We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

15

Web Intelligence for the Assessment of Information Quality: Credibility, Correctness, and Readability

Johan F. Hoorn and Teunis D. van Wijngaarden
VU University Amsterdam
Netherlands

1. Introduction

There is a shift in the use of the Internet at the cost of traditional media to access scientific information. By mere frequency of exposure, the information of Web sites seems 'more true' than the usually heavily reviewed and editorially controlled scientific venues (cf. Johnson & Kaye, 1998). In addition, the bulk of users are non-expert in a particular field (e.g., health, finance) but nevertheless use the first links Google shows to make decisions that affect their lives. Most users do not verify the information they find. Quality labels are overlooked or have little meaning to the non-expert user. A strong selection criterion for non-expert users is the readability of a site and scientific papers are not known for being easy in that respect. Speed before accuracy seems to be the doctrine for Web engineers, designers, and users alike but we wish to turn that principle around by proposing an intelligent Web service that assesses the quality of information by combining Web-page credibility through Google's PageRank algorithm, informational correctness through text mining, and over 200 formulas for readability measurement; everything under one button, rendering one simple graphical output in a 3D space.

Compared to the traditional media such as newspapers, radio, or television, its de-central nature makes the Internet non-selective in who takes the floor. Flanagin and Metzger put it like this: "Whereas newspapers, books, magazines, and television all undergo certain levels of factual verification, analysis of content, and editorial review, by and large Internet information is subject to no such scrutiny." Editorial functions now fall upon the shoulder of the media consumer (Flanagin & Metzger, 2000). Certain traditional media try to survive by maintaining a digital counterpart on the Web (e.g., the Washington Post).

The Internet provides information for the public at large and people can individually access that information without the interference of a third party such as a teacher, journalist or other expert. Nowadays, the Internet is used for information seeking more than books, television, or newspapers (Flanagin & Metzger, 2001). Web publishing happens on a global scale and bypasses the traditional media gate-keepers such as publishing houses. Sites that

provide means to shape content in a collaborative way are, for example, YouTube[1] (a movie Web site) and Wikipedia[2] (an online encyclopedia). Web-based content management systems such as Joomla![3] make Web publishing all too easy. Not a lot of technical skills are needed, for instance, to install the software, but even this can be left to specialized Web design bureaus. There are almost no boundaries, except that illegal information (e.g., porn or cracked software) can be prohibited by law.

Reality is that not many people verify information (Flanagin & Metzger, 2000). They trust the information found on the Internet, although it depends on what the user is planning to use that information for (Rieh & Belkin, 1998). In reviewing the literature, Johnson and Kaye found that young adults trust the Internet more then other media (Johnson & Kaye, 1998).

Health information is particularly wanted (Fox, 2005) and people judge, interpret, and use that information without consulting a physician (cf. self-medication), which may have considerable repercussions if done in the wrong way. Research actually shows quite some variety in the quality of health Web sites (Griffiths et al., 2005). Griffiths and Christensen evaluated the quality of health Web sites while looking at site ownership and editorship. They found that for only 40% of the sites, health professionals were involved in editing (Griffiths & Christensen, 2005).

Consumers in the early years deemed the Web as credible as traditional media (Flanagin & Metzger, 2000). Only recently, some cracks in this image occurred but not to a large extent (Flanagin & Metzger, 2007). One could argue that with the growth of the Web as indicated by the number of hosts (Internet Systems Consortium, 2007; Ministy of Economic Development of New Zealand, 2003), users became experienced and should be streetwise by now with respect to credibility of the information source. The opposite is true, however: People perceive their most used or preferred medium as the most credible source (Johnson & Kaye, 1998). Heavy users verify information the least (Flanagin & Metzger, 2000). This is in line with the repeatedly confirmed finding in cultivation theory that mere-exposure to media determines the way people look at the world (e.g., Morgan, in press).

In other words, ripe and green are made known to the world and the world consumes this information without much critique. In particular heavy users may conceive of the Internet as the most credible of all media, a problem that in the near future – the younger generations – can only become more severe.

The main question, then, is how to discern good quality information from bad quality information. In 2005, *Nature* published a Korean research paper about the cloning of a dog (Lee, et al., 2005) that in 2006, was compromised because one of the authors admitted that the results were faked.[4] Thus, highly credible sources (*Nature*) may pass on incorrect information (fake data).

Information *correctness* is one of the trickiest things to verify because it touches upon our deepest epistemic beliefs. What is 'true' in religion may not be 'true' in science or vice versa. The source of information could give a clue. That is, a university professor may be regarded as more of an expert than a lay person. Yet, professors can be wrong and lay people are sometimes right, so *credibility* may be an indicator but is not definitive. In addition, a

[1] http://www.youtube.com
[2] http://www.wikipedia.org
[3] http://www.joomla.org
[4] http://www.smm.org/buzz/blog/lies_in_korean_stem_cell_research

biochemical publication about leukemia in *The Lancet* may be correct and credible, but will be ignored by the non-expert audience because to them it is not *readable*.
In this chapter, we attempt to develop a measurement tool that helps people in determining the quality of an information Web site by indicating the estimated correctness of information, the estimated credibility of the source, and the readability of the text. We will stumble upon many hurdles and try to take them anyways in the hope that our attempts are thought-provoking enough to inspire a new generation of information-quality assessment tools.

2. Information quality

Quality of information seems to be a container term. In this section, we attempt to conceptualize 'correctness,' 'credibility,' and 'readability,' which supposedly contribute to information quality. We argue that correctness is an aspect of the information, credibility of the source, and readability of the user's level of expertise.

2.1 Conceptualization

Quality is often mentioned in health-related contexts and pertains to the actual content of a Web site in terms of correctness, readability, and completeness (e.g., Price & Hersh, 1999; Griffiths et al., 2005). The word quality is often used to indicate correctness or accuracy of information but then again, correctness is used interchangeably with credibility. In our view, this indicates that quality should be decomposed into a number of quality indicators. Moreover, that correctness and credibility may be highly related concepts but that they are not the same.
Flanagin and Metzger define credibility in terms of believability, accuracy, trustworthiness, bias, and completeness of information (Flanagin & Metzger, 2000). In comparing political Web sites with traditional media, Johnson and Kaye (1998) measured credibility as believability, fairness, accuracy, and depth (completeness). Credibility is also indicated by a Web site's domain, i.e. .com or .gov (Treise et al., 2003; Rieh & Belkin, 1998).
Credibility seems to be indicated by status and appearance factors of the source rather than correctness of information, although the latter does contribute. To measure credibility, for example, Flanagin and Metzger asked Internet users to indicate whether they checked the author of a Web site, whether contact information was provided, what the author's qualifications and credentials were, what the author's goals/objectives with the published information were, if the information itself was current, if other sources were available for validation, if there was a stamp of approval or recommendation, if the information was an opinion or fact and if the information was complete and comprehensive (Flanagin & Metzger, 2000). In other words, credibility is a surrogate for correctness of information, probably because it is easier to check and somehow is related to correctness.
If this is so, many aspects that are mentioned to indicate credibility actually indicate correctness. Accuracy and completeness are aspects of information correctness whereas believability, trustworthiness, and bias are aspects of credibility.
Quality was also indicated by readability. Readability can be approached from two sides, whether the text is easy enough that it can be accessed by lay people or whether lay people have enough reading skills to understand a text. This division is visible in the type of readability formulas available on the market. The Flesh (1948) reading ease score is a typical

example of the first and estimates how easy a text is. Its successor, the Flesh Grade Level, estimates the school grade a reader should have to be able to read a certain text. All readability measures use text properties such as syllables and sentence length to estimate a score (Hartley et al., 2004) but it is hard to decide whether such text properties indicate readability seen as reading 'ease' or as 'appropriate to the reading level of the user.'

The confusion of terms points at quite some conceptual overlap. Just like credibility and correctness may be positively correlated, readability and correctness may be negatively correlated. An explanation of a disease may be incomplete, whereas for the sake of readability certain omissions in the story may be desired. In other words, a validation of concepts and a verification of the strength of their distinctive power are most wanted.

Table 1 provides the items that in our view indicate correctness, credibility, and readability. We regard credibility an aspect of the source and correctness an aspect of the message. If readability is connected to reading level, it is an aspect of the user. Credibility is indicated by reliability, believability, trustworthiness, bias of information, and fairness (Rieh & Belkin, 1998; Flanagin & Metzger, 2000; Johnson & Kaye, 1998). Correctness is indicated by accuracy, completeness, and depth (Price & Hersh, 1999; Griffiths, et al., 2005). Readability (whether ease or level) is indicated by, among others, number of syllables and sentence length (e.g., Flesh, 1948; Hartley et al., 2004).

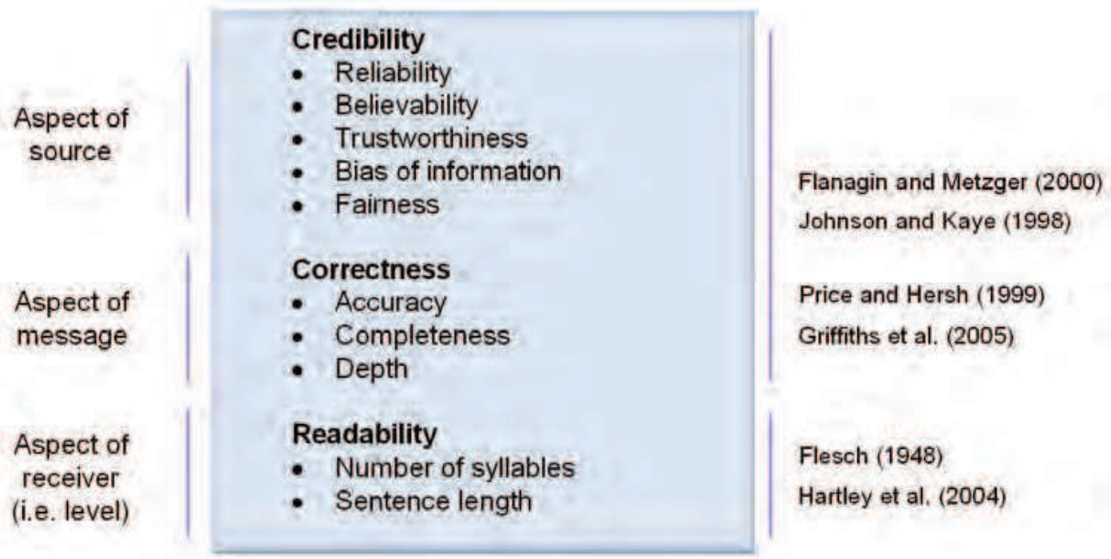

Table 1. The three dimensions of information quality and some of their indicators

In sum, information quality appears to be a container concept that ranges from believability to readability. Correctness of information comes closest to what one may regard as 'the truth.' Credibility of the source indicates how seriously the content should be taken. Readability, then, is a compound of reading ease and reading level.

3. Quality assessment

In the early days of the Internet, quality of Web sites was verified by hand. In a later stage, the user was helped by automated protocols such as AQA (Automated Quality

Assessment). Later advances made use of Google PageRank or required a semi-automatic reviewing effort as observed in the Wikipedia community.

3.1 Evaluating Web sites by hand

Over the years, assessment methods such as checklists helped experts and novices alike to evaluate Web sites by hand. For instance, the Health Information Technology Institute of Mitretek Systems, Inc. made a list of criteria that consumers could use to assess the quality of Health Web sites (Price & Hersh, 1999). DISCERN was another rating tool for health Web sites (Charmock, et al., 1999). According to Griffiths and Christensen (2005), three studies investigated the relationship between DISCERN ratings and scientific quality rated by experts and two of them found a significant relation.

The Health on the Net Foundation developed a set of principles called the Net Code of Conduct (Price & Hersh, 1999). Web sites can voluntarily comply with these principles and express their commitment through a logo. Price and Hersh (1999) proposed to have experts review Web sites and publish the reviews on the Web. Again, the site's commitment can be expressed through an examination logo. Another option is a portal with references to good quality Web sites.

The questions with these approaches are whether the user should do a checklist for each site and which list they should use? There are many logos around, but what are they worth? Checklists and logos relate to the credibility of a source, not to its contents. Who reviews the reviewer? Internet is a volatile medium – who reevaluates whether information is still up-to-date?

3.2 Early tooling

Eysenbach and Diepgen (1998) attempted to label the quality of health information by attaching metadata to each document. They argued that not only the authors should provide metadata but third parties such as rating services should do so as well. Browsers could use that metadata to filter out pages that do not meet personal quality criteria as predefined by the user. These authors concluded that an "agreed formal international standard for medial publication on the internet, enforced by appropriate peer or government organisations" was not realistic. Nevertheless, they argued for at least a standard for the labeling of health-related information. In addition, Eysenbach and Diepgen (1998) proposed that the "potential of computers to determine indirect quality indicators by means of automatic (mathematical) methods" should be explored.

Price and Hersh (1999) employed two engines to search for user-requested information. The outcomes were merged, downloaded, analyzed, the resulting URLs scored, and listed for the user. These criteria were used to assess quality in terms of relevance, credibility, absence of bias, content, currency, and value of links. The tool yielded a ranked list of URLs, but the researchers stated that evaluations remained necessary to verify that highly ranked pages were indeed more credible and that non-experts were able to use the tool. Moreover, the authors did not provide many details on the working of the tool. Although they posited that automatic analyses of Web pages for quality indicators is feasible and useful, they also stated that it is easier to identify indicators for undesirable Web pages than it is to identify indicators of high quality (Price & Hersh, 1999).

Automated Quality Assessment (AQA) as developed by Griffiths et al. (2005) consisted of six steps: Target Web sites were downloaded using web crawler software, the pages were aggregated with arbitrary pages, a previously learned relevance query was processed over the collection, a previously learned quality query was processed in the same way, site relevance and quality scores were computed and normalized, and the overall site score was computed. The relevance feedback-technique was used to learn the queries:

> A complex query consisting of weighted terms (words and phrases), is automatically generated by comparing the term frequency distributions of sets of relevant and irrelevant documents. (…) The resulting query is used by a text retrieval system to derive relevance scores for documents. (Griffiths et al., 2005)

Griffiths et al. (2005) did the same for the quality query. These authors claimed to be the first that made a customized automated tool for identifying the evidence-based quality of health information that focuses on accuracy rather than reliability. They stated that the tool is useful for quality portal maintainers to do the first selection. Their research focused on depression Web sites. To use AQA for other health topics requires a new training procedure. According to the authors, limitations of AQA are that it can be spammed (Web site owners can include terms that lead to high scores) and that the focus is solely on treatment information (Griffiths et al., 2005).

3.3 Google PageRank – indicating credibility

The Google PageRank algorithm (Brin & Page, 1998) is the central formula that ranks URLs found by Google's search engine.[5] The number is not the position in Google, but reflects the 'importance' of the page. The PageRank algorithm is based on graph theory. The Internet is represented as a directional graph (Figure 1), with every page being a node. Every link from page to page is represented by an arrow such that an incoming link is depicted as an incoming arrow.

The PageRank of page A is based on the PageRank of pages that link to page A. The more pages with a high PageRank link to A, the higher A's PageRank becomes. The assumption is that the height of Google PageRank indicates the importance of a page.

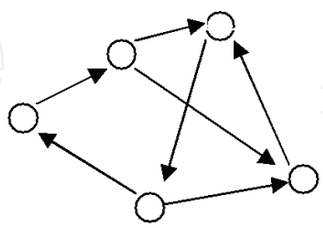

Fig. 1. Directional graph representing Internet links and targets

Chen et al. (2007) used an algorithm based on Google PageRank to assess the relative importance of all publications in the *Physical Review* family of journals from 1893 to 2003.

[5] http://www.google.com

They claimed that Google PageRank did better than simply counting the number of citations:

> We suggest that the Google number Gi of paper I (…) is a better measure of importance than the number of citiations alone in two aspects: (i) being cited by influential papers contributes more to the Google number than being cited by unimportant papers; (ii) being cited by a paper that itself has few references gives a larger contribution to the Google number, than being cited by a paper with hundreds of references. (Chen et al., 2007)

Griffiths and Christensen (2005) asserted that for consumers, Google PageRank was "as strong an indicator of evidence-based quality as DISCERN." Altogether, Google PageRank seems to provide a good indication of a Web site's credibility. We cannot regard PageRank as an indicator of information correctness, because PageRank processes hyperlinks and not contents.

3.4 Wikipedia – attaining correctness

Wiki software,[6] as developed by Ward Cunningham in 1995, allows anyone to edit a Web site from within the browser (Web-based) with a simple markup language for collaborative content creation. Wikipedia is an online encyclopedia based on the wiki principle. The English version of Wikipedia contains more than 2 million articles.[7]

The open way of content creation and editing raised questions with Stvilia, Twidale, Gasser & Smith (2005) as to why people bother to contribute at all, what the quality of the product is, and why people would trust and use it? Why does the project not disintegrate into anarchy? How is the project organized, and how do the processes change over time?

For our purposes, we would like to focus on the way Wikipedia treats information correctness. What technical facilities and social constraints are built into wiki and Wikipedia to improve and maintain the accuracy and verity of information?

3.4.1 Wikipedia's correctness

A CNET headline in December 2005 ran "Study: Wikipedia as accurate as Britannica" (Terdiman, 2005). The article referred to an investigation by *Nature* (Giles, 2005), claiming that Wikipedia came close to the traditional *Encyclopedia Britannica* in terms of accuracy. *Encyclopedia Britannica* responded that the *Nature* publication was wrong (Nature, 2006), but *Nature* still defends her findings (ibid.).

3.4.2 Vandalism on Wikipedia

Wikipedia can repair malicious edits such as mass deletion of content in a median time of 2.9 minutes (Viégas et al., 2007; Viégas et al., 2004). Because of the vandalism issue, certain pages are protected against changes. Semi-protected pages cannot be edited by anonymous and newly registered users.[8] Fully protected pages can only be edited by administrators.

[6] http://en.wikipedia.org/wiki/Wiki
[7] Based on their own statistics. See http://en.wikipedia.org/wiki/Main_Page
[8] http://en.wikipedia.org/wiki/Category:Semi-protected

Protection of a page must be requested on the talk pages and can be refused, "especially if they [the requests] are controversial, do not comply with Wikipedia policies, or do not have evidence of consensus."[9]

3.4.3 Featured articles

Wikipedia provides a list of featured articles that contains the best Wikipedia has to offer, according to the community (Figure 2 shows a screenshot). On the Wikipedia Web site, featured content is described as follows:

> These are the articles, pictures, and other contributions that showcase the polished result of the collaborative efforts that drive Wikipedia. All featured content undergoes a thorough review process to ensure that it meets the highest standards and can serve as an example of our end goals.[10]
>
> Peer reviewed material has to comply with the following criteria:[11]

- 'well written'
- 'comprehensive', in a sense that it does not neglect major facts and details
- 'factually accurate', that is: verifiable against reliable sources to be supported with citations and references
- 'neutral', without bias
- 'stable', what means that there are no significantly changes from day to day
- following style guidelines (e.g., having a lead, using the right markup tags)
- having images where they are appropriate, with captions and acceptable copyright status
- of appropriate length, meaning staying focused

Users can nominate an article for receiving the featured status. Before a user nominates an article, s/he is asked to post it on a special page that solicits for peer review. A featured article can also be nominated to be denied its status.
In a way, the list of featured articles is a portal to the high quality content that Price and Hersh (1999) were looking for. The quality is assessed by peers who use a checklist (the criteria). Unlike other checklists mentioned earlier in this chapter, the Wikipedia assessment process is different in that no individual consumer or expert evaluates the page, but a group of people.

[9] http://en.wikipedia.org/wiki/Category:Wikipedia_protected_edit_requests
[10] http://en.wikipedia.org/wiki/Wikipedia:Featured_content
[11] http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

Fig. 2. Screenshot of the Biology and Medicine category of Wikipedia´s featured articles

The Wikipedia criteria fit in nicely with our concepts of credibility, correctness, and readability. That featured articles are as unbiased as possible, refer to verifiable sources, and show copyright status of images pertains to our notion of credibility. That featured content should be comprehensive, factually accurate, and stable (no significant changes from day to day), in our view, would indicate correctness of information. Readability, then, would be indicated by following style guidelines, being well written, using appropriate images (plus captions), and having appropriate length (focused).

3.4.4 Discussion pages

Wikipedia provides so called talk pages, which discuss the quality of a page. Talk pages can be attached to every page in Wikipedia. In addition, people can also post questions or can ask for additional information (Stvilia et al., 2005). Organized and readable discussions on talk pages add to the quality of Wikipedia content (Viégas et al., 2007; Stvilia et al., 2005).
Stvilia et al. (2005) analyzed the content of 60 discussion pages of featured articles and identified ten types of quality problems that Wikipedia users mentioned. There were problems with accessibility, accuracy, authority, completeness, complexity, consistency, informativeness, relevance, verifiability, and volatility. They noted that quality assessments "… are often relative to a particular community's cultural and knowledge structures. … If the user is not aligned with those structures, his or her claim of the existence of an IQ problem may not be shared by the rest of the community and get rejected" (Stvilia et al., 2005). These authors further reported that featured pages had better discussion pages

attached to it than randomly selected (non-featured) pages. The discussions were better organized, better readable, and more polls were used (Stvilia et al., 2005).

3.4.5 Fitting in the Wikipedia approach

Wikipedia is certainly not an automatic tool but does provide the community with the means to control correctness, source credibility, and readability. The correctness of information is of particular interest to this chapter, as we can already define credibility and readability in a more automatic way. Checking a Web page for Wikipedia contents (i.e. the featured articles) at least gives some indication of correctness. Featured articles will be close to *Britannica*, are checked for vandalism, comply with strong quality criteria (not merely correctness but also credibility and readability), and are constantly scrutinized in the discussion pages. In the next section, we explore the way to automate the check-up with reviewed content so to estimate information correctness of a given Web page.

3.5 Text mining – automating correctness

Data mining, text mining, and Web mining are emerging fields in computer science, biology, and chemistry. Data mining is concerned with extracting useful information from huge amounts of (semi) structured data that are stored in databases. Text mining is a specific technique to extract information from unstructured texts, in particular, natural language. Web mining is a combination of data mining and text mining in relation to the Web. In fact, text mining is the reverse of adding metadata to documents (cf. labeling in Section 3.2).

Metadata add structured information that make a document easier to handle for a computer, whereas text mining makes the computer capable of handling unstructured data. For instance, STEMWAY is a text mining tool that was capable of extracting a general model out of a host of stem cell documents (Park et al., 2005). TAKMI (Text Analysis and Knowledge Mining) is a text mining tool for the identification of patterns in questions received by helpdesk call-centers (Nasukawa & Nagano, 2001). These tools have one thing in common: They provide a way to represent (or convert) textual data into structured knowledge.

For our quality assessment tool, text mining could help in establishing a measure for information correctness. For a given topic, a tool such as STEMWAY or TAKMI could extract a general model from the scientific literature. This becomes a reference ontology against which individual Web pages found by the user are tested.

> An ontology is an explicit specification of a conceptualization. (…) When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge. Thus, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names are meant to denote, and formal axioms that constrain the interpretation and well-formed use of these terms. (Gruber, 1993)

The World Wide Web Consortium (W3C) provides a number of languages to represent ontologies,[12] with OWL being one of the strongest Web ontology languages. A free tool to create an OWL ontology is Protégé by Stanford Medical Informatics.[13] Through text mining Protégé creates the ontology without the help of its human user.
Once the computer created a reference ontology, the knowledge of an individual Web article can be structured through text mining. The difference or tension between the reference ontology and the knowledge structure extracted from an individual paper indicates the level of correctness: The smaller the difference, the more complete and correct the Web text is. For the user, this difference could be translated into a percentage that the found article is correct.
To date, text mining tools are not as exact as hand-curated data (Rebholz-Schumann et al., 2005). Parsing natural language such as negations is still a challenge (Briscoe & Carroll, 2002; Pyysalo et al., 2004; Stavrianou et al., 2007). In the bag-of-words approach, however, negations are treated as part of the same knowledge structure because they share the same set of keywords as affirmative statements (Nasukawa & Nagano, 2001). Also rich vocabularies are harder to process than texts with limited contexts (ibid.) but "it is only a matter of time and effort before we are able to extract facts automatically" (Rebholz-Schumann et al., 2005).

3.6 Readability – multiple measures

Information may come from a highly credible source and be correct, but if a user cannot read it, the source will not be used. Readers may not have the proper level of expertise or the text is written in an obscure style.
In the area of readability formulas, many competitors exist. "By the 1980s, there were 200 formulas and over a thousand studies published on the readability formules attesting to their (…) validity." (DuBay, 2004). Often mentioned in the literature is the reading ease formula of Flesch (1948). It became the most widely used and one of the most tested and reliable formulas (DuBay, 2004; Chall, 1974; Klare, 1963). This formula uses two variables: the number of syllables and the number of sentences in a 100-word sample (DuBay, 2004).
Readability formulas should be considered rough estimates because they count linguistic forms and not content. The words 'computer' and 'freedom' are of the same length – and therefore not treated differently by formulas – but the latter word is more complex because it is an abstraction with an enormous political bias. In addition, the formulas do not account for infographics, multimedia, or any other explanatory medium besides text.
Hartley et al. (2004) evaluated the readability of magazine articles about science and compared them to articles in the field of psychology and history. They found that science articles have the shortest sentences and the highest Flesch scores. They did the same comparison with other genres, from scholarly journals to magazines and in most of the cases the same difference was found: Science is shorter (Hartley et al., 2004). In knowing, however, that most non-experts cannot read science, the abstract quality and use of formulas in scientific texts puts up a hurdle not acknowledged by a readability formula. Another unexpected finding by Hartley et al. (2004) was that passive voice does not necessarily make a text less readable.

[12] www.w3c.org
[13] www.protégé.stanford.edu

Readability formulas are based on counting text properties. Common text editors usually offer a Flesh reading ease score together with text statistics such as number of characters, words, and sentences. Surprisingly, however, automated Flesch measures sometimes vary between different tools (Hartley et al., 2004; Harris, 1996; Mailloux, et al., 1995; Sydes & Hartley, 1997). Because we want to use readability measures for relative comparisons only, this is not a problem as long as we use the same tool for all Web sites.

4. A combination of techniques

We would like to design an intelligent service that assesses the quality of information of Web pages. In this section, we suggest to estimate credibility through Google's PageRank algorithm and informational correctness through text mining. Readability may be indicated by an all-inclusive variable of over 200 available formulas. The analyses could be brought under one button, outputting a position of the Web page in a 3D graphical space.

Google PageRank can be used to calculate a *credibility value*. In line with Brin and Page (1998), Chen et al. (2007), and Griffiths and Christensen (2005), we believe that Google PageRank performs better than simply counting the number of citations in scientific journals. The importance of a Web page is reflected in the number of incoming and outgoing links. The more links point at you, the higher your rank number.

To validate this claim, we suggest creating a psychometric scale that has the following items on it, as derived from Flanagin and Metzger (2000) and Johnson and Kaye (1998): reliable, believable, trustworthy, unbiased, and fair. In employing this scale, let users rate the credibility of a large range of information Web-pages. After scale analysis, make a rank order of pages according to their level of estimated credibility. Also make a rank order for these Web pages according to Google PageRank and according to a traditional citation index. Then calculate the Spearman rho statistic between the paired rank orders of user-rated credibility vs. PageRank as well as user-rated credibility vs. citation index. The measure (PageRank or citation index) that shows the least difference with the rated-credibility ranking is the most indicative measure. The closer rho approaches 1, the higher the correlation between paired rankings. In other words, rho also indicates in how far the best measure is still away from human assessment of credibility. Credibility rating through psychometric scales should be a community effort and we could use wiki technology to do so.

To assess *correctness* of information of a Web page, we could employ AQA (Griffiths et al., 2005, Section 3.2). However, the AQA procedure is quite difficult because it takes six steps and uses multiple software programs. Also, the queries must be learned. AQA was designed and tested for depression Web sites and we do not know whether it will be successful in other fields of health or science. Therefore, we wish to try for a more generic approach.

We envision a repository of reference ontologies that relate to each lemma in, for example, the *Encyclopedia Britannica Online*. For each lemma, text mining of the relevant scientific literature supplemented with featured articles in Wikipedia provides the general pattern or semantic structure (cf. the stem cell model of Park et al., 2005) that a given Web page should provide about a topic. The difference or tension between the reference ontology and the specific page indicates the accuracy (are the proper concepts used in the proper relations?) and completeness of the page (is everything there?). This could count as the automatically generated *correctness value*.

To validate the correctness value, a psychometric scale should be made that has the following items on it as derived from Price and Hersh (1999), Griffiths, et al. (2005), and the Wikipedia criteria for featured articles: comprehensive, factual, stable, accurate, complete, and with depth. With this scale, users rate the correctness of a large number of information Web-pages. After scale analysis, the scale values can be regressed on the difference between reference ontologies and specific pages (the automatic correctness values) to permit the prediction of the most probable values of user-rated correctness. The higher the regression weights, the more the automatically extracted correctness value is indicative for human-rated correctness.

With respect to *readability*, we want to dodge the problem of arbitrarily selecting one of the readability scores. We suggest using multiple formulas so to keep from ignoring important aspects emphasized by other measures. This means that certain formulas should be automated first and that the final score is a compound of all measures. Then the body text should be assessed, that is, text satellites such as headings, sub headings, lengthy quotations, references, and other peripheral data should be discarded in the analysis (Hartley et al., 2004; Stavrianou et al., 2007). This could be done by the user but it would be better to fully automate this procedure.

Many of the available formulas seem to indicate a valid aspect of readability (DuBay, 2004). However, each readability formula makes estimations on a different scale (Figure 3, left box), so that for a compound *readability score*, we need to calculate the normalized mean or a z-score.

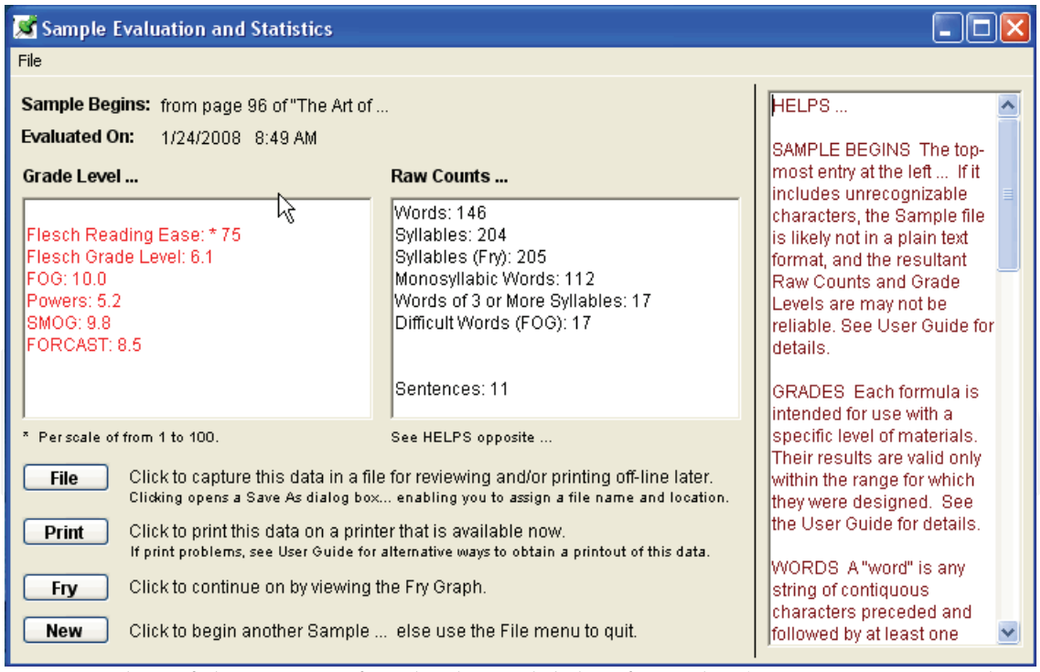

Fig. 3. Screenshot of the output of multiple readability formulas (Micro Power & Light Co.)[14]

[14] http://www.micropowerandlight.com/readability-formula-scores-screen.html

For validation purposes, a psychometric scale should be devised that follows the style guidelines for Wikipedia's featured articles: well written, appropriate images, appropriate length, and focus. Users score the readability of a number of body texts using this scale. After scale analysis, regression of the scale values on the normalized mean scores can be used to estimate user-rated readability given the automatically calculated normalized-mean readability. The higher the regression weights, the more the normalized mean is indicative for human-rated readability.

If we follow Hartley et al. (2004), the system should calculate readability for the body text, skipping text satellites such as headings and images. This restriction will mitigate the regression weights because 'appropriate images' (and captions) is one of the items on the psychometric readability scale.

As a standard, the measure yields a general readability score but this could be fine-tuned to the user's reading skills by calibrating the system first. At first use, the user could do a readability test after which the system always provides a score that is relative to the user's benchmark value.

In all, the browser could have an interface button that triggers the assessment of the information quality of an open Web page. Such a tool should be capable of positioning, for instance, a published but later on retracted paper as credible but incorrect (e.g., Lee, et al., 2005). Wikipedia featured articles will probably be positioned as correct, readable, and somewhat credible and a patient's blog as readable, somewhat correct but not too credible (Figure 4).

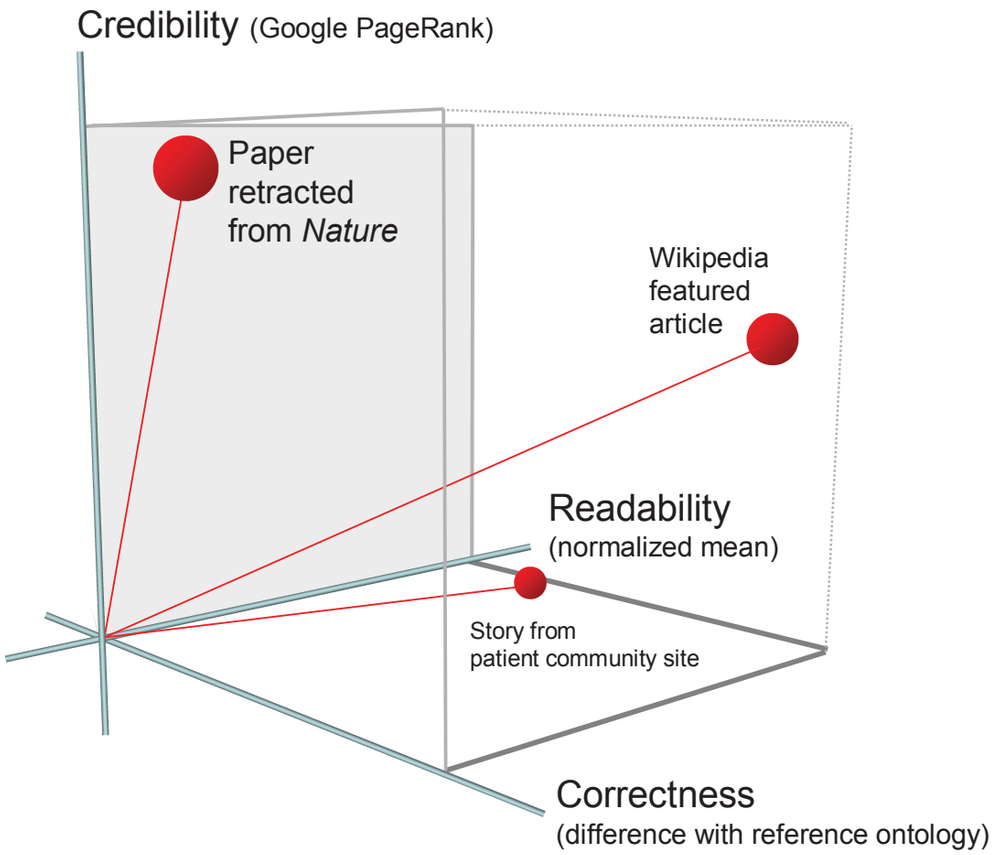

Fig. 4. Impression of a 3D space that positions Web pages on the axes of credibility, correctness, and readability

5. Conclusions

The Internet offers a vast amount of information that can be accessed and used by everyone who is online without much guidance. Heavy users such as young people judge the first hits Google offers at face value. Particularly in health issues, following the wrong advice may cause serious damage. The main question is how to separate good quality information from the bad. One could perhaps read the scientific journals but unless you are a specialist in a certain field, science is not particularly readable for non-experts.

Several methods were proposed to distinguish information quality. Checklists, quality stamps, logos, and automated quality assessment are useful tools but are either labor-intensive or simply overlooked by the non-expert. A genuine contribution would be to have an intelligent widget in the browser interface that automatically assesses the information quality of a Web site and presents the evaluation result in one easy-to-grasp representation (a number, a graph, thumbs up or down, or any other qualifier).

We observed that in the earlier approaches, difficulties were insidious in the definition of information quality as well as in the boundaries of its underlying concepts such as

credibility and correctness. Such confusion is bound to render quality estimates that are not in sync with human judgment. In this chapter, we have attempted to decompose quality into three main parts and have argued to measure them with novel Web technologies.

Credibility of the source should be assessed by Google PageRank and validated by comparing the resulting rank order of Web pages with human assessment of those pages on a psychometric scale, using the Spearman rho statistic. *Correctness* of information should be assessed by creating reference ontologies for each lemma in, for instance, the *Encyclopedia Britannica Online*. Reference ontologies could be created by text mining (e.g., STEMWAY or TAKMI) the relevant scientific literature and Wikipedia's featured articles. The semantic structure of a given Web page could then be compared with the reference ontology, yielding a difference value that indicates correctness. Again, this measure should be confronted with user assessment of correctness on a psychometric scale after which regression analysis shows whether automated estimates are predictive for human assessment of information correctness. *Readability* should be indicated by the normalized mean or a z-score for the 200+ readability measures that each in its own right assesses one or more aspects of readability (e.g., ease or grade level). In a regression analysis, the normalized mean or z-score should have predictive power for the user-estimated readability of texts as rated on a psychometric scale. Reading-level calibration could be done by letting the user do a readability test at first use (the personal readability benchmark). We realize that the readability method is a bit crude but easy to implement. In addition to the usual variables in text complexity measurement, we could look at sets of participle perfectum, embedding, priority placement, jargon, etc.

Most of what we suggested is technically feasible. A possible bottleneck lies in the state-of-the-art of text mining and ontology modeling. Stavrianou et al. (2007), for example, explain that the distribution of terms that make up a semantic structure varies across text types (e.g., abstracts, articles, or collections of articles). Word sense disambiguation is a challenge in free text (ibid.). Lastly, the text properties that need to be analyzed may vary with different text types (ibid.). Thus, modeling ontologies is not to be underestimated. The status of the technique as is may be insufficient to use instantly. Quality sometimes lies in subtle things, which a model may not perceive. Over time, the models should develop greater detail and the scope of the model should become clearer. As a cross-validation, we could look at the number of sources that provide the same information, which may indicate the acceptance of information. This could also be done cross-lingual to ensure that no duplicates are counted.

As far as we can see, a concept analysis of notions such as information quality, correctness, and credibility is new in this area. The separation between correctness as an aspect of the message and credibility as an aspect of the source is important because 'truth' is not the same as 'reputation.' To validate computer-generated estimates against human assessment on psychometric scales is a novelty in the area but important to judge whether the system is anywhere near a proper judgment. The use of text mining to create ontologies is already explored but to use ontologies as a reference to assess the correctness of a free text is a new idea. In addition, we are the first to suggest a democratic measure (all voices count) of readability instead of arbitrarily opting for, at best, a handful of measures. The same goes for the combination of all readability estimates into one measure instead of losing the overview with a host of readability scores that are all measured on a different scale. Compiling all three measures into one 3D graphic that can be generated by one button click would create a new intelligent Web service for search engines to support decisions on

information quality. Shortcomings of the present chapter are that nothing is tested yet and that there are still issues left in the domain of text mining tools.
What needs to be done, then, is to perform a large scale survey among users to scrutinize the concepts of information quality, correctness, credibility, and readability and to test the convergent and divergent validity of their indicators (e.g., believability, accuracy, depth). In addition, a large number of reference ontologies needs to be created, which will urge to look into a number of problems in text mining such as word sense disambiguation and the type of text properties that needs to be analyzed. If all is set, user studies should test the results of the automated measures against user ratings.

6. Acknowledgements

We kindly thank Piek Vossen for reviewing an earlier draft of this chapter.

7. References

Brin, S. & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Proceedings of the Seventh International Conference on World Wide Web, 7*, pp. 107-117, Brisbane, Australia, April 1998, *Computer Networks and ISDN Systems,* Vol. 30, No.1-7, (April 1998) 107-117, ISSN: 0169-7552.

Briscoe, T. & Carroll, J. (2002). Robust Accurate Statistical Annotation of General Texts, *Proceedings of the Third International Conference on Language Resources and Evaluation,* pp. 1499-1504, Las Palmas, Gran Canaria, May 2002, Canary Islands, Spain: European Language Resources Association.

Chall, J. S. (1958). *Readability: An appraisal of research and application.* Columbus, OH: Ohio State University Press. Reprinted 1974. Epping, Essex, England: Bowker Publishing Company.

Charmock, D.; Shepperd, S.; Needham, G. & Gann, R. (1999). DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J. Epidemiol. Community Health , 53*, pp. 105-111, ISSN: 0143-005X.

Chen, P.; Xie, H.; Maslov, S. & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics,* Vol. 1, No. 1 (January 2007), pp. 8-15, ISSN: 1751-1577.

DuBay, W. H. (2004). *The Principles of Readability.* Retrieved October 04, 2007, from Impact Information: Plain Language Services: http://www.impactinformation. com/impactinfo/readability02.pdf

Eysenbach, G. & Diepgen, T. L. (1998). Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information. *BMJ,* Nov 1998, pp. 1496-1502, ISSN: 09598138.

Flanagin, A. J. & Metzger, M. J. (2001). Internet use in the contemporary environment. *Human Communication Research ,* Vol. 27, No. 1, pp. 153-181, ISSN: 1468-2958.

Flanagin, A. J. & Metzger, M. J. (2000). Perceptions of Internet Information Credibility. *Journalism and Mass Communication Quarterly,* Vol. 77, No. 3, pp. 515-540, ISSN:1077-6990

Flanagin, A. J. & Metzger, M. J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New media & society,* Vol. 9, No. 2, pp. 319-342, ISSN: 1461-7315.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology,* Vol. 32, No. 3, pp. 221-223, ISSN: 0021-9010.

Fox, S. (2005). *Health Information Online.* Retrieved September 11, 2007, from PEW Internet & America Life Project: http://www.pewinternet.org/pdfs/PIP_Healthtopics_May05.pdf

Giles, J. (2005). Internet encyclopedias go head to head. *Nature,* Vol. 438, No. 15, pp. 900-901, ISSN: 0028-0836.

Griffiths, K. M. & Christensen, H. (2005). *Website Quality Indicators for Consumers*. Retrieved September 11, 2007, from Journal of Medical Internet Research: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1550688

Griffiths, K. M.; Tang, T. T.; David, H. & Helen, C. (2005). *Automated Assesment of the Quality of Depression Websites*. Retrieved September 11, 2007, from Journal of Medical Internet Research: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid =1550680

Gruber, T. R. (1993). *A Translation Approach to Portable Ontology Specifications.* Retrieved October 15, 2007, from TomGruber.org: http://tomgruber.org/writing/ ontolingua-kaj-1993.pdf

Harris, R. (1996). Variation among style checkers in sentence measurement. *TEXT Technology,* Vol. 6, No. 2, pp. 80-90.

Hartley, J.; Sotto, E. & Fox, C. (2004). Clarity across the disciplines: an analysis of texts in the sciences, social sciences, arts and humanities. *Science Communication,* Vol. 26, No. 2, pp. 188- 210, ISSN: 1824-2049.

Internet Systems Consortium. (2007). *ISC Internet Domain Survey*. Retrieved October 20, 2007, from Internet Systems Consortium: http://www.isc.org/index.pl?/ops/ds/

Johnson, T. J. & Kaye, B. K. (1998). Cruising Is Believing?: Comparing Internet and Traditional Sources on Media Credibility Measures. *Journalism and Mass Communication Quarterly,* Vol. 75, No. 2, pp.325-340, ISSN:1077-6990.

Klare, G. R. (1963). *The Measurement of Readability.* Ames, Iowa: Iowa State University Press.

Lee, B.; Kim, M.; Jang, G.; Oh, H.; Yuda, F.; Kim, H. et al. (2005). Dogs cloned from adult somatic cells. *Nature,* Vol. 436, No. 7051, pp. 641-641, ISSN: 1476-4687.

Mailloux, S. L.; Johnson, M. E.; Fisher, D. G. & Pettibone, T. J. (1995). How reliable is computerized assessment of readability? *Computers in nursing* , Vol. 13, No. 5, pp. 221-225, ISSN: 0736-8593. Ministy of Economic Development of New Zealand. (2003). *6. Size of the Internet | Statistics on Information Technology in New Zealand: updated to 2003*. Retrieved October 20, 2007, from Ministry of Economic Development: http://www.med.govt.nz/ templates/MultipageDocumentPage____9980.aspx

Morgan, M. (in press). Cultivation analysis, In: *The SAGE Handbook of Media Processes and Effects,* R. L. Nabi & M. B. Oliver, (Eds.), xx-xx, SAGE, ISBN, Thousand Oaks, CA

Nasukawa, T. & Nagano, T. (2001). Text analysis and knowledge mining system. *IBM Systems Journal,* Vol. 40, No. 4, pp. 967-984, ISSN: 0018-8670.

Nature (2006). Britannica attacks... and we respond [Editorial]. *Nature, 440,* 582. DOI: 10.1038/440582b. Retrieved April 7, 2009 from http://www.nature.com/nature/journal/v440/n7084/full/440582b.html

Park, H. S.; Kim, M. K.; Choi, E. J. & Seol, Y. S. (2005). Text Mining from Categorized Stem Cell Documents to Infer Developmental Stage-Specific Expression and Regulation Patterns of Stem Cells, In: *Lecture Notes in Computer Science: Natural Language Processing and Information Systems, Vol. 3513/2005,* pp. 353-356, Springer Berlin/Heidelberg, ISBN 978-3-540-26031-8.

Price, S. L. & Hersh, W. R. (1999). Filtering Web pages for quality indicators: an empirical approach to finding high quality consumer health information on the World Wide Web, *Proceedings of AMIA Symposium,* pp. 911-915, Washington, DC, November 1999.

Pyysalo, S.; Ginter, F.; Pahikkala, T.; Koivula, J. & Boberg, J. (2004). Analysis of link grammar on biomedical dependency corpus targeted at protein-protein interactions, *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, pp. 15-21, Geneva, Switzerland, August 2004.

Rebholz-Schumann, D.; Kirsch, H. & Couto, F. (2005). Essay: Facts from Text - Is Textmining Ready to Deliver. *PLoS Biology,* Vol. 3, No. 2, pp. 188-191, ISSN: 1545-7885.

Rieh, S. Y. & Belkin, N. J. (1998). Understanding Judgment of Information Quality and Cognitive Authority in the WWW, In*: ASIS 98: Information access in the global information economy. Proceedings of the 61st Annual Meeting of the American Society for Information Science,* pp. 279-289, ISSN: 0044-7870, Pittsburgh, PA, October 1998, Medford, NJ: Information Today.

Stavrianou, A.; Andritsos, P. & Nicoloyannis, N. (2007). Overview and semantic issues of text mining. *ACM SIGMOD Record,* Vol. 36, No. 3, pp. 23-34, ISSN: 0163-5808.

Stvilia, B.; Twidale, M. B.; Gasser, L. & Smith, L. C. (2005). Information quality in a community-based ecyclopedia. In: S. Hawamdeh (Ed.), *Knowledge Management: Nurturing Culture, Innovation, and Technology - Proceedings of the 2005 International Conference on Knowledge Management*, pp. 101-113, ISBN 978-981-256-556-3, North Carolina, October 2005, Charlotte, NC: World Scientific Publishing Company.

Sydes, M. & Hartley, J. (1997). A thorn in the Flesch: Observations on the unreliability of computer-based readability formulae. *British Journal of Educational Technology,* Vol. 28, No. 2, pp. 143-145, ISSN: 1467-8535.

Terdiman, D. (2005). Study: Wikipedia as accurate as Britannica. *CNET News.* Retrieved Oct. 1, 2007 from http://www.news.com/2100-1038_3-5997332.html

Treise, D.; Walsh-Childers, K.; Weigold, M. F. & Friedman, M. (2003). Cultivating the Science Internet Audience: Impact of Brand and Domain on Source Credibility for Science Information. *Science Communication,* Vol. 24, No. 3, pp. 309-332, ISSN: 1552-8545.

Viégas, F. B.; Wattenberg, M.; Kriss, J. & Van Ham, F. (2007). Talk Before You Type: Coordination in Wikipedia. *40th Annual Hawaii International Conference on System Sciences,* pp.78-78, ISBN: 0-7695-2755-8, Waikoloa, Big Island, Hawaii, January 2007, IEEE Computer Society 2007.

Viégas, F.; Wattenberg, M. & Dave, K. (2004). Studying Cooperation and Conflict between Authors with history flow Visualizations. *Proceedings of SIGCHI 2004,* pp. 575-582, ISBN: 1-58113-702-8, Vienna, Austria, April 2004, ACM New York, NY, USA.

Web Intelligence and Intelligent Agents

Edited by Zeeshan-Ul-Hassan Usmani

ISBN 978-953-7619-85-5
Hard cover, 486 pages
Publisher InTech
Published online 01, March, 2010
Published in print edition March, 2010

This book presents a unique and diversified collection of research work ranging from controlling the activities in virtual world to optimization of productivity in games, from collaborative recommendations to populate an open computational environment with autonomous hypothetical reasoning, and from dynamic health portal to measuring information quality, correctness, and readability from the web.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen