

# Reliability and Verification of Natural Language Text on the World Wide Web

Melanie J. Martin  
Department of Computer Science  
New Mexico State University  
P.O. Box 30001, MSC CS  
Las Cruces, NM 88003  
mmartin@cs.nmsu.edu

## ABSTRACT

The hypothesis that information on the Web can be verified automatically, with minimal user interaction, will be tested by building and evaluating an interactive system. In this paper, verification is defined as a reasonable determination of the truth or correctness of a statement by examination, research, or comparison with similar text. The system will contain modules for reliability ranking, query processing, document retrieval, and document clustering based on agreement. The query processing and document retrieval components will use standard IR techniques. The reliability module will estimate the likelihood that a statement on the Web can be trusted using standards developed by information scientists, as well as linguistic aspects of the page and the link structure of associated web pages. The clustering module will cluster relevant documents based on whether or not they agree or disagree with the statement to be verified. Relevant references are discussed.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval -- *clustering, information filtering, query formulation, relevance feedback*; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing -- *linguistic processing*

## Keywords

World Wide Web, reliability, quality, filtering

## 1. INTRODUCTION AND MOTIVATION

With the explosive growth of the World Wide Web has come, not just an explosion of information, but also an explosion of false, misleading and unsupported information. At the same time, the web is increasingly being used for tasks where information quality and reliability are vital, from legal and medical research by both professionals and lay people, to fact checking by journalists and research by government policy makers.

In this paper we define reliability as a measure of the extent to which information on a given web page can be trusted. By verifiable we mean a reasonable determination of the truth or correctness of a statement by examination, research, or comparison with similar text.

We need to go beyond retrieving relevant information to be able to determine whether the information is reliable and can be verified. For example, if Web users were able to determine a level of authority, or reliability, of the Web page they are currently viewing, they could decide whether additional verification would be necessary. A general consumer of information looking up, say, a treatment for a minor ailment might be satisfied to know the reliability of the page they are offered, or perhaps to compare the reliability rankings of two or three pages with different advice. Others, however, such as scholars, students, journalists and policy makers, would need to verify sources, check statistics and find other Web documents whose authors agree or disagree with a given text segment. This process of checking reliability and verification may need to be repeated until the user has reached a level of certainty that they deem sufficient for their use. Thus there is potential for all Web users to benefit from information about the reliability and verifiability of text on the Web.

I propose to develop a system that would verify information on the Web and return a measure of its reliability, where reliability and verify are as defined above. The four pilot studies I have conducted have shown that this is not an easy task for humans. My proposed system would work with the human user, as a team, to find the most relevant and reliable information available on the Web. Clearly constructing such a system is both valuable and desirable. What follows are some research questions, a more formal statement of the task, my proposed methodology for conducting the necessary research and constructing the system including background and related work, and some issues for discussion.

## 2. QUESTIONS

In considering the development of a system that would enable users to verify information on the Web and measure its reliability, some questions I would like to answer affirmatively arise:

- Given that the relevant documents or sections of relevant documents can be retrieved, can we provide a meaningful ranking of the reliability of the documents?

- Given that the relevant documents or sections of relevant documents can be retrieved, can we cluster them in a meaningful way, based on the extent to which they agree or disagree with a given segment of text?
- Given that some queries may be unanswerable, can we figure out when to stop searching?

To answer these questions I plan to develop a system that would enable a user, browsing the web, to highlight text they would like to verify. The user would receive a report summarizing the relevant information available on the Web, its reliability and the extent to which it agrees or disagrees with the highlighted text.

### 3. FORMAL STATEMENT OF THE PROJECT

I will develop an interactive computer system which takes a segment of text on the Web highlighted by a user as input. First, the system will return a reliability ranking for the Web page containing the segment of text. If the reliability measure or ranking is sufficient information for the user, the process may stop at this point. (See Figure 1.)

If the user elects to continue, the system will work with the user to form a query. To do so the system needs to clarify what the user wants: what is it exactly that the user wants to verify and what does the user considers relevant. That accomplished, the query will be passed to a search engine and the system will then search the Web for relevant documents and retrieve them. The retrieved documents will be ranked for reliability and presented to the user in a clustered format, based on the extent to which they agree with the information in the input text. Additional statistics about the set of documents retrieved will also be returned along with links to the documents themselves. The user will have the option to repeat this process with queries enhanced automatically or by the user, until the user is satisfied with the results.

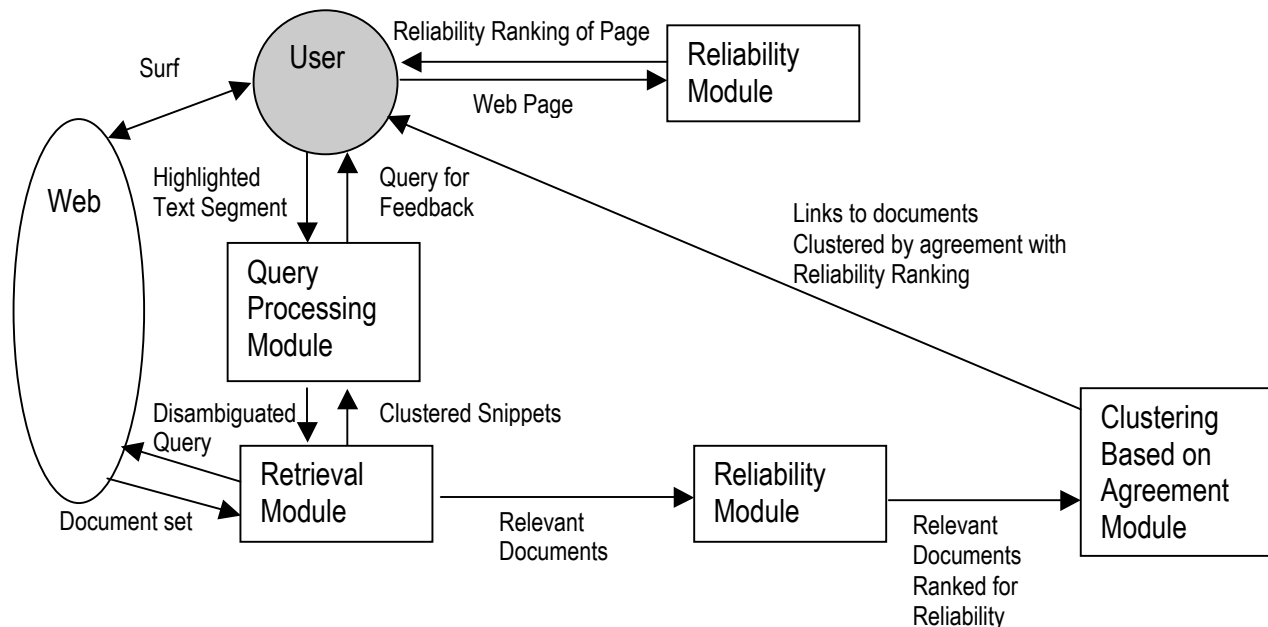


Figure 1. Architecture of Proposed System

### 4. BACKGROUND AND METHODOLOGY

The four most significant components of my system are reliability ranking, query processing, document retrieval, and clustering based on agreement. The components for ranking reliability and clustering based on agreement will be my own innovations, while the other components will make use of standard IR techniques. The development of an end-to-end system is required to properly evaluate my reliability and clustering components. In this section, I will briefly discuss the methodology for each component and provide selected background and related work to give an indication of the basis for each component.

#### 4.1 The Reliability Ranking Component

There is a growing body of work, particularly in the medical community, which highlights problems with the reliability of information on the web [7] and proposes criteria to determine reliability [5]. At present, the automatic exploration and implementation of reliability criteria is limited. Zhu and Gauch [16] implement a “quality metric” with the goal of improving search engine effectiveness using primarily superficial features. Tang et al. [13], in the context of the development of a High Quality Interactive Question Answering system, report empirical studies of machine prediction of information quality in the news domain. They used domain experts to develop quality criteria and rank their importance, resulting more linguistically oriented features.

The reliability measure, which I will develop, to estimate the likelihood that the information on a given Web page can be trusted, will build on previous work by taking into account standards developed by information

scientists, linguistic aspects of the text on the page, as well as the link structure of the Web associated with the page. Some questions I will consider include:

- Is evidence provided and does argument proceed logically?
- Are sources for the evidence cited and verifiable?
- Are sources primary (original)? If not, how far are they removed from the primary sources?
- Is the evidence provided by experts? Do non-related experts agree?

## 4.2 The Query Processing Component

Given the segment of text highlighted by the user, which the user would like to verify, there are two main concerns in understanding the segment: the user's focus [9] of interest in the text and ambiguity of individual words. Suppose a user wants to verify the statement:

*"The longest reptile is a python, which can reach 33 feet in length."*

The user's focus could be on the comparative length of reptiles, or on whether pythons 33 feet long have actually been measured. My system will work with the user to determine the focus, by presenting focus options the user can choose from. If the focus is on python length, the word "python" becomes ambiguous: a web search on "python length" will return irrelevant results about the length of strings in the Python programming language. My system will return initial retrieval results, clustered by topic, so that the user can select the relevant set.

I propose to test two standard algorithms for improving queries: relevance feedback, as in Rocchio [12], and query expansion, as in Magennis [10], to determine which method, or if a combination of the methods, works best in my system. In the case of relevance feedback, the documents returned to the user for feedback, will be clustered by topic, either using Kleinberg's HITS algorithm [8] or an appropriate clustering algorithm. A good candidate for clustering is the Scatter/Gather algorithm introduced by Cutting et al. [2] and further developed by Hearst and Pederson [6], to cluster the results of the original query by topic. The user can then choose the topic that they consider relevant for the next iteration of the search and retrieval.

## 4.3 The Document Retrieval Component

When retrieving documents from the Web it is crucial to retrieve a sufficient number of relevant documents. For retrieving documents from the web, I will use available search engines. My system will work with the user to determine relevance, which is a human determination of the similarity of a given document to the users query. One way to improve web search results and to aid the user in relevance determinations is clustering. For example, Zamir and Etzioni's Suffix Tree Clustering [15] uses the snippets from the output of the MetaCrawler search engine to create clusters of web documents based on shared phrases. For post-processing the documents determined to be relevant, I will use vector-space-based retrieval models, in particular, Latent Semantic Analysis (LSA) [3] [4]. I plan to explore modifications of LSA, such as Ando's Iterative Residual Scaling [1]. Ando's method in particular shows a great deal of promise for smaller corpora, making it ideal for processing a relatively small set of documents which the user deems relevant.

## 4.4 Clustering Based on Agreement

Clustering is the application of one of a family of techniques from multivariate statistics in order to group elements of a data set into natural categories. The basic idea is that a given data element will have high similarity to other data elements in its cluster and low similarity to data elements outside of its cluster. This component will require the development of a similarity measure that can distinguish agreement from disagreement, the choice of an appropriate clustering algorithm, and possible improvements to the clustering algorithm. Distinguishing agreement from disagreement will require the use of linguistic features and will build on computational linguistic work in the data mining of product reviews (e.g. films, books, appliances) [11] and subjectivity [14].

## 4.5 Evaluation

As each component is completed, I will evaluate it using standard evaluation measures from the discipline where the methodology is most commonly used. For example, precision and recall are standard Information Retrieval evaluation measures, which will be used to evaluate the retrieval component. The system as a whole will be evaluated through a user study.

## 5. ISSUES FOR DISCUSSION

Some of the issues raised in my paper, which I would find helpful to discuss at the Doctoral Consortium are:

- Given that the task of determining reliability and verification are difficult for humans to perform, how well can we expect a machine to do? Are we asymptotically limited by human performance? Are machines suited to perform this task better than humans?

- Are there disadvantages to viewing the human and machine as a team?
- Beyond the user making the decision, is there a way to know when to stop iterating the reliability and verification process?
- How to access parts of the web not indexed by search engines (which may be needed for verification purposes)?
- Initial work on reliability has been done in the medical domain, what steps should be taken to ensure that it generalizes as well as possible?

## 6. ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Roger T. Hartley for his patience and guidance; also the members of my doctoral committee: Dr. Peter W. Foltz, Dr. Stephen Helmreich, Dr. Enrico Pontelli and Dr. Son Cao Tran for their help and advice.

## 7. REFERENCES

- [1] R. K. Ando and L. Lee. Iterative residual rescaling: An analysis and generalization of lsi. *In Proceedings of ACM SIGIR 2001*. SIGIR, 2001.
- [2] D. R. Cutting, J. O. Pedersen, D. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. *In Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318-329, 1992.
- [3] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391-407, 1990.
- [4] S. T. Dumais, G. W. Furnas, T. K. Landauer, and S. Deerwester. Using latent semantic analysis to improve information retrieval. *In Proceedings of CHI'88: Conference on Human Factors in Computing, New York*, pages 281-285. ACM, 1988.
- [5] G. Eysenbach, J. Powel, O. Kuss, and E.-R. Sa. Empirical studies assessing the quality of health information for consumers on the world wide web. *JAMA*, 287(20):2691-2700, 2002.
- [6] M. A. Hearst and J. O. Pederson. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. *In the Proceedings of ACM SIGIR '96, Zurich*, pages 84-89. ACM SIGIR, August 1996.
- [7] P. Impicciatore, C. Pandolfini, N. Casella, and M. Bonati. Reliability of health information for the public on the world wide web: a systematic survey of advice on managing fever in children at home. *BMJ*, 314:1875-9, 1997.
- [8] J. Kleinberg. Authoritative sources in a hyperlinked environment. *In Proceedings 9<sup>th</sup> ACM-SIAM Symposium on Discrete Algorithms*, 1998. Extended version in *Journal of the ACM* 46(1999). Also appears as IBM Research Report RJ 10076, May 1997.
- [9] W. Lehnert. *The Process of Question Answering*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1978.
- [10] M. Magennis and C. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. *In the Proceedings of ACM SIGIR '97*. ACM SIGIR, 1997.
- [11] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *In Proceedings of Empirical Method in Natural Language Processing 2002*. EMNLP, 2002.
- [12] J. J. Rocchio. Relevance feedback in information retrieval. *In The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313-323, Englewood Cliffs, NJ, 1971. Prentice-Hall, Inc.
- [13] R. Tang, K. B. Ng, T. Strzalkowski, and P. B. Kantor. Automatically predicting information quality in news documents. *In Late-breaking Paper in Proceedings of Human Language Technology - North American Chapter of the Association for Computational Linguistics 2003*, 2003.
- [14] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational Linguistics*, To Appear.
- [15] O. Zamir and O. Etzioni. Web document clustering: A feasibility demonstration. *In the Proceedings of ACM SIGIR '98*, pages 84-89. ACM SIGIR, 1998.
- [16] X. Zhu and S. Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. *In Proceedings of the 23<sup>rd</sup> annual international ACM SIGIR conference on Research and development in information retrieval*, pages 288-295. ACM Press, 2000.