Foundations and Trends® in Information Retrieval
Vol. 9, No. 5 (2015) 355–475

DOI: 10.1561/1500000046

Credibility in Information Retrieval

Alexandru L. Ginsca
CEA, LIST, 91190 Gif-sur-Yvette, France
alexandru.ginsca@cea.fr

Adrian Popescu
CEA, LIST, 91190 Gif-sur-Yvette, France
adrian.popescu@cea.fr

Mihai Lupu
Vienna University of Technology, Austria
lupu@ifs.tuwien.ac.at

Contents

Abstract

Credibility, as the general concept covering trustworthiness and expertise, but also quality and reliability, is strongly debated in philosophy, psychology, and sociology, and its adoption in computer science is therefore fraught with difficulties. Yet its importance has grown in the information access community because of two complementing factors: on one hand, it is relatively difficult to precisely point to the source of a piece of information, and on the other hand, complex algorithms, statistical machine learning, artificial intelligence, make decisions on behalf of the users, with little oversight from the users themselves.

This survey presents a detailed analysis of existing credibility models from different information seeking research areas, with focus on the Web and its pervasive social component. It shows that there is a very rich body of work pertaining to different aspects and interpretations of credibility, particularly for different types of textual content (e.g., Web sites, blogs, tweets), but also to other modalities (videos, images, audio) and topics (e.g., health care). After an introduction placing credibility in the context of other sciences and relating it to trust, we argue for a quartic decomposition of credibility: expertise and trustworthiness, well documented in the literature and predominantly related to information source, and quality and reliability, raised to the status of equal partners because the source is often impossible to detect, and predominantly related to the content.

The second half of the survey provides the reader with access points to the literature, grouped by research interests. Section 3 reviews general research directions: the factors that contribute to credibility assessment in human consumers of information; the models used to combine these factors; the methods to predict credibility. A smaller section is dedicated to informing users about the credibility learned from the data. Sections 4, 5, and 6 go further into details, with domain-specific credibility, social media credibility, and multimedia credibility, respectively. While each of them is best understood in the context of Sections 1 and 2, they can be read independently of each other.

The last section of this survey addresses a topic not commonly considered under "credibility": the credibility of the system itself, in-

dependent of the data creators. This is a topic of particular importance in domains where the user is professionally motivated and where there are no concerns about the credibility of the data (e.g., e-discovery and patent search). While there is little explicit work in this direction, we argue that this is an open research direction that is worthy of future exploration.

Finally, as an additional help to the reader, an appendix lists the existing test collections that cater specifically to some aspect of credibility.

Overall, this review will provide the reader with an organised and comprehensive reference guide to the state of the art and the problems at hand, rather than a final answer to the question of what credibility is for computer science. Even within the relatively limited scope of an exact science, such an answer is not possible for a concept that is itself widely debated in philosophy and social sciences.

A. L. Ginsca, A. Popescu, and M. Lupu. *Credibility in Information Retrieval.*
Foundations and Trends® in Information Retrieval, vol. 9, no. 5, pp. 355–475, 2015.
DOI: 10.1561/1500000046.

1

Introduction

credibility

– *the quality that somebody/something has that makes people believe or trust them* (Oxford Advanced Learner's Dictionary)
– *the fact that someone can be believed or trusted* (Cambridge Advanced Learner's Dictionary & Thesaurus, 4th Edition)
– *the quality of being believed or accepted as true, real, or honest* (Merriam-Webster.com Aug 2015)
– *the quality of being believable or worthy of trust* (Dictionary.com Unabridged, Aug 2015)
– *the quality of being believed or trusted* (Collins English Dictionary, 2012 Digital Edition)
– *1. the quality, capability, or power to elicit belief / 2. A capacity for belief* (American Heritage Dictionary of the English Language, 5th Edition)
– *qualities that someone has that make people believe or trust them / a. used about things such as systems, statements, or beliefs* (Macmillan Dictionary)
– *the quality of deserving to be believed and trusted* (Longman Dictionary of Contemporary English)

Above we illustrate the fact that eight dictionaries of the English language provide eight different definitions of the object of our study. Superficially similar, the eight definitions are sometimes fundamentally different. Some refer to qualities of speaker, others to states of matter, facts. Some refer to qualities with a concrete effect (believed, trusted, accepted), others to qualities with a potential effect (eliciting, deserving belief, or “can be believed”). All use a variant of the term “belief”, implying the transfer of knowledge, but only six of eight use the term “trust” or variants thereof.

Credibility is therefore difficult to pinpoint. It is certainly not something that as computer science scholars we had imagined we would be concerned with. Computers are precise and their answers correct or formally verifiable. Yet two factors have in the past decade made credibility an issue in computer science. First, the input: a computer is only correct as long as its input is correct. With now the vast majority of content being generated by more or less hidden authors, credibility studies attempt to verify the correctness (in a very general understanding of the word) of this input. Second, the pervasiveness of statistical machine learning in many aspects of information access. The user is distant from the decision making process and generally unable to comprehend the intricacies of the decision-making process that ultimately shows him or her some information pieces but not others.

This survey will define the limits of credibility with respect to digital information access systems. Fundamentally, the discussion of credibility in the context of the digital information age is not different to that started in antiquity. From Aristotle’s Rhetoric, it is the study of the method or art by which a provider of information is able to persuade one or more listeners of the truthfulness or correctness of his or her assertion. While referring the reader to the fascinating literature on the topic, we should perhaps only remind here the three means of persuasion, according to Aristotle [1857, chap. 2, pg. 12] (paraphrased from [Ramage and Bean, 1998]):

Logos the argument itself, its clarity and logic correctness.

Pathos the emotional state of hearer (not of the speaker).

Ethos the character of the speaker, his or her trustworthiness, authority, *credibility*.

Even if Ramage uses the term “credibility” only with respect to *Ethos*, the original text (as translated in English) states that the moral character of the speaker *"carries with it the most sovereign efficacy in making credible"*. We see here the distinction between the use of the term "credibility" with reference to the speaker (and therefore a synonym to "authority" and "trustworthiness") and the use of the term with reference to the message at hand (and therefore a mix of the three factors towards the degree of belief that the hearer or reader places on the message being conveyed). This distinction will be present throughout the present survey, even if the digital medium makes it occasionally more difficult than other media to identify correctly the author, or to make the distinction between authors, publishers, endorsers, or sponsors.

This study will be limited in the philosophical discussion of the meaning of terms like "credibility", "belief", "authority", "trust", or "trustworthiness". The terms have been used differently by different authors. Whenever possible we shall make observations on possible misuse of the terms in order to bring the various studies into the same working frame, but often it is impossible to tell whether the author really meant "credibility" or "trust". It will provide the reader with the set of most up-to-date references to get his or her research started in the area.

1.1 Motivation

According to a 2011 study [Pew Research Center, 2011], about 50% of computer literate individuals, with at least a college degree take *most* of the national and international news from the Web and the trend is increasing. That is: more than television, newspapers, radio, or magazines. It is therefore easy for the reader (as a member of the computer literate population addressed in the above mentioned study) to relate to the need for credibility on the Web.

A recent EuroStat report[1] shows that within the European Union (28 countries), in 2013, 75% of all individuals had used a search engine to find information. Certainly, these percentages are likely to drop in regions under development, but Internet penetration is on the rise even in the most remote places [Talbot, 2013, Pew Research Center, 2014]. In fact, it is likely that Internet adoption will outpace e-literacy [Wy-

[1] `http://bit.ly/167xo82` Visited: August 2015, Most recent data: 2013

att, 2013], and at least some users will have the feeling of trying to quench their thirst for information from a fire hydrant. For instance, in the US, 90% of teens and young adults use the Internet [Lenhart et al., 2011], while only 83% of adults 18 to 24 have at least a high school degree [US Census, 2015]. In the case of the Web, this fire hydrant ejects an amorphous mix of useful, useless and malignant information. It is thus important to be able to differentiate good quality information in the mass of data available on the Web and an adequate understanding and/or modelling of credibility, under its different aspects, can be beneficial in this differentiation process.

This is by no means the first survey on the general topic of credibility. More on the side of communication studies, we have for instance Metzger et al. [2003], where the authors relate empowerment to the ability to determine the veracity of information in a technologically sophisticated context. The issue of youth and the digital media has been thoroughly explored in several studies, such as that of Flanagin and Metzger [2008a] and, more recently, Gasser et al. [2012]. The focus here is on communication studies, rather than the technology side. They include research on understanding users' mental models when assessing credibility, and on the development and evaluation of interventions to help people better judge online content. Moving slightly more towards technology, the field of captology [Fogg, 2003, Atkinson, 2006][2] studies how technology can be designed to persuade end-users. Much prior work in the area of credibility approaches the topic from a captology perspective, with a goal of understanding how people evaluate credibility so as, for instance, to help designers create websites that will appear more credible. Early examples include Shneiderman's [2000] guidelines for designing trust online and Ivory and Hearst's [2002] tool for high quality site design.

The current monograph complements and updates previous surveys: Golbeck [2006] and Lazar et al. [2007] examine the research literature in the area of Web credibility until the year 2007. They examine the general credibility of Web sites, online communication, such as e-

[2] The term *captology* itself is a recent creation of B.J. Fogg, as a derivation on the acronym of C*omputers* A*s* P*ersuasive* T*echnology*.

mail and instant messaging and discuss the implications for multiple populations (users, Web developers, browser designers, and librarians). We expand this with the latest works on credibility in social media and, from a technical perspective, we are mainly interested in automatic methods used for credibility predictions.

A specific focus on trust on information and communication technology (ICT) infrastructures is observable in [Cofta, 2011]. This is something we will expand on as well, particularly since the focus here is on information retrieval, which implies a non-negligible amount of automated decision making that cannot be quantifiably verified in quite the same way as security protocols or network reliability can.

In fact, search engines play an undisputed vital role in the information seeking process and statistical semantic technologies play nowadays a very important role. In addition to topical relevance[3], they also use simple and efficient metrics to estimate the importance of a Web page (e.g., PageRank, HITS algorithms). A few observations can be made at this point:

1. PageRank-like algorithms are substituting a hard problem (credibility) by an easier problem (popularity) ;

2. there is the assumption that the search engine is an impartial information indexer with the users' best interests at heart. Even if that were the case for all search engines, the Web routes search results through a variety of intermediary nodes, most of the time without encryption; The negation of the assumption, as well as the existence of third party intermediaries puts into question the credibility we can assign to a search result list;

3. for the purposes of assessing credibility, the solutions to both of the above issues feed into a recursive credibility question unless the user can develop an understanding of the results provided.

Concerning system credibility, this survey will address primarily the first two problems, and only partially the third (particularly because

[3] We include in topical relevance all methods potentially used to detect it: the variety of term-based matching methods, user click models, etc.

it includes a vast research area in Human-Computer Interaction). In doing so, we strive to focus on technical aspects employed by system developers in order to model, quantify, and assess the credibility of online digital content.

While works that tackle automatic credibility prediction for textual content already exist (and we shall cover them in the following sections), this survey will also specifically covers of credibility works in the multimedia domain. Visual content (images, but even more so videos) have the benefit of being their own proof. However, with complex image and video processing tools available on commodity, mobile hardware[4], making it significantly easier to alter visual content, this benefit will dissipate.

1.2 Definitions

Before proceeding, we should provide a definition of the two elements under discussion here: IR and credibility. In addition to definitions, the following two sections place the survey in context.

1.2.1 IR System

Figure 1.1 shows a highly schematized version of a retrieval system. The IR Engine itself may be considered to be only the *Ranking method*, which in this case includes the indexing, similarity scoring and any other components the retrieval system might have (e.g., relevance feedback). But there are other important components, particularly for the consideration of credibility:

1. Significant amount of information online is directly attributed to a person, be it the editor or author of an article, or the owner of a blog or twitter feed.

2. The data itself, generated by the above-mentioned user and to be indexed and made retrievable by the system.

[4] At the moment of writing Dell was the only producer on the market with a tablet incorporating a light-field camera, while other manufacturers, such as Lytro or Raytrix had specialised cameras available to the general public.

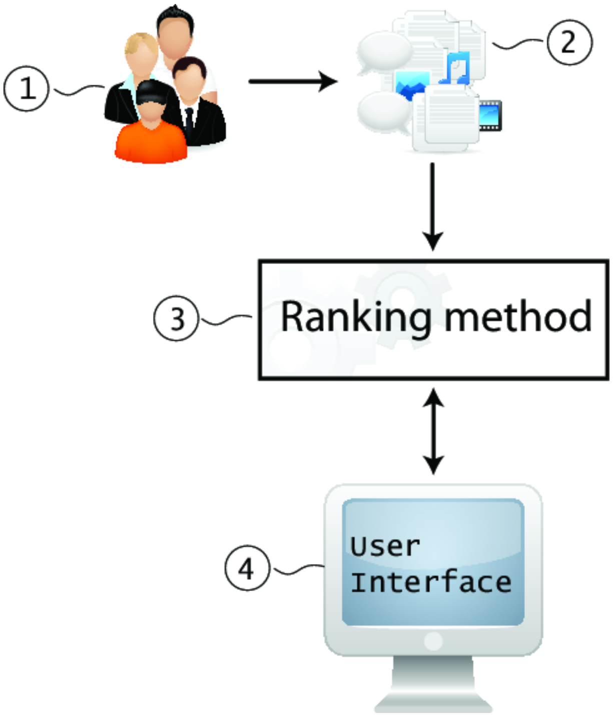


Figure 1.1: A typical information retrieval system.

3. The retrieval system itself that proposes a ranking of the documents available in its index.

4. The interface used to present results to the end-users and to provide interaction means with these results

Information Retrieval is only part of the larger process of solving work tasks involving information that the user does not possess. The credibility requirements come from the work tasks rather then being intrinsic to the IR problem.

Ingwersen and Järvelin [2005] discuss at length the common path that Information Seeking and Retrieval can and should take. Their view of information retrieval, deeply intertwined with the context in which it takes place, is depicted in Figure 1.2.

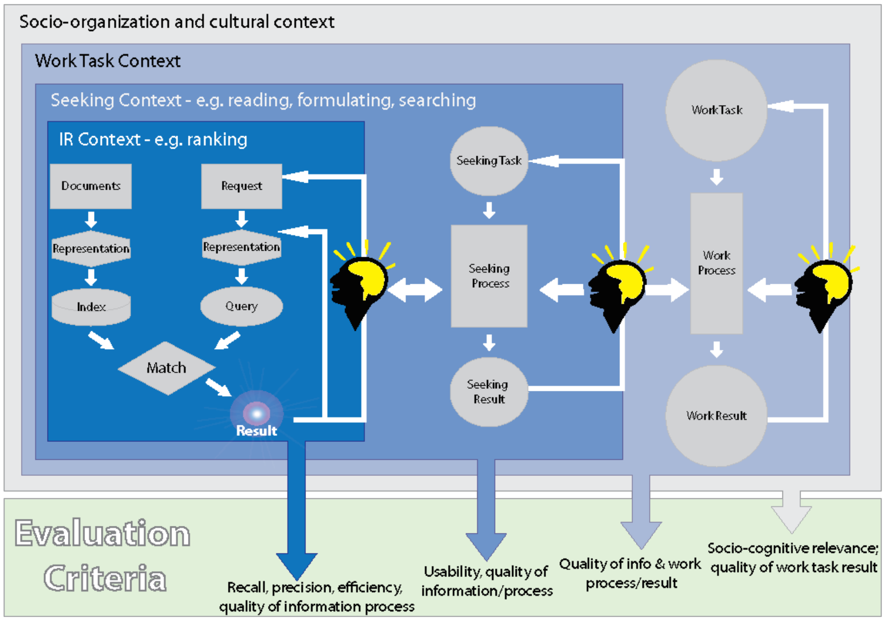

Socio-organization and cultural context
Work Task Context
Seeking Context - e.g. reading, formulating, searching
IR Context - e.g. ranking
Documents
Request
Representation
Representation
Index
Query
Match
Result
Seeking Task
Seeking Process
Seeking Result
Work Task
Work Process
Work Result
Evaluation Criteria
Recall, precision, efficiency, quality of information process
Usability, quality of information/process
Quality of info & work process/result
Socio-cognitive relevance; quality of work task result

Figure 1.2: Information seeking contexts according to [Ingwersen and Järvelin, 2005].

In this general process, credibility is present in the seeking context. Credibility requirements come from the higher levels (organisation, work contexts) and the information identified in the IR context is first assessed for credibility in the Seeking process. We will further investigate credibility in the context of information seeking in Section 3 on Credibility Research Directions.

1.2.2 Credibility

We have already seen the eight definitions provided by various dictionaries for the term “credibility”. Rather than attempting to add yet another definition to these eight, we will use this space to delineate the scope of the current survey.

Each of the four components in Figure 1.1 has its own role to play in the general assessment and study of credibility. This has been previously discussed and expressed as the difference between source, media, and message credibility [Danielson, 2006, Rieh, 2010]. Throughout this survey we shall continue to observe, whenever possible, this distinction in the various studies at hand.

For the Web domain in particular and apparently for the source credibility only, Tseng and Fogg [1999] identify another four types of credibility:

- *Presumed credibility* is based on general assumptions in the users' mind (e.g., the trustworthiness of domain identifiers).

- *Surface credibility* is derived from inspection of a website, is often based on a first impression that a user has of a website, and is often influenced by how professional the website's design appears.

- *Earned credibility* refers to trust established over time, and is often influenced by a website's ease of use and its ability to consistently provide trustworthy information.

- *Reputed credibility* refers to third party opinions of the website, such as any certificates or awards the website has won.

We would argue that these can be filed even under the three classic components of persuasion: Pathos (Surface credibility), Ethos (Reputed and Presumed credibility), and Logos (Earned credibility) and therefore we will not use this specific distinction in this survey.

In terms of constituents of credibility, a majority of researchers agree to identify two components of credibility, namely trustworthiness and expertise [Fogg and Tseng, 1999]. However, we argue that because in today's digital media the source is so much harder to pinpoint [Sundar, 2008], two additional components are of particular interest in judging credibility: quality and reliability. Section 2 will go into significantly more details on each of these. In general, trustworthiness is understood as unbiased, truthful, well intentioned, while expertise is taken to mean knowledgeable, experienced, or competent. In addition, we will discuss quality, which is often seen as an intrinsic characteristic of content shared on the Web, and reliability, which refers to the extent to which something can be regarded as dependable and consistent.

Trust

However, before moving on to the components of credibility, we cannot end this introduction without relating credibility to trust. Trustworthiness, which we agree with the literature is a component of credibility, is a characteristic of the source or of the data. Trust is a characteristic of the consumer of the information and therefore much more related to the general idea of credibility. Therefore, before going on to the computer science aspects and uses of the term, we take a moment to very briefly put trust in the sociology context.

According to Kramer and Tyler [1996], there are at least 16 definitions of *trust* in the literature. The number of associated publications is a few orders of magnitude larger. This being the case, we make no claim to be able to cover even a small part of all these references. Nevertheless, we do need a starting point, and rather than attempting a definition, we prefer an example of the term's context, taken from popular culture[5]:

> "You can't trust Melanie but you can trust Melanie to be Melanie." (Ordell Robie)

The term is used here as a verb, but it can easily be changed to a noun with the help of "having". The use in this context does not appear to refer to any particular property of the target of the trust (Melanie), but rather describes a state of the source (Ordell). This state may be described as *familiarity* of a particular situation or agent. Yet Luhmann [1988] cautions us to make the distinction between familiarity and trust: while trust can only be expressed in a familiar world, familiarity is a fact of life, whereas trust is a specific solution of problems of risk. Another way to describe Ordell's state upon issuing the statement above is *confidence*. In the scene, the character has an unmistakable confident attitude towards the situation, and towards Melanie. Yet again, Luhmann [1988] makes the distinction between trust and confidence: according to him, the first is the result of a conscious analysis of a target, while the second is to a large extent implicit and diffuse.

[5] Ordell Robie is the character played by Samuel L. Jackson in Quentin Tarantino's 1997 film "Jackie Brown"

In fact, it would appear that for any particular definition, we would need to use some terms that are either creating a circular definition, or are somehow different. The first case is of course useless, while for the second we shall probably find someone providing a reasonable and well argumented critique of why the new definition is essentially different from what we experience trust to be.

Therefore, another approach, taken among others by Ullmann-Margalit [2001], attempts to define trust by considering its apparent opposite: distrust. The problem is of course that trust and distrust are not defining the complete mental state of a person with respect to an agent: while trust implies the absence of distrust and vice-versa, it is neither the case that the absence of trust implies distrust, nor that the absence of distrust implies trust.

In his book, Deutsch [1973] analyses both trust and distrust from a psychological rather than sociological perspective, and proposes alternatives to viewing trust as confidence: trust can be despair (as the alternative of distrust), social conformity (perception of cowardness), innocence (from lack of information to cognitive defect), impulsiveness (exaggeration of benefits), virtue (related to social conformity), masochism (negative trust), or faith. This variety in definitions and perceptions led Metlay [1999] to state that attempting to provide a definition of trust conjures up former Justice Potter Stewart's oft-quoted reference to pornography—"it is something that cannot be defined precisely but one knows it when one sees it."

Trust and Knowledge

The general discussions about trust and trustworthiness, in sociology, psychology, or philosophy, are reflected in this survey only with respect to the transfer of knowledge. We mentioned in § 1.2.1 above that the focus here is Information Seeking, and Information Retrieval in particular, as methods and tools to answer an information need. Quite often these days, in both academic and non-academic life, the source of the information is separated from the consumer by the Internet. This is however not the essential difference to the time of book or print prevalence. The difference is that the information presented on the Web is dynamic (may be there one day and changed or completely removed the day after), mediated by large sets of unknown agents (re-tweets, re-posts, blogs, aggregators, and recommenders), and, most differently, potentially *created* by large sets of unknown agents.

The view from § 1.2.1—of a consumer with an information need, to be satisfied from a source of knowledge—is now to some extent undermined because the consumer of knowledge is also the creator, and the simple act of searching becomes knowledge in itself (i.e., through log analysis, for instance). The link between trust and knowledge transfer grows therefore even stronger. Hardwig expressed it most acutely:

> Modern knowers cannot be independent and self-reliant, not even in their own fields of specialization. In most disciplines, those who do not trust cannot know; those who do not trust cannot have the best evidence for their beliefs. In an important sense, then, trust is often epistemologically even more basic than empirical data or logical arguments: the data and the argument are available only through trust. If the metaphor of foundation is still useful, the trustworthiness of members of epistemic communities is the ultimate foundation for much of our knowledge (Hardwig 1991).

In her PhD thesis, Simon [2010] addresses the topic of social knowledge creation (i.e., social epistemology) in the context of today's technologies for creating and sharing knowledge (i.e., socio-technical epistemic systems). Continuing the emphasis that Hardwig placed on trust in knowledge systems, Simon states that "for epistemic content to be considered trustworthy, we further have to trust non-human epistemic agents as well as the processes involved in the creation of this epistemic content". This is also our line of attack on the issue of credibility in information retrieval: addressing both the content and its creators, but also the systems and processes that bring us to this content. This is perhaps not fundamentally different from traditional media, but the peculiarities of the digital age, and most notably of the social web, multimedia abundance, and increasing reliance on machine learning and statistical semantics, provide the research with more than enough material to warrant a new view on the topic.

1.3 Structure of the survey

We start in Section 2 by defining each of the four concepts linked to credibility and provide arguments for our particular distinction between expertise, trust, quality, and reliability. Section 3 looks at general research trends related to credibility in information access systems, making the connection with information seeking research and provides details on features, algorithms, and output of credibility estimation efforts. The following sections address different aspects or perspectives, with the aim of helping the reader jump to areas of particular interest. Section 4 looks at particular domains, such as medical, blogs, or volunteered geographic information systems. In Section 5 we present the latest works on credibility in social networks, with a focus on Twitter and Community Question Answering platforms, while in Section 6, we cover an emerging line of research, namely credibility in the multimedia domain. Finally, the last Section talks about credibility of the information system itself, rather than the data and the sources, which are the primary focus of credibility research in the literature surveyed in the previous sections. After all the different methods and studies have been presented, Appendix A summarizes the existing resources that can be used for the assessment of credibility.

2

Credibility Components

We already indicated in the previous section that the four components of credibility we consider are: expertise, trustworthiness, quality, and reliability (Figure 2.1). The first two are predominantly associated with the source, while the last two are characteristic of the content.

Expertise (the knowledge and skill of the source) and *trustworthiness* (the goodness or morality of the source), are quite well documented in literature [Fogg, 2003]. In contrast, *quality* (how good or bad something is), and *reliability* (the ability of a product, etc., to perform in a required manner, or produce a desired result consistently) need some explanations before going into more details.

It has been argued before that in online media, the ability of the consumer of information to assess the expertise and trustworthiness of the source, and therefore the credibility of the information, is impeded by the difficulty to precisely establish the actual source [Sundar, 2008]. This does not mean that expertise and trustworthiness are no longer important, nor that the user does not, implicitly or explicitly, make judgments with regards to them. It does mean however, that we need to remphasize the importance of the content itself and of the observable behaviour of the source.

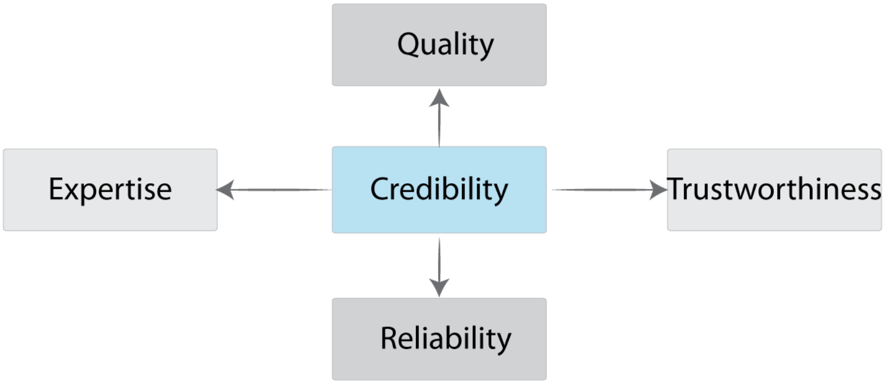


Figure 2.1: Aspects of credibility.

Finding additional characteristics that influence credibility is not difficult: in the literature we find a host of terms describing them: *official*, *truthful*, *scholarly*, *unbiased*, *accurate*, *complete*, *current*, and many others. In fact, Sundar [2008] lists 26 qualities[1] that direct the credibility judgment. These terms are not only numerous, but are used with different meanings in different papers. It therefore becomes difficult to have a meaningful discussion about each of the 26 terms. We decided to use the relatively generic *quality* to denote any observation on the fitness to purpose of the data. Then, because there exists a time component not explicitly accounted yet in our model, we use *reliability*. We do so still with respect to the content rather than the source.

In fact, the user and content axes of credibility may appear as separate research directions but, as it is a common assumption that a credible source produces credible content and vice-versa, these two axes often intertwine. This relation can be found in studies on credibility, where user profile information is analysed together with content features [Juffinger et al., 2009, Weerkamp and de Rijke, 2008] or it can be explicitly modelled, such in the case of Bian et al. [2009], where the authors propose a mutual reinforcement framework to simultane-

[1] Sundar [2008] uses the term *quality* as a synonym of *property*

ously calculate the quality and reputation[2] scores of multiple sets of entities in a network. Although in general there is a positive correlation between source and content credibility, there are examples from the community question answering domain, where the relationship between user reputation and content quality is not always evident. Users that are highly regarded in the community may provide poor answers, and users with a bad answering history may sometimes provide excellent answers [Agichtein et al., 2008].

Next, we will define and illustrate these four concepts, focusing on their relation with credibility.

2.1 Expertise

A majority of early studies describe the use of expertise finding systems within specific organizations and rely on data sources available within the organization. For example, Expert Seeker [Becerra-Fernandez, 2000] was used to identify experts within the NASA organization, relying on a human resource database, an employee performance evaluation system, a skills database, and a project resource management system.

In the social media environment, the number of studies that examine knowledge sharing and expertise increases. Expertise analysis and prediction has been applied on forums [Kardan et al., 2010], online communities [Zhang et al., 2007], blogs [Balog et al., 2008] and collaborative tagging [Noll et al., 2009]. A few studies referred to particular types of social media applications inside the enterprise. Kolari et al. [2008] presented an application for expertise location over corporate blogs that utilized the content of the blog posts, their tags, and comments. Amitay et al. [2009] presented a unified approach that allowed searching for documents, people, and tags in the enterprise. Data was derived from applications for social bookmarking and blogging, but the two data sources were not compared and the system was evaluated as a whole. Guy et al. [2013] focused on comparing a wide variety of enter-

[2] This is another example of variability in the use of terms: Bian et al. [2009] actually define reputation as *expected quality*, which we might see as a parallel to both trustworthiness and expertise

prise social media applications as data sources for expertise inference. Research covering expertise on Community Question Answering stands out among recent works on expertise in online communities. Various approaches have been proposed to automatically find experts in Question answering websites. Jurczyk and Agichtein [2007a,b] adopt the HITS algorithm [Kleinberg, 1999] for author ranking. They represent the relationship of *asker* and *answerer* as a social network and calculate each user's hub and authority value which is implicitly used as and an indicator of expertise. Liu et al. [2005] replace the use of reputation and authority values, derived from link analysis, by profiles that are built from the contents of the expert's questions and answers. They recast the problem as an information retrieval problem and use several language models to represent the knowledge of each user.

Regardless of the method of estimation, expertise information is likely to change over time. Rybak et al. [2014] introduce a temporal expertise profiling task. This task deals with identifying the skills and knowledge of an individual and tracking how they change over time. To be able to capture and distinguish meaningful changes, the authors propose the concept of a hierarchical expertise profile, where topical areas are organized in a taxonomy. Snapshots of hierarchical profiles are then taken at regular time intervals. They propose methods for detecting and characterizing changes in a person's profile, such as switching the main field of research or narrowing/broadening the topics of research.

For all of the above, we consider them to be related to expertise (even if they do not always use the term as such) because the weight is predominantly on the knowledge of the participants. Directly or indirectly, by analysing content or links between the participants, the effort is on establishing their ability to answer questions or to have an opinion on a particular topic, irrespective of their intent, or the quality of an individual content item.

2.2 Trustworthiness

In § 1.2.2 we touched upon the difficulty in defining trust as the closest term to credibility. This is in contrast to trustworthiness, which is defined as a quality of the source, rather than an attitude of the user. In computer science, most approaches to trustworthiness strongly emphasize authority, where a known source is used to inform a user's credibility determinations [Lankes, 2007]. Trustworthy sources are used as an indicator for the credibility of a given piece of information.

In the absence of explicit, external authority, the user has to rely on the content itself. This is particularly so in the context of social media. Here, many works use the concepts of credibility and trust interchangeably while studying trustworthiness in the domain of blogs [Johnson and Kaye, 2009], Wikipedia [Lucassen and Schraagen, 2010], Twitter [Thomson et al., 2012], or Social Question Answering websites [Jeon and Rieh, 2013]. More recently, Toma [2014] proposes a framework that identifies cues associated with trustworthiness in Facebook profiles. They show that cues from the Facebook friends of a user are more important for the estimation of trustworthiness of that particular user than cues from the user himself or herself.

This observation leads us to another distinct area of literature related to trustworthiness in a network environment: trust propagation. Note that the literature here uses predominantly the term "trust" in the context of "trust propagation", rather than "trustworthiness". In this context, it is particularly difficult to make the distinction because while we are considering still the properties of the source , the use of the (social) network could make the assessment dependent on the consumer, as a node in that network. When a method has as a result a unique value (as is the case for instance with PageRank), it would be justified to use the term "trustworthiness". Otherwise, when the method results in a value that is specific to a particular node, we should prefer the term "trust". Ziegler and Golbeck [2015] recently surveyed a number of algorithms for trust inference that combine the network and content features. This was an extension of some of their previous work based on spreading activation-inspired models [Ziegler and Lausen, 2005]. An interesting observation in the context of propagation is that trust and distrust have different transitivity properties. This has been shown already by Guha et al. [2004] in a study on Epinions[3] users, who may assign positive (trust) and negative (distrust) ratings to each other. They observe that, while considering trust as a transitive property makes sense, distrust cannot be considered transitive. Bachi et al. [2012] extend the work of trust on Epinions. They propose a global framework for trust inference, able to infer the trust/distrust relationships in complex relational environments in which they view trust identification as a link sign classification problem. In addition to Epinions, they also test their framework on Slashdot[4], where a user can mark another user as friend or foe, and on Wikipedia[5], where the network is extracted from the votes cast by the users in the elections for promoting users to the role of administrator.

2.3 Quality

Perceptions of quality are closely associated with credibility, with some works identifying quality as the super-ordinate concept [Hilligoss and Rieh, 2008], some viewing the two as associated with separate categories [Rieh and Belkin, 1998], and some regarding quality as subordinate to credibility [Metzger, 2007].

As mentioned above, we define quality rather broadly, as the fitness to purpose. With such a generous definition, quality can also be linked to the interest that certain content can raise (i.e., something is of *"quality"* if it is useful/interesting to the audience, because its purpose is to inform/entertain the audience). Alonso et al. [2010] study the problem of identifying uninteresting content in text streams from Twitter. They find that mundane content is not interesting in any context and can be quickly filtered using simple query independent features. Nevertheless, the primary focus when observing the literature on quality centres around stylistic analysis and spam.

[3] http://www.epinions.com/

[4] http://slashdot.org/

[5] http://www.wikipedia.org/

2.3.1 Text Quality Analysis

When dealing with textual data of any size, ranging from a few characters, such in the case of a Twitter message, to the length of a book, one of the most important features for estimating the credibility of the transmitted message is the quality of the text. This is especially important when there is little or no information about the source of the text or when the truthfulness of the content can not be easily verified.

A considerable amount of work on estimating the quality of text exists in the field of Automated Essay Scoring (AES), where writings of students are graded by machines on several aspects, including style, accuracy, and soundness. AES systems are typically built as text classification tools, and use a range of properties derived from the text as features. Some of the features employed in the systems are:

- lexical, e.g., word length;
- vocabulary irregularity, e.g., repetitiveness [Burstein and Wolska, 2003] or uncharacteristic co-occurrence [Chodorow and Leacock, 2000];
- topicality, e.g., word and phrase frequencies [Rudner and Liang, 2002];
- punctuation usage patterns;
- the presence of common grammatical errors via predefined templates [Attali and Burstein, 2006](e.g., subject-verb disagreements).

A specific perspective with regards to text quality is readability or reading level. In this case, the difficulty of text is analysed to determine the minimal age group able to comprehend it. Several measures of text readability have been proposed. For instance, unigram language models were used on short to medium sized texts [Si and Callan, 2001, Collins-Thompson and Callan, 2004]. Furthermore, various statistical models were tested for their effectiveness at predicting reading difficulty [Callan and Eskenazi, 2007] and support vector machines were used to combine features from traditional reading level measures, statistical language models and automatic parsers to assess reading levels [Petersen and Ostendorf, 2009]. In addition to lexical and syntactic features, several researchers started to explore discourse level features and examine their usefulness in predicting text readability [Pitler and Nenkova, 2008, Feng et al., 2009].

Feng et al. [2010] compared these types of features and found that part-of-speech features, in particular nouns, have significant predictive power and that discourse features do not seem to be very useful in building an accurate readability metric. They also observed that among the shallow features, which are used in various traditional readability formulas (e.g., Gunning-Fog Index or SMOG grading [McLaughlin, 1969]), the average sentence length has dominating predictive power over all other lexical or syllable-based features.

Based on an initial classification proposed by Agichtein et al. [2008], we identify the following groups of textual features used to reveal quality content:

- *Punctuation*: Poor quality text, particularly of the type found in online sources, is often marked with low conformance to common writing practices. For example, capitalization rules may be ignored, excessive punctuation particularly repeated ellipsis and question marks may be used, or spacing may be irregular. Several features that capture the visual quality of the text, attempting to model these irregularities are punctuation, capitalization, and spacing density (percent of all characters), as well as features measuring the character-level entropy of the text.

- *Typos*: A particular form of low quality are misspellings and typos. Additional features quantify the number of spelling mistakes, as well as the number of out-of-vocabulary words. These types or features are found to be useful in several tasks, such as credibility inspired blog retrieval [Weerkamp and de Rijke, 2008] or deriving credibility judgments of Web pages [Akamine et al., 2010].

- *Grammar*: To measure the grammatical quality of the text, several linguistically-oriented features can be used: part-of-speech tags, n-grams or a text's formality score [Heylighen and Dewaele, 2002]. This captures the level of correctness of the grammar used. For example, some part-of-speech sequences are typical of correctly formed questions.

- *Writing style complexity*: Advancing from the punctuation level to more complex layers of the text, features in this subset quantify the syntactic and semantic complexity of it. These include simple proxies for complexity such as the average number of syllables per word or the entropy of word lengths, as well as more advanced ones such as the readability measures [Si and Callan, 2001, Feng et al., 2009].

2.3.2 Spam as an indicator for bad quality

While the dictionary definition of spam[6] still refers exclusively to email, the term has taken a larger meaning in the last decade, referring to all means of undesired, generally commercial, communication.

When referring to Web pages, we can even differentiate between content and link spam:

- *Content spam*: Content spam refers to changes in the content of the pages, for instance by inserting a large number of keywords [Drost and Scheffer, 2005]. Some of the features used for spam classification include: the number of words in the text of the page, the number of hyperlinks, the number of words in the title of the pages, the redundancy of the content, etc. Ntoulas et al. [2006] show that spam pages of this type can be detected by an automatic classifier with a high accuracy.

- *Link spam*: Link spam may include changes to the link structure of the websites, by creating link farms [Baeza-Yates et al., 2005]. A link farm is a densely connected set of pages, created explicitly with the purpose of deceiving a link-based ranking algorithm. Becchetti et al. [2008] perform a statistical analysis of a large collection of Web pages, build several automatic Web spam classifiers and propose spam detection techniques which only consider

[6] Merriam-Webster online `http://www.merriam-webster.com/dictionary/spam`

the link structure of Web, regardless of page contents. Andersen et al. [2008] propose a variation of PageRank, Robust PageRank, that is designed to filter spam links.

Spam is not limited to Web pages and has been well studied in various applications, including blogs [Thomason, 2007], videos [Benevenuto et al., 2009b, 2012], Twitter [Benevenuto et al., 2010, Lee et al., 2010], Facebook [Gao et al., 2010], opinions [Jindal and Liu, 2008], and of course, e-mail (text based [Cormack, 2007] or using multimedia content [Wu et al., 2005]). Automatic methods for detecting spam are especially useful for exposing sources of weak credibility.

Benevenuto et al. [2009a] first approached the problem of detecting spammers on video sharing systems. By using a labeled collection of manually classified users, they applied a hierarchical machine learning approach to differentiate opportunistic users from the non-opportunistic ones in video sharing systems. In a later work, Benevenuto et al. [2012] propose an active learning algorithm that reduces the required amount of training data without significant losses in classification accuracy.

2.4 Reliability

Reliability commonly refers to something perceived as dependable and consistent in quality [Lankes, 2007], generally over a temporal axis. More specifically, text content reliability can be defined as the degree to which the text content is perceived to be true [Xu and Chen, 2006]. According to Rieh [2002], reliability of the content is a criterion that, following topic relevance, is one of the most influential aspects that should be considered for assessing the relevance of a Web publication. Connections between the field of credibility analysis and reliability can be found in works dealing with the credibility assessment of blogs [Juffinger et al., 2009, Nichols et al., 2009, Weerkamp and de Rijke, 2008]. In these works, credibility is applied to multiple concepts besides the reliability measure, and reliability is viewed as a subarea of credibility. Besides being used as a component of credibility, some works place the concept of reliability as the central object of research. Sanz et al. [2012]

use a combination of information retrieval, machine learning, and NLP corpus annotation techniques for a problem of text content reliability estimation in Web documents and Sondhi et al. [2012] propose models to automatically predict reliability of Web pages in the medical domain.

2.5 Summary

We started this section with the observation that the components of credibility can be numerous and diverse, and that literature is rarely consistent in terminology. The focus on the four components depicted in Figure 2.1 is justified, for the first two (expertise and trustworthiness) by the general agreement in the literature, and, for the second two (quality and reliability) by the need to bring to a more explicit level the content. We define quality quite broadly, but it fundamentally refers to a piece of information and its fitness to a purpose. Reliability adds a temporal dimension to the discussion on the content.

3

Credibility Research Directions

Now that we have defined our understanding of credibility, we move on to consider the different research areas at the confluence of computer science, information science and credibility. This section covers: where do credibility requirements come from (§ 3.1), what are the features of the data that can be analyzed to represent credibility (§ 3.2), how to predict credibility, or otherwise quantify, based on the features and requirements, the expectation that the user will find the information credible (§ 3.3), and, finally, how to inform the user about credibility (§ 3.4). The following sections will pick up on some of the topics described here, but the reader can already get an idea of existing and potentially future research from this section.

3.1 Credibility Effects in Information Seeking

In the context of information seeking, we see credibility as a filter or modifier between the work task at hand and the behaviour of the users in order to address the task (Figure 3.1). Manfredo and Bright [1991] argue that credibility is a fundamental cue in the decision-making process that impacts not only individuals' overall attitude but also their

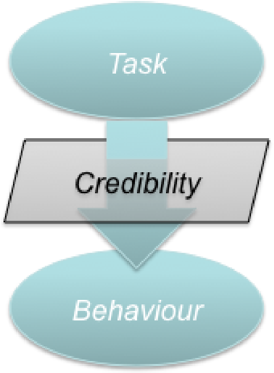


Figure 3.1: Credibility acts as a modifier between task and behaviour.

conscious readiness to engage in a behaviour, i.e., it will not only affect how the user uses the system or the information, but also whether he or she uses it at all.

In order to understand how credibility factors affect a user's information seeking behavior, a branch of research has been geared towards jointly studying the traits of the users that are searching information and the properties of information sources. In order to better understand how people form their judgments on the credibility of information, the 3S-model was introduced by Lucassen and Schraagen [2011]. In their model, three strategies of credibility evaluation are identified. The first strategy is to consider semantic features of the information, such as its accuracy or neutrality. The second one refers to surface features of the information, such as the design of the website. The third strategy is to consider previous experiences with a particular source as an indicator of credibility.

However, the relationship between user and information is difficult to pinpoint: in a revised version of the 3S-model, Lucassen et al. [2013] essentially reversed the flow of influences. If in the first version, the user was an intermediary between information and trust, now the information is an intermediary between user and trust. Even the term "Trust" is changed from *trust judgment* in the first version to simply *trust* in the second. The argument of the authors is that the 3S-model is more of an information model than a process model and therefore *trust* *judgment* might be misleading. The only constant is the relationship between “Trust” and “Source Experience”. While the authors make the argument that the second model is better than the first, merging the two models, as we did in Figure 3.2 appears to us even better, because it shows the dual relationship between users and information.

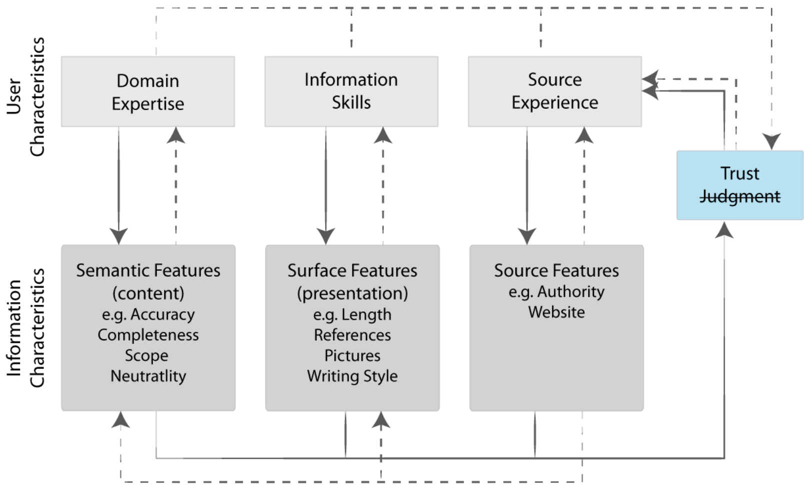

User Characteristics
Domain Expertise
Information Skills
Source Experience
Trust
~~Judgment~~
Information Characteristics
Semantic Features (content) e.g. Accuracy Completeness Scope Neutratlity
Surface Features (presentation) e.g. Length References Pictures Writing Style
Source Features e.g. Authority Website

Figure 3.2: The two versions of the 3S model, merged. Dotted lines represent the first version, Continuous lines represent the second. The first version used “trust judgement”, while the second revised to simply “trust”, to indicate that this is not a process, but a status

Lucassen et al. [2013] also look into gaining insights into the precise nature of the influence of various user characteristics on credibility evaluation with user studies of the credibility of Wikipedia data. They manipulate two key user characteristics for active credibility evaluation (domain expertise and information skills). These are controlled systematically in a think-aloud experiment, in order to better understand their relationship with credibility evaluation and, ultimately, trust. However, with only 40 participants and three variables (participant type: high-school, undergraduate, or graduate students; familiarity with topic: familiar and unfamiliar; and article quality: high and low), the experiments were inconclusive with respect to the relation between trust and information skills or domain expertise of the users. The only statistically significant result was found between quality and trust. Nevertheless, the collateral observations made during the study, about the seeking behaviour in relation to the source and quality are useful: users compensate their lack of domain expertise with information skills and/or source features in order to reach a conclusion.

This link between the source and the seeking behaviour, mediated by credibility has also been explored by Seo and Lee [2014] in a user study centred around a smartphone acquisition scenario. They also propose a general framework to explain how people perceive information credibility when they are familiar with the information source and/or when the information source seems credible. Their model analyses primarily social media (online reviews) and social networks (Facebook) and consequently includes "goal similarity" and "personal similarity" as features of the source. Fundamentally however, they confirm that different seeking behaviours are applied by users to adapt to various source types, their familiarity with the source, and ultimately, their credibility in the source and information.

Ayeh et al. [2013] perform a survey to examine online travelers' perceptions of the credibility of user generated content (UGC) sources and how these perceptions influence attitudes and intentions in the travel planning process. They report mixed results regarding a direct relationship between credibility factors and online travelers' intention to use UGC for travel planning. The direct effect of source expertise on behavioral intention was not supported, while trustworthiness only had a weak effect on behavioral intention. Their findings suggest that trustworthiness and expertise dimensions of source credibility have different importance in affecting attitude and behavioral intention and that trustworthiness is more influential. On the other hand, in a similar study, Xie et al. [2011] found perceived source credibility of online reviews to have a significant effect on participants' intention to book a hotel.

In addition to Wikipedia and e-commerce, probably the most exciting research sub-area for credibility in information access is related to the task of finding medical or health-care related information online. In their survey of the topic, Ziebland and Wyke [2012] identified seven activities that users of online health information (restricted implicitly to patients) might be engaging in: finding information, feeling supported, maintaining relationships with others, affecting behavior, experiencing health services, learning to tell the story, and visualizing disease.

Among these, the "finding information" activity is of interest here. Crawford et al. [2014] explore it, and in particular the impact of individual differences on trust in the context of information seeking behavior. They observe the relationship between health history, predilection to trust different types of online health content, and health information seeking behaviors. The study covers twenty-six subjects who participated in an online study consisting of a survey and a search task. It tests several hypotheses about individual differences in information seeking. The first hypothesis is that people vary in their trust of institutional versus peer produced data. For this purpose, the authors developed a new scale that compared trust in factual versus trust in experiential information. They found support for this hypothesis: some uses trusted institutional sources more than peer-produced data, while others vice versa. Interestingly, they also observed that healthy individuals tend to have high trust for both forums and websites. The second hypothesis tested is that people's preference for institution or peer-produced data would affect their search behavior. Data about people's search habits were collected by asking them about a recent health search. They find that information seekers' trust in websites versus forums were highly correlated. The authors consider that further studies are needed to confirm the findings and believe that the correlation was due more to the relatively healthy nature of the sample population than any issue in the method's ability to differentiate. Finally, the third hypothesis refers to the fact that people who trust peer-produced data will react differently to a forum search task than those that trust institution produced data. For this, the authors explored how individuals relate to information seeking behavior on a series of prescribed forum-search tasks. When they examined the relationship between their proposed measures of trust in forums, websites, experience, and facts and participant search activities there was no significant quantitative relationship in either the recent search question or the laboratory search task. They consider that one possible explanation for this is the relatively small amount of data in their sample or the artificial nature of the search task.

We can see here three steps: users trusting more or less a particular source (authority or community generated), users changing their search behaviour based on their preference of one source (preference presumably linked to credibility in this particular domain), and finally, a less clear step, users changing their behaviour for a specific task based on their credibility assumption about the source.

The relation between credibility, information seeking, and the different domains in which it appears is further detailed in Section 4 on Domain Specific Credibility. In summary, the user studies presented above have demonstrated that users are extremely adaptable in their information seeking behaviour as a function of the information at their disposal and their own domain expertise and information skills. Therefore, it makes sense to look a bit deeper into the features used to assess credibility. We do this next.

3.2 Analysing Credibility

Knowing that credibility affects information seeking behaviour, the next question is to understand it better in order to subsequently be able to predict it and therefore to change it.

There has been interest for providing general guidelines for improving the credibility of Web sites based on a more comprehensive set of factors. One such example is the list of 10 guidelines compiled by The Stanford Web Credibility Project[1]. The following suggestions are included in this list:

1. Make it easy to verify the accuracy of the information on your site.
2. Show that there is a real organization behind your site.
3. Highlight the expertise in your organization and in the content and services you provide.
4. Show that honest and trustworthy people stand behind your site.

[1] `http://credibility.stanford.edu/`

5. Make it easy to contact you.
6. Design your site so it looks professional (or is appropriate for your purpose).
7. Make your site easy to use and useful.
8. Update your site's content often (at least show it has been reviewed recently).
9. Use restraint with any promotional content (e.g., ads, offers).
10. Avoid errors of all types, no matter how small they seem.

This and other similar studies are based on theoretical information processing models, like the *Elaboration Likelihood Model* [Petty and Cacioppo, 1986] or the earlier *Heuristic-Systematic Model* [Chaiken, 1980], in the sense that a large component of credibility (in this case referred to as persuasion) is the ability of the user to evaluate the informational content and the intention behind it.

The importance of intention behind the informational content has been shown in a large study based on Web of Trust[2] (WOT) data covering a one year period by Nielek et al. [2013]. While they primarily investigate if the websites become more credible over time, Nielek et al. also observe that the most credible sites (among 12 categories) are weather forecast sites. They conclude that this is an indicator of the importance of intent in credibility adjudication, since weather forecast is less informationally accurate than news reports of past events, but is seen as unaffected by intentional changes motivated by potentially hidden agendas.

Building on the Elaboration Likelihood Model's two routes that affect the information readers' attitude towards information (the direct—informational route, and the indirect—information-irrelevant route) Luo et al. [2013] perform a study in which they investigate the moderating effect of recommendation source credibility on the causal relationships between informational factors and recommendation credibility. In a second step, the authors also investigate the moderating effect of source credibility on the causal relationship between recommendation

[2] `https://www.mywot.com/`

credibility and recommendation adoption. This study relates to several of the points in the above list, namely all those related to the ability of the user to identify the source of the information and the ability to assess the credibility of the source independently of the content under current examination.

Following the direct, informational route, the cognitive credibility is supported by the ability of the user to understand the content and to place it in context. One aspect here is accessibility of background information (e.g., references), as a requirement and contributor to credibility. In this sense, Lopes and Carriço [2008] present a study about the influence of accessibility of user interfaces on the credibility of Wikipedia articles. The authors looked at the accessibility quality level of the articles and the external Web pages used as authoritative references using the Unified Web Evaluation Methodology[3]. The study has shown that while Wikipedia pages themselves are fairly consistent (the lowest score is only 20% lower than the average), the referenced websites vary widely (from absolute zero to perfect scores). Based on reported results, the authors analyze the article referencing lifecycle and propose a set of improvements that can help increasing the accessibility of references within Wikipedia articles.

Following the information-irrelevant route, some of the first impressions on the credibility of a Web page are based on surface credibility which corresponds to the website's appearance: appealing, professional aspect, or the website's domain. Interestingly, an important role is played by the website's overall aesthetics [Fogg and Tseng, 1999]. Making the link between aesthetics and information source, Xu [2014] proposes a study in which she explores how two personal profile characteristics, reputation cue and profile picture, influence cognitive trust and affective trust towards the reviewer and perceived review credibility, respectively, in a combinatory manner. The findings of the study indicated that the reputation cue (a system generated indicator of reputation) contributed differently from the profile picture to users' trust towards the reviewer: the latter influenced the affective trust alone, while the former influenced both affective and cognitive trust. However,

[3] `http://www.wabcluster.org/uwem1_2/`

profile pictures are not the only factors used in assessing the personal profile of contributors. For each task, the content consumer uses all information at his or her disposal to assess the source. For instance, in the case of a travel-related task, other self disclosed personal profile information (PPI) would be the reviewer location and travel interest, in addition to the textual content of the review itself [Park et al., 2014].

The observation about the profile picture is related to long-standing observations [Patzer, 1983] associating physical attractiveness to higher credibility. Physical attractiveness, defined by Patzer as the degree to which a person's face is pleasing to observe, may be extended to the more general context of website aesthetics and logo design—the degree to which a website or logo is pleasing to observe. Lowry et al. [2014] analyze the visual content of websites as indicators for credibility, with an emphasis on logo design and propose a 4-point check-list for logo design to enhance credibility, defined by them as a combination of expertise, trustworthiness, and dynamism.

In a parallel to aesthetics, some studies focused on the influence of source demographics on the perceived credibility of user generated content on the Internet. Flanagin and Metzger [2003] analyze the impact of the gender of the source (i.e., not of the assessor/reader, but of the content creator) on the perceived credibility of personal Web pages. They found that men and women had different views of Web site credibilityand that each tended to rate opposite-sex Web pages as more credible than same-sex Web sites. A similar study was performed by Armstrong and McAdams [2009], who examine the relationship between source credibility and gender. They examine how gender cues influence perceptions of credibility of informational blogs by manipulating the gender descriptors of a blog's authors. They had participants rate the overall perceived credibility of posts and found that male authors were deemed more credible than female authors. What has not been studied in this respect is the influence of cultural background in such perceptions of credibility. Gender and its roles are perceived differently across cultures and time [Duncan and Pfau-Effinge, 2012] and it would be interesting to observe to what extent these perceptions match credibility in the relatively new, information technology world.

Nevertheless, aesthetics or gender are a more or less important function of the nature of the information to be transmitted. For instance, Endsley et al. [2014] study how different factors affect the perception of credibility of crisis information about natural disasters. In such cases, the focus is on the traditional versus social media sources, rather than the gender or aesthetics of the source.

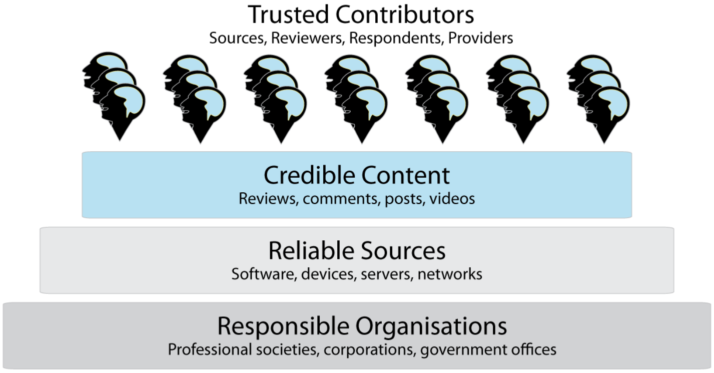

Trusted Contributors
Sources, Reviewers, Respondents, Providers
Credible Content
Reviews, comments, posts, videos
Reliable Sources
Software, devices, servers, networks
Responsible Organisations
Professional societies, corporations, government offices

Figure 3.3: A framework for analysis of social media communities proposed by Shneiderman [2015].

Very recently, Shneiderman [2015] starts from all these observations and proposes a framework for analyzing credible communities in social media platforms. As illustrated in Figure 3.3, he hypothesizes that trusted contributors provide credible content that is delivered by reliable resources, guided by responsible organizations. He also points out that contributors may be misinformed, biased, or malicious, so their content is not credible and physical resources can be undermined.

In fact, the lack of credibility on perceived commercial sales intent has long been documented in the literature [Robertson and Rossiter, 1974]. In the Web domain, the presence of intrusive advertisements has also been formally shown to be negative [Fogg et al., 2002], but the relationship is not always simple. Zha and Wu [2014] present an experimental study that explores how online disruptive advertisements affect users' information processing, feelings of intrusiveness, and news site's credibility. They find that only if ad content is suspected to co-opt with news production, media credibility suffers. Confirming previous results [Choicestream, 2013], their study indicates that users filter out Web ads and, especially for the younger generation, the authors hypothesise that the users understand the distinction between advertisements, even native ones[4], and content.

In summary, there are a variety of cues that users consider when assessing the expertise or trustworthiness of a source and information. Aesthetics and demographics are demonstrably important, in particular when the direct, informational route, is stifled (potentially by the inability of the user to judge the informational content itself).

3.3 Predicting Credibility

The studies described above relied on extensive user surveys or otherwise crowd-sourced data to understand factors affecting credibility. The next logical step is the exploitation of this information to predict what a typical user will consider credible or not. Developing models able to predict the credibility of the source or content on the Web, without human intervention, is therefore one of the most active research areas in the field of Web credibility. Approaches that have been used for this task include machine learning [Gupta and Kumaraguru, 2012, Olteanu et al., 2013, Sondhi et al., 2012], graphical models [Pasternack and Roth, 2013], link algorithms [Pal and Counts, 2011, Gyöngyi et al., 2004] and game theory [Papaioannou et al., 2012].

Olteanu et al. [2013] test several machine learning algorithms from the *scikit-learn*[5] library (SVMs, decision trees, naïve bayes) for automatically assessing Web page credibility. They first identify a set of features that are relevant for Web credibility assessment, before observing that the models they have compared performed similarly, with Extremely Randomized Trees (ERT) performing slightly better. An important factor for the classification accuracy is the feature selection step. The 37 features they initially considered, as well as those ultimately selected (22), can be grouped in two main categories:

- *Content features*: that can be computed either based on the tex-

[4] native ads are those ads designed to look like editorial content

[5] http://www.scikit-learn.org

tual content of the Web pages, text-based features or based on the Web page structure, appearance and metadata features.

- *Social features*: that reflect the online popularity of a Web page and its link structure.

Jaworski et al. [2014] also observe that there is little to no difference in predicting credibility between a simple linear regression method and a neural network model. While the authors do not discuss in great detail the precise nature of the features in [Jaworski et al., 2014], their report supports the observations made before by Olteanu et al. [2013]. Besides the features introduced in the previous two cited papers, Wawer et al. [2014] are also looking for specific content terms that are predictive or credibility. In doing so, they identify expected terms, such as *"energy"*, *"research"*, *"safety"*, *"security"*, *"department"*, *"fed"*, *"gov"*.

Such machine learning methods for predicting credibility rely on either user-study data created in the lab, or on crowdsourced data (for instance, from Web of Trust, or more generally, question answering websites). The latter method can be subjected to credibility attacks by users or by methods imitating the behaviour of correct users. Machine learning has been used also used to counter such attacks. Liu et al. [2013b] identify attackers who imitate the behavior of trustworthy experts by copying a system's credibility ratings to quickly build high reputation and then attack other Web content. They use a supervised learning algorithm to predict the credibility of Web content and compare it with a user's rating to estimate whether this user is malicious or not.

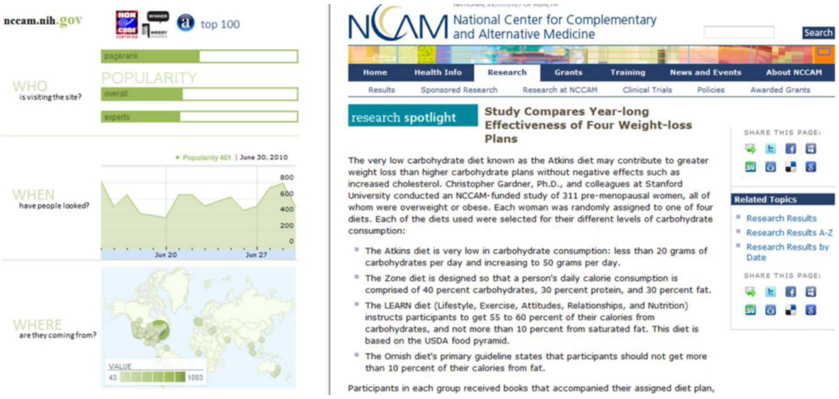
Finally, it should be noted that predictors based on content or social features are limited with respect to the transitory nature of credibility. For events rather than general information websites, the information seeking behaviour is rather reactive than proactive: events trigger a cascade of information units which have to be assessed for both informational content and credibility within a short time frame. In such cases, credibility comes as a second step, after an initial phase identifying newsworthiness. Castillo et al. [2013] use a supervised learning approach for the task of automatic classification of credible news events. In their approach, a first classifier decides if an information cascade corresponds to a newsworthy event, then, a second classifier decides if this cascade can be considered credible or not. For the credibility classifier, several learning models are tested (Bayesian methods, Logistic Regression, J48, Random Forest, and Meta Learning based on clustering.), with Random Forest, Logistic Regression and Meta Learning performing best but indistinguishably from each other from a statistical point of view.

In summary, it has been observed, as in many other domains, that while various machine learning methods can be applied, the most important factor is the set of features that are exploited to perform prediction. Credibility estimation with manual or automatic methods is further developed in the remainder of this section, as well as in Sections 5 and 6.

3.4 Informing About Credibility

Having learned something about the credibility of a website or other information units, the piece of the puzzle is how to present this information to the user in a way that is understandable and credible itself.

Schwarz and Morris [2011] present visualizations to augment search results and Web pages in order to help people more accurately judge the credibility of online content. They also describe findings from a user study that evaluates their visualizations' effectiveness in increasing credibility assessment accuracy and find that augmenting search results with information about expert user behavior is a particularly effective mean of enhancing a user's credibility judgments. In Figure 3.4, we show a sample of their visualization. The Web page visualization appears adjacent to the Web page, so that it is visible regardless of scroll positioning. The visualization uses color and font size to draw attention to a page's domain type, and includes icons to indicate whether a page has received an accredited certification. Horizontal bars indicate the relative value of the current page's PageRank score, general popularity, and popularity among topic experts.



Study Compares Year-long Effectiveness of Four Weight-loss Plans

The very low carbohydrate diet known as the Atkins diet may contribute to greater weight loss than higher carbohydrate plans without negative effects such as increased cholesterol. Christopher Gardner, Ph.D., and colleagues at Stanford University conducted an NCCAM-funded study of 311 pre-menopausal women, all of whom were overweight or obese. Each woman was randomly assigned to one of four diets. Each of the diets used were selected for their different levels of carbohydrate consumption:

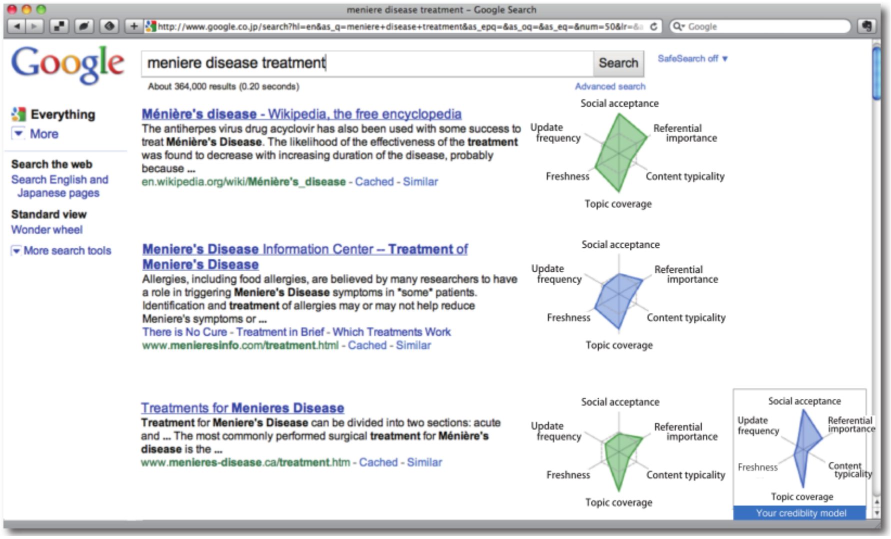
- The Atkins diet is very low in carbohydrate consumption: less than 20 grams of carbohydrates per day and increasing to 50 grams per day.
- The Zone diet is designed so that a person's daily calorie consumption is comprised of 40 percent carbohydrates, 30 percent protein, and 30 percent fat.
- The LEARN diet (Lifestyle, Exercise, Attitudes, Relationships, and Nutrition) instructs participants to get 55 to 60 percent of their calories from carbohydrates, and not more than 10 percent from saturated fat. This diet is based on the USDA food pyramid.
- The Ornish diet's primary guideline states that participants should not get more than 10 percent of their calories from fat.

Participants in each group received books that accompanied their assigned diet plan,

Figure 3.4: Example visualization taken from [Schwarz and Morris, 2011].

Yamamoto and Tanaka [2011a] present a system that calculates and provides visualizations of several scores of Web search results on aspects of credibility, predicts a of user's credibility judgment through user's credibility feedback for Web search results, and re-ranks Web search results based on user's predicted credibility model. As illustrated in Figure 3.5, when users run their system on Google's search engine result pages, the system inserts radar charts that illustrate scores of Web search results on each of credibility factors into search results. The users can also re-rank the search results in accordance with their credibility judgment model by double-clicking radar charts of credible Web search results.

A simpler visualisation of credibility is presented by Amin et al. [2009] (Figure 3.6). Their empirical user study on the effect of displaying credibility ratings of multiple cultural heritage sources (e.g., museum websites, art blogs) on users' search performance investigated whether source credibility has an influence on users' search performance, when they are confronted with only a few information sources or when there are many sources. The results of the study show that by presenting the source credibility information explicitly, people's confidence in their selection of information significantly increases, even though it does not necessarily make search more time efficient.



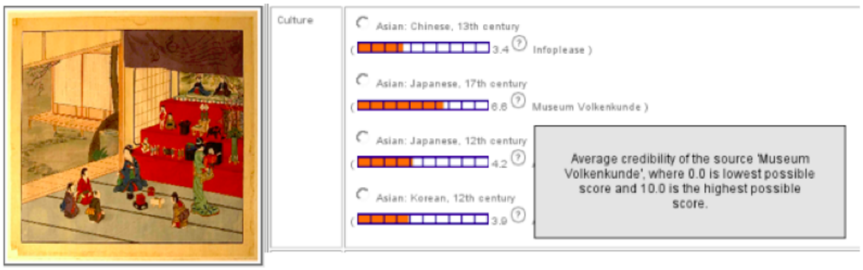
About 364,000 results (0.20 seconds)

Ménière's disease - Wikipedia, the free encyclopedia
The antiherpes virus drug acyclovir has also been used with some success to treat **Ménière's Disease**. The likelihood of the effectiveness of the **treatment** was found to decrease with increasing duration of the disease, probably because ...
en.wikipedia.org/wiki/**Ménière's_disease** - Cached - Similar

Meniere's Disease Information Center – **Treatment** of **Meniere's Disease**
Allergies, including food allergies, are believed by many researchers to have a role in triggering **Meniere's Disease** symptoms in *some* patients. Identification and **treatment** of allergies may or may not help reduce Meniere's symptoms or ...
There is No Cure - Treatment in Brief - Which Treatments Work
www.**menieresinfo**.com/**treatment**.html - Cached - Similar

Treatments for **Menieres Disease**
Treatment for **Meniere's Disease** can be divided into two sections: acute and ... The most commonly performed surgical **treatment** for **Ménière's disease** is the ...
www.**menieres-disease**.ca/**treatment**.htm - Cached - Similar

Figure 3.5: Example visualization taken from [Yamamoto and Tanaka, 2011a].



○ Asian: Chinese, 13th century
(3.4 ⓘ Infoplease)

○ Asian: Japanese, 17th century
(6.6 ⓘ Museum Volkenkunde)

○ Asian: Japanese, 12th century
(4.2 ⓘ

○ Asian: Korean, 12th century
(3.9 ⓘ

Average credibility of the source 'Museum Volkenkunde', where 0.0 is lowest possible score and 10.0 is the highest possible score.

Figure 3.6: Example visualization used in [Amin et al., 2009].

Another visualisation possibility is to show the trend of opinions and articles on news sites. Kawai et al. [2008] make the assumption that if users know the trend of the news site, users can evaluate the credibility of each news topic. Their system detects and uses the sentiment emerging from each news article (i.e., positive/negative sentiment) to resolve the trend of websites. This trend is extracted as average sentiment scores of the news articles that were written concerning a topic in each website.

The alternative to visual displays, such as those just described, is to provide the user with the necessary textual information, to enable him or her to see a variety of facts before making a judgment on credibility. For instance, Murakami et al. [2009] introduce Statement Map, a project designed to help users navigate information on the Internet and come to informed opinions on topics of interest. The proposed system mines the Web for a variety of viewpoints and presents them to users together with supporting evidence in a way that makes it clear how the viewpoints are related. The authors discuss the need to address issues of information credibility on the Internet, outline the development of Statement Map generators for Japanese and English and detail the technical issues that are being addressed. While this is a very exciting research direction, no evaluation of results is provided and it is difficult to estimate the effectiveness of the approach.

In general, the impact of methods designed to help users judge the credibility of Internet content is usually evaluated in a quantitative fashion by conducting focused surveys, either online [ODonovan et al., 2012] or in person [Alsudani and Casey, 2009]).

However, informing about credibility is not always necessarily done based on automatically computed indicators. For instance, in collaborative epistemological resources such as Wikipedia, it is generally the editors who, upon reviewing existing article, introduced credibility indicators such as *"citation needed"*,*"verification needed"*, or *"unreliable source"* [Lopes and Carriço, 2008]. Additionally, the crowd can also be used, and is in fact currently in commercial use under the Web of Trust model, where, upon installing a browser plugin, each link on a website is accompanied by a coloured logo from green to red indicating the crowd-reputation of the website on the other side of the link.

Providing all these indicators, be they automatically calculated as aggregations of credibility aspects, or simply visual cues to known credibility factors (e.g., coloured stars based on average reviews), is of course not guaranteed to trigger a specific behaviour in users. For instance, Flanagin et al. [2014] conducted a large survey and a focused experiment to assess how individuals perceive the credibility of online commercial information, particularly as compared to information available through more traditional channels. They equally evaluated the specific aspects of ratings information that affect people's attitudes toward e-commerce. The results of this survey show that consumers rely heavily on Web-based information as compared to other channels, and that rating information is critical in the evaluation of the credibility of online commercial information. The authors conclude that ratings are positively associated with perceptions of product quality and purchase intention, but that people attend to average product ratings, but not to the number of ratings or to the combination of the average and the number of ratings together. Following this direction, Rafalak et al. [2014] propose a study aimed at identifying various determinants of credibility evaluations. They had 2046 adult participants evaluate credibility of websites with diversified trustworthiness reference index and they focused on psychological factors that lead to the characteristic positive bias observed in many working social feedback systems on the Internet. They find that the level of trust and risk taking are good measures to be included in research dedicated to evaluating websites' credibility and conclude that the usage of the *need for cognition scale* (i.e., one of the scales investigated in their paper) in research connected with evaluating websites' credibility is questionable. This statement is supported by their findings, which show that results obtained in this scale do not differentiate people having tendency to overestimate and underestimate a website's credibility.

3.5 Summary

We have divided the research sphere in four. Starting where information retrieval meets information seeking, we looked at how credibility affects the behaviour of the users. Without going into details, the message here is that there is a concrete impact on the ultimate utility of the information retrieval tool, independent of its effectiveness. The following three parts we identified are: analysing, predicting, and informing about credibility. The analysis identifies features that contributed to the perception of credibility, features which can be used to create algorithms to predict it. They are fundamentally grouped in two: informational (direct) and non-informational (indirect). Finally, in terms of the conveying aspects of credibility to the user, several studies show different visualisations and prove that they generally affect positively the user experience.

4

Domain Specific Credibility

In Section 2 on Credibility Components we looked at the different aspects one may consider when exploring credibility. Section 3 then presented general research directions (factors, analysis, prediction, and informing). It is now time to consider the different requirements and expectations users have, function of the domain of their information need.

Domain specificity in information retrieval has many facets. Hanbury and Lupu [2013] proposed a definition based on a five dimensional space (subject area, modality, users, tasks, and tools): any search engine or process which restricts one or more of the dimensions being considered a domain-specific search engine or process. While they discuss the issue of domain specificity primarily in the context of IR Evaluation, it is also clear that different domains would address or require credibility differently. For instance, in healthcare, the availability of reference written for the common patient is a particularly acute issue, compared to, for instance, the domain of wine blogging, where popularity takes a predominant role.

For now, the focus in on the subject area. We leave the modality to Section 6 where we will talk about multimedia. Social Media may also be seen as a domain, according to the general definition of Hanbury and Lupu, but as its discussion is particularly important in the context of credibility, it also has its own section (Section 5 on Social Media).

With respect to subject areas, among all the possible options one has, we had to select only three, as different as possible from each other. In this context, the importance of healthcare information for both health professionals [Stefanov et al., 2013] and the general public [Peterson et al., 2006] makes this topic unavoidable for this survey. We then selected Volunteered Geographical Information (VGI) that has gained traction with the democratization of GPS-enabled devices. VGI has the particularity that it is in principle very easy to verify the correctness of the data. Healthcare information is often equally verifiable, but generally assertions pertaining to medical effects are expressed as statistics rather than certainties. Finally, this section will also cover blog information retrieval. Blogs cover various subject areas, but a common trend of the cases we survey here is the fact that they are very difficult to verify. A couple of examples in this sense are political opinions [Johnson et al., 2007]) and wine tastes [Cosenza et al., 2014]. The section on blog credibility will also make the transition to the next section, on social media.

4.1 Healthcare Domain

Extensive research has developed several manual methods to combat the propagation of unproven medical claims on the web. The Health-on-the-Net (HON) Foundation[1] is one of the most important organizations involved in the control of health related websites. The foundation applies strict criteria to websites and grants them a seal of approval if they pass their quality and impartiality checks. This is essentially done manually by members of the organisation. In another approach, experts produced rating tools that consumers themselves can apply to websites [Bernstam et al., 2005, Kim et al., 1999] in order to assess them. The two methods, institutional certification (Health-on-the-Net) and peer certification (crowdsourced) are both important, because the credibility requirements in the healthcare area are very high.

[1] http://www.hon.ch/

Price and Hersh [1999] evaluated medical Web page content by combining a score measuring quality proxies for each page. In their work, the concept of credibility is treated as an indicator of quality (i.e., contrary to how we see credibility, as containing rather than being contained in quality). Other quality proxies included relevance, bias, content, and the value of its links. The authors evaluated the algorithm on a small test collection of 48 Web pages covering nine medical topics, and even though they provide no quantitative results, it is one of the first studies that goes in the direction of credibility on medical content online. However the credibility criteria exposed in the study are very similar with those of the general domain, the only difference being the check on existence of a Health-on-the-Net logo on the website.

While in general websites that rank highly in Google are assumed of better quality than those at lower rank, several researchers tested this assumption for health websites. Frické et al. [2005] evaluated the PageRank score as one indicator of quality for 116 websites about carpal tunnel syndrome. Their results show that the PageRank score is not inherently useful for discrimination or helping users to avoid inaccurate or poor information. Griffiths et al. [2005] evaluated PageRank scores with evidence based quality scores for depression websites. The authors obtained Google PageRank scores for 24 depression websites. Two health professional raters assigned an evidence based quality score to each site. Again PageRank scores correlated weakly with the evidence based quality scores. This shows that, considered alone, an algorithm such as PageRank clearly insufficient to account for the quality of medical Websites and domain-adapted metrics or validations are needed.

In the personal healthcare domain, the amount of credibility assessment left to the information consumer is particularly important because of the target demographic of this information. If perhaps for news information the consumer demographic is skewed towards the younger generation [Keskenidou et al., 2014], personal healthcare consumers cover all age demographics and is particularly important for the higher age population. Differences in perception and use of Web content between different age groups have been documented before [Hanson, 2009]. Liao and Fu [2014] expand such studies while focusing on the healthcare domain. They showed that older adults do less deliberation about the content of the websites compared with younger adults. They are also less likely to be susceptible to contextual clues about credibility (e.g., layout, information structure, references, contact information, third-party endorsement). This is important because while studies done before 2005 had shown that Internet health information is used particularly frequently by younger people [Gray et al., 2005], more recent studies such as that of Liao and Fu [2014] point out an increase in the number of elderly adults using online information.

In an effort to assist users in understanding the credibility of websites with healthcare content, Aphinyanaphongs and Aliferis [2007] model expert opinion and build machine learning models that identify Web pages that make unproven claims for the treatment of cancer. They work on a corpus of 191 Web pages taken from 120 websites blacklisted as providing unregulated cancer treatments. Their study showed that machine learning filter models can be used to identify Web pages that make unproven claims and that they have superior performance over public ratings of medical websites and Google search rankings.

However, websites are not the exclusive source of medical information online, with a lot of content being shared on forums. Lederman et al. [2014] examine how users assess the credibility of the information in such forums. The authors discovered five criteria for credibility: reference credibility, argument quality, verification, contributor's literacy competence, and crowd consensus. They quantitatively tested the relationship between those criteria based on two types of information: scientific and experiential. Their analysis shows that different criteria, among the five above, are used by participants in online health forums for each of the two types of information. They used these findings to develop a model for how information credibility is assessed in online health forums. We present their model in Figure 4.1.

Experiential information refers to a user's first-hand experience with a condition or situation (i.e., if the user already had the disease he or she is researching). Scientific information refers to facts directly related to diseases and explains the underlying scientific mechanisms and research (e.g., medication, treatment, studies, explanation, etc.). It is usually shown in the form of referrals to other websites or citing information from external media (e.g., the Internet, books, journal, etc.).

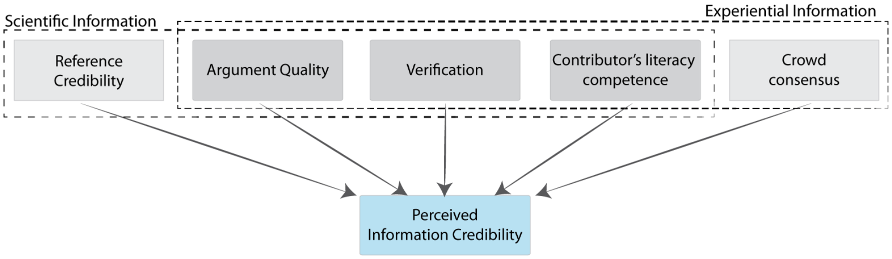


Figure 4.1: Assessment criteria of perceived information credibility of online medical forums as introduced by Lederman et al. [2014].

Bringing these observations into an automated system, Mukherjee et al. [2014] propose a method for automatically establishing the credibility of user-generated medical statements and the trustworthiness of their authors by exploiting linguistic cues and distant supervision from expert sources. They introduce a conditional random fields (CRF) model that jointly learns user trustworthiness, statement credibility, and language objectivity. Using data from `healthboards.com` they show that the most important features are affective and document length (correlating negatively with credibility).

Websites and forums are recently complemented by short messaging services, such as Twitter?. Jiang et al. [2014] explore how microblog users assess the credibility of health advice in this type of posts. Consumers of healthcare information have in this context both less content and less context to leverage in assessing the credibility of the information. In their study, in addition to the Elaboration Likelihood Model (ELM, Petty and Cacioppo [1986]) also used in other studies, Jiang and colleagues use the Protection Motivation Theory (PMT, [Rogers, 1975]), as a *Health Threat Perception*. Therefore, in addition to the usual credibility supporting tools (e.g., official identification from service provider, extensive profile information including credentials), they suggest that consumers of microblogging information are sensitive to the perceived healthcare threat and have to rely on their estimation of self-efficacy in understanding the limited amount of content available.

Research carried towards investigating the credibility of information conveyed in medical related websites also covers non English websites. Gao et al. [2015] investigate the factors influencing Chinese microblog users' perception of the credibility of health and safety information. Among others, they investigated the relation between claim extremity (i.e., how great the danger to healthcare was) and the perception of credibility for both subjective and objective claims. According to their observations, "perceived credibility was higher for objective claims than for subjective claims, and it was also higher for claims of low extremity levels than for claims of high extremity levels. For information about healthcare treatment, objective claims increased the information's credibility only when the claim extremity level was low; when the claim's extremity level was high, the information credibility for objective claims decreased more sharply."

Unfortunately, in terms of multilinguality, there is no research covering healthcare information credibility as a function of the source language. This is an important research direction that deserves exploration.

In summary, the problems of online healthcare information credibility are not fundamentally different that that of general online information. The main difference lies in the risk aversion of users of this type of information, and, to a lesser extent, their general inability to assess the correctness of the data and consequently on their reliance on indirect cues.

4.2 Credibility of Volunteered Geographic Information (VGI)

In recent years, a multitude of Web platforms and Web services have emerged that provide a rich and well maintained geographic and geo-referenced information. These type of geographic data do not come from a single entity, such as the map services provided by Google or Apple, and neither from rigorously organized outsourcing efforts. The geographic information platforms that we are covering in this Section are sustained by contributions from volunteers. This is why one of the biggest concerns is the quality of the available data and the credibility of the users that provide geographical information. Examples of volunteered geographical information include geotagged entries in Wikipedia, Websites such as OpenStreetMap or even geotaged images in Flickr. In this category, we also include user mashups added in Google Earth or Google Maps.

In one of the first efforts of studying VGI platforms, Flanagin and Metzger [2008b] propose an essay that deals with issues related to information and source credibility for volunteered geographic information. They relate the concept of credibility to previous studies focused on Web pages, but also state that credibility, in this particular scenario, is related to the degree to which people's spatial or geographic information is unique and situated, and to the extent to which its acquisition requires specialized, formal training. Since this early essay that laid the ground work and raised important work directions for the assessment of the credibility of volunteered geographical information, several advancements have been made. Idris et al. [2014] propose a conceptual model for the assessment of VGI credibility. They divide their model in two parts, one for medatada, representing the indicator of the hosting (websites) and the sources of data and the other for the actual data, which assesses absolute and relative data positioning, attribute, thematic, temporal and geometric correctness and consistency. Besides the design of the conceptual model, the authors detail its use as an evaluation tool to assess the level of credibility, correctness and consistency of data and information presented through the VGI medium. Their proposal unifies elements coming from a wide range of approaches, including machine learning, third party validation and linked data infrastructure. Senaratne et al. [2013] analyze the credibility of content provided within a visual VGI source and take Flickr as a case study. They observe which photograph/user metadata can be used to infer the credibility of contributors regarding a correct geotagging. They find that contributors who incorrectly label their photographs have on average the highest number of photos and contributors who correctly geotag and label their photos have on average the highest number of copyrighted photos, the least distance to the target, and also the least number of tags per photo. As in the case of Web page, blog or Twitter credibility, machine learning algorithms have been tested to predict the credibility of VGI. In his thesis, Kim [2013] proposes a Bayesian network approach for modeling the credibility of volunteered geographic information for emergency management.

Following the credibility related concepts schema introduced in Section 2, we also include works that focus on the data quality and user expertise aspects of VGI platforms. A conceptual workflow for automatically assessing the quality of volunteered geographic information is described in [Ostermann and Spinsanti, 2011]. Their model is designed for an automatized retrieval, processing, analysis, quality assessment and dissemination of user generated geographic content on volunteered geographic information for crisis management. The authors focus on the geographic aspects of the data, such as the value of location information for determining the credibility and relevance of the VGI, and detection and assessment of crisis events. One advantage of their method relies on the fact that it is modular and can easily incorporate research advances from other scientific disciplines, such as improved information retrieval systems or semantic analysis techniques. In a related work, Goodchild and Li [2012] propose three approaches to quality assurance for VGI contributions: crowd-sourcing, social, and geographic approaches respectively. They conclude that the crowd-sourcing approach is attractive but appears to be less effective for geographic facts than for other types of information, because such facts vary the very prominent to very obscure, and errors in geographic facts can clearly persist even in heavily populated areas. While the social approach has proven effective when projects are sufficiently popular and well-structured, and supported by extensive social networks, the geographic approach is attractive to geographers because it taps the heart of geographic knowledge and motivates a comprehensive effort to formalize such knowledge. While Goodchild and Li [2012] presented three different aspects for quality assurance of VGI, Spielman [2014] focuses on crowd-sourcing. He investigates the implications of collective intelligence in the context of crowd-sourced geographic information. He notices that quality of collectively generated maps is hard to assess because there are two aspects that should be taken consideration – the credibility perspective and the accuracy perspective. While the link between credibility and data quality comes to terms with our view of credibility, the problem of geographical data accuracy remains strictly related to this domain. The author finds support in the existing literature on collective intelligence that the structure of groups is more important than the intelligence of group members. The paper concludes with some design recommendations and by considering the implications of collectively generated maps for both expert knowledge and traditional state sponsored mapping programs. On a related note, See et al. [2013] study crowdsourced data on land cover that were collected using a branch of Geo-Wiki called Human Impact[2]. The Geo-Wiki project is a directed form of VGI creation in which volunteers are steered towards a structured reporting/scoring environment. They score the land cover in randomly selected global locations and their scores are stored in a database. Their research seeks to determine the impacts of varying landscape conceptualizations held by different groups of VGI contributors on decisions that may be made using crowdsourced data. As expected, their study found that conceptualization of cropland varies between experts and non-experts.

Remaining on the works covering expertise in the VGI domain, Karimipour and Azari [2015] look at contributions from expert users in order to help other users. The authors make the interesting statement that VGI is more based on human cognition than on measurement and that a degree of truth can be assigned to geographical information because users' experience has a direct effect on the reliability of the collected information. Similar to our view on credibility, they find the following three key components of VGI credibility: believability, trustworthiness and expertise.

Yanenko and Schlieder [2014] propose methods for improving the quality of VGI and introduce gamification principles in the context of VGI with the main focus being on the quality of the data collected in a VGI game. They present two approaches for data validation that can be included in mobile data collecting games. The approaches are compared with respect to their applicability to the VGI scenario, the

[2] `http://humanimpact.geo-wiki.org`

expected effect on the quality of the dataset collected in the game as well as on the player's in-game behaviour.

As pointed out earlier in Figure 2.1, we consider trust to be a possible proxy for credibility. Keßler and de Groot [2013] continue recent research directions towards the estimation of VGI quality through the notion of trust as a proxy measure. In their paper, they investigate which indicators influence trust, focusing on inherent properties that do not require any comparison with a ground truth dataset. The indicators are tested on a sample dataset extracted from OpenStreetMap. High numbers of contributors, versions and confirmations are considered as positive indicators, while corrections and revisions are treated as indicators that have a negative influence on the development of feature trustworthiness. They find significant support for the hypothesis that feature level VGI data quality can be assessed using a trust model based on data provenance.

The works referenced in this Section show that VGI platforms are, alongside Twitter, one of the most prolific targets for emerging credibility studies. While from a retrieval point of view, we observe a direct application of credibility models for geographic related Flickr data, in terms of information seeking, analyzing the quality of volunteered geographical data used in visualization systems remains equally important.

4.3 Blog Credibility

While blogs are social in the sense that there can be multiple contributors and the readers have the possibility to comment on every post, this is not fundamentally different from how most traditional media outlets structure their current online presence. The credibility associated with blogs is also comparable or sometimes even higher than traditional media. This has been shown in a recent survey by Johnson and Kaye [2014]: in the case of political campaigns, users find blogs more credible then online newspapers, but at the same time find Facebook and similar services less credible as a whole. This is why we position the study of blogs as domain-specific rather than social media.

As detailed in Section 2, the notion of credibility encompasses other adjacent terms. This relation between terms can also be observed when looking at the literature on blog credibility. Some works have credibility as a central subject [Rubin and Liddy, 2006, Juffinger et al., 2009, Weerkamp and de Rijke, 2008, 2012], while others approach credibility related problems through the study of trust [Johnson and Kaye, 2009] or study the impact of demographic factors (e.g., gender [Armstrong and McAdams, 2009], interest in politics [Johnson et al., 2007]) on the perceived credibility of blogs.

For blogs, a major body of work related to credibility is aimed at trying to identify blogs worth following. Therefore, in addition to general credibility features as described in the previous sections, reports on blog credibility use topic concentration, variety, or consistency. Two examples in this direction are Sriphaew et al. [2008] and Chen and Ohta [2010].

Additionally, a number of blog-specific aspects have been investigates over the years: Weerkamp and de Rijke [2008] observed that the quality of the blog (and, following our definition, credibility) can be estimated by the length of the posts and the number of comments. Hearst and Dumais [2009] continued this study and examined blogs having one or more contributors, which in addition to showing differences in average post length (more authors correlated significantly with longer posts), also showed that multi-author blogs were ranked higher than blogs with 1-2 contributors (according to a ranking based on the Technocrati blog aggregator). Also related to the contributors, Van House [2004] stressed the importance of the connection of online and offline blogger identities and enhancing the effect of personal voice.

In terms of topics, Mishne and Glance's observation [Mishne and Glance, 2006b] that bloggers often report on news events can be the basis for credibility assessment. They initially applied this observation to spam detection [Mishne, 2007] as a first step in a blog-retrieval process. Later, Juffinger et al. [2009] compared blog posts to news articles about the same topic and assign a credibility level based on the similarity between the two. Related to credibility in blogs is also the automatic assessment of forum post quality discussed by Weimer et al. [2007]. The authors use surface, lexical, syntactic and forum specific features (such as whether or not the post contains HTML, and the fraction of characters that are inside quotes of other posts) to classify forum posts as bad or good. In their analysis, the use of forum specific features gives the highest benefits to the classification.

Rubin and Liddy [2006] propose one of the first in-depth studies on features specifically related to blog credibility. They sketch a four factor analytical framework for blog-readers' credibility assessment of blog sites, based in part on website credibility assessment surveys [Stanford et al., 2002] and Van House's observations on blog credibility [Van House, 2004]. The four factors and indicators for each of them are:

- *blogger expertise and offline identity disclosure*: name and geographic location, credentials, affiliations, hyperlinks to others, stated competencies, mode of knowing;
- *blogger's trustworthiness and value system*: biases, beliefs, opinions, honesty, preferences, habits, slogans;
- *information quality*: completeness, accuracy, appropriateness, timeliness, organization (by categories or chronology), match to prior expectations, match to information need;
- *appeals and triggers of a personal nature*: aesthetic appeal, literary appeal, writing style, curiosity trigger, memory trigger, personal connection.

Weerkamp and de Rijke [2008] incorporate textual credibility indicators in the retrieval process to improve topical blog post retrieval. They consider two groups of indicators: post level (determined using information about individual blog posts only) and blog level (determined using information from the underlying blogs). We summarise the credibility indicators that they used:

- *capitalisation*: an indicator of good writing style, which in turn contributes to a sense of credibility
- *emoticons*: excessive use indicates a less credible blog post;

- *all capitalised*: words written in all caps are considered shouting in a Web environment; shouting is an indicator for non-credible posts;

- *spelling*: the more spelling errors occur in a blog post, the less credible it is;

- *length*: credible texts have a reasonable length; the text should supply enough information to convince the reader of the author's credibility;

- *timeliness*: for news related topics, a blog post is more credible if it is published around the time of the triggering news event;

- *semantic*: the semantic indicator also exploits the news-related nature of many blog posts, and prefers posts whose language usage is similar to news stories on the topic;

- *spam*: spam blogs are not considered credible and are demoted in search results;

- *comments*: readers of a blog post often have the possibility of leaving a comment for other readers or the author. When people comment on a blog post, they apparently find the post worth putting effort in, which can be seen as an indicator of credibility [Mishne and Glance, 2006a];

- *regularity*: the temporal aspect of blogs may indicate credibility: bloggers with an irregular posting behaviour are less credible than bloggers who post regularly;

- *topical consistency*: when looking for credible information, posts from bloggers that have a certain level of topical consistency should be ranked higher.

In a later work, Weerkamp and de Rijke [2012] propose to use ideas from their previous credibility framework in a re-ranking approach to the blog post retrieval problem. They introduce two methods of re-ranking the top n results of an initial run, used as baseline. The first approach, credibility-inspired re-ranking, simply re-ranks the top n of a baseline based on the credibility-inspired score. The second approach, combined re-ranking, multiplies the credibility-inspired score of the top n results by their retrieval score, and re-ranks based on this score. The evaluation shows that credibility-inspired re-ranking leads to larger improvements over the baseline than combined re-ranking, but both approaches are capable of improving over an already strong baseline.

With regards to the facets of credibility in the blogosphere, Cosenza et al. [2014] point out that these will vary based on the blog topic itself. They study the relationship between credibility and trust in wine blogs and observe that the most important factor is source (i.e., author). It is more important than the content itself or the aesthetics and quality of the website. This may appear at first hand counter-intuitive, especially with respect to the preference of source versus content. The authors motivate the observation through the nature of the blogs in general, but one could also hypothesise that the relative importance of content and source is controlled by the ability of the user to objectively verify the content. Wine, the topic of the blogs studied by Cosenza et al. [2014], is notoriously difficult to objectively evaluate for most of the people.

This reliance on the source rather than content is related to the concept of *authority*. Conrad et al. [2008] address problems related to representing, measuring, and monitoring the degree of authority and presumed credibility associated with various types of blog participants. The authors explore, in particular, the utility of authority-detection layered atop opinion mining in the legal and financial domains.

As usual, studies on credibility are not done only on behalf of the consumers, but also on behalf of the sources. Rieh et al. [2014] explain how blog owners change their behaviour to increase perceived credibility. The authors define the concept of audience-aware credibility to summarise the behaviour of blog owners changing their actions and decisions as a function of their perceived audience.

In summary, the blogosphere is at the confluence of traditional publication and social media. As such, credibility studies addressing blogs in particular look for typical credibility features, such as the identity and authority of the authors, the quality of the texts themselves, and the topical consistency as an indicator for expertise. Additionally, they take elements from social media such as the number of contributors to the same blog, or the number of comments.

4.4 Summary

The three domains covered here show three very different aspects of credibility: the healthcare domain covers users who are generally risk averse and where the content is most often understandable to experts, but not to the general public. VGI systems have on the other hand information which is generally easily verifiable, albeit sometimes also only by expert users and only with special equipment. For the general public however, the tasks are not associated with high risk and therefore credibility estimates are more lenient. Finally, the blogosphere is a mix of all possible topics, with users who predominantly use the information for very low risk tasks (e.g., entertainment, general information) and where information is most difficult, if not impossible, to verify (e.g., opinions, tastes). In this case, we observe a particular focus on authority, on topical consistency, and on convergence with the users' existing opinions or preferences.

5

Credibility in Social Media

Blogs, as discussed at the end of the previous section, mark the transition to social networks and social media. Without doubt, social media and the social networks behind it are now an important information source and have been shown to influence societal events, such as governmental elections [Johnson and Perlmutter, 2010, Geiber, 2011].

In this section we give an overview of general approaches used for assessing credibility in social media, focusing on link based methods. In the last two sections we go into more details on works on credibility focused on Twitter and question answering platforms. While microblogging services, such as Twitter, may be at first glance considered more similar to blogs than social networks, Myers et al. [2014] observe that "Twitter starts off more like an information network, but evolves to behave more like a social network". Question Answering platforms are interesting to study in this case because they are platforms where users are actively seeking information and obtain answers in a distributed, uncontrolled fashion.

Arguably, in an environment built around user contributions, the most important step towards information credibility is establishing the credibility of the user who contributed that information. In our scheme, this translates to a predominant focus on expertise and trustworthiness. Sometimes, indicators of user credibility are explicitly embedded in the website. Twitter, for example, has a set of verified accounts that are accompanied by a badge. Facebook has official pages of organisations or public figures explicitly marked as such. This helps users trust that a legitimate source is authoring the account's messages and, presumably, discover high-quality sources of information. While these initiatives are helpful, social media websites are not able to verify all their users. More than that, many users would prefer to remain unknown, and it is clearly the case that the majority of users in social media are unverified.

As an alternative, popularity is a common proxy for credibility of users and content in social media. Popularity and credibility are sometimes used interchangeably, although user studies show that subjects are aware that they are not the same [Hilligoss and Rieh, 2008]. For example, many users would trust a Twitter user who has many followers. This observation is used by spammers and popularity is then part of spam-detection methods [Lee et al., 2010]. Similarly, one might trust a video clip on YouTube if many people had already watched it [Benevenuto et al., 2009b]. Popularity is something that can be relatively easily measured. In social network analysis, link-based methods are therefore one of the most used approaches. In particular, link-based ranking algorithms that were successful in estimating the quality of Web pages have been applied in this context. The output of such algorithms can be interpreted as an indicator for expertise or trustworthiness, or a combination thereof, function of the nature of the links. If the links are of a direct, personal nature (e.g., Facebook friendships, tweeter follows), we may conjecture that the result will be predominantly based on trust. If the links are indirect, mediated by information objects (e.g., questions answered, online recommendations), we may conjecture that the result will be predominantly determined by expertise. We shall follow this line of thought in the coming two sub-sections.

However, regardless of the indicator that various systems identify, Edwards et al. [2013] remind that any such indicator is subject to the interpretation of the user. They focus on a popular influence indicator platform, *Klout.com*, a website that proposes a popular indicator of a user's online influence. Their study found that a mock Twitter page with a high Klout score was perceived as higher in terms of credibility compared with the identical mock Twitter page with a moderate or low Klout score.

5.1 Twitter

Similar to blog credibility, research dealing with credibility in microblogging environments, often represented by Twitter, targets one or several of the credibility dimensions mentioned in Section 2. In general, the amount of publications related to Twitter in the academic computer science community has been astounding: between 2007 and 2015, the digital library of the ACM indexed 8888 publications mentioning the term, with a peak of 1911 in 2013[1].

A first group of papers focuses on user studies, trying to understand what users consider as credible, both in terms of individual tweets (information credibility) and users (source credibility). Sikdar et al. [2013] propose a methodology for developing studies that introduce methods to make credible data more useful to the research community and provide interesting guidelines. For instance, they find important that the underlying ground truth values of credibility to be reliable and the specific constructs used to define credibility must be carefully described. By proposing these guidelines, they offer an important theoretical framework for future efforts in credibility ground truth construction. They also consider that the underlying network context must be quantified and documented. To illustrate these two points, the authors conduct a unique credibility study of two different data sets on the same topic, but with different network characteristics.

In a typical research setting of analyzing credibility factors through users studies, as presented in § 3.2, Westerman et al. [2014] examine how pieces of information available in social media impact perceptions of source credibility. Participants in the study were asked to view one of three mock Twitter pages that varied in the recency with which tweets

[1] That is more than five publications per day, every day of 2013. The numbers are obtained using the search interface at `http://dl.acm.org/` (September 2015)

were posted and then to report on their perceived source credibility of the page owner. In a similar work, Aladhadh et al. [2014] investigate how certain features affect user perceptions of the credibility of tweets. Using a crowdsourcing experiment, they found that users' perception of the credibility of tweets is impacted more by some features than by others. Most noticeably being the fact that displaying the location of certain types of tweets causes viewers to perceive the tweets as more credible.

Shariff et al. [2014] also examine user perception of credibility, with a focus on news related tweets. They conduct a user study on a crowdsourcing platform to judge the credibility of such tweets. By analyzing user judgments and comments, they find that eight features, including some that can not be automatically identified from tweets, are perceived by users as important for judging information credibility. Moreover, they find that distinct features like the presence of links in tweets, display name and user belief consistently lead users to judge tweets as credible and that users can not consistently judge or even misjudge the credibility for some tweets on politics news.

Kostagiolas et al. [2014] evaluate whether work-related or personal motivating factors influence the relation between perceived credibility and trust toward institutional information sources and how each factor affects this relation. Findings suggest that work-related factors have a higher impact on the relation between credibility and trust than personal motivation factors, while they are stressing the important role of hospital libraries as a dissemination point for government-sponsored information resources.

In summary, what all these studies find is that: 1. there are common points with credibility studies in any other domains (identifying the source, its expertise, trustworthiness, and consistency; the language used in each information unit); 2. there are new, Twitter specific, features: number of followers, retweets, mentions, URLs in text. ODonovan et al. [2012] and Sikdar et al. [2013] group these features in : social (the underlying network), content, and behavior (messages, retweets, mentions). The importance of each feature is topic dependent, such as personal or impersonal [Adali et al., 2012] or event-based or not (e.g., Fukushima Daiichi nuclear disaster [Thomson et al., 2012] or the Chile Earthquake from 2011 [Mendoza et al., 2010]). In particular, with respect to this last kind of topics (i.e., major events), Twitter is a prolific information breeding ground and a new definition of expertise emerges in this context: that of direct witness. Truelove et al. [2015] have a small study in this direction, where they try to classify accounts as Witness (someone who has experienced the event directly), Impact (someone who has been impacted by the event, but did not necessarily experience it), Relay (someone who is just conveying information about the event), or Other.

Given a set of features, an important problem is their usage to predict credibility, utility, newsworthiness, etc. Most of the above mentioned efforts, in addition to exploring the features,also include user- or tweet-centred predictions. For instance, Sikdar et al. [2014] propose two methods for identifying credible information in Twitter. The first one is based on machine learning and attempts to find a predictive model based on network features. Their method is geared towards assessing the credibility of messages. The second method is based on a maximum likelihood formulation and attempts to find messages that are corroborated by independent and reliable sources.

Graph-based approaches are not missing in Twitter-related studies both as a method to identify credible users and as input for further machine learning. TwitterRank [Weng et al., 2010] is an extension of PageRank, where edges in the graph (followers) are traversed taking into account the similarity in the content posted by each user, calculated using a topic modelling approach. Yeniterzi and Callan [2014a] perform a similar study comparing PageRank, HITS, and Topic-Sensitive PageRank, but after a preprocessing that creates a topic-based subgraph, which they refer to as a topic candidate (TC) graph. Based on their experiments, which focus on precision in an IR task, the best approach is using the topic-sensitive PageRank on top of TC graphs.

Finally, as a contributor to quality estimation, and therefore credibility, we need to acknowledge the various studies on Twitter spam. Grier et al. [2010] classify spam-like behaviour in Twitter in four groups:

tweet hijacking (reporting with modification), retweet (bought or otherwise forced), trend initiating (massive bot campaigns), or trend hijacking (using already trending hash-tags or users). They survey a wide number of features, many of them used before in credibility estimates, but also introduce a new one, spam specific: time behaviour. They observe that most spam is bot-driven and even visually imperceptible regularities in the tweets timelines can be detected automatically.

5.2 Community Question Answering (CQA)

Although, differently from the works focused on Twitter, where the term credibility is often directly used, in most of the papers analyzing question answering communities, we mostly encounter two out of four credibility components (expertise, quality) that we proposed at the beginning of this survey. CQA is particularly interesting for the study of credibility as it directly transforms human-to-human knowledge transfer by interposing an semi-transparent interface. On controversial topics for instance, a study of 944 answers from Yahoo! Answers has shown that of the three classical rhetoric strategies (ethos, logos, and pathos), the least used was pathos—the appeal to emotion [Savolainen, 2014]. Nevertheless, Savolainen [2012] also shows that a majority of answers were *failed openings* (a first answer without providing any arguments for or against a claim, 63%) or *mixed* (providing both arguments for and against a claim, 9%), leading to considerable difficulties on the side of a user to assess the credibility of an answer. Such a study, while interesting, is difficult to scale. Su et al. [2010] did try to detect text trustworthiness by incorporating evidentiality features (i.e., the presence of expressions denoting levels of confidence in the answer, such as *probably*, *must*, *certainly*, *think*, *suppose* and others). However, the experiments proved inconclusive with respect to the ability of such keywords to predict the credibility of the answers.

Older studies on general purpose CQA environments such as Yahoo! Answers[2], Naver[3], or Baidu Knows[4], have shown that the quality of an-

[2] `https://answers.yahoo.com/`

[3] `http://www.naver.com/`

[4] `http://zhidao.baidu.com/`

swers in question answering portals is good on average, but the quality of specific answers varies significantly [Su et al., 2007, Jeon et al., 2006]. Most of the studies, but particularly those on technical domain CQA environments (JavaForum [Zhang et al., 2007], StackOverflow [Ginsca and Popescu, 2013], TurboTax [Pal et al., 2012b], ResearchGate [Li et al., 2015]) generally ignore the question of trustworthiness altogether and address only the quality of individual answers and/or the expertise of users.

While the cited studies are limited to the detection of expertise and quality, Jeon et al. [2006] further proved that using the quality estimations in an information retrieval task significantly improved standard retrieval evaluation metrics, such as Mean Average Precision and Precision at R. Like any filtering tool, applying a filter on the quality of the answers has a positive effect on precision and, in this case very small, negative effect on recall.

As in the previous sections, the set of features can be divided in two: content- and user-related. Content features include:

- answer length,
- number of points received,
- click/view/print/recommendation counts
- topical similarity to the question
- language (readability, grammar)
- sentiment
- temporal features (time between question and answer)

User features include:

- fraction of best answers (acceptance ratio)
- number of answers given (activity level)
- profile details (location, description, user name, photo, website)
- direct recommendation by other users/sponsors/editors
- temporal features (frequency, regularity of answers)
- network features

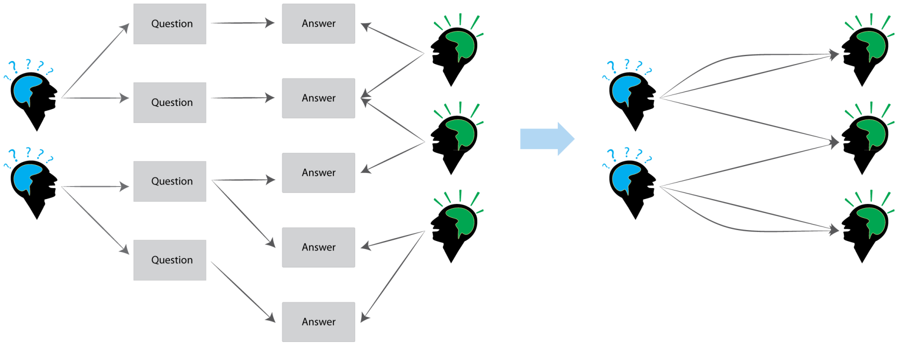


Figure 5.1: Simple user graph construction in question answering

While most of the features can be fairly straightforward mapped to the input of machine learning methods, the network features are particularly interesting here because the network is not based on friendships or follows. Instead, two users are connected via the questions and answers they provide. Therefore, in CQA, there is an added layer of variability: the particularities of the generated network. Figure 5.1 shows a typical CQA network, as a directed multigraph. Jurczyk and Agichtein [2007a] have used it and showed that keeping multiple links in the graph provided better results than reducing it to only one link between two users. However, in their experiments they had not considered weights on the links. Zhang et al. [2007] do so and show similar results. As an example aside, Bian et al. [2009] define their graph keeping Users, Answers, and Questions as distinct nodes and apply graph-based methods on top of this graph. They do so in order to model their mutual reinforcement principle, which formalises the intuition that the quality of the answer and the expertise of the user are interlinked.

Once a graph is built, there are fundamentally two algorithms that are being used or adapted across the literature on community question answering: PageRank [Page et al., 1999] and HITS [Kleinberg, 1999].

Initial work in identifying experts by monitoring communication networks was done at IBM, working with email collections from an enterprise environment. Campbell et al. [2003] and Dom et al. [2003], tested several methods, including PageRank and HITS, but also an adaptation of the positional power, the adequacy, and the successor methods [Herings et al., 2001]. Without going into the details of each method, we note that for this particular type of network, the simpler methods perform worst (adequacy and successor), HITS performs somewhat better, and Power and PageRank dispute the first place, depending on the noise in the data. We can conclude that, for this particular type of communication (emails), there are no hubs and authorities.

This observation changes once we move into web-based CQAs, and several authors show HITS to be outperforming other methods: Zhang et al. [2007] define ExpertiseRank as an extension of HITS including topicality, while Jurczyk and Agichtein [2007b] show that HITS correlates better with user rankings, though only for high-scoring users. That means it is useful for identifying high-expertise users, but not at completely ranking a list of users. This is in itself not an issue, as Zhang et al. [2007] had already observed that human assessors find it equally difficult to give a precise order to the expertise of the users, and rather group them in a few general categories (e.g., expert, professional, user, learner, novice).

Bouguessa et al. [2008] reach quite different conclusions on the relative performance of these algorithms. They compare PageRank, HITS, ExpertiseRank and In-Degree and observe that In-Degree performs best. The essential difference is in the network creation: they only consider a link between two users if the user posing the question marked the answer of the other user as the best answer.

We mentioned above that Bian et al. [2009] defined their graph also slightly differently: not only beyond users, but with all three types of nodes. The benefit of such a model is that it can be used to predict both answer quality and user expertise. Zhang et al. [2014] continue this line of thought and also identify question "interestingness". Using a similar graph to the one defined in Bian et al., Zhang et al. improve on the learning method, removing the necessity of even the small amount of training data needed in the initial study.

Throughout these studies of network topologies, the issue of topicality is constantly present. As it was for Twitter, some users will be more experienced than others on specific topics, and the aim is often to determine the user with the highest expertise related to the question at hand. This can used directly to identify experts and, for instance, Pal et al. [2012b] show that expert and non-expert CQA contributors can be differentiated based on a selection bias that is stable over time. Experts tend to choose questions for which they have a chance to make a valuable contribution. Therefore, observing the variance in the topics of their choice, together with other typical features, one can obtain a reasonable estimate of expertise. Liu et al. [2013a] report that adding domain expertise and user reputation to graph-based features improves expert identification in CQA. However, they only exploit votes given to a user's answers to derive domain expertise and reputation and disregard other relevant user data such as demographic factors or completeness of self-description.

All of the above methods have provided evaluations, implemented either using additional human assessment or with the information already present in the CQA platforms (scores or other indications of acknowledgement produced by the platform itself). Yeniterzi and Callan [2014b] recently showed that this may introduce bias into the evaluation, because assessments given by users are influenced by the system itself (the order in which answers are displayed and the time at which they are given). Whether this is indeed a problem significant enough to cast doubts over the previous results is unlikely. For instance, on the argument of time bias, one may argue that an expert is that much more of an expert if he or she provides answers quicker than others. And if time zones are involved, one may argue that a closer expert is also preferable to one that is many timezones away. Nevertheless, Yeniterzi and Callan [2014b] bring up an important issue, namely that further studies may need to address. They also provide a method to remove some of the effects.

5.3 Summary

In terms of credibility in social media, it is safe to state that the fundamental difference compared with "traditional" media comes from the underlying network. While authority in graphs has been studies for many yeas before social networks, the link between credibility and the new media is defined by the definition of nodes and links. Careful consideration must be given to the meaning of a link between two nodes, in order to be able to say that the result of PageRank or of HITS on such a network is an indicator of credibility.

We have observed only two social media here, Twitter and Community Question Answering. The first one is important because it has taken center stage in many public events, from concerts to catastrophes. In such situations we've observed that the authority of the source is the deciding factor. In more general information distribution, the network topology and individual activity starts playing a more important role, similar to that observed in CQAs. Here, the nature of the data (particularly the length of the texts) allow more in-depth studies of topical similarity, which are combined with network algorithms to identify expert users, but also quality answers and, in one case at least, interesting questions.

6

Multimedia Credibility

Multimedia information is particularly important for credibility because it addresses our senses in a more direct way than text alone. We interact with multimedia in a priori unrestricted ways: seeing, listening, touching, even smelling and the fact that the medium can influence the perception of the message is undisputed [Worchel et al., 1975]. This section does not therefore set out to explore the effects of multimedia content on persuasion or credibility. Instead, it focuses on those methods defined to assist users in their credibility assessment of multimedia content, or to estimate this assessment for a future, unknown user.

As we have seen in preceding sections, the credibility of textual content was already thoroughly studied from different angles. In contrast, we were able to find only a limited quantity of studies dealing with the credibility of multimedia content outside the area of social sciences. While recent advances in processing audio and visual content are impressive, most of existing works dealing with such content focus only on metadata. There is little prior work concerning the combination of textual and visual or audio features for automatic credibility estimation.

Before going on it is worth noting that there is a large body of research in multimedia forensics, where experts investigate both the digital traces left in multimedia content from potential alterations, as well as the use of multimedia in general criminal investigation [Battiato et al., 2012]. We consider this to be an orthogonal research avenue and thus out our immediate scope.

6.1 Video Content Credibility Analysis

Videos is afflicted by the same credibility incertitude as any other type of user generated content. Most of the time, there is no certain information regarding where it was made, who produced it, or what kind of expertise the producer has. While, the range of messages transmitted via video is as wide as that of messages transmitted via text, associated research primarily focuses on news videos. As this is the case for texts, videos are either shared by known and potentially authoritative sources, or by anonymous users. Irrespectively of the source, the credibility of a video is sometimes difficult to be established by the user alone and every so often journalistic scandals bring to attention the use and misuse of videos and images for political or economical gains.

In the news media a kind of ground truth is established by the so called *media watchdogs*: organizations proposing to monitor the information spread through online media content, including videos. Politifact[1] and FactCheck[2] are two examples, but there are others. They address issues of information quality by combing through the media and engaging in fact-checking of news and other media reports. As such, they are comparable with Health on the Net (HON), the organisation mentioned in Section 4 on Domain Specific Credibility that monitors healthcare-related websites. The difference is that while HON provides positive markers (a logo on "credible" websites), media watchdogs provide only negative markers (raise the potential fraud to the public's attention).

For media watchdogs, one method of coping with information quality is harnessing social information processing systems [Lerman, 2007] which seek to filter information and identify quality by aggregating

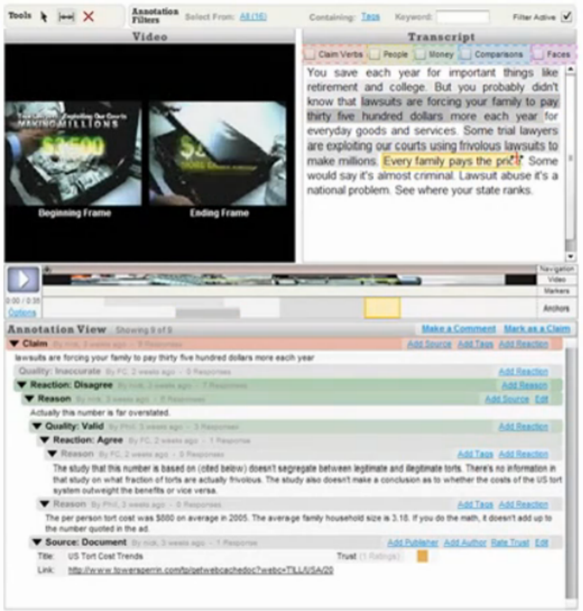
[1] http://www.politifact.com/

[2] http://www.factcheck.org/

the recommendations and ratings of many users through passive (e.g., through usage) or active (e.g., through voting or active rating) metrics of recommendation. Such social or collaborative information can be aggregated to assist video observers. For instance, Diakopoulos et al. [2009] build and study the usefulness of a tool, Videolyzer[3], designed to aid political bloggers and journalists in the activity of watchdog journalism, the process of searching though and evaluating the truthfulness of claims in the media. Videolyzer follows a video quality annotation scheme described by Diakopoulos and Essa [2008], which allows users to collectively analyze the quality of online political videos and then aggregate and share these analyses with others. Users can assess aspects of quality in the video, its transcript and annotations including bias, accuracy, and relevancy that can then be backed up with sources and reasons. We observe here the relation to our credibility framework: bias is understood as trustworthiness, accuracy as reliability and relevance as expertise. Quality, as in the quality of the video (e..g. stability, focus) was not part of the evaluation as no image processing was performed. Figure 6.1 is a sample extracted from the Videolyzer presentation video.

Diakopoulos and Essa [2010] also investigated the effects that visualisations on top of videos propose a video player augmented with simple visuals that indicate aggregated activity levels and polarity of evaluations (i.e., positive / negative) shown in-line with videos as they play. Users were able to interact with the visualization to obtain details including tags, sources, and comments. In order to understand the influence of this visualization on casual video consumption, they evaluate its impact on the credibility of the information presented in the video as compared to a control presentation of the video. They observe that for the negatively annotated videos, the graphic on credibility ratings has a stronger effect on users who engaged more with the graphic. One problem with this study is that it was run with only three videos and observations are therefore difficult to generalize. Further studies are needed to confirm the finding of Diakopoulos and Essa [2010].

[3] http://www.nickdiakopoulos.com/projects/videolyzer-information-quality-analysis-for-videos/



You save each year for important things like retirement and college. But you probably didn't know that lawsuits are forcing your family to pay thirty five hundred dollars more each year for everyday goods and services. Some trial lawyers are exploiting our courts using frivolous lawsuits to make millions. Every family pays the price. Some would say it's almost criminal. Lawsuit abuse it's a national problem. See where your state ranks.

▼ Claim By nick, 3 weeks ago · 9 Responses

lawsuits are forcing your family to pay thirty five hundred dollars more each year

Quality: Inaccurate By FC, 2 weeks ago · 0 Responses

▼ Reaction: Disagree By nick, 3 weeks ago · 7 Responses

▼ Reason By nick, 3 weeks ago · 6 Responses

Actually this number is far overstated.

▼ Quality: Valid By Phil, 3 weeks ago · 3 Responses

▼ Reaction: Agree By FC, 2 weeks ago · 1 Response

▼ Reason By FC, 2 weeks ago · 0 Responses

The study that this number is based on (cited below) doesn't segregate between legitimate and illegitimate torts. There's no information in that study on what fraction of torts are actually frivolous. The study also doesn't make a conclusion as to whether the costs of the US tort system outweight the benefits or vice versa.

▼ Reason By Phil, 3 weeks ago · 0 Responses

The per person tort cost was $880 on average in 2005. The average family household size is 3.18. If you do the math, it doesn't add up to the number quoted in the ad.

▼ Source: Document By nick, 3 weeks ago · 1 Response

Title: US Tort Cost Trends
Trust (1 Ratings)
Link: http://www.towersperrin.com/tp/getwebcachedoc?webc=TILL/USA/20

Figure 6.1: Example extracted from the Videolyzer presentation video.

Besides developing novel visualization methods to highlight credible content in videos, special attention has been given to building automatic methods to predict the quality of the content posted in online video sharing platforms.

Spamming is not restricted to email or Web pages, but also multimedia content. Bulakh et al. [2014] collect a sample of over 3,300 fraudulently promoted YouTube videos and 500 bot profiles that promote them. They characterize fraudulent videos and profiles and train supervised machine learning classifiers that can successfully differentiate fraudulent videos and profiles from legitimate ones. They find that an average fraud video has shorter and fewer comments but is rated higher (4.6 on a 5-point scale when an average legitimate video is rated only at 3.6). Also, the profiles which promote the fraudulent videos, have distinct characteristics: they are relatively new in the system but more active than legitimate profiles, they are more active in viewing and interacting with videos and rarely upload any videos.

Following the trend of the latest research on Web credibility that focuses on the credibility of the users, Benevenuto et al. [2012] aim to detect users who disseminate video pollution, instead of classifying the content itself. They use features that capture the feedback of users with respect to each other or to their contributions to the system (e.g., number of views received), exploiting their interactions through video responses. A machine learning approach is devised that explores the characteristics of manually classified users to create models able to identify spammers and promoters on YouTube. In a complementary approach, O'Callaghan et al. [2012] take advantage of the network of video propagation in YouTube and apply network analysis methods to identify spam campaigns. Content based classification is based on the exploitation of multiple features that are extracted from textual descriptions of the video such as tags, title, textual description and from the video content itself. Boll [2007] finds that these type of features are often robust for the typically low quality of user-generated videos.

So far, all of these methods have used only text (metadata) and network topology to establish the credibility of videos. As such, while they are applied to a different media, they are fundamentally similar to the techniques seen in previous sections. As a rare alternative, Xu et al. [2011] propose a method to evaluate multimedia news by comparing visual descriptions and textual descriptions. They focus their analysis on multimedia news consisting of video clips and their surrounding texts and compute the credibility score of each multimedia news by considering both visual and textual channels. To do this, they first introduce the concept of a stakeholders, i.e., important entities in the event, whose descriptions are supposed to be the most valuable parts for the comparison [Xu et al., 2010]. They then introduce a Material-Opinion model to compare any two of the multimedia news reporting the same event. The credibility score of a video news item consists of material and opinion credibility scores:

- *Material credibility score* is computed based on the idea that high credible material should be used in most items and they support similar opinions.
- *Opinion credibility score* is based on the idea that high credible

opinion should be claimed in many news items by using different materials.

Finally, they provide easily understandable to the users by ranking the multimedia news of the event by their relative credibility scores. To evaluate their method, a manual user credibility ratings (from one to five) of the news items is collected. Results show that the credibility-oriented ranking of the multimedia news correlates with the user ratings, but the size of the study makes it again difficult to claim with reasonable certainty that the method works. Nevertheless, it is one of the rare attempts to bring together video and text processing for the purposes of establishing credibility.

6.2 Credibility of Images

In Section 2 we introduced a unified credibility model that can can be seen as a conceptual framework for domain independent credibility research. We consider each of the credibility components presented in Figure 2.1 to be instantiated differently, according to the each application field of study and we can define credibility through one or several of this components. In this Section, we focus on image related credibility and propose a distinction between credibility of individual images and that of users. This last aspect is central in image sharing platforms since, if modeled and predicted correctly, users' credibility could be an important feature in image retrieval systems. Following the model described in Figure 2.1, we identify each credibility component for user image tagging credibility as follows:

- **trust:** Refers how a user is perceived by the community. Indicators of trust may include the number of users that have him/her among their contacts, or the credits the user receives for his/her photos.
- **expertise:** Either real life photography expertise or validation received by the community. Indicators of expertise may include clues in the user's description (e.g.working for a professional photography institution) or being invited to exclusive curated groups (in the case of Flickr).

- **quality:** For the credibility works dealing with user image tagging credibility, the focus is on the quality of image tagging and not the photos themselves. Imposing this restriction for the term *quality*, an image is considered to have good quality tags if they are relevant to the visual content of the image. We note here the difference from *truthfulness*. For example, a user may tag his images with the type of camera they were taken with or the date when the photos were taken. While true for the user, these tags serve no purpose for describing the content of the image and cannot be used in a retrieval scenario.
- **reliability:** The sustained quality of the annotations provided by the user over time and over different concepts that appear in her photos.

Image tagging credibility is perceived as a mixture of user specific attributes, such as expertise and the quality of her contributions. In this sense, it shares the observation made by Ye and Nov Ye and Nov [2013] who state that researchers need to take a user-centric approach to understanding the dynamics of content contribution in social computing environments. They notably study the connection between different aspects of a user's motivation and both the quantity and quality of his contributions. Their results indicate that users with more social ties, especially ties with people they have not met in the physical world, tend to contribute better content to the community. This type of observation can be quantified as user credibility features and exploited, along with other features, to attribute a credibility score to a user.

To our knowledge, the first attempt to exploit such a score in an image retrieval system is described in [Ginsca et al., 2014], where user credibility is added to more classical textual and visual features. Following this promising result, a user tagging credibility score was added to the datasets used in the MediaEval Retrieving Diverse Social Images 2014 Task [Ionescu et al., 2014]. This score is learned from a combination of features, such as: the degree of correlation between tags and pre-learned visual models of them, the amount of images contributed or the diversity of the tagging vocabulary. Some of the task participants [Calumby et al., 2014, Dang-Nguyen et al., 2014] have improved the relevance and diversity of the submitted retrieval systems by using the provided credibility estimators.

Very recently, Ginsca et al. [2015] propose a novel dataset (MTT-Cred) for evaluating the tagging credibility of Flickr users built with the aim of covering a large variety of topics. They also introduce eight user credibility features that are different from the ones from [Ionescu et al., 2015] which are derived from photo tags, titles or views. Besides the evaluation dataset, two usage scenarios are proposed in order to demonstrate the utility of the concept and stimulate future research. The first one is a user ranking task in which users are grouped in five credibility classes, ranging from *highly not credible* to *highly credible*. A multi-class supervised learning approach is proposed to automatically assign each user to a credibility class. This scenario is similar in spirit with the work that deals with predicting credibility in social media, such as the credibility of tweets [Castillo et al., 2011]. The second scenario is a credible user retrieval task inspired by expert retrieval works. Here the objective is to rank users based on their predicted credibility scores.

Yamamoto and Tanaka [2011b] exploit the same idea of deriving credibility scores from visual and textual associations on the Web. Whereas most of the works presented in this Section so far have dealt with user image tagging credibility, Yamamoto and Tanaka [2011b] propose a system that focuses on the credibility of text-image pairs. They introduce a bipartite graph model for analyzing the credibility of text-image pairs on the Web, in which one set of nodes corresponds to a set of text data, and the other corresponds to a set of images. Each text-image pair is represented by an edge. They introduce the notion of supportive relationships among edges in the bipartite graph model and postulate the hypothesis that the more supportive text-image pairs a target text-image pair has, the more credible it is.

6.3 Credibility of Online Audio Content

Tsagkias et al. [2009] present an ample study on the credibility of podcastsand implement PodCred, a framework that consists of a list of indicators that encode factors influencing listener perceptions of the credibility of podcasts. The work is performed in an information science perspective and the authors consider credibility to be a perceived characteristic of media and media sources that contributes to relevance judgments, as indicated in [Rieh and Danielson, 2007]. They incorporate quality by using an extended notion of credibility that is adapted for the purposes of the podosphere. In the context of the podosphere, similar to the works of Weerkamp and de Rijke [2008, 2012] on the credibility in the blogosphere, the components contributing to user perceptions of credibility, are four-fold. Preference is given to items published by podcasters with expertise, i.e., who are knowledgeable about the subject, who are trustworthy, i.e., have no particular motivation to deceive listeners, that are reliable, i.e., periodic, and of quality. Tsagkias et al. [2009] offer an in-depth analysis of the features used for predicting podcast credibility. The identified components are slightly different from ours: expertise, trustworthiness, *acceptability*, and *attractiveness*. *acceptability* is the "desirability or listener-appeal of a podcast arising from sources other than those that contribute to the believability of its propositional or declarative content". *Attractiveness* is not formally defined, but can be inferred to be a form of surface features. Tsagkias et al. [2009] then propose a wide set of features to assess credibility as whole, without further differentiation in the four sub-areas. Nevertheless, features are grouped into other four categories:

- *Podcast Content*: spoken content (e.g., appearance of on-topic guests, participation of multiple hosts, use of field reports, contains encyclopedic/factual information etc.) and content consistency (e.g., podcast maintains its topical focus across episodes, consistency of episode structure, presence/reliability of inter-episode references, episodes are published regularly etc.);

- *Podcaster*: podcaster speech (e.g., fluency/lack of hesitations, speech rate, articulation/diction, accent), podcaster style (e.g., use of conversational style, use of complex sentence structure, podcaster shares personal details, use of broad, creative vocabulary etc.), podcaster profile (e.g., podcaster scene name, podcaster credentials, podcaster affiliation, podcaster widely known outside the podosphere);

- *Podcast context*: podcaster/listener interaction (e.g., podcaster addresses listeners directly, podcast episodes receive many comments, podcaster responds to comments and requests, podcast page or metadata contains links to related material, podcast has a forum) and real world context (e.g., podcast is a republished radio broadcast, it makes reference to current events, podcast has a store, presence of advertisements etc.);
- *Technical execution*: production (e.g., signature intro/opening jingle, background music , editing effects, studio quality recording/no unintended background noise), packaging (e.g., feed-level metadata present/complete/accurate, episode-level metadata present/complete/accurate , ID3 tags used, audio available in high quality/multiple qualities etc.), distribution (e.g., simple domain name, distributed via distribution platform, podcast has portal or homepage, reliable downloading).

Although some ideas are borrowed from studies of blog credibility, a clear difference between blogs and podcasts is the fact that the core of a podcast is its audio content. Following the classic separation of source and content credibility, as detailed by Rieh and Belkin [1998], message credibility and source credibility overlap to a certain degree and, in the PodCred framework, it can also be noted that certain podcast content indicators could be argued to also be important podcaster credibility indicators.

6.4 Summary

The special attention given to multimedia content in the context of credibility is well justified by psychology studies on the effects of multisensorial perception. However, what we observe in the context of multimedia retrieval is that the focus is still on making sense of non-textual data. Taking this to the next step, of using content features for credibility assessment is, at best, in an incipient phase. It is most advanced in the case of image retrieval, while for video and audio the base is still the metadata analysis. One potential explanation for this situation is that, until recently, the quality content based features lagged behind that of metadata. However, the recent advances, due notably to deep learning, narrowed the gap between the two types of features and could be efficiently used to estimate credibility.

7

Credibility of IR Systems

As an essential tool for information access, IR systems were conspicuously absent from the direct observations of the previous sections. Yet they are, together with the data and the sources, amenable to inquiries of credibility. In this sense, it is interesting to note that Jean et al. [2012] observed that, among many online activities, the use of search engines was both one of the least satisfactory, as well as the one in which users were least confident.

At the beginning of this survey, in the Introduction, we motivated the necessity and opportunity of addressing issues of credibility in computer science via two factors: the predominance of user-generated data on one hand, and the increasing pervasiveness of machine learning on the other hand. All the previous sections have essentially addressed the content and source. This Section addresses the systems. We argue that ever since information retrieval has moved from the boolean model to some form of ranking, now including shallower or deeper neural networks, it has become practically impossible for the user, even an expert one, to precisely predict the outcome of a system given a query. This is, among others, the cause of the unwillingness of professional search systems to separate from the boolean model [Lupu and Hanbury, 2013].

As a product, an IR system can itself be subjected to questions of credibility. Considering the four components of credibility proposed in Section 2, expertise and trustworthiness are applied, as in the case of information, primarily to the source, in this case the producer. Expertise for IR systems is similar to that of any other product: we will find a system and its results more credible if it is provided by a source (producer) with a track record in the industry. Trustworthiness reflects the perception of the user that the search system is not filtering out potentially desirable results (e.g., censorship) or biasing the results according to an unknown agenda (e.g., hidden advertisement). Quality is everything related to system evaluation, starting from effectiveness and efficiency evaluations of the core engine, to the appropriateness for the work task (see Figure 1.2 in Section 1). Finally, reliability is the consistency of the results provided by the system over a range of different information needs of the same kind, of different kinds, or over different data sets.

This last item, reliability, reminds us of the purpose of test collection creation in evaluation campaigns: to provide the means to predict the relative performance of two or more systems, on average, over a set of future information needs of a particular kind. In fact, a lot of the work already done in IR can be cast as a conveyor of credibility in the performance of the system. Here are a few IR research areas and how they can be viewed in terms of credibility.

Probabilistic retrieval together with methods to re-generate probabilities of relevance from retrieval status values (e.g., [Nottelmann and Fuhr, 2003]) are methods to convey to the user more than the set of most relevant documents, but also how relevant these most relevant documents are. The major difficulty here is correctly estimating prior probabilities (e.g., probability of a query, of a document). Automatic explanations have appeared recently to support very specific information needs such as relations between drugs, proteins and medical conditions [Nebot Romero et al., 2011, Callahan and Dumontier, 2012], but recent work has also looked at recommender systems and attempted to provide explanations based on text similarity [Blanco and Lioma, 2012].

Diversity of search results has been studied as a method to increase the probability of a positive user satisfaction [Hawking et al., 2009] but can also be viewed as a proof of impartiality and a tool for the user to assess the IA system itself from a trustworthiness point of view.

Retrievability Azzopardi and Vinay [2008] looks at the core ability of an engine to retrieve documents and as such can be viewed as a measure of reliability (if a set of documents is unintentionally not retrievable under certain conditions), as well as trustworthiness (if a set of documents is intentionally not retrievable).

Evaluation campaigns are a direct measure of reliability of an IR method at a specific task. The difficulty in associating them with credibility as perceived by the users is the general dissociation between consumer search companies and the algorithms tested in the evaluation campaign. However, it has been shown that in professional domains, most notably in legal search, the evaluation campaigns have produced significant adoption increases of modern IR methods [Rowe et al., 2010].

Black-box studies attempt to evaluate IA systems without knowing the implementation details and even without having direct access to the system itself (e.g., access only via a Web interface) [Berendsen et al., 2011]. A part of this evaluation is testing basic quality features such as correctness of the system's behaviour (e.g., retrieving a greater or equal number of documents for a disjunctive query than for any of its components).

All this work can be cast as credibility, primarily through the trustworthiness, quality, and reliability components, but it is only a series of proxies, each independently looking at a different aspect, providing a fragmented image of credibility of IA systems. None of the areas listed above has a track record of assessing credibility directly. This has been done primarily in studies of Human-Computer Interaction (HCI), whose aspects we briefly discussed in § 3.4 (Informing about credibility). For IA systems however, all the others feed more or less directly into HCI. To the best of our knowledge, there is no study on how IR engines assist or influence the perception of credibility in the Information Access system.

7.1 Credibility components in and for IR

Modern Information Retrieval engines rely primarily on two, related, technologies: the Vector Space Model (including matrix transformations such as Principal Component Analysis or Latent Semantic Indexing) and Probabilistic models (most notably the Probabilistic Retrieval Framework and Language Models). A coarse and high-level description of them would be that they use statistical facts about the occurrence of terms to order a set of documents in decreasing order of the likelihood of their relevance to any specific request for information (query). In this process, assumptions are made and models are developed to represent both the information in the documents as well as the information requested, for the ultimate goal of ranking documents based on the match between the two. While the processes involved in this ranking are well understood by the creators of IR algorithms and academics, the users are ultimately presented with an incomplete view of the information space. It is generally difficult for the user to assess either the expertise of the system (its effectiveness, in evaluation terms) or its trustworthiness.

The fundamental scientific question of in this context is therefore: *"How can IR Methods improve or assess the credibility of IA Systems?"* We need to address this problem using a multi-angled approach along several axes, each detailed in one of the following subsections.

7.1.1 Quality

Providing more data to the user to facilitate the quality assessment is a double-edged sword because, in principle, for every new piece of data presented to the user, a new question regarding its credibility may be raised. However, in some cases, the user can take advantage of his or her own understanding of the world to assess the correctness or likelihood of a particular statement. Concept-based IR and automatic explanations are research directions in this sense. Partially, they are already present in some form (e.g., text snippets) or for specific (linked) data, but what is now left to the user to assess, could be done also automatically by the system.

Even without necessarily using concepts, the statistics about the use of terms are considerably more detailed now than they were 30 years ago when the probabilistic retrieval framework was initially developed. A number of assumptions and simplifications have been made in the original models, which have subsequently been further refined in countless works, but ultimately still with the goal of providing the set of most relevant documents, rather than a precise probability of how relevant each one is. Another research path on this axis is therefore going back to probabilities and taking into account the additional (big) data available today.

The return to the basic assumptions of the probabilistic retrieval framework and language models is also the means to automatically estimate the effectiveness of the search for each individual query. This has been researched under the title of predicting query difficulty in the recent literature [Cummins, 2012, Piroi et al., 2012].

Both of the above approaches can be used to enhance the presentation of the search results in order to increase the confidence of the user in the credibility of the information provided to him or her [Schwarz and Morris, 2011].

7.1.2 Reliability and trustworthiness

Reliability assessment as an indicator of credibility is justified in this context by studies showing that users adapt to sub-optimal systems, as long as their weaknesses are consistent. An example often given is that of patent searchers who still primarily use Boolean retrieval systems with very basic ranking functions (or even no ranking at all) [Lupu and Hanbury, 2013], while being overall satisfied with the system's performance. At the same time, they may be unwilling to change to another system because their familiarity with the Boolean model is perceived as more credible than a probabilistic model.

Currently, reliability assessment is done via evaluation campaigns such as CLEF or TREC. The primary effort in designing the benchmarks in these campaigns is ensuring that the observations made about pairs of systems against one such benchmark (e.g., *"System A is better than System B"*) are predictive of their future relative performance against similar benchmarks. Ignoring the timeline, the correlation of the comparative results can be viewed as reliability. However, this assessment is only applicable to the relative judgement of two systems and only as an evaluation of the expertise of a system. An open problem here is the investigation of methods used to assess the reliability of the results provided by one system for similar queries at the same time, or for the same query at different points in time. This is potentially based on work done earlier in the context of the PROMISE project [Berendsen et al., 2011] with the addition of an automatic assessment layer to replace the manual observations made by PROMISE.

Among others, two research points are noteworthy:

1. the definition of similar queries: The question to answer is "When do we expect results showing similar statistical features for two queries?"
2. the expected difference between results of the same system on the same queries at different points in time.

For both of the above, statistical correlation measures may provide the starting point of the study. Additionally, an inconsistent result at the level of one result list may be an indicator of censorship or hidden advertisement, which would fall under the trustworthiness component. Machine learning techniques may be used for extending the work recently done by Blanco and colleagues [Blanco and Lioma, 2012] in determining the likely algorithm used by the search engine, by studying outliers and the probability of these outliers being caused by hidden advertisement or censorship. Automatically detecting censorship is particularly difficult because it implies detecting something never observed.

7.2 New evaluation benchmarks and user studies

While the evaluation campaign benchmarks are, as we said, an indicator of quality and reliability, and therefore credibility, developments of new

IR results are linked to the existence of specific test collections and metrics. Appendix A lists a number of test collections related to data credibility, but there exist no test collections to assess the credibility of an information access system and creating them will be important for the entire community. Unlike existing test collections in IR, which, in theory, provide a set of documents D, a set of queries Q and a function

$$f : D \times Q \to R$$

defining a relevance value for each (document, query) pair, for judging a system we envision the creation of a synthetic results list targeting different evaluation objectives: A first set would be manipulated to change the quality of the results (the expertise of the system), while another set would be manipulated to remove or introduce specific documents to simulate a biased system.

User studies are ultimately needed because credibility is essentially a subjective assessment. This will have to include, or in some way factor out, the human-computer interface. However, the focus of this review lies essentially in the design and testing of automatic methods to assess credibility of IA systems based on IR methods, and therefore at this point we can only direct the reader to the best practice in the field [Sears and Jacko, 2009, Zhang, 2007].

7.3 Summary

The assessment of the quality of results, of the reliability over time, and potentially of the trustworthiness (impartiality of results) of an Information Access system is performed constantly by each user, either consciously or unconsciously. Whether the IR methods and the practices of the IR community themselves can be used to estimate and increase the credibility of an IA environment, as a product on its own, independently of the context it indexes and provides to the user, is still a research question.

8

Conclusions

In compiling this literature survey we have come against the difficulty of defining credibility, particularly in such a way as to make it amenable to automatic assessment and estimation. No two dictionaries of the English language seemed to have the same definition of the term, and the differences are sometimes startling: for instance, while the Cambridge dictionary states it to be *the fact* that someone *can* be believed or trusted, Oxford considers it to be *the quality* that somebody/something has that *makes* people believe or trust them. For the purposes of this review, we take essentially the union of all these definitions and consider all aspects that are reported in the literature to be pertaining to the relation between source, content, and consumer.

In our view, this very broad understanding of credibility can be split across four components: expertise (related to the knowledge of the source), trustworthiness (related to the intent of the source), quality (goodness of fit to a particular purpose of the content), and reliability (consistency or predictability of the quality of the content). The last two are not commonly part of the definition of credibility in the literature. In fact, many other characteristics may be considered. However, we chose these ones because, based on the papers surveyed, they are the most widely used. Even if they are not always denoted by these terms, in the absence of external assessment of expertise and trustworthiness, many studies reduce credibility to a form of quality assessment and, when possible, observe this quality over time.

We have identified four main research directions that tackle different aspects of credibility in information retrieval:

- **effects:** How does credibility affect user search behaviour (§ 3.1)?
- **analysis:** What are the features that users take into account (§ 3.2)?
- **prediction:** How can one devise general methods to use the above features for credibility prediction (§ 3.3)?
- **informing:** Can we inform the users about or otherwise assist them with the assessment of credibility (§ 3.4)?

Overall the survey shows that while studies take different approaches, the common thread is the more or less implicit reliance on the Elaboration Likelihood Model in the sense of dividing credibility perception in two categories:

- central processing route (i.e., content, proven expertise of the source)
- peripheral processing route (e.g., aesthetics, context)

We observe this for various domains, including social media or multimedia. Most surprising, in the case of multimedia, with the exception of images, there is very limited use of non-textual content features. While there are some works that briefly touch upon it in both video and audio data, this is clearly an open research issue.

From a temporal perspective, we noticed that credibility studies have adapted alongside the means of Web data generation (more or less in line with what is referred to as Web 1.0, 2.0, and 3.0). The credibility of the content provider has been an issue ever since the early ages of Web sites. The first problems raised by researchers came from a social perspective and addressed the problem of discovering credibility cues, finding efficient ways to inform but also educate users on the assessment of website credibility. With the increase in popularity of platforms grounded in user generated content, we have observed a proportional response from the Web quality community in terms of credibility studies. Also, with the growth of these new social mediums, we have seen more diversified vocabulary in papers dealing with credibility. One example is the term *influence*, that is mainly encountered when describing users belonging to a community, as opposed to Web sites. This renders finding a unified definition for Web credibility and credibility in IR an even more difficult task. Our choice of the four credibility components studied across this survey is also driven by the appropriateness of their use on most types of Web content. Although in this survey we gave special attention to the newer Web platforms, this choice in describing credibility offers a unified framework for discussing online credibility and bridges the gap between older Web site credibility research and the more recent, social oriented directions.

After Web site credibility studies, the next major ensemble of credibility works covers blogs. Blogs are often given as a reference point for Web 2.0, a term that can be briefly defined as "user generated content Web sites". In this survey, we view blogs as a specific domain of credibility studies. In terms of credibility works, when looking at the nature of cues used for either credibility estimation or prediction, we noticed that blogs are closer to Web sites than social media platforms. As it can be seen from Section 5 on Social Media, a key part in social media credibility research is the social network component. This led to the focus on graph based credibility assessment algorithms and has been used on a variety of platforms, such as Twitter, Facebook, or various Question Answering sites. On the other hand, blog credibility research focuses more on the textual or temporal aspects and less on the blog creator(s).

The third wave of credibility research covers therefor social media. Considering that a large part of credibility works that propose novel methods for the study of credibility come from this area, we dedicated a whole section for social media credibility. Among the several platforms that have been the focus of credibility related studies, we observed that Twitter stands out in regards to the number of works directly targeting credibility. Given that Twitter grew into a fast news provider in recent years, not only for users, but also for media professionals, identifying credible sources became a popular research topic.

With the increased use of image and video sharing features in social media, such as Flickr, Instagram, Facebook, Tumblr, or Youtube, multimedia content has known a rapid growth. However, as discussed in § 6, the study of credibility in the multimedia domain is still in its infancy at this point. It is however starting to receive more attention. For instance, in an article published in July 2015, it has been reported that as many as 8% of Instagram accounts are fake spam bots.[1] Thus, we can expect to see a rise of credibility works dedicated to the multimedia platforms in the following years.

On a technical note, we observed a rising trend in predicting content or source credibility, as opposed to just analysis credibility cues or informing about credibility. This is a natural shift, given the increased popularity of machine learning in a variety of areas related to social computing (e.g., subjectivity in textual data, sentiment analysis, topic recognition, demographic and gender identification, or trend identification). Having machine learning models trained for credibility prediction allows an automatic processing of large volumes of data and we observed that in an IR context they are used either as a filter or as a re-ranking approach.

When looking at the body of referenced works, we noticed the lack of a nucleus of credibility research in IR—a coherent core around which alternatives and extensions are developed. This can be justified by the large diversity of publication venues and the multiple research fields that deal with credibility, either from computer science (e.g., social computing, text mining, multimedia processing, social network analysis, visualization, human-computer interaction) or social sciences (e.g., sociology, political science, communication, journalism). However, we saw that the majority of these works touch different aspects of IR, either as the core topic (e.g., blog retrieval, expert retrieval) or as a component of an IR system (e.g., Web page credibility, Twitter post credibility, user credibility in social image retrieval).

[1] http://www.businessinsider.com/italian-security-researchers-find-8-percent-of-instagram-accounts-are-fake-2015-7?IR=T

Another research issue that is still open is the credibility of IR systems themselves. In the introduction we have argued that because the systems use large amounts of statistical information, a human being, even an expert, will no longer be able to fully understand how the system will behave. This is why in information retrieval we have standardized test collections to assess retrieval performance, and why the results of such tests are not absolute, but relative values. However, for the end user, the question of credibility of the answer provided by the system to the implicit question of “Which are the most relevant documents to my query?” has no clear solution yet. We listed a number of areas that are indirectly connected to credibility, but a direct connection is still to be made.

Appendices

A

Credibility Evaluation Datasets

A number of manually validated ground truth credibility evaluation datasets are readily available. They cover mostly textual data, website metadata but we also describe a recently introduced dataset for credibility evaluation in the multimedia domain. In § A.2, we present evaluation collections that were gathered from Web data with no or minimum human intervention.

A.1 Manually Built Datasets

Table A.1, shows the freely available datasets that are annotated with credibility judgements or were used in credibility related research with little or no alteration.

The Morris Web Credibility corpus contains a dataset of 1,000 URLs that have been manually rated for credibility on a five-point Likert scale. A score of one corresponds to "very non-credible", and five to "very credible". The URL and ratings list are available for download, as well as the page contents as cached at the time of rating. Besides these, additional expert ratings for the 21 pages used in the experiment described in [Schwarz and Morris, 2011] are available (expert raters were

Table A.1: Datasets used in credibility evaluations.

Dataset	Domain	Usage
MPI-SWS	Twitter	Influence detection [Cha et al., 2010], Spam detection [Benevenuto et al., 2010]
Morris Web Credibility	Web pages	Credibility [Schwarz and Morris, 2011]
TREC Blog06	Blogs	Credibility based ranking [Weerkamp and de Rijke, 2008, 2012]
Div150Cred	Flickr images	Landmark image retrieval and diversification [Dang-Nguyen et al., 2014, Calumby et al., 2014]
MTTCred	Flickr images	Generic user tagging credibility prediction [Ginsca et al., 2015]

two medical doctors, two banking and investment professionals, and two presidential political campaign volunteers).

The MPI-SWS[1] Twitter dataset contains 54,981,152 user accounts that were in use in August 2009 and 1,963,263,821 social (follow) links. The 54 million users are connected to each other by 1.9 billion follow links. This is based on the snapshot of the Twitter network topology in August 2009. The follow link data does not contain information about when each link was formed. The dataset also contains 1,755,925,520 tweets. For each of the 54 million users, information about all tweets ever posted by the user since the launch of the Twitter service was gathered. The tweet data contains information about the time each tweet was posted.

TREC Blog06 corpus [Macdonald and Ounis, 2006] has been constructed by monitoring around 100,000 blog feeds for a period of 11 weeks in early 2006, downloading all posts created in this period. For each link (HTML page containing one blog post) the feed id is registered.

[1] http://twitter.mpi-sws.org/

Div150Cred [Ionescu et al., 2015] represents a specially designed dataset that addresses the estimation of user tagging credibility and stems from the 2014 Retrieving Diverse Social Images Task at the MediaEval Benchmarking Initiative [Ionescu et al., 2014]. It provides Flickr photo information (the date the photo was taken, tags, user's id and photo title , the number of times the photo has been displayed, url link of the photo location, GPS coordinates) for about around 300 locations and 685 different users. Each user is assigned a manual credibility score which is determined as the average relevance score of all the user's photos. To obtain these scores, only 50 157 manual annotations are used (on average 73 photos per user).

MTTCred [Ginsca et al., 2015] is an evaluation dataset that provides a manually obtained credibility ground truth for 1009 Flickr users. This ground truth was constituted by assessing the relevance of 50 topically diversified tag-images for each user. A large number of context and content related features are computed and exploited with machine learning techniques to automatically estimate user credibility. The collection is publicly released to facilitate its reuse in the community.

A.2 Automatically Built Datasets

Building representative ground truth corpora with human annotators is a costly and time consuming process. It is common to derive evaluation corpora from existing resources (e.g., online communities, forums, social networks etc.) with minimum processing or intervention. Some of the most used resources for credibility related studies are the following:

- *Epinions*: Epinions is a Web site where users can write reviews about products and assign them a rating. It also allows the users to express their *Web of Trust*, representing users whose reviews and ratings they have consistently found to be valuable and their *Block list*, a list of authors whose reviews they find offensive, inaccurate, or in general not valuable. Works that use corpora built from Epinions data generally study trust propagation [Guha et al., 2004, Bachi et al., 2012, Liu et al., 2008, Massa and Avesani, 2005].

- *Wikipedia*: Wikipedia is the most popular source of encyclopedic information. It was used for many studies, concerning mostly data quality [de la Calzada and Dekhtyar, 2010, Suzuki and Yoshikawa, 2012], but also trust [Bachi et al., 2012, Lucassen and Schraagen, 2010, Rowley and Johnson, 2013] and credibility perceptions [Pirolli et al., 2009].
- *Yahoo! Answers*: Yahoo! Answers[2] is one of the largest question answering communities, with more than one billion posted answers. Most works on Yahoo! Answers derive their corpus by using the community votes over the answers as quality indicators. Research using Yahoo! Answers revolves around quality [Pelleg et al., 2012] , expertise [Kao et al., 2010] and trust [Jeon and Rieh, 2013].
- *StackOverflow*: Stackoverflow[3] is one of the most active and popular CQA platforms that covers a wide area of computer science topics. Similar to Yahoo! Answers, user ratings of answers and questions are used as ground truth quality scores. Works using StackOverflow generally cover the topic of expertise [Pal et al., 2012a, Hanrahan et al., 2012].
- *Websites white/black lists*: The Health on Net Foundation (HON) and Quackwatch[4] rate websites based on how credible they believe the website is. In some works, these lists are used as positive and negative examples for testing different automatic methods for estimating credibility [Sondhi et al., 2012, Aphinyanaphongs and Aliferis, 2007].

[2] http://answers.yahoo.com/

[3] http://stackoverflow.com/

[4] http://www.quackwatch.com/

References

Sibel Adali, Fred Sisenda, and Malik Magdon-Ismail. Actions Speak As Loud As Words: Predicting Relationships from Social Behavior Data. In *Proc. of the International World Wide Web Conference (WWW)*, 2012.

Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding High-quality Content in Social Media. In *Proc. of the International Conference on Web Search and Data Mining (WSDM)*, 2008.

Susumu Akamine, Daisuke Kawahara, Yoshikiyo Kato, Tetsuji Nakagawa, Yutaka I Leon-Suematsu, Takuya Kawada, Kentaro Inui, Sadao Kurohashi, and Yutaka Kidawara. Organizing Information on the Web to Support User Judgments on Information Credibility. In *Proc. of the 4th International Universal Communication Symposium (IUCS)* , 2010.

Suliman Aladhadh, Xiuzhen Zhang, and Mark Sanderson. Tweet Author Location Impacts on Tweet Credibility. In *Proc. of the Australasian Document Computing Symposium*, 2014.

Omar Alonso, Chad Carson, David Gerster, Xiang Ji, and Shubha U. Nabar. Detecting Uninteresting Content in Text Streams. In *Proc. of the the Special Interest Group on Information Retrieval (SIGIR) Crowdsourcing for Search Evaluation Workshop*, 2010.

Farah Alsudani and Matthew Casey. The Effect of Aesthetics on Web Credibility. In *Proc. of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, 2009.

Alia Amin, Junte Zhang, Henriette Cramer, Lynda Hardman, and Vanessa Evers. The Effects of Source Credibility Ratings in a Cultural Heritage Information Aggregator. In *Proc. of the 3rd Workshop on Information Credibility on the Web*, 2009.

Einat Amitay, David Carmel, Nadav Har'El, Shila Ofek-Koifman, Aya Soffer, Sivan Yogev, and Nadav Golbandi. Social Search and Discovery Using a Unified Approach. In *Proc. of the 20th ACM Conference on Hypertext and Hypermedia*, 2009.

Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcroft, Kamal Jain, Vahab Mirrokni, and Shanghua Teng. Robust PageRank and Locally Computable Spam Detection Features. In *Proc. of the 4th International Workshop on Adversarial Information Retrieval on the Web*, 2008.

Yin Aphinyanaphongs and Constantin Aliferis. Text Categorization Models for Identifying Unproven Cancer Treatments on the Web. *Studies in Health Technology and Informatics*, 129(2), 2007.

Aristotle. *Treatise on Rhetoric, Literally Translated from the Greek.* Henry G. Bohn, Theodore Buckley edition, 1857.

Cory L. Armstrong and Melinda J. McAdams. Blogs of Information: How Gender Cues and Individual Motivations Influence Perceptions of Credibility. *Journal of Computer-Mediated Communication*, 14(3):435–456, 2009.

Bernardine M.C. Atkinson. Captology: A Critical Review. In *Persuasive Technology.* Springer, 2006.

Yigal Attali and Jill Burstein. Automated Essay Scoring with E-Rater® V. 2. *The Journal of Technology, Learning and Assessment*, 4(3), 2006.

Julian K. Ayeh, Norman Au, and Rob Law. Do we Believe in TripAdvisor? Examining Credibility perceptions and Online Travelers' Attitude Toward Using User-generated Content. *Journal of Travel Research*, 2013.

Leif Azzopardi and Vishwa Vinay. Accessibility in Information Retrieval. *Advances in Information Retrieval*, 2008.

Giacomo Bachi, Michele Coscia, Anna Monreale, and Fosca Giannotti. Classifying Trust/Distrust Relationships in Online Social Networks. In *Proc. of the International Conference on Privacy, Security, Risk and Trust*, pages 552–557, 2012.

Ricardo Baeza-Yates, Carlos Castillo, Vicente López, and Cátedra Telefónica. PageRank Increase under Different Collusion Topologies. In *Proc. of the International Workshop on Adversarial Information Retrieval on the Web*, 2005.

Krisztian Balog, Maarten De Rijke, and Wouter Weerkamp. Bloggers as Experts: Feed Distillation Using Expert Retrieval Models. In *Proc. of the Special Interest Group on Information Retrieval (SIGIR)*, 2008.

Sebastiano Battiato, Sabu Emmanuel, Adrian Ulges, and Marcel Worring. Multimedia in Forensics, Security, and Intelligence. *IEEE Trans. on Multimedia*, 19(1):17–19, 2012.

Luca Becchetti, Carlos Castillo, Debora Donato, Ricardo Baeza-Yates, and Stefano Leonardi. Link Analysis for Web Spam Detection. *ACM Trans. on the Web (TWEB)*, 2(1):2, 2008.

Irma Becerra-Fernandez. Facilitating the Online Search of Experts at NASA Using Expert Seeker People-finder. In *Proc. of the International Conference on Practical Aspects of Knowledge Management (PAKM)*, 2000.

Fabrício Benevenuto, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, and Marcos Gonçalves. Detecting Spammers and Content Promoters in Online Video Social Networks. In *Proc. of the Special Interest Group on Information Retrieval (SIGIR)*, 2009a.

Fabrício Benevenuto, Tiago Rodrigues, Virgilio Almeida, Jussara Almeida, and Keith Ross. Video Interactions in Online Video Social Networks. *ACM Trans. on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 5(4), 2009b.

Fabrıcio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgılio Almeida. Detecting Spammers on Twitter. In *Proc. of the Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, 2010.

Fabrício Benevenuto, Tiago Rodrigues, Adriano Veloso, Jussara Almeida, Marcos Gonçalves, and Virgílio Almeida. Practical Detection of Spammers and Content Promoters in Online Video Sharing Systems. *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(3):688–701, 2012.

Richard Berendsen, Giorgio Maria Di Nunzio, Maria Gäde, Jussi Karlgren, Mihai Lupu, Stefan Rietberger, and Julianne Stiller. Deliverable 4.1: First Report on Alternative Evaluation Methodology. Technical report, PROMISE Network of Excellence, 2011.

Elmer V. Bernstam, Dawn M. Shelton, Walji Muhammad, and Funda Meric-Bernstam. Instruments to Assess the Quality of Health Information on the World Wide Web: What Can Our Patients Actually Use? *International Journal of Medical Informatics*, 74(1):13–20, 2005.

Jiang Bian, Yandong Liu, Ding Zhou, Eugene Agichtein, and Hongyuan Zha. Learning to Recognize Reliable Users and Content in Social Media with Coupled Mutual Reinforcement. In *Proc. of the International World Wide Web Conference (WWW)*, 2009.

Roi Blanco and Christina Lioma. Graph-based Term Weighting for Information Retrieval. *Information Retrieval*, 15:54–92, 2012.

Susanne Boll. MultiTube–Where Web 2.0 and Multimedia Could Meet. *IEEE Trans. on Multimedia*, 14(1):9–13, 2007.

Mohamed Bouguessa, Benoît Dumoulin, and Shengrui Wang. Identifying Authoritative Actors in Question-answering Forums: The Case of Yahoo! Answers. In *Proc. of Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)*, 2008.

Vlad Bulakh, Christopher W. Dunn, and Minaxi Gupta. Identifying Fraudulently Promoted Online Videos. In *Proc. of the International World Wide Web Conference (WWW)*, 2014.

Jill Burstein and Magdalena Wolska. Toward Evaluation of Writing Style: Finding Overly Repetitive Word Use in Student Essays. In *Proc. of the Conference of the European chapter of the Association for Computational Linguistics*, 2003.

Alison Callahan and Michel Dumontier. Evaluating Scientific Hypotheses Using the SPARQL Inferencing Notation. In *Proc. of the International Conference on The Semantic Web: Research and Applications*, 2012.

Jamie Callan and Maxine Eskenazi. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2007.

Rodrigo T. Calumby, Vinícius P. Santana, Felipe S. Cordeiro, Otávio A.B. Penatti, Lin T. Li, Giovani Chiachia, and Ricardo da S. Torres. Recod@ MediaEval 2014: Diverse Social Images Retrieval. *Working Notes of MediaEval*, 2014.

Christopher S. Campbell, Paul P. Maglio, Alex Cozzi, and Byron Dom. Expertise Identification Using Email Communications. In *Proc. of the International Conference on Information and Knowledge Management (CIKM)*, 2003.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information Credibility on Twitter. In *Proc. of the International World Wide Web Conference (WWW)*. ACM, 2011.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Predicting Information Credibility in Time-Sensitive Social Media. *Internet Research*, 23(5): 560–588, 2013.

Meeyoung Cha, Hamed Haddadi, Fabrıcio Benevenuto, and Krishna P Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proc. of the International Conference on Web and Social Media (ICWSM)*, 2010.

Shelly Chaiken. Heuristic Versus Systematic Information Processing and the Use of Source Versus Message Cues in Persuasion. *Journal of Personality and Social Psychology*, 39(5), 1980.

Meichieh Chen and Toshizumi Ohta. Using Blog Content Depth and Breadth to Access and Classify Blogs. *International Journal of Business and Information*, 5(1):26–45, 2010.

Martin Chodorow and Claudia Leacock. An Unsupervised Method for Detecting Grammatical Errors. In *Proc. of the North American chapter of the Association for Computational Linguistics conference*, 2000.

Choicestream. Choicestream Survey: Consumer Opinion on Online Advertising and Audience Targeting, 2013.

Piotr Cofta. The Trustworthy and Trusted Web. *Foundations and Trends® in Web Science*, 2(4), 2011.

Kevyn Collins-Thompson and Jamie Callan. A Language Modeling Approach to Predicting Reading Difficulty. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2004.

Jack G. Conrad, Jochen L. Leidner, and Frank Schilder. Professional Credibility: Authority on the Web. In *Proc. of the Workshop on Information Credibility on the Web*, 2008.

Gordon V. Cormack. Email Spam Filtering: A Systematic Review. *Foundations and Trends® in Information Retrieval*, 1(4):335–455, 2007.

Tracy Rickman Cosenza, Michael R. Solomon, and Wi-suk Kwon. Credibility in the Blogosphere: A Study of Measurement and Influence of Wine Blogs as an Information Source. *Journal of Consumer Behaviour*, 2014.

Jamie L. Crawford, Cheng Guo, Jessica Schroeder, Rosa I. Arriaga, and Jennifer Mankoff. Is it a Question of Trust?: How Search Preferences Influence Forum Use. In *Proc. of the International Conference on Pervasive Computing Technologies for Healthcare*, 2014.

Ronan Cummins. On the Inference of Average Precision from Score Distributions. In *Proc. of the International Conference on Information and Knowledge Management (CIKM)*, 2012.

Duc-Tien Dang-Nguyen, Luca Piras, Giorgio Giacinto, Giulia Boato, and Francesco De Natale. Retrieval of Diverse Images by Pre-filtering and Hierarchical Clustering. *Working Notes of MediaEval*, 2014.

David R. Danielson. Web Credibility. *Encyclopedia of Human Computer Interaction*, 2006.

Gabriel de la Calzada and Alex Dekhtyar. On Measuring the Quality of Wikipedia Articles. In *Proc. of the Workshop on Information Credibility*, 2010.

Morton Deutsch. *The Resolution of Conflict: Constructive and Destructive Processes.* Yale University Press, 1973.

Nicholas Diakopoulos and Irfan Essa. An Annotation Model for Making Sense of Information Quality in Online Video. In *Proc. of the International Conference on the Pragmatic Web: Innovating the Interactive Society*, 2008.

Nicholas Diakopoulos and Irfan Essa. Modulating Video Credibility via Visualization of Quality Evaluations. In *Proc. of the Workshop on Information Credibility*, 2010.

Nicholas Diakopoulos, Sergio Goldenberg, and Irfan Essa. Videolyzer: Quality Analysis of Online Informational Video for Bloggers and Journalists. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, 2009.

Byron Dom, Iris Eiron, Alex Cozzi, and Yi Zhang. Graph-based Ranking Algorithms for e-mail Expertise Analysis. In *Proc. of the SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2003.

Isabel Drost and Tobias Scheffer. Thwarting the Nigritude Ultramarine: Learning to Identify Link Spam. In *Proc. of the European Conference on Machine Learning*, 2005.

Simon Duncan and Birgit Pfau-Effinge, editors. *Gender, Economy and Culture in the European Union.* Routledge Research in Gender and Society, 2012.

Chad Edwards, Patric R. Spence, Christina J. Gentile, America Edwards, and Autumn Edwards. How Much Klout do You Have… A Test of System Generated Cues on Source Credibility. *Computers in Human Behavior*, 29 (5):A12–A16, 2013.

Tristan Endsley, Yu Wu, and James Reep. The Source of the Story: Evaluating the Credibility of Crisis Information Sources. *Proc. of the Information Systems for Crisis Response and Management Conference*, 2014.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. Cognitively Motivated Features for Readability Assessment. In *Proc. of the Conference of the European Chapter of the Association for Computational Linguistics*, 2009.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A Comparison of Features for Automatic Readability Assessment. In *Proc. of the International Conference on Computational Linguistics*, 2010.

Andrew J. Flanagin and Miriam J. Metzger. The Perceived credibility of Personal Web Page Information as Influenced by the Sex of the Source. *Computers in Human Behavior*, 19(6):683–701, 2003.

Andrew J. Flanagin and Miriam J. Metzger. *Digital Media, Youth, and Credibility*, chapter Digital Media and Youth: Unparalleled Opportunity and Unprecendented Responsibility. MIT Press, 2008a.

Andrew J. Flanagin and Miriam J. Metzger. The Credibility of Volunteered Geographic Information. *GeoJournal*, 72(3-4):137–148, 2008b.

Andrew J. Flanagin, Miriam J. Metzger, Rebekah Pure, Alex Markov, and Ethan Hartsell. Mitigating Risk in E-commerce Transactions: Perceptions of Information Credibility and the Role of User-generated Ratings in Product Quality and Purchase Intention. *Electronic Commerce Research*, 14(1): 1–23, 2014.

B. J. Fogg. *Persuasive Technology: Using Computers to Change What We Think and Do.* Morgan Kaufmann Publishers, 2003.

B. J. Fogg and Hsiang Tseng. The Elements of Computer Credibility. In *Proc. of the SIGCHI Conference on Human factors in computing systems*, 1999.

B. J. Fogg, Leslie Marable, Julianne Stanford, and Ellen. R. Tauber. How do People Evaluate a Web Site's Credibility? Technical report, The Stanford Persuasive Technology Lab, 2002.

Martin Frické, Don Fallis, Marci Jones, and Gianna M. Luszko. Consumer Health Information on the Internet about Carpal Tunnel Syndrome: Indicators of Accuracy. *The American Journal of Medicine*, 118(2), 2005.

Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y. Zhao. Detecting and Characterizing Social Spam Campaigns. In *Proc. of the SIGCOMM Conference on Internet Measurement*, pages 35–47, 2010.

Qin Gao, Ye Tian, and Mengyuan Tu. Exploring Factors Influencing Chinese Users' Perceived Credibility of Health and Safety Information on Weibo. *Computers in Human Behavior*, 45, 2015.

Urs Gasser, Sandra Cortesi, Momin Malik, and Ashley Lee. Youth and Digital Media: From Credibility to Information Quality. Technical Report 2012-1, Berkman Center, 2012.

Alexis Geiber. Digital Divas: Women, Politics and the Social Network. Technical Report D-63, Joan Shorenstein Center on the Press, Cambridge MA, 2011.

Alexandru L. Ginsca and Adrian Popescu. User Profiling for Answer Quality Assessment in Q&A Communities. In *Proc. of the Workshop on Data-driven User Behavioral Modelling and Mining from Social Media*, 2013.

Alexandru L. Ginsca, Adrian Popescu, Bogdan Ionescu, Anil Armagan, and Ioannis Kanellos. Toward an Estimation of User Tagging Credibility for Social Image Retrieval. In *Proc. of the International Conference on Multimedia*, 2014.

Alexandru L. Ginsca, Adrian Popescu, Mihai Lupu, Adrian Iftene, and Ioannis Kanellos. Evaluating user image tagging credibility. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer, 2015.

Jennifer Golbeck. Trust on the World Wide Web: a Survey. *Foundations and Trends® in Web Science*, 1(2):131–197, 2006.

Michael F. Goodchild and Linna Li. Assuring the Quality of Volunteered Geographic Information. *Spatial Statistics*, 1:110–120, 2012.

Nicola J. Gray, Jonathan D. Klein, Peter R. Noyce, Tracy S. Sesselberg, and Judith A. Cantrill. Health Information-seeking Behaviour in Adolescence: the Place of the Internet. *Social Science and Medicine*, 60(7), 2005.

Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @Spam: the Underground on 140 Characters or Less. In *Proc. of the Conference on Computer and Communications Security*, 2010.

Kathleen M. Griffiths, Thanh Tin Tang, David Hawking, and Helen Christensen. Automated Assessment of the Quality of Depression Websites. *Journal of Medical Internet Research*, 7(5), 2005.

Ramanthan Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Propagation of Trust and Distrust. In *Proc. of the International World Wide Web Conference (WWW)*, 2004.

Aditi Gupta and Ponnurangam Kumaraguru. Credibility Ranking of Tweets During High Impact Events. In *Proc. of the Workshop on Privacy and Security in Online Social Media*, 2012.

Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, and Inbal Ronen. Mining Expertise and Interests from Social Media. In *Proc. of the International World Wide Web Conference (WWW)*, 2013.

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating Web Spam with TrustRank. In *Proc. of Very Large Data Bases (VLDB)*, 2004.

Allan Hanbury and Mihai Lupu. Toward a Model of Domain-Specific Search. In *Proc. of the Open research Areas in Information Retrieval (OAIR)*, 2013.

Benjamin V Hanrahan, Gregorio Convertino, and Les Nelson. Modeling Problem Difficulty and Expertise in StackOverflow. In *Proc. of the Conference on Computer Supported Cooperative Work*, 2012.

Vicki L. Hanson. Cognition, Age, and Web Browsing. In *Proc. of Universal Access in HCI*, 2009.

David Hawking, Tom Rowlands, and Paul Thomas. C-TEST: Supporting Novelty and Diversity in TestFiles for Search Tuning. In *Proc. of the Special Interest Group on Information Retrieval (SIGIR)*, 2009.

Marti A. Hearst and Susan T. Dumais. Blogging Together: An Examination of Group Blogs. In *Proc. of the International Conference on Web and Social Media (ICWSM)*, 2009.

Jean-Jacques Herings, Gerard Van der Laan, and Dolf Talman. Measuring the Power of Nodes in Digraphs. *Social Science Research Network*, 2001.

Francis Heylighen and Jean-Marc Dewaele. Variation in the Contextuality of Language: An Empirical Measure. *Foundations of Science*, 7(3):293–340, 2002.

Brian Hilligoss and Soo Young Rieh. Developing a Unifying Framework of Credibility Assessment: Construct, Heuristics, and Interaction in Context. *Information Processing & Management*, 44(4):1467–1484, 2008.

Nurul H. Idris, Mike Jackson, and M.H.I. Ishak. A Conceptual Model of the Automated Credibility Assessment of the Volunteered Geographic Information. In *IOP Conference Series: Earth and Environmental Science*, 2014.

Peter Ingwersen and Kalervo Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context.* Springer, 2005.

Bogdan Ionescu, Adrian Popescu, Mihai Lupu, Alexandru L. Ginsca, and Henning Müller. Retrieving Diverse Social Images at Mediaeval 2014: Challenge, Dataset and Evaluation. In *Proc. of the MediaEval Workshop*, 2014.

Bogdan Ionescu, Adrian Popescu, Mihai Lupu, Alexandru L. Ginsca, Bogdan Boteanu, and Henning Müller. Div150Cred: A Social Image Retrieval Result Diversification with User Tagging Credibility Dataset. In *Proc. of the Multimedia Systems Conference*, 2015.

Melody Y. Ivory and Marti A. Hearst. Statistical Profiles of Highly-rated Web Sites. In *Proc. of the SIGCHI conference on Human factors in computing systems*, 2002.

Wojciech Jaworski, Emilia Rejmund, and Adam Wierzbicki. Credibility Microscope: Relating Web Page Credibility Evaluations to their Textual Content. In *Proc. of the International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2014.

Beth St. Jean, Soo Young Rieh, Yong-Mi Kim, and Ji Yeon Yang. An Analysis of the Information Behaviors, Goals, and Intentions of Frequent Internet Users: Findings from Online Activity Diaries. *First Monday*, 17(2), 2012.

Grace YoungJoo Jeon and Soo Young Rieh. Do You Trust Answers?: Credibility Judgments in Social Search Using Social Q&A Sites. *Social Networks*, 2:14, 2013.

Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, and Soyeon Park. A Framework to Predict the Quality of Answers with Non-textual Features. In *Proc. of the Special Interest Group on Information Retrieval (SIGIR)*, 2006.

Junhui Jiang, Nadee Goonawardene, and Sharon Swee-Lin Tan. Do You Find Health Advice on Microblogging Platforms Credible? Role of Self-efficacy and Health Threat in Credibility Assessment. In *Proc. of Pacific Asia Conference on Information Systems*, 2014.

Nitin Jindal and Bing Liu. Opinion Spam and Analysis. In *Proc. of the International Conference on Web Search and Data Mining (WSDM)*, 2008.

Thomas J. Johnson and Barbara K. Kaye. In Blog we Trust? Deciphering Credibility of Components of the Internet Among Politically Interested Internet Users. *Computers in Human Behavior*, 25(1):175–182, 2009.

Thomas J. Johnson and Barbara K. Kaye. Credibility of Social Network Sites for Political Information Among Politically Interested Internet Users. *Journal of Computer-Mediated Communication*, 19(4):957–974, 2014.

Thomas J. Johnson and David D. Perlmutter. The Facebook Election. *Mass Communication and Society*, 2010.

Thomas J. Johnson, Barbara K. Kaye, Shannon L. Bichard, and W. Joann Wong. Every Blog Has Its Day: Politically-interested Internet Users' Perceptions of Blog Credibility. *Journal of Computer-Mediated Communication*, 13(1):100–122, 2007.

Andreas Juffinger, Michael Granitzer, and Elisabeth Lex. Blog Credibility Ranking by Exploiting Verified Content. In *Proc. of the 3rd Workshop on Information credibility on the web*. ACM, 2009.

Pawel Jurczyk and Eugene Agichtein. Discovering Authorities in Question Answer Communities by Using Link Analysis. In *Proc. of the International Conference on Information and Knowledge Management (CIKM)*, 2007a.

Pawel Jurczyk and Eugene Agichtein. Hits on Question Answer Portals: Exploration of Link Analysis for Author Ranking. In *Proc. of the Special Interest Group on Information Retrieval (SIGIR)*, 2007b.

Wei-Chen Kao, Duen-Ren Liu, and Shiu-Wen Wang. Expert Finding in Question-Answering Websites: a Novel Hybrid Approach. In *Proc. of the Symposium on Applied Computing*, 2010.

Ahmad Kardan, Mehdi Garakani, and Bamdad Bahrani. A Method to Automatically Construct a User Knowledge Model in a Forum Environment. In *Proc. of the Special Interest Group on Information Retrieval (SIGIR)*, 2010.

Farid Karimipour and Omid Azari. Citizens as Expert Sensors: One Step Up on the VGI Ladder. In *Progress in Location-Based Services 2014*. Springer, 2015.

Yukiko Kawai, Yusuke Fujita, Tadahiko Kumamoto, Jianwei Jianwei, and Katsumi Tanaka. Using a Sentiment Map for Visualizing Credibility of News Sites on the Web. In *Proc. of the 2nd ACM Workshop on Information Credibility on the Web*. ACM, 2008.

Maria Keskenidou, Argyris Kyridis, Lina P. Valsamidou, and Alexandra-Helen Soulani. The Internet as a Source of Information. The Social Role of Blogs and Their Reliability. *Observatorio (OBS*)*, 8(1), 2014.

Carsten Keßler and René Theodore Anton de Groot. Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap. In *Geographic Information Science at the Heart of Europe*, pages 21–37. Springer, 2013.

Heejun Kim. *Credibility Assessment of Volunteered Geographic Information for Emergency Management: a Bayesian Network Modeling Approach*. PhD thesis, University of Illinois at Urbana-Champaign, 2013.

Paul Kim, Thomas R. Eng, Mary Jo Deering, and Andrew Maxfield. Published Criteria for Evaluating Health Related Web Sites: Review. *Bmj*, 318(7184): 647–649, 1999.

Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

Pranam Kolari, Tim Finin, Kelly Lyons, and Yelena Yesha. Expert Search Using Internal Corporate Blogs. In *Proc. of the Special Interest Group on Information Retrieval (SIGIR) Workshop: Future Challenges in Expertise Retrieval*, 2008.

Petros Kostagiolas, Nikolaos Korfiatis, Panos Kourouthanasis, and Georgios Alexias. Work-related Actors Influencing Doctors Search Behaviors And Trust Toward Medical Information Resources. *International Journal of Information Management*, 34(2):80–88, 2014.

Roderic M. Kramer and Tom R. Tyler. *Trust in Organizations: Frontiers of Theory and Research.* Sage Publications, Inc., 1996.

R. David Lankes. Trusting The Internet: New Approaches To Credibility Tools. *The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning*, pages 101–121, 2007.

Jonathan Lazar, Gabriele Meiselwitz, and Jinjuan Feng. Understanding Web Credibility: A Synthesis Of The Research Literature. *Foundations and Trends® in Human-Computer Interaction*, 1(2), 2007.

Reeva Lederman, Hanmei Fan, Stephen Smith, and Shanton Chang. Who Can You Trust? Credibility Assessment In Online Health Forums. *Health Policy and Technology*, 3(1):13–25, 2014.

Kyumin Lee, James Caverlee, and Steve Webb. Uncovering Social Spammers: Social Honeypots+ Machine Learning. In *Proc. of the Special Interest Group on Information Retrieval (SIGIR)*, 2010.

Amanda Lenhart, Mary Madden, Aaron Smith, Kristen Purcell, Kathryn Zickuhr, and Lee Rainie. Teens, Kindness, And Cruelty On Social Network Sites: How American Teens Navigate The New World Of "digital Citizenship". Technical report, Pew Internet and American Life Project, 2011.

Kristina Lerman. Social Information Processing In News Aggregation. *IEEE Internet Computing*, 11(6):16–28, 2007.

Lei Li, Daqing He, Wei Jeng, Spencer Goodwin, and Chengzhi Zhang. Answer Quality Characteristics and Prediction on an Academic Q&A Site: A Case Study on ResearchGate. In *Proc. of the International World Wide Web Conference (WWW)*, 2015.

Vera Liao and Wai-Tat Fu. Age Differences in Credibility Judgments of Online Health Information. *ACM Trans. on Computer-Human Interaction*, 21(1): 2:1–2:23, 2014.

Duen-Ren Liu, Yu-Hsuan Chen, Wei-Chen Kao, and Hsiu-Wen Wang. Integrating Expert Profile, Reputation And Link Analysis For Expert Finding In Question-answering Websites. *Information Processing & Management*, 49(1):312–329, 2013a.

Haifeng Liu, Ee-Peng Lim, Hady W Lauw, Minh-Tam Le, Aixin Sun, Jaideep Srivastava, and Young Kim. Predicting Trusts Among Users Of Online Communities: An Epinions Case Study. In *Proc. of the 9th ACM conference on Electronic commerce*. ACM, 2008.

Xiaoyong Liu, Bruce W. Croft, and Matthew Koll. Finding Experts In Community-based Question-answering Services. In *Proc. of the International Conference on Information and Knowledge Management (CIKM)*, 2005.

Xin Liu, Radoslaw Nielek, Adam Wierzbicki, and Karl Aberer. Defending Imitating Attacks in Web Credibility Evaluation Systems. In *Proc. of the International World Wide Web Conference (WWW)*, 2013b.

Rui Lopes and Luis Carriço. On The Credibility Of Wikipedia: An Accessibility Perspective. In *Proc. of the 2nd ACM Workshop on Information Credibility on the Web*. ACM, 2008.

Paul Benjamin Lowry, David W. Wilson, and William L. Haig. A Picture Is Worth A Thousand Words: Source Credibility Theory Applied To Logo And Website Design For Heightened Credibility And Consumer Trust. *International Journal of Human-Computer Interaction*, 30(1):63–93, 2014.

Teun Lucassen and Jan Maarten Schraagen. Trust in Wikipedia: How Users Trust Information From an Unknown Source. In *Proc. of the 4th Workshop on Information Credibility*, pages 19–26. ACM, 2010.

Teun Lucassen and Jan Maarten Schraagen. Factual Accuracy And Trust In Information: The Role Of Expertise. *Journal of the American Society for Information Science and Technology*, 62(7):1232–1242, 2011.

Teun Lucassen, Rienco Muilwijk, Matthijs L Noordzij, and Jan Maarten Schraagen. Topic Familiarity And Information Skills In Online Credibility Evaluation. *Journal of the American Society for Information Science and Technology*, 64(2):254–264, 2013.

N. Luhmann. Familiarity, Confidence, Trust: Problems and Alternatives. In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*. University of Oxford, 1988.

Chuan Luo, Xin Robert Luo, Laurie Schatzberg, and Choon Ling Sia. Impact of Informational Factors on Online Recommendation Credibility: The Moderating Role of Source Credibility. *Decision Support Systems*, 2013.

Mihai Lupu and Allan Hanbury. Patent Retrieval. *Foundations and Trends® in Information Retrieval*, 7(1), 2013.

Craig Macdonald and Iadh Ounis. The Trec Blogs06 Collection: Creating And Analysing A Blog Test Collection. *Department of Computer Science, University of Glasgow Technical Report TR-2006-224*, 1:3–1, 2006.

Michael J. Manfredo and Alan D. Bright. A Model For Assessing The Effects Of Communication On Recreationists. *Journal of Leisure Research*, 1991.

Paolo Massa and Paolo Avesani. Controversial Users Demand Local Trust Metrics: An Experimental Study On Epinions.com Community. In *Proc. of the National Conference on Artificial Intelligence*, 2005.

G Harry McLaughlin. SMOG Grading: A New Readability Formula. *Journal of Reading*, 12(8):639–646, 1969.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter Under Crisis: Can We Trust What We RT? In *Proc. of the First Workshop on Social Media Analytics*. ACM, 2010.

D. Metlay. Institutional Trust and Confidence: a Journey into a Conceptual Quagmire. In G. Cvetkovich and R. Loefstedt, editors, *Social Trust and the Management of Risk*. Earthscan, 1999.

Miriam J. Metzger. Making Sense Of Credibility On The Web: Models for Evaluating Online Information and Recommendations for Future Research. *Journal of the American Society for Information Science and Technology*, 58(13):2078–2091, 2007.

Miriam J. Metzger, Andrew J. Flanagin, Keren Eyal, Daisy R. Lemus, and Robert M. Mccann. Chapter 10: Credibility for the 21st Century: Integrating Perspectives on Source, Message, and Media Credibility in the Contemporary Media Environment. *Communication Yearbook*, 27:293–335, 2003.

Gilad Mishne. Using Blog Properties to Improve Retrieval. *Proc. of the International Conference on Web and Social Media (ICWSM)*, 2007.

Gilad Mishne and Natalie Glance. Leave a Reply: An Analysis of Weblog Comments. In *Proc. of the International World Wide Web Conference (WWW)*, 2006a.

Gilad Mishne and Natalie Glance. Leave a Reply: An Analysis Of Weblog Comments. In *The 3rd Annual Workshop on the Weblogging Ecosystem*, 2006b.

Subhabrata Mukherjee, Gerhard Weikum, and Cristian Danescu-Niculescu-Mizil. People on Drugs: Credibility of User Statements in Health Communities. In *Proc. of Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)*, 2014.

Koji Murakami, Eric Nichols, Suguru Matsuyoshi, Asuka Sumida, Shouko Masuda, Kentaro Inui, and Yuji Matumoto. Statement Map: Assisting Information Crediblity Analysis by Visualizing Arguments. In *Proc. of the 3rd Workshop on Information Credibility on the Web*. ACM, 2009.

Seth A. Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information Network or Social Network?: The Structure of the Twitter Follow Graph. In *Proc. of the International World Wide Web Conference (WWW)*, 2014.

Victoria Nebot Romero, Min Ye, Mario Albrecht, Jae-Hong Eom, and Gehard Weikum. DIDO: A Disease-determinants Ontology From Web Sources. In *Proc. of the International World Wide Web Conference (WWW)*, 2011.

Eric Nichols, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. Constructing a Scientific Blog Corpus for Information Credibility Analysis. In *Proc. of the Annual Meeting of the Association for Neuro-Linguistic Programming (ANLP)*, 2009.

Radoslaw Nielek, Aleksander Wawer, Michal Jankowski-Lorek, and Adam Wierzbicki. Temporal, Cultural and Thematic Aspects of Web Credibility. In *Social Informatics*, pages 419–428. Springer, 2013.

Michael G Noll, Ching-man Au Yeung, Nicholas Gibbins, Christoph Meinel, and Nigel Shadbolt. Telling Experts from Spammers: Expertise Ranking in Folksonomies. In *Proc. of the Special Interest Group on Information Retrieval (SIGIR)*, 2009.

H. Nottelmann and N. Fuhr. From Retrieval Status Values to Probabilities of Relevance for Advanced IR Applications. *Information Retrieval*, 6, 2003.

Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting Spam Web Pages Through Content Analysis. In *Proc. of the International World Wide Web Conference (WWW)*, 2006.

Derek O'Callaghan, Martin Harrigan, Joe Carthy, and Pádraig Cunningham. Network Analysis of Recurring YouTube Spam Campaigns. In *Procs. of the International Conference on Web and Social Media (ICWSM)*, 2012.

John ODonovan, Byungkyu Kang, Greg Meyer, Tobias Hollerer, and Sibel Adalii. Credibility in Context: An Analysis of Feature Distributions in Twitter. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 2012.

Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. Web Credibility: Features Exploration and Credibility Prediction. In *Proc. of the Annual European Conference on Information Retrieval (ECIR)*, 2013.

Frank O. Ostermann and Laura Spinsanti. A Conceptual Workflow for Automatically Assessing the Quality of Volunteered Geographic Information for Crisis Management. In *Proc. of the Annual Association of Geographic Information Laboratories for Europe*, 2011.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report 1999-66, Stanford InfoLab, November 1999. URL `http://ilpubs.stanford.edu:8090/422/`.

Aditya Pal and Scott Counts. Identifying Topical Authorities in Microblogs. In *Proc. of the International Conference on Web Search and Data Mining (WSDM)*, 2011.

Aditya Pal, Shuo Chang, and Joseph A. Konstan. Evolution of Experts in Question Answering Communities. In *Proc. of the International AAAI Conference on Weblogs and Social Media*, 2012a.

Aditya Pal, F. Maxwell Harper, and Joseph A. Konstan. Exploring Question Selection Bias to Identify Experts and Potential Experts in Community Question Answering. *ACM Trans. on Information Systems*, 30(2):10:1–10:28, 2012b.

Thanasis G. Papaioannou, Karl Aberer, Katarzyna Abramczuk, Paulina Adamska, and Adam Wierzbicki. Game-theoretic Models of Web Credibility. In *Proc. of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality.* ACM, 2012.

Heelye Park, Zheng Xiang, Bharath Josiam, and Haejung Kim. Personal Profile Information as Cues of Credibility in Online Travel Reviews. *Anatolia*, 25(1):13–23, 2014.

Jeff Pasternack and Dan Roth. Latent Credibility Analysis. In *Proc. of the International World Wide Web Conference (WWW)*, 2013.

G. L. Patzer. Source Credibility As a Function of Communicator Physical Attractiveness. *Journal of Business Research*, 11(2), 1983.

Dan Pelleg, Elad Yom-Tov, and Yoelle Maarek. Can You Believe an Anonymous Contributor? On Truthfulness in Yahoo! Answer. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Social Computing (SocialCom)*. IEEE, 2012.

Sarah E. Petersen and Mari Ostendorf. A Machine Learning Approach to Reading Level Assessment. *Computer Speech & Language*, 23(1):89–106, 2009.

G. Peterson, P. Aslani, and K. A. Williams. How Do Consumers Search for and Appraise Information on Medicines on the Internet? A Qualitative Study Using Focus Groups. *Journal of Medical Internet Research*, 5(4), 2006.

Richard E. Petty and John T. Cacioppo. The Elaboration Likelihood Model of Persuasion. *Advances in Experimental Social Psychology*, 19, 1986.

Pew Research Center. Internet Gains on Television as Public's Main News Source. Technical report, The Pew Research Center for the People and the Press, 2011.

Pew Research Center. Emerging Nations Embrace Internet, Mobile Technology. http://www.pewglobal.org/2014/02/13/emerging-nations-embrace-internet-mobile-technology/, February 2014.

Florina Piroi, Mihai Lupu, and Allan Hanbury. Effects of Language and Topic Size in Patent IR: An Empirical Study. In *Proc. of the Conference and Labs of the Evaluation Forum (CLEF)*, 2012.

Peter Pirolli, Evelin Wollny, and Bongwon Suh. So You Know You're Getting the Best Possible Information: A Tool that Increases Wikipedia Credibility. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 2009.

Emily Pitler and Ani Nenkova. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proc. of the Conference on Empirical Methods on Natural Language Processing (EMNLP).* Association for Computational Linguistics, 2008.

Susan L. Price and William R. Hersh. Filtering Web Pages for Quality Indicators: An Empirical Approach to Finding High Quality Consumer Health Information on the World Wide Web. In *Proc. of the AMIA Symposium.* American Medical Informatics Association, 1999.

Maria Rafalak, Katarzyna Abramczuk, and Adam Wierzbicki. Incredible: Is (Almost) All Web Content Trustworthy? Analysis of psychological factors related to website credibility evaluation. In *Proc. of the International World Wide Web Conference (WWW)*, 2014.

John D. Ramage and John C. Bean. *Guide to Writing.* Allyn & Bacon, 4th edition, 1998.

Soo Young Rieh. Judgment of Information Quality and Cognitive Authority in the Web. *Journal of the American Society for Information Science and Technology*, 53(2):145–161, 2002.

Soo Young Rieh. Credibility and Cognitive Authority of Information. *Encyclopedia of Library and Informaitn Sciences, 3rd Ed.*, 2010.

Soo Young Rieh and Nicholas J. Belkin. Understanding Judgment of Information Quality and Cognitive Authority in the WWW. In *Proc. of the 61st Annual Meeting of the American Society for Information Science*, volume 35. Citeseer, 1998.

Soo Young Rieh and David R. Danielson. Credibility: A Multidisciplinary Framework. *Annual Review of Information Science and Technology*, 41(1): 307–364, 2007.

Soo Young Rieh, Grace YoungJoo Jeon, Ji Yeon Yang, and Christopher Lampe. Audience-aware Credibility: From Understanding Audience to Establishing Credible Blogs. In *Proc. of the International Conference on Web and Social Media (ICWSM)*, 2014.

Thomas S. Robertson and John R. Rossiter. Children and Commercial Persuasion: An Attribution Theory Analysis. *Journal of Consumer Research*, 1(1), 1974.

Ronald W. Rogers. A Protection Motivation Theory of Fear Appeals and Attitude Change. *The Journal of Psychology: Interdisciplinary and Applied*, 1975.

B. Rowe, D. Wood, A. Link, and D. Simoni. Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program. Technical report, National Institute of Standards and Technology, 2010.

Jennifer Rowley and Frances Johnson. Understanding Trust Formation in Digital Information Sources: The Case of Wikipedia. *Journal of Information Science*, 39(4):494–508, 2013.

Victoria L. Rubin and Elizabeth D. Liddy. Assessing the Credibility Of Weblogs. In *Proc. of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (CAAW)*, 2006.

Lawrence M. Rudner and Tahung Liang. Automated Essay Scoring Using Bayes' Theorem. *The Journal of Technology, Learning and Assessment*, 1 (2), 2002.

Jan Rybak, Krisztian Balog, and Kjetil Nørvåg. Temporal Expertise Profiling. In *Advances in Information Retrieval*, pages 540–546. Springer, 2014.

Luis Sanz, Héctor Allende, and Marcelo Mendoza. Text Content Reliability Estimation in Web Documents: A New Proposal. *Computational Linguistics and Intelligent Text Processing*, pages 438–449, 2012.

Reijo Savolainen. The Structure of Argument Patterns on a Social Q&A Site. *Journal of the Association for Information Science and Technology*, 63(12): 2536–2548, 2012.

Reijo Savolainen. The Use of Rhetorical Strategies in Q&A Discussion. *Journal of Documentation*, 70(1):93–118, 2014.

Julia Schwarz and Meredith Morris. Augmenting Web Pages and Search Results to Support Credibility assessment. In *Proc. of the Special Interest Group on Computer–Human Interaction (SIGCHI)*, 2011.

Andrew Sears and Julie A. Jacko. *Human-Computer Interaction: Fundamentals*. CRC Press, 2009.

Linda See, Alexis Comber, Carl Salk, Steffen Fritz, Marijn van der Velde, Christoph Perger, Christian Schill, Ian McCallum, Florian Kraxner, and Michael Obersteiner. Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-experts. *PloS one*, 8(7):e69958, 2013.

Hansi Senaratne, Arne Bröring, and Tobias Schreck. *Assessing the Credibility of VGI Contributors Based on Metadata and Reverse Viewshed Analysis: An Experiment with Geotagged Flickr Images*. Bibliothek der Universität Konstanz, 2013.

DongBack Seo and Jung Lee. Experts versus Friends: To Whom Do I Listen More? The Factors That Affect Credibility of Online Information. In *HCI in Business*, pages 245–256. Springer, 2014.

Shafiza Mohd Shariff, Xiuzhen Zhang, and Mark Sanderson. User Perception of Information Credibility of News on Twitter. In *Advances in Information Retrieval*, pages 513–518. Springer, 2014.

Ben Shneiderman. Designing Trust into Online Experiences. *Communications of the ACM*, 43(12):57–59, 2000.

Ben Shneiderman. Building Trusted Social Media Communities: A Research Roadmap for Promoting Credible Content. In *Roles, Trust, and Reputation in Social Media Knowledge Markets*, pages 35–43. Springer, 2015.

Luo Si and Jamie Callan. A Statistical Model for Scientific Readability. In *Proc. of the International Conference on Information and Knowledge Management (CIKM)*. ACM, 2001.

Sujit Sikdar, Sarp Adali, M. Amin, Tarek Abdelzaher, Kap Luk Chan, Ji-Haeng Cho, Bing Kang, and John O'Donovan. Finding True and Credible Information on Twitter. In *Information Fusion (FUSION), 2014 17th International Conference on*. IEEE, 2014.

Sujoy Sikdar, Byungkyu Kang, John O'Donovan, Tobias Hollerer, and Sibel Adah. Understanding Information Credibility on Twitter. In *2013 International Conference on Social Computing (SocialCom)*. IEEE, 2013.

Judith Simon. *Knowing Together: A Social Epistemology for Socio- Technical Epistemic Systems*. PhD thesis, Universitaet Wien, 2010.

Parikshit Sondhi, V. Vydiswaran, and ChengXiang Zhai. Reliability Prediction of Webpages in the Medical Domain. *Advances in Information Retrieval*, pages 219–231, 2012.

Seth E. Spielman. Spatial Collective Intelligence? Credibility, Accuracy, and Volunteered Geographic Information. *Cartography and Geographic Information Science*, 41(2):115–124, 2014.

Kritsada Sriphaew, Hiroya Takamura, and Manabu Okumura. Cool Blog Identification Using Topic-Based Models. In *Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1. IEEE, 2008.

Julianne Stanford, Ellen R. Tauber, B. J. Fogg, and Leslie Marable. Experts vs. Online Consumers: A Comparative Credibility Study of Health and Finance Web Sites. *Consumer Web Watch Research Report*, 2002.

Veronika Stefanov, Alexander Sachs, Marlene Kritz, Matthias Samwald, Manfred Gschwandtner, and Allan Hanbury. A Formative Evaluation of a Comprehensive Search System for Medical Professionals. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138 of *Lecture Notes in Computer Science*, pages 81–92. Springer Berlin Heidelberg, 2013.

Qi Su, Dmitry Pavlov, Jyh-Herng Chow, and Wendell C. Baker. Internet-scale Collection of Human-reviewed Data. In *Proc. of the International World Wide Web Conference (WWW)*, 2007.

Qi Su, Chu-Ren Huang, and Helen Kai-yun Chen. Evidentiality for Text Trustworthiness Detection. In *Proc. of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*. Association for Computational Linguistics, 2010.

Shyam S. Sundar. The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. In M. J. Metzger and A. J. Flanagin, editors, *Digital Media, Youth, and Credibility*, The John D. and Cathering T. MacArthur Foundation Series on Digital Media and Learning. MIT Press, 2008.

Yu Suzuki and Masatoshi Yoshikawa. QualityRank: Assessing Quality of Wikipedia Articles by Mutually Evaluating Editors and Texts. In *Proc. of the Conference on Hypertext and Social Media*, 2012.

David Talbot. African Entrepreneurs Deflate Google's Internet Balloon Idea. *MIT Technology Review*, 2013.

Adam Thomason. Blog Spam: A Review. In *Conference on Email and Anti-Spam (CEAS)*, 2007.

Robert Thomson, Naoya Ito, Hinako Suda, Fangyu Lin, Yafei Liu, Ryo Hayasaka, Ryuzo Isochi, and Zian Wang. Trusting Tweets: The Fukushima Disaster and Information Source Credibility on Twitter. In *Proc. of the 9th International Conference on Information Systems for Crisis Response and Management*, 2012.

Catalina L. Toma. Counting on Friends: Cues to Perceived Trustworthiness in Facebook Profiles. In *Proc. of the International Conference on Web and Social Media (ICWSM)*, 2014.

Marie Truelove, Maria Vasardani, and Stephan Winter. Towards Credibility of Micro-blogs: Characterising Witness Accounts. *GeoJournal*, 80(3):339–359, 2015.

Manos Tsagkias, Martha Larson, and Maarten De Rijke. Predicting Podcast Preference: An Analysis Framework and Its Application. *Journal of the American Society for Information Science and Technology*, 61(2):374–391, 2009.

Shawn Tseng and B. J. Fogg. Credibility and Computing Technology. *Communications of the ACM*, 42(5):39–44, 1999.

E. Ullmann-Margalit. Trust, Distrust and in Between. In *Discussion Paper Series from Center for Rationality and Interactive Decision Theory.* Hebrew Universityw, Jerusalem., 2001.

US Census. Educational Attainment in the United States: 2014. http://www.census.gov/hhes/socdemo/education/data/cps/2014/tables.html, 2015.

Nancy Van House. Weblogs: Credibility and Collaboration in an Online World. In *Computer Supported Cooperative Work Workshop*, 2004.

Aleksander Wawer, Radoslaw Nielek, and Adam Wierzbicki. Predicting Webpage Credibility Using Linguistic Features. In *Proc. of the International World Wide Web Conference (WWW)*, 2014.

Wouter Weerkamp and Maarten de Rijke. Credibility Improves Topical Blog Post Retrieval. In *46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2008.

Wouter Weerkamp and Maarten de Rijke. Credibility-inspired Ranking for Blog Post Retrieval. *Information Retrieval*, pages 1–35, 2012.

Markus Weimer, Iryna Gurevych, and Max Mühlhäuser. Automatically Assessing the Post Quality in Online Discussions on Software. In *Proc. of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, 2007.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. TwitterRank: Finding Topic-sensitive Influential Twitterers. In *Proc. of the International Conference on Web Search and Data Mining (WSDM)*, 2010.

David Westerman, Patric R. Spence, and Brandon Van Der Heide. Social Media as Information Source: Recency of Updates and Credibility of Information. *Journal of Computer-Mediated Communication*, 19(2):171–183, 2014.

Stephen Worchel, Virginia Andreoli, and Joe Eason. Is the Medium the Message? A Study of the Effects of Media, Communicator, and Message Characteristics on Attitude Change. *Journal of Applied Social Psychology*, 5(2): 157–172, 1975.

Ching-Tung Wu, Kwang-Ting Cheng, Qiang Zhu, and Yi-Leh Wu. Using Visual Features for Anti-spam Filtering. In *Proc. of the International Conference on Image Processing*, 2005.

E. Wyatt. Most of U.S. Is Wired, but Millions Aren't Plugged In. *The New York Times*, August 18 2013.

Hui Jimmy Xie, Li Miao, Pei-Jou Kuo, and Bo-Youn Lee. Consumers' Responses to Ambivalent Online Hotel Reviews: The Role of Perceived Source Credibility and Pre-decisional Disposition. *International Journal of Hospitality Management*, 30(1):178–183, 2011.

Ling Xu, Qiang Ma, and Masatoshi Yoshikawa. A Cross-media Method of Stakeholder Extraction for News Contents Analysis. In *Web-Age Information Management*, pages 232–237. Springer, 2010.

Ling Xu, Qiang Ma, and Masatoshi Yoshikawa. Credibility-oriented Ranking of Multimedia News Based on a Material-opinion Model. *Web-Age Information Management*, pages 290–301, 2011.

Qian Xu. Should I Trust Him? The Effects of Reviewer Profile Characteristics on eWOM Credibility. *Computers in Human Behavior*, 33:136–144, 2014.

Yunjie Calvin Xu and Zhiwei Chen. Relevance Judgment: What Do Information Users Consider Beyond Topicality? *Journal of the American Society for Information Science and Technology*, 57(7):961–973, 2006.

Yusuke Yamamoto and Katsumi Tanaka. Enhancing Credibility Judgment of Web Search Results. In *Proc. of the 2011 Annual Conference on Human Factors in Computing Systems.* ACM, 2011a.

Yusuke Yamamoto and Katsumi Tanaka. ImageAlert: Credibility Analysis of Text-image Pairs on the Web. In *Proc. of the 2011 ACM Symposium on Applied Computing*. ACM, 2011b.

Olga Yanenko and Christoph Schlieder. Game Principles for Enhancing the Quality of User-generated Data Collections. In *Proc. of the 17th Annual Association of Geographic Information Laboratories for Europe (AGILE) Conference on Geographic Information Science*, 2014.

Chen Ye and Oded Nov. Exploring User Contributed Information in Social Computing Systems: Quantity Versus Quality. *Online Information Review*, 37(5):752–770, 2013.

Reyyan Yeniterzi and Jamie Callan. Constructing Effective and Efficient Topic-specific Authority Networks for Expert Finding in Social Media. In *Proc. of the Workshop on Social Media Retrieval and Analysis*, 2014a.

Reyyan Yeniterzi and Jamie Callan. Analyzing Bias in CQA-based Expert Finding Test Sets. In *Proc. of the Special Interest Group on Information Retrieval (SIGIR)*, 2014b.

Wei Zha and H. Denis Wu. The Impact of Online Disruptive Ads on Users' Comprehension, Evaluation of Site Credibility, and Sentiment of Intrusiveness. *American Communication Journal*, 16(2), 2014.

Jin Zhang. *Visualization for Information Retrieval (The Information Retrieval Series)*. Springer, 1st edition, 2007.

Jingyuan Zhang, Xiangnan Kong, Roger Jie Luo, Yi Chang, and Philip S. Yu. NCR: A Scalable Network-Based Approach to Co-Ranking in Question-and-Answer Sites. In *Proc. of the International Conference on Information and Knowledge Management (CIKM)*, 2014.

Jun Zhang, Mark S. Ackerman, and Lada Adamic. Expertise Networks in Online Communities: Structure and Algorithms. In *Proc. of the International World Wide Web Conference (WWW)*. ACM, 2007.

Sue Ziebland and Sally Wyke. Health and Illness in a Connected World: How Might Sharing Experiences on the Internet Affect People's Health? *Milbank Quarterly*, 90(2):219–249, 2012.

Cai-Nicolas Ziegler and Jennifer Golbeck. Models for Trust Inference in Social Networks. In Dariusz Król, Damien Fay, and Bogdan Gabryś, editors, *Propagation Phenomena in Real World Networks*, volume 85 of *Intelligent Systems Reference Library*, pages 53–89. Springer, 2015.

Cai-Nicolas Ziegler and Georg Lausen. Propagation Models for Trust and Distrust in Social Networks. *Information Systems Frontiers*, 7(4-5):337–358, 2005.