HelpfulMed: Intelligent Searching for Medical Information over the Internet

Hsinchun Chen, Ann M. Lally, Bin Zhu, and Michael Chau
Department of Management Information Systems, Eller College of Business and Public Administration, The University of Arizona Tucson, AZ 85721. E-mail: {hchen, alally, bzhu, mchau}@eller.arizona.edu

Medical professionals and researchers need information from reputable sources to accomplish their work. Unfortunately, the Web has a large number of documents that are irrelevant to their work, even those documents that purport to be "medically-related." This paper describes an architecture designed to integrate advanced searching and indexing algorithms, an automatic thesaurus, or "concept space," and Kohonen-based Self-Organizing Map (SOM) technologies to provide searchers with fine-grained results. Initial results indicate that these systems provide complementary retrieval functionalities. HelpfulMed not only allows users to search Web pages and other online databases, but also allows them to build searches through the use of an automatic thesaurus and browse a graphical display of medical-related topics. Evaluation results for each of the different components are included. Our spidering algorithm outperformed both breadth-first search and PageRank spiders on a test collection of 100,000 Web pages. The automatically generated thesaurus performed as well as both MeSH and UMLS—systems which require human mediation for currency. Lastly, a variant of the Kohonen SOM was comparable to MeSH terms in perceived cluster precision and significantly better at perceived cluster recall.

1. Introduction

The problems inherent in information retrieval from electronic systems are nothing new in the field of information science. For more than twenty-five years, efforts have been underway to understand and solve these problems, which lie not only with the indexing of documents (Bates, 1986), but with the human propensity for describing the same object in a number of different ways (Furnas et al., 1987).

The Internet and its distributed, unorganized repositories of information have compounded the problem of information retrieval. The process of browsing through over two billion unique Web pages (Lyman & Varian, 2000) looking for relevant information is a huge burden for users. In addition to the billions of Web pages currently available on the Internet there is also an "Invisible Web." Most of the Invisible Web is made up of databases which are inaccessible to search engine spiders that only find the URL of the database but not the information contained within. This adds significantly to the problem of information retrieval on the Web. It also contributes significantly to the phenomenon of information overload, a state in which the searcher no longer is able to process the volume of information retrieved effectively.

In addition to information overload, there is the vocabulary problem with respect to the retrieval of relevant information from systems. Vocabulary problems occur because information retrieval systems rely upon users to input the same words to describe a document that the designer of the system has selected to describe it. A multitude of factors contribute to this problem. Previous research indicates that different indexers who have been trained in the use of an indexing scheme assign different index terms to represent the same document and that the same indexer will often assign different index terms to the same document at different times (Bates, 1986). In addition, in research conducted by Furnas et al. (1987) it was found that when subjects were asked to spontaneously assign words to concepts in five different domains, two subjects chose the same word less the 20% of the time. Given these variables, it is a wonder that we are able to retrieve as much relevant information as we do.

While the problems of information overload and retrieval are prevalent across the many disciplines represented on the Internet, the ability to accurately search for, access, and process information is particularly pressing in the field of medicine. The availability on the Internet of vast distributed repositories of quality medical information, each with its own unique interface, has placed information retrieval at the center of research.

The goal of this paper is to describe an approach to building an architecture for a Web portal that provides information retrieval from reputable sources in the medical domain with a minimum of human mediation. To this end, we developed a knowledge portal specifically for medical information retrieval called *HelpfulMed* (http://ai.bpa.arizona.edu/helpfulmed). In developing this system we combined existing AI Lab techniques for information retrieval using spider technology, noun phrase indexing (Tolle & Chen, 2000), automatic thesaurus generation (Concept Space) (Houston et al., 2000), data visualization (Chen et al., 1998), and a meta search tool designed to search the Invisible Web of databases. These techniques have been significantly modified and enhanced for medical information retrieval.

The rest of the paper is structured as follows: Section 2 surveys the current technologies used for information retrieval on the Internet and reviews three medical information retrieval systems. Section 3 is a discussion of the HelpfulMed system, including architecture and functionalities. Section 4 presents the results of our evaluations. Section 5 presents conclusions and future directions.

2. Research Background

The problem of information retrieval, particularly as it relates to the Internet, has received a great deal of attention over the past few years. Many techniques have been developed to address the growing problem of information retrieval from large, unorganized collections. We outline these areas below.

2.1 Automatic Thesaurus Generation

Much research has been conducted in order to address the problems of keyword-based information retrieval. Most research conducted to date has used either a thesaurus or a vector space representation based on the work of Salton et al. (1975). Thesauri are used primarily to expand users' queries in order to translate them into alternative phrases that match document indexes. There are two types of thesauri: human-generated (such as Library of Congress Subject Headings (LCSH) and Medical Subject Headings (MeSH)) and automatically generated. Automatically generated thesauri contain phrases that appear within documents in a given collection. In an analysis of different query expansion techniques using the MEDLINE test collection, it was found that thesauri are a viable way to retrieve information in the medical domain (Srinivasan, 1996). The large majority of automatically generated thesauri are based on syntactic analysis using statistical co-occurrence of word types on text and vector space representation of documents (Guntzer et al., 1989; Salton, 1989; Crouch, 1990; Chen & Lynch 1992).

The use of automatically generated thesauri has helped alleviate the problem of synonymy by adding more associative phrases to keyword indexes (Chen et al., 1998; Houston et al., 2000). However, a major problem resulting from this approach concerns the introduction of noise into the indexing process since some of the phrases may have meanings that are different from those intended. According to Deerwester and colleagues (1990), this can result in rapid deterioration of precision. Improvement in document recall has been demonstrated when the thesaurus is used in a domain similar to the one in which the thesaurus was originally constructed (Crouch, 1990). However, Cimino et al. (1994) documented that there are problems associated with the automatic translation of medical terms using thesauri.

2.2 Document Analysis and Clustering

Anyone who has used the Internet to find information knows that looking through Web page after Web page can be time consuming and frustrating. In order to address this problem, much research has been focused on developing techniques and tools to analyze, categorize and visualize large collections of Web pages.

The traditional approach to creating classification systems and knowledge sources in library science and classical AI is often considered top-down since knowledge representations and formats are pre-defined by human experts or trained librarians and the process of generating knowledge is structured and well-defined. Researchers in machine learning, statistical analysis, and neural networks have suggested a complementary bottom-up approach to knowledge creation. In a bottom-up approach, based on collections, researchers develop programs that systematically segment and index documents and then identify patterns within those documents. Automatic indexing algorithms have been widely adopted to assist in this approach through their use to extract key concepts from text, and studies have shown the efficacy of this approach compared with human indexing (Salton, 1986). Linguistic techniques such as noun phrasing also have been applied to this problem with some success (Tolle & Chen, 2000).

Once these documents have been indexed they can be organized. Clustering is used to provide a richer representation of documents to a user, by grouping documents together based on their textual similarity. The presentation of these clustered documents provides a context in which to understand the relationships among retrieved documents. Document clustering is based on the Cluster Hypothesis: "closely associated documents tend to be relevant to the same requests" (van Rijsbergen, 1979). Document clustering uses one of two approaches. In the first approach, documents are categorized based on individual document attributes, such as keywords, authors, size, etc. In the second approach, documents are categorized based on inter-document similarities. This approach usually includes some type of machine learning algorithm. For example, the Self-Organizing Map (SOM) approach clusters documents into different categories which are defined during the process, using a neural network algorithm (Kohonen, 1995). Based on this algorithm, the SOM technique automatically clusters documents into different regions based on the similarity of the documents. It produces a map consisting of different regions where each region contains similar documents. This technique is discussed in further detail later.

The idea of clustering documents retrieved from the Web was reported by Mechkour et al. (1998) who applied intelligent filtering to a group of documents and thus present "specialized subsets" of the documents. Through this technique they were able to present an overall "picture" of the document relationships. Nonetheless, the issue on whether clustering techniques are useful is still under constant debate. Hearst and Pedersen (1996) and Zamir and Etzioni (1999) demonstrated that document clustering has the potential to improve performance in document retrieval, while Wu et al. (2001), on the other hand, showed that clustering only works well for some topics but not others and on average no user benefited from clustering in search tasks.

2.3 Web-based Medical Information Retrieval Systems

Numerous sites have been developed to provide access to medical information over the Internet. These services range from consumer health information sites such as MEDLINEPlus to sites which aggregate journals, books, news, clinical symposia and continuing medical education resources such as Medscape (www.medscape.com) and MDConsult (www.mdconsult.com). The National Library of Medicine's Gateway (gateway.nlm.nih.gov/gw/Cmd/) and CliniWeb (www.ohsu.edu/cliniweb/), among others, provide access to Web pages from reputable organizations and institutions. In Gateway, these Web pages are indexed according to the UMLS Metathesaurus, and in the case of CliniWeb, the MeSH tree hierarchy. One common concern of these search engines is that they provide users with only a list of ranked results without further analysis. It would be desirable to perform post-retrieval analysis on the search results for the users. In developing the HelpfulMed system, we aim to provide users with a system which combined different powerful features to form one functional, integrated system, by allowing a user to search Web pages and databases as well as use analysis tools such as automatic thesaurus and document clustering. We also try to fully automate the collection building and indexing tasks in HelpfulMed so as to lower the requirement for human intervention.

3. HelpfulMed System Architecture

In this section we present the architectural design of HelpfulMed (as shown in Fig. 1) and discuss in depth each major component. In order to illustrate the functionalities of the system and how the user interacts with the system, we use one user session as an example throughout the remainder of this discussion; our user is searching for information on lung cancer.

The major functionalities of HelpfulMed are: 1. *Intelligent Knowledge Portal*—the HelpfulMed user interface; 2. *Search Medical Web Pages*, which consists of search servlet and Web page database; 3. *Search Medical Databases*, which consists of the meta search program and online medical databases; 4. *Related Medical Terms* which consists of the concept space and term co-occurrence files; and 5. *Visual Site Browser*, which consists of SOM applet and categorization files.

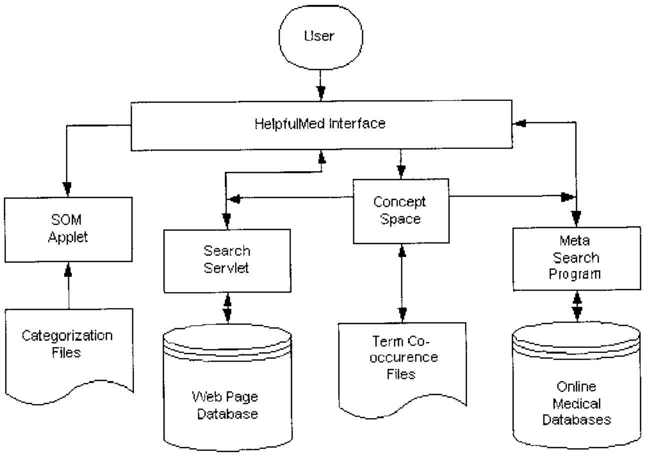


FIG. 1. HelpfulMed system architecture.

3.1 Intelligent Knowledge Portal

A single search interface ties together the Web search, related medical terms, and database search functionalities through the use of an integrated system that allows users to interact with the various technologies offered. This type of "one-stop shopping" system brings together distributed resources a user might need in one place, thus decreasing the time and cognitive analysis required of the user to sift through all of the information sources and learn the individual idiosyncrasies of each system in an attempt to capture all of the information relevant to the user's information need.

From the initial interface, shown in Figure 2, users can begin searching medical Web sites, related medical terms, and medical databases, or they can browse documents by using the visual site browser.

3.2 Searching Medical Web Pages

The HelpfulMed Web page search engine is designed to provide fine-grained medically-related Web pages; Figure 3 is a diagram of the search engine architecture.

In order to build a database of medically-related Web pages, a spider based on the Hopfield Net spreading activation algorithm is sent to collect Web pages every month. In this approach we model the Web as a Hopfield Net, which is a single-layered, weighted neural network (Hopfield, 1982). Nodes are activated in parallel and activation values from different sources are combined for each individual node until the activation scores of nodes on the network reach a stable state (convergence). The spider was designed specifically to retrieve medical Web pages and equipped with a medical vocabulary knowledge base created from the Unified Medical Language System (UMLS) Metathesaurus. An analysis algorithm, which compares the UMLS knowledge base to the text of the Web page and the Web link structure, was developed to allow the spider to assess whether a Web page in question is indeed a medical Web page.

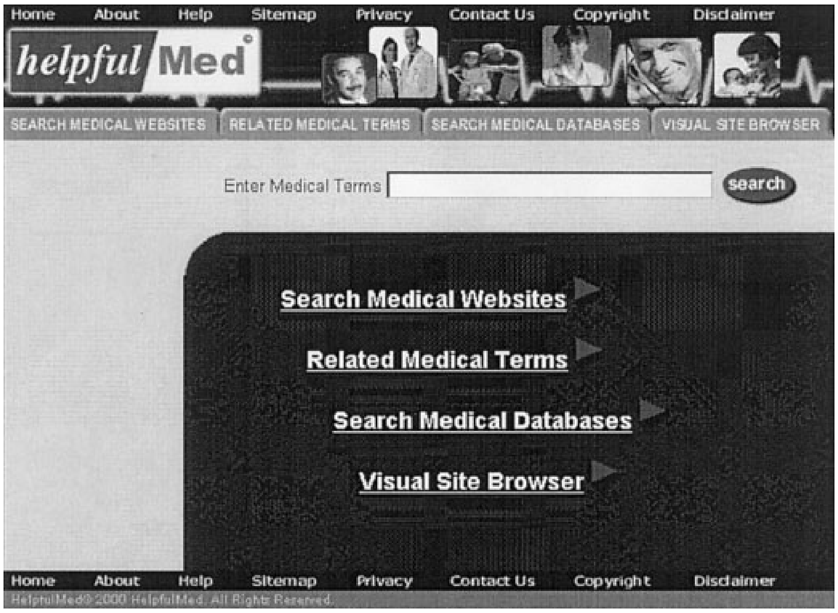


FIG. 2. The HelpfulMed intelligent knowledge portal.

In an attempt to gather a finely tuned set of search results, the spider begins its search with a pre-defined set of 354 URLs from reputable sites as determined by a medical librarian. One should note that, however, residing on reputable sites is not necessarily an indicator of high quality of a Web page (Eysenbach et al., 2002; Fallis & Frické, 2002). Beginning with these medically related URLs, the spider then begins searching the Internet and collecting related Web pages. In order to continue to collect desired Web pages, it is assumed that medical pages included in the list such as the American Medical Association or the Mayo Clinic will more likely point to sites that they consider to be useful. In addition, the pages must contain UMLS phrases consistent with the knowledge base embedded in the spider. In this manner, the spider continually casts a wider and wider net, collecting Web pages. The goal of this process is not to collect and index a large number of Web pages, but to collect and index a fine-grained set of Web pages in the domain of medical information. Once the Web pages have been collected they are stored in the database and post-processed for retrieval and display.

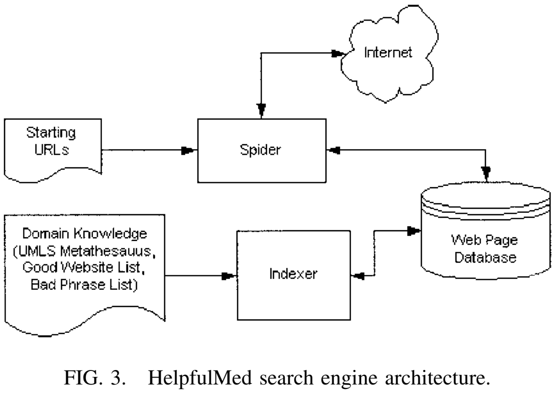
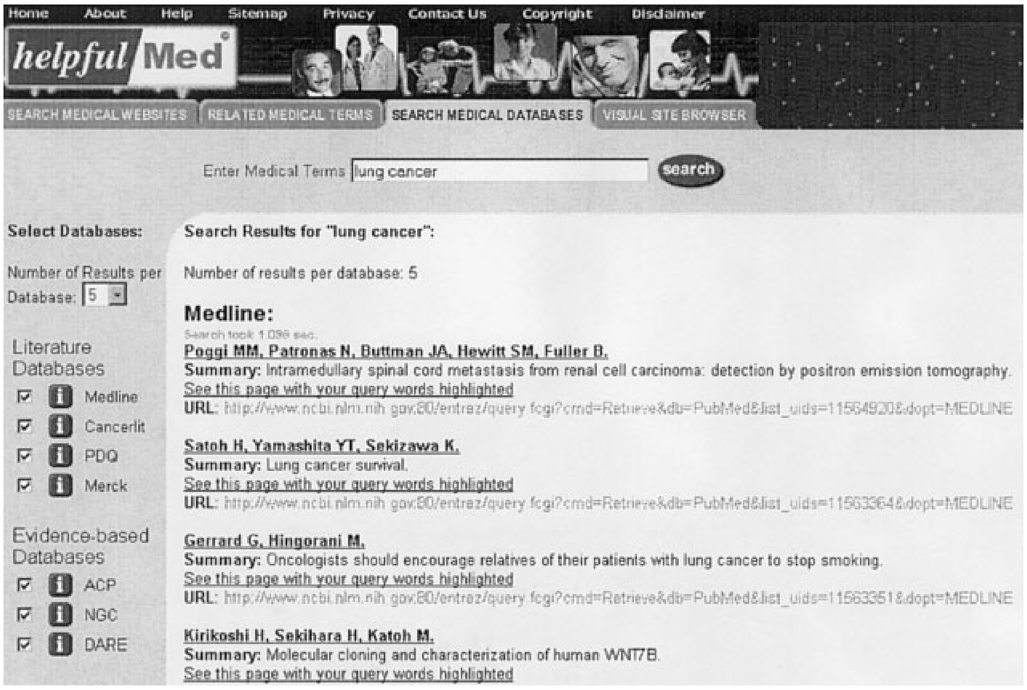

Internet
Starting URLs
Spider
Domain Knowledge (UMLS Metathesauus, Good Website List, Bad Phrase List)
Indexer
Web Page Database

FIG. 3. HelpfulMed search engine architecture.

Post-processing of the documents occurs in four steps. First, the documents are run through a noun phraser developed at the University of Arizona—the Arizona Noun Phraser (AZNP) (Tolle & Chen, 2000). This natural language processing tool is used to extract high quality noun phrases from text, and can improve retrieval precision because it allows multi-word query matching with words present in the document. The AZNP extracts noun phrases from text by first processing the raw text, removing any symbols or punctuation without interfering with the textual content and next assigning parts-of-speech to these words. This implementation of the AZNP incorporates the UMLS Specialist Lexicon in order to correctly identify the parts-of-speech contained within medical text. Studies conducted previously by the lab have confirmed the ability of this lexicon to improve the extraction of medical phrases from text (Tolle and Chen, 2000). Finally, the system converts words and part-of-speech tags into noun phrases. Next, the remaining phrases are run against a bad-phrase lexicon in order to remove irrelevant or nonsensical phrases, such as “Patient with Head.”

The terms entered by the user are compared against the terms previously extracted during the post-processing phase in order to identify the most relevant Web pages. The system displays the Web pages in a ranked order with the most relevant Web pages at the top. Each returned result has a ranking from five (very relevant) to one (marginally relevant); the title of the Web page; a description drawn from the Web page itself with the search terms highlighted; the page URL; and the number of terms matched in the page. Ranking of the results is based on the presence of query keywords in the document and the inlink score. The inlink score is computed based on the number of documents which point to a page, on the theory that the more pages that point to a given page, the greater the acceptance of the



Enter Medical Terms lung cancer

Select Databases:

Number of Results per Database: 5

☑ Medline
☑ Cancerlit
☑ PDQ
☑ Merck
☑ ACP
☑ NGC
☑ DARE

Search Results for "lung cancer":

Number of results per database: 5

Medline:

Search took 1.098 sec.

Poggi MM, Patronas N, Buttman JA, Hewitt SM, Fuller B.
Summary: Intramedullary spinal cord metastasis from renal cell carcinoma: detection by positron emission tomography.
See this page with your query words highlighted
URL: http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=11564920&dopt=MEDLINE

Satoh H, Yamashita YT, Sekizawa K.
Summary: Lung cancer survival.
See this page with your query words highlighted
URL: http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=11563364&dopt=MEDLINE

Gerrard G, Hingorani M.
Summary: Oncologists should encourage relatives of their patients with lung cancer to stop smoking.
See this page with your query words highlighted
URL: http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=11563351&dopt=MEDLINE

Kirikoshi H, Sekihara H, Katoh M.
Summary: Molecular cloning and characterization of human WNT7B.
See this page with your query words highlighted

FIG. 4. Retreived results from online medical databases.

source as reliable. Hopfield Net searching (Hopfield, 1982; Chen & Ng, 1995), phrase indexing, a UMLS knowledge base and inlink ranking combine in our system to collect and present Web pages that are more likely to be medical in nature and finely tuned to the information request.

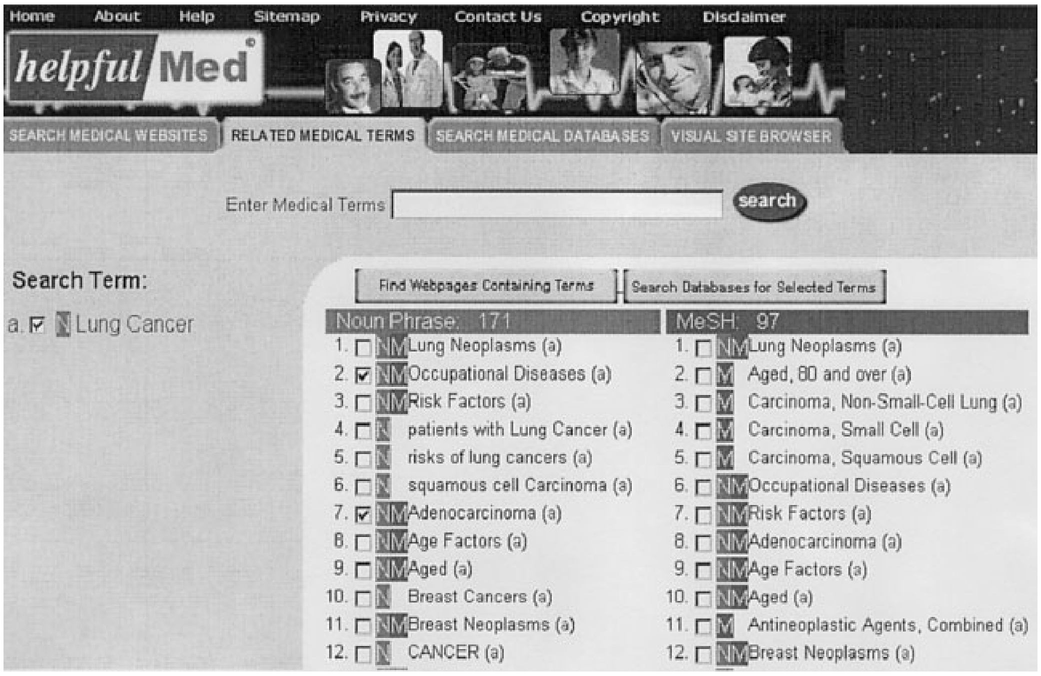
3.3 Searching Medical Databases

As mentioned previously, current search engines are able to capture only what is being called the "surface Web"—that is, those Web pages that are hyperlinked to other pages over the Web. However, there is also an Invisible Web of databases and other sites that respond only to specific queries. While an initial page for a database such as the Merck Manual of Diagnosis and Treatment or MEDLINE may appear in a collection of Web pages because it has been hyperlinked from another page, the information contained within that database is effectively lost to searchers unless they choose to search it specifically. HelpfulMED provides access to a variety of databases currently publicly available over the Internet. These include citation databases such as MEDLINE and CANCERLIT, online reference works such as the Merck Manual of Diagnosis, and Treatment and Physicians Data Query (PDQ) which provides peer-reviewed summaries on cancer treatment, screening and detection, prevention, genetics and supportive care. Access is also provided to evidence-based medical databases (EBM) such as the American College of Physicians Journal Club (ACP), National Guidelines Clearinghouse (NGC) and the York Database of Abstracts of Reviews of Effectiveness (DARE). Medical librarians at the Arizona Health Sciences Library selected these databases as being those with the most comprehensive and accurate information.

Access to these systems provides those searching for medical information with a much richer representation of information for diagnosis and treatment. EBM is a growing area for physicians and others in the field of medicine. The adoption of EBM is based on increasing costs for healthcare, variations across providers, hospitals and geographic regions in the level of healthcare service, and the desire of all concerned with patient treatment to give and receive the best care possible (Woolf et al., 1999). MEDLINE and CANCERLIT provide access to extensive collections of citations of primary research; the Merck Manual and PDQ provide information for diagnosis and treatment and ACP, NGC, and DARE allow medical professionals and other interested parties access to evidence-based medical information. Thus, searchers have a variety of types of resources that can be searched through one interface.

When searching for information in medical databases via HelpfulMED, the user is able to choose which databases are of interest. The user can also indicate how many results from each database he or she would like the system to retrieve; the default is "5," but the user can choose to see the results in increments of 5 up through 50. This gives the user some control over the retrieval process and lessens the effects of information overload.

From the retrieved results page (Fig. 4) the user is able to follow links of interest and see their search terms highlighted in the text of the Web page. This helps the user quickly identify the presence of the search term in the retrieved document for a quicker analysis as to whether the document in question might be of use.



Search Term:

a. ☑ N Lung Cancer

Noun Phrase: 171	MeSH: 97
1. □ NM Lung Neoplasms (a)	1. □ NM Lung Neoplasms (a)
2. ☑ NM Occupational Diseases (a)	2. □ M Aged, 80 and over (a)
3. □ NM Risk Factors (a)	3. □ M Carcinoma, Non-Small-Cell Lung (a)
4. □ N patients with Lung Cancer (a)	4. □ M Carcinoma, Small Cell (a)
5. □ N risks of lung cancers (a)	5. □ M Carcinoma, Squamous Cell (a)
6. □ N squamous cell Carcinoma (a)	6. □ NM Occupational Diseases (a)
7. ☑ NM Adenocarcinoma (a)	7. □ NM Risk Factors (a)
8. □ NM Age Factors (a)	8. □ NM Adenocarcinoma (a)
9. □ NM Aged (a)	9. □ NM Age Factors (a)
10. □ N Breast Cancers (a)	10. □ NM Aged (a)
11. □ NM Breast Neoplasms (a)	11. □ M Antineoplastic Agents, Combined (a)
12. □ N CANCER (a)	12. □ NM Breast Neoplasms (a)

FIG. 5. Concept space results.

3.4 Related Medical Terms

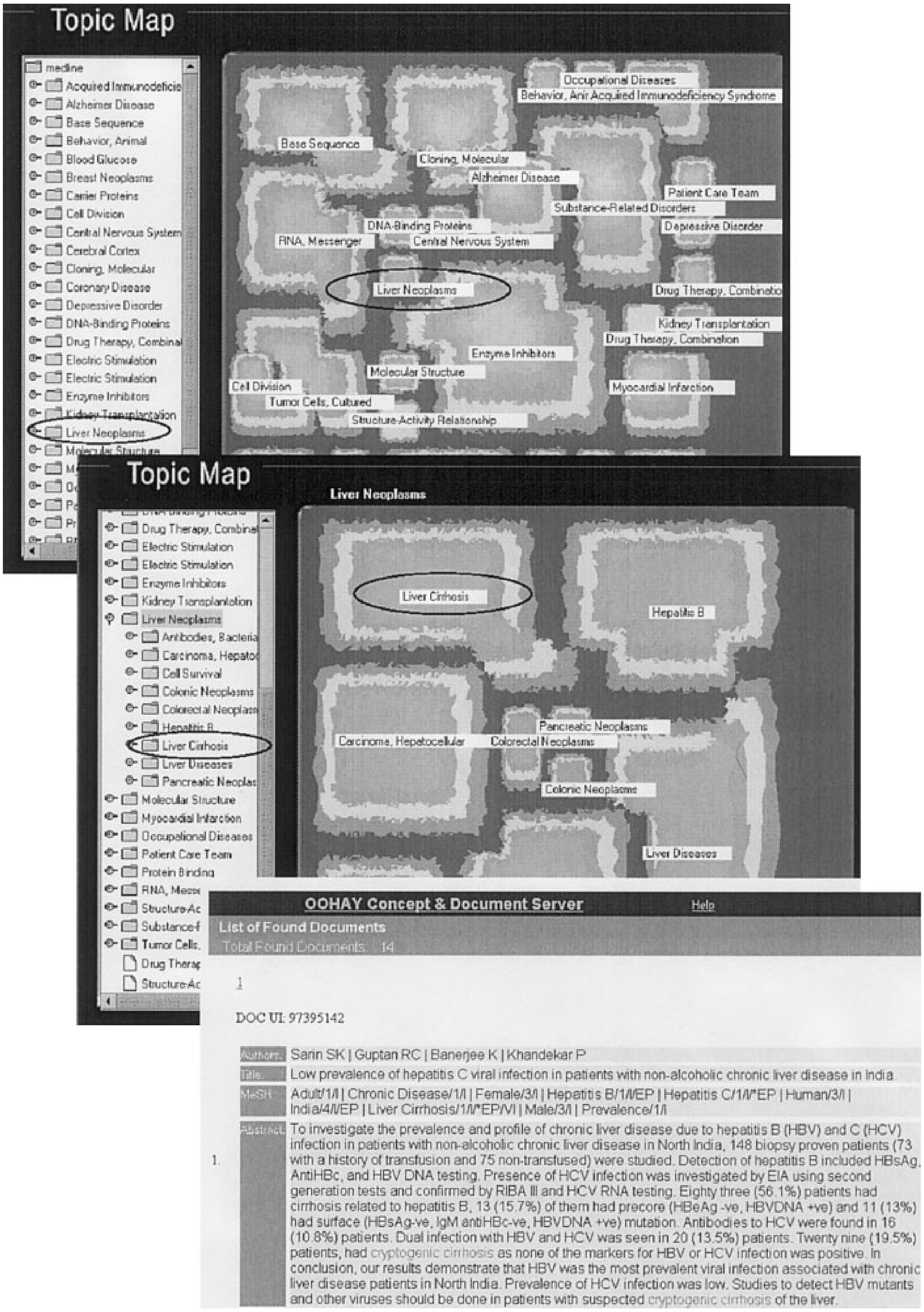
This section of HelpfulMed is a *Concept Space* of medical-related terms. A concept space is an automatic thesaurus for domain-specific terms developed by the Artificial Intelligence Lab; it is called *Concept Space* because the goal is to create meaningful and understandable domain-specific networks of terms and weighted associations that could represent the concepts (terms) and their associations for the underlying *information spaces* (i.e., documents in different domain-specific databases). This system also assists in concept-based, cross-domain information retrieval (Chen et al., 1996). What differentiates a concept space from traditional subject heading lists or thesauri is that rather than being developed and assigned by human experts, a concept space is created wholly from phrases contained within each of the documents processed. In other words, it is a thesaurus of terms drawn directly from the documents housed in the collection. Concept space uses indexing and analysis co-occurrence techniques to suggest other related and relevant search terms. These techniques are based on the computed relationships between terms. For HelpfulMed, the concept space was generated using the entire collection of Medline abstracts.

A search in the "Related Medical Terms" section provides the user with a list of additional terms that might more accurately describe the information need. If the user inputs "lung cancer" at this point (Fig. 5), the system will return a list of related noun phrases (N) drawn from the concept space, and MeSH terms (M), plus a list of authors (A). Thus, the user can decide if s/he wants to search phrases extracted from the text, related medical subjects headings, authors, or any combination of the three. In addition, on the left side of the screen the terms searched are displayed; "Lung Cancer" is search term "a". After each term returned by the concept space, there is an indication (a) that these terms are related to the lung cancer search. Additional searches would be listed on the left with subsequent letters— "b", "c", etc. and the terms to which they relate would be indicated by the appropriate letter after each one. In Figure 5, the user has chosen "Occupational Diseases" and "Adenocarcinoma" as additional terms to be searched by placing a check mark in the box next to each term. When the user has finished building the search, he or she can choose to either "Find Webpages Containing Terms" or "Search Databases for Selected Terms." The user is thus spared the onerous task of entering these search terms in Internet search engines and all of the databases he or she wants to search for information. A click of a button completes this task for the user.

3.5 Visual Site Browser

The Visual Site Browser, also called MEDMap, is a graphical system designed to facilitate the information browsing behavior of users in the domain of medical-related research. The input data to MEDMap consists of 10 million medical abstracts obtained from MEDLINE. By applying indexing, noun phrasing and self-organizing map techniques, MEDMap generates a subject hierarchy that contains 132,700 categories and 4,586 maps. The MEDMap also combines a text-based alphabetic display and a graphical approach to represent the subject categories generated (Fig. 6).

Using the Visual Site Browser, the user is able to "drill down" through the levels of a map, browsing for topics of interest until a collection of documents is eventually reached. In the case shown in Figure 6, the user chose to



List of Found Documents

Total Found Documents: 14

1

DOC UI: 97395142

1.

Authors:	Sarin SK | Guptan RC | Banerjee K | Khandekar P
Title:	Low prevalence of hepatitis C viral infection in patients with non-alcoholic chronic liver disease in India.
MeSH	Adult/1/I | Chronic Disease/1/I | Female/3/I | Hepatitis B/1/I/EP | Hepatitis C/1/I/*EP | Human/3/I | India/4/I/EP | Liver Cirrhosis/1/I/*EP/VI | Male/3/I | Prevalence/1/I
Abstract:	To investigate the prevalence and profile of chronic liver disease due to hepatitis B (HBV) and C (HCV) infection in patients with non-alcoholic chronic liver disease in North India, 148 biopsy proven patients (73 with a history of transfusion and 75 non-transfused) were studied. Detection of hepatitis B included HBsAg, AntiHBc, and HBV DNA testing. Presence of HCV infection was investigated by EIA using second generation tests and confirmed by RIBA III and HCV RNA testing. Eighty three (56.1%) patients had cirrhosis related to hepatitis B, 13 (15.7%) of them had precore (HBeAg -ve, HBVDNA +ve) and 11 (13%) had surface (HBsAg-ve, IgM antiHBc-ve, HBVDNA +ve) mutation. Antibodies to HCV were found in 16 (10.8%) patients. Dual infection with HBV and HCV was seen in 20 (13.5%) patients. Twenty nine (19.5%) patients, had cryptogenic cirrhosis as none of the markers for HBV or HCV infection was positive. In conclusion, our results demonstrate that HBV was the most prevalent viral infection associated with chronic liver disease patients in North India. Prevalence of HCV infection was low. Studies to detect HBV mutants and other viruses should be done in patients with suspected cryptogenic cirrhosis of the liver.

FIG. 6. MEDMap: the top-level map; the sub-map for Liver Neoplasm; and documents returned for Liver Cirrhosis.

browse the phrase "Liver Neoplasms," which is circled; the second screenshot of the figure displays the map for "Liver Neoplasms." The user chose to explore the phrase "Liver Cirrhosis," which is circled on the map and in the alphabetic display. The user continued refining the search until relevant documents were found. Visual Site Browser is a unique feature of HelpfulMed, and is thus far the only medically-related online retrieval systems known to provide this service. We believe this approach would be useful for the browsing of directory systems such as Yahoo! However,

TABLE 1. Summary of simulation results of the spider evaluation.

Spider System	Total number of pages visited	Number of Good Pages visited	Precision	Recall[1]	Time (minutes)
BFS Spider	100,000	36,307	36.3%	21.2%	12.7
PageRank Spider	100,000	19,630	19.6%	11.5%	1183.6
Hopfield Net Spider	100,000	40,014	40.0%	23.4%	12.6

[1] Since only 100,000 pages were visited, the recall rate of an ideal spider would have been 58.3% (= 100,000/171,405) rather than 100%.

unlike Yahoo! and other directory search engines, Visual Site Browser was created by a computer through a comprehensive analysis of a large collection of quality and time-critical documents, with a minimum of human intervention. The result is a more efficient use of time and a more fine-grained representation of content. We also believe that the map and its visual clues facilitate browsing of large collections. The size of the "islands" represents the size of the content contained below each island and the spatial proximity of the islands indicates semantic proximity between the categories. In addition, the visual clue of island "layers" represents map "levels." Thus, when clicking on each level the user is provided with visual orientation clues. None of the other systems provide the user with the ability to visually browse in this manner.

4. Evaluations

Studies of the performance of the medical spider, concept space and visual site browser have been conducted. The evaluation of the medical spider was conducted in the general medical domain; the studies of the concept space and the visual site browser were carried out on a subset of the medical domain, cancer research. The following sections outline in detail these studies and their results.

4.1 Medical Spider

In a study designed to evaluate the performance of our medical spider based on the Hopfield Net spreading activation algorithm, it was compared to both a Breadth-First Search Spider (BFS Spider) and a Best-First Search Spider using PageRank (PageRank Spider). The BFS Spider was chosen as a comparable spider due to its wide acceptance by many search engines; while PageRank, relying on the link structure of the Web as input to the relative importance of documents, is used effectively by the search engine Google for Web page ranking (Brin & Page, 1998). If a page is linked from many other pages or linked with some "good" pages (or pages that have a high PageRank score), it will receive a higher score. In our spider, the PageRank algorithm was implemented based on the description in Brin & Page (1998). Two experiments were conducted to compare the spidering systems: a simulation to analyze the spidering process and the precision/recall rates of each system.

Both experiments utilized a local database, which contained a pre-fetched "snapshot" of the Web in order to ensure that the same testbed was available for each of the three spiders. This design is based on the setup of the spider evaluation experiments performed by Cho et al. (1998). In order to create this testbed, our medical librarian identified five high-quality pages to be used as seed URLs. We ran a random-first search using these five seed URLs as the starting points. These pages were fetched from the Web and all URL links contained in them were extracted. From this set of unvisited URLs, one URL was randomly selected and the corresponding page was fetched from the Web. All the links from this page were extracted and added to the set of unvisited URLs. The process was then repeated by randomly choosing a URL from the set of unvisited URLs, fetching the page, and adding the new links to the set of unvisited URLs in each step. The resulting testbed consisted of 1,040,388 valid, unique Web pages. Similarly to the method used in the study of Cho et al. (1998), a simple concept of *good page* was used to facilitate the comparison of the performances of the three spiders. The good page concept is used to automatically determine the relevance of a page to the medical domain. In our evaluation, a page is considered a good page if the percentage of medical phrases over the total number of phrases found in the page was greater than a certain threshold. To determine whether the phrases are medical, they are compared with a medical lexicon based on the UMLS Metathesaurus. In our pilot experiment, a set of 100 randomly sampled Web pages were classified by this method and verified by an expert in the medical domain. The error rate of this simple classification method was 5.0%. Readers should note that, however, the term *good page* is used here only as a convenient "notation" and has not been tested rigorously; it just allowed us to estimate the performance of our spider in a feasible way.

Based on our classification method, the testbed in the current experiment contained 171,405 good pages. Each spider was given the same five medical URLs and each ran until 100,000 Web pages had been visited. The Hopfield Net spider, which is based on the Hopfield Net searching algorithm (Hopfield, 1982; Chen & Ng, 1995), retrieved 40,014 good Web pages (40.0% of all pages visited), compared to 36,307 (36.3%) retrieved by BFS spider and 19,630 (19.6%) by PageRank Spider. As shown in Table 1, the Hopfield Net spider had a higher precision rate than either the BFS Spider or the PageRank Spider as calculated using the following formula:

$$\text{Precision rate} = \frac{\text{number of Good Pages visited by the spider}}{\text{number of all pages visited by the spider}}$$

TABLE 2. (a) Summary of user study results of the spider evaluation; (b) *t*-tests on the results.

(a) Spider System	Average relevance score
BFS Spider	2.13
PageRank Spider	1.78
Hopfield Net Spider	2.30

(b) Comparison	*p*-value
BFS Spider versus PageRank Spider	0.0307*
BFS Spider versus Hopfield Net Spider	0.3127
PageRank Spider versus Hopfield Net Spider	0.0188*

* The difference is statistically significant at the 5% level.

$$\text{Recall rate} = \frac{\text{number of Good Pages visited by the spider}}{\text{number of Good Pages in the testbed}}$$

Because the total number of Good Pages in the testbed was 171,405 and the total number of all pages visited by the spider was fixed at 100,000, the precision rate and the recall rate were directly proportional to each other for each spider. We are more concerned with precision than recall because it is the quality of the pages in which we are interested, not the quantity; thus for the remainder of this section, we focus the discussion on the precision rate. In addition to recall and precision, we also measured the time used by each spider in order to compare efficiency.

The Hopfield Net Spider was as efficient at retrieving pages (12.6 minutes) as the BFS Spider which took 12.7 minutes. The PageRank Spider, which took 1183.6 minutes, was significantly slower because the PageRank algorithm is known to be recursive in nature and computationally expensive (Brin & Page, 1998; Haveliwala, 1999).

The second experiment sought to validate the collection created by each spider. To this end, two senior graduate students with medical training were recruited as domain experts. Each expert was assigned 100 Web pages randomly drawn from the collection of Web pages fetched by each of the three spiders during the simulation. The source of each Web page was not disclosed to the experts in order to eliminate any possible bias. For each page, each of the experts was asked to judge relevance to the medical domain based on a score of 1 to 4; 1 was the least relevant and 4 the most relevant. As shown in Table 2, the Hopfield Net spider received the highest relevance score at 2.30; the BFS relevance score was close at 2.13 and the PageRank spider score was 1.78. We also performed *t*-tests on the data and the results are shown in Table 2. The Hopfield Net Spider and BFS spider performed significantly better than the PageRank Spider at the 5% level.

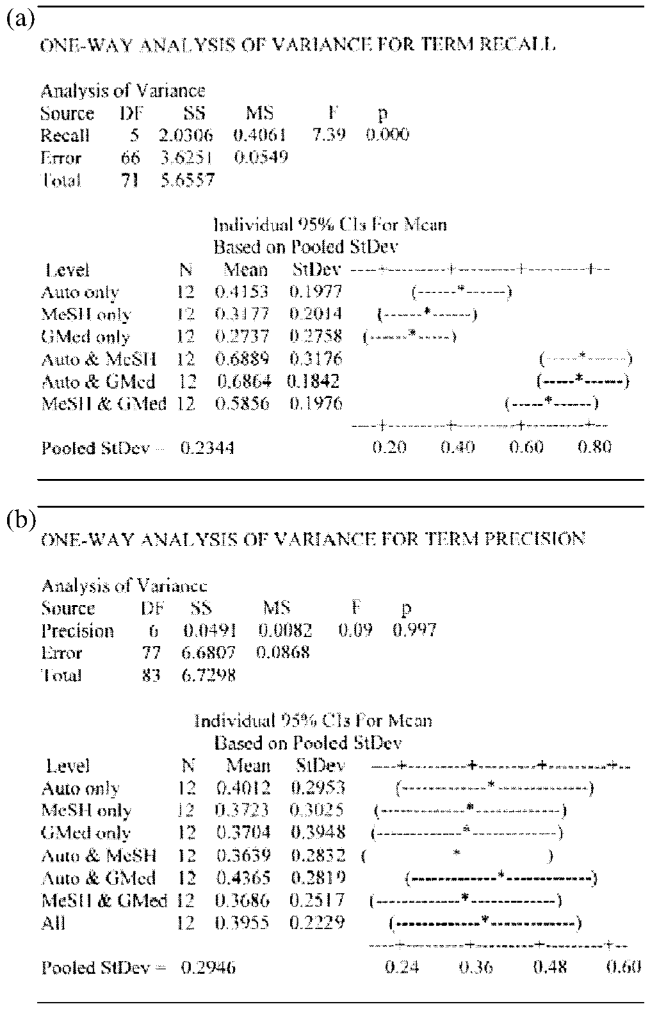
In terms of precision rate and relevance score, the Hopfield Net Spider performed best, followed by the BFS Spider. The good performance of the Hopfield Net Spider shows that the algorithm effectively combined the use of Web link structure analysis and page content analysis to locate Web pages relevant to the medical domain. The data indicate that the lesser performance of the PageRank Spider resulted from the tendency of this spider to visit more irrelevant URLs during the earlier stages of the search because some of these URLs had a high PageRank score.

4.2 Concept Space

Previous research into the development of a concept space for cancer information using the CANCERLIT collection resulted in CancerSpace, an automatically created thesaurus of cancer-related terms with 1.3 million unique terms and 52 million relationships (Houston et al., 2000). We conducted a study to assess the usefulness of suggested terms from different thesauri: our automatically generated system, the MeSH concept space and Internet Grateful Med, which at the time of the study was the most often cited online tool based on the UMLS Metathesaurus. Five cancer researchers affiliated with the University of Arizona Cancer Center and one veterinarian participated in our experiment as subjects.

The experiment was divided into two phases. During phase one twelve *directed* searches were performed on each of the three thesauri for a total of 36 searches. Each subject was asked to state one or two terms with which they would begin a document search, and to suggest five related terms for each search term.

For phase two, original search terms from phase one were entered into a thesaurus and the subjects rated the top 40 thesaurus suggested terms as to whether they were "relevant," "possibly relevant," or "not relevant." This step was repeated for each thesaurus. Precision and recall were then calculated as follows: (1) "very relevant" terms, 1 point; (2) "possibly relevant" terms, 0.5 points; and (3) "not relevant" terms, 0 points. For each search, all relevant terms from each thesaurus were combined in order to eliminate duplicates for a total relevance score. Term recall was calculated by dividing the relevance score for each thesaurus by the total relevance score for all thesauri. Term precision was calculated by dividing the total relevance score for each thesaurus by the total number of terms suggested by the thesaurus. Term recall and precision were also used to measure system performance in other research (e.g., Vélez et al., 1997; Chen & Lynch, 1992). Recall and precision rates for these systems and the results of ANOVA tests, shown in Figure 7, indicate that there were no significant differences among the three systems when used individually. The results suggest that terms returned by our tool are comparable to terms suggested by Internet Grateful Med and MeSH indexing terms. It was also discovered that the three systems rarely returned the same relevant terms. However, when systems were combined, the recall rates went up, while precision rates remained relatively static. Based on the above discovery and user feedback, which supported this notion, a combination of automatic concept space and MeSH concept space is provided to the user, with plans to add a UMLS concept space in the near future [For further discussion and evaluation of this system, please see (Houston et al., 2000)].



(a)

```
ONE-WAY ANALYSIS OF VARIANCE FOR TERM RECALL

Analysis of Variance
Source      DF       SS       MS       F       p
Recall       5   2.0306   0.4061    7.39   0.000
Error       66   3.6251   0.0549
Total       71   5.6557

                               Individual 95% CIs For Mean
                               Based on Pooled StDev
Level          N    Mean    StDev  ---+---------+---------+---------+--
Auto only     12  0.4153   0.1977          (------*------)
MeSH only     12  0.3177   0.2014      (------*------)
GMed only     12  0.2737   0.2758    (------*-----)
Auto & MeSH   12  0.6889   0.3176                          (------*------)
Auto & GMed   12  0.6864   0.1842                          (-----*------)
MeSH & GMed   12  0.5856   0.1976                      (-----*------)
                                   ---+---------+---------+---------+--
Pooled StDev = 0.2344                0.20      0.40      0.60      0.80
```

(b)

```
ONE-WAY ANALYSIS OF VARIANCE FOR TERM PRECISION

Analysis of Variance
Source        DF       SS       MS       F       p
Precision      6   0.0491   0.0082    0.09   0.997
Error         77   6.6807   0.0868
Total         83   6.7298

                               Individual 95% CIs For Mean
                               Based on Pooled StDev
Level          N    Mean    StDev   ---+---------+---------+---------+--
Auto only     12  0.4012   0.2953      (------------*-------------)
MeSH only     12  0.3723   0.3025    (------------*------------)
GMed only     12  0.3704   0.3948    (------------*------------)
Auto & MeSH   12  0.3639   0.2832  (            *            )
Auto & GMed   12  0.4365   0.2819       (------------*------------)
MeSH & GMed   12  0.3686   0.2517    (------------*------------)
All           12  0.3955   0.2229      (------------*------------)
                                    ---+---------+---------+---------+--
Pooled StDev = 0.2946                 0.24      0.36      0.48      0.60
```

FIG. 7. (a) Recall comparison by term source; (b) Precision comparison by term source.

4.3 Visual Site Browser

An earlier prototype of the MEDMap was the CancerMap, developed using the CANCERLIT collection to generate a subject hierarchy that contained 18,120 categories. An empirical study was conducted to assess whether the approach employed by CancerMap generates subject categories that are meaningful to human subjects. We designed a set of tasks to evaluate the consistency between the categories generated by the CancerMap and those expected by users during their browsing process. We also investigated the consistency between the MeSH sub-trees and users' expectations. The goal was to compare the performance of the CancerMap with that of MeSH sub-trees. Eighteen senior Ph.D. students, researchers and faculty members from the Arizona Cancer Center participated in this study. Most of them reported being familiar with the MeSH tree structure. During the evaluation process, subjects were encouraged to think aloud. Every subject finished the first-level evaluation. At the second and the third levels, some subjects quit because the categories were not in their research areas they felt they lacked the expertise to generate appropriate lists of sub-categories.

We designed a set of tasks to evaluate the consistency between the categories generated by the CancerMap and those expected by users during their browsing process. We also investigated the consistency between the MeSH sub-trees and users' expectation. The goal is to compare the performance of the CancerMap with that of MeSH sub-trees. To evaluate the *first-level labels*, we designed the task as follows:

(1) Ask a human subject to generate a list of possible sub-categories that he or she expects to see under the category of "Neoplasms."
(2) Present the lists of first-level labels generated by MeSH sub-trees and CancerMap to the subject, asking him/her to modify the list he or she has generated. During the experiment, subjects did not know from which source (the CancerMap or MeSH sub-trees) suggested labels had come.
(3) Use the list generated by the subject to evaluate the first-level labels of the MeSH sub-trees and CancerMap.

At the first level, there were six overlapping labels between the CancerMap and MeSH sub-trees. We therefore designed three tasks for each human subject in order to compare the performance of the CancerMap and MeSH sub-trees at the *second-level*. For each task, the subject was to repeat the process used for the first-level evaluation. Again, subjects did not know the source of suggested label terms.

1. We randomly selected one of the six *overlapping* labels at the first-level and asked a subject to evaluate its sub-labels. The subject was to generate his or her own desired list and modify it by reading the labels generated by the MeSH sub-trees and the CancerMap.
2. We randomly selected one of the *non-overlapping* first-level labels from MeSH sub-trees and asked a subject to evaluate its sub-labels. The subject was to generate his or her own desired list and modify it by reading the labels generated by the MeSH sub-tree.
3. We randomly selected one of the *non-overlapping* first-level labels from the CancerMap and asked a subject to evaluate its sub-labels. The subject was to generate his/her own desired list and modify it by reading the labels generated by the CancerMap.

We used *cluster recall* and *cluster precision* as defined in Roussinov & Chen (1999) as the objective measures for all the tasks. *Cluster precision* indicates the accuracy of the categories generated by a system, while *cluster recall* represents how many related categories have been captured by the system. According to Janes (1994), there are two types of relevance. Objective relevance denotes the accuracy of the categories generated by a system, whereas subjective relevance denotes a subject's perception of the accuracy of the categories created. In this study, we selected the subjective measure, because only when the categories generated are consistent with users' expectation can they be helpful in facilitating users' browsing behavior. Therefore, both the cluster precision and cluster recall results obtained

TABLE 3. Summary of the CancerMap Experiment Results (C: CancerMap, M: MeSH sub-trees)

	First Level	Level 2 (overlap)	Level 2 (non-overlap)	Level 3
Recall comparison	C: 0.557	C: 0.765	C: 0.859	C: 0.839
	M: 0.466	M: 0.113	M: 0.466	M: 0.459
	p = 0.049*	p = 0.000*	p = 0.000*	p = 0.003*
Precision comparison	C: 0.926	C: 0.826	C: 0.829	C: 0.863
	M: 0.956	M: 0.608	M: 0.904	M: 0.917
	p = 0.591	p = 0.104	p = 0.459	p = 0.808

* The difference is statistically significant at the 5% level.

in this study refer to perceived cluster precision and perceived cluster recall.

One of the characteristics of automatically generated terms compared with manually assigned terms is the likelihood that a term phrase which appears as a category in one system will appear as a subcategory in the other. Thus, evaluation of the term phrases suffers somewhat from "level confusion" in that several first-level categories on the CancerMap were sub-categories of one first-level category of the MeSH sub-trees. We counted those CancerMap categories as one first-level category. As a result, the number of first-level categories on the CancerMap was reduced. However, a one-way ANOVA test on the data revealed that the first-level of the CancerMap performed significantly better than that of the MeSH sub-trees in *perceived cluster recall* ($p = 0.049$). There was no significant difference in *perceived cluster precision* ($p = 0.591$) between the two. We found similar results in the second and third-level comparisons. Table 3 summarizes the experiment results. As displayed in Table 3, the CancerMap was comparable to MeSH sub-trees in *perceived cluster precision* at each level and was significantly better in *perceived cluster recall* at all levels. We found that the one-aspect categorization employed by manually created subject headings may restrict users' browsing activity. Overall results of the empirical study demonstrate that the approach employed by CancerMap generated a meaningful subject hierarchy to facilitate browsing behavior.

5. Conclusions and Future Directions

In this paper we discussed the development of a portal to serve the information seeking needs of the medical professionals, researchers and other advanced users. The functionalities of this system have the potential to increase user satisfaction by providing the capability to search multiple information spaces through one interface, and by providing mechanisms which allow users to refine and focus their searches through the use of an automatically generated thesaurus and maps. User studies conducted on the different components of the system indicate that they perform as well as or better than comparable systems. We are currently planning to study the performance of the system as a whole (rather than individual components) based on real search tasks. The study will be similar to those reported in the Text Retrieval Conferences (TREC) and will help determine whether real users can perform real tasks better with our system.

Future development work includes integrating additional technology that will summarize the documents retrieved from the Web and other searchable databases. We also plan to enhance the search interface such that it can process complex search queries, such as those including author names, journal titles, publication types, publication year, as well as other metadata. In addition, a dynamic self-organizing map which categorizes documents retrieved from the Web based on user-chosen attributes such as noun phrases is in development; as is a concept-space-based related terms "suggester" which will present additional terms to the user as the results from a search for Web pages are returned. Problems involved with these developments include scalability into other larger domains, and problems of "interface overload" as additional features are added.

Acknowledgements

This research is supported by:

- NSF Digital Library Initiative-2, H. Chen, "High-Performance Digital Library Systems: From Information Retrieval to Knowledge Management." IIS-9817473, April 1999–March 2002.
- NIH/NLM, H. Chen, "UMLS Enhanced Dynamic Agents to Manage Medical Knowledge." 1 R01 LM06919-1A1. February 2001–January 2004.

We thank the National Library of Medicine for making UMLS and MeSH freely available to researchers. We would also like to thank the following individuals who contributed to this project: Thian-Huat Ong, Wojciech Wyzga, Yohanes Santoso, Bill Oliver, Ming Yin, Haiyan Fan, Ye Fang, Tailong Ke, Joanne Martinez, Gondy Leroy, Alan Yip, Hui Liu, Andy Lowe, Jared Phipps, Chienting Lin, Mark Carvalho, Wai-Ki Sung, Mark Chen, Jennifer Xu, and Xiaoyun Sun of the Artificial Intelligence Lab; Polin Lei and the staff of the Arizona Health Sciences Library; and the researchers, staff and students of the Arizona Cancer Center.

References

Bates, M.J. (1986). Subject access in online catalogs: a design model. Journal of the American Society for Information Science, 37(6), 357–376.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia.

Chen, H., Houston, A.L., Sewell, R.R., & Schatz, B.R. (1998). Internet browsing and searching: User evaluation of category map and concept space techniques. Journal of the American Society for Information Science, 49(7), 582–603.

Chen, H. & Lynch, K. (1992). Automatic construction of networks of concepts characterizing document databases. IEEE Transactions on Systems, Man and Cybernetics, 22, 885–902.

Chen, H., & Ng, T. (1995). An Algorithmic Approach to Concept Exploration in a Large Knowledge Network (Automatic Thesaurus Consultation): Symbolic Brand-and Bound Search vs. Connectionist Hopfield Net Activation. Journal of the American Society for Information Science, 1995, 46(5), pp. 348–369.

Chen, H., Schatz, B.R., Ng, T.D., Martinez, J.P., Kirchhoff, A.J., & Lin, C. (1996). A parallel computing approach to creating engineering concept spaces for semantic retrieval: The Illinois digital library initiative project. IEEE Transactions on Pattern Analysis and Machine Intelligence, Special Section on Digital Libraries: Representation and Retrieval, 18(8), 771–782.

Cho, J., Garcia-Molina, H., & Page, L. (1998) Efficient crawling through URL ordering. in Proceedings of the 7th World Wide Web Conference, Brisbane, Australia, Apr 1998.

Cimino, J.J., Johnson, S.B., Peng, P., & Aguirre, A. (1994). From ICD9-CM to MeSH using the UMLS: A how-to guide. Paper presented at the Annual Symposium on Computer Applications in Medical Care.

Crouch, C.J. (1990). An approach to the automatic construction of global thesauri. Information Processing and Management, 26(5), 629–640.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41, 391–407.

Eysenbach, G., Powell, J., Kuss, O., & Sa, E.R. (2002). Empirical studies assessing the quality of health information for consumers on the World Wide Web. Journal of the American Medical Association, 287(20), 2691–2700.

Fallis, D., and Frické, M. (2002). Indicators of accuracy of consumer health information on the Internet: a study of indicators relating to information for managing fever in children in the home. Journal of the American Medical Informatics Association, 9(1), 73–79.

Furnas, G.W., Landauer, T.K., Gomez, L.M., & Dennis, S.T. (1987). The vocabulary problem in human-system communication. Communications of the ACM, 30(11), 964–971.

Guntzer, U., Juttner, G., Seegmuller, G., & Saare, F. (1989). Automatic thesaurus construction by machine learning from retrieval sessions. Information Processing and Management, 25(3), 265–273.

Haveliwala, T.H. (1999). Efficient computation of PageRank. Stanford University Technical Report [Online]. Available at: http://dbpubs.stanford.edu:8090/pub/1999–31

Hearst, M.A., & Pedersen, J.O. (1996). Reexamining the Cluster Hypothesis: Scatter/Gather on retrieval results. In Proceedings of the 19th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'96), Zurich, Switzerland, Aug 1996.

Hopfield, J.J. (1982). Neural Network and Physical Systems with Collective Computational Abilities. In Proceedings of the National Academy of Science, USA, 1982, 79(4), pp. 2554–2558.

Houston, A.L., Chen, H., Schatz, B.R., Hubbard, S.M., Sewell, R.R., & Ng, T.D. (2000). Exploring the use of concept space to improve medical information retrieval. International Journal of Decision Support Systems, 30, 171–186.

Janes, J.W. (1994). Other people's judgments: A comparison of users' and others' judgments of document relevance, topicality, and utility. Journal of the American Society for Information Science, 45(3), 160–171.

Kohonen, T. (1995). Self-Organized Maps. Berlin: Springer-Verlag.

Lyman, P., & Varian, H.R. (2000). How much information. [Online]. Available at http://www.sims.berkeley.edu/how-much-info/

Mechkour, M., Harper, D., & Muresan, G. (1998). The WebCluster project. Using clustering for mediating access to the World Wide Web. In Proceedings of the 21st annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (pp. 357–358), Melbourne, Australia.

Roussinov, D.G. & Chen, H. (1999). Document clustering for electronic meetings: an experimental comparison of two techniques. Decision Support Systems, 27(1), 67–81.

Salton, G. (1986). Another look at automatic text-retrieval systems. Communications of the ACM, 29(7), 648–656.

Salton, G. (1989) Automatic Text Processing. Addison-Wesley Publishing Company, Inc. Reading, MA.

Salton, G., Wong, A., & Yang, C.S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613–620.

Srinivasan, P. (1996). Query expansion and MEDLINE. Information Processing and Management, 32, 431–443.

Tolle, K.M., & Chen, H. (2000) Comparing noun phrasing techniques for use with medical digital library tools. Journal of the American Society for Information Science. 51(4), 352–370.

van Rijsbergen, C.J. (1979). Information Retrieval. Butterworths. London. Second Edition.

Vélez, B., Wiess, R., Sheldon, M.A., & Gifford, D.K. (1997). Fast and Effective Query Refinement. In Proceedings of the 20th ACM Conference on Research and Development in Information Retrieval, Philadelphia, Pennsylvania, July 1997.

Woolf, S.H., Grol, R., Hutchinson, A., Eccles, M., & Grimshaw, J. (1999). Potential benefits, limitations, and harms of clinical guidelines. British Medical Journal 318(7182), 527–530.

Wu, M., Fuller, M., & Wilkinson, R. (2001). Using clustering and classification approaches in interactive retrieval. Information Processing and Management, 37, 459–484.

Zamir, O. & Etzioni, O. (1999). Grouper: a dynamic clustering interface to Web search results. In Proceedings of the 8th World Wide Web Conference, Toronto, May 1999.