DEVELOPMENT OF A SOFTWARE TOOL TO AID IN THE RETRIEVAL OF CONSUMER HEALTH INFORMATION FROM THE WORLD WIDE WEB

by

Susan Loucette Price, M.D.

A THESIS

Presented to the Division of Medical Informatics and Outcomes Research

and the Oregon Health Sciences University

School of Medicine

in partial fulfillment of

the requirements for the degree of

Master of Science

May 1999

School of Medicine
Oregon Health Sciences University

CERTIFICATE OF APPROVAL

This is to certify that the M.S. thesis of

Susan Loucette Price

has been approved

[redacted]

Willia[redacted]

[redacted]

Paul Sher, M.D.

[redacted]

Paul Gorman, M.D.

[redacted]

Warren Harrison, Ph.D.

Associate Dean for Graduate Studies

TABLE OF CONTENTS

FIGURES AND TABLES

ACKNOWLEDGMENTS

I thank my advisor, Dr. William Hersh, for his generous assistance, encouragement, and enthusiasm throughout this project. I thank also the other members of my thesis committee, Dr. Paul Gorman, Dr. Warren Harrison and Dr. Paul Sher for their guidance and helpful suggestions, and for the different perspectives that each one contributed to this project. In addition, I am grateful to the OHSU librarians, particularly Dolores Judkins and Margaret Connors, for supplying the list of consumer health questions and for contributing their ideas about consumer health information needs and resources.

I offer special thanks to my husband, Scott Springer: Without his support and enthusiasm this project would not have been so much fun.

ABSTRACT

The World Wide Web (the Web) is an increasingly popular source for consumer health information, but useful information is often difficult to locate, and some information may be of dubious quality. Several approaches to the problem of poor-quality information on the Web are being investigated, including self-regulation, rating systems, and the creation of selective collections of high-quality webpages by human reviewers. Maintaining current and comprehensive human-indexed collections, or even rating lists, is a formidable task. Automatic assessment of webpages for indicators of quality may be useful both to an individual searcher and to human reviewers who index and rate webpages.

This thesis describes a prototype software tool that intervenes in the process of searching the Web for health information. The tool automatically analyzes a collection of webpages that have been returned by a general search engine and returns an ordered list of webpages to the user, ranked by likelihood of containing useful, high-quality information. Defining quality is a difficult issue, and universal agreement on a set of criteria is unlikely. This project used criteria that, although not validated, have been adapted from those proposed by respected authors in the medical field.

The general criteria that were selected for analysis were usefulness, credibility, absence or explicit disclosure of bias or conflict of interest, quantity of information, and currency of information. Because these general criteria are difficult to assess directly, the software

detects specific characteristics of webpages that may be indicators of these qualities. Some of the characteristics detected are ones that have been proposed as criteria for judging the quality of health-related webpages, such as the presence of an email address for feedback regarding the webpage, or the date the webpage was last updated. Other characteristics detected by the tool are intrinsically neither desirable nor undesirable, but are characteristics that were hypothesized to be proxies for desirable or undesirable qualities. For example, if the presence of large numbers of exclamation marks were associated with low-credibility webpages, then the number of exclamation marks might be a useful proxy for estimating the credibility of a webpage.

Preliminary bench testing of the software tool demonstrated that automatic analysis of webpages is feasible, and that the current algorithm may be somewhat successful at ranking webpages according to set criteria. Further development is likely to result in a more useful tool, and will be necessary prior to a more formal user evaluation. This study suggests that, although automated analysis will not eliminate the need for critical evaluation of information, automatic filtering is likely to be a useful adjunct to searching and may expedite a search for high-quality health information on the World Wide Web.

INTRODUCTION

Providing health information to consumers has the potential to educate, facilitate informed decisions, encourage healthy behaviors, and provide answers to patients' medical questions. The purpose of this project is to help consumers find high-quality health information on the World Wide Web (the Web) by establishing whether it is feasible to produce an improved retrieval set in response to consumer health queries by using a software tool to analyze, filter, and reorder the output from Internet search engines. More specifically, this tool accepts a query in the domain of consumer health, submits the query to two general-purpose search engines, and intercepts the output. The software analyzes the contents of the webpages that correspond to the links that are returned by the search engines in response to the query. Based on the algorithms encoded in the program, it reorders the links, ranking them in order of likely desirability, or quality, then returns the links to the user.

Background

Patients value medical information highly,[1, 2] and many patients want more information about their health than they commonly receive from their physicians.[3-5] There is a current trend toward patients taking more responsibility for their medical care[6] and a growing emphasis on the incorporation of patient preferences into medical decision making.[7] Both trends may be facilitated by enhanced access to health information.

Increased access to health information may also contribute to better health outcomes. Providing patients with health information is a necessary part of empowering them to become active participants in their medical care. Such active participation may affect patients' attitudes about illness and about their recovery. For example, Brody and colleagues studied patients' perceptions of their roles during medical visits and their subsequent attitudes and self-reported improvement. They found that patients who reported an active role reported significantly less discomfort, greater alleviation of symptoms, and more improvement in their general medical condition one week after their doctor visits than did the patients who reported passive roles.[8] Education, combined with active involvement in care, contributed to an improved outcome in diabetic patients[9] and improved functional status in ulcer patients,[10] although education alone does not necessarily lead to altered behavior[11] or better functional status.[10]

With the advent of the World Wide Web, consumers have an enormous source of health information readily available to them, and there is good evidence that they are taking advantage of it. A 1997 survey found that 36.7% of Internet users search the Internet for health and medical information.[12] A 1998 telephone survey showed 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.[13] In March 1998, the National Library of Medicine found that approximately 30% of the searches that originated in the U.S. using the free Web interfaces to MEDLINE were done by members of the general public or by students.[14] Although it is clear that there is a strong demand for health information on the Internet,

very little is known about the effects on public health outcomes of information accessed on the Internet.[15]

The availability of electronic sources of health information and electronic communication about health issues has prompted several authors to predict changes in the patient-physician relationship. McLellan suggests that increased knowledge may have a "leveling effect," changing the balance of power in the relationship between patient and physician, and may contribute to the shift toward greater patient responsibility for health care.[16] Pemberton and Goldblatt suggest that a new paradigm of patient-physician cooperation may emerge. Patients have access to detailed medical information on the Internet, and may have the time and energy to do extensive searches focused on a single medical condition, whereas their physicians may have limited time to devote to information gathering and must deal with a variety of medical problems each day. The patient or family can serve as the physician's research assistant, especially when dealing with rare diseases. The physician can then help the patient or family analyze the information and assess its relevance to the patient's situation.[17] Increased patient access to medical information may also have implications for medical journals that have a new audience, whether or not they seek it,[16] and for academic medicine, which must prepare new physicians to make use of electronic technology and to deal with changes in the patient-physician relationship.[18] Increased access to information may also lead to conflict between patients, who are aware of the latest best-practices recommendations, and physicians, who are too overworked to maintain up-to-the-minute currency about

every topic, or between patients with high expectations and health care systems with limited resources.[18, 19]

In any study regarding the retrieval of consumer health information, it is important to consider consumer preferences with regard to the type of information being retrieved. Tang and colleagues studied the information needs of patients by using patient focus groups. They found that patients not only want more information than they typically receive in a doctor's office, they want custom-tailored information, not just general brochures. They also want information at the time that they formulate their questions, which is usually after, not during, a visit to a doctor. Furthermore, patients want information that their doctor has endorsed as credible, and they want the physician to recommend additional sources of information.[20]

Coulter and colleagues also used focus groups to study patient information materials (printed, audio or video) that could be used to support patients' involvement in treatment decisions. They found that patients wanted balanced information that contains an assessment of the pros and cons of treatment and is frank when outcome probabilities are unknown. Patients wanted information about the full range of treatment possibilities, including those not proven to be effective, with an honest assessment of whether or not the treatments are known to be effective. They preferred information that was facilitative rather than prescriptive, and they wanted information about self care and prevention.[4]

Although neither Tang nor Coulter studied health information in the context of the Internet, it seems likely that the desire for tailored, timely, balanced, and credible information would apply to information on the Web as well. Ferguson believes that consumers who seek health information on the Internet want tailored information and interactive experiences. They turn to bulletin board and chat room environments to seek both practical advice and emotional support from other patients and families. After communicating with many of what he calls "online self-helpers," Ferguson believes that what online consumers find most valuable, in order of importance, are

1. Answers to questions—questions posted in their own language with a tailored response that addresses the concerns as they have been expressed.
2. Answers to other self-helper's questions, including the FAQ's (frequently asked questions) that are maintained by many support groups.
3. The results of searches for information about health concerns.
4. Prewritten "canned" text providing a general overview, often directly copied from a patient education handout (also described as "shovelware").[21]

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. The 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 one uses to retrieve and sort information. Useful information is often difficult to locate,[22, 23] and information retrieved may be of dubious quality[24-27] or even lead to harm.[28]

The primary strategies that are available to a consumer for finding health information on the Web include

1. Going to a specific website that either the consumer knows about, or someone else has recommended. This is the most efficient method if the site has the desired information.
2. Going to a website with a list of links that has been filtered and indexed manually. The collection has already been filtered by someone's criteria for quality and relevance. If the criteria are similar to the consumer's, and if the information he desires is included in the collection, the website can be a very useful resource. However, coverage of topics in human-filtered collections is necessarily incomplete.
3. Using a search engine that has been designed to cover a range of medical topics. Again, the search engine may only search a limited database of websites.
4. Using a general search engine that searches a very large portion of the entire Web. This method provides access to the greatest amount of potentially useful information, but is also likely to return the largest amount of useless information.

Although finding useful medical information on the Web can be difficult, improved search technology may make the task somewhat easier. 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,[29] and by intervening in query formation, using knowledge of the medical domain to expand the user's query.[30] A small preliminary analysis (by this author) performed at the beginning of this project suggested that top-ranked pages from general search engines are quite likely to be related to the topic of a medical query. Unfortunately, topical relevance does not guarantee usefulness. Many of the webpages retrieved during the preliminary analysis consisted of bulletin board or newsgroup postings, personal home pages with anecdotal information, or lists of links to other webpages. This is consistent with studies that reported the low percentage of webpages retrieved by general search engines that were likely to be useful to a health care professional[23] or to an orthopedic patient.[22] Furthermore, improving the precision or recall of a search does nothing to ensure the quality or credibility of the information retrieved.

Although difficult in any medium, evaluating the credibility of consumer health information on the Web is particularly challenging. The barriers to publishing information on the Web are minimal. There are no standards or requirements for peer review, for providing or referencing substantiating evidence, or even for revealing authorship of Web documents. 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. For example, Impicciatore and colleagues studied parent-oriented webpages relating to home management of children with fever. They found that only a few of the

41 websites that were examined provided complete and accurate information.[26] An accompanying editorial by Wyatt stressed the need for evaluating the quality and impact of clinical websites and discussed some of the issues and challenges that will be encountered by investigators who undertake such studies.[31] Silberg and colleagues called attention to the variable quality of health information on the Internet and suggested a core set of standards of accountability for publishing on the Web: authorship (provide authors' names, affiliations, and relevant credentials), attribution (references and sources for all content; copyright information when relevant), disclosure (website ownership and any sponsorship, advertising, underwriting, commercial funding, and potential conflicts of interest), and currency (dates that content was posted and updated).[32] Kiley cited two examples of potentially dangerous misinformation being propagated on the Internet and emphasized the need for critical appraisal of information that is encountered on the Web.[33]

Several approaches to the problem of finding high-quality health information on the Web are being explored. One approach is self-regulation. The Health on the Net Foundation in Geneva has developed a set of principles called the Net Code of Conduct.[34] Websites can display the HONcode logo to indicate 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 paraphrased summary of the Net Code of Conduct is displayed in Appendix A.

A second approach is to provide consumers with a rating tool, such as a checklist, to evaluate websites. For example, the Health Information Technology 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.[35] A potential limitation of this approach is its dependence on consumer willingness to use a rating form. The proposed criteria are summarized in Appendix B.

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 to know, the criteria used by the reviewers. Jadad and Gagliardi reviewed rating instruments that have been used to rate websites that provide health information. Rating criteria were available for only 14 of the 47 rating instruments that they identified, and of those 14 instruments, only two provided information on how the criteria had been generated. They concluded that existing tools are incompletely developed, and noted the challenges of evaluating both the content of websites and the criteria suggested by Wyatt related to structure, function and likely impact. In addition, they state that it is unclear whether rating instruments "should exist in the first place, whether they measure what they claim to measure, or whether they lead to more good than harm."[36] 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 webpages that a search may return is to use human reviewers to sift health information and to create lists of websites that have been deemed useful or credible or both. The output of this approach may range from a single list of webpages to a large collection of lists that can be searched or browsed by topic. Large collections are most commonly maintained by libraries and educational institutions. Examples of this approach include New York Online Access to Health (NOAH),[37] and NetWellness.[38] Similar sites developed for health professionals, such as Cliniweb[39] and Medical Matrix[40] 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 (data about data, or in this case, about documents) 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. Managing a decentralized rating system and controlling the quality of volunteer rating efforts could be a daunting task. Eysenbach and Diepgen

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.[41]

Previously published efforts to automatically assess webpages have focused on techniques to help the user limit exploration of webpages to those most likely to be relevant to the individual. For example, the NEC Research Institute has developed a metasearch engine (a search engine that combines the results of several search engines) that downloads and analyzes each document and displays the results showing the query terms in the context of the sentences in which they appear.[42] Other investigators have used statistical clustering techniques to subdivide Web documents into groups of similar documents based on the frequency with which words appear in those documents. The user can then choose to view documents in the clusters that appear most likely to be interesting.[43] Pazzani and Billsus used machine learning techniques and Bayesian classification to develop user profiles and to suggest Web documents likely to be of interest based on similarity to the user's profile.[44] Collaborative filtering uses the behavior of other users, either by active voting or by using automated analysis of website popularity, to predict what information might be of interest to the user.[45] A similar approach is that taken by Chakrabarti and colleagues in the design of a system that they call an Automatic Resource Compiler. Hyperlink analysis is used to identify *authority* pages that contain a lot of information about a topic, and *hub* pages that contain a large number of links to pages containing information about a topic. They compare the

hyperlink analysis they use to rank documents to citation analysis in bibliometrics.[46] Extensive use of hyperlink analysis to rank documents is also incorporated into the Google search engine developed at Stanford. Google calculates a page rank for each document using the number of other webpages with links that point to the document, the number of other links on each pointing webpage, and the page rank of each pointing webpage. The number and importance of the pointing pages comprise a measure of the document's citation importance. This page rank is combined with a modified form of word-statistical matching to prioritize search results.[47] None of these automatic approaches to analyzing webpages has been developed specifically for the medical domain, and none is explicitly designed to identify documents based on the quality of the information the documents contain.

Specific Aims

It appears that a software tool that automatically assesses the quality of consumer health webpages would be very useful. Although any such system is unlikely to be perfect, even an imperfect initial filtering would reduce the number of webpages to be critically examined, either by the consumer or by a human reviewer. The specific aim of this project is to create a prototype that automatically analyzes webpages and to perform a preliminary analysis of an algorithm that scores a collection of webpages pertaining to a consumer health query. The implementation chosen is a Web-based software tool that accepts a consumer health query, submits the query to general search engines, intercepts and analyzes the results, and returns a newly ordered set of links to the user. The goal is

to produce an improved retrieval set in response to consumer health queries by analyzing and reordering the output from Internet search engines.

For the purposes of this project, an improved retrieval set is defined as an ordered list of links in which links to higher-quality (or more desirable) webpages are ranked higher, and links to lower-quality (or less desirable) webpages are ranked lower, than in the comparison set. Since search engines commonly return links in groups of ten, the first ten links in an improved retrieval set should contain a larger proportion of links to higher-quality webpages than the comparison set. For this project, higher-quality webpages are defined as pages that

1. Contain a significant amount of information that pertains (is topically relevant) to the user's query. Judgments about quantity of information consider both completeness of the coverage of the topic of the user's query, and whether there is an appropriate level of granularity, or detail. Although relevance is clearly an attribute of webpages that is important to users, the software tool does not assess the relevance of webpages to the user's query. The general search engines that are queried select webpages for topical relevance, and any further filtering for relevance is beyond the intended scope of this project. Nevertheless, it is impossible to ignore the question of relevance when assessing the potential usefulness of a particular webpage to a user. Therefore, when assessing the overall performance of this software tool, higher-quality webpages must contain topically relevant information.
2. Contain credible information. The credibility of information may be difficult to define, but as it is used here, the term encompasses the concepts of accuracy, validity,

consistency with appropriate scientific evidence, acceptance by respected medical authorities, and adherence to published medical guidelines. This criterion is concerned with the quality of the information itself, as opposed to the quality of the webpage or the way that the information is presented.

3. Are free from bias or conflict of interest, or if bias or conflict of interest exists, it is explicitly disclosed. Possible sources of bias include commercial interests, political motivations, and intellectual beliefs. Commercial interests may be manifest either as an attempt to sell a product via the Web or as an attempt to promote the merits or image of a particular product, therapeutic approach, or practitioner in hopes of increasing future sales or use of services.
4. Contain information that is written at an appropriate level of sophistication for a reader without medical training. Highly specialized medical information may be timely and credible but incomprehensible without a detailed knowledge of human anatomy or pathophysiology. Such information may be useless to most consumers.
5. May contain one or more links to high-quality webpages that pertain to the query. Webpages that contain high-quality information (information that is useful, credible, unbiased, and comprehensible to a consumer) are defined as higher quality than those that contain only links to high-quality information, but webpages that contain links to high-quality information are defined as higher quality than webpages that neither contain high-quality information nor contain links to high-quality information.

The scope of the project is limited to aiding consumers who seek information from the World Wide Web. It is therefore limited to the subset of consumers who have access to

the Internet and who know how to use the Internet, at least at some basic level. Although it is clear that there is a demand for interactive experiences on the Web, the focus of this project is to provide information content. No attempt is made to connect users with interactive sources, such as chat rooms, bulletin boards, or mailing lists.

METHODS

This project consisted of the development of a prototype system, including preliminary bench testing of that system. The prototype consists of two parts:

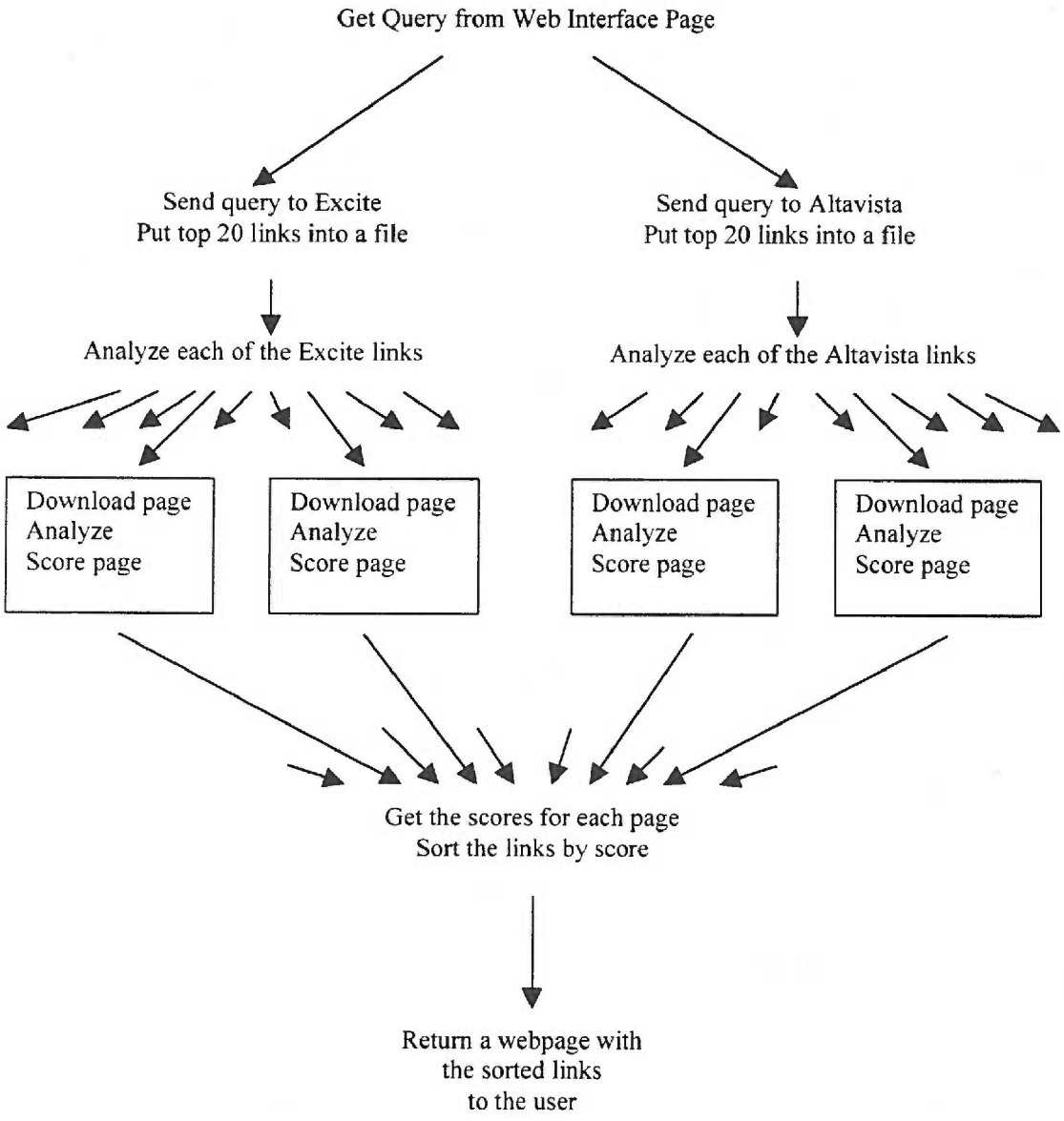
1. A shell that accepts user queries, forwards the queries to search engines, and intercepts and reorders search engine results.
2. An algorithm to analyze and score webpages.

The purpose of bench testing was to assess the usefulness of various characteristics of webpages as discriminators between higher-quality and lower-quality webpages and to assess the performance of the scoring algorithm as a whole.

Description of Prototype

The logical organization of the prototype system is shown in Figure 3. The user enters a query on a Web-based form that is displayed when the URL for the software tool is entered into a standard Web browser. The query is transmitted using the Common Gateway Interface (CGI) to the main Perl program that currently resides on a Sun Ultrasparc 140 computer that acts as a Web server. The main program passes the query through a subroutine to remove common words before forwarding it to two general search engines. The stop list of common words (a list of words, such as *a*, *and*, and *the* that are removed from queries to diminish the number of nonrelevant documents retrieved) was adapted from a list that is used by a publicly available, experimental search engine, SMART. Excite and AltaVista™ were chosen as the two search engines

Figure 1: Logical Organization of the Program



because they are popular and provide wide coverage of the Web. No single search engine covers the entire Web, but the combination of two or more search engines significantly increases coverage.[48] Additional search engines could be added in the future, but were not necessary as part of this proof-of-concept investigation.

The 20 top-ranked links are retrieved from each search engine. The number 20 was chosen arbitrarily to provide enough links to make re-ranking 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. Once the main program receives the top 20 ranked links from each search engine, it calls a second Perl program, the analyzer, to download and score the associated webpages. The analyzer places the contents of the page associated with each link into a temporary file. The file contains the HTML code needed to produce the webpage that will be displayed if the link is followed, plus any meta elements that are present. (Meta elements are HTML elements that contain information about the document that is not displayed by the Web browser, such as the document's author, its publication date, or a description of its contents.) In order to diminish the user's wait for search results, all the webpages associated with the links from Excite and AltaVista™ are requested and downloaded simultaneously and scored independently by separate computer processes. The overall time for the analyzer to complete its task is determined by the time required to download and analyze the slowest webpage.

The analyzer uses a series of algorithms to detect various components in the file for each page. It detects some of the characteristics included in the HON[34] and HITI[35] criteria. It also detects some characteristics that, although neither desirable nor undesirable themselves, serve as heuristic proxies for positive or negative attributes. The analyzer's scoring algorithm awards a variable number of points for each characteristic that is identified. In this first implementation of the scoring algorithm, the number of points awarded for each characteristic was chosen empirically. The analyzer was designed to allow scoring to be adjusted easily so that the performance of the tool can be finely tuned. The total number of points each webpage receives is the score that is used to rank the webpages. When all the webpages have been scored, the main program calls a third Perl program, the sorter, to sort the links according to score. When the links have been sorted, a new webpage that contains the ordered list of links is sent from the main program to the user's browser for display. Examples of the input form and the results page are shown in Appendix C.

In creating this prototype, flexibility and ease of development were emphasized over efficiency. The program takes about 1.5 to 2 minutes to complete from the time the user enters a query until display of the ordered results. The latency of downloading webpages from servers all over the world is responsible for much of the slowness of the application, but more efficient programming could undoubtedly improve the speed of the program.

Automatic Analysis and Scoring of Webpages

The analyzer begins by analyzing the Uniform Resource Locator (URL), which is the Web address of the webpage, such as *http://www.ohsu.edu*. First, the program identifies either the domain name of the network host (such as *www.ohsu.edu*) or the Internet Protocol (IP) address of the webpage (such as 137.53.1.40). Either the domain name or the IP address is found in the URL between the double forward slashes and the first single forward slash, or if there is no forward slash, then the end of the URL. The domain name of a network host consists of a sequence of domain labels separated by "." The rightmost domain label is the top-level domain that consists of either a country code that identifies the country of origin, or one of a few codes that identify the intended function of domain names that contain that label. For example, *.com* is intended for use by commercial users, and *.org* for use by non-profit organizations. The analyzer searches the domain name for four specific domain labels: *.edu*, *.gov*, *.org*, and *.com*. For the remainder of this thesis, these domain labels are referred to as URL extensions. Next, the analyzer compares the domain name or IP address from the webpage's URL to a list of domain names and IP addresses, such as *www.nlm.nih.gov* and *www.americanheart.org*, that were hypothesized by this author to be credible sources of information.

The analyzer then reads the file that contains the HTML code for the webpage line by line. It identifies any meta tags, it searches each line for certain patterns, such as *<a href= "mailto:">*, and it searches each line to see if it contains words or phrases from three files created by the author. The first file contains *red flag* words and phrases, such as *amazing*, *breakthrough*, *miracle*, and *revolutionary*. The second file contains *retail*

words associated with sales transactions such as *MasterCard*, *Visa*, *money order*, and *special price*. The third file contains words associated with certain forms of alternative medicine, such as *tonics*, *antioxidants*, and *magnet therapy*. The analyzer counts the total number of occurrences of words or phrases from each list.

In addition, the analyzer counts the total number of text words, the total number of hyperlinks, the total number of characters on the page, and the total number of characters in the text. It also calculates the ratio of the number of text words to the number of hyperlinks and the ratio of the number of text characters to the total number of nonwhitespace characters (i.e. all characters except those for spacing between words or sentences, and those indicating formatting such as tabs and carriage returns). The text on each page was considered to be all characters that appeared between, rather than within, HTML tags. (HTML tags are the markup indicators that determine how a Web browser displays the webpage, such as <TITLE> and </TITLE>.) Several steps comprise the identification of text words. First, the simple HTML tags
 and <P> are eliminated. The symbols "<" and ">" that are immediately followed by one or more digits are considered likely to be *less than* and *greater than* symbols, not HTML tags, and are eliminated. The program then keeps track of the number of ">" and "<" symbols as they appear on each line. When the sum of ">" and "<" symbols is odd, the characters are part of the HTML markup code that formats the page. When the sum of the ">" and "<" symbols is even, the subsequent characters constitute the text that appears between HTML tags rather than within them. As discussed below (page 36), this does not appear to be the optimum way to identify the text words in the page.

After reviewing published suggestions for evaluating health information on the Web, five criteria were chosen for automatic analysis: usefulness, credibility, absence of bias, amount of content, and currency. These criteria were chosen because they are both desirable and potentially amenable to automatic evaluation. Specific characteristics were then chosen that were hypothesized to be indicators or proxies for each of the five criteria. The analyzer awards points to each webpage based on the presence or absence of these indicators of quality. In the case of characteristics thought to be indicators that the page may be of low quality, all pages that *do not* contain the indicators are given points. Pages with the highest point total are those deemed most desirable. The point values that were assigned were chosen as empirical guesses at relative discriminatory power; they can be adjusted as more experience with automatic and manual scoring of Web pages is accumulated. The approaches to identifying the individual characteristics are discussed below by category. Table 1 shows the current point values that the analyzer awards for each characteristic.

1. *Likely usefulness*

The general search engines that are queried provide a screening for topical relevance. Useful characteristics were defined as those characteristics that might predict whether a webpage is likely to contain useful information, regardless of the specific topic. It was hypothesized that webpages such as bulletin board postings and webpages dominated by advertising and formatting code would be unlikely to convey useful health information. Therefore, webpages that contained the phrase *in reply to* or *posted by* were thought

Table 1: Initial assignment of point values for various characteristics to discriminate desirable pages from those that are less desirable:

Characteristic	Initial point values assigned
Usefulness Score Components	
Bulletin Board	Yes + 0 No + 30
Frames	Yes + 0 No + 5
Embedded app.	Yes + 0 No + 5
Applet	Yes + 0 No + 5
Formatting to text char ratio	more than .5 +15; .25 - .5 +10; .25 or less + 0
Credibility Score Components	
Number of *red flag* words	0 + 20; 1 - 2 +10; 3 or more + 0
Presence of author metatag	Yes + 5 No + 0
Email address for feedback	Yes + 5 No + 0
Cookie	Yes + 0 No + 5
HONcode	Yes + 30 No + 0
Explicit disclaimer on page	Yes + 5 No + 0
Number of exclamation marks	0 - 2 + 40; 3 - 5 + 10; 6 or more + 0
URL extension (number)	
.edu or .gov	+ 15
.org	+ 8
.com	+ 0
other	+ 5
Tilde in URL	Yes + 0 No + 5
From list of credible domains	Yes + 30 No + 0
Bias Score Components	
Number of *retail* words	0 - 1 + 20; 2 - 4 + 7; 5 or more + 0
Presence of 1-800 number	Yes + 0 No + 20
Number of *alternative* words	4 or less + 20; 5 - 9 + 10; 10 or more + 0
Content Score Components	
Number of text words	500 or more + 15; 100 - 499 + 7; less than 100 + 0
Number of hyperlinks	9 or less + 10; 10 - 19 + 3; 20 or more + 0
Ratio of text words to hyperlinks	> 25 + 30; 15 - 25 + 10; 15 or less + 0
Currency Score Components	
Metatag for publication date	1997 or later + 2; Before 1997 + 0; No + 2
Explicit last update or modification	'97 +2; After '97 +4; Before '97 +0; No +2

likely to be bulletin board or newsgroup postings and therefore are awarded 0 points; all others are awarded 30 points. Similarly, because the HTML tags *<embed src=>* and *<applet code=* are often used to display advertising or special effects such as sound, webpages that contained these tags are given 0 points and all other webpages are given 5 points. The ratio of total text characters to total nonwhitespace characters was selected as an indicator of webpages that are dominated by formatting code. If the text character to total nonwhitespace character ratio is greater than 0.5, the webpage receives 15 points. If the ratio is between 0.25 and 0.5, the page receives 10 points. Webpages with a ratio less than 0.25 receive 0 points.

2. *Likely credibility*

This category is composed of postulated indicators of whether a webpage is likely to be a credible source of information. For example, points are awarded to webpages with domain names that match the list of presumed-credible domain names and to webpages with *.edu*, *.gov* or *.org* URL extensions. Webpages with *.com* in the domain name were hypothesized to be less likely to contain credible information than other webpages, so webpages with *.com* in the URL receive no points for URL extension. Webpages that do not have *.com*, *.edu*, *.gov*, or *.org* URL extensions (usually having either *.net* or a country code domain label) were postulated to have an intermediate likelihood of containing credible information and thus receive an intermediate number of points. The program also awards points to webpages that contain few or none of the words and phrases in the *red flag* list. Personal home pages were postulated to be likely to contain personal opinion and to be less likely to contain credible information than other webpages.

Because a tilde in the URL often indicates that the webpage is a personal home page, pages without a tilde in the URL receive a small number of points. Similarly, it was postulated that webpages that set cookies on the user's browser might be more likely to exist to promote commercial gain than to convey credible information, and therefore webpages that do not set a cookie on the user's browser are awarded a small number of points. Webpages with large numbers of exclamation marks were also postulated to be less likely to convey credible information, therefore webpages are awarded points according to how many exclamation marks appear on the page. The most points are awarded to pages with no more than two exclamation marks, and no points are awarded to those with six or more. Webpages that display the HONcode, that display an explicit disclaimer, that display an email address for comments or feedback, and that contain a meta element indicating the author of the page were thought more likely to contain credible information than webpages that do not have these characteristics. The presence of each of these characteristics results in the awarding of additional points.

3. *Likely bias*

This part of the scoring algorithm consists of three components: the number of words and phrases that match the *retail* list, which that may indicate whether a commercial bias is present, the presence or absence of either a 1-800 or 1-888 telephone number on the page, and the number of words and phrases from the list of terms associated with alternative medicine. Webpages that contain many occurrences of these words and phrases may be more likely to be pages selling a large variety of nutritional supplements or promoting an indiscriminate variety of untested therapies. Less than the maximum number of points is

awarded only if five or more of these terms are detected in the meta elements and text of a webpage.

4. *Content*

An ideal filtering program would measure the quantity of information and determine whether the content is accurate, useful, and relevant to a particular user, but such analysis is not practical. Instead, the analyzer determines how much text is displayed (the number of text words), how many hyperlinks are present, and the ratio of text to hyperlinks. The supposition is that webpages with more text are more likely to be valuable to the consumer than webpages with minimal text, or webpages that consist primarily of links to other webpages.

5. *Currency*

The algorithm searches both meta elements and text for evidence indicating when the webpage was last updated. If a date is found, it determines how recently the update was done. The most points are awarded to pages for which it can be determined that the last update was 1997 or later. No currency points are given to pages that have a last update or modification prior to 1997. Because many pages do not display explicit date information, it was assumed that many pages without date information have actually been created or updated within the last three years. Therefore, if no date information is detected, a small point value is awarded to the page. The point value is less than the point value that is awarded to pages known to have been created or updated in 1997 or later, but greater

than the zero points awarded to pages known to have been created or updated prior to 1997.

Bench Testing

Stead and colleagues discussed the design of applied medical informatics projects and proposed that projects be subdivided into sequential stages, each of which builds upon successful completion of preceding stages. They also proposed that evaluation be divided into five levels, and that the level of evaluation should be appropriate to the stage of system development.[49] In the Stead classification, this project is best described as being in (a) the stages of component development, and (b) combination of components into a system. According to the Stead model, bench testing is the appropriate level of evaluation for this stage of development, and may be followed by either further development, with integration of the system into the environment in which in which it will be used, or by field testing. When discussing bench testing, Stead and associates wrote "A computer scientist may carry out bench testing using his or her own programming skills and a single subject's ideas and experiences regarding medical care. Rapid prototyping is a form of bench testing. Paper cases or scenarios may serve as the basis for initial testing, although use of actual case materials is preferred."[49] Bench testing of software might be compared to testing a drug in a laboratory prior to testing it on human subjects. The performance of the software is tested against a preliminary, substitute standard to assess the feasibility of the approach prior to more rigorous and expensive development and testing, either using human subjects or occurring in a production environment that risks financial or other consequences of failure.

This project is best regarded as the development of a prototype, and the first step in an iterative development process, rather than as production of a finished product. Whereas a completed project might be subjected to an evaluation, much like a traditional scientific experiment that produces definable results that can be subjected to statistical analysis, a different paradigm is more appropriate for the early development phase of this project. Bench testing is needed not so much to evaluate results as to gauge the feasibility of the approach and guide further development. Therefore, the goal of bench testing in this project was to assess the usefulness of the individual criteria that were used, and to assess how well the current scoring algorithm works, using manually assigned scores as the comparison.

Two test queries were used in the bench testing processes: a pilot test query and a test query. All of the pages returned by the Excite and AltaVista™ search engines in response to each query were assigned manual scores to serve as the standard of comparison. Both queries were consumer health queries not used during development and were based on search topics that have been submitted to the librarians at the Oregon Health Sciences University Consumer Health Reference Center. The librarians have created a database of approximately 400 queries. Only the topics of the requests for information have been recorded; the actual wording by the consumer, and the context in which the information needs arose, are not available. In order to create a slightly more realistic search environment, a fictitious scenario was created for each query. Medical textbooks[50, 51] and the National Guideline Clearinghouse[52] were consulted to create

summaries of the information that was to be considered desirable. The scenarios and summaries are shown in Appendix D.

Each webpage that was returned in response to the queries was assigned a manual score prior to ranking by the software tool. A modified version of the search tool submitted the query to Excite and AltaVista™, parsed the list of links returned by each search engine, and randomized the list prior to displaying the links to the user. The result was a list of 40 hyperlinks (any duplicate links would have been removed), showing the URL and the brief description of each page returned by the search engine. The format of the descriptions was altered slightly so that the human scorer was unable to determine whether a particular link was returned by Excite or by AltaVista™.

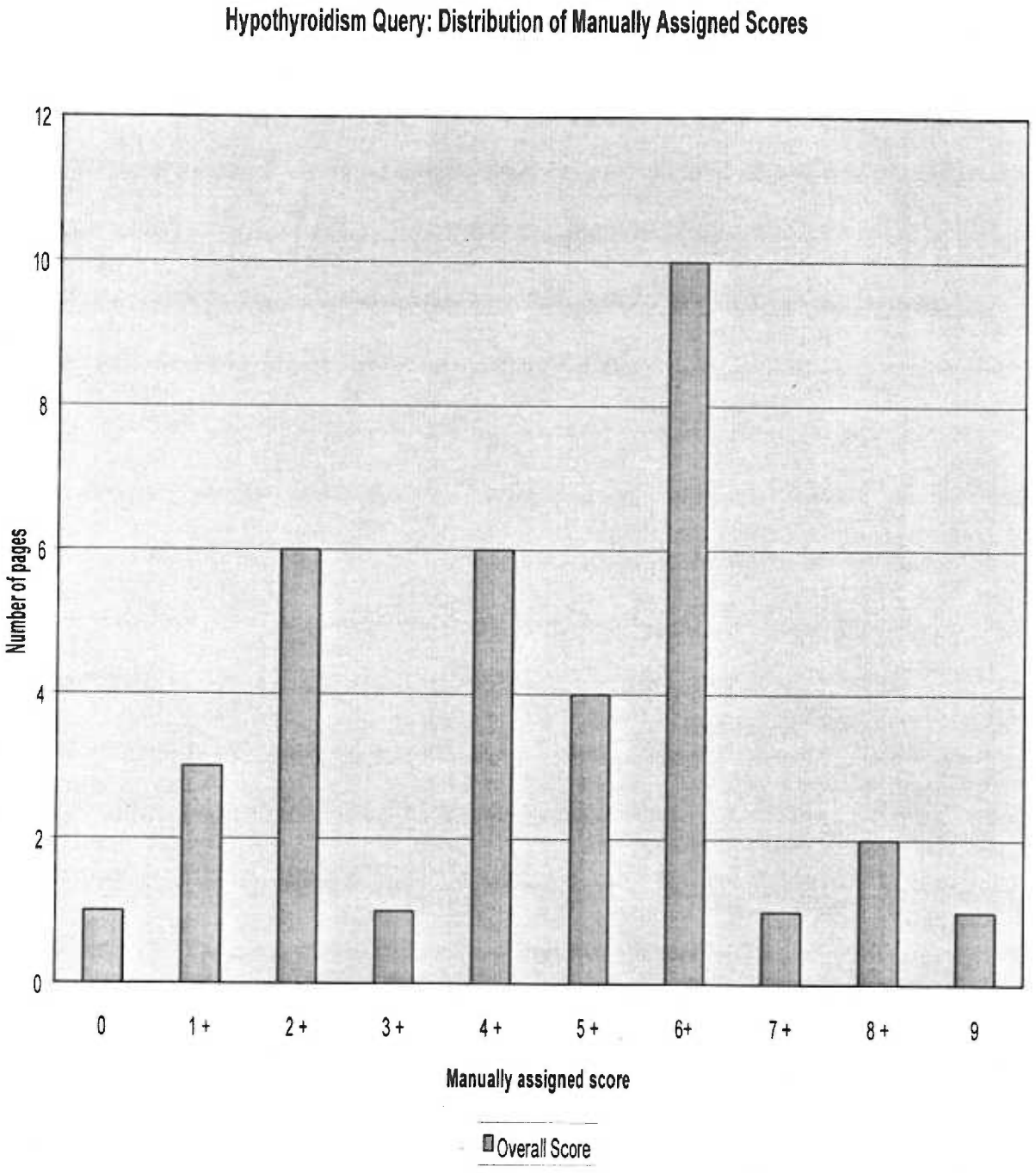
At this time, no validated tool exists to score the quality of webpages related to consumer health. The development of a validated scoring method is an important task but is beyond the scope of this thesis project. Webpages were therefore assigned a score from 1 to 9 based on the scorer's opinion about the quality of a webpage. For this preliminary stage of testing, only the investigator's opinion was used to derive a score, and thus the score cannot be considered a gold standard. The score incorporated the criteria that have been suggested by HITI[35] and HON,[34] and reflects the definition of higher-quality webpages that was given in the Introduction (pages 13–14). For the pilot query, webpages were scored on a scale of 1 to 5. The quality of the webpages retrieved was extremely low; only two pages out of 40 were given 4 or 5 points. The low number of higher-quality webpages combined with the relatively small range of scoring options

prompted the use of a score incorporating a larger range of possible values for the test query. For both queries, a score of 0 points was given to broken links because they provided no useful information. The distribution of the manually assigned scores for the test query is shown in Figure 2.

Both queries returned webpages that were completely unrelated to the subject of the query. The test query *What causes hypothyroidism?* resulted in the retrieval of five webpages about hypothyroidism in dogs, cats and horses. Because assessing webpages for topical relevance is not one of the functions of the software tool, the three options for scoring these pages were to

1. Award the pages 0 points because they contain no information that is useful to the hypothetical user, much like a broken link. However, if the webpages contain high-quality information about veterinary problems, they might exhibit the same characteristics as high-quality webpages about human health. Scoring the pages with 0 points will lead to bias against the qualities exhibited on those pages, and might result in skewing the algorithm to negatively weight characteristics that, in fact, predict high-quality webpages.
2. Score the webpages as if they are relevant to the user's query. However, it may not be possible to accurately determine the quality of the information on the webpages. In the example of this particular test query, a person qualified to assess human health information might be unable to reasonably assess the quality of veterinary information. Furthermore, if high-quality webpages unrelated to the query exhibit different characteristics from high-quality webpages related to the query, which might

Figure 2: Distribution of Manually Assigned Scores



be especially likely with webpages unrelated to human health, the analysis could again be skewed.

3. Omit the webpages from the analysis for all three retrieval techniques.

The third option, omitting unrelated webpages from the analysis, was the most appropriate choice given that the purpose of this bench test was to guide further development. A different option might be appropriate in the context of a formal evaluation. Because the tool does not filter webpages based on topical relevance, an evaluation of the performance of the scoring algorithm might score all webpages for which a quality score could reasonably be assigned by the manual scorer, regardless of the topic of the webpage. On the other hand, an evaluation from the user perspective might assign irrelevant pages 0 points if the pages were judged to be completely useless to a person posing the query.

In order to use a larger number of webpages to assess the usefulness of the individual criteria, data from the test query and the pilot query were combined. More desirable webpages were defined as those that were scored as 6 or higher (6–9), or in the case of the pilot test query, 4 or higher (4–5). Less desirable webpages were defined as those that were scored as 4 or lower (0–4), or in the case of the pilot query, 2 or lower (0–2). The occurrences (percentage or average number, depending on what was appropriate) of each characteristic were tabulated for the more desirable webpages and for the less desirable webpages.

Only the test query was used to test the overall performance of the current scoring algorithm. The scoring algorithm was assessed by examining its ability to identify both higher-quality webpages and lower-quality webpages. The number of webpages ranked in the top 10 by the tool that were of higher-quality (defined as pages with manual scores of 6 points or higher) was compared to the number of higher-quality webpages returned in the top 10 by Excite and AltaVista™. Similarly, the number of webpages ranked in the top 10 by the tool that were of lower-quality (defined as pages with manual scores of 4 points or less) was compared to the number of lower-quality webpages returned in the top 10 by Excite and AltaVista™. The number 10 was chosen because search engines commonly return 10 links on the first results page, and therefore links to high-quality webpages are most valuable if they appear on that first page of links.

RESULTS

The pilot consumer health query, *How can I avoid getting osteoporosis?*, resulted in 39 unique links being returned. There was no overlap between the links returned by the two search engines. The tool successfully eliminated the one broken link that was returned as the eighth link by Altavista™. Of the 39 webpages, eight were determined by the investigator to be topically irrelevant and were eliminated from consideration.

The test query, *What causes hypothyroidism?*, returned 40 unique links, again with no overlap between the two search engines. Of the 40 webpages, five were determined to be topically irrelevant and were eliminated from consideration. All five were veterinary pages that contained information about hypothyroidism in dogs, cats or horses. One additional page was manually judged to be a broken link because it contained the message, in bold, "This page has moved, Click here to view." Clicking on the hyperlink led to a menu of hyperlinks, none of which led to the desired page. On the original page with the error message, a table and some text were present but barely visible due to a very small font size and a very light gray color. When the page contents were downloaded into a file, however, the HTTP code was that of a normal transmission and the contents were correctly retrieved. The algorithm therefore did not categorize it as a broken link, but scored it according to the contents present in the file.

Table 2 shows the frequency of various characteristics in the more-desirable versus the less-desirable webpages. Although the number of webpages is small, 16 more-desirable webpages and 39 less-desirable webpages, it appears that the presence of a meta tag for authorship and the presence of an email address on the webpage are not likely to be useful indicators of high-quality webpages. In this small sample, they occur more frequently in the less-desirable webpages, and could eventually be found to be negative indicators of quality. The same is true, to a lesser extent, of an explicit notation of when the webpage was last updated. The number of text words, and possibly the ratio of text words to hyperlinks, does appear to be larger in the more-desirable pages. Many of the characteristics did not occur frequently enough to draw conclusions about their usefulness.

The numbers of higher-quality and lower-quality webpages returned in the top 10 spots by each method are shown in Tables 3 and 4. There is a suggestion that the tool returned more higher-quality and fewer lower-quality webpages, but the results were not statistically significant. A larger sample of queries and webpages will be needed to draw firm conclusions about the usefulness of the current algorithm.

Eight of 39 webpages (20.5 %) returned in response to the pilot query were judged to be irrelevant, and five of 40 webpages (12.5%) returned in response to the test query were judged irrelevant. If these webpages were manually scored either as zero or as if they were relevant to the queries and were included in the assessment of the overall

performance of the scoring algorithm, the numbers in Tables 3 and 4 would change, but the patterns would remain the same.

A few deficiencies were discovered in the program during bench testing. First, the one cookie event was not detected by the analyzer. If cookies are determined to be a useful part of the algorithm, a more robust means of cookie detection will be necessary. The program currently searches for messages in the HTML code indicating that a cookie will be set. Examination of the HTTP header is likely to be a more reliable technique and will be implemented in the future. Evaluation of the test and pilot test files revealed that for most webpages the text word count was accurately determined, but for six webpages sections of text were missed, resulting in a falsely low word count. This occurred on pages with complex HTML formatting and appears to have occurred due to confusion arising from the attempt to distinguish between digits occurring immediately following HTML tags (for example, HTML code to put the number 8 in italics 8 or to subscript the numeral in T_4 or T_3) and the greater or less than symbol immediately preceding a digit. More extensive analysis and debugging will be required to make the text word counting method more robust. A Perl module exists in the LWP library to parse HTML code. It was adopted in the early development stages of this project to identify text, but it did not seem to do an adequate job of identifying the text from some complex webpages. The technique that was employed was developed to try to avoid complete parsing of the HTML file, but it also was not entirely successful. A more rigorous attempt to parse HTML code may be necessary to accurately identify all the text that appears on a webpage.

Table 2: Preliminary analysis of the usefulness of various characteristics to discriminate desirable pages from those that are less desirable. The frequency of each characteristic is shown for the 16 webpages that were judged to be more desirable (of higher quality) and for the 39 webpages that were judged to be less desirable (of lower quality)

Characteristic	More Desirable (n=16)	Less Desirable (n=39)
Bulletin board posting	0 (0 %)	1 (2.6 %)
Frameset	0 (0 %)	0 (0 %)
Embedded application	0 (0 %)	1 (2.6 %)
Applet	0 (0 %)	1 (2.6 %)
Presence of author metatag	1 (6.2 %)	5 (12.8 %)
Email address for feedback	3 (18.8 %)	13 (33.3 %)
Cookie	0 (0 %)	1 (2.6 %)
HONcode	0 (0 %)	0 (0 %)
Explicit disclaimer on page	4 (25 %)	5 (12.8 %)
Tilde in URL	0 (0 %)	0 (0 %)
From list of credible sites	3 (18.8 %)	0 (0 %)
Presence of 1-800 number	0 (0 %)	0 (0 %)
Metatag for publication date	0 (0 %)	0 (0 %)
Explicit last update of mod.	3 (18.8 %)	5 (12.8 %)
Dates	'95, '98, '99	'98 = 4, '99 =1
No. of *red flag* words	0	0
No. of *retail* words	0	0.03
No. of *alternative* words	0.12	0.90
No. of exclamation marks	0.56	1.77
No. of text words	554.8	310.1
No. of hyperlinks	16.5	10.9
Ratio of text words to links	55.4	43.9
Ratio of format. to text char	0.26	0.37
URL extension		
.edu	2 (12.5 %)	0 (0 %)
.org	5 (31.2 %)	7 (18.0 %)
.com or .co*	9 (56.2 %)	28 (71.8 %)
.net	0 (0 %)	1 (2.6 %)
other	0 (0 %)	3 (7.7 %)
Usefulness score (average)	49.7	50.9
Credibility score (average)	80.1	66.4
Bias score (average)	60.0	59.2
Content score (average)	31.1	33.3
Currency score (average)	4.1	4.3
Total score (average)	227.3	214.5

*.*co* is sometimes found as a secondary domain name when the top-level domain name is a country code; it appears to be used for commercial domain names.

Table 3: Selection of higher quality webpages: Shows the number of higher-quality webpages (manual scores of six or higher on a scale of 1 to 9) that were ranked in the top 10 by the software tool, by Altavista, or by Excite.

Search Tool	Number of Webpages
Software Tool	5
AltaVista™	4
Excite	2

Table 4: Filtering out lower quality webpages: Shows the number of lower-quality webpages (manual scores of four or lower on a scale of 1 to 9) that were ranked in the top 10 by the software tool, by Altavista, or by Excite.

Search Tool	Number of Webpages
Software Tool	3
AltaVista™	4
Excite	8

DISCUSSION

This project demonstrates that automatic analysis of webpages to identify indicators of the quality of information that the pages contain is technically feasible and potentially useful. The software tool described successfully retrieves a set of links from a general search engine, analyzes and scores the associated webpages, reorders the links, and returns the new list of links to the user. Preliminary bench testing offers some insight about the role some characteristics may play in discriminating between higher-quality and lower-quality webpages and suggests that the current scoring algorithm may be somewhat successful in reordering the links according to a given set of criteria. It is not known how much additional improvement in ranking webpages could be achieved by using a different or more extensive set of criteria, or by altering the points assigned for each criterion.

Lessons Learned From Current Implementation

Analysis of the webpage scores from the pilot test and test queries led to five observations. First, some characteristics occur so infrequently that it will not be possible to assess their value as discriminators until a larger collection of manually scored pages is available. Examples of this category include: bulletin board postings, embedded programs (such as sound), Java applets, presence of *red flag* words, presence of the HONcode, presence of retail indicators, and the use of cookies. Much of the difference between the scores derived from just a few characteristics, such as the presence or absence of disclaimers, meta tags for authorship, email addresses for feedback, the URL

extension, the ratio of formatting to text characters, the text word count, the link count, and the ratio of text words to links.

Second, some characteristics that are desirable and potentially useful to the reader may not be useful discriminators between higher-quality and lower-quality webpages, and may not be correlated with the underlying characteristics of the information. When developing criteria it is important to distinguish between the desirability of a characteristic and its usefulness as a discriminator between more-desirable and less-desirable webpages. For example, from this very preliminary analysis it does not appear that the presence of an email address, or the presence of an explicit date of last modification, will predict the high quality of a webpage. There is even some suggestion that these characteristics may be more likely to occur on lower-quality webpages (see Table 2). Furthermore, 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 that are no longer supported by scientific evidence, is of low quality. Several published lists of criteria state that the date that a page was last updated should be displayed.[32, 34, 35] However, displaying the date on which a webpage was last updated may or may not be correlated with the currency of information on the page. Similarly, the presence of information about authorship on a webpage may or may not be correlated with the credibility of information on that page. The desirability of publication and modification dates and of authorship information must be evaluated separately from the desirability of currency and credibility, just as the desirability of publication and modification dates and of authorship information must be

evaluated separately from the ability of these attributes to predict the quality of a webpage.

Third, the influence of the word count, text to link ratio, and formatting to text character ratio had mixed effects. In some cases, these characteristics successfully helped to identify low-quality webpages. In a few notable cases, however, webpages with relatively little content but with links to pages with very good or excellent content received low point totals and low rankings. A good example of this phenomenon is http://www.osteo.org, the index page of the National Institutes of Health Osteoporosis and Related Bone Diseases~National Resource Center. The page itself has no information about osteoporosis, but a single click leads to a wealth of excellent information. The very low word count and low text to link ratio resulted in the page being ranked only 21st out of 31 pages.

Fourth, it is not clear that the characteristics that were analyzed truly reflect the categories in which they were placed. Although the five general criteria that were chosen for analysis, usefulness, credibility, absence of bias, amount of content, and currency, may be useful constructs, it may not be feasible to identify characteristics that are specific indicators of each of these criteria. A more successful approach may be to empirically identify individual characteristics that help discriminate between higher-quality and lower-quality webpages.

Fifth, a substantial proportion of consumer health information on the Web is found on websites with multiple pages, which makes analysis more complicated. At least two approaches to this issue are possible: a program to filter webpages can preferentially rank only pages with a significant amount of information on the page itself, or it can analyze pages as part of complete websites and confer high rankings on pages that come from, or have links to, high-quality websites. Although it is preferable to be linked immediately to a content-intensive webpage, and it is much easier to analyze a single page, analyzing isolated pages may not be the most useful approach. A single mouse click allows a user to view other pages on a website, and most Web users are probably accustomed to exploring websites. When manually ranking the webpages, it became apparent that only a few webpages contained a substantial quantity of useful information. Some webpages had links to other pages with useful information, and it seemed most reasonable to give such webpages fairly high scores. During a real search, such pages would probably be discovered and found to be useful.

Two additional observations derived from analyzing the output of Excite and AltaVista™. First, neither the pilot test query nor the test query yielded any overlap in the top 20 links returned from each of the two search engines. The lower ranking search results from each source (below the top 20) were not examined, either to determine how much overlap could be identified in a larger set of results, or to determine how many of the top 20 links from one search engine could be found among the lower-ranked links of the other. It is not known whether the pages returned by one search engine had not been indexed by the other, or whether the lack of overlap in this study merely reflects the

different ranking algorithms of the two search engines. A 1998 study of Internet search engines, not restricted to the medical domain, showed that none of the six largest Internet search engines cover more than about 34% of the indexable Web,[48] so it would not be surprising to find a substantial lack of overlap even if the entire retrieval sets were examined.

Second, the low quality of the webpages associated with the top 20 links from each search engine is striking. For the pilot test query, only two of 32 topically relevant webpages received scores of 4 or 5 points on a scale of 1 to 5. For the test query, only four of 35 topically relevant webpages received scores of 7, 8, or 9 points on a scale of 1 to 9. Although these scores merely represent the opinion of a single investigator, they are in accord with the study by Hersh and colleagues that examined the results of Web searches for professionally oriented material to answer 50 clinical questions,[23] and the study by Rose and colleagues that examined Web pages returned in response to consumer queries about knee injuries.[22] Expansion of the system to include additional search engines, or to evaluate more webpages from each search engine, may result in more high-quality pages being returned, even if the proportion of high-quality pages remains low.

Limitations

At least five problems will be encountered by any scheme to automatically analyze webpages. The first problem is the constantly changing and evolving nature of the Web and Web search engines. Less than three months after the original Perl code was written to intercept and parse the webpages containing retrieval sets from Excite and

AltaVista™, the program became inoperable due to changes in the formatting of the webpages returned by Excite. Modifications of the program corrected the problem, but the incident highlights the maintenance that will be required for any program that analyzes webpages.

A second problem, which will be encountered when evaluating the performance of an automatic webpage analyzer, is the changing retrieval sets that are returned by the search engines. The pilot test and test queries were repeated 15 and 4 days after the original bench testing was done. After 15 days, 10 of the 39 pages that were returned in response to the pilot test query had not been present in the original retrieval set. Nine of the new links were from Excite, and one was from Altavista™. The test query was repeated after just 4 days. Twenty-two of the 40 links were new: Fifteen links from Excite and 7 from Altavista™ were new additions to the retrieval set. As occurred when the queries were originally run, there was no overlap between the set of links returned by Excite and the set of links returned by AltaVista™. This means that when changes are made to the algorithms, re-entering the same query cannot be guaranteed to return the same set of webpages for analysis. Development of a large test collection of webpages that have been manually scored would expedite the process of development and incremental testing.

Third, analysis of the semantic content of webpages to filter out pages that offer misleading information is likely to remain elusive. The use of proxies for quality may be a useful substitute. More extensive analysis of potential proxies is necessary to determine

which ones will ultimately prove to be useful. Other proxies may be discovered that are more useful than those currently being used. Preliminary experience during development suggested that it will be easier to identify indicators of lower-quality webpages than to identify indicators of higher-quality webpages. In addition, the criteria are likely to be a dynamic rather than a static set. If a system using automatic analysis ever enters general use, it is likely that some webpage publishers will learn to manipulate the system, just as some webpage producers now manipulate the use of meta elements and other methods to increase the likelihood of receiving high rankings from search engines when various queries are entered.[47]

Fourth, this software tool does automatic filtering without respect to the context of an individual search by a user with unique needs and motivations. The user's query may not reveal his specific needs or related interests. For example, the tool might assign a low rating to a link to a bookseller's webpage. The link may not fit the profile of high-quality webpages, but it might be useful to a user who would like to read a book about the subject of his query as well as find useful webpages. Addition of a module that allows entry of user preferences to modify the scoring algorithm for individual queries could partially overcome this limitation.

Fifth, any attempt to rate or filter health information will be complicated by the lack of a gold standard for distinguishing high-quality information from low-quality information. The criteria that were used during development of this software tool were adapted from those proposed by respected authors, but these criteria have not been formally validated.

A related limitation is that the published criteria have not been documented to reflect the qualities that consumers desire when seeking health information on the Web. Some information is available regarding consumer preferences with respect to health information in general[4, 20] but it would be useful to know what criteria consumers apply when evaluating health information on the Web.

Related Research

As noted in the Introduction (pages 11–12), none of the automatic approaches to analyzing webpages that have been described were developed specifically for the medical domain, and none were explicitly designed to identify documents based on the quality of the information the documents contain. Nevertheless, it may be possible to adapt word-statistical and machine learning techniques to develop a profile of high quality consumer health webpages that could be used to classify Web documents as desirable or undesirable. Collaborative filtering also might be useful to identify higher-quality webpages, although popularity may not correlate with quality. An organized attempt to recruit medical professionals to participate in collaborative filtering, as suggested by Eysenbach and Diepgen,[41] might prove fruitful, although it is likely to be a complicated project that will require constant maintenance. An automated approach using hyperlink citation analysis, also suggested by Eysenbach and Diepgen in the same paper, might be easier to implement and maintain, although its usefulness for quality screening has not been demonstrated. Some form of hyperlink citation analysis might be a useful extension to the current software tool.

Future Work

Future work could involve both further development of the tool in at least five areas and more extensive evaluation of the tool. The first area of further development is to refine the scoring algorithm so that the points assigned to each characteristic are more likely to discriminate between desirable and undesirable pages. New characteristics may be added, and any characteristics that do not contribute to the discrimination should be eliminated. A second area of further development is to analyze a larger number of pages. This could be done both by adding additional search engines to provide input, and by retrieving a larger number of links from each search engine. A third area is to improve the efficiency and speed of the program. In order to conduct meaningful user evaluations, the program will need to run in a reasonably short period of time.

A fourth area is to extend analysis to the website level and develop the ability to predict the usefulness of the links presented on a webpage. In order to optimize analysis and scoring of consumer health resources on the Web, it appears that a sophisticated approach to analyzing webpages as part of larger websites, and as part of a web of external links, will be necessary. During development a preliminary approach to this problem was explored. An unused module of the program explores the links found on each page and calculates an average link score that can be used as a component of the total score. In addition, the module presents the user with a list of webpages that have at least six hyperlinks, in order of average link score. Several deficiencies in the module that were identified during development precluded inclusion of the module in the program. Additionally, it appeared that searching one level of links might not be an adequate

approach to link analysis. It may be more useful to analyze the internal links (links to webpages within the same website) separately in order to arrive at an assessment at the website level as well as the webpage level. Webpages that lead to high-quality information elsewhere on a website could be given appropriately higher scores than are currently assigned when only a single webpage is analyzed. This might also allow the program to recognize websites that maintain disclaimer messages or feedback addresses on pages separate from the original webpage being examined. A separate analysis of external links would allow recognition of webpages that lead to useful information on other websites. Webpages that consist of lists of links that lead only to more lists of links should be recognized and assigned appropriately low scores.

Website analysis, and analysis of external links from the original page, is likely to be complicated by encountering links to webpages that are irrelevant to the query. Some assessment of topical relevance may need to be incorporated into the process of link analysis, especially if more than one level of links is explored. A simple preliminary approach would be to search for the presence of query words on each linked page. The mere presence of query words, or even words related to the same concept, does not ensure that a webpage will be relevant to the user's needs, but it might provide a useful screen to filter out many irrelevant webpages. Hyperlink analysis that estimates the relative importance or popularity of a page using bibliometric techniques is a related component that could be added in the future as well. As discussed above, hyperlink citation analysis has been suggested as a means of rating medical webpages, and has been implemented on nonmedical research systems.

A fifth area of possible future development is the addition of more functionality. Users could be allowed to customize the weighting of various characteristics in order to rank preferentially particular kinds of pages. For example, webpages could be classified by reading level or by level of medical sophistication so that users could choose the type of page they wish to have ranked most highly.

At the current stage of development, it was appropriate to apply bench-level testing using the developer's judgements with respect to criteria and manual scoring of webpages. With further development the tool will need to undergo more extensive evaluation. It would be very helpful to have a validated instrument for manually rating webpages. An ideal instrument would incorporate validated criteria for evaluating Web pages, and would have acceptable inter-rater and intra-rater reliability. This would allow comparison of the rankings assigned by the tool to rankings assigned by one or more medically-trained persons (who were not involved in the development of this software) using a validated manual scoring instrument.

User evaluation is another important step. Consumers without medical training, but with some experience searching the Web, could be recruited to compare the tool to standard search engines as a means of finding health information pertaining to a particular topic. A variety of measurements could be used to compare search tools. To assess whether this tool helps consumers more efficiently find high-quality health information, the searchers could be given a list of health-related questions to answer using either the software tool

or a general search engine for each question. The number of questions that each subject successfully answered within a fixed period of time would be measured. To eliminate guessing or bias due to prior knowledge of the answers, the subject must find at least one supporting document that contains the correct answer to each question. Pilot studies would be necessary to determine the appropriate number of questions and to define useful questions that are not ambiguous and that have answers that can be found on the Web. Consumer feedback regarding usability issues could also be elicited.

SUMMARY AND CONCLUSIONS

As it continues to grow, the Web will make an increasing amount of health information available to consumers. Increased access to health information has the potential for enhancing knowledge, informing decisions, promoting healthy behaviors, and providing information to groups of consumers that are not well-served by current mechanisms. The Web also has the potential for propagating misinformation and even causing harm.

A major obstacle to obtaining high-quality health information from the Web is the tremendous volume of information that is available. Commercial search engines rank webpages based on the probability that each page is related to the topic of a query, but search engines provide no guidance as to the credibility of the information on those webpages. In the absence of controls over the information that is being published on the Web, a variety of approaches have been suggested to help guide consumers toward higher-quality sources of information on the Web. Automatic analysis and re-ranking of webpages may be one way to help consumers navigate through the maze of webpages that comprise the Web. This project demonstrates that, given a set of criteria that may indicate the quality of consumer health information, it is possible to automatically rank webpages in order of likely quality. Although it will not eliminate the need for critical evaluation of information, automatic filtering of webpages may be a useful adjunct to searching and may expedite a search for high-quality health information on the Web.

REFERENCES

1. Laine C, Davidoff F, Lewis CE, et al. Important Elements of Outpatient Care: A Comparison of Patients' and Physicians' Opinions. Ann Intern Med 1996; 125: 640-5.
2. Waitzkin H. Clinical Implications of Social Scientific Research. JAMA 1984; 252:2441-46.
3. Coulter A. Evidence Based Patient Information. BMJ 1998; 317:225-6.
4. Coulter A, Entwistle V, Gilbert D. Sharing Decisions with Patients: Is the Information Good Enough? BMJ 1999; 318:318-22.
5. Benbassat J, Pilpel D, Tidhar M. Patients' Preferences for Participation in Clinical Decision Making: A Review of Published Surveys. Behav Med 1998; 24:81-8.
6. Beecham L. Patient Education Campaign Launched. BMJ 1996; 312:531.
7. Kassirer JP. The Next Transformation in the Delivery of Health Care. N Engl J Med 1995; 332:52-4.
8. Brody DS, Miller SM, Lerman CE, Smith DG, Caputo GC. Patient Perception of Involvement in Medical Care. J Gen Intern Med 1989; 4:506-11.
9. Reichard P. To Be a Teacher, a Tutor and a Friend: The Physician's Role According to the Stockholm Diabetes Intervention Study. Patient Educ Couns 1996; 29:231-5.
10. Greenfield S, Kaplan S, Ware JE. Expanding Patient Involvement in Care. Ann Intern Med 1985; 102:520-8.
11. Herbert CP. Giving Information: Usually Necessary, but Often Insufficient to Achieve Behaviour Change. Patient Educ Couns 1996; 29:229-30.
12. Brown MS. Consumer Health & Medical Information on the Internet: Supply and Demand. Summarized at: http://etrg.findsvp.com/health/mktginfo.html (last visited 11/28/98) 1997.
13. Nammacher MA, Schmitt K. Consumer Use of the Internet for Health Information: a Population Survey, AMIA '98 Annual Symposium, Orlando, FL, 1998.
14. Lindberg DAB, Humphreys BL. Medicine and Health on the Internet. The Good, the Bad, and the Ugly. JAMA 1998; 280:1303-4.
15. Coiera E. Information Epidemics, Economics, and Immunity on the Internet. BMJ 1998; 317:1469-70.
16. McLellan F. "Like Hunger, Like Thirst": Patients, Journals, and the Internet. Lancet 1998; 352:SII39-43.
17. Pemberton PJ, Goldblatt J. The Internet and the Changing Roles of Doctors, Patients and Families. Med J Aus 1998; 169:594-5.
18. Bader SA, Braude RM. "Patient Informatics": Creating New Partnerships in Medical Decision Making. Acad Med 1998; 73:408-11.
19. Coiera E. The Internet's Challenge to Health Care Provision. BMJ 1996; 312:3-4.
20. Tang PC, Newcomb C, Gorden S, Kreider N. Meeting the Information Needs of Patients: Results from a Patient Focus Group. Proc AMIA Annual Fall Symp. 1997:672-6.

21. Ferguson T. Health Online and the Empowered Medical Consumer. J Qual Improv 1997; 23:251-7.
22. Rose S, Bruce J, Maffulli N. Accessing the Internet for Patient Information About Orthopedics. JAMA 1998; 280:1309.
23. Hersh WR, Gorman PN, Sacherek LS. Applicability and Quality of Information for Answering Clinical Questions on the Web. JAMA 1998; 280:1307-1308.
24. Bower H. Internet Sees Growth of Unverified Health Claims. BMJ 1996; 313:381.
25. Brooks A. Miracle Cures Advertised on the Internet. BMJ 1998; 317:769.
26. 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.
27. Troger U, Meyer FP. Validity of Advertising Claims for Multivitamin Preparation Vitacor 20/90 on the Internet. BMJ 1998; 317:1069-71.
28. Weisbord SD, Soule JB, Kimmel PL. Poison on Line - Acute Renal Failure Caused by Oil of Wormwood Purchased through the Internet. N Engl J Med 1997; 337:825-7.
29. Tay J, Ke S, Lun K. MediAgent: A WWW-based Scalable And Self-Learning Medical Search Engine, AMIA '98 Annual Symposium, Orlando, FL, 1998.
30. 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.
31. Wyatt J. Commentary: Measuring quality and impact of the world wide web. BMJ 1997; 314:1879-81.
32. Silberg WM, Lundberg GD, Musacchio RA. Assessing, Controlling, and Assuring the Quality of Medical Information on the Internet. JAMA 1997; 277:1244-1245.
33. Kiley R. Quality of Medical Information on the Internet. J Royal Soc of Med 1998; 91:369-70.
34. 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.
35. 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.
36. Jadad AR, Gagliardi A. Rating Health Information on the Internet. JAMA 1998; 279:611-4.
37. NOAH: New York Online Access to Health. http://www.noah.cuny.edu (last visited 2/25/99) 1999.
38. Netwellness. http://www.netwellness.org (last visited 3/02/99) 1999.
39. Cliniweb. http://www.ohsu.edu/cliniweb (last visited 3/02/99) 1999.
40. Medical Matrix. http://www.medmatrix.org (last updated 1/18/99; last visited 3/02/99) 1999.
41. Eysenbach G, Diepgen TL. Towards Quality Management of Medical Information on the Internet: Evaluation, Labelling, and Filtering of Information. BMJ 1998; 317:1496-1500.

42. Lawrence S, Giles CL. Context and Page Analysis for Improved Web Search. IEEE Internet Computing 1998; July-August:38-46.
43. Zamir O, Etzioni O. Web Document Clustering: A Feasibility Demonstration, 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998.
44. Pazzani M, Billsus D. Learning and Revising User Profiles: The Identification of Interesting Web Sites. Machine Learning 1997; 27:313-31.
45. Wiley DL. Beyond Information Retrieval: Ways to Provide Content in Context. http://www.onlineinc.com/database/DB1998/wiley8.html (last visited 3/30/99) 1998.
46. Chakrabarti S, Dom B, Raghavan P, Rajagopalan S, Gibson D, Kleinberg J. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text, 7th International World Wide Web Conference, Brisbane, Australia, April 14-18, 1998; also available at http://decweb.ethz.ch/WWW7/1898/com1898.htm (last visited 3/30/99). Elsevier Science.
47. Brin S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine, 7th International World Wide Web Conference, Brisbane, Australia, April 14-18, 1998. Elsevier Science.
48. Lawrence S, Giles CL. Searching the World Wide Web. Science 1998; 280:98-100.
49. Stead WW, Haynes RB, Fuller S, et al. Designing Medical Informatics Research and Library-Resource Projects to Increase What is Learned. J Am Med Informatics Assoc 1994; 1:28-33.
50. Fauci AS. Harrison's Principles of Internal Medicine. Vol. 2. New York: McGraw Hill, 1998.
51. Hurst WJ. Medicine for the Practicing Physician. Stamford, Connecticut: Appleton & Lange, 1996.
52. National Guideline Clearinghouse. http://www.guideline.gov (last visited 3/10/99) 1999.

Appendix A: The HONcode (Paraphrased for brevity):

1. Any medical/health advice provided and hosted will only be given by medically/health trained and qualified professionals unless clearly noted otherwise.
2. Information is designed to support, not replace, the relationship with the existing physician.
3. Confidentiality of visitors to the website, including identity, is respected.
4. Where appropriate, information will be supported by clear references to source data and, where possible, have HTML links to that data. The date when a clinical page was last modified will be clearly displayed.
5. Any claims relating to benefits/performance of specific treatment, product or service will be supported by appropriate, balanced evidence.
6. Designer will seek to provide information in the clearest possible manner and provide contact addresses for further information or support. Email address of Webmaster will be clearly displayed.
7. Support will be clearly identified, including identities of commercial and noncommercial organizations that have contributed funding, services or material.
8. If advertising is a source of funding it will be clearly stated.

Appendix B: Health Information Technology Institute: Criteria for Assessing the Quality of Health Information on the Internet (Summary)

I. Credibility
   - A. Source
     1. Source (author, institution)
     2. Credentials (qualification, credentials, personal or financial interests)
     3. Conflict of interest
     4. Bias (could be financial or intellectual slant toward a particular theory)
   - B. Context
     Examples given are (1) information provided as part of advertising should be labeled as such and (2) a treatment found effective for one disease should not be discussed as if it treats a wider spectrum of diseases (migraine headache vs. all headaches)
   - C. Currency (up to date with the present state of medical knowledge; should include date of original document and date of posting)
   - D. Relevance/Utility (content should correspond to the information the site purports to provide)
   - E. Editorial review process

II. Content
   - A. Accuracy (validity of information)
   - B. Hierarchy of evidence (should reflect the principles of evidence-based medicine, even if written at a level the general public can understand)
   - C. Original source stated
   - D. Disclaimer (describe limitations, purpose, scope, authority and currency of the information; sources of information and disclosure of any reporting errors; should emphasize that the content is general health information and not medical advice; define the relationship between original website content and links to other sites)
   - E. Omissions noted (state clearly if not all pertinent facts, negative results are included, or if advocating a particular point of view promoting a product or treatment)

III. Disclosure
   - A. Purpose of the site
   - B. Profiling (collection of data about the user)

IV. Links
   - A. Selection
   - B. Architecture (ease of navigation)
   - C. Content
   - D. Back linkages (who else links to the site)

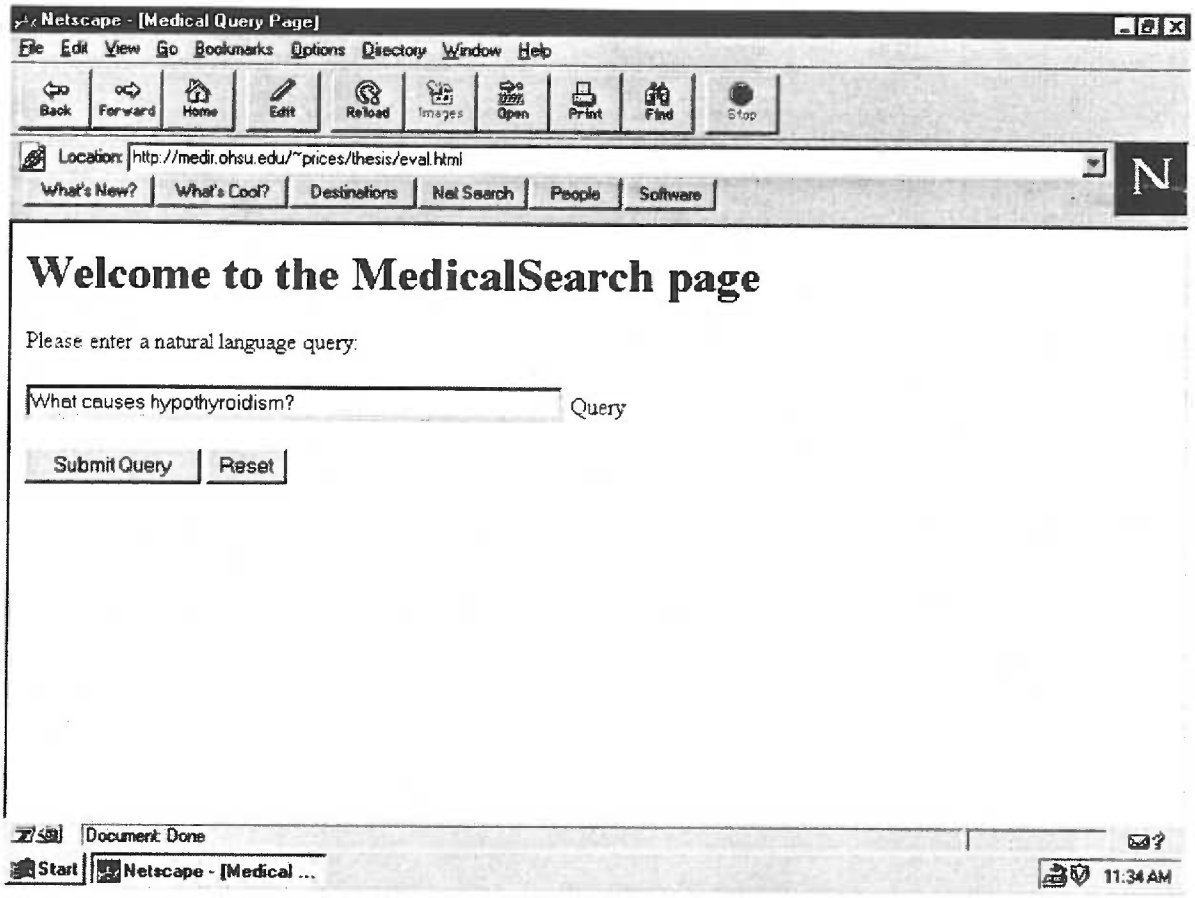
V. Design (layout)
   - A. Accessibility (includes browser technology required to view the site and any option to enable use by hearing or sight-impaired users)
   - B. Logical organization (navigability)
   - C. Internal search engine

VI. Interactivity
   - A. Mechanism for feedback (link to send comments)
   - B. Chat rooms and bulletin boards
   - C. Tailoring

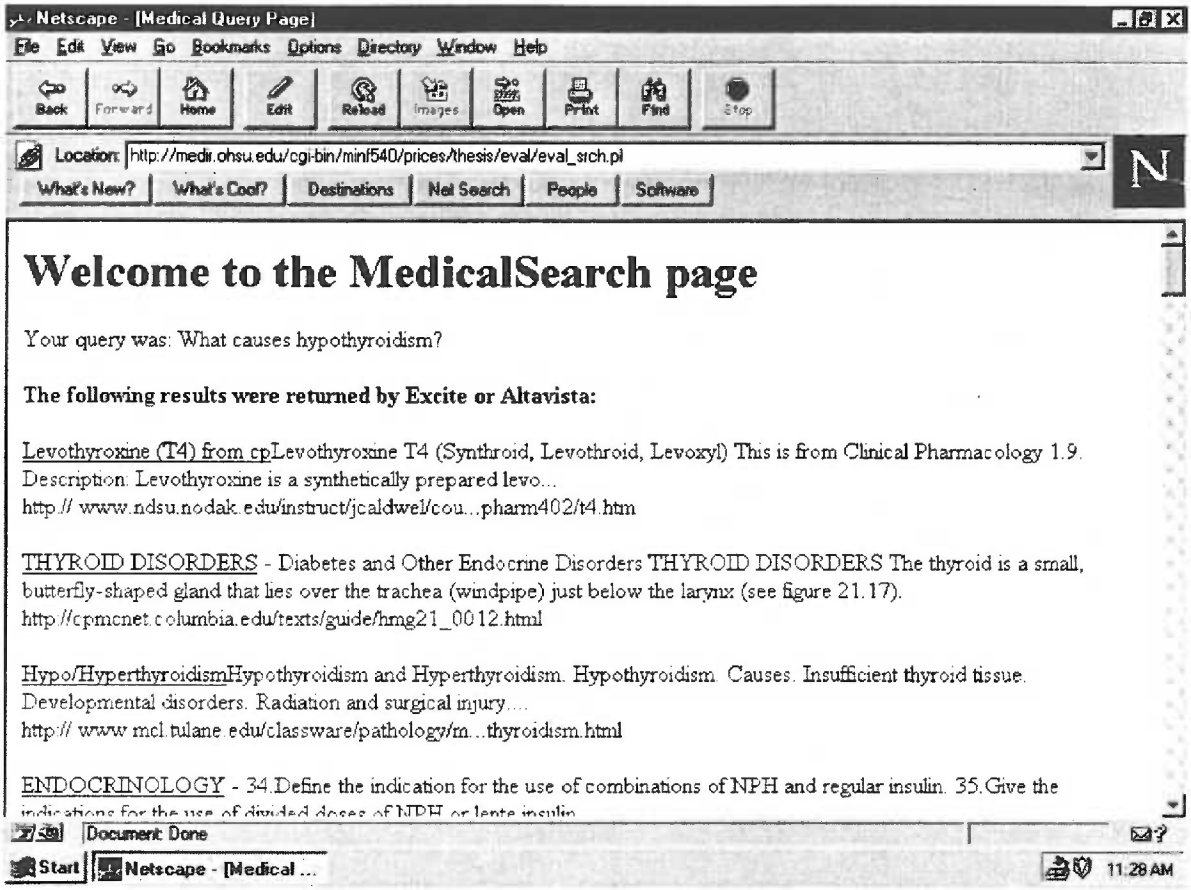
VII. Caveats (user should be aware that personal information may be collected, stored and shared without user knowledge or assent, information presented may be inaccurate)

VIII. Alerts (user should be wary of sound-alike names and prestigious sounding names which may be misleading; be skeptical of claims for "miracles" and "amazing results" or of attacks on widely accepted medical principles or public health policies)

Appendix C: Example of Input Form



Appendix C: Example of Results Page



Appendix D: Scenario and summary for Query 1

Query: *"How can I avoid getting osteoporosis?"*

Setting: *55 year old generally healthy female who enters the query "Prevention of osteoporosis." Her 63 year old friend recently fell and broke her hip. The user was told it was caused by osteoporosis. The user wants to know what osteoporosis is, whether she might be at risk for it too, how she can prevent it, and what problems it can cause.*

Information that would be useful and appropriate includes:

1. Concept of bone formation and resorption
2. Osteoporosis makes bones "brittle" and more fragile, therefore more likely to break
3. Role of menopause (estrogen deficiency)
4. Other risk factors: low body weight, short stature, smoking, low calcium intake, no weight-bearing exercise, alcohol, smoking
5. General recommendations:*
   A. Diet with adequate calcium
   B. Good general nutrition with adequate intake of vitamin D
   C. Regular weight-bearing exercise
   D. Avoid use of tobacco
   E. Consider HRT
   F. Consider calcium and vitamin D supplementation
   G. If on thyroid replacement, adjust dose if needed to avoid over-replacement
6. Bone mineral density testing can be used for diagnosis if a clinical decision to intervene with an agent will be directly influenced by the outcome of the test.*
7. Possible treatments:*
   A. HRT
   B. Calcium and Vitamin D supplements
   C. Alendronate sodium
   D. Etidronate (not FDA approved for this purpose, at least when guidelines written 1996; web page updated 12/14/98)
   E. Calcitonin
8. Possible problems caused by osteoporosis: pain, disability, deformity
   A. Hip fractures
   B. Vertebral fractures
   C. Wrist and other fractures
   D. Dowagers hump

---

* AACE clinical practice guidelines for the prevention and treatment of postmenopausal osteoporosis. Developed by the American Association of Clinical Endocrinologists and the American College of Endocrinology

Appendix D: Scenario and Summary for Query 2

Query: *"What causes hypothyroidism?"*

Setting: *42 year old moderately obese female who enters the query "What causes hypothyroidism." She has recently been diagnosed with hypothyroidism. She wants to know what causes hypothyroidism and if her family is also at risk for getting it.*

Information that would be useful and appropriate includes:[50, 51]

1. Primary Causes (thyroid):
   A. Congenital developmental defect
   B. Heritable biosynthetic defects
   C. Maternally transmitted (iodides, antithyroid agents)
   D. Iodine deficiency
   E. Primary idiopathic atrophy
   F. Post-ablative(radioiodine, surgery)
   G. Post-radiation
   H. Drug-elicited
   I. Chronic thyroiditis (Hashimoto's thyroiditis)
   J. Excessive iodine intake in susceptible individuals
   K. After treatment with interleukin 2 and lymphokine-activated killer cells
2. Secondary Causes (pituitary):
   A. Pituitary adenoma
   B. Pituitary ischemic necrosis
   C. Hypophysectomy
   D. Craniopharyngioma
   E. Trauma
   F. Hemochromatosis
   G. Hypophysitis
   H. Metastatic cancer
   I. Irradiation
3. Tertiary Causes (hypothalamic):
   A. Hypothalamic adenoma
   B. Congenital defects
   C. Trauma
   D. Infection
   E. Tumor
   F. Infiltration (sarcoidosis)

The following information might also be of interest to this person:
1. Information about symptoms that would suggest that family members be tested
2. Information about treatments
3. Information about the possible consequences of untreated hypothyroidism
4. Information about the possible consequences of excessive thyroid replacement