Filtering Web Pages for Quality Indicators: An Empirical Approach to Finding High Quality Consumer Health Information on the World Wide Web

Susan L. Price, MD, MS and William R. Hersh, MD
Division of Medical Informatics and Outcomes Research,
Oregon Health Sciences University, Portland, OR

The World Wide Web is an increasingly popular source for consumer health information, but many authors have expressed concerns about the quality of health information present on the Internet. We have developed a prototype system that responds to a consumer health query by returning a list of Web pages that are ranked according to the likely quality of the page contents. A computer program identifies some of the criteria that have been suggested for assessing the quality of health information on the Internet. It also identifies characteristics that may serve as proxies for desirable (or undesirable) qualities that are difficult to assess directly using an algorithm. Intervening in the search process and automatically analyzing the contents of each page returned by a general search engine may facilitate the search for high quality consumer health information on the Web.

INTRODUCTION

Searching the World Wide Web (the Web) for health information is now a common activity. A 1997 survey found that 36.7% of Internet users search the Internet for health and medical information.[1] A 1998 telephone survey found that 15% of the U.S. population, and 30% of the U.S. population who have Internet access, use the Web to find health information.[2] Thus the World Wide Web is delivering a substantial amount of health information to consumers.

There are numerous advantages of seeking health information on the Web. A huge volume of information is available, some of it very useful information from reputable sources, targeted at nonprofessional consumers of health and medical information. Information is available to anyone, at any time, and from anywhere that access to the Internet is available. Most of the information is free and can be viewed anonymously.

There are also disadvantages of seeking information on the Web. Publishing on the Web is very easy and very inexpensive. As a result, there are myriad webpages to be sifted by whatever human or electronic means a consumer uses to retrieve and sort information. Useful information is often difficult to locate, and information retrieved may be of dubious quality.

Investigators have tried to facilitate the search for medical information by limiting the search to a database of sites that have been identified by a computer algorithm as medical sites,[3] and by intervening in query formation, using knowledge of the medical domain to expand the user's query.[4] General search engine technology has also improved. A small preliminary analysis suggested that top ranked pages are quite likely to be related to the topic of a medical query. Unfortunately, topical relevance does not guarantee usefulness. Many of the pages retrieved during this preliminary analysis consisted of bulletin board or newsgroup postings, personal home pages with anecdotal information, or lists of links to other pages. Furthermore, improving the precision or the recall of a search does nothing to ensure the quality or credibility of the information retrieved.

Evaluating the credibility of consumer health information on the Web is particularly challenging, and important. The Web is a convenient medium to pursue an agenda reflecting political or intellectual bias as well as to seek commercial gain. Many authors have expressed concerns about the quality of medical information on the Web, and about the feasibility and desirability of rating consumer health information on the Web.[5-10]

Several approaches to this problem are being explored. One approach is self-regulation. The Health on the Net Foundation has developed a set of principles called the Net Code of Conduct.[11] Websites can display the HONcode logo to indicate their voluntary adherence to these principles. Only a small number of webpages now exhibit the HONcode logo, so limiting a search to pages that display the logo would severely limit the information provided.

A second approach is to provide consumers with a rating tool, such as a checklist, to evaluate websites. For example, the Health Information Technology

1091—8280/99/$5.00

Institute of Mitretek Systems, Inc. (HITI) has proposed an extensive list of criteria that can be used to assess the quality of health information on the Internet. They propose to use this list as the basis for a tool that consumers will be able to use.[12] A potential limitation is its dependence on consumer willingness to use a rating form.

Another approach is for third parties to publish reviews on the Web so that consumers can determine whether a website has been deemed to be of high or acceptable quality. Limitations to this approach include the introduction of the biases of the reviewers and the inability of the consumer to specify, or sometimes even know, the criteria used by the reviewers. The ratings themselves may be responsible for misleading or misinforming consumers.[6] Furthermore, websites are frequently added, removed, and changed. Maintaining a comprehensive, current list of ratings will be difficult and expensive. In addition, consumers may resist using a rating service if a separate webpage must be accessed to view the rating.

An approach that tackles both the issue of quality and the problem of the huge number of web pages that a search may return is to use human reviewers to sift health information and to create lists of sites that are deemed useful or credible or both. The results may consist of single lists of pages or of large collections of lists that can be searched or browsed by topic and are most commonly maintained by libraries and educational institutions. Examples of this approach include New York Online Access to Health (NOAH)[13], and NetWellness.[14] Similar sites developed for health professionals, such as Cliniweb[15] and Medical Matrix[16] are also available to consumers. Again, hiring reviewers with enough medical knowledge and critical expertise to review and select websites is expensive, so that keeping such sites current and comprehensive is difficult.

Eysenbach and Diepgen suggest that software residing on a user's browser could automatically filter information using both author-supplied metadata and metadata from third party rating services. The user could customize the software to filter information based on individual interests and quality requirements. They suggest that volunteer physicians could contribute to a decentralized rating system, which would solve some, but not all, of the problems associated with rating websites. They also suggest that automatic assessment of websites using indirect quality indicators, such as the number of hyperlinks to a site, the number of visitors per day from particular user groups, and user behavior statistics, could help distinguish higher quality websites from lower quality websites[17]

We propose that there is a role for a software tool that can automatically assess the quality of consumer health webpages. Although any such system is unlikely to be perfect, even an imperfect initial filtering will reduce the number of webpages to be critically examined, either by the consumer or by a human reviewer. We describe a prototype that examines webpages and assigns a score that indicates the likelihood that each page will meet quality criteria such as those proposed by HITI[12] and HON.[11]

METHODS

Description of prototype

The logical organization of the prototype is shown in Figure 1. A user enters a query on a Web-based form that is transmitted using CGI to a Perl program. A subroutine removes stop words, then forwards the

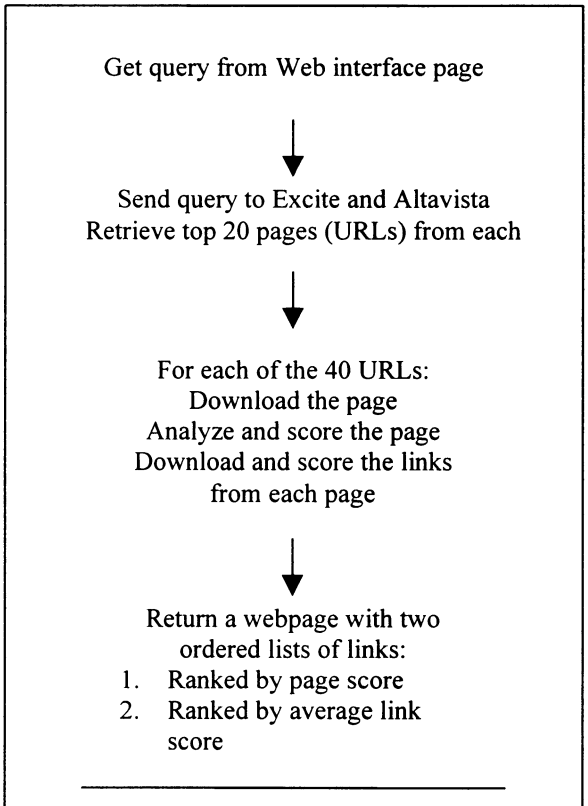

Figure 1 - Logical organization of the filtering tool. It retrieves search results, downloads and analyzes each page, then returns two ordered lists of links to the user.

query to two general search engines, Excite and Altavista. These two search engines were chosen because they are popular and provide wide coverage of the Web. Additional search engines could be added in the future, but were not necessary as part of this proof-of-concept investigation.

The twenty top ranked links are retrieved from each search engine. The number twenty was chosen as an initial arbitrary number to provide enough links to make reranking the links meaningful, yet not so many as to be unwieldy during development. The program discards any duplicate links and any links that return an error message instead of an HTML page. The contents of the page associated with each link is placed into a temporary file. The file contains the HTML code needed to produce the page that will be displayed if the link is followed plus any metatags present.

The tool analyzes the contents of each temporary file using a series of algorithms to detect various components of each page. The algorithm detects as many of the characteristics included in the HON[11] and HITI[12] criteria as possible. It also detects some characteristics that, while neither desirable nor undesirable, serve as heuristic proxies for positive or negative qualities. The tool assigns scores to each page for likelihood of exhibiting certain attributes, then calculates a weighted sum of these scores, which is used to rank the webpages.

The tool also follows the links on each webpage retrieved and downloads the page contents for analysis. The tool then assigns a score to each link and calculates an average link score for each of the originally retrieved webpages. The scores and the weighting assigned to various characteristics are chosen empirically. The modular design of the tool allows the weighting to be adjusted easily so that the performance of the tool can be finely tuned.

Once all the pages have been scored, the tool returns two ordered lists of links to the user. The first list is in order of predicted page quality. The second list consists of links to all the pages that contain at least six links (internal or external) in order of average link score. The purpose of the second list is to indicate pages likely to contain lists of potentially useful links for the user who desires a more comprehensive collection of existing Web resources.

Initial emphasis has been placed on flexibility and ease of development, not on speed or efficiency. The current implementation of the tool is a collection of Perl programs that run on a Sun Ultrasparc 140 running the Solaris 2.5.1 operating system.

Automatic Analysis of Web Pages

After reviewing published suggestions for evaluating health information on the Web, the following criteria were selected for automatic assessment because they are both desirable and potentially amenable to automatic evaluation: relevance, credibility, absence of bias, content, currency, and value of links.

1. *Likely relevance*

The general search engines that are queried provide a screening for topical relevance; a small preliminary analysis indicated that they are often successful. This part of the algorithm is designed to discover pages that are less likely to be useful, such as chat room or bulletin board postings.

2. *Likely credibility*

This module determines whether the webpage has indicators that it is likely to be a credible source of information. For example, this module contains subroutines that inspect the URL, look for authorship information, determine whether the site displays the HONcode logo, and search for particular words or phrases, such as "miracle cure."

3. *Likely bias*

This module identifies specific words and phrases, such as "mastercard" and "visa," that suggest a commercial bias may be present.

4. *Content*

An ideal program would determine whether the content is accurate, useful, and relevant to a particular user, but such analysis is not practical. Instead, this program determines how much text is displayed, and the ratio of hyperlinks to text. The hypothesis is that pages with minimal text are less likely to be valuable to the consumer than pages with more text.

5. *Currency*

The algorithm searches for evidence indicating when the page was published or last updated. If found, it determines how recently the update was done.

6. *Value of links*

The tool ranks each webpage according to the likelihood that it is a useful source of multiple links. It follows both internal and external hyperlinks, analyzes the associated webpages, and assigns a score to each link. It then calculates an average link score for each page.

Preliminary Evaluation

Two preliminary evaluations of this tool were done. The first evaluation determined whether the tool can successfully retrieve search engine results for a given query, and retrieve, analyze, score, and rank the associated Web pages. Health related queries were entered into the initial HTML form, and the Web pages the tool returned were then examined. Subroutines used during development printed the results of various parts of the scoring algorithm into separate files. The second evaluation determined whether the scoring algorithm can successfully separate webpages that are deemed desirable from those deemed less desirable. The tool was applied to a small test collection of 48 webpages covering nine different medical topics that have been labeled by the investigator as desirable or undesirable.

RESULTS

When consumer health queries were entered into the initial HTML form, the software tool successfully retrieved the query results from the two general search engines and presented the links in two new lists. Inspection of the links confirmed that duplicate links and links producing error messages had been eliminated. The test files demonstrated that scores had been successfully assigned to the links, and the links were correctly returned in order of descending score. The second list correctly contained the pages with 6 or more hyperlinks.

When the tool applied its scoring algorithm to the test set, the tool successfully separated the desirable from the undesirable pages. That is, all the scores assigned to each of the desirable pages were higher than any of the scores assigned to the undesirable pages.

DISCUSSION

Automatic analysis of Web pages for indicators of the quality of information contained is feasible and potentially very useful. As the Web continues to grow, an increasing amount of health information will be available to consumers. This has the potential for educating, facilitating informed decisions, encouraging healthy behaviors, and providing information to groups of consumers that are not well-served by current mechanisms. It also has the potential for propagating misinformation and even causing harm. A major obstacle to obtaining quality health information from the Web is the tremendous volume of information available. Although search engines return webpages based on the probability that a given page is relevant to the query, they provide no guidance as to the credibility of those webpages.

This paper explores the idea of using automatic filtering techniques to identify pages likely to be of high quality. Preliminary results demonstrate that given a set of criteria to evaluate the quality of consumer health information, it is possible to automatically rank webpages in order of likely quality. The weight given to various characteristics can easily be changed to give priority to particular criteria. Clearly further evaluations are necessary, and are being planned to show that

1) Webpages ranked highly by the tool are more likely to have credible, useful information than are the top ranked pages returned by general search engines. The comparison will use manual scores assigned by medically trained reviewers who have not been involved in the development of this tool.

2) Users who are not medically trained can use this tool to more quickly find Web resources to correctly answer specific questions about health related topics

Further development will investigate adding more functionality, such as classifying webpages by reading level, by level of medical sophistication, or by medical paradigm represented (traditional or alternative) so that users can choose the type of page they wish to have ranked most highly. In its current formative stage the prototype evaluates single webpages and is inherently somewhat biased against sites that distribute content over several pages. Because much of the health information on the Web is now presented within large integrated websites that provide other services in addition to educational content, it may be useful to adapt this methodology to automatically evaluate entire websites. It is also possible to modify the program to follow additional layers of links, and even to add highly ranked links to the original retrieval list. However, the more remote these pages are from the original link, the more likely it is that the pages will not relate to the topic of the query. Following additional layers of links may be useful only if an evaluation of topical relevance is added to the algorithm.

Automatic filtering of webpages is likely to have some limitations. The technique described provides first-pass filtering of information. The scoring occurs without respect to the context of an individual search by a user with unique needs and motivations. For example, a low rating might be assigned to a link to a bookseller's webpage. Although the link does not fit the profile of the webpages the tool seeks to return, the link might be useful to a user who would like to read a book about the subject of his query. Automatic filtering may never be able to directly assess some characteristics of webpages, such as

accuracy. Furthermore, preliminary experience with this technique suggests that it is much easier to identify indicators of undesirable webpages, than it is to identify indicators of high quality webpages.

Our preliminary results suggest that more research should be done regarding the criteria used to evaluate webpages. It is important to distinguish the quality of information from the characteristics of the medium in which it is presented. For example, obsolete information, such as recommendations no longer supported by scientific evidence, is of low quality. Several published lists of criteria for evaluating webpages state that the date that a page was updated should be displayed.[5, 11, 12] But, displaying the date on which a webpage was last updated may or may not be correlated with currency of the information on the page. Similarly, the presence of information about authorship on a webpage may or may not be correlated with the quality of information on that page. The criteria being used in development of this software are adapted from those proposed by respected authorities in the medical field. But, these criteria have not been formally validated. Nor have the published criteria been shown to reflect the qualities that consumers themselves are seeking from health information on the Web. It would be useful to compare the criteria that medical professionals and medical informatics specialists would apply to the evaluation of health information to those that consumers would apply.

In summary, although automated analysis of webpages does not eliminate the need for critical evaluation of information, automatic filtering of webpages can expedite a search for high quality information on the World Wide Web. Automatic filtering of webpages is likely to be a useful adjunct to searching, and is a technique that could be adapted to other domains as well.

References

1. Brown MS. Consumer Health & Medical Information on the Internet: Supply and Demand. Summarized at: http://erg.findsvp.com/health/mktginfo.html (last visited 11/28/98) 1997.
2. Nammacher MA, Schmitt K. Consumer Use of the Internet for Health Information: a Population Survey, AMIA '98 Annual Symposium, Orlando, FL, 1998.
3. Tay J, Ke S, Lun K. MediAgent: A WWW-based Scalable And Self-Learning Medical Search Engine, AMIA '98 Annual Symposium, Orlando, FL, 1998.
4. Suarez HH, Hao X, Chang IF. Searching for Information on the Internet Using the UMLS and Medical World Search, AMIA '97 Annual Symposium, Nashville, TN, 1997.
5. Silberg WM, Lundberg GD, Musacchio RA. Assessing, Controlling, and Assuring the Quality of Medical Information on the Internet. JAMA 1997; 277:1244-1245.
6. Jadad AR, Gagliardi A. Rating Health Information on the Internet. JAMA 1998; 279:611-614.
7. Kiley R. Quality of Medical Information on the Internet. J Royal Soc of Med 1998; 91:369-370.
8. Hersh WR, Gorman PN, Sacherek LS. Applicability and Quality of Information for Answering Clinical Questions on the Web. JAMA 1998; 280:1307-1308.
9. Impicciatore P, Pandolfini C, Casella N, Bonati M. Reliability of health information for the public on the world wide web: systematic survey of advice on managing fever in children at home. BMJ 1997; 314:1875-9.
10. Wyatt J. Commentary: Measuring quality and impact of the world wide web. BMJ 1997; 314:1879-81.
11. Health On the Net Foundation Code of Conduct for medical and health web sites. http://www.hon.ch/HONcode/Conduct.html (last updated 8/4/98; last visited 11/28/98) 1998.
12. Criteria for Assessing the Quality of Health Information on the Internet. Working draft of a White Paper. http://www.mitretek.org/hiti/showcase/documents/criteria.html Edit date: 14 October 1997 (last visited 11/28/98) 1997.
13. NOAH: http://www.noah.cuny.edu (last visited 2/25/99) 1999.
14. Netwellness: http://www.netwellness.org (last visited 3/02/99) 1999.
15. Cliniweb: http://www.ohsu.edu/cliniweb (last visited 3/02/99) 1999.
16. Medical Matrix: http://www.medmatrix.org (last updated 1/18/99; last visited 3/02/99) 1999.
17. Eysenbach G, Diepgen TL. Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information. BMJ 1998; 317:1496-1500.