of three sensors is equal to 1 and other probabilities of association are reduced to 0, so the two algorithms are equivalent. In the time interval $[75 \text{ } 125]$, the error of x_{fc} is almost always inferior to that of x_f showing the benefit of taking account of the context. After $t = 125$ the error in the two algorithms converges toward the same value, and we can see that the errors are the same a small time after $t = 140$.

The RMSE of the result without supervision is 0.63° . and with supervision it is 0.24° . The supervision of the fusion process by contextual information divides by almost three the RMSE.

24.7.2.1 Results without Contextual Information

In this subsection, the fusion process does not use external information given by contextual information but only the distance between the measurement and its

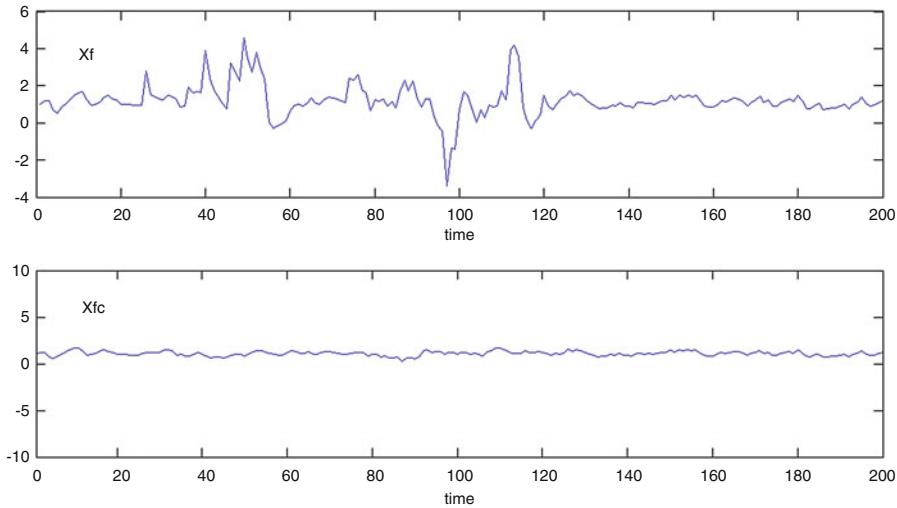


Fig. 24.9 Results of x_f and x_{fc}

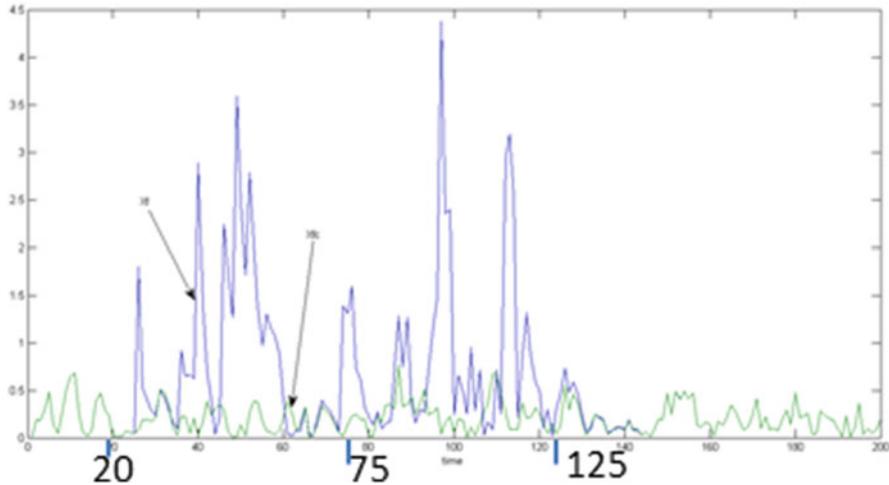


Fig. 24.10 Estimation error (X-axis is the time in second, — error x_f , — error x_{fc})

prediction given by the fusion process. We note $\hat{x}_{k/k-1}^N$ the prediction of the state after the fusion of all measurements until the time $k-1$. When a measurement y_k^i provided by the sensor i is sent to the fusion process, then the prediction of this measurement is achieved, thanks to the following equation:

$$\hat{y}_{k/k-1}^i = H_i \hat{x}_{k/k-1}^N \quad (24.32)$$

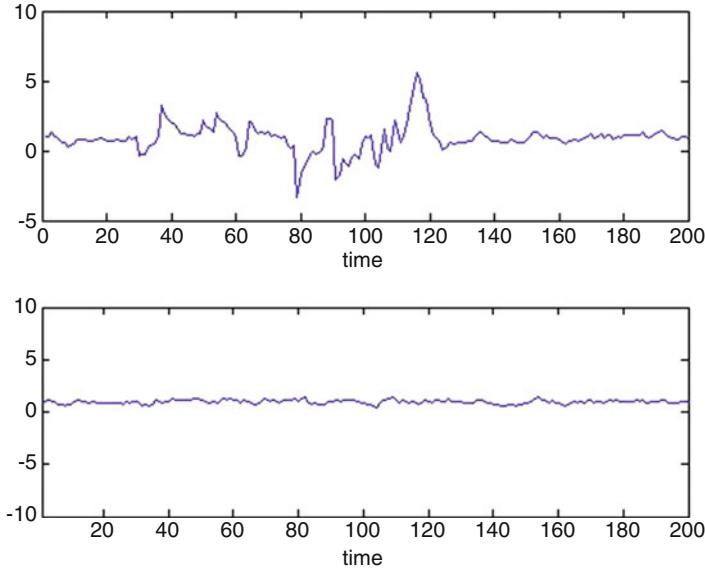


Fig. 24.11 Results upper plot x_f and lower plot x_{fc} (X-axis is the time in second)

It is possible to calculate the distance between the prediction and the measurement and define the variable z_m :

$$z_m = \left\| \hat{y}_{k/k-1}^i - y_k^i \right\|_{S_i} \quad (24.33)$$

with $S_i = H_i P(k/k) H_i^T + R_i$

We define the membership function by

$$p(s_i/z_m) = \mu(z_m) = e^{-\frac{z_m^2}{2}} \quad (24.34)$$

Using the same simulation described above, we obtain the following results (Fig. 24.11).

The upper plot shows the result of the estimate x_f without considering the context, while the lower plot shows the result of the estimate x_{fc} when the fusion process is supervised by the measurements related distance.

As it can be seen from the Fig. 24.12 below, the supervised fusion algorithm without contextual information gives a better result in term of RMSE than the one using contextual information. This result is against expectation because in the algorithm with contextual information, we use more information than in the algorithm without contextual information. This better performance can be explained

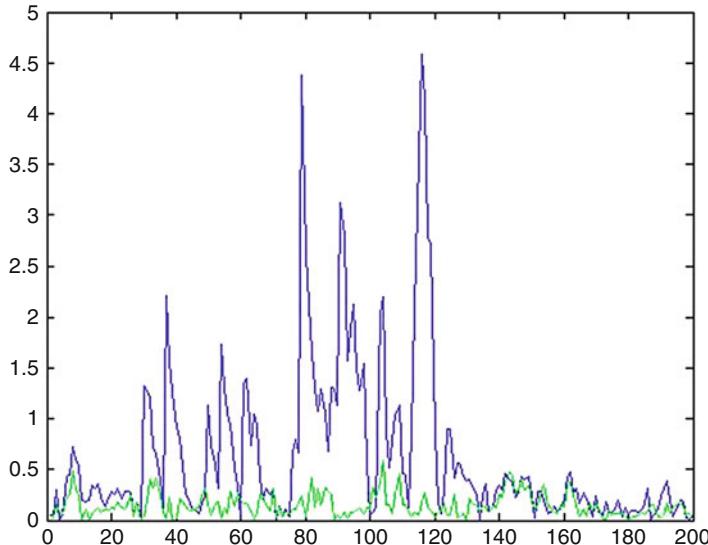


Fig. 24.12 Estimation error (X -axis the time in second, — error x_f , — error x_{fc} , RMSE = 0.14)

by the fact that in the algorithm without contextual information, the perturbed measurements are not totally discarded but have a smaller weight in the fusion process. At the same time, the supervision without contextual information gives an advantage to the prediction model. It should not be robust to target maneuvers. More studies will be conducted in the future to define the working area of each of the algorithm.

24.8 Conclusion

We have introduced an adaptive fusion algorithm, which estimates and incorporates reliability of sensor performance by utilizing supervision information. Supervision information comprises contextual variables characterizing the external sensing conditions when this information is available or endogenous sensor information. The use of the reliability of sensor performance in a multisensory system is essential because it dynamically allows selection and fusion of the measurements of the most reliable sensors.

The algorithm introduced in the chapter, as well as the simulation example, is based on simple linear observation models. The extension to the nonlinear case is straightforward. The approach developed here for a tracking application can also be adopted for numerous other applications.

Acknowledgment The authors would like to thank Dr. Galina Rogova from the State University of New York at Buffalo for her helpful advice on various technical issues examined in this chapter.

References

1. L. Snidaro, I. Visentini, Integration of contextual information for tracking refinement, in *International Conference on Information Fusion*, Chicago, 2011
2. I. Visentini, J. Llinas, G.L. Foresti, L. Snidaro, Context in fusion: Some considerations in a JDL perspective, in International Conference on Information Fusion, Istanbul, Turkey, 2013
3. V. Nimier, G. Rogova, Reliability in information fusion: Literature survey, in *International Conference on Information Fusion*, Stockholm, 2004
4. E. Duflos, D. Pomorski, P. Vanheeghe, F. Caron, GPS/IMU data fusion using multisensor Kalman filtering: Introduction of contextual aspects. *Information Fusion* **7**, 221–230 (2006)
5. J. G. Herrero, L. Snidaro, J. Linas, G. Steetharra, K. Palaniappan, E. Blasch, Overview of contextual tracking approaches in information fusion, in SPIE Signal Processing, Sensor/Information Fusion, and Target Recognition, 2013
6. L.A. Zadeh, Probability measures of fuzzy events. *J. Math. Anal. Appl.* **23**, 421–427 (1968)
7. V. Nimier, Supervised multisensor tracking algorithm, in *International Conference on Information Fusion*, Las Vegas, 1998
8. V. Nimier, Supervising the fusion process by context for target tracking, in *Context Enhanced Information Fusion*, Springer, 2016
9. C.Y. Chong, Hierarchical estimation, in *ONR Workshop on distributed Communication and Decision Problems*, Monterey, 1979
10. E. Tse, R.P. Whisner, C.Y. Chong, S. Mori, Distributed estimation in distributed sensor networks, in *IEEE American Control Conference*, Arlington, 1979
11. Y. Bar-Shalom, A. Houles, Multisensor tracking of a maneuvering target in clutter. *IIEE trans. on AES* **25**(2), 176–189 (1989)

Index

A

Acquaintance networks, 215
Acquisition-decision process, 403
Actionable intelligence (AI)
 information fusion, 472–476
 information quality, 475, 477–478
 Cloud, 514
 cost-benefit assessment, 514
 definitions, 479, 482
 HLIF domain, 479–485
 IQEs (*see* Information quality elements)
 IQMM, 487, 488
 MDL, 508–509
 mitigation, 498–500, 509
 namespaces, 488
 OODA, 514
 overloading, 487
 strategic implications, 485–487
 subjective assessment, 502–507
 integrated and holistic approach, 475,
 477–478
 intelligence community, 473–474
 JDL model, 473
 MDM, 472
Actuarial survival models, 462
Adaptive fusion
 basic idea, 594–595
 contextual information, 602–605
 filter equations, 596–597
 fusion algorithm, 598–600
 problem formulation, 595
 simulation conditions, 600–602
 supervised estimation, 593–594
 supervision information, 588–589

supervision space
 definition, 589
 exclusive validity probability, 593
 inclusive and exclusive domains,
 590–591
 inclusive validity probability, 593
 membership function, 589, 590
 probability of a sensor/subset of sensors
 validity, 591–592
 validity domain, 589–591
Advanced driver assistance systems (ADAS),
 424
Agreement weight, 303, 311
AHP, *see* Analytic Hierarchy Process
AI, *see* Actionable intelligence
Aircraft prognostics use case
 application, 393
 hypothesis, 392–393
 implementation, 394
 purpose, 390, 391
 social model, 392
 theory, 391
 verification, 394
 α -conjunctions, 42
Ambiguity measure (AM), 103
Analytic Hierarchy Process (AHP), 340
AP, *see* Average precision
Architecture principles (APs), 477
Artificial Intelligence algorithms, 188–189
Artificial neural networks (ANN)
 behavior reflex, 425–426
 camera-based detection system, 424
 direct perception, 426–427
 ego-lane estimation

- Artificial neural networks (ANN) (*cont.*)
 - architecture, 441, 442
 - ground truth acquisition, 441–443, 445
 - result of, 449–452
 - structure, 443–444
 - training, 444
- mediated perception, 427–428
- metric measures, 445
- preprocessing features, 432–433
- recording training and testing data, 444, 445
- reliability estimation
 - DST, 440–441
 - feature selection, 435–436, 446–447
 - fusion strategies, 439–440
 - hypotheses, 435
 - layers, 433, 434
 - mutual information, 434
 - ReLU, 437
 - result of, 447–449
 - SGD, 438
 - structure, 437, 438
 - training process, 436–437
- scenario features, 431–432
- sensor setup, 429–431
- types, 428
- Assumption consistency, 384–387
- Attribute-level uncertainty, 268
- Average precision (AP), 316–319
- B**
- Balanced two-layer conflict solving (BalTLCS), 547, 551–553
- Ballooning extension, 41
- Basic belief assignment (BBA), 79–80, 548
- Basic probability assignment (bpa), 100–101
- Bayesian Belief Network (BNN), 397
- Bayesian modeling, 112–113
- Behavior models
 - with ego-sources
 - discount operation, 120–122
 - joint discounting, 122–124
 - posterior distribution, 119
 - RMSE, 120–121
 - source behavior characterization method, 119–120
 - expected probabilities, 129
 - state variables/propositions, 129–130
 - without ego-sources
 - three-mode JointConDis, 126, 128–129
 - two-mode JointConDis, 125–128
- Behavior reflex, 425–426
- Behaviour-based correction (BBC), 41–44, 47
- Behaviour-based fusion (BBF) rule, 42, 44–46
- Belief Function-based Technique for Order Preference by Similarity to Ideal Solution (BF-TOPSIS), 331, 343
- DAMs
 - BBA construction, 342, 344
 - BF-TOPSIS1, 344
 - BF-TOPSIS2, 344
 - BF-TOPSIS3, 344–345
 - BF-TOPSIS4, 345
 - DMP formalization, 342, 344
 - torrential protective actions, 351–352
- Belief functions, 33, 37, 42
- DAMs (*see* Decision-aiding methods)
- DMPs (*see* Decision-making problems)
- torrential protective actions
 - BF-TOPSIS, 351–352
 - DMPs, 348–349
 - ER-MCDA, 349–351
 - FCOWA-ER, 352–353
- See also* Conflict measure
- Benchmark sets, 315
- Beta distribution, 112
- Beta reputation systems (BRS), 112
- Betweenness centrality measure, 217
- Bini Adjustment, 528
- Black campaigns, 200
- Blocking errors, 300–302
- Blocking process, 322
- Boolean binary connectives, 35–37
- Boolean rule systems, 323
- Boston Bump, 361
- Business intelligence (BI), 474, 499
- C**
- Capability, intent, and opportunity (COI), 487
- Cautious Ordered Weighted Averaging with Evidential Reasoning (COWA-ER), 331
- CCMP, *see* Common Classification of Medical Procedures
- Center hypothesis (CH), 435
- Choquet integrals, 72, 74, 143
- Classical Hartley measure, 102
- Clerical review indicators, 305
- Climate data analytics use case, 395–397
- Coalition
 - CSAR scenario vignette
 - degree by degree usefulness assessment window, 148, 150
 - global usefulness assessment presentation, 148–149

- information/support description, 148–149
in-the-loop usage, 150–151
long-term observation/intelligence task, 145
overview of, 145–146
usefulness priorities, 147–148
user information request description, 147
user profile attributes, 145–147
degrees of usefulness, 140–142
aggregation function, 142–144
assessment, 137
definition, 144
modeling information, 138–139
modeling users, 139–140
preliminaries, 138
Vector Model, 136
- Combat Search And Rescue (CSAR) scenario
vignette
degree by degree usefulness assessment window, 148, 150
global usefulness assessment presentation, 148–149
information/support description, 148–149
in-the-loop usage, 150–151
long-term observation/intelligence task, 145
overview of, 145–146
usefulness priorities, 147–148
user information request description, 147
user profile attributes, 145–147
- Common Classification of Medical Procedures (CCMP), 460
- Compactness (CW), 468
- Comprehensive quality model
challenge, 411–412
characterization, 412–413
cyber-physical systems, 412
information quality, 413–415
- Concept consistency, 383–387
- Conditional embedding, 41
- Confirmation assertion, 306
- Conflict measure, 101–102
auto-conflict, 82
based on distance, 83–86
based on inclusion degree, 84–86
conjunctive-based rules, 81
contradiction, 81–82
credibility and plausibility, 80
decision-making, 94–95
definition, 81–82
global conflict, 81
management, 86–87
- combination rule, 93
Dempster’s rule, 87–88
false assumption of closed world, 88–89
number of sources, 92–93
source reliability assumption, 91–92
source’s ignorance, 90–91
source’s independence, 90
mass function, 79–81
partial conflict, 81
pignistic probability, 80
Zadeh’s example, 81
- Conflict-Modified-DST (CMDST), 548, 551, 553
- Conflict resolution, 307
- Confusion matrix, 87
- Conjunctive query (CQ), 267
- Contact matrix, 208–209
- Context
city traffic tracking, 234–236
context discovery, 221
in information fusion
condition expectations, 222
inference/management problem, 222
interactions, 222–223
operational knowledge, 221
problem variables and context variables, 222
volcano eruptions, 221–222
- maritime surveillance systems, 237–238
- quality control, 233–234
- quality of information
abductive process, 231
accessibility, 226
consistency, 230–231
function of, 228–229
key-value models, 231–232
logic-based models, 232–233
objective quality, 223–225
observation and fusion results, 229
ontology-based models, 232
problem variables, 229–230
recognition result, 225–226
relevance, 227–228
reliability, 227
selection of variables, 229–230
subjective quality, 224–225
time delay, 231
trust, 226–227
role of, 220
- Context-dependent multisensor fusion, 427
- Contextual information quality dimensions, 413, 414
- Contradiction, 81–82
- Convolutional neural network (CNN), 426

Correction assertion, 306
 Course of action (COA), 473, 478, 486
 COWA-ER, *see* Cautious Ordered Weighted Averaging with Evidential Reasoning
 CQ, *see* Conjunctive query
 Credal sets, 100, 101
 Credibility weighting
 chained and modified evidential markers, 175–176
 cross-linguistic comparison, 172–173
 example of, 176–178
 informed decision-making, 173
 labels and expressions, 174
 lexical items, 173–174
 maximum uncertainty, 174–175
 negation, 176
 speakers of English, 172
 text analytics, 171–172
 Cross entropy, 57
 Cyber-physical systems (CPS), 539, 540

D

Damping factor, 529–531
 DAMs, *see* Decision-aiding methods
 Data acquisition, 569–570
 Data analysis process (DAP)
 aircraft prognostics use case
 application, 393
 hypothesis, 392–393
 implementation, 394
 purpose, 390, 391
 social model, 392
 theory, 391
 verification, 394
 challenge, 360
 climate data analytics use case, 395–397
 community, 361–363
 decomposition
 accuracy, 384–388
 assumption consistency, 384–387
 concept consistency, 383–387
 dimensions, 382
 external consistency, 384–388
 focal and contextual, 378
 inter-coder reliability, 389
 methodology, 381–382
 prominence, 384–388
 questionnaire, 379–380
 relationship consistency, 383–387
 utility, 384–387
 V&V checklist, 380–381
 definition, 363

epistemological hierarchy model
 analytic mechanisms, 367, 368
 composite model, 370–373
 data cleaning, 376–377
 data elements, 375, 376
 data layer, 374
 development, 375, 377
 implementation layer, 374
 instantiation, 374, 375, 378
 integration and extension, 373, 374
 MESA, 368
 Ruvinsky, Wedgwood, and Welsh model, 367–368
 scientific inquiry, 369
 “Sentiment” domain concept, 375
 falsifiability management, 366
 inconsistent assumptions, 365
 misrepresentation of data, 366
 Data fusion, *see* Entity-based data fusion (EBDF)
 Data matching, 298
 Data quality management (DQM) process
 implementation, 320
 planning, 321
 quality assurance, 321
 blocking, 322
 data rationalization, 323
 ER process, 322–323
 source information, 322
 quality control, 321
 quality improvement, 324
 Data Quality Vocabulary (DQV), 582
 Data rationalization
 errors in, 307–308
 QA process, 323
 Decision-aiding methods (DAMs)
 BF-TOPSIS, 343
 BBA construction, 342, 344
 BF-TOPSIS1, 344
 BF-TOPSIS2, 344
 BF-TOPSIS3, 344–345
 BF-TOPSIS4, 345
 DMP formalization, 342, 344
 ER-MCDA, 340–342
 FCOWA-ER, 346
 BBA construction, 347
 COWA-ER method, 346
 OWA approach, 345
 limitations, 339
 Decision-making
 under certainty, 333, 349
 evaluation of information
 accessibility, 285–286
 acquisition and processing, 287

- indicators, 290
measurement preparation, 290–291
nominal properties, 290
possibility measure, 286
problem identification, 287–288
qualitative properties, 289
quantitative necessity measure, 286
quantitative properties, 289
schema of, 288
selection of indices, 290
Shannon's model of communication, 288–289
signed probabilities, 286
situation demands, 285
software evaluation, 287
process of
 actions, 284
 descriptive models, 280
 factors, 280
 group decision-making, 284–285
 impact models, 280
 operational models, 280–281
 stages, 283
 stratification, 283–284
 structure of, 281–283
 temporal constraints, 284
- Decision-making problems (DMPs)
cost *vs.* damage reduction, 330
definition, 334
elements-at-risk, 330
formalization
 decision-making under certainty, 333
 multi-criteria, 332
 preferences, 332–333
information imperfection
 belief function theory, 336–337
 belief mass, 338–339
 fuzzy set theory, 335–336
 possibility theory, 336
 sources reliability, 337–338
 types, 334
MCDA, 331
torrential protective actions, 348–349
- Degrees of usefulness, 140–142
aggregation function, 142–144
assessment, 137
definition, 144
- Dempster's approach, 33
- Dempster-Shafer modified Theory (DSmT), 89
- Dempster-Shafer theory (DST), 41, 336–337, 440–441, 543
- See also* Theory of evidence
- Dempster's rule (DS), 338
- Dempster's rule of combination (DRC), 543
- Deng entropy, 104
“Deterministic” matching, 302
- Dezert-Smarandache Theory (DSmT), 337
- Digital literacy education programs, 201
- Directed acyclic graph (DAG), 532
- Direct perception, 426–427
- Dirichlet distribution, 114–115
- Disagreement weight, 303, 311
- Discounting, 32, 37, 46
- Disjoint-independent database, 255–257
- Dispersion, 468
- DQV, *see* Data Quality Vocabulary
- E**
- Early CHeckpoint via Epistemological Critique (E-CHEC), 381
- Echo chambers, 184–185
- Ego-lane estimation
 architecture, 441, 442
 ground truth acquisition, 441–443, 445
 result of, 449–452
 structure, 443–444
 training, 444
- Ego-lane models, 428, 429, 433
- Emergency Situation Assessment System (ESAS), 582, 584
- Endsley's model, 6
- Enterprise Engineering Review Board (EERB), 510
- Entity-based data fusion (EBDF)
 adverse events, 296
 clerical review and assertion
 correction and confirmation assertions, 306
 high-risk applications, 304, 305
 policies and procedures, 305
 robust metadata, 306
 scoring rule, 305
- DQM process (*see* Data quality management process)
- entity resolution process, 297–298
- ER accuracy
 F-measure, 309
 inferred methods (*see* Penning's method)
 precision, 308
- Pullen's stratified sampling method (*see* Pullen's stratified sampling method)
- recall, 308
- errors in, 298, 299
 blocking, 300–302
 data rationalization, 307–308
- ER systems, 302

- Entity-based data fusion (EBDF) (*cont.*)
 frequency-based, probabilistic
 matching, 302–304
 source information, 299–300
 law enforcement, education, and healthcare,
 296–297
- Entity resolution (ER)
 accuracy
 F-measure, 309
 inferred methods (*see* Penning’s
 method)
 precision, 308
 Pullen’s stratified sampling method (*see*
 Pullen’s stratified sampling method)
 recall, 308
 process, 297–298
- Environment-aware fusion approach, 428
- Epistemological Data Model Hierarchy
 (EDMH), 370–372
- Epistemological decomposition (“e-decomp”)
 accuracy, 384–388
 assumption consistency, 384–387
 concept consistency, 383–387
 dimensions, 382
 external consistency, 384–388
 focal and contextual, 378
 inter-coder reliability, 389
 methodology, 381–382
 prominence, 384–388
 questionnaire, 379–380
 relationship consistency, 383–387
 utility, 384–387
 V&V checklist, 380–381
- Epistemological hierarchy model
 analytic mechanisms, 367, 368
 composite model, 370–373
 data cleaning, 376–377
 data elements, 375, 376
 data layer, 374
 development, 375, 377
 implementation layer, 374
 instantiation, 374, 375, 378
 integration and extension, 373, 374
 MESA, 368
 Ruvinsky, Wedgwood, and Welsh model,
 367–368
 scientific inquiry, 369
 “Sentiment” domain concept, 375
- Equivalent references, 298
- Evidential Reasoning for Multi-Criteria
 Decision Analysis (ER-MCDA),
 331, 340–342, 349–351
- Exclusive validity probability, 593
- Expectation-maximization (EM) algorithm,
 113
- Exploitation support data (ESD), 494–495
- Extended Hartley measure, 102
- External consistency, 384–388
- Extract-transform-load (ETL), 472
- Intrinsic information quality dimensions, 413,
 414
- F**
- Facebook, 185, 187, 189, 201
- Fact-finding approach, 113
- Fake news
 algorithmic detection, 188–189
 automated algorithmic verification,
 187–188
 BMW X3 model, 192–193
 cognitive psychology, 184
 communication patterns, 185–186
 confirmation bias, 184
 crowdsourced fact-checking, 187
 digital literacy programs and manuals,
 187
 disinformation, 186
 echo chambers, 184–185
 Facebook users, 185
 history of, 182–184
 mainstream media, 193–195
 misinformation, 186
 particular cognitive shortcut, 184
 photo forgeries, 185
 political gains, 198–200
 professional fact-checking, 187
 red flags, 188
 clickbait writing style, 190–191
 conspiracy, 191
 lack of factual information, 190
 lack of sources/unreliable sources, 190
 manipulated images/wrongly attributed
 videos, 191
 propaganda stories, 191–192
 republishing old stories, 190
 source hacking, 191
 unnamed sources, 190
 viewpoints, 191
- Republic of Indonesia, 200–201
- sharing of information, 185
- Syrian conflict
 Arab-language publications, 197
 Atlantic Council’s Digital Forensics
 Lab, 197–198
 military source, 196–197

- on-the-ground journalistic information, 195
Russia Today, 196
SAMS, 195–196
SOHR, 195
source-hacking attempt, 197
Sputnik's information source, 198
US coalition attacks, 196–198
2016 US presidential elections, 187
Volkswagen emissions scandal, 192
Falsifiability management, 366
Fault-tolerant object estimation, 428
FCOWA-ER, *see* Fuzzy Cautious Ordered Weighted Averaging with Evidential Reasoning
Fellegi-Sunter model, 309, 310
F-measure, 309, 314, 316
Focal element, 101
Forest fire model, 216
Frame of Discernment (FoD), 336–337
Frequency-based probabilistic matching, 302–304
Frequency-based scoring rule, 302–304
Fusion-based human-machine information system
information deficiency, 4
information quality
context quality, 6–7
contextual knowledge, 7
definitions, 8–10
higher level quality, 18
information content, 10–13
information flow, 7–8
information presentation, 16–18
information sources, 13–16
inter-and intra-processing, 6
JDL model, 5–6
Omnibus Model, 6
overall quality measure, 22–23
quality attributes, 18–22
quality control, 23–25
research questions, 5
Fuzzy Cautious Ordered Weighted Averaging with Evidential Reasoning (FCOWA-ER), 331, 346
DAMs
BBA construction, 347
COWA-ER method, 346
OWA approach, 345
torrential protective actions, 352–353
Fuzzy set theory, 335–336, 546
- G**
Geographic information systems (GIS), 409–411
Gini entropy, 55–57
Global conflict, 81, 89
Google Adjustment, 528
Google Flu Trends' (GFT), 361–362
GoogLeNet, 427
Graph theory, 210
Ground truth acquisition, 441–443
- H**
High level information fusion (HLIF), 474–475, 477, 479–485
HLIF system of systems (SoS)
dynamic nature of information, 492–494
externalities considerations, 495–496
information governance, 510–511
life cycle process, 511, 512
maturity, 501–502, 512
ontology development, 513
real-time and streaming information, 494–495
resource catalog, 513
SLAs, 510
static nature of information, 491–492
Hoaxy, 187
Hospital information system (HIS), 457, 458
Hyper power set, 89
- I**
ICCT, *see* International Council on Clean Transportation
Image segmentation, 407
Implicative importance weighted ordered weighted averaging (IIWOWA), 543, 547
Inclusive validity probability, 593
Indiana University Open Networks Institute, 187
Industry 4.0
conflict measure, 541, 542
CPS, 539–540
demonstrator setup, 557
DST, 543
importance measure, 542, 554, 557
information fusion, 541
intelligent technical systems, 555

- Industry 4.0 (*cont.*)
- MACRO
 - architecture, 545
 - BalTLCS, 546, 547
 - definition, 544
 - fuzzy set theory, 546
 - OWA operator, 547
 - signal conditioning, 546
 - newspaper printing process, 540
 - requirements, 555
 - self-configuration, 554, 556
 - self-optimisation, 554, 556
 - sensor defect, 558, 559
 - solid-borne sound sensor, 557
 - system health, 558
 - system manager, 556
- Information governance (IG), 510–511
- Information product (IP) process, 299
- Information quality (IQ)
- abductive process, 231
 - accessibility, 226
 - consistency, 230–231
 - context quality, 6–7
 - contextual knowledge, 7
 - definitions, 8–10
 - function of, 228–229
 - higher level quality, 18
 - information content
 - accessibility, 10, 11
 - availability, 10–11
 - imprecision, 12–13
 - integrity/lack of imperfection, 12–13
 - relevance, 11–12
 - timeliness, 11
 - information flow, 7–8
 - information presentation, 16–18
 - information sources
 - objective sources, 14
 - quality ontology, 14
 - sensors, models, automated processes, 13–14
 - subjective sources, 14–16
 - inter-and intra-processing, 6
 - JDL model, 5–6
 - key-value models, 231–232
 - logic-based models, 232–233
 - objective quality, 223–225
 - observation and fusion results, 229
 - Omnibus Model, 6
 - ontology-based models, 232
 - overall quality measure, 22–23
- problem variables, 229–230
- quality attributes, 18
- information flow, 21–22
 - models and processed outputs, 19–20
 - priori domain knowledge, 19
 - provenance, 21
 - reliability, 21–22
 - subjective quality evaluation, 20
 - time-dependent threshold, 19
 - trust evaluation, 20–22
 - users and process designers, 20
- quality control, 23–25
- recognition result, 225–226
- relevance, 227–228
- reliability, 227
- selection of variables, 229–230
- subjective quality, 224–225
- time delay, 231
- trust, 226–227
- Information Quality Assessment in the Context of Emergency Situational Awareness (IQESA), 570, 571
- Information quality elements (IQEs)
- algorithm-information pairing, 496–497
 - characterization, 489
- HILF SoS
- dynamic nature of information, 492–494
 - externalities considerations, 495–496
 - maturity, 501–502
 - real-time and streaming information, 494–495
 - static nature of information, 491–492
 - quality attributes, 490–491
 - vectors, taxonomies and ontologies, 497–498
- Information quality meta-model (IQMM), 487, 488, 513
- Information representation, 572
- Information retrieval (IR), 315
- average precision, 316–319
 - vs. ER quality, 319
 - implementation, 320
 - MAP, 316–319
- Input-process-output (IPO) view, 479
- Intelligent vehicle, 425, 427
- International Council on Clean Transportation (ICCT), 192
- International Organization for Standardization (ISO), 410–411
- International Patent Classification (IPC), 521

- International Statistical Classification of Diseases and Related Health Problems-10th revision (ICD-10), 458, 460
- Intrinsic information quality dimensions, 413
- IQ, *see* Information quality
- Iterated filtering approach, 112
- J**
- Jaccard dissimilarity, 83
- Joint Architecture Study (JAS), 496
- Joint Directors of Laboratories (JDL) model, 5–6, 428
- K**
- Key-value models, 231–232
- L**
- Left hypothesis (LH), 435, 447
- Left lane marking (LM), 429–431
- Logic-based models, 232–233
- M**
- MACRO, *see* Multilayer attribute-based conflict-reducing observation
- Mass assignment, 100–101
- Mass media reports, 196
- Master data management (MDM), 472
- Master patient index (MPI), 303
- Matching rules, 301
- MATLAB implementations, 550
- MCDAs, *see* Multi-Criteria Decision-Aiding methods
- MCDM, *see* Multi-Criteria Decision-Making
- Mean average precision (MAP), 316–318
- Measures of performance (MOPs), 510
- Measures of uncertainty, 101–104
- Mediated perception, 427–428
- Medical encoding systems
- code analysis, 462
 - expert coders, 457
 - HIS, 457–459
 - information fusion, 463–465
 - inputs and standards, 458, 460
 - integration, 456
 - laboratory results analysis, 461
 - outputs, 460–461
 - physicians code, 457
- probabilistic analysis, 462
- quality evaluation
- code list, 467–469
 - process quality, 466–467
 - tools, 456
- Medical unit discharge summary, 458, 460
- Meta-data, 9
- Meta-knowledge, 32
- lack of truthfulness, 37
 - meta-independent sources, 42–45
 - multiple source, 41–42
 - ordered collection, 46
 - prior information, 46
 - single source, 38–41
 - source testimony, 46
 - uncertain meta-knowledge, 47
- Minimum description length (MDL), 508–509
- Mitigation techniques, information quality, 498–500
- Model Evaluation, Selection and Application (MESA), 368
- Modification of AM (MAM), 103
- Modified-Fuzzy-Pattern-Classifier (MFPC), 546
- Monte-Carlo methods, 261
- Moore-Penrose inverse regression, 443
- Moses Illusion, 184
- Moving target indicators (MOVINT), 494
- Multicriteria association, 572
- Multi-Criteria Decision-Aiding methods (MCDAs), 330, 331
- Multicriteria decision analysis, 332
- Multi-Criteria Decision-Making (MCDM), 330
- Multilayer attribute-based conflict-reducing observation (MACRO)
- architecture, 545
 - BalTLCS, 546, 547
 - definition, 544
 - fuzzy set theory, 546
 - OWA operator, 547
 - signal conditioning, 546
- Multi-source fusion process
- average conflict, 63
 - credibility function, 63–64
 - degree of conflict, 63
 - fused distributions, 66
 - pairwise fusions, 65–66
 - potential spatial locations, 65
 - quality issues, 62
 - Takagi-Sugeno approach, 65
- Muslim Cyber Army, 201
- Mutual information (MI), 433–436, 446

N

National Cyber and Encryption Agency (BSSN), 200

Natural language

- aspects, 160
- credibility weighting
 - chained and modified evidential markers, 175–176
 - cross-linguistic comparison, 172–173
 - example of, 176–178
 - informed decision-making, 173
 - labels and expressions, 174
 - lexical items, 173–174
 - maximum uncertainty, 174–175
 - negation, 176
 - speakers of English, 172
 - text analytics, 171–172

device-derived data, 157

elements, 161

human as sensor

- hearsay, 159
- hidden networks, 159–160
- intelligence communities, 160
- intention, 158
- open sources, 159
- opinions, 158–159
- subjectivity, 158

human-derived data, 157

imprecise/vague descriptions, 161

incomplete and uncertain information, 156

sentence-level uncertainty, 161–162

uncertainty about content, 166

uncertainty within content

- ambiguity, 164
- code switching, 165–166
- imprecision and vagueness, 162–163
- polysemy, 164
- speakers of English, 164–165

words of estimative probability

- evidential markers, 168–170
- hedges, 168
- Kent’s observations, 167
- passive voice and impersonal phrasing, 170–171
- time, 170–171

United States Central Intelligence Agency, 166–167

US intelligence community, 167

Neighbour matrix, 208–209

Newman’s new model, 215

Noise (N), 467, 468

Non-boolean variables and assignments, 259–260

Non-specificity, 101–102

O

Objective quality, 9

Observe-orient-decide-act (OODA), 6, 486

Omnibus Model, 6

Ontology-based models, 232

Ordered weighted averaging (OWA), 543

P

PageRank, 520

colors encode clusters, 535–536

computational method, 529

damping factor parameter, 526–527

history, 527–528

vs. reinforcement learning, 533–535

Pairwise matching, 313

Partial conflict, 81, 89

Patent citation network analysis

Penning’s method

accuracy, 315–316

F-measure, 316

information retrieval, 315

average precision, 316–319

vs. ER quality, 319

implementation, 320

MAP, 316–319

pseudo-truth sets, 315

Talburt-Wang Index, 316

Pignistic probability, 80, 87

Possibility theory, 336

Practice-oriented approach, 426

Precision measures, 308, 314, 463

Pre-match blocking, 301

Probabilistic C-Table (PC-Table), 258

“Probabilistic” matching, 302

Probabilistic query processing (PQP), 247

Probability distributions

Cross entropy, 57

Gini entropy, 55–57

maximally certain distribution, 58–62

multi-source fusion process

average conflict, 63

credibility function, 63–64

degree of conflict, 63

fused distributions, 66

pairwise fusions, 65–66

potential spatial locations, 65

quality issues, 62

Takagi-Sugeno approach, 65

Shannon entropy, 55

uncertain distributions, 58–62

unequally weighted fusions, 71–76

vector representation

degree of compatibility, 54–55

- dot/inner product, 52–53
Euclidean length, 52–53
n-dimensional vector, 52
two-dimensional case, 53–54
- on weighted average fusion
associated fused value, 67
credibility function, 68–70
dominance combinations, 67–69
non-dominated fusions, 67–69
prioritized aggregation, 70–71
Zadeh’s idea of computing, 70
- Proportional Conflict Redistribution (PCR)
rules, 338
- Public key infrastructure (PKI), 496
- Pullen’s stratified sampling method
attribute weights, 310
negative outcomes, 312–314
overview, 309–310
positive outcomes, 310–312
precision, recall, and *F*-measure, 314
- Q**
- Quality assessment
acquisition-decision process, 403
control levels, 405
fusion, 408–409
GIS, 409–411
information quality, 406–408
reasons, 403
SST, 404, 405
types, 405
- Quality assurance (QA), 321
blocking, 322
data rationalization, 323
ER process, 322–323
source information, 322
- Quality attributes (QAs), 490–491
- Quality-aware Human-driven Information Fusion Model (Quantify), 565–566, 568
- Quality control (QC), 321
- Quality evaluation
code list, 467–469
process quality, 466–467
techniques, 406, 407
- Quality indicator, 339
- Quality of context
abductive process, 231
consistency, 230–231
definition, 228
function of, 228–229
observation and fusion results, 229
problem variables, 229–230
- selection of variables, 229–230
time delay, 231
- Quality of information sources
meta-knowledge, 32
lack of truthfulness, 37
meta-independent sources, 42–45
multiple source, 41–42
ordered collection, 46
prior information, 46
single source, 38–41
source testimony, 46
uncertain meta-knowledge, 47
- relevance and truthfulness
lack of truthfulness, 32
multiple source, 35–37
single source, 33–34
- Quality transfer function, 466
- R**
- Rand index, 315–316
- Random graph model, 210
- Ranking algorithms
damping factor, 529–531
microscopic and mesoscopic level studies, 520–521
- PageRank, 520
computational method, 529
damping factor parameter, 526–527
history, 527–528
- reinforcement learning
algorithm, 532–533
DAG, 531
vs. PageRank, 533–535
- USPTO database
birth mechanism, 525
death mechanism, 525–526
decay mechanism, 523, 524
growth mechanism, 523, 524
information quality, 522
IPC system, 521–522
merge mechanism, 525–526
methods, 522–523
patent classes, temporal behavior, 522
split mechanism, 523, 525
- Recall measures, 308, 314
- Rectified linear units (ReLU), 437, 443
- Reference codes, 463
- Regional aerosol transport models, 405
- Reinforcement learning (RL)
algorithm, 532–533
DAG, 531
vs. PageRank, 533–535
- Relational algebra, 262–263

- Relationship consistency, 383–387
 Remote sensing
 comprehensive quality model
 challenge, 411–412
 characterization, 412–413
 cyber-physical systems, 412
 information quality, 413–415
 quality assessment (*see* Quality assessment)
 types, 401–402
 Responsible individual (RI), 510–511
 Revision of the JDL model II, 6
 Right hypothesis (LH), 435
 Right lane marking (RM), 429–431
 RL, *see* Reinforcement learning
 Road model, 427–428
 Root mean square error (RMSE), 118
- S**
- Scale-free graph model, 214–216
 SDM model, 283–285
 Sea surface temperature (SST), 404, 405
 Semantic illusion, 184
 Semantic information fusion, 573–574
 consistency and currentness, 582, 583
 DQV, 582
 ESAS, 582, 584
 multicriteria assessment, 581
 properties, 579
 reports, 580, 581
 risk analysis, 576–577
 Shafer’s exhaustivity assumption, 337
 Shannon entropy, 55, 101–104
 Signal conditioning (SC), 546
 Situation awareness (SAW)
 challenging, 564
 crime analysis, Brazil, 575–576
 data and information quality, 566–567
 human operator, 565
 Quantify model
 data acquisition, 569–570
 data and information quality
 assessment, 570–571
 features, 568–569
 semantic information fusion, 573–574
 consistency and currentness, 582, 583
 DQV, 582
 ESAS, 582, 584
 multicriteria assessment, 581
 properties, 579
 reports, 580, 581
 risk analysis, 576–577
 syntactic information fusion, 571–573
 reports, 577–578
- risk analysis, 576–577
 user interface, 574–575
 Social media, 185–186
 Social model, 392
 Social network analysis
 edges, 208–209
 growing networks, 214–216
 neighbour/contact matrix, 208
 random graphs, 210
 regular lattice, 209
 reliability, 216–217
 scale-free graphs, 214–216
 small world effect, 210
 clustering coefficient, 211
 communications networks, 212–213
 Erdős numbers, 212
 graph’s properties, 213
 Internet, 212–213
 local clustering coefficient, 211
 Milgram’s experiment, 212
 Poissonian distribution, 213
 short maximum distances, 212–213
 Six Degrees of Separation, 212
 Social security number (SSN)
 bank’s loan database, 245–247
 conflicting records, 244
 data curation, 245
 high-quality data, 245
 PQP, 247
 technical challenges, 247
 SOUNDEX code, 301
 Source errors, 299–300
 SST, *see* Sea surface temperature
 STANAG 2511 model, 152–153
 Stochastic Gradient Descent (SGD), 438
 Stratification method, 311
 Sub-decadal data sets, 397
 Subjective logic (SL), 111, 115
 Supervision information, 588–589
 Supervisory control and data acquisition
 (SCADA), 494
 Support Vector Machine (SVM), 557
 Syntactic information fusion
 crime emergency situation, 571
 data acquisition, 572
 information representation, 572
 multicriteria association, 572
 on-demand fusion, 573
 primary fusion, 573
 reports, 577–578
 risk analysis, 576–577
 search for synergistic information, 572
 syntactic fusion, 572–573

- Syrian American Medical Society (SAMS), 195–196
Syrian Arab News Agency (SANA), 196
Syrian Observatory for Human Rights (SOHR), 195
System of systems (SoS), *see* HLIF system of systems (SoS)
System readiness level (SRL), 501
Systems architecture (SA), 475, 477
Systems engineering (SE), 475, 477
- T**
Takagi-Sugeno approach, 65
Talburt-Wang Index, 316
Tasking, collection, processing, exploitation, and dissemination (TCPED), 493
Technology readiness level (TRL), 501
Term Frequency-Inverse Document Frequency factor (TF-IDF), 136
Text Retrieval Conference (TREC), 317
Theory of evidence (TE)
 basic probability assignment, 100–101
 belief and plausibility, 101
 total uncertainty measure
 additivity, 105
 ambiguity measure, 103
 conflict and non-specificity, 101–102
 continuous function, 101
 Hartley generalized measure and upper entropy, 102–103
 modification of AM, 103
 monotonicity, 105–106
 probabilistic consistency, 104
 range, 105
 requirements, 106
 set consistency, 104
 subadditivity, 105
Theory of Record Linking, 305
Threshold satisfaction, 20
Torrential protective actions
 BF-TOPSIS, 351–352
 DMPs, 348–349
 ER-MCDA, 349–351
 FCOWA-ER, 352–353
Total quality management (TQM), 320
TotalRank, 530
Total uncertainty (TU) measure
 additivity, 105
 ambiguity measure, 103
 conflict and non-specificity, 101–102
 continuous function, 101
- Hartley generalized measure and upper entropy, 102–103
modification of AM, 103
monotonicity, 105–106
probabilistic consistency, 104
range, 105
requirements, 106
set consistency, 104
subadditivity, 105
Transferable Belief Model (TBM), 13
Transitivity of marginalisation, 44–45
TrustServista, 187
TruthFinder, 113
Twitter, 189, 201
Two-layer conflict solving (TLCS)
 adapted conflicting factor, 548
 balancing, 551–553
 conflict relaxation parameter, 550–551
 properties, 547
 relaxed maximum conflict, 548–549
- U**
Uncertain data
 attribute-level statistics, 275
 extensional evaluation, 264–265
 incomplete database
 certain records, 249–250
 multisets, 250–251
 possible records, 250
 possible world semantics, 248–249
 intensional evaluation, 266–268
 lossless encodings
 C-Tables model, 257–260
 disjoint-independent database, 255–257
 tuple-independent database, 253–255
 U-Relations, 260
 world-set decompositions, 260–261
 missing value, 252
 possible worlds
 record identifiers, 273–274
 representational features, 272–273
 sample filtering, 274
 threshold filtering, 274
 top-k posterior, 275
 top-k prior filtering, 274–275
 probabilistic database, 251–252
 record-level statistics, 275
 relational algebra, 262–263
 sampling-based encodings
 definition, 261
 Monte-Carlo methods, 261

Uncertain data (*cont.*)
 tuple bundles, 262
 world-annotated sample sets, 261

SSN
 bank's loan database, 245–247
 conflicting records, 244
 data curation, 245
 high-quality data, 245
 PQP, 247
 technical challenges, 247

tuple identity, 271–272

VC-Tables, 268–271

Uncertain distributions, 58–62

Unequally weighted fusions, 71–76

Union of conjunctive queries (UCQ), 267

United States Patent and Trademark Office (USPTO) database
 birth mechanism, 525
 death mechanism, 525–526
 decay mechanism, 523, 524
 growth mechanism, 523, 524
 information quality, 522
 IPC system, 521–522
 merge mechanism, 525–526
 methods, 522–523
 split mechanism, 523, 525
 temporal behavior, 522

Unreliable sources
 Bayesian modeling, 112–113, 115–116
 behavior estimates (*see* Behavior models)
 belief theories, 115
 BRS, 112
 Dirichlet distribution, 114–115
 EM algorithm, 113
 fabrication, 114
 fact-finding approach, 113
 fusion of information, 111–112
 fusion problem, 117–118
 quality of information, 115
 source behaviors, types, 113–114
 source estimation, 116–117
 subjective logic, 111, 115
 TRAVOS, 112
 uncertainty, 111, 113

US Environmental Protection Agency (EPA), 192

User interface (UI), 574–575

V
 Value cluster, 313
 Value harmonization, 308
 Variable generating (VG-Terms), 270
 Variable-generating relational algebra (VG-RA), 270
 Vector Model, 136
 Vehicle hypothesis (VH), 435, 447
 Verification and validation (V&V), 364–365
See also Data analysis process (DAP)

Virtual C-Tables (VC-Tables), 268–271

“Voting” method, 307

W
 Weight-based fusion (WBF), 439
 Weighted average fusion
 associated fused value, 67
 credibility function, 68–70
 dominance combinations, 67–69
 non-dominated fusions, 67–69
 prioritized aggregation, 70–71
 Zadeh’s idea of computing, 70

Wikileaks, 193–194

Winner-take-all (WTA), 439

Words of estimative probability
 evidential markers, 168–170
 hedges, 168
 Kent’s observations, 167
 passive voice and impersonal phrasing, 170–171
 time, 170–171
 United States Central Intelligence Agency, 166–167
 US intelligence community, 167

World Intellectual Property Organization (WIPO), 521

World-set decompositions, 260–261

Y
 Yager’s approach, 143–144

Z
 Zetta Cloud, 187, 198