TOWARDS AN APPROACH BASED ON VERIFIABILITY ASPECTS TO HELP IN THE QUALITY EVALUATION OF TEXTUAL WEB PAGES

Daniel Lichtnow[1,2], Leandro Krug Wives[1] and José Palazzo Moreira de Oliveira[1]
[1]*Instituto de Informática, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre-RS, Brazil*
[2]*Centro Politécnico, Universidade Católica de Pelotas (UCPel), Pelotas-RS, Brazil*

Keywords: Information Quality, Verifiability, Reliability, Measurement.

Abstract: This work presents an approach based on verifiability aspects to evaluate Web pages with textual content. In the work, verifiability is related to the existence of references to information sources. In this sense, we take into account that textual Web pages with references to information sources use to be better than Web pages without references to information sources. Thus, aspects related to automatically identification of verifiability indicators in textual Web pages are presented. For the given context, the results of preliminary experiments show that verifiability aspects can be useful to infer the quality of texts present on the Web addressed to Web users with little knowledge about a specific subject.

1 INTRODUCTION

One considerable part of Web content consists of textual content (e.g. blogs, articles, papers, etc.). Although, some mechanisms have been created to identify the quality of Web pages, the final quality evaluation is a task that Web users must perform individually.

Take into account this scenario; the present work defines an approach to help in the evaluation process of textual Web pages addressed to Web users with little knowledge about a specific subject. The proposal approach emphasizes the use of verifiability quality indicators. Verifiability is defined as "the degree and ease with which the information can be checked for correctness" (Naumann and Rolker, 2000). In textual Web pages, verifiability is related to the existence of references to information sources represent by Web links, references to papers or even references to persons or organizations. In the work we present some preliminary experiments using verifiability indicators to identify Web pages that contain urban legends, myths, or rumors related to health information.

The paper is organized as follows. In section 2, we present the related work. In section 3, we define our approach. Section 4 describes some preliminary experiments. Section 5 presents final remarks with indications of future work.

2 RELATED WORK

Firstly, regarding to data/information quality, some aspects should be highlighted:

- There is no consensus among researchers about which quality dimensions/factors must be considered to measure or to represent data quality (Pernici and Scannapieco, 2002);
- The majority of data quality proposals are related to structured data (Batini et al., 2009);
- There are few works that emphasize data quality in the context of Web (Batini et al., 2009).

In general, the quality of Web pages is measured considering the link structure present on the Web (Brin and Page, 1998). For instance, link-based quality indicators are evaluated in some works (Amento et al., 2000).

Beyond Web links, Zhu and Gauch (2000) consider other quality metrics like currency (the time stamp of the last modification of the document), the ratio between the number of broken links on a page by the total number of links, information-to-noise (the ratio between the number of tokens present in

Lichtnow D., Krug Wives L. and Palazzo Moreira de Oliveira J..
TOWARDS AN APPROACH BASED ON VERIFIABILITY ASPECTS TO HELP IN THE QUALITY EVALUATION OF TEXTUAL WEB PAGES.
DOI: 10.5220/0003935906890694
In *Proceedings of the 8th International Conference on Web Information Systems and Technologies* (WEBIST-2012), pages 689-694
ISBN: 978-989-8565-08-2

the pre-processed main content by the number of tokens of the document) and, popularity (the number of *inlinks* – *inlinks* are the number of Web links pointing to Web page). In this work, the best results were obtained with information-to-noise metric.

In (Bethard et al., 2009), twelve dimensions of quality related to specific educational purposes are identified. For each quality dimension, some quality indicators are identified. In that work, the approach to identify quality indicators consists on using a training corpus where these indicators are previously annotated by reviewers (the indicators consist on word sequences, for instance). The process of quality identification consists in using Machine Learning techniques to predict whether a resource has good quality (contain indicators).

In addition, there are works in which the objective is to evaluate a specific type of information on the Web. One example is (Dalip et al., 2009), where the quality evaluation of Wikipedia's articles considers features like reviews per day. The problem is that some of these features are limited to Wikipedia's articles.

In (Denecke and Nejdl, 2009), the aim is to identify if a text in a Web page, related to health issues, is informative or affective. The authors consider that informative content has more value and uses Natural Language Processing techniques to identify this fact.

In (Yin et al., 2007) the authors try to identify the most reliable Web page by comparing the content (structured data extracted from Web page) of Web pages. The process uses an iterative method, where the data present in Web pages (e.g. year of publication) are compared and the most reliable source (Web page) is identified.

It is possible to identify a set of quality indicators to evaluate textual Web pages. Each quality indicator/metric must be assigned to a specific quality dimension that emphasizes a distinct quality aspect. Regarding to quality dimensions, we follow the definitions of (Wang and Strong, 1996) (Naumann and Rolker, 1999) and (Naumann and Rolker, 2000). Bellow, we present some quality dimensions and quality indicators/metrics.

- Accuracy. Spelling errors (Batini et al., 2008); number of pages in a Web site (indicates how much effort the author is devoting to the site, and more effort tends to indicate higher quality) (Amento et al., 2000); comparison of data with a reliable source (Yin et al., 2007); information-to-noise evaluation (Zhu and Gauch, 2000);
- Believability. PageRank (Brin and Page, 1998); inlinks (Amento et al., 2000); qualifications of the author or provider of the page (HONCODE, 2009).
- Timeliness. Creation date; last update.
- Relevance. Cosine (Salton et al., 1975) and metadata related to subject (Naumann and Rolker, 1999).
- Verifiability. References to information sources (e.g. Web links, references to papers or even to persons or organizations) (Naumann and Rolker, 2000).

These indicators must be considered heuristics, because since it may be difficult to evaluate some of these aspects (i.e., accuracy). We note that quality indicators related to verifiability are almost ignored in the related works. We also verify that bad textual Web pages (in general) do not indicate information sources. Besides, verifiability is an important quality criterion for Wikipedia[1]. Taking into account these facts, we define our approach in the next section.

3 THE APPROACH

In this section, we present the proposal approach to evaluate the content quality of textual Web pages using verifiability indicators. Initially, considering that quality means "fitness for use" (Wang and Strong, 1996), the context of use for our approach is defined. After, we describe a scenario of use for our approach. We also discuss how to identify some types of source references in the textual Web pages.

3.1 The Context of Use

For defining the context of use, the start point of our approach is related to Web search goals. A Web search goal is related to a user query, i.e., users construct queries to express his/her or her needs related to some task. A relevant taxonomy of Web search goals is presented in (Rose and Levinson, 2004). Another important aspect related to context of use is the user profile. Thus, based on (Garzotto et al., 1997), we define three types of Web users: Casual (user does not have knowledge about the subject); Intentional (user has some knowledge, or at least a significant interest and Specialist (user has a lot of knowledge on the subject).

In the proposed approach, we focus on Casual or Intentional Web users with the following Web search goals: Open (direct) or Advice (Rose and Levinson, 2004). In this sense, the experiments we have conducted (described in section 4) are related to getting simple textual information about a specific subject: health. Our motivation is related to the fact that there are a set of urban legends, myths, or rumours related to health on the Web, addressed to Web users with little knowledge about medicine.

3.2 Scenario

The aim of the proposal approach is to provide information about the verifiability degree of a Web page. This information gives more subsidies to users so they can better judge the information quality. Besides, they can be used to re-rank the results provided by search engines. In this sense, the proposed approach complements the analysis provided by other quality indicators like, for instance, link structure analysis.

Thus, the approach can be used in a tool (e.g. an extension of a Web Browser) that receives an information request (a set of URL's of Web pages returned by a search engine) and determines the degree of verifiability of each Web page. Details about the implementation of this tool are beyond of the scope of this paper, but, briefly, some aspects of the implementation follows the considerations stated in Section 3.3.

3.3 Identifying References

Figure 1 gives an overview of the reference's extracting process. Since a Web page may contain other types of content (e.g., user comments, advertising, legal disclaimers, etc.) the first step is to identify where its main content is. In this sense, some works address the main content extraction problem (Kohlschütter and Nejdl, 2008) (Kato et al., 2008).

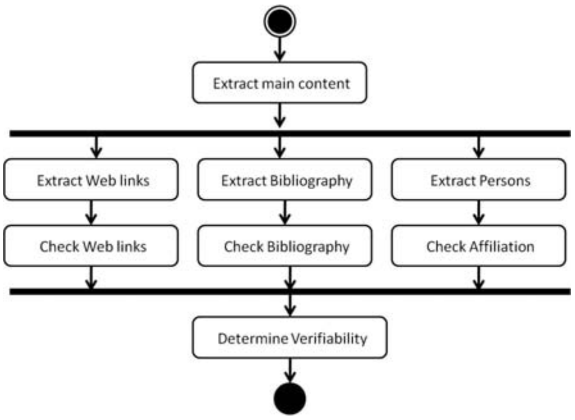

Figure 1: Reference's identification process.

The next step is to extract the references (hyperlinks, references to persons and bibliographic sources). Obviously, the process to identify hyperlinks on the main content is easier than the identification of references to papers or people because it is possible to identify Web links by HTML tags.

In the case of references to papers, when there are no specific links to them, it is necessary to identify where the references are. In this sense, some heuristics consider expressions containing the following key words: *references; selected references; see also; further reading on the subject; to learn more about; additional reading; find out more about; more information;* etc. We have observed that references are frequently placed after the main content.

For identifying these references, one possibility is to use a method similar to the one used in (Kato et al., 2008) for identifying the name of the author in a Web page. Besides, we have found that in online newspaper articles available in Web pages, in general, there are no references to papers or Web links. Then, references to persons must be considered.

For instance, in paragraphs (1) and (2), we present two texts samples regarding this problem. Text (2) has more verifiability than (1) because the affiliation is mentioned.

"[...] Several years ago, I learned of the discovery of Richard R. Vensal, D.D.S. that asparagus might cure cancer[...]"[2]. (1)

"[...] 'The results [...] can be exploited for cancer therapy,' says Dario Altieri, director of the University of Massachusetts Cancer Center in Worcester[...]"[3]. (2)

The degree of complexity to identify persons is higher than to identify references to papers or Web links in a Web page.

For identifying references to people as an information source, we use a *Named Entity Recognizer - NER* (Finkel et al., 2005). The process output is shown in (3).

"'The results […] can be exploited for cancer therapy,' says <PERSON>Dario Altieri</PERSON>, director of the <ORGANIZATION>University of Massachusetts Cancer Center</ORGANIZATION> in <LOCATION> Worcester </LOCATION>." (3)

We consider persons as information sources only when the affiliation is mentioned. For identifying affiliation, we define a set of rules in a grammar constructed with *JavaCC*. Thus, the affiliation is identified by expressions like (4) where *NE* is a Named Entity.

NE<Person> "of the" NE<Organization>
NE<Person> "director of the" NE<Organization>
NE<Person> "a"+ (("a" - "z"))+ "at the" NE<Organization> (4)

After, we check the affiliation on the Web. To do this, we use Google API[4], querying the affiliation name (the aim is to identify the Web site of the organization). This process is easy because, in general, Google returns the site of the organization as the first result. Then, a new query is submitted using the person name and the Web site of the organization as argument (e.g., see 5 bellow).

"John Smith" site:www.organization.edu (5)

When the query (5) does not return any result, we consider that the degree of verifiability is low. Considering the text examples presented before (1 and 2), it was possible to identify references to *Dario Altieri* on the Web site of *University of Massachusetts Cancer Center*. In the case of *Richard R. Vensal* there was no information about his affiliation, thus text (1) was considered as having low verifiability.

As shown in Figure 1, the degree of verifiability of each text is determined in the end of the process. At this moment, we consider that a Web page with any type of reference (according to our quality criteria – see section 4) is verifiable. In the future, we will improve this criterion.

4 EXPERIMENTS

We conducted some preliminary experiments to evaluate if references could be useful to determine the quality of the content of textual Web pages.

4.1 Experiment 1

In a first experiment, we have selected 50 Web pages associated to health related themes. From these, 25 can be considered as good (i.e., verifiable) and 25 as bad (not verifiable). The 25 bad Web pages contain known urban legends, myths, or rumours (identified in Snopes.com[5]). The good Web pages were selected from Health related Web sites (e.g., MedlinePlus[6]).

The results (Table 1) indicate that only 16% of the good Web pages do not have references to some kind of source. In the case of bad Web pages, 9 Web pages have some kind of reference, but only 3 Web pages (12%) have good references, following our criteria.

Table 1: Web pages with and without references.

Category	**Total**	**Web pages with good references**	**Web pages with bad references**	**Web pages without references**
Good Web Pages	25	19	2	4
Bad Web Pages	25	6	3	16

4.2 Experiment 2

In another experiment, we collected the first 30 Web pages returned by Google and used them in the experiment. We considered that a Web user tends to view, in general, only the first Web pages returned by a Web search engine (Hawking et al., 2001).

We decided to focus on the context of *cure of cancer based on asparagus*, since it is a known myth/rumour (according to Snopes.com[5]). To retrieve Web pages, the query was: *asparagus cancer cure*. We only considered Web pages containing textual content and discarded the ones in which the content was a video or e-mail. Regarding the user profile, we selected Web pages appropriately, to users with little knowledge about medicine (casual or intentional user).

The Web pages returned by the search engine were manually evaluated and classified as Good (do not support the myth/rumour), Medium (mentioned the myth/rumour but express doubts about) or Bad (support the myth/rumour). Table 2 shows this classification.

Table 2: Search results.

Category	**Number of Web pages**
Good	6
Medium	8
Bad	16

Using the Web pages selected, we evaluated distinct quality indicators (Cosine, information-to-noise, inlinks to Web page, size of the site and references to sources).

After extracting the main content from Web pages (manually), we have used Google API[5] to obtain the number of inlinks and the number of Web pages in a Web site. The cosine was calculated following Salton et al.'s definition (Salton et al., 1975), using the main content of each Web page and the query "asparagus cure cancer". We calculated the information-to-noise by dividing the number of words in the main content by the total number of words in the Web page.

Regarding verifiability, we analyzed the presence or absence of references to sources related to the main content. In this case, part of the process was manually performed (the identification of references to papers). For identifying references to persons, we used a Named Entity Recognition - NER (Finkel et al., 2005).

We considered the following types of references as quality indicators:

- Links to Web pages of others Web sites;
- References to papers;
- References to persons.

For each type of reference, we considered the following quality criteria:

- A Web link must point to a good Web page. In our approach, a good Web page is the one that belongs to Web sites finished by .gov or .edu, or a Web pages with good references;
- A paper related to a reference must exist on the Web (must be indexed by Scholar Google);
- A reference to a person must contain the complete name of the person and its affiliation (e.g. *John Smith of NASA*).

When a Web page has a good reference (following our criteria), we assign the value 1 to the degree of verifiability. In another hand, we assign the value 0 to the degree of verifiability when a Web page does not have any good reference.

The Table 3 contains the results of this experiment. For each metric we generated a ranking of textual Web pages and computed precision at 5. Precision indicates how many good Web pages appear near the top (precision of 0.20 at 5 means that 1 of top 5 are good Web pages).

Table 3: Experiment 2 - results.

Quality Indicator	Precision at 5
Original Ranking	0.4
Cosine	0.2
Information-to-noise	0.4
Inlinks to Web page	0.2
Size of site	0.2
References to sources	0.6

The best results were obtained using references to sources. Besides, this preliminary experiment shows that metrics like number of pages in a Web site (the size of Web site), which were considered useful in some of the previous work (Section 2), now, apparently, were not so useful in our case. In the case of the size of Web site, the problem is that some Web sites are Web applications that allow to any Web user to publish content. Thus, the number of Web pages does not represent how much effort the author is devoting to the Web site (Amento et al., 2000).

5 FINAL REMARKS

This work presented an approach that is based on verifiability to evaluate the textual Web page's content quality. Considering the results of the preliminary experiments performed, we consider that the use of these references is promising.

There are some difficulties to extract references from textual Web pages (e.g. identification of main content, person names, homonymous, name variations, check if the subject of a source is related to Web page that makes reference to this source, etc.). In this sense, we mentioned some works that can help with some of these problems, and we intend to apply some of these techniques in a more effective way.

Another future work consists on distinguishing the verifiability degree of Web pages. Besides, we are going to define how to combine metrics based on verifiability with other metrics. We also intend to do more experiments.

One limitation of the approach is that the author of Web content can include good references to increase the trustworthiness of bad Web pages. In this sense, we note that, in general, bad textual Web pages do not provide any good reference. Besides, one possibility is to give to users an explanation about the references, identified by the approach on the textual Web page. How to produce this explanation is a future work.

ACKNOWLEDGEMENTS

This work is partially supported by CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil and CAPES, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil.

REFERENCES

Amento, B., Terveen, L., and Hill, W., 2000. Does "authority" mean quality? predicting expert quality ratings of web documents. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in* *information retrieval*, pages 296–303, New York, NY, USA. ACM.

Batini, C., Cabitza, F., Cappiello, C., and Francalanci, C., 2008. A comprehensive data quality methodology for web and structured data. *Int. J. Innov. Comput. Appl.*, 1(3):205–218.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A., 2009. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3):1–52.

Bethard, S., Wetzer, P., Butcher, K., Martin, J. H., and Sumner, T., 2009. Automatically characterizing resource quality for educational digital libraries. In *JCDL ’09: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 221–230, New York, NY, USA. ACM.

Brin, S. and Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117.

Dalip, D., Gonçalves, M. A., Cristo, M., and Calado, P., 2009. Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In *JCDL ’09: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 295–304, New York, NY, USA. ACM.

Denecke, K. and Nejdl, W., 2009. How valuable is medical social media data? content analysis of the medical web. *Inf. Sci.*, 179(12):1870–1880.

Finkel, J. R., Grenager, T., and Manning, C., 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL ’05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Morristown, NJ, USA. Association for Computational Linguistics.

Garzotto, F., Mainetti, L., and Paolini, P., 1997. Designing model hypermedia applications. In *HYPERTEXT ’97: Proceedings of the eighth ACM conference on Hypertext*, pages 38–47, New York, NY, USA. ACM.

Hawking, D., Craswell, N., Bailey, P., and Griffihs, K., 2001. Measuring search engine quality. *Inf. Retr.*, 4:33–59.

HONCODE, 2009. Health on the net foundation. http://www.hon.ch/

Kato, Y., Kawahara, D., Inui, K., Kurohashi, S., and Shibata, T., 2008. Extracting the author of web pages. In *WICOW ’08: Proceeding of the 2nd ACM workshop on Information credibility on the web*, pages 35–42, New York, NY, USA. ACM.

Kohlschütter, C. and Nejdl, W., 2008. A densitometric approach to web page segmentation. In *CIKM ’08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1173–1182, New York, NY, USA. ACM.

Naumann, F. and Rolker, C., 1999. Do metadata models meet iq requirements. In *Proceedings of the International Conference on Information Quality (IQ)*, pages 99–114.

Naumann, F. and Rolker, C., 2000. Assessment methods for information quality criteria. In *Proceedings of the International Conference on Information Quality (IQ), Cambridge, MA*, pages 148–162.

Pernici, B. and Scannapieco, M., 2002. Data quality in web information systems. In *ER ’02: Proceedings of the 21st International Conference on Conceptual Modeling*, pages 397–413, London, UK. Springer-Verlag.

Rose, D. E. and Levinson, D., 2004. Understanding user goals in web search. In *WWW ’04: Proceedings of the 13th international conference on World Wide Web*, pages 13–19, New York, NY, USA. ACM.

Salton, G., Wong, A., and Yang, C. S., 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Wang, R. Y. and Strong, D. M., 1996. Beyond accuracy: what data quality means to data consumers. *J. Manage. Inf. Syst.*, 12(4):5–33.

Yin, X., Han, J., and Yu, P. S., 2007. Truth discovery with multiple conflicting information providers on the web. In *KDD ’07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1048–1052, New York, NY, USA. ACM.

Zhu, X. and Gauch, S., 2000. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *SIGIR ’00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 288–295, New York, NY, USA. ACM.

[1] http://en.wikipedia.org/wiki/Wikipedia:Verifiability

[2] http://urbanlegends.about.com/od/medical/a/asparagus_cancer.htm

[3] http://www.newscientist.com/article/dn10971-cheap-safe-drug-kills-most-cancers.html

[4] http://code.google.com/apis/ajaxsearch/web.html

[5] http://www.snopes.com/medical/disease/asparagus.asp

[6] http://www.nlm.nih.gov/medlineplus/