DAVIDE CEOLIN

TRUSTING SEMI-STRUCTURED WEB DATA

SIKS Dissertation Series No. 2014-24

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Promotiecommissie:
prof.dr. A. Th. Schreiber (promotor)
prof.dr. W. J. Fokkink (promotor)
dr. W. R. van Hage (copromotor)
dr. T. Bosse
prof.dr. J.-J. Meyer
prof.dr. L. Moreau
dr. K. O'Hara

ISBN 978-94-6259-180-6

Cover design by Andrea Morando & Alessandra Piovesan `http://a2work.it`

VRIJE UNIVERSITEIT

TRUSTING SEMI-STRUCTURED WEB DATA

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op woensdag 2 juli 2014 om 11.45 uur
in het auditorium van de universiteit,
De Boelelaan 1105

door

Davide Ceolin

geboren te San Dona di Piave, Italië

promotoren:	prof.dr. A.Th. Schreiber
	prof.dr. W.J. Fokkink
copromotor:	dr. W.R. van Hage

Contents

Preface

I like to think about my past years as of a journey that started when I left my tiny hometown in Italy heading to Amsterdam, on a cold February morning. My parents Maurizio and Milena and my brother Alberto, helped this to happen and supported it all the way, both in person and remotely. I would like to express my highest gratitude to them.

This journey allowed me to touch the realms of computer science, of statistics and of philosophy until I could dare to call myself researcher. If that has been possible, much is due to three people that I had the privilege to be supervised by. Wan Fokkink has been admiringly present and available, and I am sure that without all the discussions we had and all the suggestions he gave me, I would never have been able to achieve many results I obtained. I hope I have learnt at least a little bit from this competence, patience and determination. Willem van Hage took me by hand since the first day I arrive at the VU, and guided me until I could stand on my feet. Beside all the fruitful discussions we had and all the interesting ideas we discussed, he helped me to become more pragmatic and this is something for which I owe him a lot. Of Guus Schreiber I have always admired the wisdom and the resolution, as well as his ability to set the scene for facilitating the research work. I will always remember the teachings he gave me, I consider them as gems of distilled wisdom. About the scene where the work here presented took place, I take the opportunity to thank the Poseidon project carried by the Embedded Systems Institute (ESI) for having supported a large part of this research. The Poseidon project was partially supported by the Dutch Ministry of Economic Affairs under the BSIK program.

Riste Gligorov started his doctoral studies a few months after me. We shared the same office for many years and the discussions, the jokes and chats we had are one of my best memories of this period. Credits are due to him for having coined the name "open world opinion" that I use later on in this thesis. He and Michiel Hildebrand supported the development of the case study based on the dataset from the *Waisda?* project. *Waisda?* has been launched by the Netherlands Institute for Sound and Vision and that work has been partly supported by the PrestoPRIME project, in the European Commission FP7 ICT program, Contract No. 231161.

Two names recur in my publications more than the others: those of Archana Nottamkandath and Paul Groth. Archana is a valuable colleague and I think we both earned a lot from our collaboration. I consider Paul as a precious point of reference,

not only in the provenance domain, where he is unquestionably an expert. Part of the work done with Archana and Paul is supported by the SEALINC Media and Data2Semantics projects respectively. These projects are part of the Dutch national program COMMIT. In the context of the SEALINC Media project I would also like to thank my colleagues Mieke Leyssen, Myriam Traub and Jacco van Ossenbruggen from CWI in Amsterdam and the VU for having performed an experiment that constituted the basis for several works here presented.

The entire Web & Media Group at the VU offered me the ideal background where to grow. Marieke van Erp deserves a particular mention because from her competence and obstinacy I could set the work presented in the first part of this thesis. The realization of that work has been allowed by the National Museum of Natural History in Leiden, Naturalis, and by NWO in the CATCH programme, grant 640.004.801 that partly funded that research.

I would also like to thank Nigel Shadbolt, Luc Moreau and Kieron O'Hara. Nigel, by hosting me, allowed me to confront with another interesting, challenging and comforting environment at the University of Southampton. Luc has always been very open to me, and if I manage to harvest a lot from that period, much is due to his help. I started bothering Kieron a few months before going to the United Kingdom. The emails we exchanged and the discussions we had represent an incomparable source of inspiration for me. The work that I developed during my stay in England was supported in part under SOCIAM: The Theory and Practice of Social Machines; the SOCIAM Project is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/1.

In all this, I have left the last mention to the person that most of all allowed this journey to happen and to continue. My wife Valentina was with me in my car when I left Italy for The Netherlands, and we shared all the joys and the suffering of this experience in a place that has now become a second home. I can not count the times she helped me discussing, brainstorming, revising my writings, nor can I diminish the importance of her ability of putting up with my quirks and constantly encouraging me. Let life allow us to continue this adventure together.

Davide Ceolin

April 2014

Introduction

In this chapter we introduce the context of the research described throughout this thesis. Then, we present the four research questions addressed in the thesis, and we describe the methods and methodologies adopted to tackle them. We present an overview of the state of the art and we conclude with a list of the publications on which this thesis is based and a description of the thesis structure.

This chapter is based on an extended version of the paper entitled Trusting Web Data*, presented at the* Doctoral Symposium *at the* 10th Extended Semantic Web Conference (ESWC2013)*, held in Montpellier, France.*

Context

The Web represents an inestimable source of knowledge. Every day, huge volumes of data are published or are being created by users, for instance, by means of their tracked behavior or of their involvement in crowdsourcing activities. Can we really trust whatever piece of data we might collect on the Web? Maybe, but probably not. This problem may heavily affect the value of these data. There exist several well-known trusted sources, but not all Web data originate from them and Web users are not always aware of the trustworthiness of the sources they refer to. When we observe an unknown data source exposing potentially useful data, can we safely use them? A similar issue regards the use of data which are known to have been crowdsourced. If we want to be able to use that vast amount of data, we must first trust it.

The goal of this thesis is to describe research about methods and techniques that address this kind of problem. We restrict our attention to semi-structured data, meaning that we do not analyze the reliability and trustworthiness of pieces of free text, but rather want to determine the reliability of pieces of data which have, at least, some defined structure. The analyses and the methods that we propose make use of this knowledge about the data structures and their related meta-information at our disposal to estimate their trustworthiness. For this reason, we focus on specific domains where data are collected either from the crowd or from specific Web sources in order to address professional tasks.

Our case studies are situated in specific domains. One of these is the maritime do-

main. Here, we analyze two kinds of data: information about ships, as emerged from specific messages that ships mandatorily send to coast guard authorities, and information about piracy attacks that happened all around the world. These are two examples of data which are at the same time potentially useful and potentially unreliable (because the producers might have several reasons to disguise them). These investigations are made in the larger contexts of *Poseidon*[1] and *Poseidon Plus*[2] projects, two national Dutch research projects which researched on systems of systems suitable for the naval domain.

Another project where our research situates is the United Kingdom *SOCIAM: The Theory and Practice of Social Machines* project[3]. In the context of this project, we investigate how to estimate the reliability of open government data, in particular of police open data. Here, the reliability of these data represents the basis for deciding whether or not to trust them, and hence, it is considered as a component of the multifaceted belief that is trust.

Another domain that we focus on is professional tagging. Here, crowds of users, professionals and volunteers, annotate pictures, videos and other artifacts in order to provide useful information about their content, information that would not be available otherwise. Professional annotations are used for several important tasks by institutions collecting these artifacts, like classification, cataloging, and information retrieval, hence, it is important that these annotations are reliable. In this context, we collaborate with the *EU FP7 PrestoPRIME*[4] project, the *SEALINC Media* project, part of the COMMIT research program in The Netherlands, the *Netherlands Organization for Scientific Research (NWO) Catch*[5] programme and the *The Netherlands Centre for Biodiversity Naturalis (Naturalis)* [116] . These collaborations provide us with use cases where annotations in the video, cultural heritage and natural history domains are respectively collected. These domains, and these projects in particular, share the need for tools to assess the trustworthiness of the tags and annotations of their artifacts, although these present different characteristics: these might be pictures in the cultural heritage domain or video (and hence multi-layered) in the media domain. Moreover, all these use cases share, among each other and with the case studies we investigate in the maritime domain, the following peculiarities:

- the impossibility to determine the annotations reliability from an analysis of their content;
- the availability of meta-information about the data (e.g., the annotation author);

[1]The Poseidon project was carried out under the responsibility of the Embedded Systems Institute (ESI) in Eindhoven, The Netherlands. This project is partially supported by the Dutch Ministry of Economic Affairs, under the BSIK03021 program.

[2]The Poseidon Plus project is a follow-up project of the Poseidon project, carried out under the responsibility of the Embedded Systems Institute (ESI) in Eindhoven, The Netherlands. This project is partially supported by the Dutch Ministry of Economic Affairs, under the BSIK03021 program.

[3]The SOCIAM: The Theory and Practice of Social Machines is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/1.

[4]The PrestoPRIME project was funded by the European Commission under ICT FP7 (Seventh Framework Programme), Contract No. 231161.

[5]The CATCH programme, is funded by NWO, grant 640.004.801.

- the availability of at least some known structure in the data;
- the need for the institution that owns the data that they comply with some correctness requirements;
- the fact that these data are almost all categorical. Only in one case we analyze the trustworthiness of numerical data, but this represents an analysis of meta-trustworthiness, because these numerical values represent confidence scores resulting from a georeferencing process.

This explains why we make use of these apparently diverse use cases for tackling our research questions.

State of the Art

We present here a literature review about the topics tackled in this thesis. We start by defining trust based on the definitions we found in literature, properly adapted to our context. Then, we describe related work for each of the topics touched in this thesis.

The Oxford English Dictionary defines trust as:

> the firm belief in the reliability, truth and or ability of someone or something [128].

Such a definition has been borrowed and formulated in more precise terms to adapt to different fields of computer science and artificial intelligence. We refer the reader to the three comprehensive surveys of Sabater and Sierra [139], Golbeck [64] and Artz and Gil [4] that regard trust in computer science, in the Web and in the Semantic Web respectively. Also the extensive overview presented by Thirunarayan et al. [152] represents an important point of reference about trust in computer science, in particular about Bayesian approaches like the ones we adopt in this thesis.

In multi-agent systems, trust management is handled by means of models based on the interactions between agents. In particular, agents decide whether they can trust other agents by observing their behavior and, possibly, other evidence. After having inferred their reputation, they decide whether the reputation is high enough to trust an agent. This basic idea is then implemented and reinterpreted in different manners and nuances depending also on the specificity of the domain where it applies. In particular, we make use of the definition of trust of Castelfranchi and Falcone reported by Sabater and Sierra, that is:

> the decision that an agent x (trustor) takes to delegate a task to agent y (trustee) is based on a specific set of beliefs and goals, and this mental state is what we call trust.

In part, we reinterpret this definition, as we do not delegate a task to an agent, but we decide whether the task that has been delegated to a given agent (data creation) has been carried out satisfactorily enough. Or, otherwise, whether the piece of data that

we analyze adheres to specific conditions, that is, whether it is trustworthy. We borrow the definition of trust we make use of from the multi-agent systems domain because: (1) in our approach the reputation of an agent is crucial to assess the trustworthiness of the piece of data that she produced and we adopt a uniform approach to model both assessments (about the trustworthiness of the data and of the agent); (2) we decide whether to trust or not a piece of data based on an indirect observation of its metadata, and not on its content, and we see in this a similarity with what one does when she has to trust an agent based on indirect observations.

The above definition describes the meaning we give to our trust estimates. With respect to the process of trust management, we take inspiration from the theory defined by O'Hara [121], with some distinctions.

- We distinguish between a process of trustworthiness estimation and one of trust placement (or decision strategy). However, the targets of the trust and trustworthiness assessments might be either physical agents or pieces of data, as stated before.

- The theory proposed by O'Hara defines trustworthiness and trust as established in a context. We follow the same way of reasoning, but implicitly. Our trust estimates are made within well-defined areas, although the rules for the assessments are implicit. For instance, in the cultural heritage domain a given cultural heritage institution defines its own trust policies, which we aim to mimic with our models, but these are not explicit (neither is the context definition). Only when we evaluate the trustworthiness of cultural heritage collections annotations, we explicitly restrict the context of our assessments to the topic of the annotation.

- O'Hara, in his "Trustworthiness versus reliability" section states that "elaborating the distinction is not the purpose of this paper, so I shall not explore this thought in much detail, except to point out that the distinction is not a sharp one, and so one should expect the two concepts to blur into each other. It may be that trustworthiness is a species of reliability, in which case everyone/thing that is trustworthy is ipso facto reliable". In this thesis, since we focus mainly on the estimation of the trustworthiness of data, we consider their reliability as an essential and tightly related (if not coincidental) characteristics of their trustworthiness.

Related to this latter point, Camp [16] discerns trust in the following three overlapping facets:

Security. The act of disclosure and sharing of sensitive information directly implies that enough guarantees should be offered regarding the security level of these data. Security is a broad term. In this context it refers mainly to the intentional damage that the trustor may suffer from taking part in the process of trusting. This damage may be inflicted by either the trustee or a third party being able to somehow interfere in the process.

Privacy. While security focuses on intentional damage, privacy centers on unwanted disclosure of sensitive information beyond the boundaries of the trust process.

Reliability. Beyond the sharing needs, delegation plays a key role within trust. Delegation of tasks is needed when the trustor is not able, for many possible reasons, to deliver the task. These reasons may include the particular skills needed to deal with the tasks, or the workload implied, for instance. A reasonable belief in the trustee's reliability is therefore essential to allow the trustor to trust.

As stated before, we focus mainly on the reliability aspect of trust. Our goal is to correctly trust or distrust data that result from past interactions (e.g., ex-post analysis of tags created by users playing a video-tagging game), so given the point of view taken here, privacy and security become secondary aspects because these inhere mainly present interactions.

Related to the definition of trust, is the one of reputation. The Oxford English Dictionary defines reputation as:

> The beliefs or opinions that are generally held about someone or something. [129]

We specialize such a definition, because the opinions and beliefs we are interested in are those regarding the trustworthiness of people (e.g., data creators), behavioral stereotypes, etc. So, our definition of reputation can be outlined as follows.

> The beliefs or opinions that are generally held about someone's or something's trustworthiness. [129]

The reputations we make use of are general estimates. Depending on the case study analyzed, we can also specialize them in order, for instance, to determine someone's expertise about a given subject. Nevertheless, the reputation is meant to represent the estimate about the general trustworthiness of somebody or something.

We use uncertainty reasoning to make sense of the evidence at our disposal about the data to be trusted. We make use of subjective logic [83] because of its flexibility and its ability to cope with partial or uncertain data. Uncertainty reasoning techniques are often used to make trust assessments, like in the work of Fokoue et al. [57]. It is important to investigate further the possibility of representing these data by means of multiple layers of probabilities, because of their adequateness to deal with vast amounts of heterogenous data. Other approaches are possible as well, like the trust metrics for recommender systems collected by Massa and Avesani [108], the possibilistic approach proposed by Ceravolo and Fugazza [32] or the belief-based approaches proposed, for instance, by Richardson et al. [134], and Vu and Aberer [167]. We take inspiration from these latter examples, but we do not need simply to estimate one value representing the trustworthiness of a piece of data alone. Rather, we prefer to be able to estimate probability distributions of these trust values, to account for the uncertainty in these estimates and to be able to evaluate aggregated datasets. Thus, our approach differs from the one of Golbeck et al. [66], that use binary (Boolean) scale for trust values and

from the one of Guha et al. [70] and Kamvar et al. [89], that use binomial values (i.e., the probabilities of two mutually exclusive values, which range between zero and one). In fact, one of the shortcomings we identified is the analysis of the distribution of Web data. Thus we provide some analyses about the use of probability distributions for representing Web data and, consequently, representing the foundation for reasoning probabilistically over them.

To make trust estimates, we make use of the metadata at our disposal. These consist of provenance information, that is, recordings of how, by whom and when the data have been created. We start by studying the possibility of assessing the trustworthiness of data on the basis of the reputation of their author, hence, reputation systems are an important source of inspiration for our work. The works collected by Masum and Tovey [109] provide a remarkable overview about the subject. Then we extend the range of provenance metadata considered for our estimates. The link between provenance and trust, mentioned in the survey of Artz and Gil, has been explored by Golbeck [63] but mainly for addressing socio-related issues, while our focus is on the data trustworthiness estimation. We use provenance data to learn the peculiarities of trustworthy and non-trustworthy pieces of data, according to a given trustor. Bizer and Cyganiak [12], Hartig and Zhao [73] and Zaihrayeu et al. [182], use provenance and background information expressed as annotated or named graphs [18] to produce trust values. The provenance information we make use of is expressed in terms of provenance graphs (e.g., using the PROV Ontology [9]), and we use them as machine learning features for classifying the trustworthiness of artifacts. The same difference is valid also with respect to two works of Rajbhandari et al. [131, 130], where they quantify the trustworthiness of scientific workflows and they evaluate it by means of probabilistic and fuzzy models. Despite the work of Ebden et al. [49], we do not make network analysis on the provenance graphs we make use of, but when we deal with provenance graphs that represent the behavior of the users who produced such data, we group them in "stereotypes" (i.e., aggregations of provenance information that summarize user behaviors) and we aim to estimate the data trustworthiness from the reputation of these stereotypes, by means of support vector machines [38] and subjective logic. By using provenance for trust estimation we aim at overcoming the limitations of reputation-based approaches to be able to estimate the trustworthiness of data even if the reputation of their author is unknown or uncertain. Also, by adopting provenance stereotypes, we address a limitation we identified in the use of provenance for making trust estimation, that is the impossibility to identify patterns in the provenance traces because of the fine granularity of such traces.

Lastly, in our analyses, we employ semantic similarity measures to improve our estimations. On the one hand, we can not tell whether a piece of data is trustworthy on the basis of its content. On the other hand, we can use semantic similarity measures to measure the semantic distance between contents. The link between trust and semantic similarity measures has already been explored, for instance by Ibrahim et al. [77], who use semantic similarity measures to estimate the trustworthiness of websites, by Sensoy et al. [142] who combine semantic similarity measures with subjective logic to model the trustworthiness of information sources within specific contexts and by

Tavakolifard et al. [149], who infer new trust connections between entities (users, etc.) given a set of trust connections known a priori. We explore this link further, by integrating more the two techniques, for example extending subjective logic with semantic similarity measures and providing a theoretical definition and demonstration of such an extension. The connection between uncertainty reasoning techniques with semantic similarity measures for trust estimation is one of the gaps we have identified in the literature and that we have aimed at filling.

We employ subjective logic in combination with semantic similarity to evaluate the trustworthiness of crowdsourced cultural heritage annotations. Crowdsourcing techniques are widely used by cultural heritage and multimedia institutions for enhancing the available information about their collections. Examples include the Tag Your Paintings project [51], the Steve.Museum project [155] and the Waisda? video tagging platform [119]. The Socially Enriched Access to Linked Cultural (SEALINC) Media project investigates also in this direction. In this project, Rijksmuseum [135] in Amsterdam is using crowdsourcing on a Web platform selecting experts of various domains to enrich information about their collection. Trust management in crowdsourced systems often employs classical wisdom of crowds approaches [147]. In our scenarios we can not make use of those approaches because the level of expertise needed to annotate artifacts in the domains we consider (e.g., the cultural heritage domain) restricts the potential set of users involved, thus making this kind of approach inapplicable or less effective. Gamification is another approach that leads to an improvement of the quality of tags gathered from crowds, as shown, for instance, in the work of von Ahn et al. [165]. Our work can be considered orthogonal to a gamified environment, as it allows us to evaluate the user contributed annotations and, hence, to incentivize them. In folksonomy systems such as Steve.Museum project, traditional tag evaluation techniques such as comparing the presence of the tags in standard vocabularies and thesauri, determining their frequency and their popularity or agreement with other tags (see, for instance, the work of Van Damme et al. [157]) have been employed to determine the quality of tags entered by users. Such mechanisms focus mainly on the contributed content with little or no reference to the user who authored it. Medeylan et al. [110] present algorithms to determine the quality of tags entered by users in a collaboratively created folksonomy, and apply them to the dataset CiteULike [35], which consists of text documents. They evaluate the relevance of user-provided tags by means of text document-based metrics. In our work, since we evaluate tags, we can not apply document-based metrics, and since we do not have at our disposal large amounts of tags per subject, we can not check for consistency among users tagging the same image. Similarly, we can not compute semantic similarity based on the available annotations, like in the work of Cattuto et al. [19]. In open collaborative sites such as Wikipedia [176], where information is contributed by Web users, automated quality evaluation mechanisms have been investigated (see, for instance, the work of De La Calzada et al. [41]). Most of these mechanisms involve computing trust from article revision history and user groups (see the works of Zeng et al. [183] and Wang et al. [172]). These algorithms track the changes that a particular article or piece of text has undergone over time, along with details of the users performing the changes. In our

case studies, we do not have the revision history for the tags. Another approach to obtain trustworthy data is to find experts amongst Web users with good motivation and intentions (see the work of De Martini et al. [42]). This mechanism assumes that users who are experts tend to provide more trustworthy annotations. It aims at identifying such experts, by analyzing the profiles built by tracking users' performance, similar to what we do (especially in the media and cultural heritage domains), although we do not look for experts interactively, rather we evaluate annotations ex-post. Modeling of reputation and user behavior on the Web is a widely studied domain. Javanmardi et al. [82] propose three computational models for user reputation by extracting detailed user edit patterns and statistics which are particularly tailored for wikis, while we focus on the annotations domain. Both we and Lange and Lange [99] address the quality of crowdsourced annotations, but we focus on annotations of professional media (and cultural heritage artifacts in particular), while Lange and Lange assess the quality of product ratings, that are more constrained and structured than the data that we deal with. Finally, also the works of Aroyo and Welty [3], Inel et al. [78] and Soberon et al. [145] address the problem of assessing the quality of crowdsourcing tasks, of microtasks in particular. Despite them, however, we do not deal with natural language processing tasks and we do have a gold standard at our disposal, and these are two remarkable differences. Nevertheless, our work can be considered as complementary to theirs and the results they provide may act as a viaticum for applying our research in the natural processing domain.

Additional specific literature references will be provided in each separate chapter.

Research Questions and Thesis Statement

Here we present the research questions we investigate in order to address the problems mentioned before and the approaches we propose to address them. The key focus of this thesis can be summarized by the following research question:

> *How can the trustworthiness of semi-structured Web data be adequately estimated?*

Our position with respect to this research question can be summarized by the following thesis statement:

> The trustworthiness of semi-structured Web data can be adequately estimated by making use of uncertainty reasoning, possibly assisted by provenance analysis and semantic similarity measures.

So, we hypothesize that trustworthiness estimations for semi-structured Web data can be effectively obtained by applying uncertainty reasoning, possibly assisted by provenance analysis and semantic similarity measures. This leads us to the following four research questions, that are aimed at investigating different aspects of this hypothesis: data, metadata and reasoning techniques useful to make adequate trust estimates.

Research Question 1

The first problem that we focus on is the usage of trusted semi-structured Web data to make trust evaluations of semi-structured data, not necessarily coming from Web sources. This gives a first insight into the possibility of using Web data for assessing the trustworthiness of data. So, the first research question is:

> *Can Web data help the trust evaluation of semi-structured data?*

We propose a quantitative empirical approach for this research question, by using uncertainty reasoning to make sense of Web data to trust unknown data. This has merely explorative goals (proving the possibility of using Web data to make trust assessments).

The Naturalis Museum in The Netherlands holds a collection of annotated bird specimen, which includes information like the species these specimens belong to, and the authors of the annotations. These annotations are not fully trustworthy, either because of their inaccuracy or because of the obsolescence of the taxonomy. In Chapter 1 we map these annotations to trusted Semantic Web sources to check them and, based on a gold standard, we estimate their trustworthiness using subjective logic (see Chapter Preliminaries), that allows us to cope with uncertainty about the representativity of the sample observed. We use these trust values combined with a range of decision strategies to decide whether to trust the annotations, and we measure the accuracy of the different combinations. This approach is reprised and extended in Chapters 5, 6, 7 and 8.

Moreover, in Chapter 2, subjective logic is employed to aggregate a series of heuristics adopted to measure the trustworthiness of a series of confidence scores about the estimated geolocation of a collection of specimen also owned by Naturalis. We estimate the confidence in an (estimated) piece of data by means of uncertainty reasoning. These techniques are further explored in Chapter 4.

So, by tackling this research question, we also need to investigate the reasoning techniques necessary to properly handle Web data. By addressing the following research question, we tackle this topic more in depth.

Research Question 2

Web data present peculiar characteristics that have to be taken into account when using them to make trust evaluations. For instance, they are often accessed incrementally (e.g., by crawling) so we do not always know how representative the data that we observe are. Moreover, their reliability varies, and their source reputation is not always known. Proper reasoning techniques have to be employed to cope with this, and they will be investigated by addressing the following research question:

> *How can uncertainty reasoning be effectively used to estimate the trustworthiness of semi-structured data?*

The approach proposed for this question is quantitative and empirical, and aims at producing a description of how categorical Web data fit higher-order probability distributions.

In Chapter 3, we show how it is possible to provide extensions of subjective logic that serve the purpose of calculating the trustworthiness of Web data. In particular, these extensions comprise the possibility of using semantic similarity measures, partial observations and Dirichlet Processes within the logic, and its correctness is proven either theoretically or statistically. These extensions follow from the experiments introduced in Chapters 1 and 2, where we show the usefulness of subjective logic, and are aimed at further extending the applicability of subjective logic and its statistical foundations for the goals of this thesis. In fact these extensions are largely employed especially in Chapters 7 and 8. In Chapter 3 we also use second-order probability distributions and stochastic processes to model the data from the Linked Open Piracy dataset [163], which contains a partial collection of piracy attacks descriptions. We focus on categorical data, which are among the most popular kind of data on the Web (URI). We model the data by means of Dirichlet-multinomial distributions and Dirichlet Processes, i.e., higher-order probabilistic models for categorical data (see Chapter Preliminaries for further details about these models), and we compare their ability to cope with the lack of a full view on the data with multinomial probability distributions based on the evidence at our disposal.

In Chapter 4 we provide two procedures based on statistical techniques and subjective logic to estimate the reliability of police open data. The first procedure is aimed at measuring the quality of the open data given the corresponding closed data, from different points of view by means of different tests (for instance, by measuring the absolute and the relative error in the open data). The second procedure is aimed at predicting the reliability of open data based on analysis of open data only, and it achieves this result by running different statistical tests on consecutive open datasets, and by aggregating the results obtained by means of subjective opinions. Both procedures take advantage of the use of uncertainty reasoning techniques adopted in Chapters 1 and 2. The second procedure handles the tests on the open data similar to the georeferencing heuristics analyzed in Chapter 2.

In all the other chapters, uncertainty reasoning is used extensively, either in terms of subjective logic, or of statistical reasoning and machine learning, or both.

Research Question 3

The Web also offers a meta-level of data-related information that is useful when dealing with trust, namely provenance information, that represents by whom and how data have been produced, manipulated and exposed. Reasoning over data providing this kind of information is important because this can provide indirect evidence about the reliability of a target object. Moreover, in general, this kind of data possibly enlarges our availability of reliable sources of evidence. This subject will be explored by addressing the following research question:

How can provenance information be used for making accurate trustworthiness estimations of semi-structured data?

This research question is tackled empirically. We analyze the relationships between provenance information and the trustworthiness of pieces of data in two different manners.

First, we build a Bayesian network using subjective logic on top of provenance graphs, to derive a trust value for a data artifact from the analysis of how it has been produced, thus extending the model introduced in Chapter 1. This is validated over a set of messages (AIS) sent by ships to coast guard authorities to communicate mandatory information (e.g., their nationality), for which we compute the corresponding trust values. This is described in Chapter 5.

Second, in Chapter 6, we use machine learning techniques to make trust predictions based on the provenance graph of the target artifacts. In particular, we predict the trustworthiness of a collection of video tags provided by the gaming platform *Waisda?* [119] using support vector machines [38] and we combine this prediction with one reputation-based (that is computed using a model similar to the one introduced in Chapter 1). Accuracy, precision and recall of the predictions are computed.

Third, in Chapter 8, we use uncertainty reasoning to assess the trustworthiness level of cultural heritage annotations based on their "provenance stereotypes", that are aggregations of provenance information that summarize a given behavior of the user who generated them. The model adopted in Chapter 8 represents an extension of the one presented in Chapter 7, that addresses the following research question.

Research Question 4

Also the Web as such can be exploited for the computation of meta-information that facilitates the estimation of trust values. Web-based semantic similarity measures can be used to weigh data and metadata at the disposal of the uncertainty reasoning techniques adopted to estimate the trustworthiness of a given subject, hence the following research question:

Can semantic similarity measures improve the accuracy of trust estimates of semi-structured data based on uncertainty reasoning?

We employ a quantitative approach to tackle this question. Before doing so, we adopt a theoretical approach to incorporate semantic similarity measures in uncertainty reasoning techniques.

Semantic similarity measures (in particular, the Lin [104] and the Wu & Palmer similarity [178], see also Chapter Preliminaries) are used to improve the precision of the uncertainty reasoning techniques adopted for trustworthiness estimation. The advantage of combining these two techniques is twofold: first, their combination allows us to improve the accuracy of the estimation of trustworthiness; second, they allow to make trust assessments on probabilistic bases, that is, without the need to set an arbitrary threshold for the trustworthiness levels of artifacts. A method to decide

whether or not to trust something is to set a value (threshold); we trust only artifacts having a trustworthiness level higher than that. Such a value is often arbitrarily set; the probabilistic approach we propose avoids this arbitrariness. This method is evaluated against two datasets from the cultural heritage, as described in Chapters 7 and 8. The models proposed in these two chapters extend the models proposed in the previous chapters, especially in Chapter 1 and 6, and take advantage of the extensions of subjective logic introduced in Chapter 3.

Contributions

We outline here the main contributions presented in this thesis.

Procedures for Trustworthiness Estimation We provide a range of procedures for the estimation of the trustworthiness of semi-structured data. Starting from Chapters 1 and 2, where we provide a series of basic algorithms for applying uncertainty reasoning to estimate the trustworthiness of semi-structured Web data, we extend those procedures in order to fit with the use cases described in the other chapters and to incorporate advances earned along the way. The procedures introduced in Chapter 4 allow estimating the reliability of police open data by making use of a large set of uncertainty reasoning analyses. The procedures proposed in Chapters 5, 6 and 8 combine uncertainty reasoning with provenance analysis, while in Chapters 7 and 8, we add the use of semantic similarity measures to weigh the evidence at our disposal. The novelty of these contributions resides in the systematization of the use of evidential reasoning in combination with Web data (provenance and semantic similarity measures in particular) for making trust assessments.

Web Data Modeling We provide a first description of Web data in terms of higher order probabilities in Chapter 3. This represents a novel approach for modeling Web data. This modeling supports our uncertainty reasoning choices, since the evidential reasoning techniques that we adopt (in particular, subjective logic), rely on these high order probabilities. Hence, if at least some Web data are correctly representable by means of these distributions, we can safely reason upon them.

Subjective Logic Extensions In Chapter 3 we propose three extensions of subjective logic tailored for Web and Semantic Web data handling. These extensions are the combination of subjective logic with semantic similarity measures, the possibility to handle partial observations and the so-called "open world opinions" that is, subjective opinion based on a partially defined amount of categories. These extensions allow covering issues that are specifically related to Web data, like the fact that these are accessed incrementally (partial evidence observations, open world opinions), and allow also taking advantage of useful information derived from the analysis of Web data, like semantic similarity measures. These extensions of subjective logic represent a novel

contribution. The combination of subjective logic with semantic similarity measures is used in Chapters 7 and 8.

Provenance Analysis for Trustworthiness Estimation We have already mentioned above that one of the contributions of this thesis is represented by a series of procedures for estimating the trustworthiness of semi-structured Web data. Some of these procedures make use of provenance analysis in combination with uncertainty reasoning and semantic similarity for producing the estimates.

Another contribution regards the so-called "provenance stereotypes". These are classes of user behaviors identified by discretizing and coarsening the information contained in provenance traces, in order to group traces and identify behavioral classes (e.g., early morning weekend annotators). This is another novel contribution and, although we aim at investigating it further in the future, Chapters 6 and 8 provide two examples of uses of provenance stereotypes as a basis for trustworthiness estimation.

Trusting Web Data Website The code that implements the procedures described above is collected at the website `http://trustingwebdata.org/phdthesis/dceolin`. We created the website `http://trustingwebdata.org` with the goal to collect software, publications and other kinds of resources about the topic of trusting (semi-structured) Web data, and we aim to contribute to it also in the future.

Publications

This thesis is based on the following publications.

- D. Ceolin, W. R. van Hage, and W. Fokkink. A Trust Model to Estimate the Quality of Annotations using the Web. In *WebSci10: Extending the Frontiers of Society On-Line (WebSci 2010)*. Web Science Trust, 2010

- D. Ceolin, P. Groth, and W. R. van Hage. Calculating the Trust of Event Descriptions using Provenance. In *Proceedings of the Second International Workshop on the role of Semantic Web in Provenance Management (SWPM 2010), co-located with the 9th International Semantic Web Conference (ISWC 2010)*, volume 670 of *CEUR Workshop Proceedings*, pages 11–16. CEUR-WS.org, 2010

- D. Ceolin, W. R. van Hage, W. Fokkink, and G. Schreiber. Estimating Uncertainty of Categorical Web Data. In *Proceedings of the 7th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2011), co-located with the 10th International Semantic Web Conference (ISWC 2011)*, volume 778 of *CEUR Workshop Proceedings*, pages 15–26. CEUR-WS.org, 2011

- D. Ceolin, A. Nottamkandath, and W. Fokkink. Automated Evaluation of Annotators for Museum Collections Using Subjective Logic. In *Proceedings of Trust Management VI - 6th IFIP WG 11.11 International Conference (IFIPTM 2012)*,

volume 374 of *IFIP Advances in Information and Communication Technology*, pages 232–239. Springer, 2012

- D. Ceolin, P. Groth, W. R. van Hage, A. Nottamkandath, and W. Fokkink. Trust Evaluation through User Reputation and Provenance Analysis. In *Proceedings of the 8th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2012) at the 11th International Semantic Web Conference (ISWC 2012)*, volume 900 of *CEUR Workshop Proceedings*, pages 15–26. CEUR-WS.org, 2012

- D. Ceolin, A. Nottamkandath, and W. Fokkink. Subjective Logic Extensions for the Semantic Web. In *Proceedings of the 8th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2012), co-located with the 11th International Semantic Web Conference (ISWC 2012)*, volume 900 of *CEUR Workshop Proceedings*, pages 27–38. CEUR-WS.org, 2012

- D. Ceolin, W. R. van Hage, G. Schreiber, and W. Fokkink. Assessing Trust for Determining the Reliability of Information. In *Situation Awareness with Systems of Systems*, pages 209–228. Springer, 2013

- D. Ceolin. Trusting Semi-structured Web Data. In *Proceedings of The Semantic Web: Semantics and Big Data, 10th International Conference (ESWC 2013)*, volume 8219 of *Lecture Notes in Computer Science*, pages 676–681. Springer, 2013

- D. Ceolin, A. Nottamkandath, and W. Fokkink. Semi-automated Assessment of Annotation Trustworthiness. In *Proceedings of the Eleventh Annual International Conference on Privacy, Security and Trust (PST 2013)*, pages 325–332. IEEE Computer Society, 2013

- D. Ceolin, L. Moreau, K. O'Hara, G. Schreiber, A. Sackley, W. Fokkink, W. R. van Hage, and N. Shadbolt. Reliability Analyses of Open Government Data. In *Proceedings of the 9th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2013), co-located with the 12th International Semantic Web Conference (ISWC 2013)*, volume 1073 of *CEUR Workshop Proceedings*, pages 34–39. CEUR-WS.org, 2013

- D. Ceolin, A. Nottamkandath, and W. Fokkink. Efficient Semi-automated Assessment of Annotations Trustworthiness. *Journal of Trust Management*, 2014. To appear

- D. Ceolin, L. Moreau, K. O'Hara, G. Schreiber, A. Sackley, W. Fokkink, W. R. van Hage, N. Shadbolt, and V. Maccatrozzo. Two Procedures for Analyzing the Reliability of Open Government Data. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2014)*. Springer, 2014. To appear

- M. van Erp, R. Hensel, D. Ceolin, and M. van der Meij. Georeferencing Animal Specimen Datasets. Accepted for publication in Transactions in GIS, John Wiley & Sons, Inc., 2014

Thesis Structure

This thesis is structured as follows. First, in Chapter Preliminaries, we introduce some techniques that have not been developed by us, but that are extensively used throughout the thesis, namely subjective logic, conjugate priors and semantic similarity measures. In Part I, Chapters 1 and 2 describe two preliminary works on the use of Web data to make trustworthiness estimations and provide some insights about which reasoning techniques to use and which metadata to rely on. Part II regards deeper investigations about the use of uncertainty reasoning for making trustworthiness estimations. In particular, Chapter 3 presents subjective logic extensions aimed at improving the handling of Web data and investigates the use of probability distributions to represent categorical Web data, while Chapter 4 presents two procedures for analyzing the reliability of Web data and applies them on police open data while dealing with the inner uncertainty of these estimates. Part III focuses on the use of provenance metadata as a basis for making trust assessments. Chapters 5 and 6 present two analyses of provenance using uncertainty reasoning techniques to estimate the trustworthiness of semi-structured Web data. In Part IV we focus on the use of semantic similarity measures for making trust assessments. In Chapter 7 we present applications of the combination of subjective logic and semantic similarity for making trust assertions of cultural heritage annotations through the estimation of the reputation of their author. Chapter 8 makes use of similar techniques in the same context, but based on the estimation of the reputation of provenance stereotypes. Lastly, Chapter Conclusion and Discussion provides a final discussion and indication of future work perspectives.

Preliminaries

This chapter introduces a series of definitions, methods and tools that will be extensively used throughout the rest of the thesis. This chapter aims to group descriptions and references of all the methods that are used in the thesis, but are prior to my research and were developed by others.

Introduction

The goal of this chapter is to lay the foundations for the rest of the thesis. We describe subjective logic, which is a probabilistic logic that constitutes one of the leitmotifs of this thesis, conjugate priors, that represent the statistical background of subjective logic, and semantic similarity, that is employed as a weighing factor for evaluating annotations using, again, subjective logic. The reason why this logic is so relevant for the purposes of the research presented here is at least threefold. First, this logic allows us to represent the truth value of propositions in probabilistic terms, and allows us also to account for uncertainty in the estimation of such a value. This is an important feature, because in a Web environment, we often estimate the truth value of a proposition based on small samples of observations that may be representative of the entire data population, but may also be misleading. Thus when estimating the truth value of a proposition based on evidence, subjective logic reserves part of the probability mass for the "uncertainty" value (that represents probability mass that is assigned neither to the *true* nor to the *false* value). As long as new evidence is collected, part of the uncertainty probability mass is assigned either to the probability of the *true* or of the *false* value, depending on the evidence collected (positive or negative pieces, respectively). The second reason why we employ this logic is that it allows us to keep track of the subject that made an assertion about the truth value of a proposition. This is important because we make trust estimations by using data being provided by several Web sources having different trust levels. Being able to track the provenance of a given assertion helps us to adequately estimate trust. As a consequence, data or proposition evaluations provided by unreliable sources can be tracked and their weight reduced when reasoning. Lastly, the logic is important for our needs because it offers us a wide range of operators for combining proposition arguments (called "opinions"). These operators comprise a probabilistic extension of Boolean logic operators that allow us to logically combine propositions along with their estimated probability to

be true. Moreover, the logic offers operators to weigh opinions provided by sources with different trust levels (so as to take into account also the trustworthiness of the source when using the opinion it provides) and operators to merge opinions provided by different sources, so as to aggregate opinions despite of their possible disagreement.

In particular, we refer to "crisp" propositions (as opposite to "vague" ones). We do not consider propositions that contain linguistically vague statements (such as "tall" or "hot"). We consider propositions reporting arbitrary crisp and exact measures about facts (such as how tall somebody is or how hot something is). In fact, despite fuzzy logic [181] that allows assigning graded truth values to propositions, subjective logic allows propositions to have only a Boolean truth value, i.e. to be either *true* or *false*. Nevertheless, it allows us to express the belief that somebody has about a proposition being true, and this belief can be graded and uncertain. A typical proposition we focus on is "Annotator *xyz* is trustworthy" as stated by an institution (e.g. museum) by relying on a series of observations about *xyz*. This has to be interpreted in a "crisp" manner, since if that proposition is true, the contributions of that annotator will be always accepted by the institution. Despite the fact that the interpretation of the proposition is not vague or fuzzy, the institution can estimate the truth value of that proposition in probabilistic terms (to express its belief in the proposition being true, its disbelief and its uncertainty). Based on this estimate, the institution can, for instance, decide to accept only part of the contributions of *xyz* (proportional to the belief in *xyz* being trustworthy) or accept contributions from *xyz* only if the belief of the institution in *xyz* is high enough (depending on the policy adopted by the institution).

The rest of this chapter is structured as follows. First we introduce subjective logic, its elements and the rationale behind them. Then, we describe the subjective operators that are relevant for the rest of this thesis. The third section regards the most relevant probability distributions employed in this thesis, together with an important property that connects them, that is conjugacy. These distributions are deeply analyzed in Chapter 3 and constitute the probabilistic foundation of subjective logic. Lastly, we introduce semantic similarity measures, that are mostly used in the last part of this thesis as a means to manipulate the evidence used in our statistical reasoning.

Subjective Logic

Subjective logic [83] is a type of probabilistic logic, that is, a formalism that combines Boolean logic with probability theory to express uncertainty about the truth value of proposition. This logic allows us to represent the fact that the truth value of propositions may be uncertain and that different sources may disagree about the truth value of a given proposition. Therefore, in subjective logic, arguments are represented by means of so-called "opinions" that are tuples composed by the belief owner (or "source"), the proposition (or "target" or "object") and the truth value assigned by the source to the proposition. Propositions are assumed to belong to a "frame of discernment" (or "frame" or "state space", represented, for instance, as X). All the propositions that belong to the same frame are assumed to be mutually disjoint and

exhaustive. Subjective opinions are represented as:

$$\omega_x^A \tag{1}$$

where ω is the generic symbol used to denote subjective opinions, A is the belief owner and x is a proposition. An alternative notation is:

$$\omega(A : x) \tag{2}$$

The first notation is the one we adopt throughout the thesis, since it is the one most commonly used. In some cases we want to represent also the context c in which the agent A expresses his opinion. We can represent this case as follows:

$$\omega_{x:c}^A \tag{3}$$

$$\omega(A : x : c) \tag{4}$$

It is also possible to represent a subjective opinion about a whole frame of discernment X, that is an opinion about the whole set of all the mutually exclusive and exhaustive propositions we focus on as:

$$\omega_X^A \tag{5}$$

Estimated truth values of propositions are expressed in terms of beliefs (owned by the source with respect to the proposition). Beliefs specify the distribution of the probability mass over the set

$$\mathcal{R}(X) = \mathcal{P}(X) \setminus \{\emptyset, X\}. \tag{6}$$

where $\mathcal{P}(X)$ represents the power set of X. $\mathcal{R}(X)$ is called "reduced power set of X". Beliefs are compatible with Dempster-Shafer belief assignment functions [43, 143], that are represented as $m(x)$. Since $m(\emptyset) = 0$, then $\emptyset$ is excluded from $\mathcal{R}(X)$. Also, beliefs are intended to represent the probability mass that is assigned to the plausible propositions. Since the frame of discernment is assumed to be exhaustive, it is not possible that some probability mass is assigned to none of the propositions belonging to the frame. X is excluded from $\mathcal{R}(X)$ because the probability mass to that particular element is represented by means of a specific value, called "uncertainty". The separation between belief and uncertainty allows us to represent the uncertainty in the opinion, and this representation facilitates a mapping between the opinion and a Beta (or Dirichlet) probability density function representation.

If $|X| > 2$ and beliefs are attributed only to singleton elements of X, opinions are called "multinomial". If beliefs are attributed to all the proper subsets of X, opinions are called "hyperopinions". We focus now on the case when $|X| = 2$, that is the case of binomial opinions that are the main type of opinion employed in this thesis.

The value of a subjective opinion that an agent A has about a single proposition x is represented as follows:

$$\omega_x^A = (b, d, u, a) \tag{7}$$

where b, d, u and a represent the belief, disbelief, uncertainty and a priori value (or "base rate" or "prior probability") owned by A with respect to x. For brevity, the opinion is represented also as

$$\omega_x^A(b, d, u, a) \tag{8}$$

or

$$\omega_y(b, d, u). \tag{9}$$

In these cases, we assume the source to be unknown or implicit, and the base rate a equal to $\frac{1}{2}$ which, as we will see, is a neutral value for this parameter. When referring to one of the belief elements individually, we can refer to them as indexed elements. For instance, as b_x^A.

The four numeric elements (b, d, u, a) are constrained as follows:

$$b \in [0, 1] \quad d \in [0, 1] \quad u \in [0, 1] \tag{10}$$

$$b + d + u = 1 \tag{11}$$

$$a \in [0, 1]. \tag{12}$$

The a priori value (or base rate) represents the prior probability that x owns about y, while belief and disbelief represent the probability mass that x attributes to y being true or false respectively. The uncertainty represents unassigned probability mass (due, indeed, to uncertainty motivated, for instance, by the fact that a given opinion is determined by observing little evidence).

Considering X, i.e. the whole frame x belongs to, we can also represent the opinion that A owns about all the propositions of X as:

$$\omega_X^A(\overrightarrow{b}, u, \overrightarrow{a}) \tag{13}$$

where beliefs are grouped in a vector ($\overrightarrow{b}$), as well as base rates ($\overrightarrow{a}$). There is one belief and one base rate for each proposition in the frame of discernment. In this case, the belief elements are constrained as follows:

$$\Sigma_i b_{x_i} + u = 1 \tag{14}$$

$$\Sigma_i a_{x_i} = 1 \tag{15}$$

The disbelief is not present in the opinion over the whole frame because it is a peculiarity of the opinions about single propositions: the disbelief of a single proposition corresponds to the sum of the beliefs in all the other propositions belonging to the same frame.

Let us illustrate this by means of an example. Suppose that a museum M wants to model and reason upon whether one of its annotator is trustworthy or not. Then, M can define a frame of discernment as:

$$X = \{x_1, x_2\} \tag{16}$$

where:

$$x_1 = \text{Davide is a trustworthy annotator} \tag{17}$$

$$x_2 = \text{Davide is an untrustworthy annotator} \tag{18}$$

These two propositions are exhaustive and mutually exclusive, since we assume that Davide can only be either trustworthy or not (propositions are interpreted as "crisp"; we will see later how it is possible to attribute partial belief in them). Hence, if we compute the opinion that M holds with respect to x_1, then we will have a belief, a disbelief, an uncertainty and a base rate. The same we will have if we compute an opinion on x_2 but the disbelief on x_1 actually corresponds to the belief in x_2 and vice versa. Therefore, if we compute the opinion on the whole frame, we do not have disbeliefs, but we have a vector of beliefs corresponding to the beliefs in each of the propositions that form the frame. In symbols:

$$\omega_{x_1}^{M}\left(b_{x_1}, d_{x_1}, u, a_{x_1}\right) \tag{19}$$

$$\omega_{x_2}^{M}\left(b_{x_2}, d_{x_2}, u, a_{x_2}\right) \tag{20}$$

Since

$$\Sigma_i b_{x_i} + u = 1 \tag{21}$$

$$b_{x_1} + d_{x_1} + u = 1 \quad b_{x_2} + d_{x_2} + u = 1 \tag{22}$$

then

$$b_{x_1} = d_{x_2} \quad b_{x_2} = d_{x_1} \tag{23}$$

so

$$\omega_X^M(\{b_{x_1}, b_{x_2}\}, u, \{a_{x_1}, a_{x_2}\}) \tag{24}$$

Subjective Logic Compatibility with Boolean Logic

If either $b = 1$ (and $d = u = 0$) or $d = 1$ (and $b = u = 0$), then the corresponding opinion represents the case when x has full belief or disbelief in y. In these cases, the subjective opinions are equivalent to Boolean propositions (see Table 1) and the logical operators of subjective logic behave, in this case, equivalently to Boolean logical operators (see Table 2 for an example about the conjunction operator).

ω_x^A	x
$\omega_x^A(1,0,0)$	1
$\omega_x^A(0,1,0)$	0

Table 1: Equivalence between Boolean logic and subjective logic

This fact is important for our needs because it allows us to rely on logical reasoning, when necessary. Throughout the thesis, we do make a rather limited use of this kind of reasoning. We employ it mainly in Chapters 1 and 5. Nevertheless, given that we

ω_y^x	ω_z^x	$\omega_{(y\wedge z)}^x$
$\omega_y^x(1,0,0)$	$\omega_z^x(1,0,0)$	$\omega_{(y\wedge z)}^x(1,0,0)$
$\omega_y^x(1,0,0)$	$\omega_z^x(0,1,0)$	$\omega_{(y\wedge z)}^x(0,1,0)$
$\omega_y^x(0,1,0)$	$\omega_z^x(1,0,0)$	$\omega_{(y\wedge z)}^x(0,1,0)$
$\omega_y^x(0,1,0)$	$\omega_z^x(0,1,0)$	$\omega_{(y\wedge z)}^x(0,1,0)$

x	y	$x \wedge y$
1	1	1
1	0	0
0	1	0
0	0	0

Table 2: Equivalence between the conjunction operator of subjective logic and the one of Boolean logic.

$P(x)$	$P(y)$	$P(x) \times P(y) = P(x \wedge y)$
0.4	0.3	$0.6 \times 0.3 = 0.12$
ω_x	ω_y	$\omega_{(x\wedge y)}$
$\omega_x(0.4, 0.6, 0)$	$\omega_y(0.3, 0.7, 0)$	$\omega_{(x\wedge y)}(0.6\times0.3, 0.6+0.7-0.6\times0.7, 0)=$ $\omega_{(x\wedge y)}(0.12, 0.88, 0)$

Table 3: Equivalence between the conjunction of propositions in probabilistic logic and in probability theory (above), and in subjective logic (below).

focus on semi-structured data, which comprise, for instance, also semantic web data, having the possibility to use this kind of reasoning may reveal to be useful to make use of logical and ontological [7] relations between data in future work.

Subjective Logic Compatibility with Probability Theory and Probabilistic Logic

In some cases, source A may want to express some uncertainty about the possibility of proposition x to be true. For instance, A may believe that x has 60% probability to be true (and hence 40% probability to be false). In subjective logic, such a fact is easily represented by assigning the corresponding values to b and d, respectively, as follows:

$$\omega_x^A(0.6, 0.4, 0) \tag{25}$$

This is equivalent with the following probability theory statements:

$$P(x) = 0.6 \quad P(\neg x) = 0.4 \tag{26}$$

This exemplifies the compatibility of subjective logic with probability theory and other probabilistic logic (see, for instance, the work of Nilsson [120] and of Hájek [2] for examples of other probabilistic logics). In this case subjective logic behaves like other probabilistic logic, as exemplified in Table 3.

This peculiarity of subjective logic is important for our needs because in a Web environment we have to deal with propositions which are believed to be true (up to

a certain extent) and that are neither fully believed nor fully disbelieved. Having the possibility to apply a sound reasoning over such kind of proposition is crucial for our needs because in many situations we are not able to derive Boolean truth values about propositions from Web data. In all the cases where the truth value of propositions can only be estimated probabilistically, we employ the power of probabilistic reasoning provided by subjective logic.

Peculiarities of Subjective Logic

We described in the previous sections how subjective logic is compatible with both Boolean and probabilistic logic. However, subjective logic extends Boolean and probabilistic logic in two manners. First, it keeps record of the belief owner, thus allowing different sources to provide different opinions on the same proposition, and allowing us to take into account the provenance of each opinion. Second, it allows to account for the uncertainty in the estimation of the probability of a given proposition to be true or false.

The first peculiarity is necessary to reason over data coming from different Web sources that present different reliability levels. By using the PROV Ontology [9], we can model the provenance of Web data, and subjective logic offers a means to make sense of these metadata by applying logical and evidential reasoning over them.

Also the second peculiarity is crucial in an open Web environment. In fact, we know that the amount of data available from the Web is huge, and we also know that often times we base our estimates on a small portion of these data. On the one hand, this is often our best choice to rely on, but on the other hand, these samples have to be carefully treated because they may not be representative of the entire data population. We saw in Section that subjective logic addresses this issue by assigning part of the probability mass to the uncertainty value. Here we describe more in detail how the probability mass is assigned to the possible truth values of a proposition given the evidence observed. Recall that the uncertainty value represents probability mass that is neither assigned to the *true* nor to the *false* value. Moreover, it is inversely proportional to the size of the sample set: as long as the set of observations grows, the uncertainty decreases, and vice versa. Also, the rest of the probability mass is still divided between belief (probability that the proposition is *true*) and disbelief (probability that the proposition is *false*), proportionally to the ratio between positive and negative evidence.

A proposition can be either true or false, and two propositions belonging to the same frame cannot be true at the same time, so the probability mass assigned to the uncertainty has to be divided between belief and disbelief in order to obtain an expected truth value for the proposition. This is done using the a priori value. This helps compensating with the possible lack of representativity of small samples: in these cases, in fact, part of the probability mass is assigned according to the composition of the evidence, but a relevant part (the one corresponding to the uncertainty), is allocated to belief and disbelief (to obtain the expected probability) according to the prior. If the sample is large, it is possible to safely rely on it, so the probability mass

assigned using the prior is very small.

$$E = b + a \cdot u \tag{27}$$

We claimed before that $\frac{1}{2}$ is a neutral value for the a priori value. Here we explain why this is the case: by setting the a priori value to $\frac{1}{2}$, we split equally the probability mass assigned to the uncertainty between belief and disbelief, with no bias towards one of them. If $a = 1$, then all the probability mass assigned to the uncertainty would be put in the expected probability (bias toward the belief), and vice versa if $a = 0$.

In the previous subsections we showed that subjective logic is compatible with Boolean and probabilistic logics. The examples analyzed there assume that the belief owner assigns his own truth value in a "dogmatic" way. In fact those opinions are called also "dogmatic opinions" since they do not imply any uncertainty in the probability value (or Boolean value) assigned to the opinion. There may be some uncertainty due to the fact that the truth value of a proposition is expressed in terms of a probability, but that probability is certain. This is the kind of certainty we are focusing on here.

In some situations, the belief owner is not in a position to express a dogmatic opinion about a proposition. Rather, he can only estimate the truth value of the proposition based on a set of observations. In case the belief, disbelief and uncertainty of an opinion are based on a set of evidence, then they are computed as follows:

$$b = \frac{p}{p+n+W} \quad d = \frac{n}{p+n+W} \quad u = \frac{W}{p+n+W} \quad a = \frac{1}{2} \tag{28}$$

where p and n represent the count of positive and negative observations in the set of evidence. The weighing factor W is usually set equal to $|\mathcal{R}|$, since this implies the useful consequence that the probability distribution that is equivalent to a subjective opinion based on zero observations is a uniform distribution. So, in the case of binomial opinions, we obtain:

$$b = \frac{p}{p+n+2} \quad d = \frac{n}{p+n+2} \quad u = \frac{2}{p+n+2} \quad a = \frac{1}{2} \tag{29}$$

In this case, we see how the uncertainty is always higher than zero, and its value is inversely proportional to the amount of evidence observed.

Recall the example of an annotator (Davide) being evaluated by a museum M. We want to estimate the belief of the museum in Davide being trustworthy based on the evaluation of a set of annotations provided by Davide. Suppose that the set comprises only five good annotations, that is, five positive observations. This is the only knowledge at the disposal of the museum M, so it has to rely on these observations. On the other hand, this evidence set is so small that fully relying on it would be risky: what if these five positive pieces of evidence are followed by three (not yet known) negative ones? A museum that considers an annotator as fully trustworthy accepts all his contributions, but if the trustworthiness estimation is fallacious, the authoritative position of the museum could be affected by a wrong annotation of its artifacts, resulting in a wrong cataloging, retrieval, etc. Following the formulas above,

we can estimate an opinion about the trustworthiness of Davide in a prudent manner, that is, by accounting the fact the estimation is partially uncertain.

$$\omega^{M}_{(\text{Davide is a trustworthy annotator})}(0.71, 0, 0.29) \tag{30}$$

The resulting expected probability for the trustworthiness of Davide is then:

$$E = 0.71 \times 0.29 \times \frac{1}{2} = 0.86 \tag{31}$$

So the museum puts a high trust level on Davide, although it does not fully trust him yet.

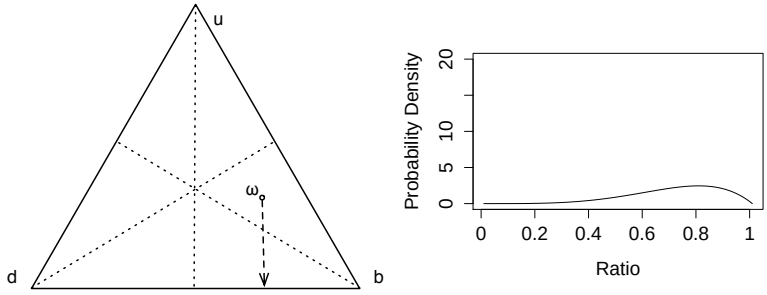
A subjective opinion is equivalent to a Beta probability distribution (binomial opinion) or to a Dirichlet distribution (multinomial opinion). The expected probability (E) is indeed the expected value of such a probability distribution. In fact, such a distribution represents the probability for each of the values in the interval $[0, 1]$ to represent the right probability for a given proposition.

A subjective opinion can be graphically represented as one of the two representations in Figure 1. First, we explain the triangular representation. Later in this section we will explain the Beta distribution representation of opinions, together with a clarification about the statistical implications of this probabilistic representation. The triangle depicted in Figure 1(a) is not a triangle in a Cartesian space, rather it represents the space of all the possible opinions. Each of the dashed axes that connects a vertex with the mid point of the opposite edge represents the geometric dimension of that vertex. The value of that dimension is zero in the edge and one in the vertex. For instance, an opinion with uncertainty 0.5 is positioned halfway between the vertex "u" and the edge "b-d". The values of belief and disbelief determine the exact position of the opinion. An opinion that has uncertainty one (and hence belief and disbelief zero) is positioned in the "u" vertex. If we take opinion $\omega(0.4, 0.1, 0.5)$, this opinion will be situated at distance 0.4 from the edge "d-u", since its belief value is 0.4. It will be at distance 0.1 from "b-u" and 0.5 from "b-d". The intersection of all these three distances determines the position of the opinion.

The lower edge, that links belief and disbelief, represents also the interval of expected probabilities for a given proposition since it is the edge where opinions with zero uncertainty situate. Given an opinion, it is possible to determine the expected probability of the proposition by projecting the opinion onto the lower edge of the triangle. The inclination of the projection is determined by the base rate a. If $a = 0.5$, the projection is orthogonal with respect to the lower edge, because there is no bias. If $a > 0.5$, there is positive bias, so the projection will be inclined towards the "b" vertex of the triangle, and vice versa if $a < 0.5$.

The equivalence between the subjective opinions and the corresponding probability distribution is determined by the following formulas. In particular, we focus on the equivalence between binomial opinions and Beta distributions, that are the opinions (and distributions) we adopt throughout the thesis.

$$\begin{aligned} \alpha &= \tfrac{2*b}{u} + 2 \times a \\ \beta &= \tfrac{2*d}{u} + 2 \times (1 - a) \end{aligned} \tag{32}$$



(a) Opinion (b) Beta distribution

Figure 1: ω is an opinion based on four positive pieces and one negative piece of evidence. Figure (a) represents ω in the triangular space. Figure (b) represents the corresponding Beta distribution (Beta(4+1,1+1) = Beta(5,2)). The parameters of the Beta are equal to the amount of positive and negative evidence respectively, increased by one.

With respect to the evidence counts, the two parameters of the Beta distribution are computed as follows:

$$\begin{aligned} \alpha &= p + 2 \times a \\ \beta &= n + 2 \times (1 - a) \end{aligned} \tag{33}$$

The meaning of this interpretation is the following. When uncertainty is zero, we assign a probability to a proposition as to say that its truth value is implied by a Bernoulli distribution:

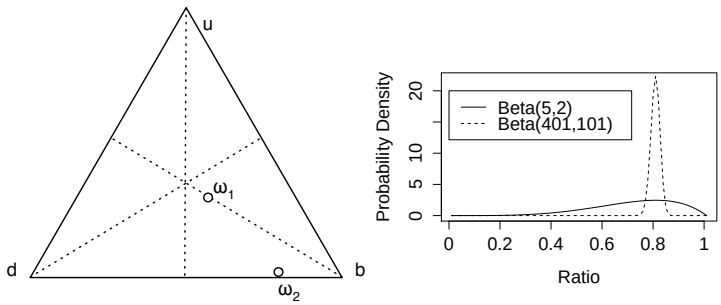
$$\omega_x(b, d, 0) \equiv P(x) = b \equiv x \sim Bern(b) \tag{34}$$

But what happens if uncertainty is higher than zero? In that case the expected probability of the proposition is still determined by the belief and the disbelief, but the higher is the uncertainty, the more this probability is, in fact, uncertain. In other words, this probability is yet to be determined (given the evidence observed), and the variance of the distribution of this probability is proportional to the uncertainty of the opinion.

$$\omega_x(b, d, u) \equiv x \sim Bern(p), p \sim Beta(\alpha, \beta) \tag{35}$$

$$var(Beta(\alpha, \beta)) \propto u \tag{36}$$

The fact that the Bernoulli and the Beta distribution (as well as the Multinomial and the Dirichlet distributions for the case of multinomial opinions) belong to the same exponential family simplifies our computations. Given a prior distribution, we obtain the posterior distribution based on our evidence by just updating the parameters of the prior. The third section of this chapter explains this fact more in depth. Figure 2 and 3 show how two opinions based on observations with the same ratio (4:1) but with



(a) Opinion (b) Beta distribution

Figure 2: ω_1 is an opinion based on four positive pieces and one negative piece of evidence, while ω_2 is based on four hundred positive and one hundred negative pieces of evidence. The left picture locates the opinions in the triangular space. The right picture shows the two distributions describing the opinions. Opinion ω_2 has less variance than opinion ω_1, since it is based on more evidence.

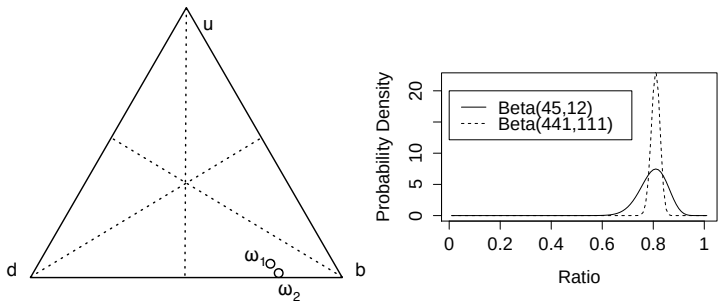
two different observation set sizes (500 vs. 5) are affected by 50 new observations, still keeping the same ratio. Of course, the opinion based on the smaller set of data is the most susceptible, since it is more uncertain. This explains also why opinions based on fewer observations are more susceptible to the influence of a biased prior: if we base our opinion only on few observations, then we rely more on our prior knowledge.

Subjective Logic Operators for Combining Opinions

In several situations, single opinions are not enough to answer questions that involve many atomic facts related to each other in disparate manners. We saw before that, thanks to the compatibility of subjective logic with Boolean logic, we can combine opinions by logically combining the corresponding propositions. Also, we hinted at how it is also possible to manipulate opinions with respect to the belief owner. In particular, we can merge opinions provided by different sources and weigh them according to the trust level of the source, when this is known. Here we describe the operators of subjective logic we make use of in the rest of the thesis: the fusion, the discounting and the conjunction operator.

Fusion Operator

Suppose that we collect different disagreeing opinions about the same proposition from different independent Web sources. We do not have any prior knowledge about the proposition, nor about the reliability of sources. To handle these conflicts, we can merge all the opinions into a global one, trusting that if the amount of opinions



(a) Opinion (b) Beta distribution

Figure 3: ω_1 and ω_2 after update. Now ω_1 is based on 44 positive and 11 negative pieces of evidence. ω_2 is now based on 440 positive and 110 negative pieces of evidence. ω_1 is much more affected by this change than ω_2.

considered is large enough, the majority of the opinions will provide us a reliable result (wisdom of the crowd [147]). If the amount of opinions considered is not large, then the uncertainty in the "merged" opinion is high. This merge is done by means of the "cumulative fusion operator" ($\oplus$):

$$\omega_x^A \oplus \omega_x^B = \omega_x^{A \diamond B} = \begin{cases} b_x^{A \diamond B} = \frac{u_x^B \times b_x^A + b_x^B \times u_x^A}{u_x^A + u_x^B - u_x^A \times u_x^B} \\ d_x^{A \diamond B} = \frac{u_x^B \times d_x^A + d_x^B \times u_x^A}{u_x^A + u_x^B - u_x^A \times u_x^B} \\ u_x^{A \diamond B} = \frac{u_x^B \times u_x^A}{u_x^A + u_x^B - u_x^A \times u_x^B} \\ a_x^{A \diamond B} = \frac{2 \times a_x^A \times a_x^B}{a_x^A + a_x^B} \end{cases} \tag{37}$$

This operator merges two opinions by treating them as independent. The resulting opinion is equivalent to an opinion based on the union of the sets of evidence of the two input opinions. The belief owner of the resulting opinion is an imaginary agent that binds together the believed owners of the fused opinions. The beliefs (and also disbeliefs and uncertainty) are averaged with respect to the mutual uncertainties. In this way, the operator lets the contribution of the opinion with lower uncertainty count more on the aggregated opinion. The precise name for this operator is "cumulative fusion operator" (represented by the symbol $\oplus$). There exists also a so-called "averaging fusion operator", that averages the evidence counts instead of cumulating them (hence assuming dependence between opinions). We do not make use of the averaging fusion operator. For the sake of simplicity, we often refer to the "cumulative fusion operator" as the "fusion operator". Note that the merged opinion is exactly equivalent to an opinion based on the amount of evidence used to build the input opinions. This

fact allows us to compute opinions in an incremental and possibly distributed way. If we get opinions from different sources, we do not need to look for the evidence that led to these opinions in order to obtain a global opinion that takes all this evidence into consideration. It is possible to combine the opinions that we see and obtain an opinion that is equivalent to an opinion based on that evidence. The same holds in case evidence is not all available at the same time, but rather, is collected progressively.

Discounting Operator

Suppose that a Web source provides us with an opinion about something we do not know. Suppose, also, that we know that such a source is not always reliable. The discounting operator allows us to weigh the received opinion in order to take into account the opinion about the source that provided it. If the source is known to be malicious, then we can make use of the so-called "opposite belief favoring" operator. In the case studies analyzed in this thesis, we did not have explicit indication of malicious behaviours, so we refer to two other discounting operators: the "uncertainty favoring" discounting and the "base rate sensitive" discounting. Both operators favor the uncertainty in the resulting opinion, as a consequence of the uncertainty in the source that provides us the opinion. So, in both cases we start from an opinion we (we are represented by means of A in the opinion) own with respect to a source B:

$$\omega_B^A = (b_B^A, d_B^A, u_B^A, a_B^A) \tag{38}$$

and with an opinion that B provides us about a proposition x:

$$\omega_x^B = (b_x^B, d_x^B, u_x^B, a_x^B). \tag{39}$$

The operator is denoted by the same symbol in both cases, because the symbol $\otimes$ denotes generically the discounting operation. Case by case, we will make explicit the kind of discounting operator we consider. We see now how the two operators allow us to weigh the opinion provided by the third party, by taking into account the opinion on it. A small note regarding the resulting, weighed opinion regards its notation $\omega_x^{A:B}$. Recall from Equation 2 that opinions can be represented also as $\omega(A:x)$. The fact that the subject of the discounted opinion is $A:B$ stands exactly for the fact that the belief owner of this opinion is A through its opinion on B.

Uncertainty Favoring Discounting

The uncertainty favoring discount operator computes the discounted opinion as follows:

$$\omega_B^A \otimes \omega_x^B = \omega_x^{A:B} = \begin{cases} b_x^{A:B} = b_B^A \times b_x^B \\ d_x^{A:B} = b_B^A \times d_x^B \\ u_x^{A:B} = d_B^A + u_B^A + b_B^A \times u_x^B \\ a_x^{A:B} = a_x^B \end{cases} \tag{40}$$

This operator is designed in such a manner that in the resulting opinion, belief and disbelief in x are the same as the belief owned by B, but in both cases they are weighed with respect to the belief in B. In case the belief in B is less than 1, $b_x^{A:B} + d_x^{A:B} < b_x^B + d_x^B$. The remaining mass is assigned to the uncertainty, that is thus favored.

Base Rate Sensitive Discounting

The base rate sensitive discounting works as follows:

$$\omega_B^A \otimes \omega_x^B = \omega_x^{A:B} = \begin{cases} b_y^{A:B} = E(\omega_B^A) b_y^B \\ d_y^{A:B} = E(\omega_B^A) d_y^B \\ u_y^{A:B} = 1 - E(\omega_B^A)(b_y^B + d_y^B) \\ a_y^{A:B} = a_y^B \end{cases} \tag{41}$$

At first glance, this operator looks natural since it relies on the expected probability about the trust level of the source. Nevertheless, it has to be adopted with caution. In fact, recall that

$$E = b + u \times a$$

and suppose that the base rate about the source is high (e.g. because the source belongs to a class of trustworthy sources, even if we do not know if actually the source is trustworthy). Paradoxically, opinions based on very few observations, and thus highly uncertain, have a high expected value, thanks to the high base rate. So, in this case, the opinion provided by B would be highly weighed, despite the very few observations about B's trustworthiness. Therefore, usually we prefer the uncertainty favoring discounting operator.

Conjunction Operator

Another operator that we use is the conjunction operator, also called "product" or "multiplication", which is an extension of the Boolean logical AND operator and is represented by means of the symbol $\cdot$. Recall from Equations 11 and 27 that

$$b + d + u = 1$$

and

$$E = b + u \times a$$

Now, following Table 3

$$b_{x \wedge y} = b_x \times b_y$$

$$d_{x \wedge y} = d_x + d_y - d_x \times d_y$$

Given the constraints above, we obtain the so-called "simple multiplication":

$$\omega_x^A \cdot \omega_y^A = \omega_{x\wedge y}^A = \begin{cases} b_{x\wedge y}^A = b_x b_y \\ d_{x\wedge y}^A = d_x + d_y - d_x d_y \\ u_{x\wedge y}^A = b_x u_y + u_x b_y + u_x u_y \\ a_{x\wedge y}^A = \frac{b_x a_y u_y + a_x u_x b_y + a_x u_x a_y u_y}{b_x u_y + u_x b_y + u_x u_y} \end{cases} \tag{42}$$

However, this operator has the unpleasant consequence to let the prior $a_{x\wedge y}$ be dependent on the belief, disbelief and uncertainty of ω_x^A and ω_y^X. In other words, the derived prior is dependent on the actual observations. In order to correct this problem, we change our constraints, and we set:

$$a_{x\wedge y} = a_x \times a_y$$

$$d_{x\wedge y} = d_x + d_y - d_x \times d_y$$

In this way we obtain the so-called "normal multiplication" operator:

$$\omega_x^A \cdot \omega_y^A = \omega_{x\wedge y}^A = \begin{cases} b_{x\wedge y}^A = b_x b_y + \frac{(1-a_x)a_y b_x u_y + a_x(1-a_y)u_x b_y}{1-a_x a_y} \\ d_{x\wedge y}^A = d_x + d_y - d_x d_y \\ u_{x\wedge y}^A = u_x u_y + \frac{(1-a_y)b_x u_y + (1-a_x)u_x b_y}{1-a_x a_y} \\ a_{x\wedge y}^A = a_x a_y \end{cases} \tag{43}$$

When uncertainty is zero, this operator preserves the compatibility with the boolean conjunction operator. Because of this, and because of the reason introduced above (prior of the conjuncted propositions dependent only on the priors of the two propositions), we prefer this operator over the "simple" one introduced before.

One last remark regards the formula for computing $b_{x\wedge y}^A$. This is similar to the product of the two probabilities of x and y that is adopted in probability theory to obtain the probability of two independent events, but there is an additional addend. This addend is due to the uncertainty, since in probability theory all the probability is attributed to the events or their negation, while here we reserve some probability, indistinguishably attributable to the belief and the disbelief. So that additional addend compensates this fact by partially attributing this probability mass.

For more details about the logic, its foundations and its operators, see the works of Dempster [43], Shafer [143], Jøsang [83], and Jøsang et al. [88].

Conjugate Priors

Conjugate priors provide one of the probabilistic foundations of this thesis. In particular, in subjective logic it is possible to easily update subjective opinions exactly because the prior distribution (e.g. Beta) is conjugated with the distribution describing the actual observations (e.g. Bernoulli).

The basic idea of conjugate priors starts from the Bayes theorem (44): given prior knowledge and our data, we update the knowledge into a posterior probability.

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)} \tag{44}$$

This theorem describes how it is possible to compute the posterior probability, $P(A \mid B)$, given the prior probability of our data, $P(A)$, the likelihood of the model, given the data, $P(B \mid A)$, and the probability of the model itself, $P(B)$.

When dealing with continuous probability distributions, the computation of the posterior distribution by means of Bayes theorem can be problematic, due to the need to possibly compute complicated integrals. Conjugate priors allow us to overcome this issue: when prior and posterior probability distributions belong to the same exponential family, the posterior probability can be obtained by updating the prior parameters with values depending on the observed sample (see also the work of Fink [56]). Exponential families are classes of probability distributions having their density functions that share the form $f(x) = e^{a(q)b(x)+c(q)+d(x)}$, with q a known parameter and a, b, c, d known functions. Exponential families include many important probability distributions, like the Normal, Binomial, Beta, etc. (see the work of Andersen [52] for further details about the exponential family of probability distributions). So, if X is a random variable that distributes as defined by the function $P(p)$ (for some parameter or vector of parameters p) and, in turn, p distributes as $Q(\alpha)$ for some parameter (or vector of parameters) α called "hyperparameter", and P belongs to the same exponential family as Q,

$$p \sim Q(\alpha),\ X \sim P(p) \tag{45}$$

then, after having observed *obs*,

$$p \sim Q(\alpha') \tag{46}$$

where $\alpha' = f(\alpha, obs)$, for some function f.

For example, the Beta distribution is the conjugate of the Binomial distribution. Suppose we observe *pos* positive pieces of evidence and *neg* negative ones. We want to estimate a probability distribution that allows us to predict whether the next observation will be positive or negative, so we use a Bernoulli distribution (which is a special case of the Binomial distribution)

$$Bernoulli(p)$$

and suppose that we derive its parameter from the sample observed:

$$p = \frac{pos}{pos + neg} \tag{47}$$

That would be quite risky, because the sample we observe is small, and we have no assurance about its representativity. Thus, we refer to conjugacy, and since the Beta distribution is conjugated with the Binomial (and hence with the Bernoulli), we let the parameter p of the Bernoulli distribution be determined by the prior Beta distribution.

$$p \sim Beta(\alpha, \beta)$$

With no observations and no prior bias, our Beta is:

$$p \sim Beta(1, 1)$$

and, the posterior will be

$$p \sim Beta(\alpha + pos, \beta + neg)$$

that is,

$$p \sim Beta(1 + pos, 1 + neg)$$

This means that the Beta, shaped by the prior information and by the observations, defines the range within which the parameter p of the Bernoulli is probably situated, instead of directly assigning to it the most likely value. This is exactly how a binomial subjective opinion handles the evidence observed. Other examples of conjugate priors are: Dirichlet distribution, which is conjugate to the Multinomial distribution, and Gaussian distribution, which is conjugate to itself. Conjugacy guarantees ease of computation, which is a desirable characteristic when dealing with very big data sets as Web data sets often are. Moreover, the model is incremental, and this makes it fit the crawling process with which Web data are obtained, because crawling, in turn, is an incremental process. Both the heterogeneity of the Web and the crawling process itself increase the uncertainty of Web data. The probabilistic determination of the parameters of the distributions adds a smoothing factor that helps to handle this uncertainty.

In this thesis we focus mainly on categorical distributions because these are intended to represent categorical data, that comprise also the semi-structured Web data that constitute the focus of the research here presented. Being able to estimate the data distribution is crucial for our trust estimates because, based on an estimate of the data distribution, we can decide whether the data we face are trustworthy or not. In Chapter 3 we show that using conjugated priori we obtain a better approximation of the data distribution than relying only on the sample observed.

Dirichlet Process

Dirichlet processes [55] are a generalization of Dirichlet distributions, since they correspond to probability distributions of Dirichlet probability distributions. In Chapter 3 we show that it is possible to use Dirichlet processes to model Web data, and we provide an extension of subjective logic that makes use of Dirichlet processes ("open world

opinions"). In fact Dirichlet processes extend the conjugated Beta-binomial distributions and Dirichlet-multinomials by letting the amount of categories to be potentially infinite, and by allocating part of the probability mass for categories that have not yet been observed.

From the formal point of view, Dirichlet processes are stochastic processes, that is, sequences of random variables (distributed as Dirichlet distributions) which value depends on the previously seen ones. Using the so-called "Chinese Restaurant Process" representation (see the work of Pitman [125]), a Dirichlet process can be described as follows:

$$X_n = \begin{cases} X_k^* & \text{with probability } \frac{num_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } H & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases} \tag{48}$$

where H is the continuous probability measure ("base distribution") from which new values are drawn, representing our prior best guess. Each draw from H will return a different value with probability 1. α is an aggregation parameter, inverse of the variance: the higher α, the smaller the variance, which can be interpreted as the confidence value in the base distribution H. So, the higher the α value is, the more the Dirichlet process resembles H. The lower the α is, the more the value of the Dirichlet process will tend to the value of the empirical distribution observed. Each realization of the process is discrete and is equivalent to a draw from a Dirichlet distribution, because, if

$$G \sim DP(H, \alpha) \tag{49}$$

is a Dirichlet process, and $\{B\}_{i=1}^n$ are partitions of the category set S, we have

$$(G(B_1), \ldots, G(B_n)) \sim Dirichlet(\alpha H(B_1), \ldots, \alpha H(B_n)) \tag{50}$$

If our prior Dirichlet process is distributed as in Equation (49), given (50) and the conjugacy between Dirichlet and Multinomial distribution, our posterior Dirichlet process (after having observed n values θ_i) can be represented as one of the following two representations:

$$(G(B_1), \ldots, G(B_n)) \mid \theta_1, \ldots, \theta_n \sim Dirichlet(\alpha H(B_1) + n_{\theta_1} \ldots \alpha H(B_n) + n_{\theta_n}) \tag{51}$$

$$G \mid \theta_1 \ldots \theta_n \sim DP\left(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \frac{\Sigma_{i=1}^n \delta_{\theta_i}}{n}\right) \tag{52}$$

where δ_{θ_i} is the Dirac delta function (see [45]), that is, the function having density only in θ_i. The new base function will therefore be a merge of the prior H and the empirical distribution, represented by means of a sum of Dirac delta's. The initial status of a Dirichlet process posterior to n observations, is equivalent to the nth status of the initial Dirichlet process that produced those observations (see De Finetti theorem, [98]).

The Dirichlet process, starting from a possibly non-informative "best guess", as long as we collect more data, will approximate the real probability distribution. Hence,

it will correctly represent the population in a prudent (smoothed) way, exploiting conjugacy like the Dirichlet-Multinomial model, that approximates well the real Multinomial distribution only with a large enough data set. The improvement of the posterior base distribution is testified by the increase of the α parameter, proportional to the number of observations. It is evident that the mechanism of update of the Dirichlet process based on a series of observations is the same of the conjugate priors described before. In Chapter 3 we show that this model is useful to model uncertain categorical Web data.

Semantic Similarity Measures

We use semantic similarity measures to increase the availability of evidence when evaluating the trustworthiness of tags and annotations, especially in the cultural heritage domain. Semantic similarity measures quantify the likeness between the meaning of two given terms. Whenever we evaluate a tag, we take the evidence at our disposal, and tags that are more semantically similar to the one we focus on are weighed more heavily. There exist many techniques for measuring semantic relatedness, which can be divided into two groups. First, we have so-called "topological" semantic similarity measures, which are deterministic measures based on the graph distance between the two words examined, based on a word graph (e.g., WordNet [112]). Second, there is the family of statistical semantic similarity measures, which includes, for instance, the Normalized Google Distance [34]. These latter measures are characterized by the fact that the similarity of two words is estimated on a statistical basis from their occurrence and co-occurrence in large sets of documents.

We focus on deterministic semantic relatedness measures based on WordNet or its Dutch counterpart Cornetto [166]. In particular, we use the Wu and Palmer [178] and the Lin [104] measure for computing semantic relatedness between tags, because both provide us with values in the range $[0, 1]$, but other measures are possible as well. WordNet is a directed and acyclic graph where each vertex is an integer that represents a synset (set of word synonyms), and each directed edge from vertex v to vertex w implies that w is a hypernym of v. In other words w shares a "type-of" relation with v. For instance, if v is the word "winter" (hyponym), w can be the word "season" (hypernym). If a synset is a generalization of another one, we can measure the depth, that is the distance between the two. The first ancestor shared by two nodes is the Least Common Subsumer. The Wu and Palmer measure calculates semantic relatedness between two words by considering the depths between two synsets in WordNet, along with the depth of the Least Common Subsumer, as follows:

$$score(s1, s2) = 2 * \frac{depth(lcs(s1, s2))}{depth(s1) + depth(s2)} \tag{53}$$

where $s1$ is a synset of the first word and $s2$ of the second.

We compute the similarity of all synsets combinations and pick the maximum value, as we adopt the upper bound of the similarity between the two words. The Lin

measure considers the information content of the Lowest Common Subsumer and the two compared synsets, as follows:

$$2 * \frac{IC(lcs(s1, s2))}{IC(s1) + IC(s2)} \tag{54}$$

where IC is the information context, defined as:

$$IC(s) = -log\left(\frac{freq(s)}{freq(root)}\right) \tag{55}$$

and *freq* is the frequency of the synset in a given document corpora.

By choosing to use these measures we limit ourselves in the possibility of evaluating only single-word tags and only common words, because these are the kinds of words that are present in WordNet. However, we choose these measures because almost all the tags we evaluate fall into the mentioned categories and we validate the use of these similarity measures together with subjective logic in Chapters 3, 7 and 8. Moreover, almost all the words used in the annotations that form the datasets we use in our evaluations are single-word tags and common words, hence, this limitation does not affect our evaluation significantly.

Part I

Using Web Data to Make Trust Assessments

*This part collects the chapters that address the first research question (*Can Web data help the trust evaluation of semi-structured data?*). As a consequence of the need to process Web data for making trust assessments, also the second research question (*How can uncertainty reasoning be effectively used to estimate the trustworthiness of semi-structured data?*) is touched here, although it will be deeper addressed in Part II. Chapter 1 introduces a trust model that makes use of Web data that will be adopted and extended in several other chapters of this thesis, namely Chapters 3, 5, 6, 7, 8. Chapter 2, instead, proposes a statistical approach to validate the use of (Web-based) heuristics to validate semi-structured data, that lays the foundations for the work presented in Chapter 4.*

1

Estimating Trust in Annotations using Web Data

A Bird Specimen Annotations Case Study

*This chapter presents a research about the use of Web data to make trust assessments about uncertain semi-structured data, thus addressing the first research question (*Can Web data help the trust evaluation of semi-structured data?*). In particular, we focus on professional semi-structured data: annotations of a collection of bird specimen held by the Naturalis Museum in the Netherlands. Web data are used to enrich the internal data and since subjective logic (see Chapter Preliminaries) is used to reason upon the enriched data to obtain a trust value for each annotation, we partially address also the second research question (*How can uncertainty reasoning be effectively used to estimate the trustworthiness of semi-structured data?*). We propose five strategies for trust assessments, based on different use of the sources and of the subjective operators at our disposal, and we evaluate them against a gold standard provided by the Naturalis Museum. The work presented here provides an exploration that is further extended in the following chapters. In fact, the enrichment of internal data is one of the possible gains deriving from the use of Web data: the Web offers a vast amount of data and, especially when we are able to choose authoritative data sources like in this case, we can benefit from data that would be unavailable otherwise. In Chapters 6, 7 and 8 we extend the reasoning model proposed here, and we benefit from the use of Web data in other manners as well, like the possibility to crowdsource data. In Chapter 2, instead, we tackle the problem of using Web data to make trust assessments from a different point of view, i.e., instead of enriching the data to be evaluated, we evaluate them by making use of heuristics based also on Web data.*

The work presented in this chapter refers to the paper A Trust Model to Estimate the Quality of Annotations using the Web *coauthored with Willem*

Robert van Hage and Wan Fokkink and presented at the Web Science Conference 2010 (WebSci10) *in Raleigh, USA.*

1.1 Introduction

Professional media include all digital media, such as images, video and audio used for professional purpose. The term "professional" refers to all media used for business (both profit and non-profit) such as in entertainment, culture, science and product catalogs. In professional media there is a strong *supplier* point of view. Typical suppliers are TV broadcasters, museums and digital libraries. In this chapter we focus on examples of museums as professional media providers. In Chapters 6, 7 and 8 we encounter other datasets from the media and the cultural heritage domain, and the work presented here is further extended in those chapters.

Authority and quality are two key issues for professional media. Museums invest heavily in building expertise on the items in their collection. Metadata creation is therefore a core and knowledge-intensive activity in the management of professional media. A museum plays the role of authority within its field of competence. This means that it has the responsibility to keep and protect the artifacts it owns (including digital representations of the work), to guarantee their preservation over time.

However, museums often have large collections which are only partly properly catalogued. They do not have the resources to cover the complete collection. For this reason museums are looking with great interest at the current Web 2.0 trend of "tagging". An example is the Steve.Museum [155]. But tagging of museum artifacts is a different ball game when compared to tagging of say family photos. Tagging of museum artifacts is a task that requires a high level of skill for the annotator, because of the precision and quality needed for tags. In this context, quality of the tags is what ensures the authority to keep its authoritative position. Both the selective amount of skills needed to annotate properly and the consequences of a low-quality tagging are important issues for professional media.

We situate the trust definition adopted in Chapter Introduction within the context of professional media. The role of the trustor is played by the authority, that is, the museum or the TV broadcaster. This authority owns the media and wants to share it with the public, without running the risk of compromising its authoritative position. Situated between the trustor and the public is the trustee: the actor who allows the delivery of content, for instance by properly annotating it. Reliability evaluation is necessary because the trustor's authoritative position may be seriously damaged by wrong annotations. For instance, museums could use annotations (provided by internal or external experts) to manage artifacts and present them to the public. Their high reputation is achieved through the delivery of trustworthy information, in particular trustworthy metadata, so they must avoid using low quality annotations. Therefore, museums need to focus on trust modeling and to evaluate trust levels of annotations before delivering them. However, because of the workload and specific skills required,

it may turn out to be infeasible. So, we introduce a model which aims to automatically assess these trust levels.

As we saw before, quality is a key feature within a professional media environment. Precision of the terms that are used plays a crucial role in determining the quality of annotations. Its achievement is mainly due to two factors: the annotator, who needs have the necessary skills, and the thesauri or knowledge repository from which the information for annotating is chosen, which should ensure a minimum level of reliability. Although a high reputation of the expert and of the source of information can be an important assurance about the correctness of the annotation, we still need to talk of trusted instead of correct annotations, since these evaluations are made by reasonably confident inference and not by a direct manual check, which implies the possibility that the annotation is not really correct. So, before deciding whether a piece of data is trusted or not, we compute a "reputation" or "trust value" (also "trustworthiness level") also for the piece of data itself. This value reflects the reputation of the data creator and possibly other information, as described in the following sections.

The remainder of this chapter is structured as follows: Section 1.2 presents an overview of the approach and related work; Section 1.3 describes a trust model; Section 1.4 describes the application of the trust model in a concrete case study, while Section 1.5 presents a final discussion.

1.2 Approach and Related Work

Our approach is to use RDF [171]/OWL [7] in association with subjective logic (see Chapter Preliminaries). RDF/OWL is a family of languages that is commonly used for metadata management and we use it to uniformly represent our data and metadata and hence facilitate their aggregation in a unique graph. By means of RDFS [168] reification we can easily refer to the single data items that compose such a graph, and refer to each annotation separately. For each annotation, we consider the metadata associated with it, we estimate their reputation and, from this reputation, we estimate the trustworthiness of the annotation.

We use subjective logic to reason on the evidence about the metadata at our disposal. We refer the reader to Chapter Preliminaries for an introduction to this probabilistic logic. Moreover, in Chapter Introduction is presented a description literature relevant for this work, from trust models for the Semantic Web, to the use of uncertainty reasoning for making trust assessments.

1.3 Trust Model

We propose a model based on semantic web technology for the representation and ontological reasoning part, and subjective logic for the probabilistic reasoning part.

The aim of the model is to provide a tool for the automatic estimation and evaluation of trust levels of annotations, by pursuing to objectives. The first goal is connected

to the primary need which the model tries to satisfy: try to avoid or at least reduce the amount of human work needed to accept annotations. If the authority would have to manually validate each annotation, this would imply a great overhead. As a consequence, one aim of the model is try to make use of the smallest amount of manual work to safely evaluate the annotations. "Safely" means with a low margin of error. The second goal is to have a trustworthy model. Clearly, the model becomes useless if its accuracy is very low, because its evaluations become completely unreliable. Therefore, although this may imply an increase of the manual work needed, we also need to achieve a maximum overall accuracy to ensure the usability of the model. Indeed, we need to gather a significant sample of data to use it to create reliable predictions. Thus, our overall goal is to find a good balance between these two different needs.

1.3.1 Data Representation and Ontological Reasoning

Trust data is, in fact, a special form of metadata. As said before, we therefore use RDF/OWL for the representation of trust data. In particular, annotations are represented in RDF, and through RDFS we reify them in order to record metadata. Typical examples of metadata are the author of the annotation, which is linked to the reified annotation, or the author of a taxonomy used in the annotation, which is linked to the object of the annotation, in case of an annotation using taxonomies. When possible, we use standard ontologies like FOAF [15] and Dublin Core Metadata Terms [46] to represent these metadata. However, to fulfill all our requirements, we developed also the "annotationTrust ontology"[1]. For instance, we need to represent specific annotations which make use of taxonomies, and this implies the need both to represent meta-information regarding the taxonomy itself and to reason about the connection between the annotated object and the taxonomy elements (since, e.g., genus may be correct but species not, we avoid treating the taxonomy as a unique entity). The main classes included in the ontology are the following ones:

- Annotation
- AnnotatedObject
- EstimateValue
- Thesaurus
- AnnotationCreation
  - InternalThesaurus
  - ExternalThesaurus
- TrustValue
  - SLTrustValue

[1]The ontology is available at `http://trustingwebdata.org/phdthesis/dceolin`.

    - BeliefValue
    - DisbeliefValue
    - UncertaintyValue

- Annotator

and the following properties:

- hasAnnotation
- hasAuthor
- hasBelief
- hasTrustLevel
- hasBeliefLevel

This small ontology is aimed at cover the main concepts regarding annotations and their trust levels. It extends the Simple Event Model [162], because it aims at modeling both annotations, trust values, and the moments when these are created (e.g., for allowing to determine whether a trust value may or may not be outdated). This ontology has been superseded by the Open Annotation Model [14] (that is adopted, for instance, in Chapter 7 that is based on a most recent publication). However, this ontology models directly trust values and beliefs, while in the Open Annotation Model those elements are modeled by using annotations of annotations.

Moreover, by exploiting Linked Open Data [153], we can enlarge the availability of metadata and, therefore, increase the number of possible sources of information about the trustworthiness of annotations. For example, if we consider the annotation of an artwork or an animal specimen, then meta-information about the term or taxonomy used to annotate could be limited when simply relying on data internal to the authority. With Linked Open Data we can gather information regarding the painter used to annotate the artifact or the taxonomy used to annotate the specimen. Using this additional evidence, we can more confidently check the correctness of the annotation.

1.3.2 Evidential Reasoning

Once we have gathered enough semantically significant metadata, we can merge all contributions in order to obtain a single value representing the probability that the evaluated annotation is correct. Subjective logic is the method we choose to tackle this issue. Chapter Preliminaries provides an introduction to subjective logic. This logic can also be used to control the behavior of the system. By sampling and controlling the reliability of the system, we can build an opinion about its reliability and weight opinions on annotations according to these opinions. This can be seen as a web of trust, since by adding this layer, we build a reputation for the system that is returning us reputations about annotations.

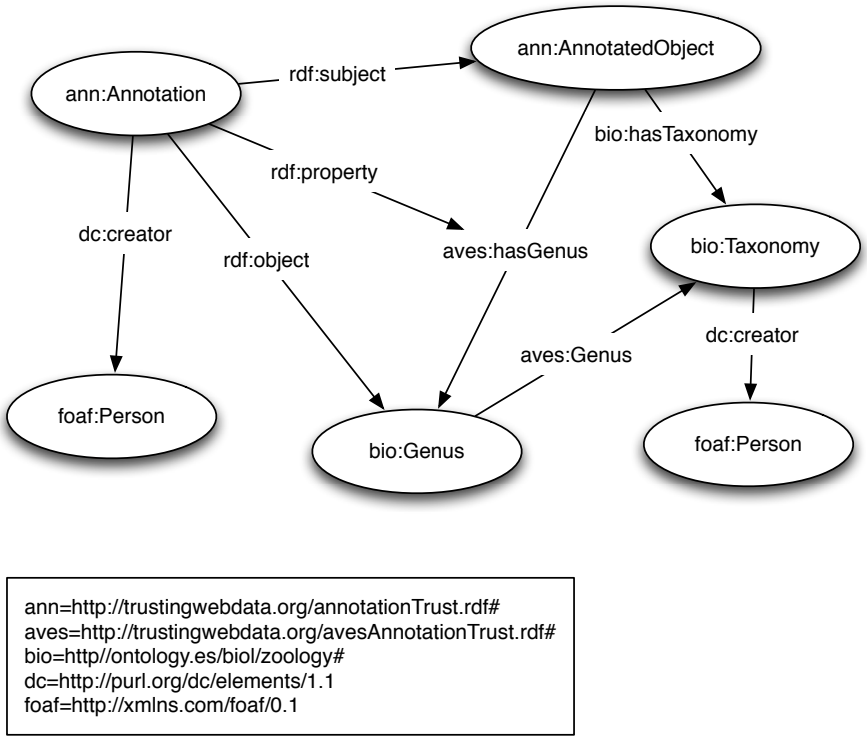


Figure 1.1: Annotation representation with RDF. By the reification of the annotation, that is, by the treatment of the annotation as an object, we can easily enrich it with the meta-information we collect.

1.3.3 Implementation

For data manipulation we use the SWI-Prolog Semantic Web package [174]. We also developed trust-management procedures in Prolog. The model has been developed according to the following structure:

Subjective logic module: This module has been developed as a generic subjective logic module and leaves aside any domain-related issues. This module, therefore, contains all the tools needed to represent subjective logic opinions, merge and discount them and record evidence. Moreover, it includes a set of predicates that allows different kinds of evidence management. For instance, these include the possibility of counting all the evidence available, or to give more importance to the most recent ones, by giving less weight to the less recent ones, and the usage of a so-called sliding window. The sliding window allows one to take into consideration only the last pieces of evidence when evaluating a reputation. This

module is available online[2].

Domain-related module: This file collects all the domain-related predicates. Here are defined the conditions for positive and negative evidence, as well as more generic strategies for evidence management and error handling. Finally, within this module the implementation choices are taken regarding evidence management, by choosing a suitable strategy among those offered by the subjective logic module.

1.3.4 Decision Strategies

Our model aims at calculating the probability that a certain annotation is correct. Taking a decision always implies a degree of error, but errors may not always have the same importance. For instance, within some contexts a false positive can be less desirable than a false negative. Different strategies are suitable for different application domains. We propose some decision strategies:

Fixed threshold: Once we have decided the maximum level of error acceptable, we will accept all annotations with a trust level above it. Clearly, there may be two kinds of threshold, one for acceptance and one for refusal. In the latter case, annotations are refused only when the trust level is below such a threshold. In case these different thresholds coexist, we have to take into account the fact that they outline a middle section, below the threshold of acceptance, but above the threshold of refusal, where our model is not able to evaluate annotations. For instance, we could decide that we can accept a maximum level of error of 10% due to acceptance of false positives. Therefore, we will accept all annotations with a trust level above 0.9. Since a trust level of 0.9 means that on average no more than one annotation out of ten is wrong (but we do not know which one), our false positive rate should be at most 10%.

Probability distribution simulation: The previous strategy guarantees a certain maximum error rate, but on the other hand, it does not leave room for improving the error rate, since it accepts any annotation which trust level is beyond the threshold without trying to discover wrong annotations among those with a sufficiently high trust level (which may exist since their trust level is still lower than 1). In order to try to do this, we can simulate the probability distribution determined by the trust level and use such a simulation to take decisions. Suppose that an annotator has a reputation of 0.9. This means that, on average, he will make one wrong annotation out of ten. If our function accepts one annotation out of ten, when they are made by this author, then our error rate may reach 0%, in case we are able to match the wrong annotation with the refusal by the function, or diverge otherwise. By running this function multiple times, checking the expected value and variance of the results, and by trying to limit

[2]The code is available at `http://trustingwebdata.org/phdthesis/dceolin`.

its deviation, we can at least infer useful information on which we can base our decisions.

Speed of variation: Within certain domains, positive opinions coming from different sources "sustaining" each other may lead to a decision, although the final opinion resulting from their merge may be slightly different from 0 or 1. In particular, when we face an opinion which is positive or negative, but still far from acceptance or refusal, and by merging it with another one regarding the same subject our total opinion moves rapidly to one extreme value, this may be enough to take a decision.

Another important aspect that has to be taken into account is the choice to reuse evaluations made by the model as evidence. On the one hand, this may be an optimal choice looking at the dependency of the data, since it allows us to reinforce the strength of opinions without the need for more manual evaluations by the authority. On the other hand, this may also be a risky choice, since in case we make evaluations based on a not completely sure reputation, this increases the error rate.

These are the approaches analyzed so far, but clearly, this is not an exhaustive selection. However, the decision strategy prescinds from the calculation of the trust levels, which is the primary aim of the model, unless we do not reuse evaluations as evidence.

In the case study presented in Section 1.4, we will use only the fixed threshold strategy.

1.3.5 Usage of the Model

The model can uniformly deal with heterogeneous metadata about the annotations. This uniform representation of trust evaluation leads to two important advantages. The first is clearly the possibility of merging all these various contributions into a unique value. The second is reusability of this value. By clearly defining the context, the authority creating it, the metadata used and the methods applied, we facilitate their reuse. Another authority needing to evaluate the handiwork of the same author may directly make use of such evaluations, taking into account the reputation of the assessing authority and the methods used for the assessment. This way we implicitly allow the creation of a so-called "web of trust" [84]. A concrete implementation of such a web is something we will investigate in the future. Besides the uniform representation of trust assessments, we will also need to keep track of which authority made such statement, when it made them and how. These provenance information will be easily tracked by means of the PROV Ontology [9] or of a similar model.

1.4 Case Study: Naturalis Data

The case study we face regards the annotation of bird specimens curated by the National Museum of Natural History in Leiden, Naturalis. Here we implement and apply

the model we propose to a concrete use case. The museum has to deal with a vast amount of annotations, and their quality and trustworthiness are crucial for its business. In principle, these annotations can be subject to imprecision, inaccuracy or, in general, errors, because of the high expertise needed to produce them. The vastness of the annotations that the museum deals with makes an automated approach to the annotation evaluation particularly valuable. Such a model overcomes also the need for highly specialized knowledge that is required to review these annotations. The model learns from previously evaluated annotations their peculiar characteristics and uses the knowledge acquired to assess the trustworthiness of other annotations. In fact, the model implements a supervised machine learning approach. In this way the museum is also relieved from the burden of having to explicitly define the policy for the annotation acceptance or rejection.

After having introduced the dataset and the case study setup, we describe a list of alternative strategies for deciding on the evaluated annotations, using the trust values produced, and we present the results obtained with these.

1.4.1 Dataset

The Naturalis Museum has at its disposal a database of annotations of its bird specimens. This database records information about taxonomies, specimens, and how these are classified using taxonomies. Experts annotated each specimen using a taxonomy recorded in the database. The result of such a linking is a "one-to-many" relation since, in general, more specimens of the same species are present. However, these annotations are not always correct, and this may be due to many reasons: for instance a mistake by the annotator, or the fact that the taxonomy became obsolete after a certain period. Therefore, in such a database, in a second moment, the museum experts created a second set of annotations which, because of their recent creation and because of additional checks, are considered as correct by the museum, and hence treated as a gold standard in this case study. Since the museum is the authority we refer to, this series of annotations is our landmark: our model should assign a high trust value to annotations produced by an annotator and confirmed by the museum, and a low trust value to the others. From a comparison between the trust values and the judgement by the museum, we are therefore able to evaluate our model. For reasons of confidentiality, we cannot expose this dataset in full detail. However we can outline the structure and basic information contained in the three tables that form the database as follows:

AvesRegister Main table of the database. Contains data about the specimen (age, sex, collection date, etc.), its classification, by means of an external code referring to table AllNames, and its recorder (i.e., annotator creator).

AllNames Table describing the taxonomies adopted. It reports the values for genus, species, subspecies, etc. of each taxonomy.

AdditionalInfo Table collecting additional information about the specimens, like color, weight, etc.

At the "Netherlands Biodiversity Information Facility" portal [118], it is possible to see examples of correct annotations exposed by Naturalis Museum.

1.4.2 Case Study Setup

Data are provided in the form of a classical relational database. Through the use of D2RQ [13], these are easily converted into RDF. Once converted into RDF, we reify annotations, in order to associate also the creator with the annotations themselves. The same process is performed when enriching the taxonomy with additional information. Since taxonomy authors are not recorded in a homogeneous way in the Naturalis dataset, we refer instead to the U.S. National Biological Information Infrastructure [115] to collect this kind of information. This infrastructure exposes an authoritative and exhaustive database of taxonomies which, once converted into RDF, has been used to annotate the annotations we are evaluating. In order to improve the representation of reified bird annotations, we developed a small ontology available online[3], called "avesAnnotationTrust ontology", which extends the one cited in Section 1.3.1 in order to accurately represent taxonomic annotations of bird specimens. This ontology contains the following classes and subclasses:

- AvesAnnotation
  - AvesGenusAnnotation
    - NewAvesGenusAnnotation
    - OldAvesGenusAnnotation
  - AvesSpeciesAnnotation
    - NewAvesSpeciesAnnotation
    - OldAvesSpeciesAnnotation

The AvesAnnotation class is a subclass of the Annotation class defined in the AnnotationTrust ontology presented before. Also the following properties are defined:

- hasGenus
- hasSpecies

These properties are used to link respectively the value of the genus and of the species to a given specimen. This ontology can be seen as a specialized and smaller precursor of the Open Annotation Model [14], that is currently used as a standard model for representing annotations. To represent taxonomies, we use the Biological Taxonomy Vocabulary [79].

Figure 1.2 shows the overview of the case study. We start by joining four tables in our database (Aves register, Unchecked annotations, Correct Annotations and Taxonomies). Then, we enrich the data at our disposal, we refer to the NBII database, and

[3]The ontology is available at `http://trustingwebdata.org/phdthesis/dceolin`.

in order to maintain a uniform representation of our data, we make use of RDF and of specific ontologies, as described above in this section, in Section 1.3.1 and shown in Figure 1.1.

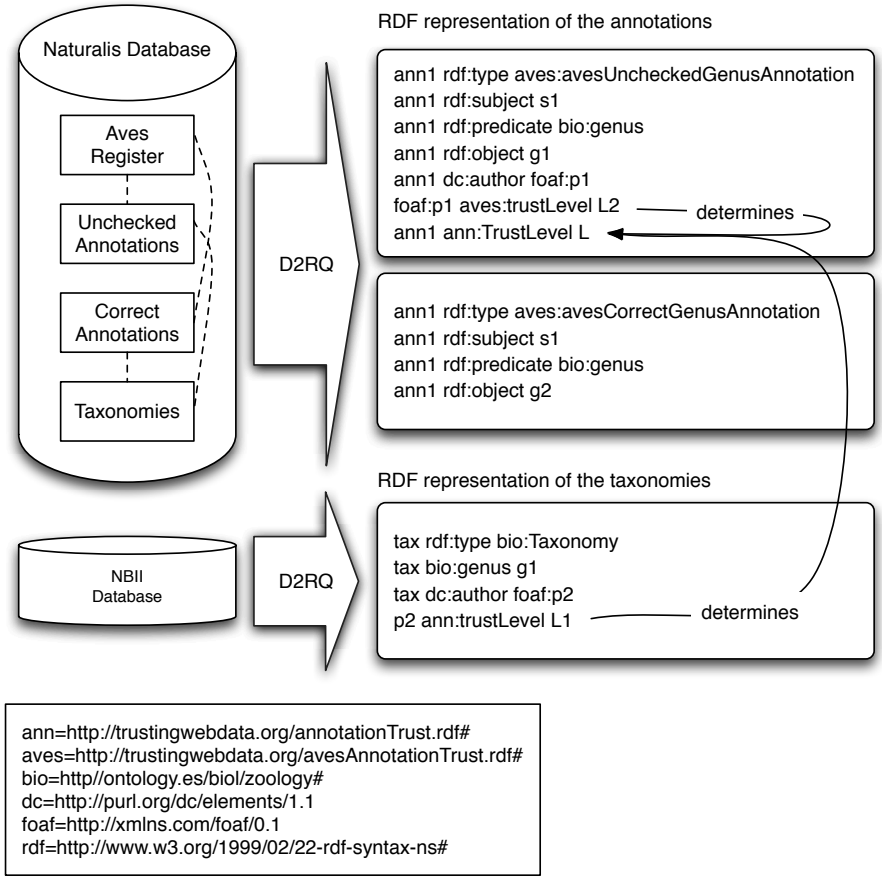


Figure 1.2: Case study overview. We represent in RDF the annotation and taxonomies available in the initial databases. Then, we compute and merge the corresponding trust values, to obtain one trust value per annotation.

Once the data are prepared, we create a series of Prolog procedures, available online[4], which allow us to build reputations for each kind of information source and compute trust levels of annotations. We aim at estimating whether the value of the properties "hasGenus" and "hasSpecies" is trustworthy. These values are entities from controlled vocabularies (e.g., NBII), so here we estimate the correct association be-

[4]The code is available at `http://trustingwebdata.org/phdthesis/dceolin`.

tween specimen and genus or species value. Different implementation strategies are adopted and the results are reported in Section 1.4.3. The subjective logic predicates used by these procedures are those contained in the module described in Section 1.3.3. The data thus obtained are then split into a training and a test set. We use the training set to build a subjective opinion from a given point of view (e.g., an opinion about the author trustworthiness), and whenever we find a matching entry in the test set, we apply a decision strategy based on the value of the opinion. Splitting strategy, opinion subject and decision strategy can be implemented in several ways. The following subsection describes five joint strategies and implementations.

1.4.3 Results and Analysis

We analyzed a set of 65,600 annotations made by ten authors. We adopted different implementation strategies both to compare them and to simulate different scenarios. The results are presented in Table 1.1. We report only the accuracy as a metric for evaluating our estimates. By adopting this metric, we want to measure the ability of the system to produce correct results, without attributing a higher weight either to the positive (trustworthy annotations) or to the negative (untrustworthy annotations) estimates. False positive estimates, in fact, imply that the museum accepts annotations that are wrongly classified as trustworthy, and this would affect its authoritativeness. False negative estimates would mean that the museum would discard trustworthy annotations, hence implying a waste of effort and an increase of the workforce demand for this task. The reason why the museum may want to adopt this automated method for analyzing annotations is exactly because of the workforce needed to do it manually. Both scenarios are equally undesirable for the museum, and the accuracy takes both of them into account.

Nr.	Training set	Information sources	Error handling	Accuracy
1	30% Data	Author	No	43%
2	10 per source	Author	No	53%
3	10 per source	Author $\otimes$ Taxonomy	No	60%
4	10 per source	Author $\wedge$ Taxonomy	No	76%
5	10 per source	Author $\wedge$ Taxonomy	Yes	82%

Table 1.1: Results with different strategies. The information source is the object of the subjective opinion computed in each strategy.

Each strategy works on the same data, splitting them in a training set and a test set, but the way these subsets are built changes for each strategy: for instance, some strategies take the first 30% of data as training set, others consider a fixed amount of data for each information source. This, and the other differences explained in the following paragraphs, lead us to different results.

Strategy 1 The first solution adopted is the simplest one: a fixed threshold (see Section 1.3.4). The strategy is presented as follows.

1. Order the annotations by creation date.
2. Use the first 30% of the ordered dataset as a training set.
3. Build an opinion per author in the training set.
4. Use the opinions to evaluate the annotations in the test set: every time we find an annotation created by a given author, we accept it if the author's reputation passes a threshold.

This leads to a poor result, that is, an accuracy of 43%, mainly due to two reasons. First, considering only one kind of metadata and a fixed threshold, once it is established where the author is situated (above or under the threshold, that is, accepted or not), there is no way to adjust his evaluation, since no different point of view is taken into account. Second, since authors are not uniformly distributed in the dataset (some authors started working on the dataset earlier, some later), we cannot gather enough evidence for all annotators. This leads to a conservative consequence: since we have no information to evaluate such annotations, these will always be refused (since false negatives are preferred to false positives), decreasing accuracy.

Strategy 2 The second solution solves one of the previous problems, that is, the non-homogeneous distribution of the authors over the dataset. Here we collect a fixed amount of evidence before using reputations to evaluate the annotation. The strategy can be described as follows.

1. Collect n annotations and their evaluations for each author and compute an opinion about the author.
2. Collect all the remaining annotations of each author.
3. Evaluate the remaining annotations using the opinion computed.

This means that each reputation is used only after having collected a reasonable amount of evidence (all the reputations now have the same uncertainty), and clearly this helps to improve the results. The improvement is quite significant (accuracy 53%) but, although the performance is slightly better than what we would have obtained by tossing a coin, we are still far from a positive result.

Strategy 3 The third solution uses two sources of information, the reputation of the author of the annotation and the reputation of the author of the taxonomy. This second source of information is chosen because a typical reason for refusing this kind of annotation is that the taxonomy used may have become out of date. By looking at the author of the taxonomy, we implicitly take into account the period when the taxonomy was created and the methods used for assessment, which are important

indicators whether the taxonomy is out of date. Moreover, it incorporates the previous improvement and takes a fixed amount of evidence for each source of information. This strategy is implemented as follows.

1. Collect n annotations and their evaluations per annotation author.
2. Collect m annotations and their evaluations per taxonomy author.
3. For each annotation left, build an opinion by merging the opinion about the author with the opinion about the taxonomy author, and accept it if the expected value of the resulting opinion passes a given threshold.

These improvements lead us to an accuracy of 60% which, although far from an optimal result, again shows a substantial improvement. This solution is important because it shows how it is possible to successfully merge contributions from different sources in order to obtain a more precise result.

Strategy 4 The fourth solution reaches 76% of accuracy. This variant builds opinions based on the performance of each author with each taxonomy. Compared to the previous version, which took the two reputations and merged them, this is more precise, since it evaluates the contribution given by these reputations, taking also into account the existing relation between the subjects to which these reputations belong. This strategy is implemented as follows.

1. Collect n annotations and their corresponding evaluations for each combination (annotation author, taxonomy author).
2. For each annotation left, build an opinion based on the evidence available about the author of the annotation intersected with the taxonomy author, and accept it if the expected value of the resulting opinion passes a given threshold.

So, when an author *a1* has a certain reputation, this is computed according to his behavior over time. The same can be said about the taxonomy *t1*. By analyzing the annotation made by *a1* using *t1*, in the previous strategy we merged their reputations, which were considered two distinct inputs. This approach is quite realistic, since it simulates the case when we collect opinions coming from different sources about different metadata of the same annotation. But using this strategy we can be more precise, by looking at the reputation of the author with a particular opinion, that is, the reputation of $a1 \wedge t1$.

Strategy 5 The fifth solution gives the best result: 82% of accuracy. It starts from the improvement achieved with the previous strategy and adds an error handling procedure. This procedure monitors the behavior of the system, and checks if annotations accepted by the model are really correct annotations and vice-versa. So, beyond evidence about authors of annotations and taxonomies, the procedure collects also this kind of evidence and, in case the accuracy goes below a certain threshold, then it firstly

improves the reputation of the considered sources by collecting new evidence about them and secondly collects new evidence about the behavior of the system, in order to see if the more accurate reputations did actually improve the system behavior. This strategy is implemented as follows.

1. Collect n annotations and their corresponding evaluations for each combination (annotation author, taxonomy author).

2. Evaluate k annotations created by a given author using the taxonomies created by a given taxonomy author using the opinion based on the evidence of the intersection of the two (annotation author $\wedge$ taxonomy author, as in Strategy 4).

3. Evaluate the performance of the algorithm (using some spare evaluations): if the accuracy is above a given value, then continue to point 2, else go to point 1 (and hence collect new evidence and improve the accuracy of the opinion).

1.5 Discussion

This work represents a first investigation on the issues related to the estimation of Web data trustworthiness. Because of this explorative nature, the conclusions we can draw from it are rather limited. We can note that the main focus of the analyses is, given a subject and a property, the value of the object of RDF triples, in particular when this value is a categorical value (URI). Also, another important aspect is the relevance of metadata in the estimation process, in particular the identity of the annotator creator. These are basic elements of the method proposed that will be further investigated in the next chapters, and that constitute founding elements of the methods for trustworthiness estimation developed throughout the thesis. We propose a series of strategies for trustworthiness estimation and decision strategies, because we consider it as an open problem, and we propose different plausible solutions for this. In particular in Chapters 6, 7 and 8 we address these issues again, and we propose a decision strategy that reduces to one the amount of arbitrary parameters to set. In general, although in this chapter we evaluate the procedures proposed in one specific domain (natural history), throughout the thesis we propose variations of these procedures and we apply them on disparate domains (e.g. cultural heritage and naval). Lastly, we can suggest best practices which could help to reason about trust and to represent it.

1.5.1 Best Practices

From the Naturalis Museum case study, we see how our model could be improved by the adoption of some best practices by the authority. These may include:

- The use of RDF as a standard language for metadata representation. Although any database can be easily "triplified" (using, for instance, D2RQ [13] and

Triplify [1]), since RDF is the standard technology for metadata representation, its usage is desirable.

- The usage of references (URIs) to standard knowledge repositories for annotations. Instead of building an internal knowledge repository for annotations, if possible, it is preferable to refer to repositories offered by authorities in the field. For instance, a taxonomy used to annotate may frequently be taken from standard authorities. In the case of biological taxonomies, for instance, the U.S. National Biological Information Infrastructure offers an authoritative database of known taxonomies. This helps to keep the meta-information about annotations consistent and uniform.

- Keep a log and profile for each annotator. From the profile, for instance, we can retrieve information useful to assess a precise *a priori* probability for an annotator's ability to annotate.

- Record physical information about the annotated object. Any kind of evidence useful to assess the correctness of annotations should be recorded and evaluated. In particular, this kind of data can reveal a direct link between an annotated object and its annotation, by the coincidence of e.g., shape, color or dimensions of the object and, for instance, the species represented by the taxonomy.

1.6 Conclusion

This chapter introduces a model for deciding whether to trust museum annotations. The model uses Web sources to enrich the data to be evaluated and makes use of Semantic Web technologies to uniformly represent the enriched data. By enriching the data we obtain meta-information that allows us to discover regularities in the data that can be used as a basis for trustworthiness estimations. The model relies on a combination of data enrichment with subjective logic: Web data offer the information needed to find the missing links in the internal dataset (e.g., the fact that two records share the same taxonomy author), and subjective logic allows us to easily apply probabilistic reasoning over it, while taking into account the fact that, possibly, the set of observations at our disposal is limited. We propose a series of strategies for data selection, opinion computation and opinion handling. We also propose an error handling strategy, and our model reaches up to 82% accuracy.

2

Estimating Trust in Confidence Heuristics

Analyzing Georeferenced Animal Specimens

*This chapter continues the exploratory work started in Chapter 1 and analyzes the use of (Web-based) heuristics for estimating the confidence in georeferenced animal specimen entries. We saw in Chapter 1 that a Web source can be used to derive data trustworthiness. Here we analyze the reliability of different heuristics, part of which are Web-based, as confidence indicators, thus further addressing the first research question (*Can Web data help the trust evaluation of semi-structured data?*). Like in Chapter 1, here we use subjective logic to handle the evidence available. We extend the set of statistical techniques adopted (tackling the second research question,* How can uncertainty reasoning be effectively used to estimate the trustworthiness of semi-structured data?*), as to include statistical hypothesis tests to evaluate the reliability of the scores computed by means of subjective logic. We use these tests also in Chapters 4 and 6.*

This chapter is based on the paper Georeferencing Animal Specimen Datasets *coauthored with Marieke van Erp, Robert Hensel from the VU University Amsterdam and Marian van der Meij from the Naturalis Museum in The Netherlands and accepted for publication in the journal* Transactions in GIS. *I report a significant part of that paper, and in Section 2.7 I extend the analysis of the heuristics for determining the confidence in the georeferencing process that was the main contribution I provided in that paper.*

2.1 Introduction

In a previous work [159] we proposed a method for georeferencing the location of a collection of specimen findings of The Netherlands Centre for Biodiversity Naturalis

(Naturalis) [116] and, together with it, we present an analysis of the confidence in the results of the geodisambiguation process. Here we report the main findings of that work, we further extend such analysis, and we evaluate it. The majority of objects in the Naturalis collection were collected a long time ago, often in countries that were previously colonies of the Netherlands. As the oldest parts of the Naturalis collection date back to the 18th century, most collection records consist of a textual description indicating location(s) and offset(s) such as "Anti-Atlas, 10-20 km S. Ait-Baha, Morocco" rather than precise geographical coordinates. With semi-automatic approaches such as the MaPSteDI method [114], georeferencing a record reportedly takes approximately five minutes per record. As Naturalis harbours 37 million objects with each their own record, manually georeferencing each record would be a time-consuming and costly undertaking. This is an important similarity with the problem faced in Chapter 1. To address this challenge, we have developed an automatic georeferencing approach that uses domain knowledge about species, that is, their geographical distribution from the online Global Biodiversity Information Facility [61]. This approach has been realized in a prototype currently being tested at Naturalis. Moreover, we also developed a series of heuristics aimed at quantifying the reliability of the results of the georeferencing process. This chapter aims at determining the trustworthiness of these heuristics and, for this purpose, it presents an extensive analysis of their trustworthiness, separately and combined with each other. In Chapter 1 we use Web data as a basis for estimating the trustworthiness of semi-structured data. Here we progress on that direction and we analyze the trustworthiness of different heuristics (including Web data-based ones) and their combination as confidence indicators for semi-structured data. To do so, we make use of subjective logic as we do in Chapter 1 and of statistical methods that are analogue to the statistical hypothesis tests that we use in Chapters 4 and 6.

The remainder of this chapter is organized as follows. In Section 2.2, we describe the challenges in georeferencing biodiversity data, followed by previous work in Section 2.3. In Section 2.4, we describe the datasets we used. Our georeferencing approach is described in Section 2.5, followed by the results in Section 2.6. Our confidence measure is described and analyzed in Section 2.7. Conclusions are discussed in Section 2.9.

2.2 Georeferencing Challenges

The problem of georeferencing natural history collections is not new: the different types of challenges have been categorised and described by Beamann and Conn [6]. In Table 2.1, we illustrate each of the challenges by an example from the Naturalis collection. It is not possible to georeference all types of localities with equal precision. Vague localities, such as "Southeast Michigan", simply contain too little information to pinpoint a spot within a small range (<5 km) of the actual finding location, but for localities containing for example linear feature measurements such as "16 km N of Murtoa" this is feasible. However, although the result of the georeferencing process

can be of higher or lower precision, it can be accompanied by a confidence score that facilitates to understand how precise the measure is. Since the confidence score is estimated using the same data used for georeferencing in combination with Web data, the challenge here is to determine how trustworthy such a score is. The georeferencing process is based on the analysis of several geographical types of information (the name of a region, of a city, the distance from a point, etc.). See Table 2.1 for a complete list of possible geographical information at our disposal. The confidence score can be based on estimations of the availability and the quality of this information. Understanding which of them or which combination of them is the most trustworthy as a basis for computing a confidence score is the challenge we tackle in this chapter.

Challenge posed	Example textual locality
Two or more locations that share the same name	"Amsterdam"
Two or more location descriptors	"Wakarusa, 24 mi WSW of Lawrence"
Topological nesting	"Moccasin Creek on Hog Island"
Complex interpretative description	"Bupo [?Buso] River, 15 miles [24 km] E of Lae"
Linear feature measurement	"16 km (by road) N of Murtoa"
Linear ambiguity	"On the road between Sydney and Bathurst"
Vague localities	"Southeast Michigan"
Political borders change over time	"Yugoslavia"
Historical place names	"British North Borneo"

Table 2.1: Georeferencing challenge and example of corresponding geographical information from the Naturalis collection.

2.3 Related Work

There is a fair body of research on georeferencing both outside and inside the domain of natural history. Within the natural language processing community georeferencing is treated as follow-up task to named entity recognition [100, 105], or possibly as complementary to it [62]. However, these approaches assume full text, whereas the datasets in the natural history domain are part of structured database records, making them suboptimal for this domain.

Most approaches for structured data use some sort of gazetteers combined with some form of reasoning to disambiguate and ground location names [92, 101, 103]. These assume that the location names have been identified, skipping the step of recognizing the location name and possible extra locality information. Other approaches derive toponyms disambiguation from combining the analysis of events with a gazetteer [137]. In our case study we do not have at our disposal event descriptions, but the procedure of information extraction, gazetteer lookup and geodisambiguation

is very similar to the one described in that work.

Another emerging relevant research thread regards the use of crowsourcing for acquiring information useful for georeferencing [44, 91]. This kind of approaches assumes the availability of a crowdsourcing platform and of a population of information contributors we do not have at our disposal. However, these approaches are complementary to the procedure described in this chapter, since the crowdourced information might help in improving the disambiguation heuristics. Hence, we will possibly investigate their integration in the future.

Within the natural history domain, several attempts have been undertaken to automatically assign coordinates to textual descriptions of locations in specimen datasets, with BioGeoMancer [72] as its most well-known application [6, 72]. BioGeoMancer provides an application for text processing, interpreting, gazetteer querying (using a variety of sources [10]), intersecting spatial descriptions and as a result returning a standardized geographical reference including uncertainty levels. The initial version of BioGeoMancer supported interpretation of localities in English, Spanish and Portuguese. However, the latest available version of BioGeoMancer supports English queries only.

Also developed for georeferencing natural history data is GeoLocate [136]. It uses similar gazetteer data as the BioGeoMancer project[1]. GeoLocate uses different georeferencing heuristics as well as additional linear features to its gazetteer such as rivers, road, legal land descriptions and river miles. These additional information sources can lead to more accurate results, but are only available for the United States, Canada, and Mexico. A comparison of automated georeferencing tools found that, at the time, GeoLocate was the best software tool to efficiently georeference large datasets [114].

Although both focused on the biodiversity domain, neither BioGeoMancer nor GeoLocate makes use of domain-specific knowledge, such as species occurrence data. Also neither can deal with non-English data.

The use of heuristics has been widely employed for trust estimation. In Chapter 1 we show an example of heuristics-based trust estimation, and others are presented in the next chapters. Trust itself, and trustworthiness estimation in particular, can be seen as a heuristics supporting a decision rule. In fact, if one decides to accept or reject something based on trustworthiness estimations, she actually takes a decision by relying on an experience-based technique, that is, a heuristic. Chapter Introduction presents an extensive review of computational trust models, that covers in particular approaches related to those used in this thesis.

2.4 Data

In this section, we describe our primary dataset as well as the resources used for georeferencing and the development of the gold standard.

[1]Since 2006, GeoLocate is part of the BioGeoMancer Workbench, but the current status of the integration of the projects is unclear.

2.4.1 Reptiles and Amphibians Database

Several large datasets of animal specimen datasets are maintained at Naturalis. The information in these datasets comes from the field logs and registers in which biologists who made these finds recorded them manually, usually during expeditions. Part of the information from these sources has been converted to electronic datasets over the course of time by many different biologists working with the specimens. Creating these databases was not a top-down organized undertaking, but rather taken up by the researchers themselves to improve access to the data for themselves.

For our case study, we used the reptiles and amphibians database containing 29,752 records, each referring to a specific animal find. Among the information in these records, one typically finds locality information indicating where these specimens were found, its species, the name of the collector, information about when it was entered into the database and by whom it was entered into the database. The location find information is divided over several different database fields, namely "Town/City", "Province/State", "Country", "Location"[2], "Altitude", and "Coordinates" (only filled in 3.4% of the records). In this contribution, we shall mostly focus on the information from the Town/City, Province/State, Country and Location fields.

Gold Standard

To test our system, we created a gold standard consisting of 200 records, from which we kept 50 records for development and 150 records for final testing. Records were selected with two aspects in mind: common challenges and internal representativeness. The first aspect ensures that the different types of locality information present in the database are represented in the gold standard. The second aspect balances for the fact that some types of locality descriptions are more frequent than others.

Due to the limited resources for annotating the gold standard dataset, we decided to focus on four categories from the initial nine categories in Table 2.1 (presented in Section 2.2). These four categories were selected because they are mostly influenced by the use of background information and thus provide the most suitable types of challenges for the research at hand. In Table 2.2, we show our categories, as well as an example, the distribution of records pertaining to this category in the gold standard and in the entire database.

2.4.2 Gazetteers and Biodiversity Resources

Two geographical gazetteers were used to look up place names: **GeoNames** [60] and **Google Maps** [67]. GeoNames contains about 10 million place names and information about those places, such as coordinates, alternative names, elevation levels and population numbers.

[2]This sometimes contains the town or city value, but more often it is used to describe offsets or particularities of the find, such as that the specimen was found under a branch or in a puddle.

Category	Example	#in gold standard	# in full set
A. Single Place	"Maastricht"	90 (45%)	10,750 (42,7%)
B. Single Place with offset	"18 mls. E. of Kumasi"	20 (10%)	2,363 (9.4%)
C. Two or more places	"Sibil, Sterrengebergte"	62 (31%)	9,150 (36.4%)
D. Two or more places with offset	"Alachua Co., 10 mi S. Gainesville on Wachahoota rd."	28 (14%)	2,856 (11.3%)
Total	-	200	25,119

Table 2.2: Categories for different types of textual descriptions. For each type of textual description, we report also its relative frequency in the gold standard and in the overall dataset.

For biodiversity background data, we use **The Global Biodiversity Information Facility (GBIF)** [61]. GBIF is the largest online portal for biodiversity data. As of November 2011, the portal contains 312 million records, of which 271 million also contain coordinates. These records come from the combination of many individual datasets provided by institutions from around the world. A study on the accuracy of geographical data in GBIF records [180] showed that the majority of the records were annotated with correct coordinates (83%), but the relatively large amount of incorrectly georeferenced records is something that has to be taken into account when using this data.

2.5 Georeferencing Approach

Our georeferencing approach consists of five automatic, rule-based modules that form the pipeline through which each record from the gold standard is processed:

1. **Record Retrieval.** This module filters the database record to include only those database fields used by the system, that are: "Town/City", "Province/State", "Country", "Location", "Altitude", "Collection Date", "Genus", and "Species".

2. **Text Parsing** In the parsing module, sentences are split and tokenized. Then tokens are matched against patterns and keywords to recognize indicators for offsets (such as cardinal directions and units of measurement), place names and common words in Dutch and English.

3. **Gazetteer Lookup.** Identified location name candidates from the text parsing module are looked up in GeoNames [60] and Google Maps [67].

4. **Offset Calculation.** If an offset, such as "112 km S El Dorado", is encountered, coordinates retrieved from the gazetteer for the place of reference ("El Dorado")

need to be combined with the offset ("112 km South") to calculate the final coordinates. For the calculations of the coordinates we use the Perl Geo::Calc [126] module.

5. **Disambiguation Heuristics.** As many place names share the same name ("Amsterdam, the Netherlands" vs. "Amsterdam, MO, USA") or similar names ("York, UK" vs. "New York, USA"), several disambiguation heuristics were selected to disambiguate location names.

In the remainder of this section, we will detail each of our disambiguation heuristics.

Spatial Minimality The spatial minimality heuristic is a fairly standard statistic in georeferencing and relies on co-occurrence of geographic entities within the same discourse. This heuristic assumes that, in a text which mentions more than one location, the cluster of physical locations in the world that are most closely related by distance are the most likely candidates to be actually referred to. We start with a list of potential candidates for each place name and their corresponding coordinates and match each candidate to every possible combination of candidates from the other place names. For each of these combinations the system creates a polygon that encloses these candidates. The system selects the smallest polygon, and the set of candidates used to create that polygon are seen as the most likely candidates.

Expedition Clusters The spatial minimality heuristic uses only information from within individual records. However, specimen database records are not independent. The Expedition Clusters heuristic assumes that information from similar records can be used to aid georeferencing. Work on this same dataset by van Erp [158] shows that it is possible to use information available in the dataset to rediscover expeditions from a dataset. Information about which expedition a record belongs to is only explicitly available in a small number of records, but it is "re-discovered" by using data such as collection date and country. Enriching the data in such a way enables comparison between records which would otherwise not be possible. For example, it is very unlikely that two records from the same expedition are in entirely different locations. Thus, if such an anomaly was to be detected it would be a clear signal that one of the records is incorrectly georeferenced. Furthermore, the information can be used for disambiguation of place names as also suggested in the work of Guo et al. [71], to increase confidence in the outcome of the georeferencing process. A candidate for a place name that is close to the previous georeferenced location record (when that record belongs to the same expedition) will be assigned a higher confidence measure.

Species Occurrence Data Occurrence data from existing specimen finds can be used to check if new data fits the currently known locations for species. In the current implementation, this data is retrieved solely from GBIF as, at the time of writing, this is the only openly available resource containing such information. This information is used to disambiguate location descriptions and validate results in much the same way

	Accuracy @5km	Accuracy @25km	Accuracy @100km	Mean distance off	Not Found
Baseline	38.9%	47.0%	58.4%	251.1km	26.2%
+ Google Maps & Fuzzy match	53.0%	65.1%	74.5%	244.1km	8.7%
+ Spatial Heuristics	59.1%	71.8%	77.2%	171.1km	7.4%
+ Expeditions	59.1%	71.8%	77.2%	171.1km	7.4%
+ GBIF	61.7%	74.5%	79.9%	114.5km	7.4%

Table 2.3: Accuracy of the georeferencing heuristics within 5 km, 25 km and 100 km of the gold standard coordinates, compared to baseline in percentages. The table also shows the mean distance the different heuristics were off, as well as the percentage of cases for which no coordinates were found by the system.

	Precision	Recall	F_1
Baseline	64%	47%	54%
+ Google Maps & Fuzzy match	71%	65%	68%
+ Spatial Heuristics	78%	72%	75%
+ Expeditions	78%	72%	75%
+ GBIF	80%	74%	77%

Table 2.4: Precision, recall and F-measure of the different heuristics at 25 km from the coordinates in the gold standard.

as the expedition heuristic. By querying GBIF data, coordinates are retrieved for all currently known finds of the species in the record. Each coordinate for a previously found specimen find is then compared to each place candidate, and based on the closest specimen find to a candidate a confidence measure is assigned to the candidate; the smaller the distance to a candidate the higher the confidence. The confidence measure is detailed in Section 2.7. Note that GBIF can only be used to give approximate locations, as species' localities may change over time. However, together with the geo-information within the database record, it helps the system disambiguate between, for example, different continents.

2.6 Geodisambiguation Results and Discussion

All presented results are measured by applying the heuristics in our knowledge-driven georeferencing approach to the 200 records that were manually georeferenced for the gold standard (see Subsection 2.4.1). We computed a baseline score to compare our approach to a simple look-up approach by retrieving the coordinates of the first location name found in the record, looking up this name in the GeoNames gazetteer, filtering by country and province and returning coordinates of the first candidate. Ta-

ble 2.3 presents the accuracy results of the different modules on the test set. Table 2.4 presents the precision, recall and F-measure of the best system at 25 km. Application of the t-test shows that all modules provide significant improvement over the baseline at $p\text{-}value < 0.005$.

The spatial minimality heuristic improves results for records that contain more than one place name (50%), but with some caveats. The first implementation included each location found in the record (Place, Location and Province/State). Because the location field is a free text field, it contains long sentences in a number of records, negatively affecting the rule-based system to recognise location names. However, in many other cases the location field does contain useful information, so it was decided not to parse any location fields with a length exceeding 60 characters. Also, the spatial minimality heuristic performs better if the Province/State field is not considered. Provinces and states generally cover larger areas, but the gazetteer will return only one single point that does not represent this fact. As such, these points do not add much information on a smaller scale and pollute the created polygons. Since the country is usually known, this already dramatically decreases the area that has to be searched. As a result, the heuristic mainly improves results that were not too far off to begin with.

As our data are in Dutch, we could not run our data in BioGeoMancer and GeoLocate. For GeoLocate it is also the case that only georeferencing in the USA is supported. We could also not get hold of the data they tested their systems with. Therefore, an exact comparison of our system to BioGeoMancer and GeoLocate is not possible, but we have strived to set up our experiments in a similar fashion. We therefore assume that our results for the spatial heuristics are in the same ballpark as those reported in the work of Murphey et al. [114].

Although the results in Tables 2.3 and 2.4 seem to indicate that the expedition heuristic does not improve the results, manual inspection of the records showed that the heuristic does add valuable information. For now this information mostly affects the confidence score (see Section 2.7), and we attribute the lack of improved scores to the configuration of our gold standard dataset. As our gold standard contains a sample of random records from across the entire dataset, the number of records belonging to the same expedition in this sample is small, and as such these small clusters add little evidence to support the disambiguation process. However in this chapter we focus on measuring the trustworthiness of the heuristics, and in the future we will address the issue of improving this and possibly other heuristics (e.g., by increasing the number of records belonging to the same expedition in our gold standard).

The use of GBIF Species Occurrence Data is especially useful in situations for disambiguation of location names in a large geographical area (notice that this is why the mean distance off improves more than the percentages of correctly georeferenced localities). If a specimen find is only annotated with the place name “Sibil”, a list of 20 possible candidates would be retrieved from the GeoNames gazetteer in different continents. By cross-referencing these candidates with existing finds of the species (“Sphenomorphus schultzei”), only two likely candidates remain: “Ok Sibil, Papua, ID” and “Sibil, Papua New Guinea”, greatly decreasing the search space.

Category	Accuracy @5km	Accuracy @25km	Accuracy @100km	Mean Distance Off	No Result
A: Single Location (67)	58.2%	64.7%	68.7%	140.1 km	16.4%
B: Single Location + offset (15)	86.7%	100%	100%	1.7 km	0%
C: Multiple Locations (46)	60.9%	73.9%	82.6%	146.4 km	0%
D: Multiple Locations + offset(s) (21)	57.1%	85.7%	95.2%	54.9 km	0%

Table 2.5: Results split out per category based on best results from Table 2.3 (GeoNames + Google Maps + fuzzy search + spatial heuristics + expeditions + GBIF). The numbers behind the categories indicate the number of records in that category.

Care needs to be taken however that on a smaller scale, the heuristic should not be used too rigorously, since it will only favour locations that fit within the existing data model and many species occurrences are spread out across an area. Furthermore, a significant part (16%) of the geographical data in GBIF records was found to contain errors, as demonstrated by Yesson et al. [180]. Species occurrence records for "Sphenomorphus schultzei" show that the species was found on multiple locations across the island "New Guinea", in an area of almost 600,000 km. In this case, the occurrence data should not be used for disambiguation of the two remaining candidates on this island.

The results for different categories presented in Table 2.5 show that records that are annotated with one single location name and an offset (category B) are georeferenced with a much higher accuracy than other categories. Obviously, the textual complexity of these records is limited, but there are two other points of interest. In each of these cases, there is no problem with the distinction between administrative areas (provinces, states) and populated places (cities, villages) since it is obvious that an offset will always be from a populated place and not from a province. Secondly, the offsets usually appear to be from a well-known (or important) place. A major difficulty in geo-referencing biological collections is the use of place names that are only locally known. The location of place names such as "Meyers' farm" or "Base Bivouac" might be very well known during expeditions and to local inhabitants. However, it is nearly impossible to use this information on its own without the use of very specific information sources such as the field logs and maps created for specific expeditions. In specific implementations, one could consider manually creating an additional gazetteer for such places.

As can be seen in the third column, the other categories (A, C and D) have an almost similar score for correct matches within 5 km. However, results for single location names (category A) show that the number of additional places found within 25 or 100 km is limited, whereas records with more than one location show improvements. The georeferencing process produces more precise results if records that are annotated with more than one location are provided with contextual information. For example,

when encountering a description such as “Lake Jaroe, Kampong Gariau, Indonesia”, “Lake Jaroe” does not occur in any generic gazetteer. The record can still be georeferenced using the more generic location “Kampong Gariau”, but this means the record is georeferenced to a location several kilometres away from the correct location, decreasing accuracy.

2.7 Measuring Georeferencing Confidence and Heuristics Trustworthiness

There is a large number of potential uncertainties in the georeferencing process. These stem from the data itself, external data-sources used, and the process of linking data to these external sources. It is important that these sources of uncertainty are identified and recorded, so as to be able to calculate a confidence score (CS) for the resulting georeferenced locality. Although Graham et al. [69] found that “species distribution modeling approaches in general are fairly robust to locational error”, not having information about the uncertainty of georeferenced localities makes it impossible to know if this geospatial data is suitable for a specific purpose and it may thus be of little use as also suggested in the work of Wieczorek et al. [173] and of Guo et al. [71].

Inspired by a manual confidence value system used in the MaPSteDI method [114], a scale from -12 to 12 is used to *automatically* indicate the confidence in a georeferenced locality (12 indicating the highest degree of confidence, -12 lowest). This automatic measure represents the confidence that the returned coordinates for a georeferenced location are accurate. The confidence measure is based on several different indicators presented in Table 2.6. Each heuristic can increase or decrease the confidence. For example, based on the spatial minimality heuristic, the confidence will be increased if the polygon describing the area of co-occurring place names is very small or decreased if very large. If a record belonging to the same expedition is georeferenced to a location that is close to other specimen finds from that same expedition, the confidence is increased. For instance, in our dataset we have a record for which: the country is known (+2), the province is unknown (0), the location description contains unknown words (-1), the place description is found in GBIF (+1), but a fuzzy search in the gazetteers does not return a positive result (-3) and the place description is found only in Google Maps (-3), gets a confidence score of -4. The fact that the distance of the georeferenced location of this entry from its actual location is 1,204.72 meters, confirms the indication given by the low confidence score. In the rest of this section we analyze the trustworthiness of the heuristics in depth. In principle, a high confidence score may not indicate a high georeferencing error, but just a georeferencing entry that is certainly correct. Still, we assume that there should exist a negative correlation between error and confidence score because in many situations the result of the geodisambiguation process is an approximation (we always return a georeferenced value, also when the confidence is low), and the confidence score derives the quality of the georeferencing from the degree of approximation adopted.

Level	Indicator	Points
Record	Country known	$+2$
Record	Province known	$+1$
Record	Unknown parts in description	-1
GeoNames Result	Place not part of province	-1
GeoNames Result	Fuzzy string search	-3
Candidate	(SM) Close together	$+x, x \in [0, 2]$
Candidate	(OD) Close to GBIF	$+x, x \in [-1, 2]$
Candidate	(EXP) Close to previous find	$+x, x \in [-2, 2]$
Candidate	GeoNames Candidate very close to Google Maps	$+x, x \in [0, 2]$
Candidate	Only found on Google Maps	-3
Candidate	GeoNames first candidate	$+1$
Candidate	Administrative area	-1

Table 2.6: Calculation of the Confidence Score. SM denotes the spatial minimality heuristic, EXP denotes the expedition heuristic, and OD indicates the use of species occurrence data. Each heuristic contributes to the final score based on the points indicated. These values have been determined manually. The final, aggregated score ranges between -12 (low confidence) and +12 (high confidence) and, therefore, the average confidence score is zero.

2.7.1 Confidence Heuristics as Subjective Opinions

The most important component of the confidence score is the amount of information available in a record. A single place name with structured additional information about the province and country such as "Santa Bárbara, Amazonas, Brazil" can usually be retrieved with a higher confidence than a single description such as "Forest between 20-10 km from Ambohaobe". Therefore, the latter record receives a lower confidence score based on absence of country and province information. Secondly, the confidence score is based on the consistency and type of input data from gazetteers and biodiversity resource. For example, if no direct match in a gazetteer is found but a result is found using fuzzy matching, that result will still be used but with decreased confidence. If a georeferenced location is consistent with existing occurrence data from GBIF, this will increase the confidence.

The extent to which certain variables influence the accuracy cannot always be determined and, as such, makes the method fallible. In some cases, there is simply not enough information to determine an indicative confidence measure. To estimate the reliability of our confidence measure, we treat it as an estimated observation about the correctness of the corresponding georeferenced entry. Similarly to the work described in Chapter Preliminaries, this estimated evidence is used to build a Beta probability distribution (represented by means of a "subjective opinion") that describes the probability of each confidence score in the interval $[0, 1]$ to represent the trustworthiness of the entry.

The contribution of each heuristic has a different weight on the computation of the final confidence score. We treat each heuristic as a subjective opinion for three reasons:

- Each heuristic can be seen as an opinion from a different point of view about the quality of the disambiguation process. If, for instance, a location description contains unknown words, then it will probably be hard to geolocate it. The same holds for all the other heuristics. An aggregation of all the points of view hopefully provides a comprehensive view about the geodisambiguation confidence.

- Each heuristic weighs differently to the final aggregated confidence score. This fact is easily encoded by a subjective opinion, as opinions from heuristics that have a heavier weight will provide less uncertain and hence stronger opinions.

- By representing the heuristics as subjective opinions we take into account the fact that the heuristics base their score on the presence or lack of specific evidence (e.g., the presence of unknown words in the description). However, the lack of a specific evidence constitutes a piece of evidence per se (since, for instance, it is easier to disambiguate descriptions that do not contain unknown parts).

For the conversion to subjective opinions we adopt the procedure described below.

Mapping 1 The representation of a heuristic score as a subjective opinion is made as follows. Given a heuristic, we consider its value and its range: if the value of the heuristic equals the upper bound of the heuristic, then the heuristic provides only positive evidence, if it is equal to the lower bound only negative evidence. That is straightforward. Two issues are left open: how many pieces of evidence does each heuristic provide and how are the heuristic scores between the two bounds converted. This issue is solved by considering that the heuristics provide only integer results and, therefore, we interpret them as "counts of evidence" and we compute the corresponding evidence as follows:

$$positive\_evidence = h - min$$

$$negative\_evidence = max - h$$

where min and max are the lower and upper bound respectively, and h is the actual value of the heuristic. So, the opinion expressed by a heuristic h about a given *item* is:

$$\omega^{h}_{item}\left(\frac{h-min}{max-min+2}, \frac{max-h}{max-min+2}, \frac{2}{max-min+2}\right)$$

This representation of the heuristic as a subjective opinion takes into account the weight expressed by the heuristic itself.

Mapping 2 In case we want only to test the ability of the heuristic to grossly highlight big errors, without considering the actual weight of the heuristic, we adopt the following mapping:

$$positive\_evidence = pos(h)$$
$$negative\_evidence = neg(h)$$

where *pos* is defined as follows:

$$pos(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

and *neg* is its symmetric variant:

$$neg(x) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{otherwise} \end{cases}$$

In this case the opinion becomes:

$$\omega^h_{item}\left(\frac{pos(h)}{pos(h)+neg(h)+2}, \frac{neg(h)}{pos(h)+neg(h)+2}, \frac{2}{pos(h)+neg(h)+2}\right)$$

A heuristic h obtained by summing up different single heuristics ($h = h_1 + \cdots + h_n$) is mapped as follows:

$$\omega^h_{item}\left(\frac{\Sigma_i pos(h_i)}{\Sigma_i pos(h_i)+\Sigma_i neg(h_i)+2}, \frac{\Sigma_i neg(h_i)}{\Sigma_i pos(h_i)+\Sigma_i neg(h_i)+2}, \frac{2}{\Sigma_i pos(h_i)+\Sigma_i neg(h_i)+2}\right)$$

where $\Sigma_i pos(h_i)$ is the sum of the *pos* function applied to all the heuristics and $\Sigma_i neg(h_i)$ the sum of the application of *neg*.

Mapping 3 This mapping is a simple variant with respect to Mapping 2, as it applies *pos* and *neg* directly to the value h of the heuristic and computes the corresponding subjective opinion independently of the heuristic being the result of the aggregation of other heuristics or not.

Each heuristic is utilized as a piece of evidence for the correctness of the estimate: when the heuristic provides a positive value (e.g., the country of a record is known), this counts as a positive piece of evidence; when the heuristic provides a negative score (e.g., in case of unknown parts in the record), this counts as a negative piece of evidence. In fact, each heuristic can be seen as an indication about the possibility to correctly georeference the record. The more heuristics positively indicate the possibility to correctly geolocate, the more confident we are about the geolocation. Of the resulting subjective opinion (or of the corresponding Beta distribution), the expected value represents the value having the highest probability to be the right confidence score, while the variance is a measure of the uncertainty in choosing that value as a correct confidence score.

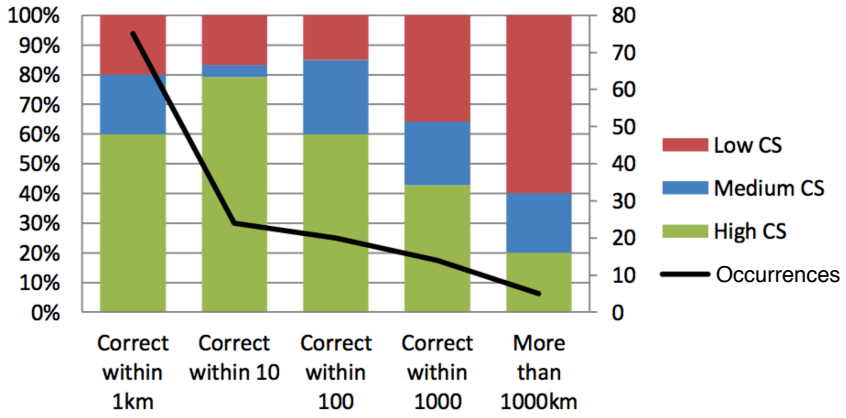


Figure 2.1: Distribution of confidence scores for georeferenced locations of the gold standard. The black line displays the number of occurrences in each category (right axis scale).

2.7.2 Measuring the Quality of the Confidence Heuristics

We start by analyzing the aggregation of all the heuristics. A Shapiro-Wilk normality test at 95% confidence level shows that both the error in the georeferencing process and the expected values of the Beta distributions computed using the heuristics are not normally distributed. Therefore, we use Spearman's rank correlation test [146] at 95% confidence level to check the existence of a linear correlation between the two series of values. In particular, since Spearman's test compares the rank between variables (without taking into account their differences), we standardize the distances and we round them (to 7 decimal digits) because we do not expect our confidence scores to be extremely precise; rather they should help us to distinguish between good and bad georeferences. The test results in a weak negative correlation (-0.15 using Mapping 1, -0.22 using Mapping 2 and -0.12 using Mapping 3), as shown in Figure 2.1. This suggests that the procedure is not always able to compute a confidence score that resembles the real trustworthiness of the result of the georeferencing process, and there is still considerable room for improvement. Also, another Spearman correlation test at 95% confidence level shows a weak positive correlation (0.21) between the variance of the Beta distribution based on the heuristics and the error of the georeferencing process, indicating that the more certain the score is, the lower the error.

So, the first results are moderately positive. However, we want to go deeper in our analyses, to better understand the following aspects.

- What is the trustworthiness of each heuristic when used alone. We hypothesize that the different heuristics have a different capability to predict the precision of the georeferencing results, at least for the considered dataset. By analyzing the

trustworthiness of the heuristics alone, we understand how important they are and what is their impact on the final aggregated score that was analyzed before.

- What is the best combination of heuristics in terms of trustworthiness. By aggregating the different heuristics we might end up in a situation where their contribution is summed up (in case all the heuristics have the same correlation with the actual error in the georeferencing results), annihilated (in the opposite case), or are situated in between these two extremes. By looking at the trustworthiness of the aggregations of all the possible combinations of heuristics, we want to understand whether it is better to consider all of them together or whether by using only a subset of them we obtain a better performance.

The following subsections address these issues.

2.7.3 Single Heuristics Analysis

As a first analysis, we want to understand the trustworthiness of each heuristic when used alone. We computed Spearman's Rank-correlation test of each heuristic using both Mapping 1 and Mapping 2 (Mapping 2 and Mapping 3 coincide when non-aggregated heuristics are considered alone). Computed on the single heuristics, the two mappings provide results having a very small difference. These results in the two mappings providing the same correlation results that are reported in Table 2.7. For instance, a heuristic having value 1 within [-3,3] is converted in 0.625 using Mapping 1 and in 0.67 using Mapping 2. The fact that we adopt a rank-correlation test makes these differences even more negligible, since this test compares the ranks and not the actual values. We can see from Table 2.7 that some heuristics are better than others to highlight the error of the geodisambiguation. In particular, the presence of unknown parts in a description is a relatively good indicator of the confidence in the georeferencing outcome. On the other hand, even if these heuristics do not present a high correlation, some heuristics present a very small correlation. In all but one case the correlation is negative, as to correctly indicate that a low heuristic value corresponds to a high error and vice versa. However, the correlation estimates are quite low, and this means that only a small portion of the heuristic values of the georeferenced entries actually behaves as wished. The following subsection investigates the performance of combinations of heuristics. A combination of the heuristics might present an improvement with respect to the single heuristics or not, depending on whether different heuristics are able to highlight the confidence of different items (and hence their combination extends the number of correlated items) or not.

2.7.4 Analysis of Heuristics Combination

We saw in the previous subsection that there exists a negative correlation between some heuristics and the errors in the georeferencing process. Now we want to analyze the performance of combinations of heuristics. In fact, a weak negative correlation means that only few of the heuristics value correctly indicate the presence or absence of a

Heuristics	Mapping 1 & Mapping 2
Country known	-0.04
Province known	-0.01
Unknown parts in description	-0.23
Place not part of province	-0.22
Close together	-0.12
GeoNames Candidate very close to Google Maps	-0.20
Close to GBIF	-0.03
Fuzzy string search	+0.05
Correct province	-0.01
Only found on Google Maps	-0.07
Geonames first candidate	-0.05
Close to previous find	-0.11

Table 2.7: Correlation between each heuristic and the georeferencing error. Here, there is no difference between the results using Mapping 1 or Mapping 2, so the outcomes are reported together. This lack of difference is due to the fact that, especially when focusing on a single heuristic, the two mappings do not produce significantly different results because they produce a non-significantly different amount of evidence.

high error in the georeferencing result. If the heuristics are able to correctly indicate the trustworthiness of the georeferencing of different items, then by merging them we might be able to extend this "coverage" and hence obtain a higher correlated heuristic. If this is not the case, we do not obtain any improvement from the aggregation (or, in principle, we could even obtain a worsening).

We calculated the correlation of the entire powerset of the heuristics. For the sake of simplicity, we report only the combinations that have the highest correlation (the correlation value is reported between parentheses). These combinations are:

Mapping 1 (-0.36) and Mapping 2 (-0.32) {Unknown parts in description, Place not part of province, Close together, GeoNames Candidate very close to Google Maps, Close to previous find}.

Mapping 3 (-0.35) {Unknown parts in description, Place not part of province, GeoNames Candidate very close to Google Maps, Only found on Google Maps, Close to previous find}.

There is a significant overlap between the two results, although they do not coincide. These combinations comprise the heuristics having highest correlation when analyzed alone. The correlations of these combinations, however, are lower than the sum of the correlation (considering the absolute value of the correlation), because some items that correlate with one heuristic correlate also with another one, but these are still higher

than the heuristics alone. This tells us also that some heuristics chosen are correct and useful, especially if aggregated. However, these would be better integrated with other heuristics capable to correctly indicate the confidence of the items that are not correlated by the existing heuristics. Also, as it was already predictable, the heuristics that alone do not correlate with the errors are useless also when aggregates with other heuristics.

2.8 Discussion

We propose three different mappings between the confidence score used in the case study and trust levels expressed in terms of subjective opinions (and hence in probabilistic terms). Then, we measure the correlations between these probabilistic scores and the actual error, in order to determine the reliability of the scores themselves, that is, to check if the scores are actually able to identify geodisambiguated items with a high chance to be incorrect. Several implementations for the mappings are possible, as well as it is possible to adopt different correlation metrics to test the reliability of the confidence scores. These different possibilities are intended to leave to the analyst the freedom to choose the implementation that best suits his needs and assumptions. However, despite the different implementations, the method adopted is uniform and can be described as follows. First, we identify the following elements:

- estimates (or predictions) obtained from a georeferencing process;
- confidence scores for the estimates (or predictions).

Then, we identify the following procedure:

Select the relevant features or heuristics. If different features or heuristics are available, in first place it is necessary to select those of interest.

Map the confidence scores into probabilistic scores. We adopt a uniform representation for the scores. If we adopt subjective opinions, we allow to measure not only the belief or disbelief in the correctness of a given measurement, but also the possible uncertainty in its correctness.

Measure the errors in the estimates. In order to check if the confidence score is reliable, we need to check if it is really able to hint at the correctness of the measurement.

Measure the correlation between the scores and the errors. It is possible to use different correlation coefficients. All are intended to measure the correlation between the confidence score represented in probabilistic terms and the actual error, in order to evaluate the first based on the latter.

This generic procedure allows to measure the reliability of confidence scores expressed in disparate manners. Also, it is easily extensible to be used in closely related

fields (e.g., in recommender systems, user ratings are expressed in one to five or one to ten scale. We could use an adapted version of the procedure above to measure the reliability of recommendations made to the user). In Chapter 4 we propose an extension of this procedure to measure the reliability of police open data.

2.9 Conclusion

We have presented a method to automate georeferencing of records in animal specimen datasets, and an extensive analysis of the use of heuristics to estimate the confidence in the georeferencing process. Several heuristics for the disambiguation of location names that use domain knowledge from external resources and reasoning were implemented and tested.

The complexity of the georeferencing task is not to be taken lightly. A substantial amount of specimen finds are not annotated with enough information to return accurate coordinates, and generic gazetteers are only partially suited for the natural history domain as they often lack information on location names mentioned in locality descriptions. For this reason, we have developed a series of heuristics aimed at indicating the confidence in the georeferencing result and we have evaluated them in depth. Our confidence measure proves useful in some of cases, pointing experts at Naturalis to the most problematic records so they can focus their attention on those cases that require input from a human expert. We have demonstrated that there exists a correlation between the heuristics scores and the actual errors made in the georeferencing process, and that this correlation is strengthened by aggregating more heuristics. However, not all the heuristics are informative, and although the performance of the georeferencing approach can be considered as satisfactory, the performance of the aggregated confidence measure needs further future improvement, in order to strengthen its correlation with the real georeferencing error.

So, there is still much to be gained by combining a domain specific knowledge for georeferencing, however, this knowledge alone is not sufficient to provide a fully reliable set of heuristics for estimating the confidence in the georeferencing results. It is helpful to refer to Web sources in combination with domain-specific knowledge to tackle this task, but a proper validation of such heuristics is always necessary, as to avoid relying on deceptively illusory heuristics, since using data from trustworthy sources does not always guarantee the reliability of trust estimates.

Part II

Uncertainty Reasoning for Assessing Trust

*The study of uncertainty reasoning as a means to assess trust in semi-structured Web data is the leitmotiv of the second part. Having shown in Part I the usefulness of selected Web data for making trust assessments, here we tackle the second research question of this thesis (*How can uncertainty reasoning be effectively used to estimate the trustworthiness of semi-structured data?*) more in depth. Chapter 3 extends subjective logic (that has been adopted in both Chapter 1 and Chapter 2) and proposes statistical models for handling Web data, while Chapter 4 proposes several analyses of Web data using uncertainty reasoning. The extension of subjective logic proposed in Chapter 3 are extensively adopted in Part IV.*

3

Uncertainty Reasoning for Handling Web Data

Statistical Models and Subjective Logic Extensions

This chapter explores the use of statistical techniques for handling uncertain semi-structured Web data (hence addressing the second research question, How can uncertainty reasoning be effectively used to estimate the trustworthiness of semi-structured data?*). Chapters 1 and 2 used subjective logic to make trustworthiness estimations about semi-structured data because this probabilistic logic allows us to estimate the shape of the probability distribution underlying the observations we face. Estimating such a distribution is important to understand how much we can rely on our observations for making trust assessments. This chapter extends Chapters 1 and 2 by using the probability distributions underlying subjective logic (Beta and Dirichlet distributions) and their natural extension (Dirichlet process) to model uncertain semi-structured Web data suitable for trustworthiness estimation. These kinds of probability distributions, also called higher-order probability distributions because these abstract over the available sets of observations, allow us to still use the uncertain data while compensating for their uncertainty. We evaluate the use of these distributions using a dataset from the maritime domain.*
Moreover, we propose three extensions of subjective logic: one to make use of Dirichlet processes, one to deal with partial observations and one to incorporate semantic similarity measures. We provide a theoretical validation of these extensions. In particular, the use of semantic similarity measures within subjective logic will be extensively adopted in Chapters 7 and 8.

This chapter results from the merge of the paper Estimating Uncertainty of Categorical Web Data *coauthored with Willem Robert van Hage, Wan Fokkink and Guus Schreiber, presented at the* 7th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2011) *at the* 10th

International Semantic Web Conference (ISWC 2011) *in Bonn, Germany, and of the paper* Subjective Logic Extensions for the Semantic Web *, coauthored with Archana Nottamkandath and Wan Fokkink and presented at the* 8th International Workshop on Uncertainty Reasoning for the Semantic Web *at the* 11th International Semantic Web Conference (ISWC 2012).

3.1 Introduction

The World Wide Web and the Semantic Web offer access to an enormous amount of data and this is one of their major strengths. However, the uncertainty about these data is quite high, due to the multi-authoring nature of the Web itself and to its time variability: some data are accurate, some others are incomplete or inaccurate, and generally, such a reliability level is not explicitly provided. We have already seen in Chapters 1 and 2 that Web sources are particularly useful in the process of trust assessment.

In this chapter, we start by focusing on the real distribution of Web data, in particular of categorical Web data, regardless of whether these are provided by documents, RDF [171] statements or other means. Categorical data are the among the most important types of Web data, because these also include URIs. We do not look for correlations among data, but we stick to estimate how category proportions distribute over populations of Web data.

We assume that any kind of reasoning that might produce new statements (e.g., subsumption) has already taken place. Hence, unlike for instance Fukuoe et al. [57], that apply probabilistic reasoning in parallel to OWL [7] reasoning, we will propose some models to address uncertainty issues on top of that kind of reasoning layers. These models, namely the parametric Beta-binomial and Dirichlet-multinomial, and the non-parametric Dirichlet process, will use first and second order probabilities and the generation of new classes of observations, to derive prudent conclusions about the overall populations of our data, given that we are deriving those from possibly biased samples.

Then, we propose three extensions of subjective logic (see Chapter Preliminaries) to handle partial observation, semantic similarity measures and Dirichlet processes. On the one hand these extensions allow the logic to handle also Dirichlet processes, that the first part of this chapter shows are useful to model specific classes of Web data (categorical data which categories are only partially known). On the other hand, these extensions provide useful Web-based tools that will be used in other chapters of this thesis (mainly Chapters 7 and 8), like the possibility of handling partial observations and to use semantic similarity measures within the logic.

The rest of this chapter is structured as follows: Section 3.2 defines the scope of this work; Section 3.3 describes the use of parametric and non-parametric models for managing categorical Web data; Section 3.4 describes the extension of subjective logic to incorporate semantic similarity measures, the use of partial evidence and so-called

"open world opinions". Section 3.5 discusses the analyses and the extensions described. Lastly, Section 3.6 presents conclusions.

3.2 Scope of this Chapter

3.2.1 Empirical Evidence from the Web

Uncertainty is often an issue in case of empirical data. This is especially the case with empirical Web data, because the nature of the Web increases the relevance of this problem but also offers means to address it, as we have already seen in Chapters 1 and 2 and we further demonstrate in this section. Of course, even within the Web it can be hard to find multiple sources asserting about a given fact of interest. However, the growing dimension of the Web makes it reasonable to believe in the possibility of finding more than one data set about a given subject, at least by means of implicit and indirect evidence. This chapter aims to show how it is possible to address the described issues by handling such empirical data, categorical empirical data in particular, by means of the Beta-binomial, Dirichlet-multinomial and Dirichlet process models (see Chapter Preliminaries).

3.2.2 Requirements

Our approach will need to be quite elastic in order to cover several issues, as described below. The non-triviality of the problem comes in a large part from the impossibility of directly handling the sampling process from which we derive our conclusions. The requirements that we will need to meet are:

Ability to handle incremental data acquisition. The model should be able to handle data that are acquired incrementally. As long as we collect more data (even by crawling), our knowledge will grow.

Prudence. It should derive prudent conclusions given all the available information. In case not enough information is available, the wide range of possible conclusions derivable will clearly make it harder to set up a decision strategy.

Cope with biased sampling. The model should deal with the fact that we are not managing a supervised experiment, that is, we are not randomly sampling from the entire population. We are using an available data set to derive safe conclusions, but these data could, in principle, be incomplete, inaccurate or biased, and we must take this into account.

Ability to use samples drawn from mixture distributions. Data at our disposal may have been drawn from diverse distributions, so we can not rely on the central limit theorem, because it relies on the fact that the sequence of variables is identically distributed. This implies the impossibility of making use of estimators that approximate by means of the Normal distribution.

Ability to handle temporal variability of parameters. Over time, data distributions can change, and this variability has to be properly accounted.

Complementarity with higher order layers. The aim of the approach is to quantify the intrinsic uncertainty in the data provided by the reasoning layer, and, in turn, to provide to higher order layers (time series analysis, decision strategy, trust, etc.) reliable data and/or metadata.

3.2.3 Related Work

Chapter Preliminaries describes the theoretical foundations for this chapter, that are subjective logic, conjugate priors and semantic similarity measures.

The models adopted here are applied in a variety of fields. For the parametric models, examples of applications are: topic identification and document clustering [106, 50], quantum physics [75], and combat modeling in the naval domain [97]. What these heterogeneous fields have in common is the presence of multiple levels of uncertainty (for more details about this, see Section 3.3.1). Also non-parametric models are applied in a wide variety of fields. Examples of these applications include document classification [40] and haplotype inference [179]. These heterogeneous fields have in common with the previous application the presence of several layers of uncertainty, but these also show lack of prior information about the number of parameters. These concepts will be treated in Section 3.3.2.

To our knowledge, the chosen models have not been applied to categorical Web data yet. We propose to adopt them, because, as the following sections will show, these fit the requirements previously listed.

The connection between subjective logic and the Web has been investigated in Chapter 1 and 2. We refer the reader to Chapters Introduction and Preliminaries for an extensive review on subjective logic. Kaplan et al. [90] focus on the exploration of uncertain partial observations used for building subjective opinions. Unlike their work, we restrict our focus on partial observations of Web-like data and evaluations, which comprise the number of "likes", links and other similar indicators related to a given Web item. The weighing and discounting based on semantic similarity measures can resemble the work of Jøsang et al. [85], although the additional information that we include in our reasoning (which is semantic similarity) is related only to the frame of discernment in subjective logic, and not to the belief assignment function.

3.3 Modeling Categorical Web Data

In this section we evaluate the use of higher-order probability distributions for modeling categorical Web data. Higher-order probability distributions (Beta and Dirichlet distributions used as priors of Binomial and Multinomial distributions respectively), represent the statistical foundations of subjective logic. By performing this evaluation, we aim at demonstrating the capability of subjective logic to provide a framework to reason on Web data, especially for trust assessment.

3.3.1 Parametric Bayesian Models for Categorical Web Data

In this subsection we will handle situations where the number of categories is known a priori, by using the Dirichlet-multinomial model and its special case with two categories, i.e., the Beta-binomial model [56]. As generalized versions of the Binomial and Multinomial distribution, these describe the realization of sequences of mutually exclusive events. Categorical data can be seen as examples of such events. These models are parametric, since the number and type of parameters is given a priori, and these can also be classified as "empirical Bayesian models". This further classification means that these can be seen as an approximation of a full hierarchical Bayesian model, where the prior hyperparameters (that are the parameters of the prior distribution, see Chapter Preliminaries) are set to their maximum likelihood values according to the analyzed sample.

Case Study: Deciding Between Alternatives - Ratio Estimation

Suppose that a museum has to annotate a particular item I of its collection. Suppose further, that the museum does not have expertise in house about that particular subject. For this reason, in order to correctly classify the item, it seeks judgments from outside people, in particular from Web users who provide evidence of owning the desired expertise.

After having collected judgements, the museum faces two possible classifications for the item, C1 and C2. C1 is supported by four experts, while C2 by only one expert. We can use these numbers to estimate a probability distribution that resembles the correct distribution of C1 and C2 among all possible annotations. A basic decision strategy that could make use of this probability distribution, could accept a certain classification only if its probability is greater or equal to a given threshold (e.g., 0.75). If so, the Binomial distribution representing the sample would be treated as representative of the population, and the sample proportions would be used as parameters of a Bernoulli distribution about the possible classifications for the analyzed item: $P(class(I) = C1) = 4/5 = 0.8$, $P(class(I) = C2) = 1/5 = 0.2$. A Bernoulli distribution describes the possibility that one of two alternative events happens. One of these events happens with probability p, the other one with probability $1-p$. A Binomial distribution with parameters n, p represents the outcome of a sequence of n Bernoulli trials having all the same parameter p. However, this solution shows a manifest leak. It provides to the decision strategy layer the probabilities for each of the possible outcomes, but these probabilities are based on the current available sample, with the assumption that it correctly represents the complete population of all existing annotations. This assumption is too ambitious. (Flipping a coin twice, obtaining a heads and a tails, does not guarantee that the coin is fair, yet.) In order to overcome such a limitation, we should try to quantify how much we can rely on the computed probability. In other words, if the previously computed probability can be referred to as a "first order" probability, what we need to compute now is a "second order" probability (see Chapter Preliminaries and the work of Hilgevoord and Huffink [75]). Given that

#C1	$\#C2$	$P(p \geq 0.75)$ $p \sim Beta(\#C1+1, \#C2+1)$
4	1	0.4660645
8	2	0.5447991
12	3	0.8822048

Table 3.1: The proportion within the sample is kept, so the most likely value for p is always exactly that ratio. But the probability of p being correct passes the 0.75 threshold only if the sample size is at least 15.

the conjugate prior for the Binomial distribution representing our data is the Beta distribution, the model becomes:

$$p \sim Beta(\alpha, \beta),\ X \sim Bin(p, n) \tag{3.1}$$

where $\alpha = \#evidence_{C1} + 1$ and $\beta = \#evidence_{C2} + 1$.

By analyzing the shape of the conjugate prior Beta(5,2), we can be certain enough about the probability of C1 being safely above our acceptance threshold. In principle, our sample could be drawn by a population distributed with a $40\% - 60\%$ proportion. If so, given the threshold of acceptance of 0.75, we would not be able to take a decision based on the evidence. However, the quantification of that proportion would only be possible if we know the population. Given that we do not have such information, we need to estimate it, by computing Equation (3.2), where we can see how the probability of the parameter p being above the threshold is less than 0.5. This manifests the need for more evidence: our sample suggests accepting the most popular value, but the sample itself does not guarantee to be representative enough of the population.

$$P(p \geq 0.75) = 0.4660645,\ p \sim Beta(5, 2) \tag{3.2}$$

Table 3.1 shows how the confidence in the value p being above the threshold grows as long as we increase the size of the sample, when the proportion is kept. By applying the previous strategy (0.75 threshold) also to the second order probability, we will still choose C1, but only if supported by a sample of size at least equal to 15. Finally, these considerations could also be done on the basis of the Beta-binomial distribution, which is a probability distribution representing a Binomial which parameter p is randomly drawn from a Beta distribution. The Beta-binomial summarizes the model presented in Equation (3.1) in one single function (Equation (3.3)). We can see from Table 3.2 that the expected proportion of the probability distribution approaches the ratio of the sample (0.8), as the sample size grows. If so, the sample is regarded as a better representative of the entire population and the Beta-binomial, as sample size grows, will converge on the Binomial representing the sample (see Figure 3.1).

$$X \sim BetaBin(n, \alpha, \beta) = p \sim Beta(\alpha, \beta), X \sim Bin(n, p) \tag{3.3}$$

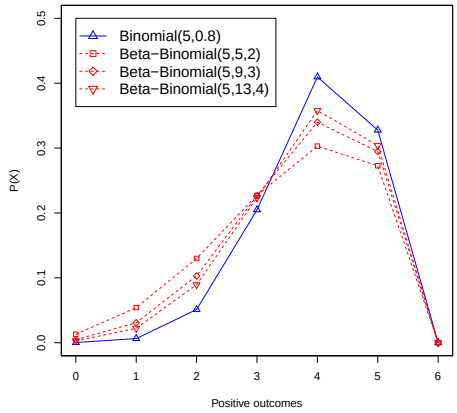


Figure 3.1: Comparison between Binomial and Beta-binomial with increasing sample size. As the sample size grows, Beta-binomial approaches Binomial.

X	$E(X)$	$p = E(X)/n$
BetaBin(5,5,2)	3.57	0.71
BetaBin(5,9,3)	3.75	0.75
BetaBin(5,13,4)	3.86	0.77

Table 3.2: Even though the sample composition is similar, the expected proportion (p) derived by Beta-binomials based on different samples passes a given threshold (0.75) only if the sample is large enough.

Case Study: Deciding Proportions - Confidence Intervals Estimation

The Linked Open Piracy [163] is a repository of piracy attacks that happened around the world in the period 2005 - 2011, derived from reports retrieved from the ICC-CCS website [80]. Attack descriptions are provided, in particular covering their type (boarding, hijacking, etc.), place, time, as well as ship type.

Data about attacks is provided in RDF format, and a SPARQL (see [169]) endpoint permits to query the repository. Such a database is very useful, for instance, for insurance companies to properly insure ships. The premium should be related to both ship conditions and their usual route. The Linked Open Piracy repository allows an insurance company to estimate the probability to be victim of a particular type of attacks, given the programmed route. Different attack types will imply different risk levels.

However, directly estimating the probability of a new attack given the dataset, would not be correct, because, although derived from data published from an official entity like the Chamber of Commerce, the reports are known to be incomplete. This

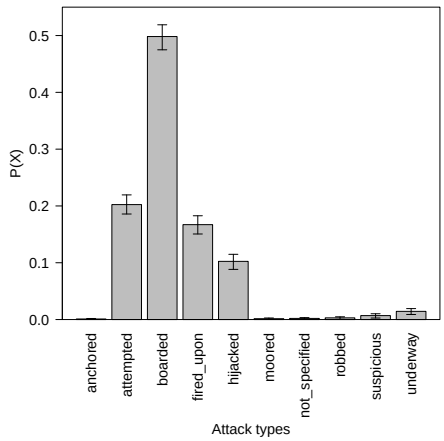


Figure 3.2: Attack type proportion and confidence intervals

fact clearly affects the computed proportions, especially because it is likely that this incompleteness is not fully random. There are particular reasons why particular attack types or attacks happening in particular zones are not reported. Therefore, beyond the uncertainty about the type of next attack happening (first order uncertainty), there will be an additional uncertainty order due to the uncertainty in the proportions themselves. This can be handled by a parametric model that will allow us to estimate the parameters of a Multinomial distribution. The model that we are going to adopt is the multivariate version of the model described in Section 3.3.1, that is, the Dirichlet-multinomial model (see also Chapter Preliminaries and the work of Elkan [50], Kvam and Day [97] and Madsen et al. [106]):

$$Attacks \sim \text{Multinom}(params),\ params \sim \text{Dirichlet}(\alpha) \tag{3.4}$$

where α is the vector of observations per attack type (incremented by one unit each, as the α and β parameters of Beta probability distribution). By adopting this model, we are able to properly handle the uncertainty carried by our sample, due to either time variability (over the years, attack type proportions could have changed) or biased samples. Drawing the parameters of our Multinomial distribution from a Dirichlet distribution instead of directly estimating them, allows us to compensate for this fact, by smoothing our attacks distribution. As a result of the application of this model, we can obtain an estimate of confidence intervals for the proportions of the attack types (with 95% of significance level, see Equation (3.5)). These confidence intervals depend both on the sample distribution and on its dimension (Figure 3.2).

$$\forall p \in param, CI_p = (p - \theta_1, p + \theta_2), P(p - \theta_1 \leq p \leq p + \theta_2) = 0.95 \tag{3.5}$$

3.3.2 Non-parametric Bayesian Models

In some situations, the previously described parametric models do not fit our needs, because these set a priori the number of categories, but this is not always possible. In the previous example, we considered and handled uncertainty due to the possible bias of our sample. The proportions showed by our sample could be barely representative of the entire population because of a non-random bias, and therefore we were prudent in estimating densities, even not discarding entirely those proportions. However, such an approach lacks in considering another type of uncertainty: we could not have seen all the possible categories and we are not allowed to know all of them a priori. Our approach was to look for the prior probability to our data in the n-dimensional simplex, where n is the number of categories, that is, possible attack types. Now such an approach is no more sufficient to address our problem. What we should do is to add yet another hierarchical level and look for the right prior Dirichlet distribution in the space of the probability distributions over probability distributions (or space of simplexes). Non-parametric models differ from parametric models in that the model structure is not specified a priori but is instead determined from data. The term non-parametric is not meant to imply that such models completely lack parameters, but that the number and nature of the parameters are flexible and not set in advance. Hence, these models are also called "distribution-free".

Case Study: Piracy Attacks Classification - Unseen Types Generation

We aim at predicting the type distributions of incoming attack events. To build an "infinite category" model, we need to allow for event types to be randomly drawn from an infinite domain. Therefore, we map already observed attack types with random numbers in $[0, 1]$ and, since all events are a priori equally likely, then new events will be drawn from the Uniform distribution, $U(0, 1)$, that is our base distribution (and is a measure over $[0, 1]$). The model is:

- $type_1 \sim DP(U(0,1), \alpha)$: the prior over the first attack type in region R;
- $attack_1 \sim Categorical(type_1)$: type of the first attack in R during $year_y$.

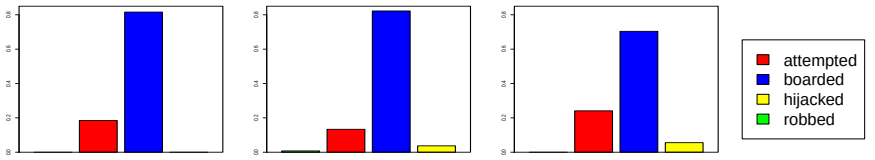
After having observed $attack_{1...n}$ during $year_y$, our posterior process becomes:

$$type_{n+1} \mid attack_{1...n} \sim DP\left(\alpha + n, \frac{\alpha}{\alpha+n}U(0,1) + \frac{n}{\alpha+n}\frac{\sum_{i=1}^{n}\delta_{attack_i}}{n}\right)$$

where α is a low value, given the low confidence in $U(0, 1)$, and $type_{n+1}$ is the prior of $attack_{n+1}$, that happens in $year_{y+1}$. A Categorical distribution is a Bernoulli distribution with more than two possible outcomes (see Section 3.3.1).

	Simulation	Projection
Average distance	0.29 △	0.35
Variance	0.09 △	0.21

Table 3.3: Averages and variances of the errors. The simulation performs best.



(a) Year 2005 (b) Simulation (c) Year 2006

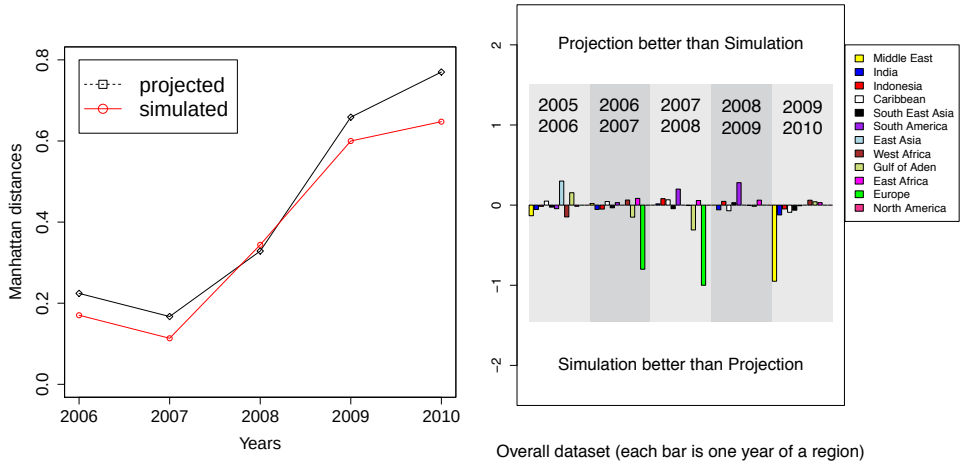
Figure 3.3: Comparison between the projection forecast and the simulation forecast with the real-life year 2006 data of region India.

Results Focusing on each region at time, we simulate all the attacks that happened there in $year_{y+1}$. Names of new types generated by simulation are matched to the actual $year_{y+1}$ names, that do not occur in $year_y$, in order of decreasing probability. The simulation is compared with a projection of the proportions of $year_n$ over the actual categories of $year_{n+1}$. The comparison is made by measuring the distance of our simulation and of the projection from the real attack types proportions of $year_{y+1}$ using the Manhattan distance (see [96]). This metric simply sums, for each attack type, the difference between the real $year_{y+1}$ probability and the one we forecast, so it can be used as an error measure. Table 3.3 summarizes the results over the entire dataset.[1] Our simulation reduces the error with respect to the projection, as confirmed by a Wilcoxon signed-rank test [177] at 95% significance level. (This statistical hypothesis test determines whether the two population means differ significantly.) The simulation improves when a large amount of data is available and the category cardinality varies, as in case of Region India (see Figure 3.3 and 3.4a).

3.4 Subjective Logic Extensions for the Semantic Web

We have seen in the previous section that higher-order probability distributions can help modeling Web data, and account for their uncertainty. In Chapters 1 and 2 we adopted subjective logic as a basic uncertainty reasoning framework. Subjective logic relies on higher-order probabilities, and the fact that these distributions model well

[1]The code can be retrieved at `http://trustingwebdata.org/phdthesis/dceolin`.



(a) Errors (b) Distances differences

Figure 3.4: Error distance from real distribution of the region India (Figure 3.4a) and differences of the error of forecast based on simulation and on projection (Figure 3.4b). Positive difference means that the projection predicts better than our simulation.

Web data strengthens our motivation for adopting subjective logic for estimating the trustworthiness of Web data. In this line of thinking, we focus here on the gaps shown by subjective logic with respect to Web and Semantic Web data handling. One of these gaps is the possibility to adopt semantic similarity measures to compare pieces of evidence. Thus we extend the logic in order to be able to weigh pieces of evidence (in form of Web data) based on semantic similarity values, expressed as subjective opinions. This allows increasing the availability of evidence, as well as the possibility to estimate trust levels tailored for each piece of evidence, as we show in Chapters 7 and 8. Then, we extend subjective logic for handling partial evidence observations, in order to allow partial evidence (e.g., "likes" or "retweets" counts) to be used to derive subjective opinions about statements. Another gap emerges also in the previous sections: there exists subjective opinions based on Beta and Dirichlet distributions, but no subjective opinion is based on Dirichlet Processes. Since these have revealed to be useful in representing Categorical Web data, we propose an extension of subjective logic that covers them.

3.4.1 Subjective Logic and Semantic Similarity Measures

Semantic similarity measures can be used to aid subjective logic to handle textual evidence and treat it qualitatively (so that not only the amount of evidence counts, but also its semantic relevance). In this section we tackle the problem of combining

subjective logic and semantic similarity measures.

Deriving Opinion about a New or Unknown Context

Recall from Chapter Preliminaries that a subjective opinion is represented as

$$\omega_y^x(b, d, u, a)$$

where x is a given source, y the target of the subjective opinion and b, d, u and a represent the belief, disbelief, uncertainty and a priori belief encoded by the opinion. To be more precise, we compute the opinions based on context $c \in C$ where C is the set of all the possible context. A context provides a different instantiation for the observations regarding y: in a given context c, x can have at her disposal some observations about y that are unavailable in context c' or vice versa. It is possible that evidence required to compute the opinion for a particular context is unavailable. For example, suppose that source x owns observations about an assertion in a certain context (e.g., the expertise of an agent about tulips), but needs to evaluate them in a new context (e.g., the agent's expertise about sunflowers), of which it owns no observations. The semantic similarity measure between two contexts (or, more precisely, between the context definitions), $sim(c, c')$ can be used for obtaining the opinion about an assertion y on an unknown or new context through two different methods. In order to derive an opinion about a new or unknown context we can use either the weighing (on the evidence) or the discounting operation (on the opinion); both the approaches are described below. We will show that the discounting and the weighing operators are theoretically but not statistically different.

Weighing the Evidence We derive evidence for a new context c by counting all the evidence available in the other contexts c' in the set of context C for the source x and the opinion target y. Before summing up the evidence counts, we weigh the evidence from each context on the basis of the semantic similarity with the new context c. So, if x wants to compute an opinion about y being an expert on *Tulips*, but owns only evidence about y being expert on *Sunflowers*, x uses the evidence about y being an expert on *Sunflowers*, weighted on the similarity between *Tulips* and *Sunflowers*, *sim(Tulips, Sunflowers)* (see Chapter 7 for a case study implementation). Formally, the evidence counts that x owns about y in context c are computed as follows, where C is the set of all the contexts of x.

$$p_c = \sum_{c' \in C} (sim(c, c') \cdot p_{c'}) \qquad n_c = \sum_{c' \in C} (sim(c, c') \cdot n_{c'})$$

.

Discounting the Opinion In the second approach, we use the subjective logic fusion operator (see Chapter Preliminaries) to merge all the opinions that source x has about y in contexts c', where $c' \in C$ about target y. Before merging the opinions, each

of them is discounted with the corresponding semantic similarity measure $sim(c, c')$ using the discounting operator in subjective logic. The discounted opinions, once aggregated form the final opinion of x about y in the new context c.

Subjective logic offers a variety of operators for "discounting", i.e., for weighing opinions given by third parties, provided that we have at disposal an opinion about the source itself. "Smoothing" is meant as reducing the belief provided by a third party, depending on the opinion on the source (the worse the opinion, the higher the reduction). Moreover, since the components of the opinion always sum to one, reducing the belief implies an increase of (one of) the other components: hence there exists a discounting operator favoring uncertainty and one favoring disbelief. Finally, there exists a discounting operator that makes use of the expected value E of the opinion. Following this line of thought, we can use semantic similarity as a discount factor for opinions imported from contexts related to the one of interest, in case of a lack of opinions in it, to handle possible variations in the validity of the statements due to the change of context.

Choosing the Appropriate Discounting Operator We need to choose the appropriate discounting operator that allows us to use the semantic similarity value as a discounting factor for opinions. The disbelief favoring discounting is an operator that is employed whenever one believes that the source considered might be malicious. This is not our case, since discounting is used to import opinions owned by ourselves but computed in different contexts than the one of interest. Hence we do not make use of the disbelief favoring operator.

In principle, we would have no specific reason to choose one between the uncertainty favoring discounting and the base rate discounting. Basically, having that only rarely the belief (and hence the expected value) is equal to 1, the two discounting operators decrease the belief of the provided opinion, one by multiplying it by the belief in the source, the other one by the expected value of the opinion about the source. In practice, we will see that, thanks to Theorem 3.4.1 these two operators are equivalent in this context. A dogmatic opinion is an opinion having *uncertainty* equal to zero.

Theoretical Foundations

Theorem 3.4.1 (Semantic Similarity Measure is a Dogmatic Opinion)**.** *Let $sim(c, c')$ be the semantic similarity between two contexts c and c' obtained by computing the semantic distance between the contexts in a graph through deterministic measurements (e.g., the Wu and Palmer semantic similarity measure [178]). Then, $\forall sim(c, c') \in [0, 1]$,*

$$\omega_{c=c'}^{measure} = (b_{c=c'}^{measure}, d_{c=c'}^{measure}, u_{c=c'}^{measure}, a_{c=c'}^{measure})$$

is equivalent to a dogmatic opinion in subjective logic.

Proof: A binomial opinion is a dogmatic opinion if the value of *uncertainty is 0*. The semantic similarity measure can be represented as an opinion about the similarity of two contexts c and c'. However, since we restrict our focus on *WordNet*-based

measures, the similarity is inferred by graph measurements, and not by probabilistic means. This means that, according to the source, this is a "dogmatic" opinion, since it does not provide any indication of uncertainty: $u^{measure}_{c=c'} = 0$. The opinion is not based on evidence observation, rather on actual deterministic measurements.

$$E(\omega^{measure}_{c=c'}) = b^{measure}_{c=c'} + u^{measure}_{c=c'} \cdot a = sim(c, c') \tag{3.6}$$

where *measure* indicates the procedure used to obtain the semantic distance, e.g., the Wu and Palmer measure. The values of belief and disbelief are obtained as:

$$b^{measure}_{c=c'} = sim(c, c') \qquad d^{measure}_{c=c'} = 1 - b^{measure}_{c=c'} \tag{3.7}$$

■

Corollary 3.4.2 (Discounting an Opinion with a Dogmatic Opinion)**.** *Let A be a source who has an opinion about y in context c', $\omega^A_{y:c'} = (b^A_{y:c'}, d^A_{y:c'}, u^A_{y:c'}, a^A_{y:c'})$, and let the semantic similarity between the contexts c and c' be represented as a dogmatic opinion $\omega^{measure}_{c=c'} = (b^{measure}_{c=c'}, d^{measure}_{c=c'}, 0, a^{c'}_{c=c'})$. Since, the source A does not have any prior opinion about the context c, we derive the opinion of A about c represented as $\omega^{A:c'}_c = (b^{A:c'}_c, d^{A:c'}_c, u^{A:c'}_c, a^{A:c'}_c)$ using the base rate discounting operator on the dogmatic opinion.*

$$\begin{gathered} a^{A:B}_y = a^B_y \qquad b^{A:B}_y = sim(c, c') \cdot b^B_y \\ u^{A:B}_y = 1 - sim(c, c') \cdot (b^B_y + d^B_y) \qquad d^{A:B}_y = sim(c, c') \cdot d^B_y \end{gathered} \tag{3.8}$$

Weighing Operator Let C be the set of contexts c' of which a source A has an opinion derived from the positive and negative evidence in the past. Let c be a new context for which A has no opinion yet. We can derive the opinion of A about facts in c, by weighing the relevant evidences in set C with the semantic similarity measure $sim(c, c')$ $\forall c' \in C$. The belief, disbelief, uncertainty and a priori obtained through the weighing operation are expressed below.

$$\begin{gathered} b^A_c = \frac{sim(c,c') \cdot p^A_{c'}}{sim(c,c')(p^A_{c'} + n^A_{c'}) + 2} \qquad d^A_c = \frac{sim(c,c') \cdot n^A_{c'}}{sim(c,c')(p^A_{c'} + n^A_{c'}) + 2} \\ u^A_c = 1 - \frac{sim(c,c') \cdot (p^A_{c'} + n^A_{c'})}{sim(c,c')(p^A_{c'} + n^A_{c'}) + 2} \qquad a^A_c = a^A_{c'} \end{gathered} \tag{3.9}$$

Theorem 3.4.3 (Approximation of Weighing and Discounting Operators)**.** *Let $\omega^{A:c'}_{y:c} = (b^{A:c'}_{y:c}, d^{A:c'}_{y:c}, u^{A:c'}_{y:c}, a^{A:c'}_{y:c})$ be a discounted opinion which source A has about y in a new or unknown context c, derived by discounting A's opinion on known context $c' \in C$ represented as $\omega^A_{c'} = (b^A_{c'}, d^A_{c'}, u^A_{c'}, a^A_{c'})$ with the corresponding dogmatic opinions (e.g., sim(c,c')). Let source A also obtain an opinion about the unknown context c based on the evidence available from the earlier context c', by weighing the evidence (positive and negative) with semantic similarity between c and c', sim(c,c') $\forall c' \in C$. Then the difference between the results from the weighing and from the discount operator in subjective logic are statistically insignificant.*

Proof: We substitute the values of belief, disbelief, uncertainty values in Equation (3.10) for Base Rate Discounting with the values from Equation (28) and the expectation value from Equation (3.6). We obtain the new value of the discounted base rate opinion as follows:

$$b_c^{A:c'} = \frac{sim(c,c') \cdot p_{c'}^A}{(p_{c'}^A + n_{c'}^A + 2)} \qquad d_c^{A:c'} = \frac{sim(c,c') \cdot n_{c'}^A}{(p_{c'}^A + n_{c'}^A + 2)}$$
$$u_c^{A:c'} = 1 - \frac{sim(c,c') \cdot (p_{c'}^A + n_{c'}^A)}{(p_{c'}^A + n_{c'}^A + 2)} \qquad a_c^{A:c'} = a_{c'}^A \tag{3.10}$$

Equations (3.10) and (3.9) are similar, except for the $sim(c, c').(p_{c'}^A + n_{c'}^A)$ factor in the weighing operator. In the following section we use a 95% Student's t-test and Wilcoxon signed-rank statistical test to prove that the difference due to that factor is not statistically significant for large values of $sim(c, c')$ (at least 0.5).

Evaluations

We prove statistically the similarity between weighing and discounting.[2]

First Validation: Discounting and Weighing in a Real-Life Case For the purpose of our evaluations, we use the "Steve Social Tagging Project" [155] data (in particular, the "Researching social tagging and folksonomy in the ArtMuseum"), which is a collaboration of museum professionals and others aimed at enhancing social tagging. In our experiments, we used a sample of tags which the users of the system provided for the 1784 images of the museum available online. Most of the tags were evaluated by the museum professionals to assess their trustworthiness. We use only the evaluated tags for our experiments. The tags can be single words or a string of words provided by the user regarding any objective aspect of the image displayed to them for tagging.

We select a set of tags highly semantically related, by using a Web-based *WordNet* interface [127]. We then gather the list of users who provided the tags regarding the chosen words and count the pieces of positive and negative evidence. The opinions are calculated using two different methods. The first method weighs the evidence with the semantic distance using Equation (3.9), and the second method is by discounting the evidence with the semantic distance using Equation (3.10). We consider the *Chinese-Asian* pair (semantic similarity 0.933) and the *Chinese-Buddhist* pair (semantic similarity 0.6667).

We employ the Student's t-test and the Wilcoxon signed-rank test to assess the statistical significance of the difference between two sample means. At 95% confidence level, both tests show a statistically significant difference between the two means. This difference, for the *Chinese-Asian* pair is 0.025, while for the *Chinese-Buddhist* pair is 0.11, thanks also to the high similarity (higher than 0.5) between the considered topics. Having removed the average difference from the results obtained from discounting

[2]Complete results are available at `http://trustingwebdata.org/phdthesis/dceolin`.

(which, on average, are higher than those from weighing), both tests indicate that the results of the two methods distribute equally.

Second Validation: Weighing on a Large Simulated Dataset In order to validate our hypothesis that weighing with semantic distance produces results that are highly similar to those obtained with the discounting operator of subjective logic, we perform the Student's t-test and the Wilcoxon signed-rank test on a larger dataset consisting of 1000 samples. For semantic distance values $sim(c, c') > 0.7$, the mean difference between the belief values obtained by weighing and discounting is 0.092. Thus with 95% confidence interval, both tests ensure that both the weighing operator and the discounting operator produce similar results. The semantic similarity threshold $sim(c, c') > 0.7$ is relevant and reasonable, because it becomes more meaningful to compute opinions for a new context based on the opinions provided earlier for the most semantically related contexts, while also in case of lack of evidence for a given context, evidence about a very diverse context can not be much significant.

3.4.2 Partial Evidence Observation

The Web and the Semantic Web are pervaded of data that can be used as evidence for a given purpose, but that constitute partially positive/negative evidence for others. Think about the *Waisda?* tagging game [119]. Here, users challenge each other about video tagging. The more users insert the same tag about the same video within the same time frame, the more the tag is believed to be correct. Matching tags can be seen as positive observations for a specific tag to be correct. However, consider the orthogonal issue of the user reputation. User reputation is based on past behavior, hence on the trustworthiness of the tags previously inserted by him/her. Now, the trustworthiness of each tag is not deterministically computed, since it is roughly estimated from the number of matching tags for each tag inserted by the user. The expected value of each tag, which is less than one, can be considered as a partial observation of the trustworthiness of the tag itself. Vice versa, the remainder can be seen as a negative partial observation (see Chapter 6 for additional information about the dataset and an extensive analysis of user reputation and tag trustworthiness). After having considered tag trustworthiness, one can use each evaluation as partial evidence concerning the user reliability: no tag (or other kind of observation) is used as a fully positive or fully negative evidence, unless its correctness has been proven by an authority or by another source of validation. However, since only rarely the belief (and therefore, the expected value) is equal to one, these observations almost never count as a fully positive or fully negative evidence. We propose an operator for building opinions based on indirect observations, i.e., on observations used to build these opinions, each of which counts as a piece of evidence.

Theorem 3.4.4 (Partial Evidence-Based Opinions)**.** *Let p be a vector of positive observations (e.g., a list of "like" counts) about distinct facts related to a given subject s. Let l be the length of p. Let each opinion based on each entry of p have an a priori*

value of $\frac{1}{2}$. Then we can derive an opinion about the reliability of the subject in one of these two manners.

- *By cumulating the expected values (counted as partial positive evidence) of each opinion based on each element of p:*

$$b = \frac{1}{l+2}\sum_{i=1}^{l}\frac{p_i+1}{p_i+2} \quad d = \frac{1}{l+2}\sum_{i=1}^{l}\frac{1}{p_i+2} \quad u = \frac{2}{l+2} \tag{3.11}$$

- *By averaging the expected values of the opinions computed on each of the elements of p:*

$$b = \frac{1}{l(l+2)}\sum_{i=1}^{l}\frac{p_i+1}{p_i+2} \quad d = \frac{1}{l(l+2)}\sum_{i=1}^{l}\frac{1}{p_i+2} \quad u = \frac{2}{l(l+2)} \tag{3.12}$$

Proof: The expected value of each opinion is computed as:

$$E = b + a \cdot u = \frac{p}{p+2} + \frac{1}{2}\cdot\frac{2}{p+2} = \frac{p+1}{p+2} \tag{3.13}$$

E is considered as partial positive evidence. Hence $1 - E$ is considered as partial negative evidence. Given that we have l pieces of partial evidence (because we have l distinct elements in $\overrightarrow{p}$), we compute the opinion about s following Equations (28). Having that p (positive evidence of ω_s) is equal to $\frac{p'+1}{p'+2}$, we obtain Equation (3.11). If we choose to average the evidence (and hence, the expected values) instead of cumulate them, what we obtain is

$$p = \frac{1}{l}\sum_{i=1}^{l}\frac{p_i+1}{p_i+2} \tag{3.14}$$

hence

$$b = \frac{1}{l+2}\cdot\frac{1}{l}\sum_{i=1}^{l}\frac{p_i+1}{p_i+2} \tag{3.15}$$

and therefore we obtain Equation (3.12). ■

3.4.3 Open World Opinions

Having to deal with real data coming from the Web, which are accessed incrementally, the possibility to update the relative probabilities of possible outcomes might not be enough to deal with them. We may need to handle unknown categories of data which should be accounted and manageable anyway. Here, we propose a particular subjective opinion called "open world opinion" which accounts for partial knowledge about the possible outcomes. A subjective opinion resembles a personal opinion provided by sources about a fact. Open world opinions represent the case when something about

a given fact has been observed, but the evidence allows also for some other (not yet observed) outcome to be considered as plausible. With this extension we allow the frame of discernment to have infinite cardinality. In practice, open world opinions allow us to represent situations when the unknown outcome of an event can be equal to one among a list of already observed values (proportional to the amount of observations for each of them), but it is also possible that the outcome is different from what has been observed so far, and is drawn from an infinitely large domain (and so some probability mass is reserved for this case).

Open World Opinion Let X be a frame of infinite cardinality, $\alpha \in \mathbb{R}^+$, k the number of categories observed, $\overrightarrow{p}$ the array of evidence per category, and $\overrightarrow{B}$ a belief function over X. We define the open world opinion ω_x:

$$\omega_x(\overrightarrow{B}, U, H) \quad B_{x_i} = \frac{p_{x_i}}{\alpha + \sum_{x=1}^{k} p_{x_i}} \quad U = \frac{\alpha}{\alpha + \sum_{x=1}^{k} p_{x_i}} \quad 1 = U + \sum_{x_i} B_{x_i} \tag{3.16}$$

Expected Value of Open World Opinion The expected value of an open world opinion is computed as follows:

$$E(p(x_i) \mid r, H) = \frac{p_{x_i} + H(x_i)}{\alpha + \sum p_{x_t}} = \frac{p_{x_i}}{\alpha + \sum p_{x_t}} \tag{3.17}$$

Theorem 3.4.5 (Equivalence between the Subjective and Dirichlet process Notation)**.** *Let $\omega_X^{bn} = (\overrightarrow{B}, U, H)$ be an opinion expressed in belief notation, and $\omega_X^{pn} = (E, \alpha, H)$ be an opinion expressed in probabilistic notation, both over the same frame X. ω_X^{bn} and ω_X^{pn} are equivalent when the following mappings holds:*

$$\begin{cases} B_{x_i} = \dfrac{p_{x_i}}{\alpha + \sum_{x=1}^{k} p_{x_i}} \\ U = \dfrac{\alpha}{\alpha + \sum_{x=1}^{k} p_{x_i}} \end{cases} \Leftrightarrow \quad \begin{cases} p_{x_i} = \frac{\alpha B_{x_i}}{U} \\ 1 = U + \sum B_{x_i} \end{cases} \tag{3.18}$$

Proof: Each step of the Dirichlet process can be seen as a Dirichlet distribution. Hence the mapping between Dirichlet distributions and multinomial opinions [86] holds also here. ■

Theorem 3.4.6 (Mapping Open World Opinion and Multinomial Opinion)**.** *Let $\omega 1_y^x(\overrightarrow{B}, U, H)$ be an open world opinion and let $\omega 2_y^x(\overrightarrow{B}, U, \overrightarrow{a})$ be a multinomial opinion. Let X_2 and Θ_2 be the frame and the frame of discernment of $\omega 2_y^x$. Let $\{B_i\}_{i=1}^k$ be the result of the partition of $dom(H)$ such that:*

1. $\mid \Theta_2 \mid = \mid \{B_i\} \mid$
2. $\bigcup \{B_i\}_{i=1}^{k} = dom(H)$
3. $\forall \{x_i\}[(\{x_i\} \in X_2 \wedge \mid \{x_i\} \mid = 1 \wedge x_i \in B_j) \Rightarrow \nexists x_{k \neq j} \in B_i]$
4. *$W = k$, where W is the non-informative constant of multinomial opinions*

Then there exists a function $D : Dom(H) \rightarrow \{B_i\}$ such that $D(\omega 1_y^x) = \omega 2_y^x$.

Proof: The equivalence between the discretized open world opinion and the multinomial opinion is proven by showing that:

- given Equation (49), since the partition $\{B_i\}_{i=1}^{k}$ covers the entire $dom(H)$, the partition distributes like the corresponding Dirichlet distribution;
- to each category of $\omega 2_y^x$ corresponds one and only one partition of $\{B_i\}$ in view of item 2 of Theorem 3.4.6. ■

In other words, open world opinions extend multinomial opinions by allowing the frame of discernment Θ to be infinite. However, by properly discretizing an open world opinion, what we obtain is an equivalent multinomial opinion.

3.4.4 Using Open World Opinions

We saw in Section 3.3.2 that every year, several ships are attacked, hijacked, etc. by pirates. Here we refer to the Linked Open Piracy dataset (LOP) [163], the Semantic Web repository of piracy attacks there described, and to the Dirichlet process, that is a useful means to model and predict the distribution of upcoming attacks. Having the possibility to represent this information by means of an open world opinion adds the power of subjective logic to the Dirichlet process-based representation. We can merge contributions from different sources, taking into account their reliability. Moreover, we can combine these facts with others in a logical way and estimate the opinion (and the corresponding probability to be true) of the consequent facts. By using open world opinions, we can easily apply standard subjective operators to these data and easily represent them in a way that takes into account basic provenance information (e.g., the data source) when applying fusing or discounting operators. For instance, if according to LOP, in Asia in 2010 we had ten hijacking events and ten attempted boarding, then we would represent this as:

$$\omega^{LOP}_{Attacks\ in\ Asia\ in\ 2010}([0.48, 0.48], 0.04, U(0, 1))$$

If our opinion about LOP is that is a reliable but not fully accountable source (e.g., $\omega^{us}_{LOP}(0.8, 0.1, 0.1)$), then we can take this information into account by weighing the opinion given by LOP as follows:

$$\omega^{us}_{LOP}(0.8, 0.1, 0.1) \otimes \omega^{LOP}_{Attacks\ in\ Asia\ in\ 2010}([0.48, 0.48], 0.04, U(0, 1)) =$$

$$= \omega^{us:LOP}_{Attacks\ in\ Asia\ in\ 2010}([0.384, 0.384], 0.232, U(0,1))$$

The resulting weighed opinion is more uncertain than the initial one, because, even though the two observed types are more likely to happen, the small uncertainty about the source reliability makes the other probabilities to rise.

A difference with respect to multinomial opinions arises in case of fusion, because the fusion operator requires that the *a priori values* have to be merged (averaged). Since the a priori values in the case of the open world opinions are represented by the distribution H (supposedly, H_1 and H_2 for two opinions to be merged). The averaging is still performed, and in this case the averaged distribution corresponds to the distribution Z having $E(Z) = b{\cdot}E(X_1){+}a{\cdot}E(X_2)$ and $VAR(X) = b^2{\cdot}(\,VAR(X_1)){+}$ $a^2 \cdot (\,VAR(X_2))$, where a, b are the two weights (e.g., u_1 and u_2 in case of cumulative fusion).

3.5 Discussion

This chapter lies the foundations for the use of subjective logic for trustworthiness estimation. In the first part, we show that higher-order probability distributions, that constitute the statistical foundation of subjective logic, are a valuable means for modeling categorical Web data. This is important because by using subjective logic for estimating the trustworthiness of semi-structured Web data, we focus on categorical data and we use exactly those probability distributions as a basis for modeling the observations at our disposal and making estimates based on them.

Having shown that the statistical background of subjective logic is useful to model categorical Web data, we devote the second part of the chapter to the development of three extensions of subjective logic aimed at covering gaps we identified in the logic, when using it to assess the trustworthiness of Web data. These extensions, that allow incorporating semantic similarity measures, partial evidence and Dirichlet processes in subjective logic, can be adopted to increase the capability of subjective logic to represent and reason upon Web data to determine their trustworthiness. In particular, in Chapters 7 and 8, we make use of the combination of subjective logic with semantic similarity to improve the quality of trustworthiness estimations of cultural heritage and media annotations by weighing the different pieces of evidence at our disposal about a given user based on their semantic similarity with a new annotation to be evaluated. In Chapter 6 we make use of partial evidence observations to compute the trust value of tag entries provided by a video tagging platform. In fact, in that use case, to estimate the trustworthiness of tag entries, we have at our disposal only the number of matches that a given tag entry received. We treat them as partial evidence observations about the tags because they represent partial and always positive pieces of evidence about the tag entries. In this chapter we show also how it is possible to use open world opinions to model uncertain Web data, and we will investigate additional applications in future research.

These extensions exemplify the flexibility of the logic. For instance, if an analyst needs to make use of a type of measure that is not present in the logic, he may follow

the same approach adopted in the definition of the combination of semantic similarity measures and subjective logic to incorporate it. It is possible that other measures are harder to incorporate, but the approach adopted in this chapter may be a valid starting point. This approach can be summarized as follows:

- normalize the measure values to let them belong to the $[0, 1]$ range;
- identify relations between the parameters of the measure (e.g., in a measure of semantic similarity, a parameter could be the number of occurrences of a term in a corpus of documents) and the components of a subjective opinion (belief, disbelief, uncertainty, base rate, but also source and statement);
- compute or approximate the formulas that bind the parameters of the measure with the components of the subjective opinion.

As said, this latter item may be particularly difficult. We will investigate this further in the future. Likewise, we will investigate in the future possible applications of the other two extensions of subjective logic that we propose, the use of partial observations and open world opinions. The fact that we included them in the logic after having shown their utility in modeling categorical Web data (in particular, the Dirichlet process, that constitutes the background of open world opinions), opens up for their use for assessing the trustworthiness of Web data and other possible applications. In order to allow a full usability of open world opinions, we need first to extend the subjective logic operators in order to handle them. That is another issue we will address in the future.

3.6 Conclusion

We have proposed three statistical models for representing Web data, namely the Beta-binomial distribution, the Dirichlet-multinomial distribution and the Dirichlet process. We have shown, by means of three applications that these are effective in modeling such data, mostly thanks to the smoothing factor these provide. The case studies we provide a comforting corroboration, since the Beta and the Dirichlet distribution, tightly connected with the Beta-binomial and the Dirichlet-multinomial distributions, represent the statistical foundation of subjective logic that has been successfully employed in Chapters 1 and 2 to reason on Web data, especially from the statistical point of view. The same logic is adopted also in the following chapters.

Moreover, we have shown that it is possible to incorporate also Dirichlet processes in subjective logic, and this will open further possibilities in the future. We also provide other two extensions of subjective logic, namely one to deal with partial observations and one to handle semantic similarity. The latter one is extensively employed in both Chapters 7 and 8. We foresee that other extensions will be possible as well like, for instance, the usage of hyperopinions [87] to handle subsumption reasoning about uncertain data, and the extension of subjective logic to cover the candidate statistical models, especially once these have shown to be effective for modeling Web data.

Also, since the use of probability distributions for modeling Web data looks promising, in the future the set of models adopted will be extended to deal with concrete domain data (e.g., time intervals, measurements), for instance, by adopting the Normal or the Poisson process (see the work of Fink [56]). Moreover, automatic model selection will be investigated, in order to choose the best model also when the limited information about our problems could make more models suitable. From a pure Web perspective, our models will be extended to properly handle contributions coming from different sources together with their reputation. This means, on one hand, considering also provenance (like in Chapter 5) and, on the other hand, using Mixture Models [132], Nested [138] and Hierarchical Dirichlet processes [150], eventually employing Markov Chain Monte Carlo algorithms [53, 117] to handle lack of conjugacy.

4

Reliability Analyses of Web Data

A Police Open Data Case Study

This chapter continues the exploration of the use of uncertainty reasoning techniques for assessing trust values started in Chapter 3 (hence revisiting the second research question, How can uncertainty reasoning be effectively used to estimate the trustworthiness of semi-structured data?*). We saw in Chapter Introduction how trust is a multi-faceted relationship; here the reliability facet is prevalent. We use statistical techniques to both measure the reliability of a set of police open datasets based on the comparison with the corresponding closed datasets and to estimate reliability changes in the open datasets alone. Firstly, we show that uncertainty reasoning techniques are useful to assess the reliability of open data when a gold standard is available, by measuring the possible discrepancies in the open data by means of different methods. This extends the work we presented in Chapter 2. Secondly, we use subjective logic to merge the results of tests run to identify reliability changes in the data when a gold standard is unavailable, thus extending the application of subjective logic presented in Chapters 1 and 3.*

This chapter extends the paper Reliability Analyses of Open Government Data*, coauthored with Luc Moreau, Kieron O'Hara, Guus Schreiber, Alistair Sackley, Wan Fokkink, Willem Robert van Hage and Nigel Shadbolt, presented at the* 9th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2013) *at the* 12th International Semantic Web Conference (ISWC 2013) *in Sydney, Australia, and the paper* Two Procedures for Analyzing the Reliability of Open Government Data*, coauthored with Luc Moreau, Kieron O'Hara, Alistair Sackley, Wan Fokkink, Nigel Shadbolt, Valentina Maccatrozzo, Willem Robert van Hage and Guus Schreiber, to be presented at the special session* Uncertainty and Imprecision on the Web of Data *at the* 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2014) *in Montpellier, France, in July 2014.*

4.1 Introduction

The goal of the work presented in this chapter is to cope with the lack of tools and methodologies to measure and compare a specific class of Web data, namely open government data.

Open Government Data are unquestionably a valuable source of information for empowering the citizens, boosting the economy and enhancing the transparency of public administration, but when published on the Web, these need to be properly processed in order to reduce the amount of personal information exposed. This process consists of aggregation and so-called "smoothing" procedures which introduce some imprecision in the data, to avoid the reconstruction of the identity of a citizen from a piece of data. The value of this data might be affected by such procedures, as they limit the extent to which we can rely on them. Throughout the chapter, we will refer to the published Open Government Data as "open data" and to the original data as "closed data".

The processes applied on these data increase the uncertainty that one faces when dealing with them: even though they are exposed by authoritative sources, the fact that some information are obfuscated voluntarily introduces some error in the data. Consequently, it is necessary to quantify the extent to which one can rely on them or, in other words, their reliability needs to be quantified.

For instance, Crime Reports [39] and `data.police.uk` [154] both publish information about crimes occurring in the UK, but these data differ in terms of format (maps versus CSV files), level of aggregation and smoothing, timeliness (daily versus monthly update). So, if the smoothing process unavoidably introduces some error in the data, there might be other reasons as well for possible reliability differences among these datasets, like the fact that a given dataset is not based on timely data (or, in general, it is generated from questionable data sources) or that an erroneous aggregation process inadvertently introduced some mistakes. For the police, as well as for citizens, it is important to understand how different two sources are, in order to understand how much they can rely on the data they expose. The police, who can access the original, raw data, is interested in measuring the reliability of the open data in order to know how much they can rely on them, e.g., when establishing projects involving citizens, and making use of such data. For citizens, on the other hand, it is important to understand the reliability of the different datasets because, although they are exposed by authoritative sources, the possible discrepancies between them give rise to questions about their reliability.

We address this problem by means of a twofold contribution: first we propose a procedure for computing the reliability of an open dataset, when having at disposal both the open and the closed data. We apply this procedure on a set of UK police data. We show that the reliability of these data is not highly affected by the smoothing and aggregation procedures applied to them and that this procedure, once properly instantiated, allows us to guide the analyst to the discovery of points of open data creation policy changes and reliability variations. Second, we show how it is possible

to estimate variations in the reliability of the open data by comparing them to each other, when the closed data are not at our disposal. The application of this procedure makes use of subjective logic (that is described in Chapter Preliminaries), and the approach adopted can be seen as an extension of the analysis of heuristics presented in Chapter 2. Both procedures are aimed at allowing us to measure and compare these datasets from the reliability point of view, and to guide an analyst to the discovery of possible critical points (e.g., policy changes, relevant errors) in these datasets.

Here, we still devote to the estimation of the trustworthiness of the data we deal with, but despite the works presented in Chapter 1 and Chapter 2, data are known to be trustworthy, at least up to a certain extent, because these are smoothed and aggregated, but these are exposed by authoritative sources. Also, despite the use of data as a heuristic for the evaluation of data trustworthiness made before, here there is no uncertainty related to the effectiveness of using the data to accomplish the task they are intended for (e.g., reporting crime counts). However, trust is still a crucial point in this case because of the importance of the reliability that comes into play because of the reasons previously mentioned (we saw in Chapter Introduction how reliability is a component of trust). Reliability is contemplated throughout the thesis, but this is one of the chapters where it plays a key role.

The rest of this chapter is structured as follows: Section 4.2 describes related work; Section 4.3 describes a procedure for determining the reliability of open data given closed data and Section 4.4 presents a procedure for analyzing open data. In Section 4.5 we put forward a case study implementation of both procedures. Section 4.6 provides a final discussion.

4.2 Related Work

The analysis of open data is increasingly being spread, for instance, by the leading Open Data Institute [151]. Koch-Weser [94] presents an interesting work on the analysis of the reliability of China's Economic Data which, although focused on a different domain, shares with this work the goal of understanding the reliability of open data. In general, tools for the quality estimation of open data are being developed (see for instance Talend Open Studio for Data Quality [148] and Data Cleaner [76]). These tools are designed to understand the adherence of data to particular standards, similar to our goals, but they aim at constituting a proper middleware component of the entire business process of data management and curation. These tools are not limited to monitoring data quality, but they aim also at quantifying the risk and the financial impact of these data, as well as how to intervene in the business process in case of any problem discovered. Our goal is less business-oriented and more targeted, as we aim at developing procedures for measuring and estimating open data reliability. However, this can be seen as a step towards the development of a more comprehensive tool (which, in principle, might have business implications as well).

From the point of view of the methods adopted, relevant for this work are also two chapters of this thesis, Chapter 3 and Chapter 5. The first of these chapters shares

with the work here presented the statistical approach in modeling categorical Web data and the use of the Wilcoxon signed-rank test to measure the reliability of these data (either real data or predicted ones). With the second mentioned chapter, the work here presented has in common the use of provenance information to make reliability estimates. The provenance graphs at our disposal are rather limited, but they still play a relevant role in this work as the different processes employed to produce the open data determine their reliability.

Closer to the topic of the case studies analyzed, i.e., the reliability of police open data, this work can be seen as complementary to the one of Cornelli [37], who researches on the reasons citizens have to trust police.

4.3 Comparing Closed and Open Data

Closed data need to be manipulated in order not to expose sensitive information when publishing them. There are two main categories of processes that serve this need. The first one is aggregation, that is to present the data at a coarser, higher level than available. This way the correctness of the data is preserved, while their granularity is reduced. For example, instead of presenting the counts of different crime categories in a specific area, one might aggregate such categories in some broader ones. However, even if this sort of procedures is not intended to introduce imprecisions, a faulty aggregation process or the wrong use of heterogeneous data sources might unexpectedly affect the data reliability. The second kind of procedure applicable is the so-called "smoothing" operation, a data aggregation procedure which, on purpose, introduces some error in order to anonymize the data. This kind of manipulation is necessary, for instance, because aggregation does not sufficiently anonymize data about low-populated areas. By smoothing, authorities voluntarily introduce some small errors in the data so that they remain reliable at coarse level, but it is not possible (or at least, hard) to reconstruct the details of the single items. Figure 4.1 exemplifies the process of open data creation. We use the W3C Recommendation PROV Ontology [9] to represent the open data creation process. PROV will facilitate the future quantification of the impact of each single process in the reliability of the resulting data. Here we describe a procedure that allows us to evaluate the reliability gap existing between open and closed data, if any. The procedure manipulates the closed data in order to make them comparable to the open data, and compares the two. It consists of four steps:

Select the relevant data Closed data might be spurious. Therefore the first step is the selection of the data items that are relevant for our analyses. This selection might involve temporal aspects (i.e., only data referring to the relevant period are considered), or their geographical location (select only the data regarding the area of interest). Other constraints and their combination are possible as well.

Roll up categorical data The categories used to classify the categorical data are

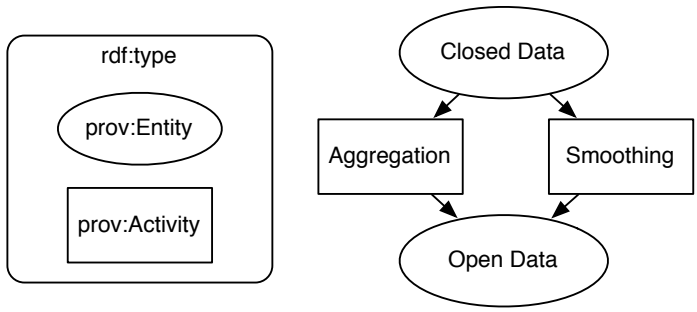


Figure 4.1: Open data creation.

ordered in hierarchies. Hierarchies are created to define different categories for different refinement levels when presenting categorical data. For instance, when speaking about crime data, items can be classified with respect to the type of crime they describe and the category of the crime can be more or less refined: we might have a refined categorization (e.g., "Anti-Social Behaviour - Noisy neighbourhood") or a coarser one (e.g., "Anti-Social Behaviour"). In order to make comparisons possible, we need to bring the data to the same level of refinement. Given that we can not increase the refinement of the data categorized in a coarser manner (since each coarser category contains many subcategories and we do not have any evidence about which is the correct subcategory), we decrease the granularity level. The category used to represent the data is then the *least common subsumer* of the categories of an open item and of the corresponding closed item: $lcs(category(item_{open}), category(item_{closed}))$.

Roll up smoothed categorical data This step is similar to the previous one, besides the fact that the expected result is not necessarily coincident with the original one because the smoothing procedure might have caused a loss of precision in the data.

Compare the corresponding counts Here a few different measures are possible. For instance, the ratio of the correct items over the total amount or the Wilcoxon signed-rank test [177].

This procedure is quite generic by purpose, because it aims at predisposing all the necessary calculations to make the open and closed data comparable, and leaves to the analyst the freedom to make the most appropriate analyses. The procedure actually allows us to measure the reliability distance between open and closed data, but this distance can be defined in different manners according to the precision needs or the point of view of the analyst. Therefore, we propose some comparison methods, but the procedure can be easily instantiated with others as well.

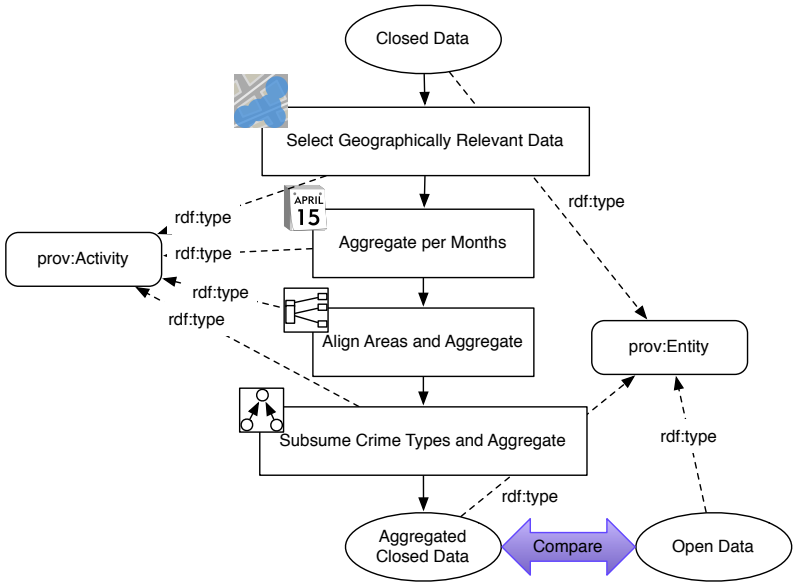


Figure 4.2: Procedure for comparing police open and closed data.

4.4 Analyzing Open Data

The previously described procedure is useful to compute the reliability of the open data. That is an important step to take for people in the sector willing to understand the loss of precision of the published data. However, a necessary condition for being able to perform that kind of analysis is to have closed data at our disposal. This latter condition is not always easily met, especially by layman people (e.g., normal citizen). In our view it does still make sense to perform analyses of open data alone, although these lead to weaker but still useful results. If we compare each dataset with each consecutive one (i.e., with the dataset containing data about the subsequent time interval), measure their differences, and analyze the variations in these differences over time, we can pinpoint occurrences of events possibly affecting the reliability of the datasets (e.g., the change in the policy of creation of the open data).

Open data counts might differ with respect to different points of view. For instance, there might be relevant absolute differences, but it can also be the case that the relative differences are more important than the absolute ones. We do not know a priori which is the best manner to compare the data counts, so we propose a new similarity measure for comparing datasets, resulting from the aggregation of different similarity "tests" performed on couples of datasets. We developed an application (described in Section 4.5.2) to make this kind of analyses.

When analyzing open datasets, we can compare only data about related but differ-

ent facts. In the case of a comparison between open and closed datasets, we compare data about the same facts, happening at the same time in the same place. One of these is considered as correct (closed data), and we estimate how much the other one (open data) differs from it. Now we can, at most, compare each dataset of the same typology of facts (e.g., crimes), but, for instance, corresponding to different times. As a consequence, the results that we can expect from this kind of analyses is much less detailed and definite than before: since we do not have at our disposal a gold standard, we can not test properly our hypothesis. However, we try to estimate points of reliability change using the following procedure, based on the idea that from the analysis of the similarity of the datasets using different similarity measures, these changes can emerge.

Choose one or more dataset similarity score. In general, given two datasets d_1 and d_2, we compute their similarity as:

$$sim(d_1, d_2) = avg(t(i_1, j_1), \ldots, t(i_n, j_n)) \tag{4.1}$$

where avg computes the average (possibly weighted) of the results of the similarity scores resulting from the test t on the n items composing d_1 and d_2, $i_1, \ldots, i_n$ are the items in d_1 and $j_1, \ldots, j_n$ the items in d_2. We propose the following tests, although we are not limited to these:

Wilcoxon signed-rank test. As in the case of the comparison between open and closed data it is possible to use this test to check if the data counts are likely to be drawn from two significantly different distributions.

Linear regression test. Apply a linear regression analysis [58] on one test and measure the average error when the resulting linear function is applied on the other dataset. So we can understand if the two datasets share a linear relation.

Support Vector Machines test. Similar to the linear regression test, we can also build a support vector machines model [38] over one of the two dataset, apply it on the other one and then evaluate the performance. This test takes into account even more than the linear regression test the existing relations between the counts of the different categories in the two datasets.

Other. Following the same line of thought as the previous examples, it is possible to apply plenty of other statistical test and/or data mining techniques to model and compare the data.

The results of the tests can be obtained either in terms of Boolean values or by means of a value in the $[0, 1]$ range. In both cases, 0 means no similarity and 1 means equality. These tests are then aggregated (or merged) in order to express the overall similarity between two datasets. These tests can be aggregated in different ways, for instance:

Average. The values resulting from the different tests can be averaged, e.g., using a simple or a weighted average, if one or more of the tests are to be considered more or less relevant.

Subjective opinion. We can consider the different tests as "subjective opinions" (see Chapter Preliminaries) about the similarity of the two datasets, and we can merge them using the "fusion" operator of subjective logic. The resulting opinion is equivalent to a Beta probability distribution representing the probability for each value in the $[0, 1]$ interval to be the correct value for the similarity. The expected value of the Beta is close to the arithmetical average, but the variance represents the uncertainty in our calculation: the more tests we consider, the smaller the variance of the resulting Beta distribution.

Compute the similarity, with one or more scores. Measure the pairwise similarity, for all the datasets about consecutive time intervals. For brevity, we refer to these datasets as "consecutive datasets".

Identify change points in the similarity sequence. Change points in the similarity sequence are likely to indicate policy changes in the data creation and hence reliability changes resulting from these policy modifications.

Aggregate the evidence about the sequence of datasets. The change points identified in the previous step are pieces of evidence of policy changes. By running more similarity analyses, we can reduce the risk to have false positives (datasets wrongly identified as policy changes initiators) or false negatives (missed policy changes). Since we are dealing with uncertain observations, we suggest adopting subjective opinions also in this step and to compute a binomial opinion for each dataset, based on the evidence available. Since this evidence might be limited, by using subjective opinions we avoid overweighing it.

Despite the previous approach, the similarity between datasets alone is not sufficient to say anything about the data themselves. There can be natural reasons that explain a variation in the data (e.g., a new law, or a rare event) without necessarily implying a lack of reliability in one of the two datasets. Moreover, a similarity value taken alone might be difficult to interpret: what does it mean that the similarity between two datasets is, for instance, 0.8? We propose to overcome this problem by analyzing the similarity of a series of consecutive datasets (i.e., datasets about consecutive time intervals) by using different similarity measures. By doing so, we can pinpoint the similarity values that differ most from the rest: still we do not have a warranty that these values indicate a change in the data reliability (there can still be other reasons for such a change), but we flag the datasets having a higher chance to present a reliability variation.

4.5 Case Study - Police Open Data Analyses

We evaluate the procedures that we propose over police data counts for the Hampshire Constabulary. As open data we adopt the corresponding entries from the `data.police.uk` website, in particular in the interval from April 2011 until December 2012. We focus on the datasets presenting the counts aggregated per police neighbourhood because this kind of classification, although not as detailed as the classification per address, allows an easy comparison between entries and reduces the burden of having to geolocate and disambiguate addresses. As closed data, we have at our disposal a series of datasets from the Hampshire Police Constabulary, covering the interval from October 2010 until September 2012. The two datasets do not perfectly overlap but, as described as follows, we focus mainly on the intersection between the two intervals covered, which still is the largest part of both datasets.

The reason why we evaluate the two procedures together is that they are tightly connected each other. We demonstrate that they allow us to provide similar findings, even though the first procedure is clearly less uncertain than the second one.

4.5.1 Analyzing the Reliability of Police Open Data

We start from an analysis of the reliability of open data. Therefore, we focus on the intersection between the open and the closed data at our disposal (that is, the period from April 2011 until September 2012). Here we interpret the reliability in terms of coverage. The procedure we proposed in Section 4.3 is quite flexible. Here we define the reliability of the open dataset as the statistical similarity between two datasets:

The reliability of an open dataset is measured as the percentage of non-significantly different entries from the corresponding closed dataset.

So, we compare the distribution of the crime counts among the different category for each police neighbourhood. The comparison is made by means of a Wilcoxon signed-rank test at 95% confidence level. If a neighbourhood is significantly different from another neighbourhood, then we count it as a negative evidence, otherwise as a positive one. We run the test over all the neighbourhoods and we aggregate them in a subjective opinion. The reason why we do not simply average the counts of positives over the total number of neighbourhoods is that we consider the outcomes of the tests run over the neighbourhoods as pieces of evidence about the reliability of the open data, and we treat them as error-prone observations. We provide a detailed description of the procedure below.

Case Study Setup

To make the analyses possible, we preprocess the closed data in order to select only the relevant closed data items and to bring the data at the same level of aggregation as the open data. The data at our disposal contain the following information:

- crime category;

- crime date;
- geographic Cartesian coordinates (or Grid Reference coordinates, or Easting and Northing) of the crime.

We implement the procedure described in Section 4.3 to align the open and closed data at our disposal as described as follows.

1. **Data preprocessing.** This part is performed in two steps.
   (a) **Convert the coordinates to WGS84.** Coordinates are converted from the Cartesian system, that is, from Easting and Northing coordinates, to World Geodetic System 84 coordinates, that is, Latitude and Longitude, using the R Geospatial Data Abstraction Library library [11].
   (b) **Estimate the postal code of the crime location.** This is performed by looking for the postal code that is closer to the point analyzed.

   It is necessary to preprocess the data in order to retrieve from the Northing and Easting coordinates the geographic coordinates (latitude and longitude) that are used in the analyses. However, this step potentially introduces some error in the analyses. One possible cause of this error is the approximation in the coordinates conversion. The other cause is that looking up the closest postal code to the point that we are analyzing is the best approximation we can make, but this is not necessarily always correct. We manually checked some sample items to confirm the robustness of this procedure, although it is not perfect. Also, in our results we show how the impact of these imperfections is limited. Lastly, we must stress how it was not possible to compute the postal code of all the points that we had at our disposal, for several reasons like, for instance, the fact that some data items were partially incomplete or presented wrong coordinates. We now report about this procedure in more detail.

2. **Select the relevant data.** This step is performed in three steps:
   (a) **Query the MapIt API [107].** This aims to retrieve the police constabulary each postal code belongs to and allows us to discard all the crime items belonging to constabularies other than the Hampshire Constabulary in the closed datasets.
   (b) **Select the relevant open data.** Here we select the open data for the months for which closed data are available.
   (c) **Exclude crime counts of categories not shared.** Categories that are not shared between open and closed data are excluded. These include, for instance, counts belonging to the "Anti-social behaviour" category that are present in the open data but do not have a counterpart in the closed data.

3. **Aggregate the data.** Also data aggregation is performed in three steps: temporal, geographical and categorical.

(a) **Temporal aggregation.** This is made in order to group together data about crimes occurring in the same month.

(b) **Geographical aggregation.** This is made to aggregate the data at police neighbourhood level. Open data are aggregated at police neighbourhood level. A police neighbourhood contains several postal codes. To aggregate the data at neighbourhood level, we match zip code and neighbourhood by querying the MapIt API [107].

(c) **Categorical aggregation.** This is performed by aligning the classifications of the crimes in the open and closed datasets. Since the open data are presented more coarsely, the closed data are brought at the same level of detail. For instance, raw data crimes in the "Theft of Motor Vehicle" category are reclassified as "Vehicle crime".

4. **Compare the aggregated data.** Once the items are brought to the same level of aggregation, open and closed data are compared to check the reliability of the open data. The comparison is made as follows.

(a) **Select each neighbourhood.** Each neighbourhood is selected in the closed and in the open data set.

(b) **Compare the crime counts.** In particular, we adopted the following comparison methods:

Apply a Wilcoxon signed-rank test. This is applied to this vector to check the estimated location of the error distribution and its significance. So, we check the significance of possible differences between the two datasets, at 95% confidence level. If the test outcome is to accept the null hypothesis (that is, that the two crime counts are not significantly different), we count one positive observation, otherwise a negative one. This test bases the comparison on the ranks of the crime categories.

Apply a χ^2 test. Despite the previous test, this one allows us to compare the distribution of the crime counts in the open and closed datasets to check if the frequencies of the two distributions are significantly similar or not.

Measure the differences between the crime counts. This difference can be absolute or relative.

(c) **Aggregate** the estimated errors over all the neighbourhoods. The aggregation is made by means of two alternative methods:

Binomial subjective opinions.

$$\omega\left(\frac{p}{p+n+2}, \frac{n}{p+n+2}, \frac{2}{p+n+2}, \frac{1}{2}\right)$$

where

$p = \#$ not significantly different entries

$$n = \#\text{ significantly different entries}$$

The actual value that we use in our measurements is the expected value of the subjective opinion (and of the corresponding beta distribution), that is:

$$E = \frac{p}{p+n+2} + \frac{2}{p+n+2} \cdot \frac{1}{2} = \frac{p+1}{p+n+2}$$

Arithmetic mean.

However, given that the total number of observations is higher than 200, the choice of this method of aggregation does not yield us results much different from what we would have obtained using the arithmetic mean.

Results

We analyze the reliability of all the datasets in the interval between April 2011 and September 2012. We know that the open datasets might differ each other and might differ with respect to the closed data, for instance, in terms of neighbourhoods represented. We start by comparing the distribution of crime counts per category on the intersection of neighbourhoods in the closed and open datasets. Indeed, because of smoothing, data might have been moved from a neighbourhood to another one, and here we want to check if the distribution of the crime counts in the matching neighbourhoods is affected by data manipulation procedures. We present a series of graphs (and corresponding descriptions) that allow us to visually understand the results of the analyses performed. Each analysis is performed by means of an R script.

The closed data at our disposal are quite complete, although they do not match perfectly the open data, as we can see from Figure 4.3. Also, the closed data at our disposal regard only the crime figures, so the counts about "Anti-social behaviour" that are reported in the open data are excluded from these analyses.

We apply the Wilcoxon-signed rank test and the χ^2 test on the crime counts of each matched neighbourhood. The first test checks if the order of the crime categories ranked in terms of crime occurrences is preserved. The second checks if the distributions are significantly different or not. We use these tests to compute the percentage of neighbourhoods that are significantly different (at 95% confidence level) between the two datasets. Figure 4.4 and Figure 4.5 show the results for the two tests. In both cases, the open datasets score quite high, as to confirm the high similarity between the crime count distributions in the open and closed datasets, in the overlapping neighbourhoods, although the χ^2 is more sensitive to the variations. We can see from the plots how the aggregation computed by means of subjective opinions is similar to the one computed by means of arithmetic average.

We know that smoothing introduces some error. In principle, this error might move the geolocation of one crime item, and this might cause a shift of the crime occurrence from the police neighbourhood it belongs to, to another one. Now, we want to understand the impact of these manipulations on the reliability of the resulting open data. So, we extend the open and the closed datasets in order to have them covering the

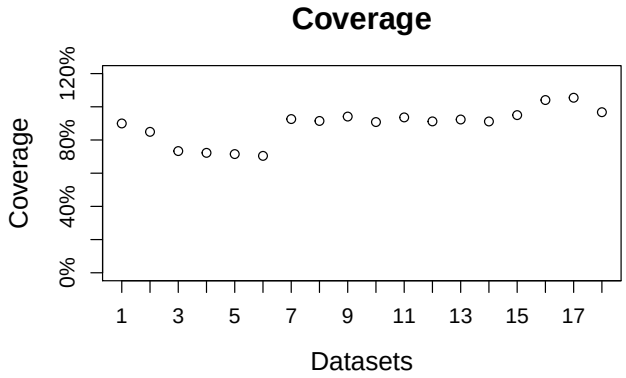


Figure 4.3: Ratio of crime items in the open dataset present in the closed data at our disposal.

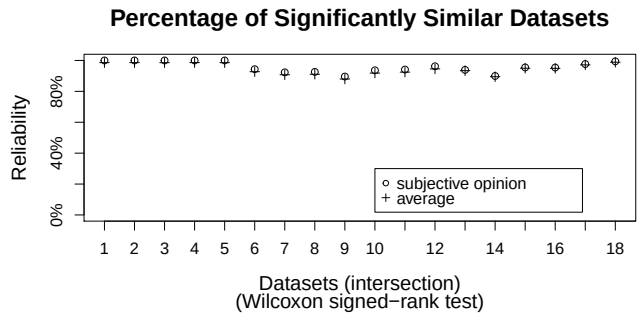


Figure 4.4: Percentage of neighbourhoods in the open data set that are significantly different from the corresponding closed data entries, according to a Wilcoxon signed-rank test at 95% confidence level and considering only the neighbourhoods in the intersection between open and closed datasets.

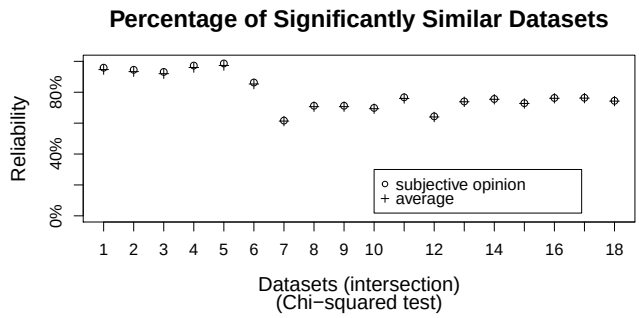


Figure 4.5: Percentage of neighbourhoods in the open data set that are significantly different from the corresponding closed data entries, according to a χ^2 test at 95% confidence level and considering only the neighbourhoods in the intersection between open and closed datasets.

same neighbourhoods: when a neighbourhood is not present in one of the two datasets, we assume that this dataset presents zero crimes in that neighbourhoods. This aims at computing the reliability of the datasets from the point of view that the open datasets are interpreted by the laymen people as the report of all the crimes happening in the area surveilled by a given constabulary. So, given the actual distribution of crimes in that area, how is the open dataset representative of it? Figure 4.6 and Figure 4.7 address this issue. Here the results are quite different from before. We can see that there are at least two different trends which, we suspect, correspond to policy changes. One of the possible policy changes regards the smoothing technique adopted, which determines the neighbourhood a crime belongs to. So, we compute the percentage of neighbourhoods shared between the open and the closed datasets. Figure 4.8 shows that there are two distinct trends of matching neighbourhoods: initially only about 30% of the neighbourhoods were present in both the open and closed datasets, and then this percentage suddenly rose to 100%. This is clearly due to a policy change in the smoothing algorithm that makes the more recent open data more reliable and similar to the closed data. This also explains the "step" shown in Figure 4.6. Indeed, starting from the thirteenth dataset, the reliability of the extended datasets corresponds to the percentage of matching neighbourhoods: when the open data present the same neighbourhoods as the corresponding closed data, the reliability of the counts in those neighbourhoods is high. When a neighbourhood is present in the open data and not in the closed data or vice versa, its reliability is zero, since a series of zeros is much different from a series of non-negative crime counts. As a result, the overall reliability of the corresponding open dataset is high only if the open and closed datasets share the same set of neighbourhoods.

However, in Figure 4.6 and in Figure 4.7 it seems that there is also another discontinuity point, between the fifth and the sixth dataset. This leads us to analyze our

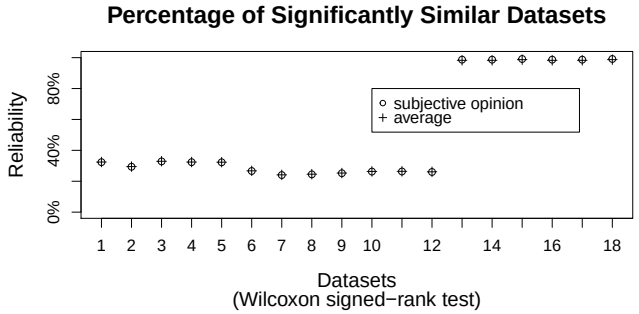


Figure 4.6: Percentage of neighbourhoods in the open data set that are significantly different from the corresponding closed data entries, according to a Wilcoxon signed-rank test at 95% confidence level and considering all the police neighbourhoods in the constabulary.

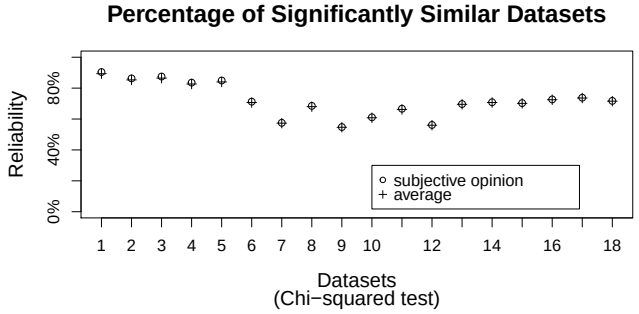


Figure 4.7: Percentage of neighbourhoods in the open data set that are significantly different from the corresponding closed data entries, according to a χ^2 test at 95% confidence level and considering all the police neighbourhoods in the constabulary.

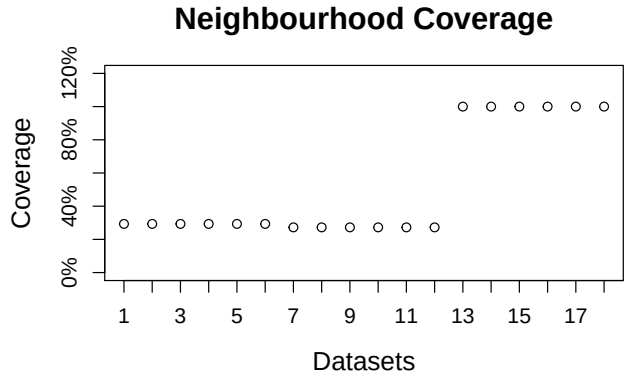


Figure 4.8: Percentage of matching neighbourhoods between open and closed datasets.

datasets more deeply, so we continue with our analyses and we take another perspective. One of the characteristics of the analyses based on the two statistical tests is that they consider the crime counts in a given neighbourhood as a whole, as a probability distribution, and they focus on the shape of this distribution. This is useful and important to discard rare and sparse errors (that might be due to sporadic and limited episodes) and to look for changes in the data that affect larger category sets, that are more likely to be caused by policy changes or similar events. However, another perspective that might result as relevant is the analysis of the absolute errors in the data. In Figure 4.9 we report the plot of the average error per neighbourhood and per crime category. Here we can notice two things: first, there are also here two trends, and the first trend breaks approximatively were we expected it to break (at the sixth month instead of at the fifth), and second, there is also a correspondence between these two trends and the trends shown in Figure 4.3. These three graphs considered together led us to focus our attention on the fifth and sixth datasets, that are those containing the crime counts for August and September 2011, and three important facts emerged.

1. The closed datasets at our disposal do not contain crime counts for the "Drugs" category for July, August and September 2011. In principle, this might be because actually no drug crime occurred in that period. However, this looks unlikely, given that the corresponding open datasets present positive figures for that category and that the category itself presents relevant figures in the other months (and so it is unlikely that no crime of that sort occurred in that interval). So, we hypothesize that simply the closed data at our disposal lacks counts for that category in that time interval. This would explain the following two observations.

2. In the September 2011 open dataset, there is one entry, relative to the police

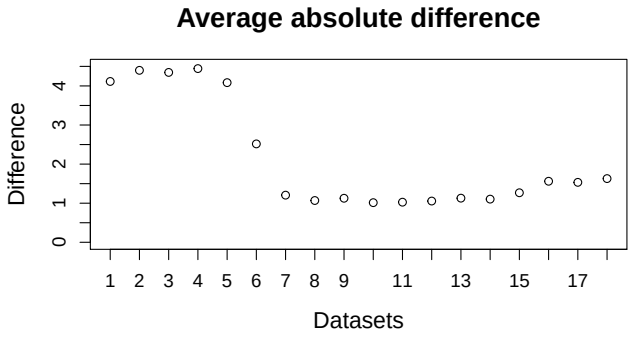


Figure 4.9: Average of absolute differences between open and closed datasets. The differences are computed per crime count and averaged per neighbourhood and dataset.

neighbourhood "2LW02", that presents two very high figures for "Drugs" and "Other.theft", 205 and 319 respectively. The average counts for these crime categories in that dataset excluding this particular neighbourhood are 8.44 and 1.68, so those figures are clearly outliers. We suspect that they are caused by an error in the open data production or to a rare event, but we cannot validate this hypothesis with the information at our disposal. Exactly one year after, September 2012, these two categories present a similar pattern for that neighbourhood, that is, 174 and 316, and in July 2012 their counts are 88 and 131. Because of the similarity of these patterns and since for the other months these categories present much lower figures in this neighbourhood (22 at most, in one case, and less than 20 in the rest of the cases), we suspect that those high counts are not random.

3. Between August and September 2011 a policy change occurred. Until August 2011, the categories adopted in the open data were: {Burglary, Robbery, Vehicle.crime, Violent.crime, Anti.social.behaviour, Other.crime}. This set was extended from September 2011 as: {Burglary, Robbery, Vehicle.crime, Violent.crime, Anti.social.behaviour, Criminal.damage.and.arson, Shoplifting, Other.theft, Drugs, Public.disorder.and.weapons, Other.crime}. This change made the "Other.crime" category more narrow, since the newly introduced categories contain crimes that before were generically classified as "Other.crime". Of course, if we align the closed data to the second crime classification and we compare them with crime counts generated using the first policy, then the error in the open data looks more significant than it actually is.

We reclassify the crimes in the datasets belonging to the first trend so that all the crimes previously classified as belonging to a category not present in the open data are now classified as "Other crime", and we recompute the average absolute differences

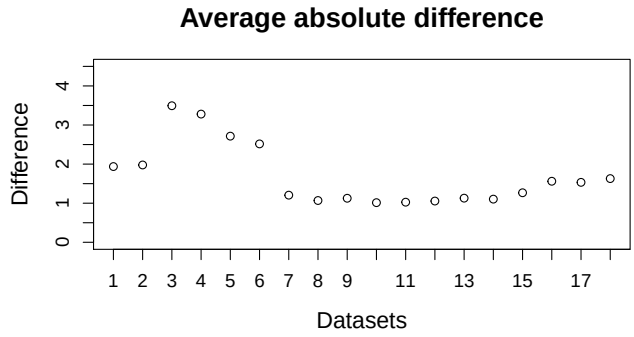


Figure 4.10: Average of absolute differences between open and closed datasets, after a reclassification of the crime counts in the first six datasets. The differences are computed per crime count and averaged per neighbourhood and dataset.

with the new classification. Figure 4.10 shows the results. As we hypothesize, by correctly classifying the crime counts in the first five datasets, the average absolute error decreases (on average, of 1.56 counts per month). Still, the error is different from the rest of the dataset because, since the correct crime classification contains fewer categories, in this case the error is spread across fewer categories, and so the same error, in this classification, weighs more.

Thanks to the analyses made, we discover two changes in the policy for open data creation: one regards the classification of the crimes, the other regards the crime geolocation. Each of these changes affects the data and their reliability. These changes could possibly have been detected manually, but at a high cost in terms of time and effort spent, given the size of the dataset and the sparsity of the crucial points. In fact, we are aware of the fact that some policy change happened, although we have no knowledge about their entity. However, by manually looking at the data, we noted two changes in the datasets, one regarding the number of crime categories shown, and one about the number of police neighbourhoods reported. We assume that these changes constitute at least a subset of the set of changes actually happened. The procedure that we propose allows us to: (1) measure the reliability of these datasets; (2) semi-automatically determine when a policy change occurs and guide the analysis of the datasets; (3) demonstrate that the variations identified (variations in the number of crime categories and neighbourhood reported) correspond in changes in the crime data reported (the reliability changes when the changes we identified take place).

4.5.2 Estimating the Reliability of Police Open Data

Now we apply the procedure described in Section 4.4, that is, the procedure for estimating reliability variations by analyzing open data only. We first define a similarity

measure, then we look for variations in the similarity of consecutive datasets over time.

Case Study Setup

To facilitate the analysis of these data, we develop an application[1] that allows us to visualize the similarities between the consecutive datasets to allow a visual analysis. Figure 4.11 shows a screenshot of the application, that allows us to load a set of CSV files containing the datasets. In particular, in Figure 4.11 we can see the plot of the similarities of the datasets considering only the "Other crime" category, which is one of the crime categories which is most likely to be affected by open data policy changes, as we describe in the previous subsection. In the figure, the presence of a peak is likely to indicate one of the policy changed highlighted before. In general, here the challenge is twofold: on the one hand we should let these changes emerge, and on the other hand, we should try to run tests from different points of view so that possibly all changes are likely to emerge. Since in this case we do not have at our disposal the closed data to validate our hypotheses, the fact that more tests confirm a given variation might be a good indication of its high quality.

In general, the procedure that we apply is the following.

Choose one or more similarity measures. Compute the similarity of the neighbourhoods of datasets about consecutive months. If more than one similarity measure has been chosen, then aggregate the scores for each neighbourhood.

Aggregate the similarity scores of the neighbourhoods. In this way we obtain a global similarity value. Like in Subsection 4.5.1, the aggregation can be made by means of arithmetic average or subjective opinion. Here the amount of evidence is very limited, so it makes sense to use a subjective opinion, which allows us to be more prudent: having two positives on a neighbourhood, the average returns 1 as similarity value, while the opinion is 0.67, both in the $[0, 1]$ range.

Analyze the series of similarities. We look for variations in the similarity that might signal a policy variation. A change point is detected by means of the changepoint package in R [93], so that the change point is detected automatically and without the need to set thresholds or other arbitrary settings. In particular, we use the *multiple.mean.cusum* function, that allows us to detect multiple change points in the series, based on variations of the cumulative sum.

Aggregate all the evidence per dataset couple. The results of this aggregation tell us which couple is likely to be represent a change point and which not.

Results

Here we describe the results obtained by analyzing the datasets at our disposal. We apply the procedure introduced above, and we show the results obtained from the

[1] The source code is available at `http://trustingwebdata.org/phdthesis/dceolin`.

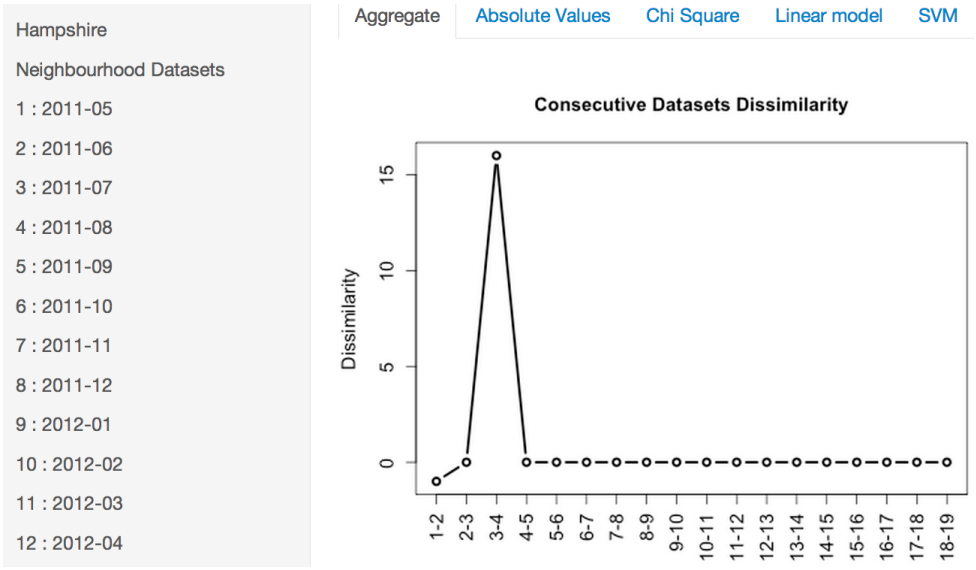


Figure 4.11: Screenshot of the application developed to visually analyze the open datasets. In the x-axis, the datasets are represented by means of numeric identifiers.

application of four tests by means of four graphs. All the tests are run by means of a script in the R programming environment. From each test we extrapolate a series of change points, by analyzing the variations in the mean of the cumulative sums. All these candidate change points are aggregated, again by means of an R script and the results are shown in Figure 4.16.

We start by applying to these datasets an analysis which is similar to one performed before. We compare, on neighbourhood basis, the distribution of the crime counts among the crime categories, and we check, using a Wilcoxon signed-rank test, how similar the two are: we represent the similarity between two datasets as the percentage of neighbourhoods that are statistically similar.

The results of the comparison are reported in Figure 4.12. The datasets are indicated by means of a sequential number (the first circle corresponds to the similarity between the first and the second dataset, and so on). The plot highlights that the twelfth comparison constitutes a change point: before that, the datasets are highly similar each other, and similarly after it. But at that point, the similarity trend breaks and starts a new one: that is likely to be a point where the reliability of the datasets diverges. We have found one of the discontinuity points we discovered in Section 4.5.1.

A second analysis consists of combining the results of the Wilcoxon signed-rank test and of the χ^2 test for each neighbourhood, averaged. Each test gives a positive piece of evidence if the neighbourhoods are not statistically different, at 95% confidence level. Figure 4.13 reports the results. Here we can see that this test highlights the

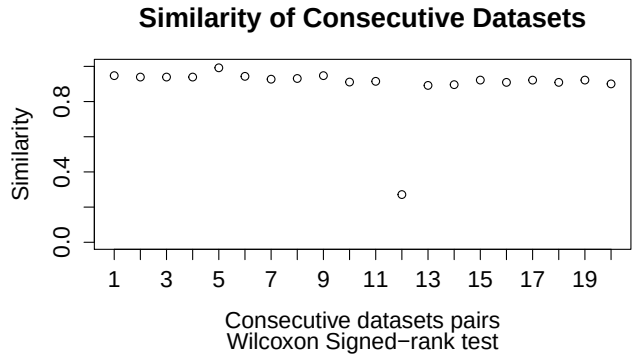


Figure 4.12: Plot of the similarity of consecutive datasets of crime counts for the Hampshire Constabulary from the `data.police.uk` website, computed by means of the Wilcoxon signed-rank test.

first change point we discovered before, between the fourth and the fifth dataset.

Lastly, we compute the sum of the differences (Figure 4.14) and the sum of the absolute differences of the neighoburhoods (Figure 4.15).

Some of the tests we run agree with each other, some not, and we do not have any prior information about which to trust more. Therefore we aggregate all the evidence that we obtain, and what results is the graph represented in Figure 4.16, where the two peaks correspond exactly to the two change points identified before. There is also a third peak, less pronounced, but we do not have at our disposal enough information to say whether it is due to actual policy changes, imprecision of the tests or particular events happened. Moreover, despite the previous case, here we can not say whether a change point indicates the start of an increase or decrease in reliability. However, these results are useful to pinpoint these events in order to facilitate an analyst to understand the eventual magnitude of the reliability variation. As we noted earlier, in the open data there are two structural variations, that is one regarding the number of crime categories reported, and (at least) one about the number of neighbourhoods represented. The procedure that we propose here does not rely on an analysis of the structure and of the schema of the dataset (although we may be able to add that in the future, as an additional item of evidence). The procedure identifies candidate points of policy change by analyzing the data themselves, thus suggesting that a policy change has probably happened in correspondence to the structural changes (the two points coincide).

4.6 Conclusions

In this chapter we present two procedures for the computation of the reliability of open data: one based on the comparison between open and closed data, the other one

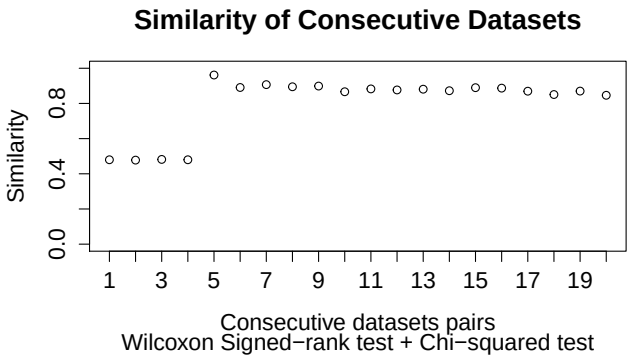


Figure 4.13: Plot of the similarity of consecutive datasets of crime counts for the Hampshire Constabulary from the `data.police.uk` website, computed by summing up the results of the Wilcoxon signed-rank test and of the χ^2 test.

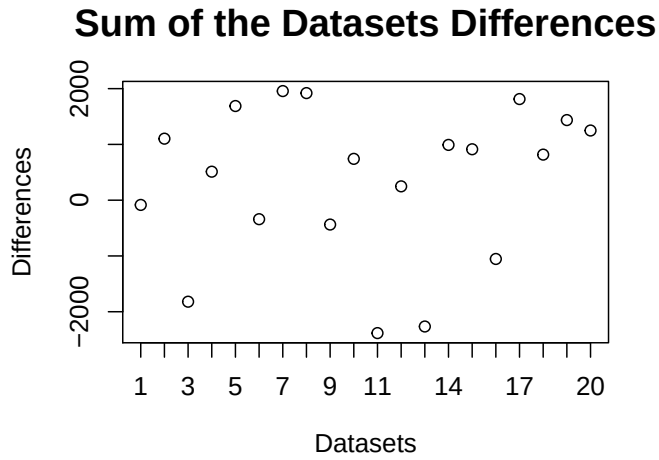


Figure 4.14: Plot of the similarity of consecutive datasets of crime counts for the Hampshire Constabulary from the `data.police.uk` website, expressed in terms of sum of dataset differences, computed at neighbourhood level.

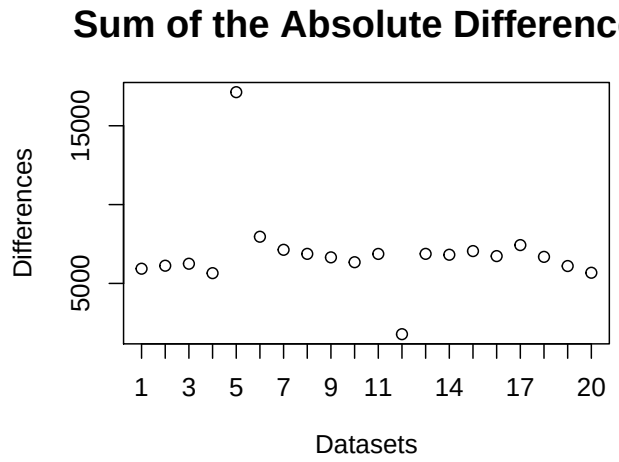


Figure 4.15: Plot of the similarity of consecutive datasets of crime counts for the Hampshire Constabulary from the `data.police.uk` website, expressed in terms of sum of the absolute value of the differences between crime counts, at neighbourhood level.

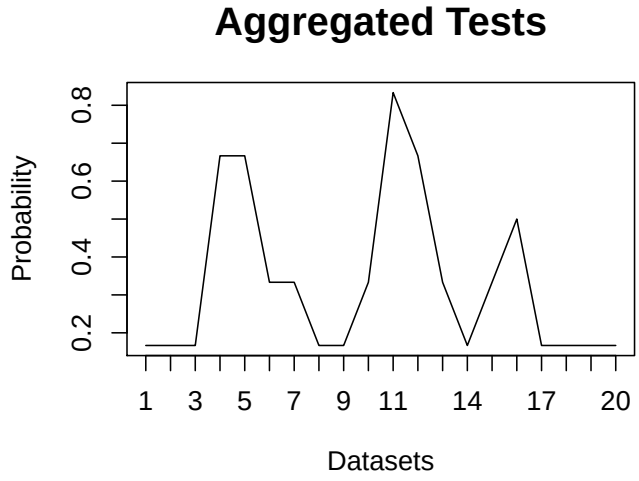


Figure 4.16: Plot of the similarity of consecutive datasets of crime counts for the Hampshire Constabulary from the `data.police.uk` website.

based on the analysis of open data. Both procedures are evaluated using data from the `data.police.uk` website and from the Hampshire Police Constabulary in the United Kingdom. The first procedure shows to be effective in allowing us to estimate the reliability of open data, showing also that smoothing procedures preserve a high data reliability, while allowing authorities to anonymize them. Different tests show different reliability levels for these data, however the overall reliability is high. Also, the procedure adopted to produce these data effectively affect data reliability, and the most recent policies adopted show a higher ability to preserve data reliability. The second procedure shows to be useful to grasp indications about data reliability and to identify the same critical points detected using the first procedure. Nevertheless, the quality of the results achieved with this method is lower than the one achieved when the closed data are available, because it does not allow us to directly and explicitly estimate the reliability of the data analyzed, although it produces very useful results to achieve this result in a semi-automated manner.

These two procedures provide additional value to the open data, as they allow us to enrich the data with information about their reliability: even though these data are already provided by authoritative institutions, these procedures can increase the confidence both of insider specialists (the first procedure, which relies on closed data) and of common citizens (the second procedure, which relies only on open data) who deal with them. These procedures show that the statistical approach adopted in Chapter 2 has a big potential when using Web data to make trust assessments. These also confirm the usefulness of subjective logic when dealing with uncertain observations and continues the adoption of statistical reasoning on Web sources for trust assessments outlined in Chapter 3.

Part III

Provenance Analyses for Assessing Trust

The use of provenance for making trust assessments is the main topic of the chapters collected in this part. Provenance has already been touched in Chapter 4, but here we propose two works on the use of provenance information for estimating trust evaluations. The work here presented is built upon the research introduced in Parts I and II, because it is based on the use of Web data, handled by means of uncertainty reasoning, however the focus is the use of a particular kind of data, namely provenance for our trust estimates. We present two approaches for linking provenance graphs to trust estimations. One is presented in Chapter 5, where we build a Bayesian network using subjective logic on top of a provenance graph and, in this way, we derive an estimate for the trustworthiness of a ship message. The other one is presented in Chapter 6, where we use support vector machines to classify provenance graphs of video tags according to the estimated trustworthiness of the tags they refer to.

5

Provenance Analyses for Trust Assessment

A Maritime Domain Case Study

*In Parts I and II we point out the importance of using uncertainty reasoning techniques to handle Web data for making trust assessments of semi-structured data. Here we continue on this direction, by making use of a specific kind of Web data, provenance data. So, we start addressing the third research question (*How can provenance information be used for making accurate trustworthiness estimations of semi-structured data?*). The work presented here extends the research described in Chapter 1 and in Chapter 3 by providing an algorithm capable of estimating the trust in data by making use of uncertainty reasoning (and, thus, touching again the second research question,* How can uncertainty reasoning be effectively used to estimate the trustworthiness of semi-structured data?*) over provenance graphs that describe how data are produced. Such an algorithm takes advantage of the semantics of the graph that uses, and combines the subjective opinions (see Chapter Preliminaries) computed about the artifacts in the graph by using subjective operators that reflect the operations applied on the data. These operations are reported in the graph and recorded by means of provenance information. We apply this algorithm in the naval domain, and we use it to estimate the trustworthiness of a set of ship messages.*

This chapter is based on the paper Calculating the Trust of Event Descriptions using Provenance *coauthored with Paul Groth and Willem Robert van Hage and presented at the* 2nd Workshop on the Role of the Semantic Web in Provenance Management (SWPM 2010) *at the* 9th International Semantic Web Conference (ISWC 2010)*, and the chapter coauthored with Willem Robert van Hage, Guus Schreiber and Wan Fokkink* Assessing Trust for Determining the Reliability of Information *in the book* Situation Awareness in Systems of Systems*, published in 2013 by Springer.*

5.1 Introduction

In the naval domain, particular messages, called "Automated Identification System" (AIS) messages, are periodically exchanged between ships and captured by particular receivers that allow ship and land based naval authorities to avoid collisions, to locate and to identify ships. These messages contain important information, about the identity of the ship (identification code, name, flag, ship dimension, etc.), about its location (latitude, longitude, timestamp of the message, that is, temporal identification of the moment when the message is sent) and about kinematic data (for instance, speed and heading) and allow authorities to keep track of the position of the ships, together with their identity. However, these messages can, in principle, be intentionally or unintentionally manipulated by the senders. For instance, there exist episodes of ships willing to impersonate the identity of others to evade controls[1]. Trust plays a crucial role when dealing with these messages, because the information that they provide is not always certain, but a naval operator that reads them, would like to know whether he can trust them.

This chapter describes how it is possible to estimate appropriate trust in the information that the message exposes. We do not consider the explanations for a possible error in a message (e.g., intentional versus unintentional information manipulation). We limit ourselves to collect the information at our disposal and to check if these are trustworthy. We treat a message as an "annotation of a ship", providing us with information about the vessel considered. This approach makes the work presented here resemble and extend the research presented in Chapter 1. We base our evaluations on two factors: the reputation of the sender (and, more generally, the "provenance" of the message, that is, who produced it and how), and the co-occurrence of multiple observations supporting or contrasting the information provided by the message itself. Roughly speaking, this means that we trust messages when they are sent by well-reputed agents and we trust information that is confirmed by many agents.

AIS messages can be seen also as event descriptions, as they describe the ship movements over time. Therefore, in this work, we investigate the generation of trust ratings for event descriptions. These trust ratings are calculated with respect not only to the original sources but also to the data integration process itself. Thus, the trust calculations consider the whole of an event description's provenance. The trust algorithms presented here rely on the novel combination of two existing representations, the Simple Event Model (SEM) for event representations and the Open Provenance Model (OPM) for representing the data integration process itself (in Chapters 6 and 8 we make use of the PROV Ontology [9] to represent provenance. OPM is a precursor of PROV). Based on a mapping of these models, we develop a trust algorithm using subjective logic (see Chapter Preliminaries) that extends the algorithm presented in Chapter 1. We apply our trust algorithm to a use case from maritime shipping. The contributions of this chapter are twofold:

[1]For instance, in January 2009 the fleet of an Iranian company tried to disguise its identity. See: `http://www.nytimes.com/2010/06/08/world/middleeast/08sanctions.html?pagewanted=all`

1. a mapping of SEM to OPM;
2. an algorithm for computing trust ratings for event descriptions based on their provenance.

The rest of this chapter is structured as follows. In Section 5.2 we describe related work. In Section 5.3 we describe the mapping between OPM and SEM; in Section 5.4, we describe how to represent AIS messages by means of subjective opinions. In Section 5.5 we present the trust rating algorithm and in Section 5.6 we apply it and we describe the results obtained. Finally Section 5.8 presents a final discussion.

5.2 Related Work

For general references about trust computation and about the use of provenance to make trust estimations, we refer the reader to Chapter Introduction. An introduction to subjective logic, that is widely employed in this work, can be found in Chapter Preliminaries. This chapter proposes also a mapping between two ontologies, OPM and SEM. For additional references about ontology mapping, see the work of Euzenat and Shvaiko [54], while a detailed description of mappings between OPM and other ontologies can be found in the work of Sahoo et al. [140].

5.3 Mapping SEM and OPM

In order to connect the description of an event to how that description was created, we need to be able to interpret the event description with respect to its provenance. To do so, we provide a mapping from the model used for event descriptions (SEM) to the model used for describing provenance (OPM). To facilitate the explanation of this mapping, we first briefly introduce both SEM and OPM.

5.3.1 SEM, the Simple Event Model

SEM [160, 161, 162] is a schema for the semantic representation of events. It does not deal with the way data about events is stored, but only with the events themselves. SEM focuses on modeling the most common facets of events: who, what, where, and when. These are represented respectively by the SEM core classes sem:Actor, sem:Place, sem:Object and sem:Time. SEM is a model that takes into account the inherent messiness of the Web by making as little semantic commitment (e.g., disjointness statements, functional properties) as possible. Every instance of one of the core classes can be assigned types from domain vocabularies. For example, the sem:Event instance ex:world_cup_2010 can be assigned a sem:eventType dbpedia:FIFA_Club_-World_Cup. Any property of SEM, including the type properties, is optional and duplicable. SEM and Simple Knowledge Organization System (SKOS) [170] mappings

to related models can be accessed online[2]. Additionally, through sem:View an event can have multiple, perhaps conflicting, descriptions.

5.3.2 OPM, the Open Provenance Model

OPM is a community developed model for the exchange of provenance information [113]. It stems from a series of interoperability challenges (Provenance Challenges) held by the provenance research community to understand and exchange provenance information between systems. While not as comprehensive as some other provenance models such as ProPreO [141] , OPM provides a common technology-agnostic layer of agreement between systems. OPM was used by 15 teams during the Third Provenance Challenge [113]. These teams used a variety of provenance management systems ranging from those focused on workflow systems to those concentrating on operating systems. Thus, by using OPM, we aim to be able to apply our trust algorithm to a variety of systems. OPM has been superseded by the PROV ontology [9], that is used in Chapters 6 and 8. Since the two models are compatible each other, in this chapter we focus on OPM.

OPM represents the provenance of an object as a directed acyclic graph with the possibility for annotations on the graph. The graph is interpreted as being causal. An OPM graph captures the past execution of a process. The graph consists of three types of nodes:

- An *opm:Artifact*, which is an immutable piece of state, for example, a file.
- An *opm:Process*, which perform actions upon artifacts and produce new artifacts. An example of a process would be the execution of the Unix command cat on two files to produce a new concatenated file.
- An *opm:Agent*, which controls or enables a process. An example of an agent would be the operating system that a process runs in or the person who started the process.

These nodes are linked by five kinds of edges representing dependency between nodes. An opm:Process used and generated opm:Artifacts, represented by opm:used and opm:wasGeneratedBy edges. These artifacts can be given an opm:Role with respect to an opm:Process distinguishing it from other artifacts. Note, an opm:Process can only produce one opm:Artifact. Dependency between opm:Artifacts is represented using opm:wasDerivedFrom while dependency between opm:Processes is represented using the opm:wasTriggeredBy edge. Finally, the control of an opm:Process by an opm:Agent is expressed using the opm:wasTriggeredBy edge.

Each part of an OPM graph can be labeled with an *account*, which allows the same execution to be explained from different perspectives. For example, one could describe the generation of an event description with more or less detail.

[2]The mappings are available at `http://semanticweb.cs.vu.nl/2009/11/sem/`.

SEM	SKOS relation	OPM
opm:Process	skos:closeMatch	sem:Event
opm:Artifact	skos:closeMatch	sem:Actor
opm:Agent	skos:broadMatch	sem:Actor
opm:Artifact	skos:closeMatch	sem:Place
opm:Agent	skos:broadMatch	sem:Place
opm:Role	skos:closeMatch	sem:Role
opm:Account	skos:closeMatch	sem:View
opm:Observer	skos:closeMatch	sem:Authority
opm:Role	skos:closeMatch	sem:Role

Table 5.1: Mapping between OPM and SEM classes.

5.3.3 Mapping

Given an event description in SEM, we would like to determine how its facets should map to OPM so that we can describe the facet's provenance using OPM. For example, if an event occurred at a sem:Place, we could consider that place an opm:Artifact. This idea is in-line with the notion of sub-typing within OPM [113]. We could say that a particular opm:Artifact has a type of sem:Place. To represent the mapping, we use SKOS, a W3C standard for describing and mapping vocabularies (i.e., concept schemes). The use of SKOS follows the practice of the W3C Provenance Incubator Group in defining a set of Provenance Vocabulary Mappings [140]. We refer the readers to the work of Bechhofer et al. [170] for the exact definitions of skos:closeMatch, skos:relatedMatch and skos:broadMatch. Since SKOS Mappings support mappings between concepts from different schemes, we use RDFS [168] to align SEM and OPM properties.

Our mapping focuses mainly on the nodes within the OPM graph, since SEM nodes find a straightforward correspondent in OPM nodes, while OPM edges capture information that is not explicitly considered by SEM. Our aim is to describe the provenance of both the event description described using SEM, and its facets. In Table 5.1 we report the mapping at class level, while in Table 5.2 we report the core elements of the mapping at property level. A more comprehensive description of the mapping is available on the Web.[3]We now discuss the mapping shown in Tables 5.1 and 5.2 in more detail.

Each sem:Event is an action with some duration, this maps very closely with the notion of an opm:Process. SEM has the notion of a sem:Actor, the entities or people *who* take part or are involved in an event. If a sem:Actor is directly a cause or is vital for an event to take place, we would model this as an opm:Artifact used by an opm:Process. For people who were not directly involved but enabled the event to take place, the sem:Actor would be mapped to an opm:Agent. By way of example, the crew on board a ship would be modeled as opm:Artifacts while the CEO of the shipping

[3]The complete mapping is available at `http://trustingwebdata.org/phdthesis/dceolin`.

OPM	RDFS relation	SEM
opm:used	rdfs:subPropertyOf	sem:hasActor
opm:wasGeneratedBy	rdfs:subPropertyOf	sem:hasActor
opm:wasTriggeredBy	rdfs:subPropertyOf	sem:subEventOf
opm:refinement	rdfs:subPropertyOf	sem:subEventOf
opm:wasControlledBy	rdfs:subPropertyOf	sem:hasActor

Table 5.2: Mapping between OPM and SEM properties.

company can be seen as an opm:Agent controlling the event of sending an AIS message. Similar reasoning applies to mapping sem:Place to OPM. The sem:Role signifies the role a particular SEM facet plays in an event, just as an opm:Role signifies the role a particular opm:Artifact plays with respect to an opm:Process. Additionally, a sem:View allows for multiple descriptions of the same event, which maps naturally to an opm:Account describing different descriptions of the same execution. Finally, the time of a sem:Event can be easily mapped to the time annotations present on OPM edges. The OPM properties reported in Table 5.2 are all treated as subproperties of corresponding SEM properties. This is because we see provenance as a class of information regarding a particular class of events. Since SEM is designed to model events in general, OPM can in part be considered as a specialization of SEM. OPM allows also modeling interactions between agents and artifacts, but this kind of information is not modeled by SEM.

The mappings that we propose have been manually created. The main goal of these mappings is to align two vocabularies that are aimed at modeling two classes of tightly related and partly overlapping information from two points of view. These mappings can serve as a basis for deriving provenance descriptions starting from event descriptions modeled using SEM, and hence contribute in the estimation of the trustworthiness of such event descriptions by means of the trust assessment algorithm that we define as follows. However, in order to tackle this problem, SEM event descriptions should be assisted by additional information, in order to cover the gaps identified in the mapping between OPM and SEM. For instance, SEM models artifact and agents that take part in events, and this is an important information modeled by OPM as well and used to determine the trustworthiness of artifacts like event descriptions. To properly estimate such trustworthiness, we need to know information such as who created a given artifact. OPM allows modeling this, SEM does not (at least explicitly).

5.4 Subjective Logic for Trusting AIS Messages

Opinions are the basic element of subjective logic (see Chapter Preliminaries), because they are the means to link logical statements to probabilities and to contextualize them. We can encode AIS messages as opinions exposed by particular sources. For instance, if our source (or "subject") is AIS_1 and we are determining whether the

name of the ship is "Beauty" (this statement constitutes our "object"), then we can represent it as an opinion (represented by the symbol $\omega_{object}^{subject}$) as:

$$\omega^{AIS_1}_{\textit{the name of Ship}_{123}\ \textit{is}\ \text{“Beauty”}}\left(\frac{1}{3}, 0, \frac{2}{3}, \frac{1}{2}\right)$$

Suppose, that one of the messages is retrieved through a receiver that we know is not always reliable. This means that at least the uncertainty of the opinion computed on the basis of such a message should be increased (because we do not know if the receiver was working properly or not, when it recorded the message). This is obtained by using the "discount" operator ($\otimes$) that weighs the opinion on the message itself according to the opinion on the receiver, that is, on the reputation of the receiver. This allows us to "smoothen" strong opinions coming from subjects (that is, sources) of which the reputation is not surely positive, while allowing us to incorporate opinions about facts of which we do not have direct observations, but that are "told us" by third parties (in this case, the receiver). Here is an example. If our opinion about AIS_1 was

$$\omega^{we}_{AIS_1}(0.4, 0.4, 0.2)$$

and the opinion given by AIS_1 is the one we have seen before,

$$\omega^{AIS_1}_{\textit{the name of Ship}_{123}\ \textit{is}\ \text{“Beauty”}}(0.333, 0, 0.667)$$

we can weigh this opinion on the basis of AIS_1's reputation by applying the discount operator ($\otimes$) and the result is:

$$\omega^{we}_{AIS_1}(0.4, 0.4, 0.2) \otimes \omega^{AIS_1}_{\textit{the name of Ship}_{123}\ \textit{is}\ \text{“Beauty”}}(0.333, 0, 0.667) =$$

$$\omega^{we:AIS_1}_{\textit{the name of Ship\_123}\ \textit{is}\ \text{“Beauty”}}(0.133, 0, 0.867)$$

Other operators are available, in order to allow different logical operations to be applied to the statements of our interest and to have the corresponding beliefs, disbeliefs and uncertainties properly updated. The choice of the correct operator to be applied on the opinions at our disposal depends on the relations between objects and subjects and is usually related to domain knowledge. Chapter Preliminaries shows a range of subjective logic operators. In Section 5.5 we propose an algorithm that makes use of such operators to compute the trust of AIS messages, and we explain how these operators are chosen.

5.5 Estimating the Trust in AIS Messages Using Provenance

First we present the algorithm, then we describe it in detail.

Algorithm 5.1: Trust Rating Algorithm

```
1  tv(A_i)
2      res ← null
3      for P_k : A_i opm : wasGeneratedBy P_k do
4          for A_j : P_k opm : used A_j do
5              if A_i opm : wasDerivedFrom A_j then
6                  if res = null then
7                      res ← tv (A_j)
8                  else
9                      res ← F(P_k)(res, tv(A_j))
10     for s_i : ∃v_{s_i}(A_i) ≠ ∅ do
11         if res = null then
12             res ← opinion_source(A_i)
13         else
14             res ← res ⊕ opinion_source(A_i)
15     return res
16 opinion_source(A_i)
17     for s_i : v_{s_i}(A_i) ≠ null do
18         record_evidence(v_{s_i}(A_i))
19     return ω^{x:s_i}_{v(A_i)}
20 π(t, s_i, A_i)
21     e : e ∈ domain∧ dist(e,v_{s_i}(A_i)) = min_{∀e'∈domain}(dist(e', v_{s_i}(A_i)))
22     d ← dist (e, v_{s_i}(A_i))
23     record ω^{s_i}_{v_{s_i}(A_i)=e}(b'_{s_i} · 1/d, 0, (d'_{s_i} + u'_{s_i}) · (1 − 1/d), a'_{s_i})
24 dist
       distance between two points (e.g., Euclidean)
25 record_evidence
       stores evidence in memory
26 record
       stores opinion in memory
27 ω
       returns an opinion based on stored evidence
   Possible values for F:
   F(concat) = ∧
   F(lookup(t)) = ∧ · π(t)
```

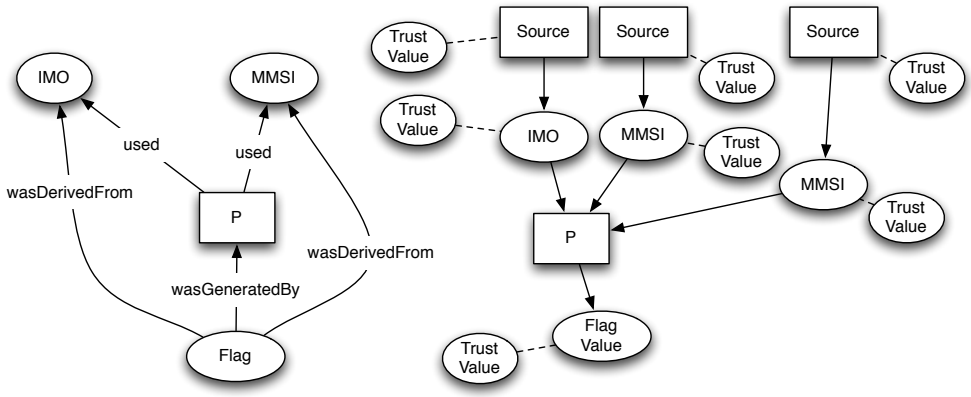


Figure 5.1: Provenance and Trust graphs about the flag value of a ship. The left graph reconstructs the provenance of the flag field. The graph on the right, starting from the first ancestors of the flag field, collects all the evidence about all the artifacts involved in the provenance trail (of the left graph) and gradually merges them.

5.5.1 Trust Rating Algorithm

Here we present a formal definition of the algorithm for calculating the trust value of an event facet, represented by artifacts which is also applicable to AIS messages.

Given an artifact to calculate the trust value of, our first step is to determine the opinion of any source that directly generates the artifact's value. Then we:

- take the amount of evidence given by each source about each possible value for the artifact.
- weigh the opinions given by the sources according to the opinion on the source itself (in turn, based on previous evidence about its trustworthiness);
- merge all the opinions.

Generalizing, we can say that:

- given an artifact A;
- given a set of sources: $s_1, \ldots, s_n$
- given a function $v(s_i, A) = v_{s_i}(A)$
- given opinions on the sources $\omega^x_{s_i}(b_{s_i}, d_{s_i}, u_{s_i}, a_{s_i})$

we compute the opinion on an event facet from each source:

$$\omega^{x:s_i}_{v_{s_i}(A)}(b_{s_i}, 0, d_{s_i} + u_{s_i}, a_{s_i})$$

Once we have the opinions about the values from each source, we merge them in order to obtain an opinion for each value from all sources:

$$\bigoplus_{v_{s_i}} \omega^{x:s_i}_{v_{s_i}(A)}(b_{s_i}, 0, d_{s_i} + u_{s_i}, a_{s_i})$$

5.5.2 Integration Process

We want to consider not only sources that directly provide the artifact value but also which process is used during integration to generate the artifact. Therefore, in case the artifact is not a leaf node, we need to merge the (eventual) opinions computed, taking into account the provenance of the artifact. For example, considering the example of Figure 5.1, we see that the trust level of the root node depends on the trust levels of the leaf nodes, combined according to how the process manipulates them. Therefore, we should use a functor that allows us to apply proper functions to the trust values of the input artifacts, according to the kind of process that manipulates them.

Two examples are provided in Algorithm 5.1: in case of a concatenation process (that takes as inputs two strings and outputs their concatenation), all the trust values equally contribute to determine the outcome and therefore are merged by conjunction. In case of a lookup process (that takes as inputs a key and a value table, and outputs the value in the table corresponding to the key), then before calculating the conjunction of the trust values, we project them into the space of the possible values, possibly smaller than the space of plausible ones. Moreover, in case the value we face does not fall into the range of possible values, we consider the value or values closer to it that belong to the set of possible values. Clearly, we weigh these contributions according to the distance to the given value.

Algorithm Advantages

We now discuss how, by taking advantage of both provenance and background knowledge, the trust algorithm can produce precise trust ratings.

By means of provenance, the algorithm incorporates semantic information. In this way, it restricts the domain of possible value for each field to the range of real, meaningful values. For instance, if the nationality field of a MMSI is a three-digit code, then there are 10^3 possible values, since any cypher would be equally probable in each of the three positions. By taking into account the meaning (semantics) of the MMSI, the cardinality of the set of the plausible values would restrict to 35 (considering the countries that own 99% of the ships). So, if we own 10 positive pieces of evidence and we restrict the plausibility set from 1000 to 35, then the trust value rises from

$$E = \frac{10}{1010} + \frac{1}{1000} \cdot \frac{1000}{1010} = 0,0189$$

to

$$E = \frac{10}{45} + \frac{1}{35} \cdot \frac{35}{45} = 0,3143.$$

The MMSI field is retrieved traversing the provenance graph.

Also thanks to the use of provenance, we enlarge the availability of evidence at our disposal for calculating trust values. In fact, we do not limit to the use of direct evidence about the facets we have to evaluate, but we consider and properly handle also evidence about elements used in the process that lead us to our facets. Therefore, we check whether these initial elements were correct and whether they were combined properly in order to produce the facet we are analyzing. Once we have this result, we can compare it with evidence directly referring to the facet we are evaluating, obtaining an improvement of the precision of the trust value. Continuing the previous example, if we have also sources that provide a value for the nation, knowing that the national code is determined by looking it up into a trusted table, then by applying the Trust Ranking Algorithm, we obtain the following trust value:

$$E = \frac{20}{45} + \frac{1}{35} \times \frac{35}{45} = 0,4667.$$

Since we adopt a conservative approach and accept only facets which trust value is above a certain threshold, then this latter advantage reduces the number of errors due to false negatives.

5.6 Algorithm Application

An AIS message contains, amongst others, the following fields:

- IMO: unique identification code from the International Maritime Organization;
- MMSI: the Maritime Mobile Service Identity code is a nine digits code used for communication purposes. Its first three digits are determined on national basis;
- CallSign: four or five digits communication code. Its first two digits are determined on the basis of the nationality of the ship;
- Name of the ship;
- Flag of the ship.

We apply subjective logic reasoning via the algorithm introduced before on the fields reporting static information about the ship (like IMO, MMSI and CallSign), and not on the fields reporting kinematic information (like speed or heading). For each field we compute an opinion based on all the available evidence, that is, AIS messages and information crawled from the Web. In particular, we crawled `www.vesseltracker.com` [164] and `www.shipais.com` [144] websites. Then, we merge all the opinions taking into account provenance information, that is, how information contained in the fields is produced. Since there exists also a dependency relation between certain fields, provenance allows us to encode dependencies between them, as we will see later. The trust level of the whole static part of a message is determined by combining the trust

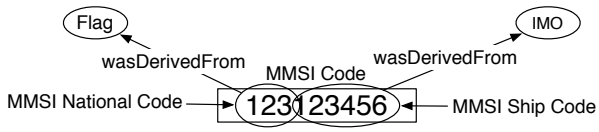


Figure 5.2: Graph representing the provenance of the MMSI code of an AIS message. The MMSI depends on the Flag field (national part) and on the IMO code (ship part).

level computed for each field separately. These pieces of information are combined by the "AND" operator. The reason why we use this operator is that, in order to be considered "trustworthy", a message should carry only correct (or "trustworthy") information. If one or more pieces of information are not, then the whole rating should be affected by this.

There are two categories of fields: independent and dependent (or partially dependent) fields. Fields like, for instance, the width or the name of the ship, are not bound to any other information within the message itself. So, to compute the trust value for these fields, we gathered all the evidence available and properly count them to build an opinion. Other fields, like the MMSI code and the CallSign are dependent on the flag field, which represents the nationality of the ship. The Open Provenance Model can help us to record this information. Assuming that the IMO is a code able to uniquely identify the ship, we can record the following relations:

```
MMSI_national_code opm:wasDerivedFrom Flag .
MMSI_ship_code opm:wasDerivedFrom IMO .
MMSI opm:wasDerivedFrom MMSI national code .
MMSI opm:wasDerivedFrom MMSI_ship_code.
```

Listing 5.1: RDF representation of the provenance of the MMSI code using OPM.

The CallSign field is defined exactly in the same way. For each field we have a small graph (for instance, see Figure 5.2) with the field itself being dependent on two components. We know from the domain that the process that produced the codes is the concatenation process. From the trust perspective, it means that the two input elements do not influence each other, because they determine the value of the two elements that, once concatenated, lead to the overall code. Therefore, these two elements need to be both true so that the whole message can be true ("AND" operation). So, we computed the opinion for the second part of the MMSI, of the CallSign and of the flag, based on the available evidence. Figure 5.3 shows the network of information that we have just described.

We do not have the possibility of determining the MMSI local part given the IMO code from a reliable service, but by employing subjective logic and exploiting provenance information we could obtain reliable estimates for this 6-digit code (that is, the ship code of the MMSI). The national part, instead, we have the possibility of mapping it into the nation that it represents. Therefore, we can merge all the evidence we have about the flag, the MMSI national part and the CallSign national part into a

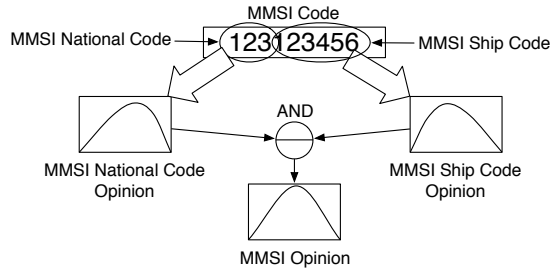


Figure 5.3: The computation of an opinion regarding the communication code of $Ship_{123}$ is made by merging opinions on the national and ship components of the code. The national part is evaluated by considering the nationality of the ship.

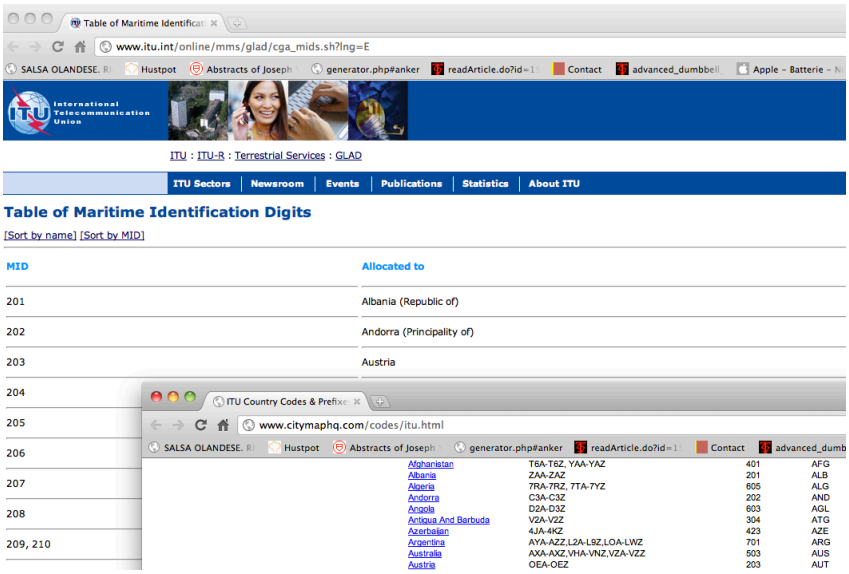
single opinion about the nationality of the ship. The national part of the MMSI code is a three-digit code. Before doing this, we need to have a map that collects all the national codes for MMSI and CallSign. We retrieved these maps from Web repository of places-related information and of communication codes. In particular we crawled the Table of Maritime Identification Digits from International Telecommunication Union website [81] and the Citymap HQ website [36] (see Figure 5.4(a)).

Finally, in order to determine the trust values of the AIS messages, we compute:

- the trust values for all the "independent" fields (e.g., the flag). Note that the AIS messages also report a timestamp, i.e., a field indicating when they are sent. We use all the evidence available at a given point: messages arrived before the one that we are analyzing, and all the Web data available (e.g., AIS-related data);
- the trust values for all the "dependent" fields (e.g., CallSign and MMSI code), by applying the AND operation on the input elements (national and local codes);
- the trust value of the whole message, by computing the AND operation of the trust values of all the fields.

5.7 Results

We compute the trust values for all the messages in our dataset, that covers one week period. In addition to these data, we consult a few Web sources [144, 164] to increase the amount of data at our disposal. An example of the visualization of the results of these calculations is available in Figure 5.4(b). Table 5.3 reports some summarizing statistics about the trust levels computed. We do not have any information about the reputation of the sources at our disposal, so the belief average concentrates around the middle of the range, because of the initial situation of high uncertainty. Moreover, when different disagreeing values are proposed by different sources for a given field, we compute the trust values for all of them. This explains why the range of the beliefs is so



(a) Web sites screenshots

(b) Trust visualization

Figure 5.4: Screenshot of International Telecommunication Union and Citymap HQ websites (Figure 5.4(a)) and trust value visualization (Figure 5.4(b)). The ship is localized thanks to the data contained in an AIS message.

	min value	max value	average	median
belief	0.0005	0.9985	0.5834	0.5
uncertainty	0.0015	0.5	0.2578	0.1667

Table 5.3: Statistics about the belief and uncertainty of the trust level computed.

wide: correct values are very popular and so have a high trust value, and consequently, wrong or non-comforming values (like messages reporting MMSI code value "0") get a lower trust value. The maximum value for uncertainty is 0.5 because it corresponds to the uncertainty of an opinion based on one observation, that is, on the first message. Opinions are computed incrementally, so we compute an opinion for each message, considering all the messages observed up to that moment. So, consecutive opinions manifest decreasing uncertainty and the belief in rare values decreases, while the belief in common values increases. For instance, a belief of 0.0005 corresponds to one positive piece of evidence over 1998 pieces of evidence in total ($\frac{1}{(1998+2)} = 0.0005$; 2 is the range of possible evidence (true, false)). Vice-versa, a belief of 0.9985 corresponds to 1997 positive pieces of evidence over 1998 observations in total ($\frac{1997}{(1998+2)} = 0.9985$).

5.8 Discussion

The choice of the model is driven by the clear requirements that the problem has. One important requirement is the impossibility to assume that the evidence at our disposal is the result of a random sampling process. This is because the evidence considered by us is not the result of a controlled drawing process, similar to the situation described in Chapter 3. Rather, we use all the observations at our disposal without any information about their reliability or representativity. This is an important consideration and motivates why, for instance, we do not take a classical Bayesian approach, and, more precisely, why we do not assume that our data are normally distributed: if there is biased manipulation in the messages, their distribution could have taken any shape. This fact leads us to the following differences with respect to a classical Bayesian approach, similar to the consideration proposed in Chapter 3:

- we can not assume that our observations are identically independently distributed, for the reasons that we have just explained, so we can not make use of estimators based on normality assumptions;

- sources' reputations have to be explicitly recorded and incorporated in our evaluations. If we have no prior information, their reputations are neutral (neither positive nor negative), but the data they provide are considered as uncertain;

- we do not know what is the representativity of the model that we infer from the observations, hence we estimate the likelihood of the various possible models, instead of directly estimating the most likely values. We infer two orders of

probability: one about the possible models and one about the outcomes, given the most likely model.

- finally, the uncertainty component of opinions as such is a typical characteristic of subjective logic that concisely quantifies the lack of information. There is no parameter in Bayesian models directly corresponding to it.

The use of first and second order probabilities (see Chapter Preliminaries) is useful when dealing with multiple levels of uncertainty (uncertainty about the outcomes and uncertainty about the model representing the data), because it prudently computes a probability distribution based on the actual observations. This probability distribution can be used, for instance, by a decision strategy with the aim of deciding whether a message is trustworthy or not. The prudence of the model is due to the limiting assumptions on which it is based (for instance, it does not assume that the observations are randomly obtained). The results presented in Section 5.7 highlight at least part of these considerations (e.g., the prudence of the estimates and the fact that these are built incrementally). The dataset at our disposal allows us only to verify that the algorithm proposed permits to estimate the trust level of AIS messages. We perform a first qualitative manual validation of the trust values computed, by verifying that the messages with lower trust level are incomplete or wrong and those with higher trust value are actually likely to be correct. Future work will be dedicated to a more extensive validation of the trust levels computed, provided that the necessary information will be then available. Lastly, one of the limitations of the algorithm is the need to map each process in the provenance information to a subjective operator that properly handles the opinions about the input facets of the process. Chapter 6 proposes an alternative approach to use provenance for trust assessments.

5.9 Conclusion

In this chapter we propose a mapping between OPM and SEM and an algorithm for computing the trustworthiness of event descriptions using provenance information. The mapping plays a facilitating role in this, because it bridges the description of events (represented by means of AIS messages encoded by means of SEM) with their provenance, expressed in OPM. Once we have a reliable representation of the provenance of the event description, we can estimate whether a message is trustworthy or not. The algorithm makes use of the provenance information at our disposal about the message we evaluate, together with the corresponding evidence. It combines the evidence using subjective operators that reflect the process that have been performed on the facets (or fields) that compose the message. The rationale behind the design of the algorithm is close to the principles described in Chapter 3. The algorithm makes use of a prudent statistical representation of Web data, by adopting probabilistic models that use smoothing. We provide a validation of the algorithm by evaluating a set of AIS messages using such an algorithm and a limited set of Web sources.

6

Combining Provenance with User Reputation for Trust Assessment

A Video Tagging Game Case Study

*This chapter tackles the problem of estimating the trustworthiness of media annotations by combining a reputation- and a provenance-based approach. Here we tackle the third research question presented in Chapter Introduction (*How can provenance information be used for making accurate trustworthiness estimations of semi-structured data?*) in a different fashion than we did in Chapter 5. In fact, the reputation of the annotation author is computed by means of subjective logic (like in Chapter 1), while the provenance-based annotation trust levels are computed by means of a machine learning approach, namely support vector machines. Moreover, the two values are combined by means of a procedure that we developed. The fact that here we use a machine learning approach quite successfully, provides another addressing to the second research question (*How can uncertainty reasoning be effectively used to estimate the trustworthiness of semi-structured data?*), already tackled in Chapters 3, 4 and 5. The algorithm proposed here is evaluated over a dataset of crowdsourced tag entries provided by a video-tagging game platform. The trust values of these tag entries are computed by means of the extension of subjective logic for handling partial evidence observations introduced in Chapter 3.*

This chapter is based on the paper Trust Evaluation through User Reputation and Provenance Analysis *, coauthored with Paul Groth, Willem Robert van Hage, Archana Nottamkandath and Wan Fokkink and presented at the* 8th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2012) *at the* 10th International Semantic Web Conference (ISWC 2012).

6.1 Introduction

Reputation is an important mechanism in our set of strategies to determine trust, and we have already employed it in the work presented in Chapter 1. However, we may base our assessment on a variety of other factors as well, including prior performance, a guarantee, or knowledge of how something was produced, as we have seen in Chapters 1 and 5. Nevertheless, many systems, especially on the Web, choose to reduce trust to reputation estimation and analysis alone. In this chapter, we take a multi-faceted approach. We look at trust assessment of Web data based on reputation, provenance (i.e., how data has been produced), and the combination of the two.

We first propose a procedure for computing reputations that uses basic evidential reasoning principles and is implemented by means of subjective opinions (see Chapter Preliminaries) that is similar to the procedure for computing the reputation of annotation authors proposed in Chapter 1. Secondly, we propose a procedure for computing trust assessments based on provenance information represented in the W3C PROV Ontology [9], that is a continuation of the OPM Model adopted in Chapter 5. Here, PROV plays a key role, both because of the availability of provenance data over the Web recorded by using this standard, and because of its role as an interchange format: having modeled our procedure on PROV, any other input format can be easily treated after having mapped it to PROV. The increasing effort spent in recording and sharing provenance information using PROV makes this procedure particularly important. Moreover, by showing that trust assessments based on combinations of reputation and provenance are more accurate than those based only on reputation, we show how a solution to trust issues can be found on the Web itself, as indicated also in Chapters 1, 2 and 5. We implement these procedures by discretizing the trust values and applying a support vector machine classification. Finally, we combine these two procedures in order to maximize the benefit of both. The procedures are evaluated on data provided by the *Waisda?* [119] tagging game[1], where users challenge each other in tagging videos. If the tags of two or more users regarding the same video are matched within a given time frame, they both get points. User consensus about tags correlates with tag trustworthiness: the more users agree on a given tag, the more likely it is that the tag is correct. We show how it is possible to predict tag consensus based on who created the tag, how it was created and a combination of the two. In particular, we show that a reputation-based estimation is not significantly different from a provenance-based estimation and, by combining the two we obtain a small but statistically significant improvement in our estimations. We also show that reputation- and provenance-based assessments correlate.

The rest of the chapter is organized as follows: Section 6.2 describes the dataset used for our evaluations, Section 6.3 introduces the trust assessment procedures based on reputation, provenance and their combination, including example associated experiments. Section 6.4 describes and discusses the results obtained with the three

[1] A zip file containing the R and Python procedures used, together with the dataset, is retrievable at `http://trustingwebdata.org/phdthesis/dceolin`

procedures. Section 6.5 provides conclusions.

6.2 The *Waisda?* Dataset

Waisda? is a video tagging gaming platform launched by the Netherlands Institute for Sound and Vision in collaboration with the public Dutch broadcaster KRO. The logic of the game is simple: users watch video and tag the content. Whenever two or more players insert the same tag about the same video in the same time frame (10 seconds, relative to the video), they are both rewarded. The number of matches for a tag is used as an estimate of its trustworthiness. However a tag that is not matched by others, it is not considered to be untrustworthy, because, for instance, it can refer to an element of the video not noticed so far by any user, or it can belong to a niche vocabulary, so it is not necessarily wrong. In the game, when counting matching tags, typos or synonymity are not taken into consideration.

We validate our procedures by using them to estimate the trustworthiness of tag entries produced within the game. Our total corpus contains 37,850 tag entries corresponding to 115 tags randomly chosen. These tag entries correspond to about 9% of the total population. We have checked their representativity of the entire dataset. First, we compared the distribution of each relevant feature that we will use in Section 6.3.2 in our sample with the distribution of the same feature in the entire dataset. A 95% confidence level Chi-squared test [124] confirmed that the hour of the day and the day of the week distribute similarly in our sample and in the entire dataset. The typing duration distributions, instead, are significantly different according to a 95% confidence level Wilcoxon signed-rank test [177]. However, the mode of the two distributions are the same, and the mean differs only 0.1 seconds which, according to the KLM-GOMS model [17], corresponds, at most, to a keystroke. So we conclude that the used sample is representative for the entire data set. A second analysis shows that, by randomly selecting other sets of 115 tags, the corresponding tag entries are not statistically different from the sample that we used. We used 26,495 tag entries (70%) as a training set, and the remaining 11,355 (30%) as a test set.

6.3 Procedures for Trust Estimation

6.3.1 Computing User Reputation

Reputation is an abstraction of a user identity that quantifies his reliability as artifact author. Here, we use it to estimate the trustworthiness of the artifact.

Procedure

We present a generic procedure for computing the reputation of a user with respect to a given artifact produced by him or her.

Algorithm 6.1: Procedure for Reputation Computation

```
1 reputation(user, artifact)
2     evidence ← evidence_selection(user,artifact)
3     weighted_evidence ← weigh_evidence(user,artifact,evidence)
4     reputation ← aggregate_evidence(weighted_evidence)
5     return reputation
```

Evidence Selection Reputation is based on historical evidence, hence the first step is to gather all pieces of evidence regarding a given person and select those relevant for trust computation. Typical constraints include temporal (evidence is only considered within a particular time frame) or semantic traits (evidence is only considered when is semantically related to the given artifact). By *evidence* we denote the set of all evidence regarding *user* about *artifact*.

Algorithm 6.2: Procedure for Evidence Selection

```
1 evidence_selection(user, artifact)
2     for i ← 1 to length(observations) do
3         if observations[i].user = user then
4             evidence.add(observation[i])
5     return evidence
```

Evidence Weighing Given the set of evidence considered, we can decide if and how to weigh its elements, that is, whether to count all the pieces of evidence as equally important, or whether to consider some of them as more relevant. This step might be considered as overlapping with the previous one since they are both about weighing evidence: evidence selection gives a Boolean weight, while here a fuzzy or probabilistic weight is given. However, keeping this division produces an efficiency gain, since it allows computation to be performed only on relevant items.

Algorithm 6.3: Procedure for Weighing Evidence

```
1 weigh_evidence(user, artifact, evidence)
2     for i ← 1 to length(evidence) do
3         weighted_evidence.add(weigh(evidence[i],artifact))
4     return weighted_evidence
```

Aggregate Evidence Once the pieces of evidence (or observations) have been se-

lected and weighed, these are aggregated to provide a value for the user reputation that can be used for evaluation. We can apply several different aggregation functions, depending on the domain. Typical functions are: *count*, *sum*, *average*. Also subjective opinion represent a means to aggregate evidence. We refer the reader to Chapter Preliminaries for an extensive description of this probabilistic logic, which probabilistic reasoning we use in the application of this procedure.

Application Evaluation

First, we convert the number of matches that each tag entry has into trust values:

tag selection For each tag inserted by the user, we select all the matching tags that belong to the same video. In other contexts, the number of matching tags can be substituted by the number of "likes", "retweets", etc.

tag entries weighing For each matching entry, we weigh the entry contribution on the time distance between the evaluated entry and the matched entry. The weight is determined from an exponential probability distribution, which is a "memory-less" probability distribution used to describe the time between events. If two entries are close in time, we consider it highly likely that they match. If they match but appear in distant temporal moments, then we presume they refer to different elements of the same video. Instead of choosing a threshold, we give a probabilistic weight to the matching entry. 85% of probability mass is assigned to tags inserted in a 10-second range.

tag entries aggregation In this step, we determine the trustworthiness of every tag. We aggregate the weighed evidence in a subjective opinion about the tag trustworthiness. We have at our disposal only positive evidence (the number of matching entries). The more evidence we have for the same tag entry, the more certain our estimate of its trustworthiness will be. Non-matched tag entries have equal probability to be correct or not.

Then, we repeat this for each entry created by the user to compute his reputation.

user tag entries selection Select all the tag entries inserted by *user*.

user tag entries weighing Tag entries are weighed by their corresponding trust value previously computed. If an entry is not matched, it is considered as half positive (trust value 0.5) and half negative (1-0.5 = 0.5) evidence (it has 50% probability to be incorrect), as computed by means of subjective opinions. In fact, we use the subjective logic extension for handling partial evidence observation we introduce in Chapter 3 to compute the trust values of the tag entries:

$$tv(tagentry) = \frac{\#m + 1}{\#m + 2}$$

where $\#m$ is the number of matches obtained by *tagentry*.

user tag entries aggregation The overall user reputation is computed by cumulating (by means of the subjective logic cumulative fusion operator [86]) all the subjective opinions about each of the tags he contributed. This opinion represents the user reputation and can be summarized even more by the corresponding expected value or trust value (a particular average over the evidence count).

6.3.2 Computing Provenance-based Trust

We focus on the "how" part of provenance, that is, the modality of production of an artifact. (For simplicity, in the rest of the chapter, we will use the word "provenance" to refer to the "how" part of it). We learn the relationships between PROV and trust values through machine learning algorithms. This procedure allows us to process PROV data and, on the basis of previous trust evaluations, predict the trust level of artifacts. PROV is suitable for modeling the user behavior and provenance information in general.

Provenance Stereotypes

The domain where we situate, that is the video tagging domain, is such that each provenance graph is likely to differ from all the others, at least to some extent. It is indeed difficult that several artifacts have been created at the same time, using the same inputs, in the same amount of time. So, if we considered the provenance graphs as they are, we would have only one piece of evidence per graph while, to estimate its "reputation", we would need several pieces of evidence. It is necessary to apply a reasoning similar to the one we adopted for the users: given a tag, estimate its trustworthiness based on the reputation of its provenance graph.

We tackle this issue by grouping provenance traces so that each group represents coarsely a user behaviour or "stereotype", and by computing the reputations of these groups, rather than focusing on single traces. These groups are identified by extracting a sequence of features from the provenance graphs describing the artifacts which trustworthiness we want to estimate, and by using these features, eventually after having coarsened them or extracted information from them. For instance, from the starting or ending time of an activity, we extract the day of the week and the hour of the day when a given artifact has been created. This allows us to identify patterns in the activities that describe the creation of artifacts. Likewise, we coarsen the duration of the activities, by creating "classes" of durations, instead of considering the actual duration. In the use cases described below, we employ all the provenance information at our disposal for defining user stereotypes. This is possible because the provenance graphs at our disposal are limited in size, but most of all, because the information in the graph is explicitly related to the artifacts which trustworthiness we want to estimate (the inputs from which the artifact is derived, the activity that generate it, etc.). In case the provenance graph is larger, we may define the stereotype based on a selection of provenance features. Since we hypothesize that the reputation of the stereotype is an indicator for the trustworthiness of the artifact, we could create the

stereotype starting by using the features that are directly related to the artifact evaluated and extend our selection when necessary (e.g., to allow creating stereotypes also in graphs without provenance derivations). We will address this issue in future research. Concerning the representation of stereotypes, in the case study presented below, we represent stereotypes by means of sequences of (possibly processed) provenance features, since this representation fits the machine learning approach used to estimate trustworthiness based on them. In Chapter 8 we propose another representation of provenance stereotypes, by means of provenance bundles. Also the representation of provenance stereotypes will be addressed in future research.

Procedure

We present the procedure for computing trust estimates based on provenance.

Algorithm 6.4: Procedure for Making Provenance-based Estimations.

```
1 provenance_estimation(artifact_provenance, artifact)
2     attribute_set ← attribute_selection(artifact_provenance)
3     attributes ←attribute_extraction(attribute_set)
4     trainingset, testset ← trust_levels_aggregation(trainingset,testset)
5     classified_testset ← classify(testset,trainingset)
6     return classified_testset
```

attribute_selection Among all the provenance information, the first step of our procedure chooses the most significant ones: agent, processes, temporal annotations and input artifacts can all hint at the trustworthiness of the output artifact. This selection can lead to an optimization of the computation.

attribute_extraction Some attributes need to be manipulated to be used for our classifications, e.g., temporal attributes may be useful for our estimates because one particular date may be particularly prolific for the trustworthiness of artifacts. However, to ease the recognition of patterns within these provenance data, we extract the day of the week or the hour of the day of production, rather than the precise timestamp. In this way we can distinguish, for instance, between day and night hours (when the user might be less reliable). Similarly, we might refer to process types or patterns instead of specific process instances.

trust_level_aggregation To ease the learning process, we aggregate trust levels in n classes. Hence we apply classification algorithms operating on a nominal scale without compromising accuracy.

classification Machine learning algorithms (or any other kind of classification algorithm) can be adopted at this stage. The choice can be constrained either from the data or by other limitations.

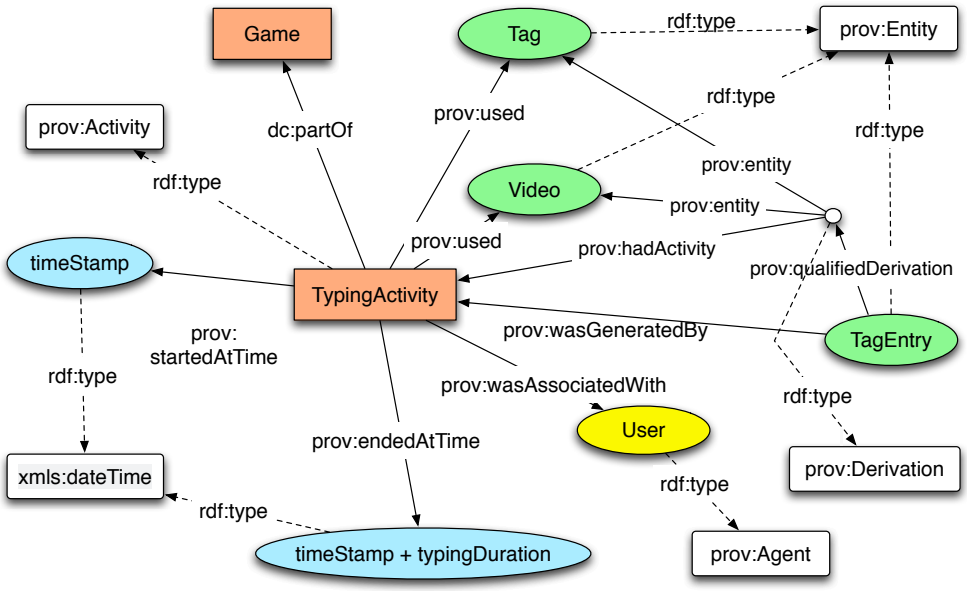


Figure 6.1: Graph representation of the provenance information about each tag entry. A tag entry is derived from the tag and the video to which it is associated. Typing is the activity that produces a tag entry given a tag and a video.

Application Evaluation

We apply the procedure to the tag entries from the *Waisda?* game as follows.

attribute selection and extraction The provenance information available in *Waisda?* is represented in Figure 6.1, using the W3C PROV Ontology. First, for each tag entry we extract: *typing duration*, *day of the week*, *hour of the day*, *game_id* (to which the tag entry belongs), *video_id*. This is the provenance information at our disposal describing how the tag entry has been produced. Here we want to determine the trustworthiness of a tag given the modality with which it was produced (stereotype), rather than the author reputation. Some videos may be easier to annotate than others, or, as we mentioned earlier, user reliability can decrease during the night. For similar reasons we use all the other available features. These features are closely related to the artifact evaluated: they are either entities from which the tag entry has been derived, or the features that describe the activity that generated it. Since we want to identify relations between these features and the tag entry evaluated, these features are particularly relevant. If we had other, less relevant features, we would have selected the most relevant ones. This selection is necessary to avoid the so-called "curse of cardinality", that is the deterioration of performance consequent to

the increase of the availability of features. Provided the availability of a gold standard, this selection could, for instance, be performed by identifying a subset of features that strongly correlates with the entity (tag entry) trustworthiness.

trust level classes computation In our procedure we are not interested in predicting the exact trust value of a tag entry. Rather we want to predict the range of trust within which the entry locates. We start by computing the trust values as we did in the case of user reputations, that is, as:

$$tv(tagentry) = \frac{\#m + 1}{\#m + 2}$$

Then, we split the $[0, 1]$ interval (that is, the trust value range) into 20 classes of length 0.05: from $[0, 0.05[$ to $[0.95, 1]$. This allows us to increase the accuracy of our classification algorithm without compromising the accuracy of the predicted value or the computation cost. The values in each class were approximated by the middle value of the class itself. For instance, the class $[0.5, 0.55]$ are approximated as 0.525. This discretization increases the reported accuracy of the algorithm because by discretizing the interval introduce an approximation of the results: we do not check if a given estimate matches the correct value, but we only check if it falls in the same interval.

regression/classification algorithm We use a regression algorithm to predict the trustworthiness of the tags. Having at our disposal five different features (in principle, we might have more), and given that we are not interested in predicting the "right" trust value but the class of trustworthiness, we adopt the "regression-by-discretization" approach [95], which allows us to use support vector machines algorithm (SVM) [38] to classify our data. The training set is composed by 70% of our data, and we predict the trust level of the test set. We used the SVM version implemented in the e1071 R library [175]. In the future, we will consider alternative learning techniques.

6.3.3 Combining Reputation and Provenance-based Trust

We combine reputation- and provenance-based estimates to improve our estimations. If a certain user has been reliable so far, we can reasonably expect him/her to behave similarly in the near future. However, reputation has an important limitation. To be reliable, a reputation has to be based on a large amount of evidence, which is not always available. So, both in case the reputation is uncertain, or in case the user is anonymous, other sources of information should be used in order to correctly predict a trust value. The trust estimate based on provenance information, as described in Section 6.3.2, is based on behavioral patterns which have a high probability to be shared among many users. Hence, if a reputation is not reliable enough, we substitute it with the provenance-based estimation. In general we prefer reputation- over provenance-based estimates, because they are determined from an analysis of the user behavior and

not derived from the "stereotype" the user belongs to. However, if a reputation is not based on enough evidence (i.e., if the amount of evidence on which it is based is lower than a threshold we set), we prefer to substitute it, since it is highly uncertain and possibly unreliable. In principle, we could also combine the reputation- and the provenance-based estimates. However, assigning a weight to the two estimates may be non-trivial: we could simply average the two values or use the uncertainty of the reputation as a weight (the provenance-based estimate is not expressed as a subjective opinion and hence its uncertainty is not quantified). We will investigate this issue in the future.

Procedure

The algorithm is as follows:

Algorithm 6.5: Procedure for combining reputation- and provenance-based trust

```
1 provenance_reputation_estimation(artifact_provenance,artifact)
2     q_ev ← evaluate_user_evidence(user,artifact)
3     if q_ev > min_evidence then
4         trust_value ← predict_reputation(user,artifact)
5     else
6         trust_value ← predict_provenance(artifact_provenance,artifact)
7     return trust_value
```

evaluate_user_evidence This function quantifies the evidence. Some implementation examples are: (1) *count*; (2) compute a subjective opinion and check if the uncertainty is low enough.

Application Evaluation

We analyze the tags provided by the *Waisda?* platform by making use of the two computations previously performed. In fact, this procedure chooses to adopt the results obtained by the reputation- or be provenance-based procedure, depending on whether the reputation is based on enough evidence. The results are combined as follows: if the reputation is based on a minimum number of observations, then we use it, otherwise we substitute it with the estimation based on provenance. We run this procedure with different values for both the threshold and the minimum number of observations per reputation.

evaluate user tags We instantiate the *evaluate_user_evidence(user,artifact)* function as a *count* function of the evidence of *user* with respect to a given *tag*.

6.4 Results and Discussion

We implement the abstract procedure for reputation computation and we evaluate its performance by measuring its ability to make use of the available evidence to compute the best possible trust assessment. Our evaluation does not focus on the ability to predict the exact trust value of the artifact by computing the user reputation, because these two values belong to a continuous space, and they are computed on a different basis. What we expect is that these two values hint at trustworthiness in a similar fashion: when a tag is trustworthy, then both trust value and reputation should be higher than a certain threshold and vice-versa.

We proceed in the evaluation as follows:

1. We use as gold standard the matches that each tag entry obtained in the *Waisda?* tagging game and we represent them by means of the expected value of subjective opinions, through the formula

$$tv(tagentry) = \frac{\#m + 1}{\#m + 2}$$

   where $\#m$ represents the number of matches. This formula corresponds to the subjective logic extension for handling partial evidence observations introduced in Chapter 3.

2. We split the *Waisda?* dataset into training and test set, where the training set represents the 30% of the whole dataset, and the test set the remaining 70%. We use the training set to train both provenance- and reputation-based models. The reputation-based model uses only part of the training set, as it uses a fixed amount of evidence per user and if a given user produced more than that amount of tag entries in the training set, we do not use them. We set different values for the amount of evidence used to compute user reputations (e.g., 2, 4, 6, 8, 10).

3. We predict the trust value for a tag entry in the test set, by means of the chosen method (reputation, provenance, and their combination).

4. We set a sequence of possible thresholds and, for each of them, we check if the actual and the predicted value are both above or below the threshold.

5. We check the performance of the method by evaluating with how many different thresholds the two values behave in the same manner.

The validation, then, depends upon the choice of the threshold. We run the procedure for computing reputation-based estimates with different thresholds as presented in Figure 6.2. Low thresholds correspond to low accuracy in our estimations. However, as the threshold increases, the accuracy of the estimation rises. Moreover, we should consider that:

1. it is preferable to obtain "false negatives" (reject correct tags) rather than "false positives" (accept wrong tags), so high thresholds are more likely to be chosen (as suggested by the work of Gambetta [59]), in order to reduce risks;

2. a Wilcoxon signed-rank test at 95% confidence level proves that the reputation-based estimates outperform blind guess estimates (having average accuracy 50%). The average improvement is 8%, the maximum improvement is 49%.

In Figure 6.2 we can see also that our method outperforms a blind guess. Unfortunately we have to limit ourselves to this kind of comparison, as we do not have at our disposal more significant data to compare with. We adopt the same procedure for estimating reputation-based assessments to compute the trustworthiness of tags on the Steve.Museum artifacts in Chapter 7.

Also the accuracy of the estimations of the provenance-based procedure depends on the choice of the threshold. If we look at the ability to predict the right (class of) trust values, then the accuracy is of about 32% (which still is twice as much as the average result that we would have with a blind guess), but it is more relevant to focus on the ability to predict the trustworthiness of tags within some range, rather than the exact trust value. Depending on the choice of the threshold, the accuracy in this case varies in the range of 40% - 90%, as we can see in Figure 6.2. For thresholds higher than 0.85 (the most likely choices), the accuracy is at least 70%. We also compared the provenance-based estimates with the reputation-based ones, with a 95% confidence level Wilcoxon signed-rank test that proves that the estimates of the two algorithms are not statistically different. *For the Waisda? case study, reputation- and provenance-based estimates are equivalent: when reputation is not available or it is not possible to compute it, we can substitute it with provenance-based estimates.* This is particularly important since the ever growing availability of PROV data will increase the ease for computing less uncertain trust values.

Since we apply the "regression-by-discretization" approach for making provenance-based assessments, we approximate our trust values. This is necessary because we use support vector machines for their ability to learn a reliable model from the provenance features we use, however support vector machines are classification models. Since we use them to make our estimations, we need to define trust classes, and these are exactly the twenty classes defined by splitting the $[0, 1]$, interval in twenty subintervals. In the reputation approach we do not employ a classification algorithm, thus it is not necessary to discretize the interval in that case. Had we applied the same approximation to the reputations as well, then provenance-based trust would have performed better, as proven with a 95% confidence level Wilcoxon signed-ranked test we run, because reputation can rely only on evidence regarding the user, while provenance-based models can rely on larger data sets. Anyway, we have no need to discretize the reputation from an accuracy point of view but, in general, we prefer it for its lightweight computational burden.

Finally, the performance of the algorithm that combines reputation- and provenance-based estimates depends both on the choice of the threshold for the decision and on the number of pieces of evidence that make a reputation reliable, so we run the

algorithm with several combinations of these two parameters (Figure 6.2). The results converge immediately, after having set the minimum number of observations at two. We compare these results with those obtained before. Two Wilcoxon signed-rank tests (at 90% and 95% confidence level about respectively reputation- and provenance-based assessments) show that *the procedure which combines reputation and provenance evaluations in this case performs better than each of them applied alone.* The improvement is, on average, about 5%. Although most of the improvement regards the lower thresholds, which are less likely to be chosen (as we explain in Subsection 6.3.1), even at 0.85 threshold there is a 0.5% improvement (which is small, but still, present). Moreover, we would like to stress how the combination of the two procedures performs better than (in a few cases, equal to) each of them applied alone, regardless of the threshold chosen.

Combining the two procedures allows us to go beyond the limitation of reputation-based approaches. Substituting estimates based on poorly reliable reputations with provenance-based ones improves our results without significantly increasing our risks, since we have previously proven that the two estimates are (on average) equivalent. This explains the similarity among the results shown in Figure 6.2. Hence, when a user is new in a system (and so his/her history is limited) or anonymous, we can refer to the provenance-based estimate to determine the trustworthiness of his/her work, without running higher risks. We proved the existence of a little positive correlation (0.16) between the reputation- and provenance-based trust assessments, by using a Pearson's correlation test [123] with a confidence level of 99%. This tells us that in some cases, reputation- and provenance-based estimates behave alike. If the correlation was higher, we would not have needed to refer to provenance-based estimates, because provenance-based estimates would not have added much information. Moreover, the reputation-based estimates that are most likely not to correlate with the provenance-based estimates are those based on less evidence, since those are heavily affected by smoothing. Those are exactly the estimates that we substitute with provenance-based estimates.

6.5 Conclusion

This chapter explores two important components of trust assessments: reputation and provenance information. We propose and evaluate a procedure for computing reputation and one for computing trust assessments based on provenance information represented with the W3C standard PROV Ontology. We show that it is important to use reputation estimation for trust assessment, because it is simple and accurate. We also show the potential of provenance-based trust assessments: these can be at least as accurate as reputation-based ones and can be used to overcome the limitations of a reputation-based approach. In *Waisda?* the combination of the two methods revealed to be more powerful than each of the two alone.

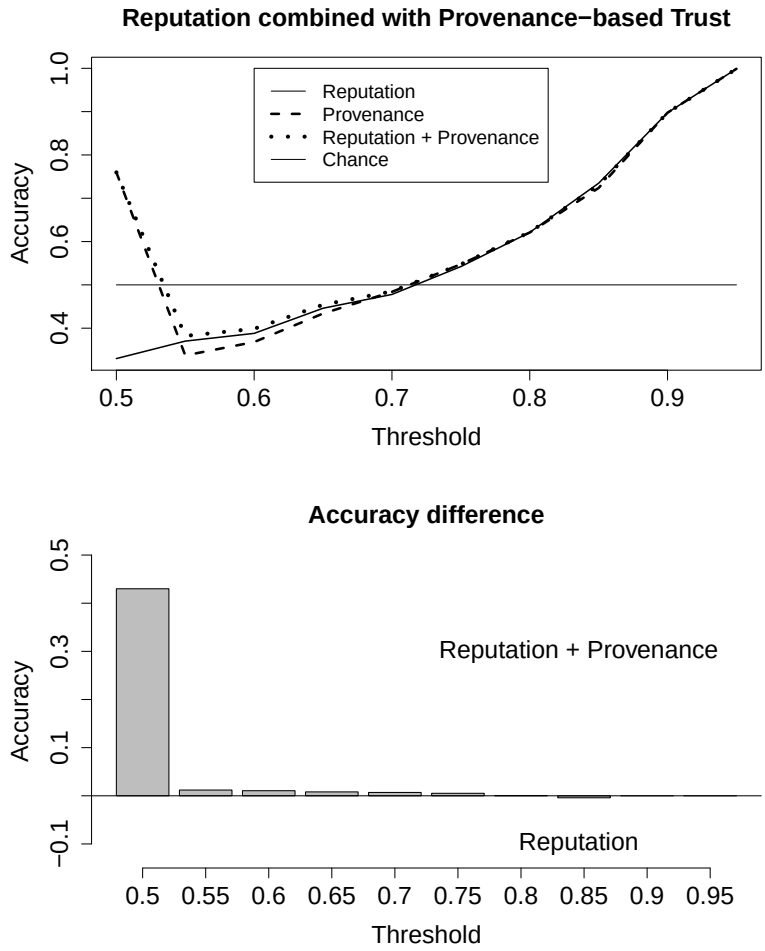


Figure 6.2: Absolute and relative (Reputation+Provenance vs. Reputation) accuracy. The gap between the estimation (provenance-based) and the real value of some items explains the shape between 0.5 and 0.55: only very low or high thresholds cover it.

Part IV

Semantic Similarity to Improve Trust Estimation

*This last part has a twofold goal. The main topic of this part is the use of semantic similarity measures to support uncertainty reasoning for making trust estimations. In other words, here are collected the chapters that aim to tackle the fourth research question (*Can semantic similarity measures improve the accuracy of trust estimates of semi-structured data based on uncertainty reasoning?*). Moreover, this part, being the last part of the thesis, presents two works that touch all the themes covered throughout the thesis. Of course these two chapters can not satisfactorily address all the issues addressed before, nevertheless these are built on top of all the previous research. Chapter 7 addresses the research question by proposing an extension of the models presented in Chapters 1 and 6, by combining user reputation built using subjective logic with semantic similarity measures. Chapter 8 extends this model and bases the computation on the reputation of so-called "provenance stereotypes".*

7

Assessing Annotation Trustworthiness Using Semantic Similarity

Two Cultural Heritage Case Studies

*This chapter addresses the last research question described in Chapter Introduction (*Can semantic similarity measures improve the accuracy of trust estimates of semi-structured data based on uncertainty reasoning?*), so it shows an exploration of the use of semantic similarity measures to improve the trustworthiness estimation of semi-structured Web data and to better manage the evidence at our disposal for making trust inference. The targets of our assessments are crowdsourced cultural heritage annotations. We use subjective logic to reason upon the evidence at our disposal for making trust estimates, like in Chapters 1 and 6, but the probabilistic logic is combined with semantic similarity measures (thanks to the extensions introduced in Chapter 3) to improve the accuracy of the predictions and to limit the need to set arbitrary thresholds in our algorithms, that is a critical point of the algorithms proposed in Chapter 6. Furthermore, we show that semantic similarity measures can not only be effectively used to improve the accuracy of trust predictions, but also to improve their computational time. Finally, we also show an alternative, interactive version of the algorithm, with corresponding annotation representation.*

This chapter is based on the paper Automated Evaluation of Annotators for Museum Collections using Subjective Logic *coauthored with Archana Nottamkandath and Wan Fokkink and presented at the* 6th IFIP Trust Management Conference (IFIPTM 2012) *in Surat, India and on the paper* Semi-Automated Assessment of Annotation Trustworthiness *coauthored with Archana Notthamkandath and Wan Fokkink and presented at the* 11th Privacy, Security and Trust Conference (PST 2013) *in Tarragona, Spain. The latter paper was awarded as "Best Student Paper ex-aequo". My contribution to these papers regards the data representation, the design of the*

algorithms, the preliminary statistical evaluation over the Steve.Museum dataset and the two evaluations of the algorithm for semi-automated assessment of annotation trustworthiness.

7.1 Introduction

The goal of the work described in this chapter is to show how it is possible to semi-automate in an optimized way the process of evaluation of crowdsourced annotations (we could also use the term "tag" in place of "annotation", since here we deal with non-hierarchical terms assigned to cultural heritage artifacts, but for the sake of uniformity, we prefer the term "annotation"). This is done by first collecting manual evaluations about the quality of a small part of the annotations contributed by a user and then learning a statistical model from them. On the basis of such a model, the system automatically evaluates the annotations further added by the same user. We employ Semantic Web technologies to represent and store the annotations and the corresponding reviews, as in Chapter 6. Like in several chapters so far we use subjective logic (see Chapter Preliminaries) to build a reputation for users who contribute to the system, and moreover semantic similarity measures to generate assessments on the annotations entered by the same users at a later point in time. This approach is an extension, in particular, of the works presented in Chapters 1 and 6. In order to improve the computation time, we cluster the evaluated annotations to reduce the number of comparisons, and our experiments show that this preprocessing does not seriously affect the accuracy of the predictions. The proposed algorithms are evaluated on two datasets from the cultural heritage domain. In our experiments we show that it is possible to semi-automatically evaluate the annotations entered by users in crowdsourcing systems into binomial categories (good, bad) with a level of accuracy above 80%.

The novelty of this research lies in the automation of annotation evaluations on crowdsourcing systems by coupling subjective logic opinions with measures of semantic similarity (as described in Chapter 3). The sole variable parameter that we require is the size of the set of manual evaluations that are needed to build a useful and reliable reputation. Moreover, our experiments show that varying this parameter does not substantially affect the performance (resulting in about 1% precision variation per five new observations considered in a user reputation). Using our algorithms, we show how it is possible to avoid asking the system administrators to set a threshold in order to make assessments about the trustworthiness of annotations (e.g., accept only annotations which have a trust value above a given threshold).

The rest of the chapter is structured as follows: Section 7.2 presents related work; Section 7.3 describes the framework that we propose; Section 7.4 provides two different case studies where the system has been evaluated and Section 7.5 discusses them. Section 7.6 describes an interactive version of the model, and finally Section 7.7 provides conclusions.

7.2 Related Work

In this chapter we make use of subjective logic and semantic similarity for trust estimation. We refer the reader to Chapter Preliminaries for a detailed description to both techniques and to Chapter Introduction for an extensive literature review about them. We also use semantic similarity measures to cluster related annotations to optimize the computations. In the work of Cilibrasi et al. [33] hierarchical clustering has been used for grouping related topics, while Ushioda et al. [156] experiment with clustering words in a hierarchical manner. Begelman et al. [8] present an algorithm for the automated clustering of annotations on the basis of annotation co-occurrences in order to facilitate more effective retrieval. A similar approach is used by Hassan-Montero and Herrero-Solana [74]. They compute annotation similarities using the Jaccard similarity coefficient and then cluster the annotations hierarchically using the k-means algorithm. In our work, to build the user reputation, we cluster the annotations contributed by the users, along with their respective evaluations (e.g., accept or reject). Each cluster is represented by a medoid (that is, the element of the cluster which is the closest to its center), and in order to evaluate a newly entered annotation by the same user, we consider clusters which are most semantically relevant to the new annotation. This helps in selectively weighing only the relevant evidence about a user for evaluating a new annotation.

7.3 System Description

7.3.1 High-level Overview

The system that we propose aims at relieving the institution personnel (reviewers in particular) from the burden of controlling and evaluating all the annotations inserted by users. The system asks for some interaction with the reviewers, but tries to minimize it. Figure 7.1 shows a high-level view of the model. For each user, the system asks the reviewers to review a fixed number of annotations, and on the basis of these reviews it builds user reputations. A reputation is meant to express a global measure of trustworthiness of the corresponding user as artifact creator. Here, "global" indicates the fact that the reputation is not intended to measure the user expertise about different annotation subjects. Rather, it expresses the trustworthiness of the user as an annotator. So, for instance, a highly reputed user should provide high quality annotations about subjects in his area of expertise. The reviews are also used to assess the trustworthiness of each annotation inserted afterwards by a user: given an annotation, the system evaluates it by looking at the evaluations already available. The evaluations of the annotations semantically closer to the one that we evaluate have a higher impact. So we have two distinct phases: a first training step where we collect samples of manual reviews, and a second step where we make automatic assessments of the trustworthiness of annotations (possibly after having clustered the evaluated annotations to improve the computation time). The more reviews there are,

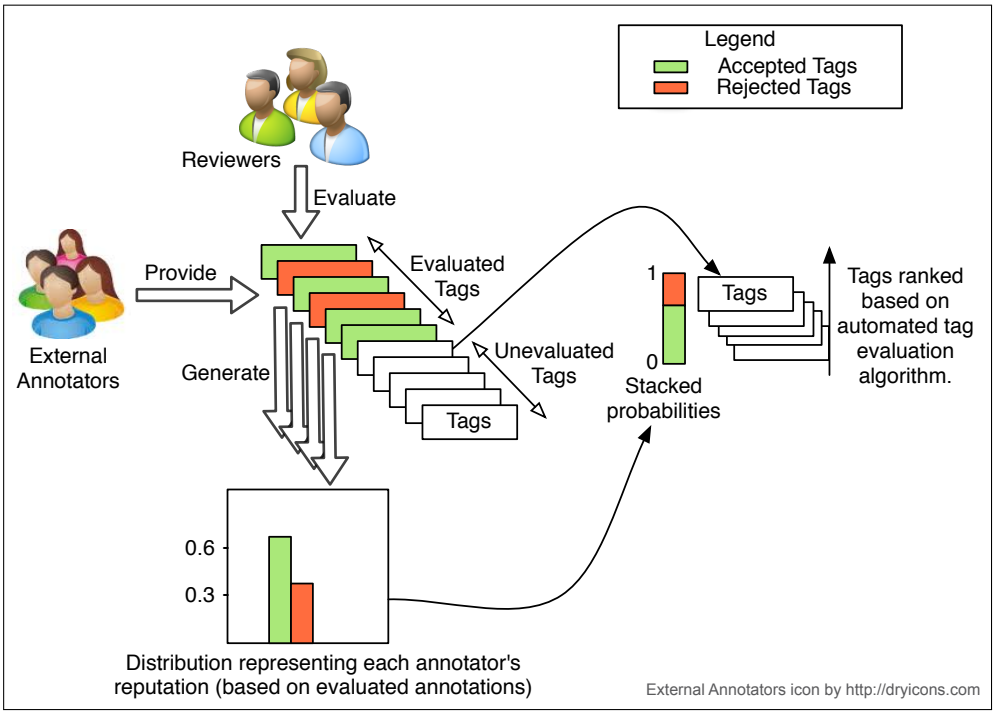


Figure 7.1: High-level overview of the system.

the more reliable the reputation is, but this number depends also on the workforce at the disposal of the institution. On the other hand, as we will see in Section 7.4, this parameter does not affect significantly the accuracy obtained. Moreover, we do not need to set an "acceptance threshold" (e.g., accept only annotations with a trust value of say at least 0.9, for trust values ranging from zero to one), despite the work presented in Chapter 6. This is important, since such a threshold is arbitrary, and it is not trivial to find a balance between the risk to accept wrong annotations and to reject good ones.

Suppose that a user, Alex (whose profile already contains three annotations which were evaluated by the museum), newly contributes to the collection of a museum by annotating five artifacts. Alex annotates one artifact with "Chinese". If the museum immediately uses the annotation for classifying the artifact, it might be risky because the annotation might be wrong (maliciously or not). On the other hand, if the museum had enough employees to check the external contributed annotation, then it would not have needed to crowdsource it. The system that we propose relies on few evaluations of Alex's annotations by the museum. Based on these evaluations, the system: (1) computes Alex's reputation; (2) computes a trust value for the new annotation; and (3) decides whether to accept it or not. We describe the system in the following sections.

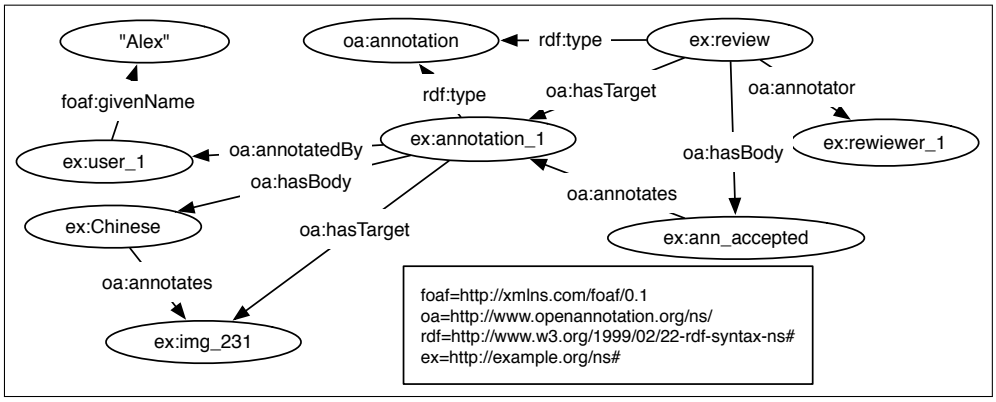


Figure 7.2: We represent annotations and their reviews using the Annotation class from the Open Annotation model.

7.3.2 Annotation Representation

We use the Open Annotation model [14] as a standard model for describing annotations, together with the most relevant related metadata (e.g., the author and the time of creation). The Open Annotation model allows us to reify the annotation itself, and by treating it as an object, we can easily link to its properties like the annotator URI or the time of creation. Moreover, the review of an annotation can be represented as an annotation which target is an annotation and which body contains a value of the review about the annotation. To continue with our example, Figure 7.2 and Listing 7.1 show an example of an annotation and a corresponding review, both represented as "annotations" from the Open Annotation model.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix oa: <http://www.w3.org/ns/openannotation/core/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix ex: <http://example.org/ns#> .
ex:user_1 foaf:givenName "Alex" .
ex:annotation_1 oa:hasBody ex:Chinese;
                oa:annotatedBy ex:user_1;
                oa:hasTarget ex:img_231;
                rdf:type oa:annotation .
ex:review oa:hasBody ex:ann_accepted;
        oa:annotatedBy ex:reviewer_1;
        oa:hasTarget ex:annotation_1;
        rdf:type oa:annotation .
ex:annotation_accepted oa:annotates ex:annotation_1 .
```

Listing 7.1: Example of an annotation and respective evaluation.

7.3.3 Trust Management

We employ subjective logic for representing, computing and reasoning on trust assessments. Chapter Preliminaries provides a description of this logic.

Trust is context-dependent (see Chapter Introduction), since different users or annotations (or, more in general, agents and artifacts) might receive different trust evaluations, depending on the context from which they originate and the reviewer. In our scenarios we do not have at our disposal an explicit description of trust policies by the museums. Also, we do not aim at determining a generic annotation (or user) trust level. Our goal is to learn a model that evaluates annotations as closely as possible to what that museum would do, based on a small sample of evaluations produced by the museum itself.

User Reputation Computation and Representation

We define a user reputation as a global value representing the user's ability to annotate according to the museum policy. Global, since we do not relate the user reputation to a specific context, because this value should represent an overall trust level about the user production: a highly reputed user is believed to have the ability to produce high-quality annotations and to choose annotations/artifacts related to his/her domain of expertise. Also, the possible number of topics is so high that defining the reputation to be topic-dependent would bring manageability issues. Expertise will be considered when evaluating a single annotation, as we will see later in this section.We require that a fixed amount of user-contributed annotations is evaluated by the museum. Based on those evaluations we compute the user reputation using subjective opinions.

To continue with the previous example, suppose that Alex contributed three annotations: {Indian, Buddhist} where evaluated as accepted and {tulip} as rejected. His reputation is:

$$\omega_{Alex}^{museum} = \left(\frac{2}{5}, \frac{1}{5}, \frac{2}{5}, \frac{1}{2}\right) \quad E(\omega_{Alex}^{museum}) = 0.6 \tag{7.1}$$

Annotation Trust Value Computation and Representation

Annotation trust values are represented by means of subjective opinions, as in Chapter Preliminaries.

Here, we still use the annotations created by the user and the corresponding evaluations to compute the trust value, but despite the computation of the user reputation, evidence are weighed with respect to the similarity to the annotation to be evaluated. So p and n (the counts of positive and negative pieces of evidence respectively) are determined as in Equation 7.2, where sim is a semantic relatedness measure, t is an annotation to be evaluated, m represents the museum and $train$ is the set of evaluated annotations (training set). Here, we still use the tags created by the user and the corresponding evaluations to compute the trust value, but despite the computation of the user reputation (see Chapter Preliminaries and Equation 7.1), evidence is weighed

with respect to the similarity to the tag to be evaluated before being used.

$$\begin{aligned} p_t^m &= \Sigma_{t_i \in train} sim(t, t_i) \ \ if evaluation(t_i) = true \\ n_t^m &= \Sigma_{t_i \in train} sim(t, t_i) \ \ if evaluation(t_i) = false \end{aligned} \tag{7.2}$$

The annotation "Chinese" inserted by Alex is evaluated as:

$$p^m_{Chinese} = sim(\text{Chinese}, \text{Indian}) + sim(\text{Chinese}, \text{Buddhist}) = 1.05 \tag{7.3}$$

$$n^m_{Chinese} = sim(\text{Chinese}, \text{tulip}) = 0.1 \tag{7.4}$$

$$\omega^m_{Chinese} \left(\tfrac{1.05}{1.05+0.1+2}, \tfrac{0.1}{1.05+0.1+2}, \tfrac{2}{1.05+0.1+2}, \tfrac{1}{2} \right) \tag{7.5}$$

$$E(\omega^m_{Chinese}) = 0.95 \tag{7.6}$$

Annotation Evaluation

In order to evaluate annotations (i.e., decide to accept or reject them), we define an ordering function on the set of annotations based on their trust values (see Equation 7.7). The ordered set of annotations is represented as $\{t\}_1^{|annotations|}$, where $|annotations|$ is the cardinality of the set of annotations. For annotations t_1 and t_2,

$$t_1 \leq t_2 \iff E(\omega^m_{t_1}) \leq E(\omega^m_{t_2}) \tag{7.7}$$

Recall that $E(\omega^m_u)$ is the user reputation, that is, the expected percentage of correct annotations created by the user. Hence, we accept the last $E(\omega^m_u) \cdot |annotations|$ annotations in $\{t\}_1^{|annotations|}$ (see Equation 7.8) ($\{t\}_1^{|annotations|}$ is in increasing order, so accept the tags having higher trust value).

$$evaluation(annotation) = \begin{cases} rejected & \text{if } t \in \{t\}^1_{E(\omega^m_u) \cdot |annotations|} \\ accepted & \text{otherwise} \end{cases} \tag{7.8}$$

We saw how the reputation of Alex was 0.6. He inserted five new annotations, so $0.6 \cdot 5 = 3$ will be accepted. The annotation "Chinese" had a trust value of 0.95, which ranks it as first in the ordered list of annotations. Therefore, the annotation "Chinese" is *accepted*.

7.3.4 Algorithm

We provide here a pseudocode representation of the algorithm that implements the annotation evaluation procedures, and we explain it in detail.

build_user_reputation Builds a reputation for each user in the training set using subjective logic, so to obtain Equation 7.1. A reputation is represented as a vector of probabilities for possible annotation evaluations.

trust_values Trust values are represented as vectors of probabilities of possible annotation evaluations, following Equation (28) in Chapter Preliminaries.

Algorithm 7.1: Algorithm to compute trust values of annotations

Input: Two finite sets, $Training\_set = \{\langle annotation, evaluation, UserID\rangle\}$ and $Test\_set = \{\langle annotation, UserID\rangle\}$

Output: A finite set $Result\_Test\_set = \{\langle annotation, trust\_values\rangle\}$

```
1  for UserID ← UserID_1 to UserID_n do
2      forall the annotations in Training_set do
3          rep[UserID] ← build_reputation(Training_set)
4  for UserID ← UserID_1 to UserID_n do
5      forall the users in Test_set do
6          for Annotation ← annotation_1 to annotation_n do
7              forall the annotations in Test_set do
8                  trust_values[Annotation] = comp_tv(Training_set)
9              s_annotations ← sort_annotations(trust_values)
10             Result ← assess(s_annotations, rep[UserID])
11 return Result
```

comp_tv Implements Equations 7.3 and 7.4 to compute the annotation trust value represented in Equation 7.5. The value actually stored is the expected value of the opinion, that is

$$E(\omega_t^m) = \frac{p_t^m}{p_t^m + n_t^m + 2} + \frac{2}{p_t^m + n_t^m + 2} \cdot \frac{1}{2}$$

.

sort_annotations The annotations are sorted according to their trust value, following the ordering function in Equation 7.7.

assess The assess function assigns an evaluation to the annotation, by implementing Equation 7.8.

7.3.5 Clustering Semantically Related Annotations

Reputations built using large training sets are likely to be more accurate than those built using smaller ones. On the other hand, the larger the set of annotations used for building the reputation, the higher the number of comparisons we will have to make to evaluate a new annotation. In order to reduce this tension, we cluster the annotations in the training set per user on a semantic basis, and for each resulting cluster we compute the medoid (that is, the element of the cluster which is, on average, the closest to the other elements), and record the evidence counts. The clustering allows us to reduce the computation time of the trust assessments, and the fact that clustering is performed on a semantic basis, that is, annotations are clustered in order to create

subsets of annotations having similar meanings, aims at preserving the accuracy of the estimation algorithm. We check this in our evaluation.

After having clustered the annotations, we adapt the algorithm so that we compute a subjective opinion per cluster, but we weigh it only on the semantic distance between the new annotation and the cluster medoid. In this way we reduce the number of comparisons (we do not measure the distance between the new annotation and each element of the cluster), but we still account for the size of the training set, as we record the evidence counts of it. We use hierarchical clustering [68] for semantically clustering the words, although it is computationally expensive, because: (1) we know only the relative distances between words, and not their position in a simplex (the semantic distance is computed as $1 - \mathit{similarity}(\mathit{word}_1, \mathit{word}_2)$), and this is one of the algorithms that requires such kind of input; (2) it requires only one input argument, a real number "cut", that determines the number of clusters of the input set S of words: if cut=0, then there is only one cluster, if cut=1, then there are n clusters, where n is the cardinality of S. Clustering is performed offline, before any annotation is evaluated, and here we focus on the improvement of the performance of the newly introduced annotations. Algorithm 7.2 incorporates these optimizations.

To continue with the previous example, the museum can cluster the annotations inserted by Alex before making any estimate. We have only three annotations in the training set, which result in two clusters, {Indian, Buddhist} and {tulip}.

$$p^m_{Chinese} = \mathit{sim}(\text{Chinese}, \text{Indian}) \cdot 2 = 1.75$$

$$n^m_{Chinese} = \mathit{sim}(\text{Chinese}, \text{tulip}) = 0.1$$

$$\omega^m_{Chinese}\left(\frac{1.75}{1.75+0.1+2}, \frac{0.1}{1.75+0.1+2}, \frac{2}{1.75+0.1+2}, \frac{1}{2}\right)$$

$$E(\omega^m_{Chinese}) = 0.72$$

This result is different from the previous trust value computed in a non-clustered manner (0.95). However, this variation is likely to affect all the computed trust values. In the two case studies presented in Section 7.4, the accuracy of the algorithm is not affected by semantically clustering the training set, and the computational time is reduced.

7.3.6 Implementation

The code for the representation and assessment of the annotations with the Open Annotation model has been developed using SWI-Prolog Semantic Web Library [174] and the Python libraries rdflib [133] and hcluster [48], and is available on the Web.[1]

[1] Available at `http://trustingwebdata.org/phdthesis/dceolin`.

Algorithm 7.2: Algorithm to compute annotation trust with clustering.

Input: Two finite sets, $Training\_set = \{\langle annotation, evaluation, UserID\rangle\}$ and $Test\_set = \{\langle annotation, UserID\rangle\}$
Output: A finite set $Result\_Test\_set = \{\langle annotation, trust\_values\rangle\}$

```
for UserID ← UserID_1 to UserID_n do
    for all annotations in Training_set do
        rep[UserID] ← build_reputation(training_set)
        clusters[UserID] ← build_clust(training_set)
        medoids[UserID] ← get_med(clusters,UserID)
for UserID ← UserID_1 to UserID_n do
    for all users in Test_set do
        for Annotation ← annotation_1 to annotation_n do
            for all annotations in Test_set do
                trust_values[Annotation] =
                comp_tv(medoids[UserID],rep[UserID])
        sort_annotations ← sort(trust_values)
        Result ← assess(sort_annotations,rep[UserID])
return Result
```

7.4 Evaluation

7.4.1 Datasets

We introduce the two datasets that are used in this evaluation, namely the Steve.Museum dataset and a dataset from an experiment of the SEALINC Media project.

Steve.Museum Dataset

Steve.Museum is a project involving several museum professionals in the cultural heritage domain. Part of the project focuses on understanding the various effects of crowdsourcing cultural heritage artifact annotations. Their experiments involved external annotators annotating musea collections, and a subset of the data collected from the crowd was evaluated for trustworthiness. In total, 4,588 users annotated the 89,671 artifacts using 480,617 annotations from 21 participating museums. Part of these annotations consisting of 45,860 annotations were manually evaluated by professionals at these museums and were used as a basis for our second case study. In this project, the annotations were classified in a more refined way, compared to the previous case study, namely as:

- Todo
- Judgement-negative, Judgement-positive

- Problematic-foreign, Problematic-misperception, Problematic-misspelling, Problematic-no_consensus, Problematic-personal, Problematic-huh
- Usefulness-not_useful, Usefulness-useful

There are three main categories: judgement (a personal judgement by the annotator about the picture), problematic (for several, different reasons) and usefulness (stating whether the annotation is useful or not). We consider only "usefulness-useful" as a positive judgement, all the others are considered as negative evaluations. The annotations classified as "todo" are discarded, since their evaluation has not been performed, yet.

SEALINC Media Project Experiment Dataset

As part of the SEALINC Media project, Rijksmuseum is crowdsourcing annotations of artifacts in its collection using Web users. An initial experiment was conducted to study the effect of presenting pre-set annotations on the quality of annotations on crowdsourced data [102]. In the experiment, the external annotators were presented with pictures from the Web and prints from the Rijksmuseum collection along with a pre-set annotations about the picture or print, and they were asked to insert new annotations, or remove the pre-set ones which they did not agree with. A total of 2,650 annotations resulted from the experiment, and these were manually evaluated by trusted personnel for their quality and relevance using the following scale:

- 1 : Irrelevant
- 2 : Incorrect
- 3 : Subjective
- 4 : Correct and possibly relevant
- 5 : Correct and highly relevant
- typo : Spelling mistake

These annotations, along with their evaluations, were used to validate our model. We neglect the annotations evaluated as "Typo" because our focus is on the semantic correctness of the annotations, so we assume that such a category of mistakes would be properly avoided or treated (e.g., by using autocompletion and checking the presence of the annotations in dictionaries) before the annotations reach our evaluation framework.

This section provides three evaluations of our method. We start by introducing a first preliminary evaluation about the effectiveness of combining subjective logic with a semantic relatedness measure. We test this combination over the Steve.Museum project [155] dataset. Then we test the overall algorithm that we propose over the SEALINC Media project dataset and, again, over the Steve.Museum dataset.

Topic1	Topic2	SemDist
Asian	Chinese	0.9333
Asian	Buddhist	0.7143
Chinese	Buddhist	0.6667

Table 7.1: Semantic relatedness between words in Cluster1 and Cluster2

Topic1	Topic2	SemDist
Piano	Instrument	0.9
Piano	Music	0.4286
Piano	String	0.9091
Instrument	Music	0.7059
Instrument	String	0.9
Music	String	0.5

Table 7.2: Semantic relatedness between words in Cluster2.

7.4.2 Preliminary Evaluation: Steve.Museum Dataset

We provide here a first limited evaluation that motivates the use of subjective logic in combination with a semantic relatedness measure as a mean to effectively use assessments of annotations as evidence for trustworthiness prediction. Here we focus on the Steve.Museum [155] dataset, for validating our proposed approach. For this experiment, we compute the semantic relatedness by using the Wu & Palmer measure (see Chapter Preliminaries) on WordNet using an online service [127]. This gives us a $measure \in]0..1]$.

A first empirical overview of the dataset hints at the presence of possible semantic clusters. We then manually select the candidate set of single words and prove that the semantic relatedness among those words is high. An example of clusters found is available in Figure 7.3. After having shown the existence of these semantic clusters, we compare the expertise of people using words from those clusters and notice that people having a high amount of positive (or negative) evidence regarding one word in a particular cluster also have a high amount of positive (or negative) evidence about the other words in the same cluster. Positive and negative evidence is derived from the evaluation by the museum: annotations evaluated as useful are counted as positive evidence, non-useful as negative. This manual and empirical analysis gives us a first concrete indication about the relatedness between reputation based on evidence and semantic similarity.

We also build each user's reputation using a subset of the evaluations made by the museum and, based on this, we predict the usefulness of future annotations inserted by each user[2]. Annotations having a trust level of at least 0.7 are labelled as "useful". As a side effect of weighing, *uncertainty* of reputations rises, since weighing reduces

[2]The complete set of analyses is available at: `http://trustingwebdata.org/phdthesis/dceolin`

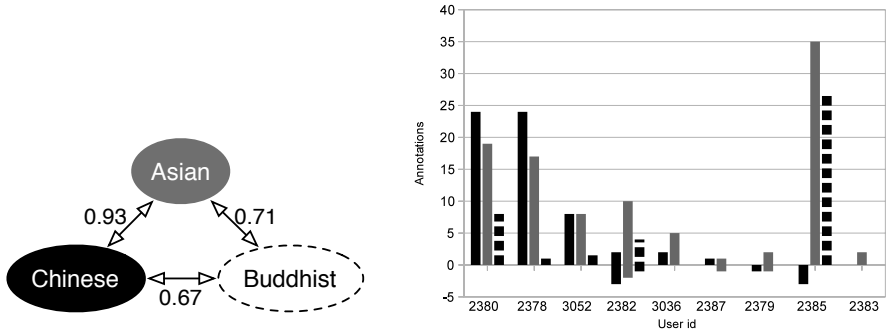


Figure 7.3: Cluster and corresponding positive/negative evidence per user.

the amount of evidence considered. However, often, this consequence does not worsen our results, especially when the reputation is already quite high (e.g., the reputation of an annotator reduced to 0.97 from 0.72). On the contrary, our approach allows us to be prudent in our evaluations, so we could avoid accepting as useful annotations with high *uncertainty*. Weighing improves the accuracy of subjective logic in a statistically significant manner, as proven by applying the sign test with a confidence interval of 95% on the compared errors. This is due to the fact that semantic similarity allows us to weigh more the semantically relevant pieces of evidence, hence making their contribution prevail in the computation. If, instead, like we do in the following case studies, semantic similarity is employed to cluster the annotations to gain computational efficiency, then this may have a negative impact on the accuracy of the computation. However, the fact that the clustering is made on semantic similarity-basis is likely to limit such an impact. To test these hypotheses, we evaluate our model against two aforementioned datasets from cultural heritage crowdsourcing projects.

7.4.3 Case Study 1: Dataset from a SEALINC Media Project Experiment

We use here the dataset from the SEALINC Media project experiment described above. We build our training set using a fixed amount of evaluated annotations for each of the users, and form the test set using the remaining annotations. The number of annotations used to build the reputation and the percentage of the dataset covered is presented in Table 7.3. The behavior of an annotator is classified as either correct or wrong, based on the positive and negative evidence available. The positive evidence is constituted by the annotations classified as category 4 and 5, while the negative evidence comprises annotations from category 1, 2 and 3. We run the previously described algorithm for different numbers of annotations used as a basis for building user reputations, in order to analyze the impact of different sizes of training sets. The results of the experiment are reported in Table 7.3, where correct annotations are

considered as a target to be retrieved, so that we can compute metrics such as precision, recall and F-score. This first case study provides us with interesting insights about the model that we propose. The evaluation shows positive results, with a level of accuracy higher than 80% and a level of recall higher than 85%. Clustering brings a clear reduction of the computation time without compromising accuracy (with two different values for the cut parameters, chosen to split almost evenly the $[0, 1]$ interval). The shape of the dataset and the high variance for measurements of small execution times determine a non-linear pattern in the execution times. An important consideration regards the fact that some errors can be due to intrinsic limitations of the experiment rather than the imprecision of the algorithm. For instance, since training and test set are part of the same dataset, the bigger the training set is, the smaller the test set is. Since our prediction is probabilistic, a small training set forces us to discretize our predictions, and this increases our error rate. Also, while an increase of the number of annotations used for building a reputation produces an increase of the reliability of the reputation itself, such an increase has the downside to reduce our test set size, since only few annotators produced a large number of annotations. It is important to stress that, on the one hand, the increase of the size of the training set brings an improvement of the performance, and on the other hand, performance is already satisfactory with a small training set (five observations per user). Also, this improvement is small. This is important because: (1) the sole parameter that we did not set (i.e., size of the training set) does not seriously affect our results; and (2) when the size of the training set is small, the performance is relatively high, so the need of manual evaluation is reduced. The results are satisfactory even with a small training set, also thanks to the smoothing factor of subjective logic, that allows us to compensate for the possibly limited representativity (with respect to the population) of a distribution estimated from a small sample. In some cases, the performance obtained after having clustered the training set is even slightly higher than that obtained without clustering. This could be due, for instance, to a peculiarity of the dataset or to the fact that by clustering and comparing each annotation to be evaluated only against each cluster medoid, we distribute the weights among the evidence in a more accurate way than by comparing it with each piece of evidence alone. In Subsection 7.4.4 we present another case study that may provide useful hints for this issue as well.

7.4.4 Case Study 2: Steve.Museum Project Dataset

We partition the Steve.Museum dataset into a training and a test set, as shown in Table 7.4, along with their percentage coverage of the whole dataset, together with the results obtained. This second case study focuses on a larger dataset than the first one. The average accuracy attests around 70%. This shows that our algorithm can be trained to different museum policies, because the accuracy, although lower than before, can still be considered satisfactory. The decrease in accuracy with respect to the previous case is possibly due to the different annotation distribution (of positives and negatives) of the dataset and different domains. Different distributions might make it harder to discriminate between trustworthy and non-trustworthy annotations

Annotations per reputation	Training set coverage	Accuracy	Precision	Recall	F-score	Time (sec.)
non-clustered results						
5	8%	0.73	0.88	0.81	0.84	87
10	19%	0.76	0.87	0.84	0.86	139
15	31%	0.76	0.86	0.86	0.86	221
20	41%	0.84	0.87	0.96	0.86	225
clustered results (cut=0.6)						
5	8%	0.73	0.88	0.81	0.84	43
10	19%	0.82	0.87	0.93	0.90	24
15	31%	0.83	0.87	0.95	0.91	14
20	41%	0.84	0.87	0.96	0.91	18
clustered results (cut=0.3)						
5	8%	0.78	0.88	0.88	0.88	43
10	19%	0.82	0.87	0.93	0.90	14
15	31%	0.84	0.87	0.95	0.91	16
20	41%	0.84	0.87	0.96	0.92	21

Table 7.3: Performance on the data from the SEALINC Media project experiment.

(as one might encounter mostly one type of observations). Different domains might lead to a different variability of the topics of the annotations. This fact affects the reliability of clusters computed on a semantic basis (since clusters will tend to contain less uniform annotations, and medoids will be, on average, less representative of their corresponding clusters), and consequently affects the accuracy of the algorithm. Moreover, one underlying assumption of the algorithm is the existence of a correlation between an artifact author and its reliability. This correlation, apparently, does not always have the same strength in all domains. However, by clustering the training set per user (in Table 7.4 we report the most significant results, with cut equal to 0.3), we almost always halve the computation time, and this gain, together with the relatively satisfactory accuracy, underlines the strength of our approach. Also, despite the previous case study, here in the clustered version of the algorithm, the accuracy is preserved but not improved. This supports the hypothesis that the small accuracy of the clustered version in the previous case is due to dataset peculiarities.

7.5 Discussion

In the two case studies described above, the algorithm that we propose achieves satisfactory results. This is due to the fact that two basic assumptions that we make actually hold, at least in part. The first assumption is that there exists a probabilistic

Annotations per reputation	Training set coverage	Accuracy	Precision	Recall	F-score	Time (sec.)
non-clustered results						
5	18%	0.68	0.79	0.80	0.80	1254
10	27%	0.70	0.79	0.83	0.81	1957
15	33%	0.71	0.80	0.84	0.82	2659
20	39%	0.70	0.79	0.84	0.81	2986
25	43%	0.71	0.79	0.85	0.82	3350
30	47%	0.72	0.81	0.85	0.83	7598
clustered results (cut=0.3)						
5	18%	0.71	0.80	0.84	0.82	707
10	27%	0.70	0.79	0.83	0.81	1004
15	33%	0.70	0.79	0.84	0.82	1197
20	39%	0.70	0.79	0.84	0.82	1286
25	43%	0.71	0.79	0.85	0.82	3080
30	47%	0.72	0.79	0.86	0.82	3660

Table 7.4: Performance on the Steve.Museum project dataset.

relation between the identity of users and the trustworthiness of their annotations. By knowing who made an annotation, we can make a probabilistic estimate about the trustworthiness of the annotation. As a consequence, we can make use of the user reputation as a basis for accepting or rejecting his annotations. The second assumption is that the reputation of a user can be estimated based on a small sample of observations about him. This is possible because the variance of the trustworthiness of the annotations provided by a given user is usually low. In other words, the trustworthiness level of the annotations provided by the same user is, in general, homogeneous. Thus, a small sample of user annotation is a significant sample about his performance. In order to apply the algorithm that we propose to other case studies, we need to verify that the assumptions above are satisfied and, otherwise, adapt the algorithm consequently. Another important requirement is that, to build our model, we need a gold standard composed, for instance, by a set of annotation evaluations provided by the institutions that manages the annotations. Future research will try to address the possibility to make trust assessments when a gold standard is unavailable. One restriction regards the fact that the algorithm we describe restricts its focus to single word annotations, since it makes use of the Wu & Palmer semantic similarity measure. It would be possible to relax this requirement by making use of semantic similarity measures that are able to handle small sentences (e.g., latent semantic analysis [47]). One last restriction of the algorithm regards the fact that it applies offline, that is, after all the annotations have been collected and part of them evaluated. In the next section we propose a variant of the algorithm that uses trust assessments and reputation management to guide interactively the process of annotation gathering.

7.6 Algorithm - Interactive Version

The algorithm we adopt assesses the trustworthiness of user contributed annotations based on a limited amount of evaluations. However, such an evaluation takes place offline, once the annotations have all been collected. We present here a small variant of the algorithm that aims at making it applicable online as well, and allows the cultural heritage institution to interactively choose the estimated best candidate for the task. This algorithm aims at increasing the efficiency of the system for crowdsourcing annotations. We saw in the previous section that the algorithm for evaluating the annotations offline performs reasonably well in terms of precision, recall and F-measure. However, since not all the annotators have a high reputation, a relevant amount of annotations is rejected, and this can be seen as an inefficiency for the overall crowdsourcing system. In fact, these annotations are discarded only at a later stage, after they are collected and evaluated. So, we propose an interactive version of the algorithm, that takes advantage of the good performance of the previously described evaluation algorithm and tries to reduce the amount of annotations rejected. This algorithm works as follows. Every time the museum needs to annotate a new artifact, it selects the best candidate annotator based on a preliminary classification of the artifact. The selection is based on the annotator profile, and the only additional requirement with respect to the previous algorithm is a preliminary classification of the artifact (e.g., impressionist painting). By selecting the most qualified annotator per painting, the algorithm tries to reduce the overall amount of annotations rejected. We introduce a pseudo-code version of the algorithm, presented in Algorithm 7.3, and we provide a qualitative description of it together with a workflow diagram (Figure 7.4).

Algorithm 7.3: Interactive version of the algorithm for trust estimation.

Input: Two finite sets, $Request\_set = \{\langle Request \rangle\}$ and $Users\_set = \{\langle User \rangle\}$
Output: A finite set $Output\_Test\_set = \{\langle annotation, trust\_values \rangle\}$

```
for Request ← Request_1 to Request_n do
    Users ← select_users(Request)
    for User ← User_1 to User_n do
        Result ← append_value(user,Request)
    Output ← evaluate_results(Result)
    update_expertise(User)
return Output
```

select_users Selects a set of annotators to whom we forward a *request* containing:

- A reference to the artifact to be annotated.
- A first, high-level classification of the item, that facilitates the annotator selection (e.g., the decade when it was made).

- The requested "facet", necessary to obtain comparable candidate values (e.g., the "what" facet, i.e., the artifact content).

The selection procedure depends on internal policies of the museum deploying the system, so we do not make it explicit. Some examples:

- Select the n highest ranked experts about the requested topic.
- Consider all the experts. Weigh their reputation with respect to the distance from the request. Order and select them.
- Also consider the belief and uncertainty (and impose some conditions on them) when selecting annotators.

append_value Collects the contributions obtained from the selected annotators. *result* is a list of pairs like (*value*, *annotators_opinions*).

evaluate_results Aggregates results and takes a decision about them. Subjective logic's cumulative fusion operator is a possible aggregation function. A decision strategy has to select a candidate value (possibly the highest-rated), while reducing the risk of taking a wrong decision and solving possible controversies, such as when multiple candidate values all share the highest rank.

update_expertise After having evaluated the candidate values for the annotation, annotators will be "rewarded" (if their candidate was selected) or "penalized" (otherwise). In principle, this means adding a positive piece of evidence to the first ones and a negative piece of evidence to the last ones, but once again, this may depend on museum policies.

Output The annotations selected can be directly accepted by the museum, or ranked qualitatively according to their trust level (e.g., "accept" when the trust level is higher than 0.9, "review" otherwise), so that appropriate actions are taken.

To maximize its efficiency, such an algorithm needs to have at its disposal specific information about the expertise of users and their reliability. Hence, it would benefit from an extensive representation of expertise by means of subjective logic and Semantic Web technologies. We represent the expertise of each annotator using the hoonoh ontology, by linking the URI representing the user to the one representing the concept of expertise. In RDF statements, it is represented as follows:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex: <http://example.org/ns#> .
@prefix hoonoh: <http://hoonoh.com/ontology#> .
ex:T1 a hoonoh:Topic, skos:Concept.
ex:user rdf:type foaf:Person .
ex:E1 a hoonoh:ExpertiseRelationship;
    hoonoh:from ex:user;
    hoonoh:toTopic ex:T1 .
```

Listing 7.2: Representing the user expertise by means of the hoonoh ontology.

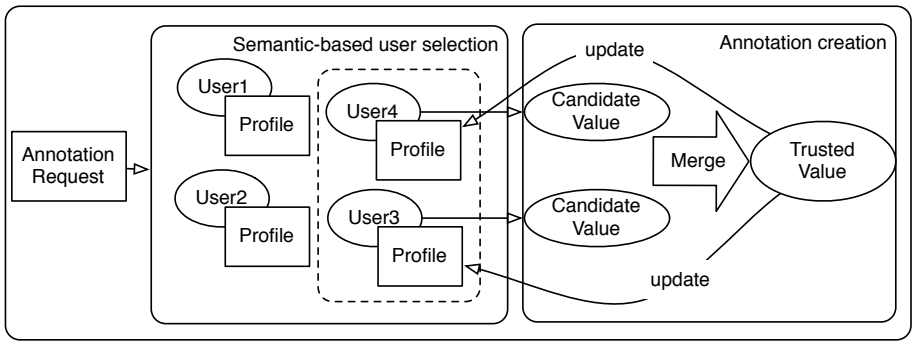


Figure 7.4: Algorithm workflow.

We define a data structure representing a subjective opinion, we link it to the corresponding hoonoh:ExpertiseRelationship and populate it with opinion elements, i.e., belief, disbelief, uncertainty and a priori value:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex: <http://example.org/ns#> .
@prefix hoonoh: <http://hoonoh.com/ontology#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix qb: <http://purl.org/linked-data/cube#> .
@prefix dcterms: <http://purl.org/dc/terms/> .

ex:Opinion rdf:type qb:DataStructureDefinition;
    qb:component
        [ qb:measure ex:belief; ],
        [ qb:measure ex:disbelief; ],
        [ qb:measure ex:uncertainty; ],
        [ qb:measure ex:apriori; ] .
ex:dataset rdf:type qb:DataSet;
    qb:structure ex:Opinion;
    dcterms:subject ex:E1 .
ex:obs1a rdf:type qb:Observation, prov:Entity;
    qb:dataSet ex:dataset;
    prov:wasAttributedTo ex:Museum;
    ex:belief 0.4;
    ex:disbelief 0.2;
    ex:uncertainty 0.4;
    ex:apriori 0.5.
```

Listing 7.3: Representing a subjective opinion about the user expertise.

Museum artifacts are annotated by means of Dublin Core [46] subjects, that are of type skos:Concept. For instance:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex: <http://example.org/ns#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
ex:item1 dcterms:subject ex:T1.
ex:T1 rdf:Type skos:Concept .
```

Listing 7.4: Representing an annotated museum artifact.

We are interested in determining the user expertise about a given topic, so, we compute opinions about objects of type hoonoh:ExpertiseRelationship. Opinions are recorded by means of objects of type qb:Observation defined before. If **eg**:E1 is of type **hoonoh**:ExpertiseRelationship, an opinion is:

$$expertise(user, T1) = \omega \begin{array}{l} \textbf{eg}\text{:E1 } \textbf{hoonoh}\text{:from } \textbf{eg}\text{:user} \\ \textbf{eg}\text{:E1 } \textbf{hoonoh}\text{:toTopic } \textbf{eg}\text{:T1} \end{array} (b, d, u, a) \tag{7.9}$$

7.7 Conclusion

In this chapter we present a framework which helps in partially, but efficiently and accurately automating the process of annotation evaluation in crowdsourced systems that extends the models presented in the previous chapters, in particular in Chapters 1 and 6. One of the major advantages of our system is that it does not require to set any particular parameter regarding decision strategies, hence, the final result does not rely on our ability to choose precise values for such parameters. The only parameter we need to set is the size of the training set used to build user reputations, but we observed that it does not substantially affect our performance, thanks to the smoothing factor introduced by subjective logic: smoothing helps to compensate for the fact that small training sets might diverge substantially from the whole population they are sampled from, and this limits the decrease in accuracy. This represents an important achievement with respect to the model described in Chapter 6 that is a precursor of this one. In addition, the use of semantic relatedness measures as weighing factors for the evidence allows us to make precise estimations. This is obtained thanks to the theoretical foundation for this combination presented in Chapter 3. The use of probability distributions to represent reputations allows us to make estimates taking into account that high reputations do not necessarily imply a perfect performance by the user. Clustering helps to make the computation affordable, without compromising accuracy. Lastly, we present a variant that makes it possible for museums to intervene in the process of annotation acquisition, by requesting annotations to the most reliable and expert users.

8

Provenance-based Assessment of Annotations Trustworthiness

Two Cultural Heritage Case Studies

The work presented in this chapter is built upon the results presented so far, in particular those in Chapters 6 and 7. Here we adopt uncertainty reasoning in combination with semantic similarity measures to semi-automatically predict the trustworthiness of annotations, similar to what we have described in Chapter 7, but we adapt the algorithm thus obtained to make provenance-based and not user-based predictions. In fact, we group the provenance information about these annotations in so-called "provenance-sterotypes" and we base our predictions on the stereotypes reputation. A stereotype is a set of provenance traces meant to coarsely describe a user behavior, similarly to what we described in Chapter 6. By evaluating the annotations on the basis of the reputation of the stereotyped user behavior rather than of the user itself, we show a means to possibly overcome the limitation of the classic user reputation-based trust management systems. We evaluate our algorithm over two datasets from the cultural heritage domain.

This chapter is based on the article Efficient Semi-automated Assessment of Annotations Trustworthiness*, coauthored with Archana Nottamkandath and Wan Fokkink, and published by Springer in the* Journal of Trust Management *in 2014. My contribution to this work regards the design, the implementation and the evaluation of the algorithm.*

8.1 Introduction

The goal of the work described in this chapter is to automate the process of evaluation of annotations obtained through crowdsourcing in an effective way and prescinding

from the author of these contributions. In Chapter 7 we proposed a model based on user reputations. Here we go beyond such a paradigm, and we adapt our algorithm to rely on the annotation provenance, similar to the work presented in Chapter 6. So, apart from the leitmotifs subjective logic and semantic similarity measure already introduced, we also use provenance to evaluate the quality of user-contributed annotations. Since we focus on annotations, provenance reports information on its creation such as time of day, day of the week, typing speed, etc., obtained by tracking user behavior. We use provenance information to group annotations according to the "stereotype" or "behavior" that produced them. So, we group them depending on whether they are produced by, for instance, early-morning or late-night users, because we hypothesize that this information, properly analyzed, can act as a heuristic for the trustworthiness of the annotations produced. Once the annotations have been grouped per stereotype, we compute a reputation for each stereotype, based on a sample of evaluations provided by an authority: we learn the policy adopted by the authority in evaluating the annotations and we apply the learnt model on further annotations. This allows leveraging the advantages of the algorithm proposed in Chapter 7 and, at the same time, evaluate annotations whose author is unknown or lacks a sure reputation (because not enough evidence about her reliability has been collected). We test our hypothesis by applying our algorithm on two datasets, one from the SEALINC Media project experiment described in Chapter 7, and the other from the Steve.Museum project. We show that the algorithm we propose is dependable and not solely dependent on the availability of information about the author of an annotation (as shown in Chapter 7). We assume that when the identity of the author is not known or when a reliable reputation about the author is not available, we can base our estimates on provenance information, that is, on a range of information about how the annotation has been created (e.g., the timestamp of the annotation). By properly aggregating such information (for manageability reasons), we derive a "stereotypical description" of a user's behavior. Users are often constrained in their behavior by the environment and other factors. For instance, they produce annotations within certain periodic intervals, such as the time of the day or day of the week. Being able to recognize such stereotypes, we can compute a reputation per stereotype rather than per user. This approach guarantees the availability of evidence, as typically multiple users belong to the same stereotype, while compensating for the lack of evidence about specific users. We evaluate our hypothesis over the two datasets mentioned before by splitting them into two parts, one to build a provenance-based model and the other to test it.

The chapter continues as follows: in Section 8.2 we describe the system proposed, and in Section 8.3 we present its evaluation. We conclude in Section 8.4.

8.2 System Description

The system that we propose aims at estimating the trustworthiness of annotations based on a set of evaluated ones (per user or per provenance stereotype, as we will see). To make the estimates, we make use of subjective logic (see Chapter Preliminaries)

to reason about the evidence at our disposal. Also, since this consists of textual annotations, we use semantic similarity measures (see again Chapter Preliminaries) to understand the relevance of each piece of evidence when analyzing each different annotation.

8.2.1 High-level System Overview

We propose a system that aims at relieving the institution personnel from the burden of evaluating all the annotations inserted by users, similar to what we propose in Chapter 7, although here we evaluate on user stereotype-basis and not on user-basis. The system asks for some interaction with the reviewers, but tries to minimize it.

For each provenance stereotype, the system asks the reviewers to review a fixed number of annotations, and on the basis of these reviews it builds stereotype reputations. A reputation is meant to express a global measure of trustworthiness and accountability of the corresponding stereotype. The reviews are also used to assess the trustworthiness of each annotation inserted afterwards: given an annotation, the system evaluates it by looking at the evaluations already available for its stereotype. The evaluations of the annotations semantically closer to the one that we evaluate have a higher impact. So we have two distinct phases: a first training step where we collect samples of manual reviews, and a second step where we make automatic assessments of annotations trustworthiness (possibly after having clustered the evaluated annotations to improve the computation time). We already saw in Chapter 7 that the more reviews there are, the more reliable the reputation is, and exactly as in that chapter, also here the amount of evidence used to build a reputation does not affect significantly the accuracy obtained. This is the only parameter we need to set in our algorithm.

8.2.2 Annotation Representation

Similarly to Chapter 7, we adopt the Open Annotation model [14] as a standard model for describing annotations, together with the most relevant related metadata (like the author and the time of creation). The Open Annotation model allows us to reify the annotation itself, and by treating it as an object, we can easily link to it properties like the annotator URI or the time of creation. Also, using the PROV ontology [9], we attribute a given annotation to a precise provenance bundle (as a means to represent a group of provenance traces). Moreover, the review of an annotation can be represented as an annotation which target is an annotation and which body contains a value of the review about the annotation.

Listing 8.1 shows an example of an annotation and a corresponding review, both represented as "annotations" from the Open Annotation model. Listing 8.1 represents a tag ("Chinese") created by a user, "Alex", about a given image ("img_231"). The tag has been created at a given time of the day and on a given day of the week, which determine the provenance stereotype the tag belongs to, that is represented by means of the bundle "c1".

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix oa: <http://www.w3.org/ns/openannotation/core/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix ex: <http://example.org/ns#> .
ex:user_1 foaf:givenName "Alex" .
c1 rdf:type prov:bundle .
ex:annotation_1 oa:hasBody ex:Chinese; oa:annotatedBy ex:user_1 ;
                oa:hasTarget ex:img_231; rdf:type oa:annotation ;
                prov:mentionOf c1.
ex:review oa:hasBody ex:ann_accepted ; rdf:type oa:annotation ;
                oa:annotatedBy ex:reviewer_1 ;
                oa:hasTarget ex:annotation_1 .
ex:annotation_accepted oa:annotates ex:annotation_1 .
```

Listing 8.1: Example of an annotation and respective evaluation.

8.2.3 Provenance Stereotypes

The algorithms described in the previous chapters, and in particular in Chapter 7, are based on the fact that there exists a relationship between the identity of an author and the trustworthiness of his annotations, or that the user reputation is a meaningful estimate of his ability to annotate. However, there might be cases when the user reputation is not available, for instance if there is not enough evidence about his trustworthiness or in case his identity is not known. We show that the algorithm presented here is not entirely dependent on the user reputation and, in case this is not available, other classes of information can be used as well. This class of information is so-called provenance information about how an artifact (in this case, an annotation) has been produced, and represents, therefore, an extension of the information about the sole author of the annotation.

We follow a reasoning similar to Chapter 6, as we use "provenance stereotypes" to group annotations. By stereotype we mean a class of provenance traces classified according to the user behavior they hint at. For instance, we could have "Monday early morning users" or "Saturday night users". We suppose that a given behavior should be associated with a particular reputation and hence with a given degree of trustworthiness of the annotations created in that manner, for two reasons:

- The trustworthiness of a given annotation might be affected by when it is created. For instance, late at night, users may on average be more tired and hence less precise than on other moments of the day.
- Users tend to follow a regular pattern in their behavior, because, for instance, their availability for annotating is constrained by their working time. Therefore, by considering their behavior, we implicitly consider their identity as well, even when they act as anonymous users.

In order to apply this kind of reasoning, we need to refer to the provenance information at our disposal about the annotations. In particular, these include only the day of the week and the time of creation for the datasets considered, but other information, when available, might be used as well (e.g. the typing duration for a given annotation). Since annotations are hardly created at the same time, in general do not coincide, we need to group them in order to be able to identify patterns in the data that allow us to link specific provenance information to the trustworthiness of the tags. In fact, the creation time of a tag may be recorded as a timestamp, but since tags are probably created at different times, we need to increase the granularity of this piece of information and analyze the part of the day or the day of the week when the tag was created, rather than the exact moment (tracked by the timestamp). Of course, this grouping introduces some uncertainty in the calculations because it introduces an approximation and because, in principle there are several possible groupings that we can apply, with different granularity and semantics (e.g. the days can be distinguished in weekdays and weekends, or simply be kept as single days of the week). In the next section, we report the results we obtained and we provide a possible explanation of why the grouping we propose allowed us to obtain the results we achieved, in the case studies we analyzed. In general, as we discussed in Chapter 6, we identify stereotypes with sequences of possibly processed provenance features. In case the provenance features at our disposal are numerous, then a preselection of these features is advisable, to avoid excessive computational burden. Such a preselection could be made by selecting the provenance features that correlate most with the trust values we aim at predicting by using a gold standard, so that only the most significant features are considered. This strategy can also be employed to choose the best processing method for data that are too fine grained. We saw that dates are too rarefied to allow having enough evidence per stereotype. So, we can extract part of the dates and coarsen them, but deciding the correct coarsening level may be hard. On the one hand, we need to coarsen these values in order to obtain high amounts of evidence for each stereotype. On the other hand, a finer granularity guarantees the possibility to identify stereotypes that correspond to behaviours that are significantly different from the trustworthiness point of view. For instance, if we do not make any distinction based on the hour of the day for a given day in the annotation dataset, all the observations produced in one day would belong to the same stereotype, and this would guarantee availability of evidence for that stereotype (provided that some observations are available). However, if we split the hours of the day in, for example, three parts (morning, afternoon, night), then it is possible that the corresponding stereotypes differ in trustworthiness of the associated annotations (and this actually happens in the case study below). Also, the method for coarsening these features can be implemented in different manners. For instance, hours of the day could be split in three even classes or split according to quantiles. In the case study presented below we create stereotypes manually, based on their expected ability to act as proxy for annotation trustworthiness, but in future research we will investigate the possibility to identify them automatically.

Lastly, from the modeling point of view, each group or stereotype can be thought of as a **prov:bundle** from the PROV Ontology [9], that is a “named set of provenance

descriptions", where each set groups provenance traces according to the day of the week and the part of the day they belong to. In fact, a stereotype can be seen as an identifier for a group of provenance traces. By annotating annotations in this manner, we can easily retrieve all the annotations belonging to a given stereotype without having to make use of specific SPARQL queries. This is useful, because despite the work presented in Chapter 6, we do not apply support vector machines to learn the trustworthiness of the annotations created with a given stereotype. Rather, we collect a predefined amount of evidence (i.e. of evaluated annotations) per group, and we evaluate the remaining annotations of the same group based on the reputation estimated using the evidence collected, so as to exploit the provenance semantics instead of using it only as a statistical feature. Hence, having at disposal a shortcut to identify all the provenance traces (and corresponding annotations) belonging to a given stereotype, facilitates the handling of stereotypes.

For representing provenance information we adopt the W3C Recommendation PROV-O Ontology [9], which provides founding types and relations for representing this specific kind of information, like entities and activities, which coincide with tags and tag creation processes respectively.

8.2.4 Trust Management

We employ subjective logic for making our trust assessments, as described in Chapter Preliminaries.

Trust is context-dependent (see Chapter Introduction), since different users and annotations (or, more in general, agents and artifacts) may receive different trust evaluations, depending on the context from which they originate and the reviewer. In our scenarios we do not have at our disposal an explicit description of trust policies by the museums, so we aim at learning a model that evaluates annotations as closely as possible to what that museum would do, based on a small sample of evaluations produced by the museum itself.

We define a stereotype reputation as a global value representing the expected quality of the annotations obtained by users behaving particularly (e.g., annotating at late night).

Once we decide how to group the provenance traces, we start collecting evidence per group. We fix a limit to the amount of evidence needed to create the opinion representing the stereotype's reputation. (In the experiment described in the next section we vary this limit to evaluate the impact it has on the accuracy of the reputation itself.) The reputation is computed as in the **build_reputation()** procedure described in Algorithm 8.2. First, we determine which stereotype the annotation belongs to. Then we increment the evidence count for the evaluation of the current annotation until we reach the limit per stereotype. Lastly, we convert the list of evidence counts in subjective opinions.

Once the training set has been built, we evaluate the trustworthiness of the annotations in the test set for each group. We compare each annotation to be evaluated against each piece of evidence in the training set, and we use the semantic similarity

emerging from that comparison to weigh the evidence and compute an opinion per annotation.

Once we obtain one trust value per annotation, we have to decide whether or not to accept the annotation itself. To be more precise, for each annotation we compute an entire opinion, representing the probabilities for each annotation to be correctly evaluated with one of the possible evaluations. Now we must decide which evaluation to assign to the annotation. One strategy would use, for each annotation, the evaluation having the higher probability. We do not adopt this strategy because by doing so we would most likely tend to evaluate all annotations of a given stereotype with the same dominant evaluation. For instance, if 95% of the training set annotations of one stereotype are useful, we will most likely evaluate all its annotations in the test set as useful. In turn, this implies that we do not take into account that we estimated that 5% of the annotations are not useful.

So we use an approach that combines the stereotype reputation with the trust values of the annotations, because we want to take fully into account the probabilities that are estimated by means of the reputation, and trust values estimate the trustworthiness of annotations. Algorithm 8.2 presents the algorithm for annotation evaluation.

Algorithm 8.1: Procedure for computing user reputation.

```
1  procedure build_reputation()
       Input: A list of annotations with their corresponding provenance
              information
       Output: A list of reputations, one per provenance stereotype
2      for annotation in training_set_annotations do
3          i ← annotation.get_stereotype_id()
4          if length(trainingset[stereotypes[i]])<n then
5              trainingset[length(trainingset[stereotypes[i]]) + 1] ←
               get_eval(annotation)
6          else
7              testset[length(testset[stereotypes[i]]) + 1] ← get_eval(annotation)
8      for s in stereotypes do
9          rep[s] ← compute_reputations(s)
10     return rep
```

Suppose that a user, Alex, created an annotation ("Chinese") on Monday at 13.00. Suppose, further, that in the group Monday-afternoon already the annotations {Japanese, Christian} have been evaluated as useful, while {rose} has been evaluated as not useful. Now the trust value of the annotation Chinese is evaluated as before, with as only difference that the evaluation is made on the basis of the provenance group it belongs to, and not of the author.

Algorithm 8.2: Algorithm to compute trust values of annotations using provenance stereotypes.

Input: A list of annotations with their corresponding provenance information
Output: A finite set of evaluated annotations
$Result\_Test\_set = \{\langle annotation, trust\_values\rangle\}$

```
for s in trainingset[stereotypes] do
    rep[s] ← build_reputation(Training_set)
Result ← {}
for s in testset[stereotypes] do
    for Annotation ← annotation[s][1] to annotation[s][n] do
        trust_values[Annotation] ← compute_tv(Training_set)
    s_annotations ← sort_annotations(trust_values)
    Result ← Result ∪ assess(s_annotations,rep[s])
return Result
```

$$p^m_{Chinese} = sim(\text{Chinese}, \text{Japanese}) + sim(\text{Chinese}, \text{Christian}) = 0.9 + 0.63 = 1.53$$

$$n^m_{Chinese} = sim(\text{Chinese}, \text{rose}) = 0.57$$

$$\omega^m_{Chinese}\left(\frac{1.53}{1.53+0.57+2}, \frac{0.57}{1.53+0.57+2}, \frac{2}{1.53+0.57+2}, \frac{1}{2}\right)$$

$$E(\omega^m_{Chinese}) = 0.62$$

The reputation of the group is as follows.

$$\omega^m_{Group}\left(\frac{2}{5}, \frac{1}{5}, \frac{2}{5}, \frac{1}{2}\right)$$

$$E(\omega^m_{Group}) = 0.6$$

So the annotation inserted by Alex will be accepted only if it is one of the 60% best annotations belonging to that group.

Implementation

The code for the representation and assessment of the annotations with the Open Annotation model has been developed using SWI-Prolog Semantic Web Library [174] and the Python libraries rdflib [133] and hcluster [48], and is available on the Web[1].

[1]The code is available at `http://trustingwebdata.org/phdthesis/dceolin`.

8.3 Evaluation

8.3.1 Datasets Adopted

We validate the algorithm we propose over two datasets of annotations of images. The first is a dataset from a SEALINC Media project experiment, that consists of 2,650 annotations of artifacts collected by the Rijksmuseum in Amsterdam evaluated using a scale ranging from one to five. The second is the Steve.Museum project dataset, that consists of 45,860 crowdsourced painting annotations manually evaluated by museum professionals as "useful", "problematic" or "non-useful". Both these datasets have already been introduced in Chapter 7. We refer the reader to Section 7.4 for additional details. In the SEALINC Media dataset, we consider only the annotations having an evaluation higher than three as positive pieces of evidence, while in the Steve.Museum case, only those classified as "useful" are considered as positive. We build our training set using a fixed amount of evaluated annotations for each of the users, and form the test set using the remaining annotations. The number of annotations used to build the reputation and the percentage of the dataset covered are presented in Tables 8.1 and 8.2.

8.3.2 Results and Discussion

We evaluate the algorithm that we propose by running it on Steve.Museum and SEALINC Media experiment datasets. As described before, we split each dataset into a training and a test set, learn a model based on the training set, and evaluate it on the test set. There is a tradeoff between complexity and performance. On the one hand, a larger training set in general produces a more accurate model. On the other hand, an increased size of the training set induces a larger number of comparisons for each estimate, and hence an increased computation cost. To determine an optimal size for the training set in each case study, we ran the algorithm with different training set sizes, expressed in terms of annotations per user reputation, and tracked their performance. Table 8.1 and Table 8.2 present the results for the SEALINC Media and the Steve.Museum dataset, respectively. We run this evaluation with the same setting as before. Since we are interested only in checking whether the trustworthiness estimations based on provenance stereotypes perform as well as those based on user reputations in terms of precision and recall, we do not report the execution times of the algorithm.

In Table 8.1 precision is about 88% and recall ranges between 73% and 88%. The decrease in accuracy for the training set built with 20 annotations per reputation is plausibly due to the fact that many provenance stereotypes do not have 20 or more annotations available, so these groups cannot contribute to the overall accuracy measurement, while they do with 5, 10 or 15 annotations per reputation. So, some errors can be due to intrinsic limitations of the experiment rather than imprecision of the algorithms. For instance, since training and test set are part of the same dataset, a larger training set means a smaller test set, and vice versa. Since our prediction

is probabilistic, a small training set forces us to discretize our predictions, and this increases our error rate. Also, while an increase of the number of annotations used for building a reputation produces an increase of the reliability of the reputation itself, such an increase has the downside to reduce our test set size, since often only few annotators produce a large number of annotations.

Moreover, the amount of evidence needed to make these assessments is low, as demonstrated by the percentage covered by the training set over the dataset. In Table 8.2 the performance is even higher than in Table 8.1. First, this is due to the existence of a correlation between the provenance group an annotation belongs to and its trustworthiness. Second, the fact that the provenance stereotypes that we consider for this experiment are 21, which is much less than the number of users, together with the imbalance between useful and non-useful annotations in the Steve.Museum dataset (the first are much more plentiful than the latter) compensates a collateral effect of smoothing. In fact, smoothing helps in allocating some probability to unseen events (for instance, possible future mistakes of good users). So, because of smoothing, we predict the existence of non-useful annotations for users who actually did not produce them (the dataset contains only relatively few non-useful annotations). Since there are many more users than provenance stereotypes, this error is higher with user-based estimates, where there are many more smoothed probability distributions (one per author), which causes many more annotations to be wrongly evaluated as non-useful. On the other hand, with provenance stereotypes, this error is much more limited, because the corresponding smoothed reputations introduce fewer wrong non-useful evaluations. Still, we will continue employing smoothing, as these are posterior considerations based on the availability of privileged information about the test set (i.e., its evaluation), and smoothing allows us to compensate the lack of this information. On the other hand, the specific Steve.Museum dataset possibly shows a limitation of smoothing.

Annotations in each reputation	Accuracy	Training set coverage	Precision	Recall	F-measure
5	0.68	1.69%	0.88	0.73	0.80
10	0.71	3.35%	0.87	0.80	0.83
15	0.78	4.97%	0.88	0.88	0.88
20	0.72	6.45%	0.87	0.80	0.83

Table 8.1: Results of the evaluation of Algorithm 8.2 over the SEALINC Media dataset for training sets formed by aggregating 5, 10, 15 and 20 reputations per user. We report the percentage of dataset covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

In the previous section, we hypothesized that the time of creation of an annotation may implicitly affect its trustworthiness and that the users follow approximatively reg-

Annotations in each reputation	Accuracy	Training set coverage	Precision	Recall	F-measure
5	0.84	0.25%	0.84	0.99	0.90
10	0.84	0.45%	0.84	0.99	0.90
15	0.84	0.66%	0.84	0.99	0.90
20	0.84	0.86%	0.84	0.99	0.90
25	0.84	1.04 %	0.84	0.99	0.90
30	0.84	1.22 %	0.84	0.99	0.90

Table 8.2: Results of the evaluation of Algorithm 8.2 over the Steve.Museum dataset for training sets formed by aggregating 5, 10, 15, 20, 25 and 30 reputations per user. We report the percentage of dataset covered by the training set, the accuracy, the precision, the recall and the F-measure of our prediction.

ular patterns in their behaviors. To support these statements, we made the following analyses:

- we computed the average of the user reputations per provenance group. The averages vary from 0.73 to 0.84 in the Steve.Museum case study and from 0.75 to 0.91 in the SEALINC Media case study. Each user that took part in the SEALINC Media experiment participated only once. Moreover, their contributions are concentrated in the mid part of the weekdays, so we could not make additional checks. In the Steve.Museum dataset, instead, we also run a series of Wilcoxon signed-ranked tests at 95% confidence level (since the data distribution is not always normally distributed, as shown by a Shapiro-Wilk test at 95% confidence level, we prefer not to use a t-student test), and we discovered that:
  - there is no significant difference within user reputations in the morning, afternoon, and night slots respectively across the week. For instance, we took the reputations in the morning slots for Monday, Tuesday, etc. and the Wilcoxon signed-rank test showed no significant difference. The same holds for the afternoon and the night ones;
  - there is a significant difference between the morning and the afternoon slots and the afternoon and night slots. Here we compared the series of reputations per slot across the week;
  - if we compare the averages of the reputations with respect to the days (for instance, considering the three slots of Monday versus the three slots of Tuesday, etc.) we see no significant difference;
  - there is no significant difference between weekends and weekdays.

  The first two points support our hypothesis because they show that actually there are some relevant differences between groups and actually these depend on

the time of creation of an annotation. The third and the fourth point show that, at least in this case study, it is not useful to keep track of the day of the week when the annotation was created. On the other hand, the fact that we recorded the day of the week allowed us to check if there is any difference both among days and between weekend and weekdays, while if we started directly with this latter distinction, we could not have decreased the granularity.

- as we stated in the previous item, the average number of provenance groups a user contribution belongs to is 1 in the SEALINC Media dataset. In the Steve.Museum dataset, instead, the average number of groups a user contributions belongs to is 1.17, variance 0.56. This means that most of the users' contributions belong to one group. So we can say that, approximatively, there exists a one-to-many relation that links the groups with the users: given a group, we can identify a group of users that provide annotations mostly in that group. This means that, when we analyze the annotations that belong to a given group, then we implicitly analyze the annotations produced by a group of users that annotate mostly in that time interval. So the provenance group acts as a proxy to this group of users, and hence, in practice, we analyze the annotations in that group based on the reputations of the users linked to that group. In principle, there may be a high variance among the users belonging to a given provenance group. However, in the case studies analyzed in this paper, this does not happen to be the case, since the variance of the users reputation belonging to a given group is low.

- in the Steve.Museum case study, the variance of the user reputations ranges between 0.12 and 0.15. This shows that, even if the averages of user reputations per group range between 0.73 and 0.84, the reputations are not sparsely distributed. Rather, within provenance groups users tend to be rather homogeneous in terms of reputation. The same holds for the SEALINC Media case study, where the variance of user reputation per provenance group ranges between 0.004 and 0.01;

- the time that we used in our computation is the server time and the fact that, in principle, the annotations are collected worldwide, this might imply that our calculations are misleading. However, since: (1) as shown before, there is a consistent distinction between morning, afternoon and night reputations (which is determined by user performance, and users tend to contribute at fixed times), (2) the amount of tags annotated as "problematic-foreign" is very small (about 1.9%) and (3) the artifact annotated in the case study belong mainly to U.S. cultural heritage institutions, we assume that the annotations are approximatively provided by users in the same time zone or in the neighboring ones.

When grouping the tags based on time, the choice between coarser and finer granularity is not trivial and, in general, affects the uncertainty of the final result. Grouping the tags at a coarser granularity allows easily collecting evidence for a given group and finding a semantic justification for the differences between groups. If we find a difference between morning and afternoon tags, we can easily suppose (and possibly test)

that this is due to the influence that different parts of the day have on the user conditions (tired, sleepy, etc.). If we find a difference between tags made at 8.00 a.m. and at 9.00 a.m., we may need additional information to justify semantically the reasons of such differences. On the other hand, a finer granularity may reveal to be useful to avoid to group together heterogeneous tags. All these are generic considerations, and the choice of the best granularity depends on the peculiarities of the single use case evaluated. In our cases, as is evident from the considerations above, we chose a coarser granularity for the hours of the day and a finer one for the days of the week, because this combination was the most significant and gave us the highest accuracy. Future work will investigate the possibility to automatically determine the best granularity level for this grouping.

8.4 Conclusion

We present an algorithm for automatically evaluating the trustworthiness of user-contributed annotations by using subjective logic and semantic similarity measures to learn a model from a limited set of annotations evaluated by an institution. In particular, our model predicts the annotation trustworthiness based on the provenance stereotype that the annotation belongs to, that is, on the temporal frame when the annotation was created. This represents an attempt to go beyond the classical reputation-based trust management (like those presented in Chapters 1, 6 and 7) and move the focus from the reputation of the user to the reputation of the behavior assumed by him. It turns out to be useful, for instance, in cases when the user or his reputation is not known. This kind of approach also might prove to be useful to prevent some prejudice against particular classes of users who might affect the quality of the annotation assessment: for instance, a young annotator might not be highly trusted because of his apparent low expertise, but this kind of analysis might help in adjusting such an estimate, if necessary. Finally, this approach can also be used to suggest to the user to adopt particular behaviors that have shown to lead to particularly trustworthy annotations. Also in Chapter 6 we propose a model for making trust predictions based on provenance stereotypes, but here we merge that achievement with the improvements provided in Chapter 7. Hence, here we merge successfully the accuracy of the probabilistic approach proposed in Chapter 7, that relies on the combination of subjective logic with semantic similarity measures proposed in Chapter 3, with the efficiency of the provenance-based approach proposed in Chapter 6. We evaluate each algorithm on two different datasets of annotations from the cultural heritage domain. Our algorithm satisfactorily allows us to estimate the annotation trustworthiness with a level of accuracy of about 80% in one case and 73% in the other one.

Conclusion and Discussion

This chapter presents conclusions we draw about the work presented in this thesis. We revisit the research questions presented in Chapter Introduction, then we present a final discussion and future work indication.

The overall problem addressed in this thesis is represented by the following research question: *How can the trustworthiness of semi-structured Web data be adequately estimated?* We develop our research around the following statement, that represents the core message of the thesis:

> ***The trustworthiness of semi-structured Web data can be adequately estimated by making use of uncertainty reasoning, possibly assisted by provenance analysis and semantic similarity measures.***

Based on this statement, we decompose the main research question in four subquestions and, for each of them, we perform a series of experiments intended to validate a subcomponent of the thesis statement. First, we analyze whether ***uncertainty reasoning is a key element for determining the trustworthiness of semi-structured Web data***. In Chapters 1 and 2, we perform two preliminary studies about the use of uncertainty reasoning together with Web data to make trust assessments on two datasets provided by the Naturalis Museum in the Netherlands. From these two case studies emerge the basic procedures adopted through the rest of the thesis as the starting point for determining the trustworthiness of semi-structured Web data. Then, in Chapters 3 and 4 we explore in depth the use of uncertainty reasoning techniques, and in particular of higher-order probability distributions and of subjective logic to model, represent and estimate Web data and their trustworthiness. From these analyses we derive a representation of Web data with higher-order probability distributions that we apply on a dataset of piracy attacks. Also, we provide an advanced procedure for estimating the reliability of Web data, that extends the procedures introduced before and is evaluated on a set of Police open and closed data. Then, we analyze the role played by ***provenance*** in assisting uncertainty reasoning for trust assessment of semi-structured Web data. In Chapter 5 we propose a first procedure for applying uncertainty reasoning over provenance graphs to estimate the trustworthiness of Web data (AIS ship messages in the specific case study analyzed).

In Chapter 6 we introduce the concept of "provenance stereotype" (that we further developed in Chapter 8) and we propose a series of procedures that employ provenance stereotypes to estimate the trustworthiness of video tags. Lastly, we analyze the use of ***semantic similarity measures*** to assist the trustworthiness estimation of Web data. First we propose a procedure for estimating the trustworthiness of museum annotations that makes use of user reputation in combination with semantic similarity measures in Chapter 7 and then, in Chapter 8 we substitute users with provenance stereotypes, and we show how it is possible to combine uncertainty reasoning, provenance analysis and semantic similarity measures to estimate the trustworthiness of Web data. Another important characteristics of these two latter procedures is that they not only allow estimating the trustworthiness of data, but they also allow distinguishing the trusted data from the untrusted ones (i.e., they allow "placing trust", following the theory of O'Hara [121]) in an automated manner. These procedures are evaluated on two datasets from the cultural heritage domain.

The Research Questions Revisited

Research Question 1 - Can Web data help the trust evaluation of semi-structured data?

This research question is tackled throughout the thesis, but in particular Chapters 1 and 2 present insights about this issue. First, in Chapter 1 we demonstrate the value of enriching data internally curated by an authority with trusted Web data. The use of Web data permits to increase the performance of the trust algorithm that we propose, and the use of Semantic Web technologies facilitates the data aggregation process. This result is only based on a limited case study and makes use of selected Web sources that are trusted and known to be authoritative. Therefore, on the one hand, it represents a positive partial answer to the research question, while on the other hand the result can not be generalized for all Web data. These data need to be properly selected and handled in order to fruitfully and safely use them for trust assessments but, when this selection is applied correctly, linking the data to be evaluated to Web data is useful to find patterns that help in the process of trust assessment. Moreover, the use of uncertainty reasoning techniques reveals to be a positive choice to handle this kind of data, for two reasons. First, it allows us to take into account the inner uncertainty in these data and in their representativity (often we observe just a sample data, and we do not know how representative it is of the entire data population). Second, this probabilistic logic allows us to reason over trust while taking advantage of the graph structure of the aggregated data.

Chapter 2 shows another case study that makes use of Web data to estimate the confidence of semi-structured data. Here we evaluate the use of a series of heuristics (some of which are Web-based) to make assessments about the confidence in some georeferenced records. Despite the work presented in the first chapter, in Chapter 2 we evaluate a series of heuristics developed to estimate the confidence in a series of

georeferenced entries, and some of these heuristics turn out not to be correct indicators of the confidence in the georeferenced records. So, this confirms that it is possible to use Web data to make trust assessments of semi-structured data (some of the heuristics worked decently), but the correct choice of these sources is crucial for the success of the trust estimation process and again, uncertainty reasoning techniques are found to be useful to properly handle Web data used for trust assessments. Also, the choice of the Web data for making trust assessments depends not only on the authoritativeness of the sources, but also on their relevance in a specific context, confirming that trust is context-dependent, as stated in the theory of O'Hara [121].

In general, Web sources could have been too noisy or unreliable for us. Moreover, the Web is so vast that it could have been hard to find out reliable data sources. These chapters show that, when we choose sources that are known to be authoritative (Chapter 1) or when we select the most useful and reliable sources after having analyzed them (Chapter 2), we can use the data that these expose for making trust assessments. Therefore, the work presented in these chapters is fundamental for the rest of the thesis because it shows that we can extract useful data from Web sources, either directly or after filtering them. We need to handle such data with care, and we address this issue by tackling the following research question, but the fact that we can use them (at least, in specific use cases) is a crucial achievement.

Research Question 2 - How can uncertainty reasoning be effectively used to estimate the trustworthiness of semi-structured data?

Chapter 3 starts by showing that second-order probability distributions can be effectively used to model categorical Web data. The idea behind a statistical representation of Web data is that usually we observe only small portions of these, and to be able to effectively use them for making trust assessments, we need to understand how representative these data are. The probability distributions used in that chapter are shown to be particularly useful in plausibly depicting the sample representativity. Also, some of the probability distributions adopted in that chapter represent the statistical background of subjective logic that we use in Chapters 1 and 2. These facts motivate us to: (1) continue adopting this logic, at least for the representation of trust values; and (2) tighten the connection between the probabilistic logic and the probability distributions. From this effort results the development of open world opinions, that are subjective opinions based on Dirichlet processes, which extend subjective logic. That chapter moreover describes other subjective logic extensions aimed at integrating the logic and other kinds of Web-based knowledge, being partial evidence observations and semantic similarity measures. The first extension is particularly useful in Chapter 6, since it provides a method for computing trust values based on number of matches for tag entries from the video tagging platform *Waisda?*. The second extension, instead, creates a bridge between Part I and Part IV, because it motivates the choice of uncertainty reasoning techniques for handling Web data for trust assessment, and lays the foundations for using semantic similarity measures in combination with subjective

logic as done in Chapters 7 and 8.

Chapter 4 addresses this issue from a different perspective. In Chapter 3 uncertainty reasoning is employed to estimate the data representativity and use them accordingly, and subjective logic is extended to provide a uniform framework for trust assessments using Web sources. In Chapter 4, instead, uncertainty reasoning is used for two reasons. In that chapter trust is interpreted in terms of reliability and the analyses proposed are aimed at assessing the reliability of a set of police open data derived from a corresponding closed dataset. Uncertainty reasoning techniques are used here, in the first instance, as a means to measure the reliability of open data. In fact, we know that some differences between open and closed data have been introduced by purpose (for privacy reasons), so choosing uncertainty reasoning techniques to measure the difference between open and closed data is a means to identify relevant (i.e., statistically significant) differences without dwelling on small differences that were intentionally introduced. Then, uncertainty reasoning is used to test the similarity between consecutive open datasets, and to aggregate the results thus obtained. The search for reliability changes in a series of open datasets is subject to high degrees of uncertainty because its outcome can not be confirmed by closed data. Uncertainty reasoning here helps in dealing with such uncertainty. Subjective logic allows us to aggregate the evidence collected by means of tests that may be informative or not.

Chapter 5 tackles the research question with the use of provenance data, hence establishing a link between Part II and Part III. In that chapter we use the logical facilities of subjective logic to combine opinions about elements of the graphs describing the provenance of artifacts. By aggregating all the opinions using subjective logic operators, we obtain a final opinion representing the estimated trustworthiness of the artifact of interest.

Summarizing, we show that uncertainty reasoning helps in estimating the trustworthiness of semi-structured data in three manners.

- By giving a means to estimate the data distribution and the sample representativity, hence allowing us to handle the data prudently.
- By giving a means to estimate data similarity, to identify regularities in the data.
- By giving a means to combine observations about different facts that might affect the data trustworthiness. And by aggregating those observations in a properly weighted manner.

We show in the previous subsection that there exist Web sources that expose reliable semi-structured data that are useful for our computations. Nevertheless, these data present also regularities and patterns, despite a degree of uncertainty. This is crucial because it means that we can build model to mimic, predict or represent at least part of them. We had no prior assurance that the data were regular enough to be modeled by means of the relatively simple uncertainty reasoning techniques that we adopt. The choice of a suitable model could have been not trivial, and such a model could have been particularly complicated and computationally heavy. However,

we show that it is possible to use the techniques mentioned above to model and to handle semi-structured data (and, in particular, Web data), and the fact that these are effective in dealing with the uncertainty in the data that we use, and that their computation burden is light, makes these models ideal for our needs.

Research Question 3 - How can provenance information be used for making accurate trustworthiness estimations of semi-structured data?

Chapter 5 addresses the problem of trusting semi-structured data by using subjective logic. That chapter also makes use of a particular useful class of meta-information of these data, that is provenance. Provenance describes how, by whom, when and where artifacts (semi-structured data in this case) have been produced. It is an important class of information that needs to be considered when estimating the trustworthiness of semi-structured data. Especially when we can not judge data from their content, knowing the reputation of whom produced them, or of the processes that led to them, might give useful hints about their trustworthiness.

Provenance gives relevant support to the process of trust estimation of semi-structured data. However, provenance needs to be properly managed in order to be effective. How could otherwise the fact that a given piece of data was produced by a given agent help us to estimate the data trustworthiness? Per se, that information does not provide enough warranties, because that kind of information needs to be interpreted in order to furnish indications about the data trustworthiness or untrustworthiness. Uncertainty reasoning techniques are employed in Chapter 5 to compute the reputation of a set of provenance artifacts belonging to the provenance traces of a given piece of data, and particular operators are used to aggregate such trust values in order to reflect the processes that actually took place over those artifacts. Subjective logic is used to its full potential in this application. In fact, the trustworthiness of each artifact is determined on the basis of the observations at our disposal, and the logic's operators are a handy tool to correctly combine the opinions about the artifacts from which the artifact of our interest is derived. A case study presented in Chapter 5 shows how it is possible to determine the trustworthiness of AIS messages based on estimates of the trustworthiness of each subcomponent of these messages used in the naval domain.

However, the approach presented in Chapter 5, despite its high potential testified by the ability to produce trust values for a set of AIS messages, has two possible problems that might limit its generalizability. One is that it relies on the availability of evidence for each artifact involved in the provenance trace. Although subjective logic allows us to compute a trust value even when no evidence is available, a good outcome depends on the availability of observations for each artifact. The second limitation regards the choice of the subjective operators that aggregate the trust values of the artifact according to the process that was operated over these artifacts. In some cases, the choice of this operator could not be trivial, especially in case of particularly complex provenance processes.

In Chapter 6 we tackle the research question with another method that aims also at overcoming the limitations just described. In that chapter we use machine learning algorithms to predict the trustworthiness of crowdsourced media annotations based on similarities in their provenance graphs. This approach is useful because it allows us to "blindly" use provenance to make trust assessments, since the machine learning algorithm is used as a "black box" which does not require any knowledge about how the artifacts and agents involved in the process of annotation creation affects the trustworthiness of the annotation. The results are satisfactory, although clearly this approach, even if it solves some of the previous limitations, achieves this result by renouncing to consider some of the semantics captured by the provenance graphs. There is, therefore, room for future research in this direction, and we elaborate on this in the following section.

One last remark regards the fact that we can also reinterpret the results presented in Chapter 1 in the light of the considerations above. In that chapter we linked the data at our disposal to Web data and, even if that was not explicit, the type of information that we retrieved from the Web was actually provenance information. For instance, the author of a bird specimen annotation is easily encoded as a *prov:Agent* in the PROV Ontology [9]. Amongst others, in the case study described in Chapter 1 we make use of the Dublin Core ontology [46], and a mapping with the PROV Ontology is also officially proposed by the W3C [111], as to testify that part of the information used in that case is provenance information.

Provenance reveals to be very useful in the process of trust assessment. Either because we are able to interpret the process of data creation in trust terms, or because provenance gives a means to identify regularities in the data that would otherwise be hidden. Our evaluations of provenance-based trust estimates suggest that this is an interesting direction to take also in the future. Provenance alone does not provide the necessary knowledge to make useful trust estimates, but the coupling with uncertainty reasoning shows positive results. The fact that this coupling is effective is important, because it tells us which direction to take when using provenance for estimating trust. Moreover, it tells us that there exists a correlation between specific provenance features and the trustworthiness of a given artifact. It is intuitive that knowing, for instance, who made something can help us to estimate how good that artifact is. However, here we show that such an intuitive fact is exploitable for making reliable trust assessments because, at least in some contexts, it holds consistently.

Suppose that users produce tags with random trustworthiness: in that case we would not have been able to use uncertainty reasoning techniques to estimate the trustworthiness of these tags. This reasoning can be extended also to the other provenance features we make use of: we could make our provenance-based estimates because there exists a relationship between provenance and artifact trustworthiness. Suppose, further, that the provenance information that hints at the trustworthiness of the artifacts that we analyze is so complex and specific to require tailored methods for each specific estimation. This would have been a relevant problem that could have compromised the applicability of our solutions. In part we run into this problem in Chapter 5, but we solved it by manually selecting the subjective logic operators to use in that

case and by employing more generic methods in the subsequent case studies presented in Chapters 6 and 8.

Research Question 4 - Can semantic similarity measures improve the accuracy of trust estimates of semi-structured data based on uncertainty reasoning?

Semantic similarity measures represent a complementary means for improving the management of evidence at our disposal for making trust assessments. This kind of measures takes advantage of Web sources, either in case we use deterministic semantic similarity measures (like the Wu & Palmer similarity measure, that is based on WordNet) or in case we use probabilistic ones (like the Normalized Google Distance, that is based on the results given by the Google search engine).

We employ semantic similarity measures for assessing trust only in one specific domain, the cultural heritage domain, but other developments are possible as well. In Chapters 7 and 8 we use the Wu and Palmer semantic similarity measure to properly weigh the evidence used in the uncertainty reasoning techniques adopted for assessing the trustworthiness of annotations. We combine semantic similarity with subjective logic and demonstrate how it is possible to extend the logic to include that measure in a theoretically sound manner in Chapter 3. The use of the similarity measures allows us to improve the management of evidence about the performance of annotators, hence adapting the interpretation that we make of this evidence according to the context of the assessment. So, if we are evaluating a given annotation provided by an annotator, we weigh all the observations (i.e., an evaluated annotation) we already have about that annotator in order to give higher weight to the observations that are semantically closer to the annotation to be evaluated.

We run our evaluation on two datasets, using an annotator-centered algorithm in Chapter 7 and provenance-based one in Chapter 8, and the results we obtain are promising. The use of semantic similarity measures in combination with uncertainty reasoning can be seen as a refinement of the use of those reasoning techniques over provenance. Semantic similarity measures have the goal to help to estimate the expertise of a given annotator (or in general, of an artifact creator), depending on a set of evaluated artifact created by her. The expertise is implicitly determined every time an artifact of the same author is evaluated.

Besides their accuracy, semantic similarity measures are used in Chapter 7 to improve the computation time of the uncertainty reasoning-based algorithms. This is orthogonal to the research question tackled here, but the fact that the accuracy is not compromised by this addition, shows another potential for the use of semantic similarity measures in trust assessment.

Finally, already in Chapter 4 we show how statistical similarity measures are a helpful means to obtain hints about the trustworthiness of categorical data. These and the semantic similarity measures adopted in Chapters 3, 7 and 8 perform a similar task in two domains: semantic similarity measures allow us to identify similarity in textual evidence, while statistical semantic similarity allow us to identify similarity

in categorical evidence. Both kinds of similarity measures are important in the trustworthiness estimation process. In the same way as provenance allows us to identify links in our data that would otherwise be hidden, similarity measures allow us to identify regularities in the data and refine our estimates. We plan to further investigate the use of both kinds of similarity measures and their combination for making trust assessments, for instance by measuring the semantic similarity between categories in categorical data and comparing such similarity with the results of statistical similarity measures run on such categorical data.

Tackling the previous research questions, we saw that when using uncertainty reasoning techniques and in particular, subjective logic, the more pieces of evidence we have, the more reliable our estimates are. In principle, the use of semantic similarity measures is aimed at extending the set of evidence at our disposal. Semantic similarity measures allow considering also evidence about different subjects than those of interest. On the other hand, since this evidence is weighed, we run the risk to reduce the accuracy of our predictions, since these measures inevitably reduce the evidence counts at our disposal, because we multiply semantic similarity measures by that evidence, and the semantic similarity measures we use range between zero and one. If a piece of evidence is semantically identical to the target of our evaluation, then it is counted as one. Otherwise, it is counted as a number between zero and one, corresponding to the value of the similarity. The ability to extend the set of evidence considered and, at the same time, to understand the qualitative importance of each piece of evidence, case by case, rewarded us. This explains why the use of semantic similarity measures is helpful to improve the accuracy of our estimates.

Discussion and Future Work

Reflection on the Applicability of the Procedures for Trustworthiness Estimation

In this thesis we propose a set of procedures for estimating the trustworthiness of semi-structured Web data and deciding which of them to trust. The procedures are designed to address specific research questions identified in the case studies described, and they differ with respect to the aspects they emphasize (e.g., provenance analysis). However, based on them, we identify a common methodology that a data analyst can adopt to determine the trustworthiness of Web data. We present it as follows.

1. Identify a set of evaluations for at least a subset of the data to be evaluated. Trust assessments are subjective, so it is crucial to have at our disposal a sample set of evaluation to be used to train our models. If such evaluations are not available, then an estimate of such a gold standard is needed, e.g. based on user agreement and disagreement [3, 78, 145]. Given that we estimate the trustworthiness of data based on metadata and, in a limited manner, on their content (by means of semantic similarity measurements), it is necessary that this requirement is fulfilled to be able to apply this methodology.

2. Identify a creator for each piece of data. If such information is unavailable, then define user stereotypes based on the provenance information available and identify which stereotype each piece of data belongs to. Again, the availability of metadata is crucial for being able to apply our method. If no metadata is available, then a possibility could be to move to a content-based approach, if applicable.

3. If creators or provenance stereotypes have been identified, then compute a reputation for each of them, by using the evaluations at our disposal. We propose to use subjective opinions to represent user reputations because it allows aggregating and smoothing the evidence at our disposal, depending on the size of the evidence set. Other approaches may be possible as well.

4. Are the data comparable by means of a similarity measure (e.g., semantic similarity in the case of annotations)? If yes, then compute a trust value per piece of data to be evaluated, by weighing the relevant evidence with respect to the similarity with it. The relevant evidence is the evidence related, for instance, to a given author or provenance stereotype. The combination of subjective logic with semantic similarity measures we propose in Chapter 3 that is used in Chapters 7 and 8 exemplifies this approach. Once we computed such a trust value, we can use it to decide which data to trust and which not, for instance by setting a threshold and accepting only the values above it (see Chapter 6), or through ranking (see Chapters 7 and 8). If data are not comparable by means of a similarity measure, then one can use a probabilistic strategy to evaluate them by starting from a user or stereotype reputation estimated using the training set. Chapter 1 offers a range of decision strategies that are applicable in this case. Other strategies will be investigated in the future.

This approach may be subject to additions and changes in the future, as new research will be devoted to the topic. However, it has been effectively applied throughout the thesis, so we consider it at least as a starting point with respect to the methodology for trust assessment of semi-structured Web data.

Depending on which data and metadata are available, there is a progression of results that is possible to obtain with the method above: user or stereotype reputations, data trustworthiness estimations, trusted data selection. In principle, the role played by semantic similarity measures in assisting the computation of trustworthiness estimates tailored for each piece of data may be played by metrics based on provenance analysis as well. We will investigate this use of the analysis of provenance information in the future.

Future Research Outline

We identify here some limitations and open research questions for the work presented in this thesis.

1. The analyses of provenance for trust estimation that we propose represent only a starting point about the potential offered by this class of information.

2. The methods presented make a bland use of semantics. Since we make use of Web data in our analyses, it is worth investigating the use of semantically enriched data in our estimations.

3. Trust representation is mainly adopted for storing the results of computations, but a deeper study of this aspect might improve it and enhance the sharing and reuse of the results.

4. The human aspect is almost neglected in the analyses made. It might be useful to incorporate more social and human-centered aspects in the evaluation of data trustworthiness.

5. As a consequence of the previous point, security needs to be considered. It was not a focal point of the research presented here, but it becomes more relevant in a social perspective.

We address these limitations in detail and, by doing so, we describe future directions that we envision for this research.

1. Provenance We use provenance analysis for trust estimation in three chapters of this thesis. The results are promising, but the use of provenance can be further investigated, in order to identify more precisely the provenance features one needs to focus on to estimate the trustworthiness of pieces of data, and to fully take advantage of the semantics of provenance graphs. If general rules could be learnt, we might be able to evaluate the trustworthiness of a piece of data based on the characteristics of its provenance graph without having to learn, again, a model that fits particular requirements, as we do in our case studies. It would be sufficient to adjust the evaluation to the required constraints, and apply the generic rules to obtain it. Otherwise, a deeper analysis would at least allow us to guide the selection of provenance features to look at. In fact, having to deal with complex provenance graphs might lead to computational issues. Being able to identify links between specific provenance features and the trust evaluation of specific artifacts would be heavily beneficial for our computations, so we believe this is one direction worth investigating.

Another interesting related issue is the combination of provenance- and reputation-based estimates that has been touched in Chapter 6. User reputation-based trust estimates combined with provenance-based ones show to be particularly effective in the case study presented in that chapter. We will apply that combination in other domains as well, and we will study deeply the methods for combining the two estimates.

2. Semantics The use of semantics might be further enhanced in our calculations, in order to grasp the meaning of the data analyzed and process them specifically. For instance, by interpreting the fact that some data belong to a specific class might lead

to the use of a specific probability distribution to describe them or to a particular weighing of the observations considered in our estimates. Moreover, if two categorical data belong to sibling classes, then we can take advantage of such dependencies in our predictions. In general the potential is still big, because so far in our computations we make a prevalent use of statistics to obtain trust estimates, but by properly combining it with logics and semantics, we might be able to further improve the results achieved and to fully take advantage of all the information at our disposal. For instance, the use of subjective logic has mostly been limited to its ability to represent contextual trust values and its statistical representation of trust values. Only in a few cases we employ the logics of subjective logic, and this technique can be further applied and better tightened with ontological reasoning. By moving our focus to the semantics aspects, we might take full advantage of the huge volume of data available in the Linked Open Data cloud (similar to what we do in Chapter 1) and start evaluating the data not only based on their metadata, but also considering their meaning.

Semantics alone is not sufficient to show whether a given piece of data is correct or not, because, for instance, although a semantic interpretation can help to tell whether a given tag is correctly spelled and used in the right context, semantics alone can hardly tell if the annotation is, for instance, correctly linked to the right specimen. However, the combination of this knowledge with the techniques already presented in this thesis is promising and will be further investigated.

3. Social and Psychological Aspects In this thesis we focus on the analysis of the data themselves: where they come from, how they are produced, what is their expected representativity, etc., and all these elements are considered when evaluating their trustworthiness. However, there is a human factor in all this that can become relevant, for two reasons. First, by properly analyzing the social interactions existing between people involved in the process of data creation we can improve the trust estimation, like Golbeck suggests [65]. Considering the social implications of trust can be done by reusing at least some of the techniques already adopted: for instance, in this context, subjective logic has already been explored by Bakar et al. [5]. At the same time, the uncertainty reasoning techniques used in this thesis need to be combined with social network analysis techniques in order to properly tackle this interesting angle. Moreover, by considering more the social aspect in our analysis, security needs to be taken into account as well. Our definition of trust does not need to consider the reasons that make a piece of data untrustworthy: whether it is because of malicious intentions or because of lack of competence of its author, that neither is known to us, nor is the focus of our research. However, by moving in this direction, we must also consider the incentives that move people, and study them deeply in order to improve the accuracy of our predictions and avoid us to misinterpret our observations.

There is also another aspect that is worth considering when looking at the human aspects of trust estimation of semi-structured data. Social analyses give another means to understand the data trustworthiness, but it might be useful and interesting also to consider how people interpret the trustworthiness of these data. We adopt one specific trust definition but, although there are also other definitions of trust, and

different people (or entities) with different constraints and goals in mind can implement such a definition in different manners. For instance, in Chapters 7 and 8, we take the point of view of specific institutions, learnt from their evaluations and built a model from a sample of evaluations. Different institutions and different people might apply different evaluations strategies: some consistent and some not, etc. In order to facilitate the adoption of the methods and techniques developed in this thesis, it would be useful to evaluate their adaptability in different contexts in order to measure also the trustworthiness of the trust management systems themselves and their ability to really understand user requirements in terms of trustworthiness.

Pariser [122] describes the problem that exacerbated personalization prevents people from accessing contents that they might enjoy but that are not presented to them because they are too far from their known profiles. Here, although the intentions are different, we run the risk to encounter a similar problem. If some classes of content are mistakenly considered to be untrustworthy or if some sources are erroneously evaluated as more trustworthy than they actually are, we might stumble on a problem similar to *The Filter Bubble*: user overexposure to content from limited sources, at the expense of other, potentially useful content. Here recommender systems might help: although these are aimed at a different goal than trust management systems, the latter ones might borrow from recommender systems methods that are necessary to employ to prevent filter bubbles.

4. Trust Representation We employ the RDF Data Cube and the Open Annotation ontologies to represent trust values. Our goal is merely to store the results of our computations for further use. However, considering a possible extension of this research in the social domain, the trust representation might become even more important. We represent the trust values we compute by means of subjective opinions because we find them simple and complete, but by moving to the social domain, this representation may be sided by others and, in general, the sharing of these results might become more important. Webs of trust have already been widely studied, but a more data-oriented (instead of user-oriented) sharing of subjective opinions by means of Semantic Web technologies may be an interesting direction to take. By sharing subjective opinions about different facts we might allow reusing them and, hence, reduce the dependence from observation availability in our computations, because it allows reusing a second-hand opinion instead of looking for first-hand evidence. Of course, this requires the ability to properly weigh the opinions that one might retrieve from a social network, as well as a deep study of the requirements for representing subjective opinions in such a scenario: RDF Data Cube and Open Annotation might be ideal tools for accomplishing this task, but it is important to validate this aspect to avoid pitfalls or mistakes. A standardization of the trustworthiness and trust representation is desirable and such a representation should be, in our opinion, lightweight, complete and precise. URIs are the ideal tool to represent the elements involved in the expression of trust (trustor, trustee, context) in order to make the trust evaluation reusable and interpretable by other entities. Research needs to be done about the best means to represent the trust evaluations, to balance the need for them to be easily understood

but not bound to a specific calculation method. Moreover, we propose to handle these shareable trust statements at RDF level in order to allow sharing subjective trust assessments about single RDF triples. A given triple may be believed to be true by one or more sources, based on evidence at their disposal. This would not affect the semantics of the triple itself, rather, the fact that a possibly unreliable proxy refers us about its trustworthiness, and we rely on it, with some possible precautions. Having the possibility of encoding this information with an extended version of RDF would increase the compactness and the portability of the trust representation. However, this probably requires a deep research effort and, in the short run, solutions based on the Open Annotation model and RDF Data Cube are a good compromise. Hence, a first standardization round needs probably to start from these vocabularies and, in the long run, hopefully a more compact and precise representation will be possible.

5. Security Although trust is tightly connected with security issues, we do not tackle this aspect in this thesis. We do not investigate the motivations that lead actors involved in the process of data creation to the production of bad data, when it happens. Whether it is because an actor does not hold the necessary expertise or knowledge, because of chance, or because of malicious intentions, we do not treat the error differently, because we can not validate any result about this. The only thing we can do is to check and predict the data trustworthiness per se, and this is what we focus on. However, as soon as we extend our focus to social and psychological aspects of trust, security issues become more relevant, because when people are incentivized to be more involved in the system, then the possible gain from a system deception rises, as highlighted several times in the volume of Hasum and Tovey [109]. So, in that perspective, the incentives and the motivations that pull people to produce and share data and trust assessments become more important and need to be researched. This will tighten the link with computer security, or simply pose more attention to the security aspects of trust. As a consequence, more attention will have to be paid to the robustness of the trust management techniques, since these will have to resist security attacks from malicious users. Also, in this context, the trust management mechanism becomes part of an incentive system. Therefore, it will not only have to correctly estimate the trustworthiness of the data and resist the users attack, but also help to avoid such attacks by correctly incentivizing the users.

Looking Ahead

The importance of trust management is increasingly being considered, as testified, for instance, by the recent developments of the IFIP Trust Management Working Group. Also, since the beginning of the work presented in this thesis, the provenance models evolved from OPM to the recent W3C recommendation PROV Ontology. This work tries to bridge the gaps between these fields. Moreover, the use of semantic similarity enables more free text-oriented applications of this work. Natural language processing is not a trivial task per se, but there may be interesting parallels between the methods

used here to analyze provenance and possible natural language processing analyses for trust estimation. So, there is still a long way ahead, but the work developed here hopefully poses useful indications about the directions to pursue.

To conclude, this work has focused on the estimation of the trustworthiness of semi-structured Web data. With a growing amount of data being created and dispatched every day, and with the growing complexity of our society, the ability to correctly estimate the trustworthiness of data (and implicitly, of people) will help to improve the society we belong to. As we have seen, this road is still full of challenges, and some are also yet to be faced. In some situations, people consciousness about these problems is still limited, in other situations the methods for coping with these threats are limited and complex. Taking these issues seriously will lead to a better use of our resources, to an elimination of the inefficiencies due to a lack of trust and, therefore, to a lot of potential advances that will beneficial to us all.

Bibliography

[1] Agile Knowledge and Semantic Web research group (AKSW). Triplify.org. `http://triplify.org`, 2010.

[2] H. Alan. Probability, Logic, and Probability Logic. In *The Blackwell Guide to Philosophical Logic*, pages 362–384. Oxford: Blackwell, 2001.

[3] L. Aroyo and C. Welty. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *WebSci '13: Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 2013.

[4] D. Artz and Y. Gil. A survey of trust in computer science and the Semantic Web. *Journal of Web Semantics*, 5(2):58–71, 2007.

[5] A. A. Bakar, R. Ismail, A. R. Ahmad, and J.-L. A. Manan. Trust Formation Based on Subjective Logic and PGP Web-of-Trust for Information Sharing in Mobile Ad Hoc Networks. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing (SOCIALCOM 2010)*, pages 1004–1009. IEEE Computer Society, 2010.

[6] R. S. Beaman and B. J. Conn. Automated geoparsing and georeferencing of malesian collection locality data. *Telopea*, 10(1):43–52, 2003.

[7] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. OWL Reference. Technical report, W3C, 2004. W3C Recommendation.

[8] G. Begelman. Automated Tag Clustering: Improving search and exploration in the tag space. In *Proceedings of Collaborative Web Tagging Workshop, co-located with the 15th International World Wide Web Conference (WWW 2006)*, 2006.

[9] K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. PROV-O: The PROV Ontology. `http://www.w3.org/TR/prov-o`, 2012. W3C Recommendation.

[10] BioGeomancer Working Group. BioGeomancer Metadata. `http://www.biogeomancer.org/metadata.html`, 2013.

[11] R. Bivand, T. Keitt, B. Rowlingson, E. Pebesma, M. Sumner, and R. Hijmans. *RGDAL - R Geospatial Data Abstraction Library*, 2010. `https://r-forge.r-project.org/projects/rgdal`.

[12] C. Bizer and R. Cyganiak. Quality-driven information filtering using the WIQA policy framework. *Journal of Web Semantics*, 7(1):1–10, 2009.

[13] C. Bizer and R. Cyganiak. The D2RQ Plattform - Treating Non-RDF Databases as Virtual RDF Graphs. `http://www4.wiwiss.fu-berlin.de/bizer/d2rq`, 2010.

[14] S. Bradshaw, D. Brickley, L. J. G. Castro, T. Clark, T. Cole, P. Desenne, A. Gerber, A. Isaac, J. Jett, T. Habing, B. Haslhofer, S. Hellmann, J. Hunter, R. Leeds, A. Magliozzi, B. Morris, P. Morris, J. van Ossenbruggen, S. Soiland-Reyes, J. Smith, and D. Whaley. Open Annotation Core Data Model. `http://www.openannotation.org/spec/core`, 2012. W3C Community Draft.

[15] D. Brickley and L. Miller. FOAF Vocabulary Specification 0.97. `http://xmlns.com/foaf/spec/20100101.html`, 2013.

[16] L. J. Camp. Designing for Trust. In *Proceedings of Trust, Reputation, and Security: Theories and Practice, International Workshop, co-located with the First International Joint Conference on Autonomous Agents & Multiagent Systems (AAMAS 2002)*, volume 2631 of *Lecture Notes in Computer Science*, pages 15–29. Springer, 2002.

[17] S. Card, T. P. Moran, and A. Newell. *The Psychology of Human Computer Interaction*. Lawrence Erlbaum Associates, 1983.

[18] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Named Graphs, Provenance and Trust. In *Proceedings of the 14th International Conference on World Wide Web (WWW 2005)*, pages 613–622. ACM, 2005.

[19] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In *Proceedings of the 7th International Semantic Web Conference (ISWC 2007)*, volume 4825 of *Lecture Notes in Compter Science*, pages 615–631. Springer, 2008.

[20] D. Ceolin. Trusting Semi-structured Web Data. In *Proceedings of The Semantic Web: Semantics and Big Data, 10th International Conference (ESWC 2013)*, volume 8219 of *Lecture Notes in Computer Science*, pages 676–681. Springer, 2013.

[21] D. Ceolin, P. Groth, and W. R. van Hage. Calculating the Trust of Event Descriptions using Provenance. In *Proceedings of the Second International Workshop on the role of Semantic Web in Provenance Management (SWPM 2010), co-located with the 9th International Semantic Web Conference (ISWC 2010)*, volume 670 of *CEUR Workshop Proceedings*, pages 11–16. CEUR-WS.org, 2010.

[22] D. Ceolin, P. Groth, W. R. van Hage, A. Nottamkandath, and W. Fokkink. Trust Evaluation through User Reputation and Provenance Analysis. In *Proceedings of the 8th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2012) at the 11th International Semantic Web Conference (ISWC 2012)*, volume 900 of *CEUR Workshop Proceedings*, pages 15–26. CEUR-WS.org, 2012.

[23] D. Ceolin, L. Moreau, K. O'Hara, G. Schreiber, A. Sackley, W. Fokkink, W. R. van Hage, and N. Shadbolt. Reliability Analyses of Open Government Data. In *Proceedings of the 9th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2013), co-located with the 12th International Semantic Web Conference (ISWC 2013)*, volume 1073 of *CEUR Workshop Proceedings*, pages 34–39. CEUR-WS.org, 2013.

[24] D. Ceolin, L. Moreau, K. O'Hara, G. Schreiber, A. Sackley, W. Fokkink, W. R. van Hage, N. Shadbolt, and V. Maccatrozzo. Two Procedures for Analyzing the Reliability of Open Government Data. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2014)*. Springer, 2014. To appear.

[25] D. Ceolin, A. Nottamkandath, and W. Fokkink. Automated Evaluation of Annotators for Museum Collections Using Subjective Logic. In *Proceedings of Trust Management VI - 6th IFIP WG 11.11 International Conference (IFIPTM 2012)*, volume 374 of *IFIP Advances in Information and Communication Technology*, pages 232–239. Springer, 2012.

[26] D. Ceolin, A. Nottamkandath, and W. Fokkink. Subjective Logic Extensions for the Semantic Web. In *Proceedings of the 8th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2012), co-located with the 11th International Semantic Web Conference (ISWC 2012)*, volume 900 of *CEUR Workshop Proceedings*, pages 27–38. CEUR-WS.org, 2012.

[27] D. Ceolin, A. Nottamkandath, and W. Fokkink. Semi-automated Assessment of Annotation Trustworthiness. In *Proceedings of the Eleventh Annual International Conference on Privacy, Security and Trust (PST 2013)*, pages 325–332. IEEE Computer Society, 2013.

[28] D. Ceolin, A. Nottamkandath, and W. Fokkink. Efficient Semi-automated Assessment of Annotations Trustworthiness. *Journal of Trust Management*, 2014. To appear.

[29] D. Ceolin, W. R. van Hage, and W. Fokkink. A Trust Model to Estimate the Quality of Annotations using the Web. In *WebSci10: Extending the Frontiers of Society On-Line (WebSci 2010)*. Web Science Trust, 2010.

[30] D. Ceolin, W. R. van Hage, W. Fokkink, and G. Schreiber. Estimating Uncertainty of Categorical Web Data. In *Proceedings of the 7th International Workshop*

on Uncertainty Reasoning for the Semantic Web (URSW 2011), co-located with the 10th International Semantic Web Conference (ISWC 2011), volume 778 of *CEUR Workshop Proceedings*, pages 15–26. CEUR-WS.org, 2011.

[31] D. Ceolin, W. R. van Hage, G. Schreiber, and W. Fokkink. Assessing Trust for Determining the Reliability of Information. In *Situation Awareness with Systems of Systems*, pages 209–228. Springer, 2013.

[32] P. Ceravolo, E. Damiani, and C. Fugazza. Trustworthiness-related Uncertainty of Semantic Web-style Metadata: A Possibilistic Approach. In *Proceedings of the 3rd Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2007), co-located with the 6th International Semantic Web Conference (ISWC 2007)*, volume 327 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.

[33] R. Cilibrasi and P. M. B. Vitányi. Automatic Meaning Discovery Using Google. In *Kolmogorov Complexity and Applications*, number 06051 in Dagstuhl Seminar Proceedings. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2006.

[34] R. L. Cilibrasi and P. M. B. Vitanyi. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.

[35] CiteULike. CiteULike. `http://www.citeulike.org`, 2012.

[36] CityMapHQ.com. CityMapHQ. `http://www.citymaphq.com/codes/itu.html`, 2013.

[37] R. Cornelli. *Why people trust the police. An empirical study.* PhD thesis, Università degli Studi di Trento, International Ph.D. in Criminology, 2003.

[38] C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[39] CrimeReports. CrimeReports. `https://www.crimereports.co.uk`, 2013.

[40] M. Davy and J.-Y. Tourneret. Generative Supervised Classification Using Dirichlet Process Priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1781–1794, 2010.

[41] G. De la Calzada and A. Dekhtyar. On Measuring the Quality of Wikipedia Articles. In *Proceedings of the 4th Workshop on Information Credibility (WICOW 2010), co-located with the 19th International World Wide Web Conference*, pages 11–18. ACM, 2010.

[42] G. Demartini. Finding Experts Using Wikipedia. In *Proceedings of the 2nd International ISWC+ASWC Workshop on Finding Experts on the Web with Semantics (FEWS 2007), co-located with the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference*, volume 290 of *CEUR Workshop Proceedings*, pages 33–41. Springer, 2007.

[43] A. P. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of the Mathematical Statistics*, 2(38):325–339, 1967.

[44] D.-P. Deng, T.-R. Chuang, K.-T. Shao, G.-S. Mai, T.-E. Lin, R. Lemmens, C.-H. Hsu, H.-H. Lin, and M.-J. Kraak. Using Social Media for Collaborative Species Identification and Occurrence: Issues, Methods, and Tools. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information 2012 (GEOCROWD 2012), co-located at the 20th ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL 2012)*, pages 22–29. ACM, 2012.

[45] P. Dirac. *Principles of Quantum Mechanics*. Oxford at the Clarendon Press, 1958.

[46] Dublin Core Metadata Initiative. DCMI Metadata Terms. `http://dublincore.org/documents/dcmi-terms`, 2012.

[47] S. T. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230, 2004.

[48] D. Eads. scipy-cluster: An extension to Scipy for generating, visualizing, and analyzing hierarchical clusters. `http://scipy-cluster.googlecode.com`, 2008.

[49] M. Ebden, T. D. Huynh, L. Moreau, S. Ramchurn, and R. Stephen. Network Analysis on Provenance Graphs from a Crowdsourcing Application. In *Proceedings of the 4th International Provenance and Annotation Workshop (IPAW 2012)*, volume 7525 of *Lecture Notes in Computer Science*, pages 168–182. Springer, 2012.

[50] C. Elkan. Clustering Documents with an Exponential-Family Approximation of the Dirichlet Compound Multinomial Distribution. In *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, volume 148 of *ACM International Conference Proceeding Series*, pages 289–296. ACM, 2006.

[51] A. Ellis, D. Gluckman, A. Cooper, and A. Greg. Your Paintings: A Nation's Oil Paintings Go Online, Tagged by the Public. In *Museums and the Web 2012*. Museums and the Web LLC., 2012.

[52] A. Erling. Sufficiency and Exponential Families for Discrete Sample Spaces. *Journal of the American Statistical Association*, 65:1248–1255, 1970.

[53] M. D. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

[54] J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer, 2007.

[55] T. S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[56] D. Fink. A Compendium of Conjugate Priors. Technical report, Montana State University, 1997.

[57] A. Fokoue, M. Srivatsa, and R. Young. Assessing Trust in Uncertain Information. In *Proceedings of the 9th International Semantic Web Conference (ISWC 2010)*, volume 6496 of *Lecture Notes in Computer Science*, pages 209–224. Springer, 2010.

[58] F. Galton. Regression Towards Mediocrity in Hereditary Stature. *Journal of the Anthropological Institute*, 15:246–263, 1886.

[59] D. Gambetta. *Can We Trust Trust?* Basil Blackwell, 1988.

[60] GeoNames. GeoNames Ontology. `http://www.geonames.org`, 2013.

[61] Global Biodiversity Information Facility. GBIF. `http://ontologi.es/biol/zoology`, 2013.

[62] J. Godoy, J. Atkinson, and A. Rodriguez. Geo-referencing with semi-automatic gazetteer expansion using lexico-syntactical patterns and co-reference analysis. *International Journal of Geographical Information Science*, 25(1):149–170, 2011.

[63] J. Golbeck. Combining Provenance with Trust in Social Networks for Semantic Web Content Filtering. In *Proceedings of the International Provenance and Annotation Workshop (IPAW 2006)*, volume 4145 of *Lecture Notes in Computer Science*, pages 101–108. Springer, 2006.

[64] J. Golbeck. Trust on the World Wide Web: A Survey. *Foundations and Trends in Web Science*, 1(2):131–197, 2006.

[65] J. Golbeck. Weaving a Web of Trust. *Science*, 321(5896):1640–1641, 2008.

[66] J. Golbeck and J. Hendler. Accuracy of Metrics for Inferring Trust and Reputation in Semantic Web-Based Social Networks. In *Proceedings of the 14th Engineering Knowledge in the Age of the Semantic Web Conference (EKAW 2004)*, volume 3257 of *Lecture Notes in Computer Science*, pages 116–131. Springer, 2004.

[67] Google. Google Maps. `http://maps.google.com`, 2013.

[68] J. C. Gower and G. J. S. Ross. Minimum Spanning Trees and Single Linkage Cluster Analysis. *Journal of the Royal Statistical Society*, 18(1):54–64, 1969.

[69] C. H. Graham, J. Elith, R. J. Hijmans, A. Guisan, T. Peterson, and B. A. Loiselle. The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, 45(1):239–247, 2008.

[70] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of Trust and Distrust. In *Proceedings of the 13th International Conference on World Wide Web (WWW 2004)*, pages 403–412. ACM, 2004.

[71] Q. Guo, Y. Liu, and J. Wieczorek. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, 22(10):1067–1090, 2008.

[72] R. P. Guralnick, J. Wieczorek, R. Beaman, and R. J. Hijmans. BioGeomancer: Automated Georeferencing to Map the World's Biodiversity Data. *PLoS Biology*, 4(11):1908–1909, 2006.

[73] O. Hartig and J. Zhao. Using Web Data Provenance for Quality Assessment. In *Proceedings of the First International Workshop on the role of Semantic Web in Provenance Management (SWPM 2009), co-located with the 8th International Semantic Web Conference (ISWC 2009)*, volume 526 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.

[74] Y. Hassan-Montero and V. Herrero-Solana. Improving Tag-Clouds as Visual Information Retrieval Interfaces. In *Proceedings of the International Conference on Multidisciplinary Information Sciences and Technologies (INSCIT 2006)*. ACL, 2006.

[75] J. Hilgevoord and J. Uffink. Uncertainty in prediction and in inference. *Foundations of Physics*, 21:323–341, 1991.

[76] Human Inference. DataCleaner. `http://datacleaner.org`, 2013.

[77] H. Ibrahim, P. K. Atrey, and A. El Saddik. Semantic Similarity Based Trust Computation in Websites. In *Workshop on Many Faces of Multimedia Semantics (MS 2007)*, pages 65–72. ACM, 2007.

[78] O. Inel, L. Aroyo, C. Welty, and R.-J. Sips. Domain-independent quality measures for crowd truth disagreement. In *Proceedings of the 3rd International Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2013) co-located with 12th International Semantic Web Conference (ISWC 2013)*, CEUR Workshop Proceedings. CEUR-WS.org, 2013.

[79] T. Inkster. Biological Taxonomy Vocabulary 0.2 (Zoölogy). `http://ontologi.es/biol/zoology`, 2013.

[80] International Chamber of Commerce. ICC Commercial Crime Services. `http://icc-ccs.org`, 2013.

[81] International Telecommunication Union. Table of Maritime Identification Digits. `http://www.itu.int/online/mms/glad/cga_mids.sh?lng=E`, 2013.

[82] S. Javanmardi, C. Lopes, and P. Baldi. Modeling user reputation in wikis. *Statistical Analysis and Data Mining*, 3(2):126–139, 2010.

[83] A. Jøsang. A Logic for Uncertain Probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–212, 2001.

[84] A. Jøsang and T. Bhuiyan. Optimal Trust Network Analysis with Subjective Logic. In *Proceedings of the Second International Conference on Emerging Security Information, Systems and Technologies (SECURWARE 2008)*, pages 179–184. IEEE Computer Society, 2008.

[85] A. Jøsang, M. Daniel, and P. Vannoorenberghe. Strategies for Combining Conflicting Dogmatic Beliefs. In *Proceedings of the 6th International Conference on Information Fusion (FUSION 2003)*. IEEE Computer Society, 2003.

[86] A. Jøsang, J. Diaz, and M. Rifqi. Cumulative and Averaging Fusion of Beliefs. *Information Fusion*, 11(2):192–200, 2010.

[87] A. Jøsang and R. Hankin. Interpretation and Fusion of Hyper Opinions in Subjective Logic. In *Proceedings of 15th IEEE International Conference on Information Fusion (FUSION 2012)*. IEEE Computer Society, 2012.

[88] A. Jøsang and D. McAnally. Multiplication and comultiplication of beliefs. *International Journal of Approximate Reasoning*, 38(1):19–51, 2005.

[89] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The Eigentrust Algorithm for Reputation Management in P2P Networks. In *Proceedings of the 12th International Conference on World Wide Web (WWW 2003)*, pages 640–651. ACM, 2003.

[90] L. Kaplan, S. Chakraborty, and C. Bisdikian. Subjective Logic with Uncertain Partial Observations. In *15th International Conference on Information Fusion (FUSION 2012)*, pages 565–572. IEEE Computer Society, 2012.

[91] R. Karam and M. Melchiori. Improving Geo-spatial Linked Data with the Wisdom of the Crowds. In *Proceedings of the 3RD International Workshop on Linked Web Data Management (LWDM 2013), co-located with the 16th International Conference on Extending Database Technology (EDBT 2013)*, pages 68–74. ACM, 2013.

[92] T. Kauppinen, R. Henriksson, R. Sinkkilä, R. Lindroos, J. Väätäinen, and E. Hyvönen. Ontology-based Disambiguation of Spatiotemporal Locations. In *Proceedings of the 1st International Workshop on Identity and Reference on the Semantic Web (IRSW2008), co-located with the 5th European Semantic Web Conference (ESWC 2008)*, volume 422 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[93] R. Killick and I. A. Eckley. *changepoint: An R Package for Changepoint Analysis*, 2013. `http://cran.r-project.org/package=changepoint`.

[94] I. N. Koch-Weser. The Reliability of China's Economic Data: An Analysis of National Output. `http://www.uscc.gov/sites/default/files/Research/TheReliabilityofChina'sEconomicData.pdf`, 2013.

[95] I. Kononenko. Naive Bayesian Classifier and Continuous Attributes. *Informatica*, 16(1):1–8, 1992.

[96] E. F. Krause. *Taxicab Geometry*. Dover, 1987.

[97] P. Kvam and D. Day. The Multivariate Polya Distribution in Combat Modeling. *Naval Research Logistics (NRL)*, 48(1):1–17, 2001.

[98] L. Accardi (originator). De Finetti Theorem. In *Encyclopedia of Mathematics*. Springer, 2001.

[99] R. Lange and X. Lange. Quality control in crowdsourcing: An objective measurement approach to identifying and correcting rater effects in the social evaluation of products and services. In *AAAI Spring Symposium: Wisdom of the Crowd*, volume SS-12-06 of *AAAI Technical Report*. AAAI Press, 2012.

[100] J. L. Leidner and M. D. Lieberman. Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. *SIGSPATIAL Special*, 3(2):5–11, 2011.

[101] J. L. Leidner, G. Sinclair, and B. Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, volume 1 of *HLT-NAACL-GEOREF'03*, pages 31–38. ACL, 2003.

[102] M. H. R. Leyssen, M. C. Traub, J. R. van Ossenbruggen, and L. Hardman. Is It A Bird Or Is It A Crow? The Influence Of Presented Tags On Image Tagging By Non-Expert Users. CWI Tech. Report INS-1202, CWI, 2012.

[103] H. Li, R. K. Srihari, C. Niu, and W. Li. InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References (HLT-NAACL-GEOREF'03)*, pages 39–44. ACL, 2003.

[104] D. Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th International Conference on Machine Learning (ICML 1998)*, pages 296–304. Morgan Kaufmann Publishers Inc., 1998.

[105] V. Loureiro, I. Anastácio, and B. Martins. Learning to Resolve Geographical and Temporal References in Text. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM-GIS 2011)*, pages 349–352. ACM, 2011.

[106] R. E. Madsen, D. Kauchak, and C. Elkan. Modeling Word Burstiness Using the Dirichlet Distribution. In *Proceedings of the 22nd International Conference on Machine Learning (ICML 2005)*, volume 119 of *ACM International Conference Proceeding Series*, pages 545–552. ACM, 2005.

[107] Mapit. Mapit. http://mapit.mysociety.orgs, 2013.

[108] P. Massa and P. Avesani. Trust Metrics in Recommender Systems. In *Computing with Social Trust*, Human Computer Interaction Series, pages 259–285. Springer, 2009.

[109] H. Masum and M. Tovey, editors. *The Reputation Society.* MIT Press, 2012.

[110] O. Medelyan, E. Frank, and I. H. Witten. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 1318–1327. ACL, 2009.

[111] S. Miles, C. M. Trim, and M. Panzer. Dublin Core to PROV Mapping. http://www.w3.org/TR/2012/WD-prov-dc-20121211, 2012. W3C Working Draft.

[112] G. A. Miller. WordNet: a Lexical Database for English. *Commununications ACM*, 38(11):39–41, 1995.

[113] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, and J. Myers. The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems*, 2010.

[114] P. C. Murphey, R. P. Guralnick, R. Glaubitz, D. Neufeld, and J. A. Ryan. Georeferencing of museum collections: A Review of the problems and automated tools, and the methodology developed by the Mountain and Plains Spatio-Temporal Database-informatics initiative (MaPSTeDI). *Phyloinformatics: Journal for Taxonomists*, 1:1–29, 2004.

[115] National Biological Information Infrastructure. NBII. https://www.nbii.gov, 2010.

[116] Naturalis. Naturalis Biodiveristy Center. https://www.naturalis.nl, 2013.

[117] R. M. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and graphical statistics*, 9(2):249–265, 2000.

[118] Netherlands Biodiversity Information Facility. Netherlands biodiversity information facility. https://www.nlbif.nl, 2013.

[119] Netherlands Institute for Sound and Vision. Waisda? http://waisda.nl, 2012.

[120] N. J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28(1):71–88, 1986.

[121] K. O'Hara. A General Definition of Trust. Technical report, University of Southampton, 2012.

[122] E. Pariser. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin Press, 2011.

[123] K. Pearson. Mathematical Contributions to the Theory of Evolution. In *Proceedings of the Royal Society of London*, volume 60, pages 489–498. Royal Society Publishing, 1896.

[124] K. Pearson. On the Criterion that a Given System of Deviations from the Probable in the Case of Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling. *Philosophical Magazine*, 50:157–175, 1900.

[125] J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995.

[126] S. A. Pop. Geo::Calc - simple geo calculator for points and distances. `http://search.cpan.org/~asp/Geo-Calc-0.12/lib/Geo/Calc.pm`, 2013.

[127] Princeton University. Wordnet::Similarity Web Service. `http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi`, 2012.

[128] M. Proffitt, editor. *The Oxford English Dictionary*. Oxford University Press., 2 edition, 1989. Trust.

[129] M. Proffitt, editor. *The Oxford English Dictionary*. Oxford University Press., 2 edition, 1989. Reputation.

[130] S. Rajbhandari, O. F. Rana, and I. Wootten. A Fuzzy Model for Calculating Workflow Trust using Provenance Data. In *Proceedings of the 15th ACM Mardi Gras Conference: From lightweight mash-ups to lambda grids: Understanding the spectrum of distributed computing requirements, applications, tools, infrastructures, interoperability, and the incremental adoption of key capabilities (MG 2008)*, ACM International Conference Proceeding Series. ACM, 2008.

[131] S. Rajbhandari, I. Wootten, A. S. Ali, and O. F. Rana. Evaluating Provenance-based Trust for Scientific Workflows. In *Proceeding of the Sixth IEEE International Symposium on Cluster Computing and the Grid (CCGRID 2006)*, volume 1, pages 365–372. IEEE Computer Society, 2006.

[132] C. E. Rasmussen. The Infinite Gaussian Mixture Model. In *Advances in Neural Information Processing Systems 12 (NIPS 1999)*, pages 554–560. MIT Press, 1999.

[133] RDFLib Team. Python library RDFLib. `http://www.rdflib.net`, 2013.

[134] M. Richardson, R. Agrawal, and P. Domingos. Trust Management for the Semantic Web. In *Proceeding of the Second International Semantic Web Conference (ISWC 2003)*, volume 2870 of *Lecture Notes in Computer Science*, pages 351–368. Springer, 2003.

[135] Rijksmuseum Amsterdam. Rijksmuseum. `https://www.rijksmuseum.nl`, 2013.

[136] N. E. Rios and H. L. J. Bart. GEOLocate (Version 3.22). `http://www.museum.tulane.edu/geolocate`, 2010. Belle Chasse, LA: Tulane University Museum of Natural History.

[137] K. Roberts, C. A. Bejan, and S. M. Harabagiu. Toponym Disambiguation Using Events. In *Proceedings of the Twenty-Third International FLAIRS Conference (FLAIRS-10)*. AAAI Press, 2010.

[138] A. Rodriguez, D. B. Dunson, and A. E. Gelfand. The Nested Dirichlet Process. *Journal of the American Statistical Association*, 103(483):1131–1144, 2008.

[139] J. Sabater and C. Sierra. Review on Computational Trust and Reputation Models. *Artificial Intelligence Review*, 24(1):33, 2005.

[140] S. Sahoo, P. Groth, O. Hartig, S. Miles, S. Coppens, J. Myers, Y. Gil, L. Moreau, J. Zhao, M. Panzer, and D. Garijo. Provenance Vocabulary Mappings. W3C Working Draft, W3C, 2010.

[141] S. S. Sahoo, A. Sheth, and C. Henson. Semantic provenance for eScience: Managing the deluge of scientific data. *IEEE Internet Computing*, 12(4):46–54, 2008.

[142] M. Sensoy, J. Z. Pan, A. Fokoue, M. Srivatsa, and F. Meneguzzi. Using Subjective Logic to Handle Uncertainty and Conflicts. In *Proceedings of 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-12)*, pages 1323–1326. IEEE Computer Society, 2012.

[143] G. Shafer. *A mathematical Theory of Evidence*. Princeton University Press, 1976.

[144] ShipAis. ShipAis. `http://www.shipais.com`, 2013.

[145] G. Soberón, L. Aroyo, C. Welty, O. Inel, H. Lin, and M. Overmeen. Measuring crowd truth: Disagreement metrics combined with worker behavior filters. In *Proceedings of the 1st International Workshop on Crowdsourcing the Semantic Web (CrowdSem 2013) co-located with 12th International Semantic Web Conference (ISWC 2013)*, volume 1030 of *CEUR Workshop Proceedings*, pages 45–58. CEUR-WS.org, 2013.

[146] C. Spearman. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology*, 15:88–103, 1904.

[147] J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Anchor, 2004.

[148] Talend. Talend Open Studio for Data Quality. `http://www.talend.com/products/data-quality`, 2013.

[149] M. Tavakolifard, P. Herrmann, and S. J. Knapskog. Inferring Trust Based on Similarity with TILLIT. In *Proceeding of the Third IFIP WG 11.11 International Conference on Trust Management (IFIPTM)*, volume 300 of *IFIP Advances in Information and Communication Technology*, pages 133–148. Springer, 2009.

[150] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2005.

[151] The Open Data Institute. The Open Data Institute. `www.theodi.org`, 2013.

[152] K. Thirunarayan, P. Anantharam, C. Henson, and A. Sheth. Comparative trust management with applications: Bayesian approaches emphasis. *Future Generation Computer Systems*, 31:182–199, 2014.

[153] Tom Heath. Linked data - Connect Distributed Data across the Web. `http://linkeddata.org`, 2010.

[154] United Kingdom Police Home Office. data.police.uk. `data.police.uk`, 2013.

[155] United States Institute of Museum and Library Service. Steve Social Tagging Project. `http://www.steve.museum`, 2012.

[156] A. Ushioda. Hierarchical Clustering of Words and Application to NLP Tasks. In *Proceedings of the Fourth Workshop on Very Large Corpora (WVLC-4)*, pages 28–41. ACL, 1996.

[157] C. van Damme, M. Hepp, and T. Coenen. Quality Metrics for Tags of Broad Folksonomies. In *Proceedings of the International Conference on Semantic Systems (I-SEMANTICS 2008)*. Journal of Universal Computer Science, 2008.

[158] M. van Erp. Retrieving Lost Information from Textual Databases: Rediscovering Expeditions from an Animal Specimen Database. In *Proceedings of the ACL 2007 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), co-located with the 45th Annual Meeting of the Association for Computational Linguistics*, LaTeCH 2007. ACL, 2007.

[159] M. van Erp, R. Hensel, D. Ceolin, and M. van der Meij. Georeferencing Animal Specimen Datasets. Accepted for publication in Transactions in GIS, John Wiley & Sons, Inc., 2014.

[160] W. R. van Hage and D. Ceolin. The Simple Event Model. In *Situation Awareness with Systems of Systems*, pages 149–169. Springer, 2013.

[161] W. R. van Hage, V. Malaisé, G. de Vries, G. Schreiber, and M. van Someren. Combining Ship Trajectories and Semantics with the Simple Event Model (SEM). In *Proceedings of the 1st ACM International Workshop on Events in Multimedia (EiMM '09), co-located with the 17th International Conference on Multimedia (ACM Multimedia 2009)*, EiMM '09, pages 73–80. ACM, 2009.

[162] W. R. van Hage, V. Malaisé, R. Segers, and L. Hollink. Design and Use of the Simple Event Model (SEM). *Journal of Web Semantics*, 9(2):128–136, 2011.

[163] W. R. van Hage, V. Malaisé, and M. van Erp. Linked Open Piracy: A Story about e-Science, Linked Data, and Statistics. *Journal of Data Semantics*, 2012.

[164] Vesseltracker GMBH. Vesseltracker. `http://www.vesseltracker.com`, 2013.

[165] L. von Ahn and L. Dabbish. Labeling Images with a Computer Game. In *Proceedings of the 2004 Conference on Human Factors in Computing Systems (CHI 2004)*, pages 319–326. ACM, 2004.

[166] P. Vossen, K. Hofmann, M. de Rijke, E. T. K. Sang, and K. Deschacht. The Cornetto Database: Architecture and User-Scenarios. In *Proceedings of the 7th Dutch-Belgian Information Retrieval Workshop (DIR 2007)*, pages 89–96. K.U. Leuven's professional press and publishing cooperation ACCO, 2007.

[167] L.-H. Vu and K. Aberer. Effective Usage of Computational Trust Models in Rational Environments. In *Proceedings of the IEEE / WIC / ACM International Conference on Web Intelligence (WI 2008)*, pages 583–586. IEEE Computer Society, 2008.

[168] W3C. RDF Vocabulary Description Language 1.0: RDF Schema. `http://www.w3.org/TR/rdf-schema`, 2004. W3C Recommendation.

[169] W3C. SPARQL Query Language for RDF. Technical report, W3C, 2008. W3C Recommendation.

[170] W3C. SKOS Simple Knowledge Organization System Reference. `http://www.w3.org/TR/2009/REC-skos-reference-20090818`, 2009. W3C Recommendation.

[171] W3C. Resource Definition Framework. `http://www.w3.org/RDF`, 2011. W3C Recommendation.

[172] S. Wang and M. Iwaihara. Quality Evaluation of Wikipedia Articles through Edit History and Editor Groups. In *Proceedings of the 13th Asia-Pacific Conference on Web Technologies and Applications (APWeb 2011)*, volume 6612 of *Lecture Notes in Computer Science*, pages 188–199. Springer, 2011.

[173] J. Wieczorek, Q. Guo, and R. J. Heijmans. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, 18(8):745–768, 2004.

[174] J. Wielemaker. SWI-Prolog Semantic Web Library. `http://www.swi-prolog.org/pldoc/package/semweb.html`, 2013.

[175] T. Wien. e1071: Misc Functions of the Department of Statistics (e1071). `http://cran.r-project.org/web/packages/e1071`, 2012.

[176] Wikimedia Foundation. Wikipedia. `http://www.wikipedia.org`, 2013.

[177] F. Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[178] Z. Wu and M. Palmer. Verbs Semantics and Lexical Selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL '94)*, pages 133–138. Morgan Kaufmann Publishers Inc. / ACL, 1994.

[179] E. P. Xing, M. I. Jordan, and R. Sharan. Bayesian Haplotype Inference via the Dirichlet Process. *Journal of Computational Biology*, 14(3):267–284, 2007.

[180] C. Yesson, P. W. Brewer, T. Sutton, N. Caithness, J. S. Pahwa, M. Burgess, W. A. Gray, R. J. White, A. C. Jones, F. A. Bisby, and A. Culham. How Global Is the Global Biodiversity Information Facility? *PLoS ONE*, 2(11):e1124, 2007.

[181] L. A. Zadeh. Fuzzy logic. *IEEE Computer*, 21(4):83–93, 1988.

[182] I. Zaihrayeu, P. Pinheiro da Silva, and D. L. McGuinness. IWTrust: Improving User Trust in Answers from the Web. In *Proceedings of the 3rd International Conference on Trust Management (iTrust2005)*, volume 3477 of *Lecture Notes in Computer Science*, pages 384–392. Springer, 2005.

[183] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness. Computing Trust from Revision History. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services (PST 2006)*, volume 380 of *ACM International Conference Proceeding Series*, page 8. ACM, 2006.

Summary

Trusting Semi-structured Web Data

This thesis tackles the problem of trusting semi-structured data, in particular in the Web context. Different definitions of trust are in use in different areas. However, by constraining the context of application of our analyses, we also constrain the definition of trust that we adopt. In particular, we follow the definition of Castelfranchi and Falcone, the definition of Camp and the theory of O'Hara.

We focus on how to properly make use of metadata to estimate the trustworthiness of the corresponding data, given that this is the kind of information at our disposal and the knowledge or time is lacking to determine the trustworthiness of the content of these data. Moreover, we assume the existence of a correlation between metadata and the trustworthiness of the data themselves. For instance, we can estimate the trustworthiness of a piece of data by knowing who created it, and by observing part of his or her behavior. In fact, there exists a probabilistic relation between the user identity and the user trustworthiness, which is summarized by the user reputation. By using appropriate smoothing techniques, the user reputation can be reliably estimated based on a set of sample observations. The strength of this correlation is not the same for all metadata, but our evaluations confirm that this approach (possibly after combining data to strengthen the correlation) can be effective in estimating data trustworthiness.

Chapter 1 gives a first hint about the use of metadata enriched with Web data to assess the trustworthiness of museum annotations. That chapter presents an algorithm for assessing the trust of annotations, as well as an evaluation of it on a collection of bird specimen annotations from the Naturalis museum in Leiden in The Netherlands. The algorithm employs subjective logic, an uncertainty reasoning technique, to learn from a training set of evaluated annotations the relations between trust levels and annotations metadata, some of which are derived from Web sources. That chapter provides some useful insights. One is the fact that in order to make a trust assessment, one can distinguish two distinct phases: trustworthiness estimation and trust assessment (or decision strategy). Another insight is the fact that it shows how metadata are relevant features to be used to learn the trustworthiness of data, and how we can leverage the metadata at our disposal to enlarge their availability by means of Web sources, hence extending the information that we can use to accomplish our goal.

In this line of thinking, Chapter 2 shows that not all the metadata-based and Web-based heuristics that we can formulate to help in our trustworthiness predictions are

actually meaningful, and hence these need to be accurately selected. In that chapter we evaluate a series of heuristics used to estimate the confidence in a set of georeferenced records from specimen annotations from the Naturalis museum, and we show how the best combination of these heuristics is not the one that includes all of them. Also, the heuristics adopted in that case study present ample margins of improvement.

Both Chapters 1 and 2 share the use of statistical as well as evidential reasoning, in particular as a means to handle the uncertainty that is present in our computations. We use some information sources for our estimates, but we do not have any prior warranty, neither about their utility, nor about the representativity of the samples we face. We still use such information because it is the only handhold at our disposal, but we treat it prudently. Evidential reasoning allows us to take into account that the evidence that we did not observe about given metadata could have led us to different conclusions. So, we adopt evidential reasoning as a means to draw prudent conclusions about the data at our disposal, and in Chapter 3 we propose a series of case study applications where uncertainty reasoning is used to model piracy attacks recorded in a Web database. Also, we present a series of extensions of subjective logic which aim at enhancing the logic's usability in the Web environment, thus allowing the logic to deal with semantic similarity, partial evidence observations and open world opinions. Chapter 4 shows, by means of a case study based on police open data, that uncertainty reasoning can also help to prudently extrapolate indications of reliability changes in the data at our disposal, even when it is not possible to have sure indications about this issue. By analyzing the data from different points of view (relative, absolute, etc.), we may be able to obtain insights about their reliability. If this is not possible, uncertainty reasoning can be employed again for aggregating such analyses and deriving weaker but still prudent conclusions about data reliability.

Having understood that statistical and evidential reasoning is particularly important to make trust assessments, we devote part of our research to the analysis of an important class of metadata, that is, provenance information. This leads to two different approaches: on the one hand we build Bayesian networks on top of small provenance graphs to determine the trustworthiness of naval messages, based on the trustworthiness of each single component used to build them (Chapter 5); on the other hand we run supervised learning algorithms on top of aggregated provenance graphs of media tags to predict the trustworthiness of these tags based on how they were produced (Chapter 6). Both approaches share the attempt to learn statistically some models that link information about how, when and by whom artifacts have been created as well as about their trustworthiness. Provenance information has a great potential for helping the trust assessment, and by using statistical techniques we are able to connect some feature combinations with the trustworthiness of the artifacts. This is a necessary step to rely on this class of information for our estimates, because provenance describes how artifacts have been produced, not how trustworthy they are. We first group provenance graphs in classes called "provenance stereotypes", which represent user behaviors, and then use these stereotypes as a basis for trust assessments via machine learning. In this way, we obtain meaningful models, while overcoming possible problems due, for instance, to the complexity of the graphs. This

comes at a cost, since we do not focus deeply on the semantics of these graphs, and therefore we may have neglected useful knowledge. The results that we obtained are already satisfactory. We have not investigated deeper into the link between provenance semantics and trust, but we will in the future.

The last part of the thesis regards the use of semantic similarity measures in combination with uncertainty reasoning (as defined in Chapter 3) to make reputation-based (Chapter 7) and provenance-based (Chapter 8) trust assessments, in particular in the cultural heritage domain. Semantic similarity measures are useful at least for two reasons. Firstly, the uncertainty reasoning techniques we use rely on the availability of evidence, and the more evidence we get the better results we obtain. Semantic similarity measures allow us to enlarge the evidence set at our disposal, while keeping its relevance high. Secondly, semantic similarity measures allow us to reduce the computation complexity in our estimates by avoiding repetition of computations for pieces of evidence that are syntactically different but semantically similar. These are important achievements. Although they are currently limited to specific domains, like cultural heritage, in principle they may be adapted to other domains as well.

Samenvatting

Semi-gestructureerde Web Data Vertrouwen

Dit proefschrift behandelt het probleem van vertrouwen in semi-gestructureerde gegevens, in het bijzonder in verband met het Web. Verschillende definities van vertrouwen zijn in gebruik op verschillende gebieden. Onze definitie hiervan is toegesneden op de toepassingsgebieden van onze analyses. Met name volgen we de definitie van Castelfranchi en Falcone, de definitie van Camp en de theorie van O'Hara.

Onze aandacht richt zich op het op een juiste wijze gebruiken van metadata bij een schatting van het vertrouwen in de corresponderende gegevens, aangezien dit soort informatie tot onze beschikking staat en de kennis of tijd ontbreekt om de betrouwbaarheid van de inhoud van deze gegevens te bepalen. We zijn uitgegaan van het bestaan van een correlatie tussen metadata en de betrouwbaarheid van de gegevens zelf. Bijvoorbeeld, we kunnen de betrouwbaarheid van een stukje informatie afschatten als we weten wie het heeft gecreëerd, en door zijn of haar gedrag te bestuderen. Er bestaat een probabilistische relatie tussen de identiteit en de betrouwbaarheid van een gebruiker, die wordt samengevat door de reputatie van deze gebruiker. Door geschikte transformaties uit te voeren kan de reputatie van de gebruiker betrouwbaar worden afgeschat, gebaseerd op een verzameling onservaties. De kracht van deze correlatie is niet hetzelfde voor alle metadata, maar onze evaluaties bevestigen dat deze aanpak (mogelijk na het combineren van gegevens om de correlatie te versterken) geschikt is voor onze doeleinden.

Hoofdstuk 1 geeft een eerste aanwijzing over het gebruik van metadata verrijkt met gegevens van het Web bij de beoordeling van de betrouwbaarheid van museumaantekeningen. Dit hoofdstuk presenteert een algoritme voor het bepalen van vertrouwen in aantekeningen, en een evaluatie hiervan op aantekeningen bij de vogelverzameling van het Naturalis museum in Leiden. Het algoritme hanteert subjectieve logica om van een trainingverzameling bestaande uit geëvalueerde aantekeningen de relaties te leren tussen niveaus van betrouwbaarheid en metadata van annotaties, deels afkomstig uit nieuwsbronnen op het Web. Dit hoofdstuk geeft een aantal nuttige inzichten. Eén daarvan is het feit dat het maken van een beoordeling van vertrouwen bestaat uit twee verschillende fasen: afschatting van betrouwbaarheid en bepaling van betrouwbaarheid (oftewel de beslissingsstrategie). Een ander inzicht is dat het laat zien hoe metadata gebruikt kunnen worden om het vertrouwen in de gegevens te leren, en hoe we het gebruik van de metadata die ons ter beschikking staan kunnen vergroten door middel van bronnen op het Web, om aldus de informatie uit te breiden die kan worden ingezet voor het bereiken van ons doel. Hierop voortbouwend laat Hoofdstuk 2

zien dat niet alle metadata- en Web-gebaseerde heuristieken voor het bepalen van vertrouwen daadwerkelijk zinvol zijn, en bijgevolg dat deze heuristieken heel precies moeten worden geselecteerd. In dat hoofdstuk evalueren we een serie heuristieken voor het schatten van vertrouwen in een verzameling van geo-informatie voorziene annotaties bij voorwerpen uit het Naturalis museum, en laten we zien dat de combinatie van al deze heuristieken niet optimaal is. Ook zijn er voor de heuristieken die in deze studie worden toegepast ruime marges voor verbetering mogelijk.

Hoofdstukken 1 en 2 delen het gebruik van zowel statistisch redeneren als analyse van gegeven metadata, vooral om met de onzekerheid om te gaan die aanwezig is in onze berekeningen. We gebruiken een aantal informatiebronnen voor onze schattingen, zonder voorafgaande garantie over het nut of de representativiteit van de steekproef waarmee we worden geconfronteerd. Desondanks gebruiken we zulke informatie, omdat dit het enige houvast is dat ons ter beschikking staat, maar we behandelen het omzichtig. Analyse van gegeven metadata maakt het mogelijk om er rekening mee te houden dat gegevens die we niet hebben gezien zouden kunnen leiden tot andere conclusies. Aldus kunnen we prudente conclusies trekken over de gegevens die tot onze beschikking staan. Hoofdstuk 3 bevat een aantal studies waarin redeneren met onzekerheid wordt gebruikt om aanvallen van piraten die zijn geregistreerd in een Web-database te analyseren. Ook presenteren we een aantal uitbreidingen van subjectieve logica om de logica beter bruikbaar te maken voor het Web, doordat de logica om kan gaan met semantische gelijkenis, gedeeltelijke waarnemingen en opinies in een open wereld. Hoofdstuk 4 toont, door middel van een studie gebaseerd op open data afkomstig van de politie, dat redeneren met onzekerheid ook kan helpen om zorgvuldig indicaties te extrapoleren van betrouwbaarheidsveranderingen in de gegevens die ons ter beschikking staan. Indien we niet vast kunnen stellen of deze gegevens betrouwbaar zijn, kunnen we ze analyseren vanuit verschillende gezichtspunten (relatief, absoluut, etc.), om aldus mogelijk inzicht te krijgen in hun betrouwbaarheid. Als we aldus niet kunnen vaststellen of de gegevens die ons ter beschikking staan betrouwbaar zijn, kunnen uiteindelijk technieken voor het redeneren met onzekerheid opnieuw worden toegepast, om resultaten van dergelijke analyses te aggregeren en zwakkere maar nog steeds prudente conclusies te trekken over hun betrouwbaarheid.

Nadat we aldus het inzicht hebben verworven dat statistisch redeneren en analyse van gegeven metadata belangrijk zijn bij het beoordelen van vertrouwen, besteden we een deel van ons onderzoek aan de analyse van een belangrijke klasse van metadata, namelijk herkomst. Dit leidt tot twee verschillende aanpakken: enerzijds bouwen we Bayesiaanse netwerken bovenop kleine grafen van herkomstrelaties voor het bepalen van vertrouwen in nautische berichten, gebaseerd op de betrouwbaarheid van elk afzonderlijk onderdeel dat wordt gebruikt om ze te bouwen (Hoofdstuk 5); anderzijds passen we algoritmen voor leren onder supervisie toe op geaggregeerde herkomstgrafen van media-annotaties om de betrouwbaarheid van deze annotaties te voorspellen op basis van hoe ze zijn geproduceerd (Hoofdstuk 6). Beide aanpakken pogen statistisch modellen te leren die informatie relateren over hoe, wanneer en door wie artefacten zijn gemaakt zowel als hun betrouwbaarheid. Informatie over herkomst heeft een groot potentieel om het beoordelen van betrouwbaarheid te ondersteunen, en met behulp

van statistische technieken zijn we in staat om sommige combinaties van eigenschappen te verbinden met de betrouwbaarheid van artefacten. Dit is een noodzakelijke stap om afhankelijk te kunnen zijn van dit soort informatie voor onze schattingen, omdat herkomst beschrijft hoe artefacten zijn geproduceerd, en niet hoe betrouwbaar ze zijn. We groeperen eerst herkomstgrafen in zogeheten "herkomst stereotypes", die gedrag van gebruikers representeren, en gebruiken vervolgens deze stereotypes als een basis voor het beoordelen van vertrouwen via machinaal leren. Op deze manier verkrijgen we zinvolle modellen, terwijl we mogelijk problemen overwinnen ten gevolge van bijvoorbeeld de complexiteit van de grafen. Dit heeft een prijs, want aangezien we niet diep kijken naar de semantiek van deze grafen, laten we mogelijk nuttige kennis buiten beschouwing. De resultaten die we hebben verkregen zijn al bevredigend. In de toekomst zullen we de connectie tussen herkomstsemantiek en vertrouwen dieper onderzoeken.

Het laatste deel van het proefschrift betreft het gebruik van methoden voor semantische gelijkenis in combinatie met het redeneren over onzekerheid (zoals gedefinieerd in Hoofdstuk 3) om beoordelingen van vertrouwen te geven op basis van reputatie (Hoofdstuk 7) en herkomst (Hoofdstuk 8), met name in het domein van cultureel erfgoed. Methoden voor semantische gelijkenis zijn nuttig vanwege ten minste twee redenen. Ten eerste zijn de technieken voor het redeneren met onzekerheid afhankelijk van de beschikbaarheid van gegevens, en des te meer gegevens er zijn des te beter de resultaten die we verkrijgen. Semantische gelijkenis maakt het mogelijk de hoeveelheid gegevens waarover we beschikken te vergroten, terwijl de relevantie ervan hoog blijft. Ten tweede staat semantische gelijkenis het ons toe om de complexiteit van berekeningen in onze ramingen te verminderen door herhaalde berekeningen te vermijden voor gegevens die syntactisch verschillen maar semantisch vergelijkbaar zijn. Dit zijn belangrijke resultaten. Alhoewel ze momenteel nog beperkt zijn tot specifieke domeinen, zoals cultureel erfgoed, kunnen ze in principe worden aangepast voor andere toepassingsgebieden.

SIKS Dissertatiereeks

2009-1 Rasa Jurgelenaite (RUN)
Symmetric Causal Independence Models
2009-2 Willem Robert van Hage (VU)
Evaluating Ontology-Alignment Techniques
2009-3 Hans Stol (UvT)
A Framework for Evidence-based Policy Making Using IT
2009-4 Josephine Nabukenya (RUN)
Improving the Quality of Organisational Policy Making using Collaboration Engineering
2009-5 Sietse Overbeek (RUN)
Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality
2009-6 Muhammad Subianto (UU)
Understanding Classification
2009-7 Ronald Poppe (UT)
Discriminative Vision-Based Recovery and Recognition of Human Motion
2009-8 Volker Nannen (VU)
Evolutionary Agent-Based Policy Analysis in Dynamic Environments
2009-9 Benjamin Kanagwa (RUN)
Design, Discovery and Construction of Service-oriented Systems
2009-10 Jan Wielemaker (UVA)
Logic programming for knowledge-intensive interactive applications
2009-11 Alexander Boer (UVA)
Legal Theory, Sources of Law & the Semantic Web
2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin)
Operating Guidelines for Services
2009-13 Steven de Jong (UM)
Fairness in Multi-Agent Systems
2009-14 Maksym Korotkiy (VU)
From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)
2009-15 Rinke Hoekstra (UVA)
Ontology Representation - Design Patterns and Ontologies that Make Sense
2009-16 Fritz Reul (UvT)
New Architectures in Computer Chess
2009-17 Laurens van der Maaten (UvT)
Feature Extraction from Visual Data
2009-18 Fabian Groffen (CWI)
Armada, An Evolving Database System
2009-19 Valentin Robu (CWI)
Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets
2009-20 Bob van der Vecht (UU)
Adjustable Autonomy: Controling Influences on Decision Making
2009-21 Stijn Vanderlooy (UM)
Ranking and Reliable Classification
2009-22 Pavel Serdyukov (UT)
Search For Expertise: Going beyond direct evidence
2009-23 Peter Hofgesang (VU)
Modelling Web Usage in a Changing Environment
2009-24 Annerieke Heuvelink (VUA)
Cognitive Models for Training Simulations
2009-25 Alex van Ballegooij (CWI)
RAM: Array Database Management through Relational Mapping
2009-26 Fernando Koch (UU)
An Agent-Based Model for the Development of Intelligent Mobile Services
2009-27 Christian Glahn (OU)
Contextual Support of social Engagement and Reflection on the Web
2009-28 Sander Evers (UT)
Sensor Data Management with Probabilistic Models
2009-29 Stanislav Pokraev (UT)
Model-Driven Semantic Integration of Service-Oriented Applications
2009-30 Marcin Zukowski (CWI)
Balancing vectorized query execution with bandwidth-optimized storage
2009-31 Sofiya Katrenko (UVA)
A Closer Look at Learning Relations from Text
2009-32 Rik Farenhorst (VU) and Remco de Boer (VU)
Architectural Knowledge Management: Supporting Architects and Auditors
2009-33 Khiet Truong (UT)
How Does Real Affect Affect Affect Recognition In Speech?
2009-34 Inge van de Weerd (UU)
Advancing in Software Product Management: An Incremental Method Engineering Approach
2009-35 Wouter Koelewijn (UL)
Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling
2009-36 Marco Kalz (OUN)
Placement Support for Learners in Learning Networks
2009-37 Hendrik Drachsler (OUN)
Navigation Support for Learners in Informal Learning Networks
2009-38 Riina Vuorikari (OU)
Tags and self-organisation: a metadata ecology for learning resources in a multilingual context
2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin)
Service Substitution – A Behavioral Approach Based

on Petri Nets
2009-40 Stephan Raaijmakers (UvT)
Multinomial Language Learning: Investigations into the Geometry of Language
2009-41 Igor Berezhnyy (UvT)
Digital Analysis of Paintings
2009-42 Toine Bogers
Recommender Systems for Social Bookmarking
2009-43 Virginia Nunes Leal Franqueira (UT)
Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients
2009-44 Roberto Santana Tapia (UT)
Assessing Business-IT Alignment in Networked Organizations
2009-45 Jilles Vreeken (UU)
Making Pattern Mining Usefu
2009-46 Loredana Afanasiev (UvA)
Querying XML: Benchmarks and Recursion

2010-1 Matthijs van Leeuwen (UU)
Patterns that Matter
2010-2 Ingo Wassink (UT)
Work flows in Life Science
2010-3 Joost Geurts (CWI)
A Document Engineering Model and Processing Framework for Multimedia documents
2010-4 Olga Kulyk (UT)
Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments
2010-5 Claudia Hauff (UT)
Predicting the Effectiveness of Queries and Retrieval Systems
2010-6 Sander Bakkes (UvT)
Rapid Adaptation of Video Game AI
2010-7 Wim Fikkert (UT)
Gesture interaction at a Distance
2010-8 Krzysztof Siewicz (UL)
Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments
2010-9 Hugo Kielman (UL)
A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging
2010-10 Rebecca Ong (UL)
Mobile Communication and Protection of Children
2010-11 Adriaan Ter Mors (TUD)
The world according to MARP: Multi-Agent Route Planning
2010-12 Susan van den Braak (UU)
Sensemaking software for crime analysis
2010-13 Gianluigi Folino (RUN)
High Performance Data Mining using Bio-inspired techniques
2010-14 Sander van Splunter (VU)
Automated Web Service Reconfiguration
2010-15 Lianne Bodenstaff (UT)
Managing Dependency Relations in Inter-Organizational Models
2010-16 Sicco Verwer (TUD)
Efficient Identification of Timed Automata, theory and practice
2010-17 Spyros Kotoulas (VU)
Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications
2010-18 Charlotte Gerritsen (VU)
Caught in the Act: Investigating Crime by Agent-Based Simulation
2010-19 Henriette Cramer (UvA)
People's Responses to Autonomous and Adaptive Systems
2010-20 Ivo Swartjes (UT)
Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative
2010-21 Harold van Heerde (UT)
Privacy-aware data management by means of data degradation
2010-22 Michiel Hildebrand (CWI)
End-user Support for Access to Heterogeneous Linked Data
2010-23 Bas Steunebrink (UU)
The Logical Structure of Emotions
2010-24 Dmytro Tykhonov
Designing Generic and Efficient Negotiation Strategies
2010-25 Zulfiqar Ali Memon (VU)
Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective
2010-26 Ying Zhang (CWI)
XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines
2010-27 Marten Voulon (UL)
Automatisch contracteren
2010-28 Arne Koopman (UU)
Characteristic Relational Patterns
2010-29 Stratos Idreos(CWI)
Database Cracking: Towards Auto-tuning Database Kernels
2010-30 Marieke van Erp (UvT)
Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval
2010-31 Victor de Boer (UVA)
Ontology Enrichment from Heterogeneous Sources on the Web
2010-32 Marcel Hiel (UvT)
An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems
2010-33 Robin Aly (UT)
Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval
2010-34 Teduh Dirgahayu (UT)
Interaction Design in Service Compositions
2010-35 Dolf Trieschnigg (UT)
Proof of Concept: Concept-based Biomedical Information Retrieval
2010-36 Jose Janssen (OU)
Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification
2010-37 Niels Lohmann (TUE)
Correctness of services and their composition
2010-38 Dirk Fahland (TUE)
From Scenarios to components
2010-39 Ghazanfar Farooq Siddiqui (VU)
Integrative modeling of emotions in virtual agents
2010-40 Mark van Assem (VU)
Converting and Integrating Vocabularies for the Semantic Web
2010-41 Guillaume Chaslot (UM)
Monte-Carlo Tree Search
2010-42 Sybren de Kinderen (VU)
Needs-driven service bundling in a multi-supplier

setting - the computational e3-service approach
2010-43 Peter van Kranenburg (UU)
A Computational Approach to Content-Based Retrieval of Folk Song Melodies
2010-44 Pieter Bellekens (TUE)
An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain
2010-45 Vasilios Andrikopoulos (UvT)
A theory and model for the evolution of software services
2010-46 Vincent Pijpers (VU)
e3alignment: Exploring Inter-Organizational Business-ICT Alignment
2010-47 Chen Li (UT)
Mining Process Model Variants: Challenges, Techniques, Examples
2010-48 Milan Lovric (EUR)
Behavioral Finance and Agent-Based Artificial Markets
2010-49 Jahn-Takeshi Saito (UM)
Solving difficult game positions
2010-50 Bouke Huurnink (UVA)
Search in Audiovisual Broadcast Archives
2010-51 Alia Khairia Amin (CWI)
Understanding and supporting information seeking tasks in multiple sources
2010-52 Peter-Paul van Maanen (VU)
Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention
2010-53 Edgar Meij (UVA)
Combining Concepts and Language Models for Information Access

2011-1 Botond Cseke (RUN)
Variational Algorithms for Bayesian Inference in Latent Gaussian Models
2011-2 Nick Tinnemeier(UU)
Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language
2011-3 Jan Martijn van der Werf (TUE)
Compositional Design and Verification of Component-Based Information Systems
2011-4 Hado van Hasselt (UU)
Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference learning algorithms
2011-5 Base van der Raadt (VU)
Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.
2011-6 Yiwen Wang (TUE)
Semantically-Enhanced Recommendations in Cultural Heritage
2011-7 Yujia Cao (UT)
Multimodal Information Presentation for High Load Human Computer Interaction
2011-8 Nieske Vergunst (UU)
BDI-based Generation of Robust Task-Oriented Dialogues
2011-9 Tim de Jong (OU)
Contextualised Mobile Media for Learning
2011-10 Bart Bogaert (UvT)
Cloud Content Contention
2011-11 Dhaval Vyas (UT)
Designing for Awareness: An Experience-focused HCI Perspective
2011-12 Carmen Bratosin (TUE)
Grid Architecture for Distributed Process Mining
2011-13 Xiaoyu Mao (UvT)
Airport under Control. Multiagent Scheduling for Airport Ground Handling
2011-14 Milan Lovric (EUR)
Behavioral Finance and Agent-Based Artificial Markets
2011-15 Marijn Koolen (UvA)
The Meaning of Structure: the Value of Link Evidence for Information Retrieval
2011-16 Maarten Schadd (UM)
Selective Search in Games of Different Complexity
2011-17 Jiyin He (UVA)
Exploring Topic Structure: Coherence, Diversity and Relatedness
2011-18 Mark Ponsen (UM)
Strategic Decision-Making in complex games
2011-19 Ellen Rusman (OU)
The Mind ' s Eye on Personal Profiles
2011-20 Qing Gu (VU)
Guiding service-oriented software engineering - A view-based approach
2011-21 Linda Terlouw (TUD)
Modularization and Specification of Service-Oriented Systems
2011-22 Junte Zhang (UVA)
System Evaluation of Archival Description and Access
2011-23 Wouter Weerkamp (UVA)
Finding People and their Utterances in Social Media
2011-24 Herwin van Welbergen (UT)
Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior
2011-25 Syed Waqar ul Qounain Jaffry (VU)
Analysis and Validation of Models for Trust Dynamics
2011-26 Matthijs Aart Pontier (VU)
Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots
2011-27 Aniel Bhulai (VU)
Dynamic website optimization through autonomous management of design patterns
2011-28 Rianne Kaptein(UVA)
Effective Focused Retrieval by Exploiting Query Context and Document Structure
2011-29 Faisal Kamiran (TUE)
Discrimination-aware Classification
2011-30 Egon van den Broek (UT)
Affective Signal Processing (ASP): Unraveling the mystery of emotions
2011-31 Ludo Waltman (EUR)
Computational and Game-Theoretic Approaches for Modeling Bounded Rationality
2011-32 Nees-Jan van Eck (EUR)
Methodological Advances in Bibliometric Mapping of Science
2011-33 Tom van der Weide (UU)
Arguing to Motivate Decisions
2011-34 Paolo Turrini (UU)

Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations
2011-35 Maaike Harbers (UU)
Explaining Agent Behavior in Virtual Training
2011-36 Erik van der Spek (UU)
Experiments in serious game design: a cognitive approach
2011-37 Adriana Burlutiu (RUN)
Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference
2011-38 Nyree Lemmens (UM)
Bee-inspired Distributed Optimization
2011-39 Joost Westra (UU)
Organizing Adaptation using Agents in Serious Games
2011-40 Viktor Clerc (VU)
Architectural Knowledge Management in Global Software Development
2011-41 Luan Ibraimi (UT)
Cryptographically Enforced Distributed Data Access Control
2011-42 Michal Sindlar (UU)
Explaining Behavior through Mental State Attribution
2011-43 Henk van der Schuur (UU)
Process Improvement through Software Operation Knowledge
2011-44 Boris Reuderink (UT)
Robust Brain-Computer Interfaces
2011-45 Herman Stehouwer (UvT)
Statistical Language Models for Alternative Sequence Selection
2011-46 Beibei Hu (TUD)
Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
2011-47 Azizi Bin Ab Aziz(VU)
Exploring Computational Models for Intelligent Support of Persons with Depression
2011-48 Mark Ter Maat (UT)
Response Selection and Turn-taking for a Sensitive Artificial Listening Agent
2011-49 Andreea Niculescu (UT)
Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

2012-1 Terry Kakeeto (UvT)
Relationship Marketing for SMEs in Uganda
2012-2 Muhammad Umair(VU)
Adaptivity, emotion, and Rationality in Human and Ambient Agent Models
2012-3 Adam Vanya (VU)
Supporting Architecture Evolution by Mining Software Repositories
2012-4 Jurriaan Souer (UU)
Development of Content Management System-based Web Applications
2012-5 Marijn Plomp (UU)
Maturing Interorganisational Information Systems
2012-6 Wolfgang Reinhardt (OU)
Awareness Support for Knowledge Workers in Research Networks
2012-7 Rianne van Lambalgen (VU)
When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
2012-8 Gerben de Vries (UVA)
Kernel Methods for Vessel Trajectories
2012-9 Ricardo Neisse (UT)
Trust and Privacy Management Support for Context-Aware Service Platforms
2012-10 David Smits (TUE)
Towards a Generic Distributed Adaptive Hypermedia Environment
2012-11 J.C.B. Rantham Prabhakara (TUE)
Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
2012-12 Kees van der Sluijs (TUE)
Model Driven Design and Data Integration in Semantic Web Information Systems
2012-13 Suleman Shahid (UvT)
Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
2012-14 Evgeny Knutov(TUE)
Generic Adaptation Framework for Unifying Adaptive Web-based Systems
2012-15 Natalie van der Wal (VU)
Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes
2012-16 Fiemke Both (VU)
Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
2012-17 Amal Elgammal (UvT)
Towards a Comprehensive Framework for Business Process Compliance
2012-18 Eltjo Poort (VU)
Improving Solution Architecting Practices
2012-19 Helen Schonenberg (TUE)
What's Next? Operational Support for Business Process Execution
2012-20 Ali Bahramisharif (RUN)
Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
2012-21 Roberto Cornacchia (TUD)
Querying Sparse Matrices for Information Retrieval
2012-22 Thijs Vis (UvT)
Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
2012-23 Christian Muehl (UT)
Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
2012-24 Laurens van der Werff (UT)
Evaluation of Noisy Transcripts for Spoken Document Retrieval
2012-25 Silja Eckartz (UT)
Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
2012-26 Emile de Maat (UVA)
Making Sense of Legal Text
2012-27 Hayrettin Görkök (UT)
Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
2012-28 Nancy Pascall (UvT)
Engendering Technology Empowering Women
2012-29 Almer Tigelaar (UT)

Peer-to-Peer Information Retrieval
2012-30 Alina Pommeranz (TUD)
Designing Human-Centered Systems for Reflective Decision Making
2012-31 Emily Bagarukayo (RUN)
A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
2012-32 Wietske Visser (TUD)
Qualitative multi-criteria preference representation and reasoning
2012-33 Rory Sie (OUN)
Coalitions in Cooperation Networks (COCOON)
2012-34 Pavol Jancura (RUN)
Evolutionary analysis in PPI networks and applications
2012-35 Evert Haasdijk (VU)
Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
2012-36 Denis Ssebugwawo (RUN)
Analysis and Evaluation of Collaborative Modeling Processes
2012-37 Agnes Nakakawa (RUN)
A Collaboration Process for Enterprise Architecture Creation
2012-38 Selmar Smit (VU)
Parameter Tuning and Scientific Testing in Evolutionary Algorithms
2012-39 Hassan Fatemi (UT)
Risk-aware design of value and coordination networks
2012-40 Agus Gunawan (UvT)
Information Access for SMEs in Indonesia
2012-41 Sebastian Kelle (OU)
Game Design Patterns for Learning
2012-42 Dominique Verpoorten (OU)
Reflection Amplifiers in self-regulated Learning
2012-43 Withdrawn
2012-44 Anna Tordai (VU)
On Combining Alignment Techniques
2012-45 Benedikt Kratz (UvT)
A Model and Language for Business-aware Transactions
2012-46 Simon Carter (UVA)
Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
2012-47 Manos Tsagkias (UVA)
Mining Social Media: Tracking Content and Predicting Behavior
2012-48 Jorn Bakker (TUE)
Handling Abrupt Changes in Evolving Time-series Data
2012-49 Michael Kaisers (UM)
Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
2012-50 Steven van Kervel (TUD)
Ontology driven Enterprise Information Systems Engineering
2012-51 Jeroen de Jong (TUD)
Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

2013-1 Viorel Milea (EUR)
News Analytics for Financial Decision Support
2013-2 Erietta Liarou (CWI)
MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
2013-3 Szymon Klarman (VU)
Reasoning with Contexts in Description Logics
2013-4 Chetan Yadati(TUD)
Coordinating autonomous planning and scheduling
2013-5 Dulce Pumareja (UT)
Groupware Requirements Evolutions Patterns
2013-6 Romulo Goncalves(CWI)
The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
2013-7 Giel van Lankveld (UvT)
Quantifying Individual Player Differences
2013-8 Robbert-Jan Merk(VU)
Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
2013-9 Fabio Gori (RUN)
Metagenomic Data Analysis: Computational Methods and Applications
2013-10 Jeewanie Jayasinghe Arachchige(UvT)
A Unified Modeling Framework for Service Design.
2013-11 Evangelos Pournaras(TUD)
Multi-level Reconfigurable Self-organization in Overlay Services
2013-12 Marian Razavian(VU)
Knowledge-driven Migration to Services
2013-13 Mohammad Safiri(UT)
Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
2013-14 Jafar Tanha (UVA)
Ensemble Approaches to Semi-Supervised Learning
2013-15 Daniel Hennes (UM)
Multiagent Learning - Dynamic Games and Applications
2013-16 Eric Kok (UU)
Exploring the practical benefits of argumentation in multi-agent deliberation
2013-17 Koen Kok (VU)
The PowerMatcher: Smart Coordination for the Smart Electricity Grid
2013-18 Jeroen Janssens (UvT)
Outlier Selection and One-Class Classification
2013-19 Renze Steenhuizen (TUD)
Coordinated Multi-Agent Planning and Scheduling
2013-20 Katja Hofmann (UvA)
Fast and Reliable Online Learning to Rank for Information Retrieval
2013-21 Sander Wubben (UvT)
Text-to-text generation by monolingual machine translation
2013-22 Tom Claassen (RUN)
Causal Discovery and Logic
2013-23 Patricio de Alencar Silva(UvT)
Value Activity Monitoring
2013-24 Haitham Bou Ammar (UM)
Automated Transfer in Reinforcement Learning
2013-25 Agnieszka Anna Latoszek-Berendsen (UM)
Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
2013-26 Alireza Zarghami (UT)

Architectural Support for Dynamic Homecare Service Provisioning
2013-27 Mohammad Huq (UT)
Inference-based Framework Managing Data Provenance
2013-28 Frans van der Sluis (UT)
When Complexity becomes Interesting: An Inquiry into the Information eXperience
2013-29 Iwan de Kok (UT)
Listening Heads
2013-30 Joyce Nakatumba (TUE)
Resource-Aware Business Process Management: Analysis and Support
2013-31 Dinh Khoa Nguyen (UvT)
Blueprint Model and Language for Engineering Cloud Applications
2013-32 Kamakshi Rajagopal (OUN)
Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development
2013-33 Qi Gao (TUD)
User Modeling and Personalization in the Microblogging Sphere
2013-34 Kien Tjin-Kam-Jet (UT)
Distributed Deep Web Search
2013-35 Abdallah El Ali (UvA)
Minimal Mobile Human Computer Interaction
2013-36 Than Lam Hoang (TUe)
Pattern Mining in Data Streams
2013-37 Dirk Börner (OUN)
Ambient Learning Displays
2013-38 Eelco den Heijer (VU)
Autonomous Evolutionary Art
2013-39 Joop de Jong (TUD)
A Method for Enterprise Ontology based Design of Enterprise Information Systems
2013-40 Pim Nijssen (UM)
Monte-Carlo Tree Search for Multi-Player Games
2013-41 Jochem Liem (UVA)
Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning
2013-42 Léon Planken (TUD)
Algorithms for Simple Temporal Reasoning
2013-43 Marc Bron (UVA)
Exploration and Contextualization through Interaction and Concepts

2014-1 Nicola Barile (UU)
Studies in Learning Monotone Models from Data
2014-2 Fiona Tuliyano (RUN)
Combining System Dynamics with a Domain Modeling Method
2014-3 Sergio Raul Duarte Torres (UT)
Information Retrieval for Children: Search Behavior and Solutions
2014-4 Hanna Jochmann-Mannak (UT)
Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation
2014-5 Jurriaan van Reijsen (UU)
Knowledge Perspectives on Advancing Dynamic Capability
2014-6 Damian Tamburri (VU)
Supporting Networked Software Development
2014-7 Arya Adriansyah (TUE)
Aligning Observed and Modeled Behavior
2014-8 Samur Araujo (TUD)
Data Integration over Distributed and Heterogeneous Data Endpoints
2014-9 Philip Jackson (UvT)
Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language
2014-10 Ivan Salvador Razo Zapata (VU)
Service Value Networks
2014-11 Janneke van der Zwaan (TUD)
An Empathic Virtual Buddy for Social Support
2014-12 Willem van Willigen (VU)
Look Ma, No Hands: Aspects of Autonomous Vehicle Control
2014-13 Arlette van Wissen (VU)
Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains
2014-14 Yangyang Shi (TUD)
Language Models With Meta-information
2014-15 Natalya Mogles (VU)
Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare
2014-16 Krystyna Milian (VU)
Supporting trial recruitment and design by automatically interpreting eligibility criteria
2014-17 Kathrin Dentler (VU)
Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
2014-18 Mattijs Ghijsen (VU)
Methods and Models for the Design and Study of Dynamic Agent Organizations
2014-19 Vincius Ramos (TUE)
Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
2014-20 Mena Habib (UT)
Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
2014-21 Kassidy Clark (TUD)
Negotiation and Monitoring in Open Environments
2014-22 Marieke Peeters (UU)
Personalized Educational Games - Developing agent-supported scenario-based training
2014-23 Eleftherios Sidirourgos (UvA/CWI)
Space Efficient Indexes for the Big Data Era