MetaSpider: Meta-Searching and Categorization on the Web

Hsinchun Chen, Haiyan Fan, Michael Chau, and Daniel Zeng
Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721. E-mail: {hchen, fan, mchau, zeng} @bpa.arizona.edu

It has become increasingly difficult to locate relevant information on the Web, even with the help of Web search engines. Two approaches to addressing the low precision and poor presentation of search results of current search tools are studied: meta-search and document categorization. Meta-search engines improve precision by selecting and integrating search results from generic or domain-specific Web search engines or other resources. Document categorization promises better organization and presentation of retrieved results. This article introduces MetaSpider, a meta-search engine that has real-time indexing and categorizing functions. We report in this paper the major components of MetaSpider and discuss related technical approaches. Initial results of a user evaluation study comparing MetaSpider, NorthernLight, and MetaCrawler in terms of clustering performance and of time and effort expended show that MetaSpider performed best in precision rate, but disclose no statistically significant differences in recall rate and time requirements. Our experimental study also reveals that MetaSpider exhibited a higher level of automation than the other two systems and facilitated efficient searching by providing the user with an organized, comprehensive view of the retrieved documents.

1. Introduction

The dynamic, unregulated nature and rapid proliferation of the World Wide Web have made finding useful information on it increasingly difficult. Development of information retrieval (IR) systems such as Web search engines to a great extent has alleviated this problem of information and cognitive overload. By 1998, search engines numbered in the hundreds (possibly thousands) and had come to include specialized search engines within narrow domains (Chignell, Gwizdka, & Bodner, 1999). However, their effectiveness and usefulness have been limited by low search precision and poor presentation. Search engines often retrieve a large number of documents, many of them not relevant to user queries. Users themselves must manually explore suggested links and judge their relevance. A further complication is that each search engine maintains its own query format, searching strategies (often hidden from the user), output format, and relevance ranking strategies. From the user's point of view, dealing with an array of different interfaces and understanding each one's idiosyncrasies adds much confusion and presents an additional layer of information and cognitive overload. Meta-search, which leverages the capabilities of multiple Web search engines and other types of information sources, has provided a simple, uniform user interface that promises significant advances in coping with overload and low-precision issues.

Traditional search engines present search results as ranked lists, ordered by estimated relevance to the query. A major drawback of this presentation is that it fails to give users a quick "feel" for the retrieval. Users know little about a document's content until they click on and read it. This can be very time consuming and disruptive in a dynamic, fast-changing electronic information environment. In a browsing scenario, it is highly desirable for an IR system to provide such a feel for a summarization of the retrieved document set so the user can explore a specific topic and gain general view of the particular area of interest. An ideal IR system should categorize retrieved documents automatically and give the user easy access to various aspects of the subject of interest. In a searching scenario, the user needs immediate assistance in locating places that might contain useful information. Such assistance may take the form of determining the relevance of retrieved document sets or helping reformulate queries based on feedback from the previous search. In both cases, traditional ranked-list presentation lacks the immediate responsiveness desired for high quality IR.

To address these problems with existing Web search approaches, we developed MetaSpider, a meta-search engine that performs real-time post-retrieval document clustering, an IR technique that has been shown to produce superior results (Hearst & Pedersen, 1996; Zamir & Etzioni,

Received July 3, 2000; Revised February 6, 2001; Accepted May 8, 2001

1999). The main strength of MetaSpider is that it combines meta-search and categorization in an integrated manner, differentiating itself from all other meta-search or document analysis tools available on the Web. Our goal is to aid users in interfacing with multiple search engines, gaining an overview of the retrieved documents, quickly sifting through noise, and locating useful information.

One major component of our research is evaluation of the proposed integrated approach. The existing IR evaluation methodology based on precision and recall measured by the relevancy of the retrieval documents is suitable for traditional retrieval, in which an IR system is used to help the user answer a specific query (e.g., Leighton & Srivastava, 1999). Nevertheless, many users also want to understand the context and various aspects of the topic of interest and to identify major relevant themes. For these tasks, new measures and an evaluation framework are called for. In this paper, we describe in details such an evaluation framework developed to study the effectiveness of MetaSpider, and report initial findings of a user study guided by this metric.

The article starts with a brief review of related fields. Section 2.1 discusses using meta-search engines to increase precision. Section 2.2 presents different clustering paradigms and existing post-retrieval document clustering visualization techniques. Section 2.3 brings in an important dimension of IR, the evaluation of IR systems. In section 3, the architectural design and major components of MetaSpider are illustrated. Section 4 discusses the evaluation framework of the user study, and section 5 reports and discusses the findings of the user experiments. Finally, in section 6, we summarize our research contributions and discuss future work.

2. Related Work

MetaSpider provides both meta-search and post-retrieval clustering functionality. In this section, we review literature in both areas. Since user evaluation constitutes a significant portion of our research, we also present a brief review of IR system evaluation.

2.1 Meta-Search

Empirical studies show that every Web search engine returns a different set of documents for the same query. In addition, each search engine maintains its own query formats and freshness criteria. For example, some search engines claim to be able to do *natural language processing*, some support Boolean logic and scope limitation, and others do not. Various ranking algorithms utilized by search engines to decide in which order the retrieved documents will be presented to the user make the results even more disparate. The idiosyncrasies and diversity of existing search engines have left bewildered users having to sift through piles of information.

Selberg and Etzioni (1995) suggested that by relying solely on one search engine, users could miss over 77% of the references they would find most relevant because no single search engine is likely to return more than 45% of relevant results. A study by NEC Research Institute drew some similar conclusions, revealing an alarming fact about Internet search engines: they cannot keep up with the net's dynamic growth, and each search engine covers only about 16% of the total Web sites (Lawrence & Giles, 1999).

The emergence of meta-search engines provides a credible resolution of divergence by triangulating output from several engines to arrive at relevant results. By sending queries to multiple search engines and collating only the highest-ranking subset of the returns from each, meta-search engines can greatly improve search results.

Meta-search engines vary widely in their combination of primary search engines, query formation, results processing, and display. For example, SavvySearch (www.savvysearch.com) supports up to 100 engines and allows the searcher to customize a selection of engines in which to search and in what order. Users can then save the customized selection for future use. Some meta-search engines may simply be "mega engines" that run long lists of general or topic-specific search engines. An example is Beaucoup Search Engine (www.beaucoup.com/engines.html), which offers the user only an assemblage of boxes; users have to enter each of its 14 search engines separately (Garman, 1999). Some recent meta-search engines put the user in control of results by allowing him or her subjectively to filter and rate results rather than relying on generic relevance rating. Satyam Spark Solutions' SearchPad 1.6 is reported to have this interactive feature (Morgan, 1999). Besides the Web, some more recent systems also search other parts of the Internet such as Usenet, newswires, FTP, business news, quotes, weather, white and yellow pages, etc. Dogpile (www.dogpile.com) is an illustration of this type of system. In terms of content, there are general meta-search engines as well as those that feature specific subject categories. For instance, BuildingOnline (www.buildingonline.com) specializes in searching in the building industry domain on the Web, and CollegeBot (www.collegebot.com) searches for educational resources.

Typically, a meta-search engine has to deal with the following set of issues: 1) It has to handle different query and output formats supported by underlying search engines. MetaCrawler (www.metacrawler.com) is an example of such catering to its primary search engine's syntax. 2) In addition to collating search results from various search services, a quality meta-search engine has to eliminate duplicate pages as well as poor, outdated pages. 3) Based on the returned results, a meta-search engine should be able to post-process and re-rank the results according to its own ranking criteria. 4) A meta-search engine should present a unified user interface that displays search results in an integrated and intuitive manner.

2.2 Document Categorization and Visualization

Manually browsing through Web pages to locate useful information can be mentally exhausting and time consum-

ing. In order to address this problem, much research has been devoted to developing techniques and tools to analyze, categorize, and visualize large collections of Web pages, among other text documents.

Documents have to be indexed before they can be retrieved in response to any given user query. Many automatic indexing algorithms have been developed to extract key concepts from text, and it has been shown that automatic indexing is as effective as human indexing (Salton, 1986). One effective approach, the Arizona Noun Phraser (AZNP) developed by our research group, performs indexing based on meaningful noun phrases rather than mere keywords (Tolle & Chen, 2000).

Based on an index, categorization tools allow users to classify documents into different categories, which in turn can be visually presented to facilitate the elicitation of meaning and understanding. Categorization and visualization of search results in recent years has been shown to be a powerful post-retrieval document processing tool that can cluster similar documents into a category and present to the user the resulting clusters in an intuitive and sensible way. The use of categorization in IR is based on the cluster hypothesis: "closely associated documents tend to be relevant to the same requests" (Van Rijsbergen, 1979). Although categorization techniques and visualization metaphors vary vastly from system to system, the purported goals are the same: to help users better comprehend the returned documents, identify interesting documents more quickly, and gain a quick overview of the documents' contents.

Web document clustering techniques can be classified into two broad categories. The first approach aims to provide additional information about the retrieved documents and can be further broken down into three subcategories. The first of these is the query term's distribution; it shows how the retrieved documents relate to each of the terms used in the query and displays how the internal subtopic structure of the documents relates to the query (Veerasamy & Belkin, 1996; Hearst, 1995). The second subcategory focuses on predefined document attributes such as size, source, topic, or author. For instance, the Boston-based search engine NorthernLight (www.northernlight.com) organizes its retrieved documents in what is marketed as a "custom search folder." Such a folder is based on type (e.g., press release, current news, special collection), subject, language, or source (e.g., government site, educational site) following library science classification methods. Electric Library (www.elibrary.com) organizes documents according to "recurring themes" and is another example of an IR system using predefined document attributes. The third subcategory uses user-specified attributes to show how the retrieved documents relate to items such as query history, user profile, etc. (Zamir, 1998).

The second approach is based on interdocument similarities and attempts to reduce the multidimensional document space to a 2-D or 3-D space by aggregating similar documents under the same theme. It provides users with a quick overview of the whole collection, making it useful not only for browsing but also for searching, because once users find an interesting document or theme, they can easily locate useful information in the vicinity. Unlike attribute-based clustering, clustering based on interdocument similarities classifies documents without predefined categories. Category labels will be determined based on the keywords that appear in the documents collected. This approach usually includes some machine learning components. For example, the self-organizing map (SOM) approach classifies documents into different categories that are automatically defined on the fly using neural network algorithms (Kohonen, 1995). These categories then are mapped into different regions, given the similarity of the documents. Regions (each of which contains similar documents) that are conceptually related are located close to each other. Lin et al. were the first to apply this algorithm to document sets (Lin, Soergel, & Marchionini, 1991; Lin, 1997). Internet-based systems employing this algorithm include WEBSOM (Kohonen, 1997) and ET-MAP (Chen, Schufels, & Orwig, 1996).

2.3 IR System Evaluation Methodology

Empirical evaluation of IR systems is critical, yet difficult to perform, partly because of the difficulty arising from the interplay of many variables such as the IR system, retrieval tasks, search topics, user sophistication, clustering methods, etc. In this section, we review recent evaluation methodologies that are relevant to this research.

Empirical evaluation looks for both quantitative data and qualitative data. Retrieval effectiveness (measured by search precision and recall) and efficiency (represented by time and effort expended) are the most commonly used criteria. Most IR system evaluations calculate precision and recall rate based on relevance (Leuski & Allan, 1999; Leuski, 1998; Leighton & Srivastava, 1999; Chignell et al., 1999; Chen, Houston, Sewell, & Schatz, 1998). Besides precision and recall, a categorization system can be evaluated on the basis of its usefulness as a browsing tool (Chen et al., 1998). Other measures, such as *term relevance* and *term association*, have also been used (Orwig, Chen, & Nunamaker, 1997). However, no previous studies have attempted to evaluate the combination of Web document retrieval and categorization.

Relevance-based evaluation is well suited to relevance retrieval, but since IR systems extend a tool exclusively designed for retrieving to accommodate an integrated information management environment, relevance-based evaluation become less applicable. Zamir & Etzioni (1999) take a different evaluation approach by analyzing the log file of the search engine and computing the number of documents clicked on by users. The assumption is that clicking on fewer documents in a given amount of time suggests that a better clustering functionality has been provided by the system. However, for many reasons, this is a very coarse measure. First, the number of clicks does not reflect how

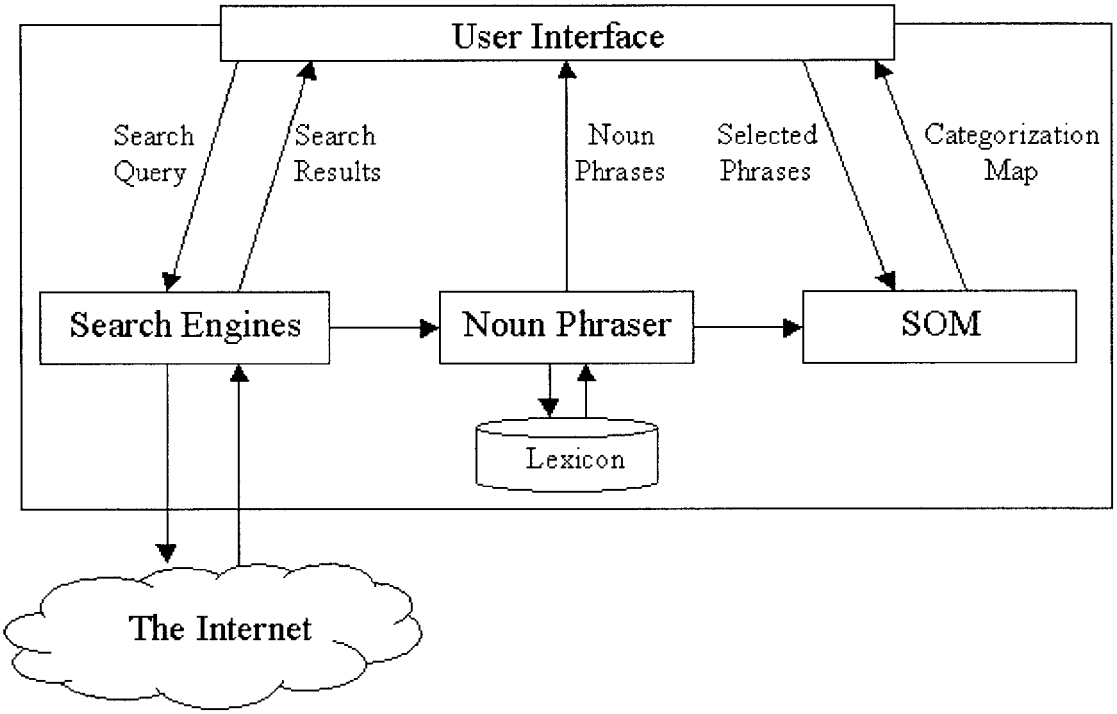

FIG. 1. MetaSpider architecture.

well the user actually understood the content of the retrieved documents. Second, fewer clicks do not necessarily imply high or low quality of the content of the documents. Third, uncertainty about user behavior also greatly calls into question the validity of this type of evaluation. Some users tend to click on many documents; others are more selective.

In addition to quantitative measures, qualitative data are also important to IR system evaluation. Qualitative data are ordinarily collected through recording users' "think aloud" protocols and through questionnaires. During typical experiments, subjects are encouraged explicitly to express their likes and dislikes concerning the system, as well as to give reasons behind their navigation choices. They are also asked to complete questionnaires regarding the experiment. Comments are usually recorded by the experimenters and later subjected to protocol analysis. In our research, we have used both qualitative and quantitative data for system evaluation.

3. MetaSpider System Architecture

In this section, we present the architectural design of MetaSpider (as shown in Figure 1) and discuss in details each major component and related technical issues. We focus on MetaSpider both as a meta-search tool and as a document categorization tool.

We first provide a brief description of an example task that the user is trying to accomplish using MetaSpider. This task will be used to illustrate the user-system interaction and the functionalities of MetaSpider. In this example, the user is searching for Web pages that are relevant to both "health" and "computer terminals."

The major components of MetaSpider are: 1) user interface, 2) Internet searching, 3) fetching, 4) Noun Phraser, and 5) Self-Organizing Map (SOM). We discuss below each component and how these components relate to each other.

3.1 User Interface

The user interface is primarily used to configure the settings for searching and fetching. The *Search* panel provides users with an array of key search engines (e.g., AltaVista, Infoseek, Lycos), as shown in Figure 2.

All six search engines have been selected by default. The system supports multiple search phrases. The relationship among these phrases can be defined as either *AND* or *OR*. These two Boolean operators determine how query phrases are going to be considered as a match.

The *Options* panel (see Figure 3) allows the user to set searching and fetching parameters such as the number of returns from each search engine, the number of spiders requested for fetching, and the upper limit for time allowed for fetching. Consideration has also been given to limiting the search scope of the search in terms of domain and location. For example, the user can specify that Web pages from the military domain (.mil) not be included in the search result. The *Stop Terms* panel lists words and phrases that are not going to be indexed. These words are mostly common words such as "month," "comment," etc. Sophisticated users can modify the list to suit their own needs by adding new terms to the list or deleting existing ones. The user can also save the search session. If the user is behind a firewall, MetaSpider can be configured such that it accesses the Web

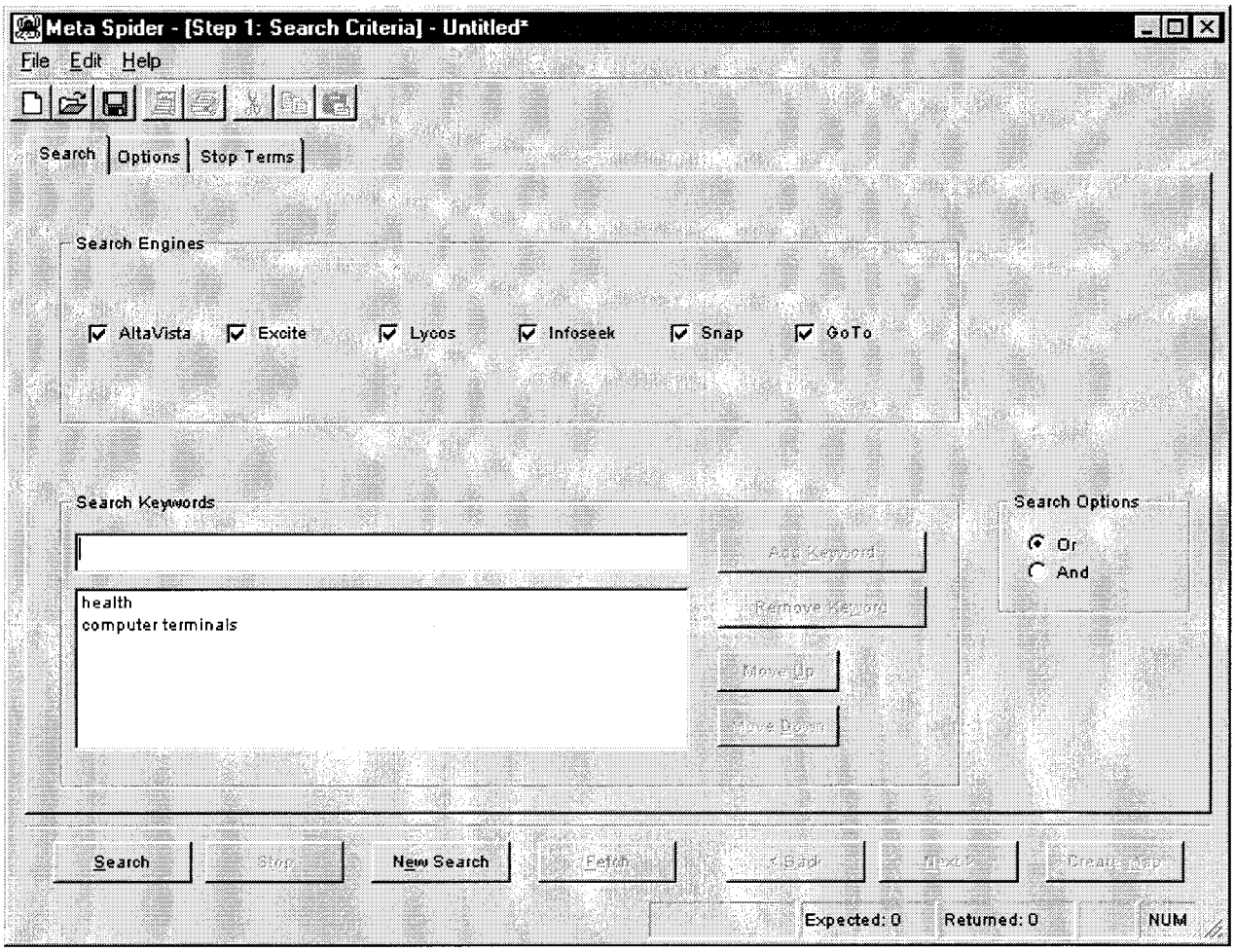

FIG. 2. MetaSpider: User interface.

through a proxy server. A *Help* file describing how to use MetaSpider is provided to users.

3.2 Searching Component

MetaSpider sends out queries to the multiple search engines selected and collects the results. On the *Search Result* panel, users can see the URL address of each document, its ranking in the search engine that returns it, and the title of the document. Clicking on the *Rank* tab, users can get search results in ascending or descending rank order. A click on any document listed will allow the user to view the actual Web page.

The current version of MetaSpider performs a weeding routine to eliminate "bad" pages from the set of documents returned by the underlying source search engines. The ranking of the remaining "good" pages is based on the relevance scores provided by source search engines; no re-ranking operations are performed. From our experience, this approach seems to work well on small document sets. Whether re-ranking will have significant impact on system performance on large document collections remains a future research topic.

3.3 Fetching Component

MetaSpider has a *GoodURL* mechanism that performs a real-time check of each candidate Web page returned by the search engines to sift out invalid pages. Invalid pages are candidate pages that do not contain the exact phrase supplied by the user. In our example, as the user queries "health" and "computer terminals," the *GoodURL* mechanism will filter out a large proportion of documents, such as those referring to just "health" or "computer" in general. MetaSpider uses different icons to distinguish between "Good" URLs and all others. As shown in Figure 4, documents with no exact phrase match are preceded by a globe icon with a red X sign, whereas good URL documents are marked by a regular globe icon.

To achieve high categorization quality, MetaSpider chooses to fetch from the Internet only "good" Web pages, selected by Web spiders running in a multithread mode designed for time efficiency. The robot exclusion protocol is also implemented such that the spiders will not access sites where the Web master indicates that robots are not welcome. Despite all our efforts to improve response time, the downloading process is inherently slow (largely influenced

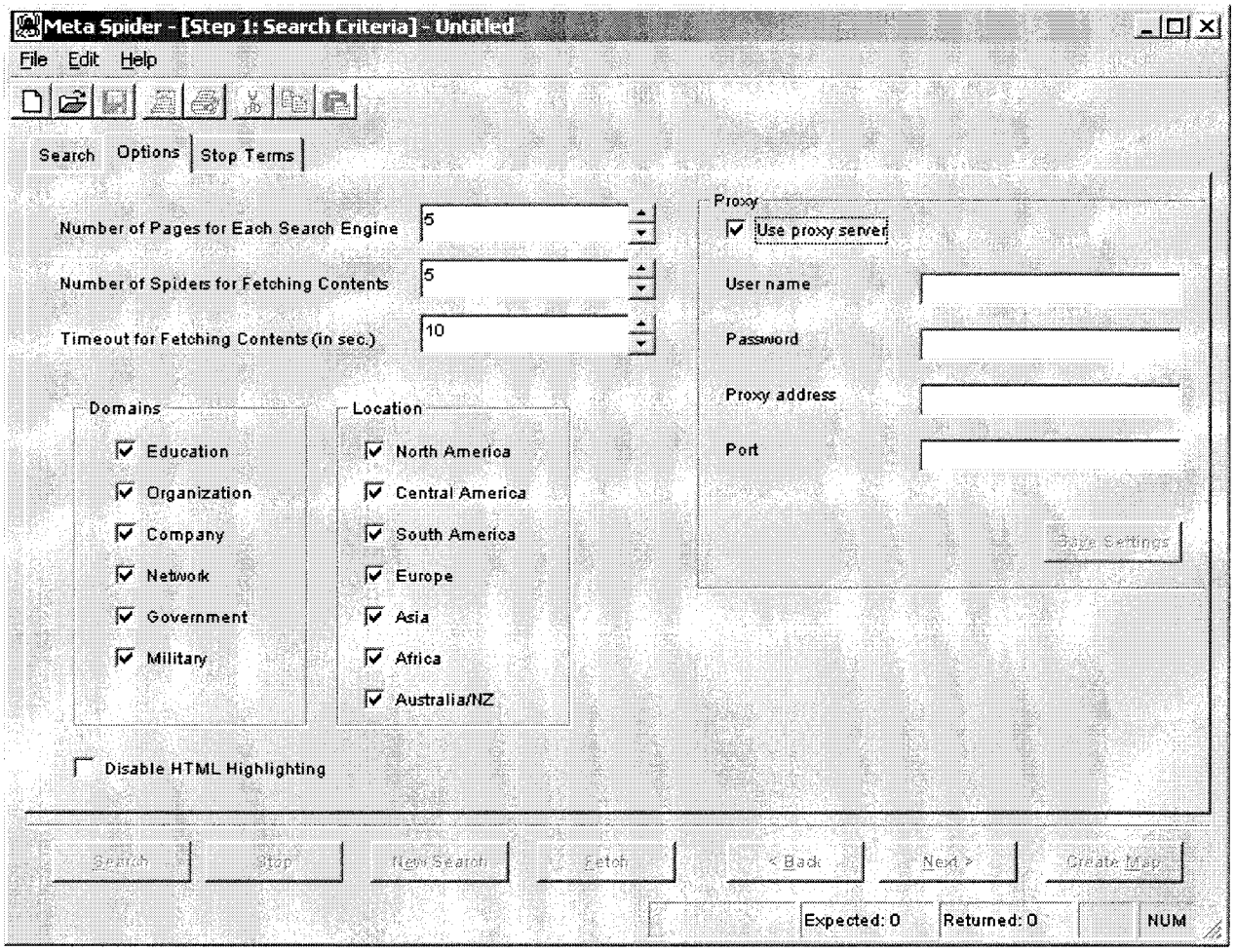

FIG. 3. MetaSpider: Spider options.

by server responsiveness and network connectivity). As a result, time responsiveness of MetaSpider, at a certain stage of search, is compromised. In our research, high clustering quality is given priority over speed. Zamir and Etzioni's research on *Grouper* (1999) shows the same design trade-off but a different design decision. Their approach clusters only the snippets returned by the search engines, therefore enabling quick response time.

3.4 Noun Phraser

Arizona Noun Phraser (AZNP), developed by our research group, is the tool used in MetaSpider to index the key phrases that appear in each document retrieved by the Internet spiders. It extracts all the noun phrases from each document, based on part-of-speech tagging and linguistic rules (Tolle & Chen, 2000). AZNP has three components. The tokenizer takes Web pages as text input and creates output that conforms to UPenn Treebank word tokenization rules by separating all punctuation and symbols from text without interfering with textual content. The tagger module assigns every word in the document a part-of-speech designation. The phrase generation module converts the words and associated part-of-speech tags into noun phrases by matching tag patterns to a noun phrase pattern established by linguistic rules. For example, the phrase "visual display terminal" would be considered a valid noun phrase because it matches the rule that an adjective-noun-noun sequence forms a noun phrase. The occurrence frequency of every phrase is recorded and sent to the user interface.

Phrases are listed in descending order of frequency, as shown in Figure 5. Clicking on any phrase, the user can view a list of documents that contain the key phrases from the entire document set retrieved, as shown in Figure 6. A further click on the document title will take the user to the actual Web site. By default, all these phrases will be sent to the SOM for automatic categorization. However, users are allowed to deselect any of the phrases.

3.5 Self-Organizing Map

In order to give users an overview of the set of Web pages collected, MetaSpider employs the Kohonen SOM to automatically cluster the Web pages collected into different regions on a 2-D map (Fig. 7) (Chen et al., 1998). Each region is labeled by the phrase best describing the key

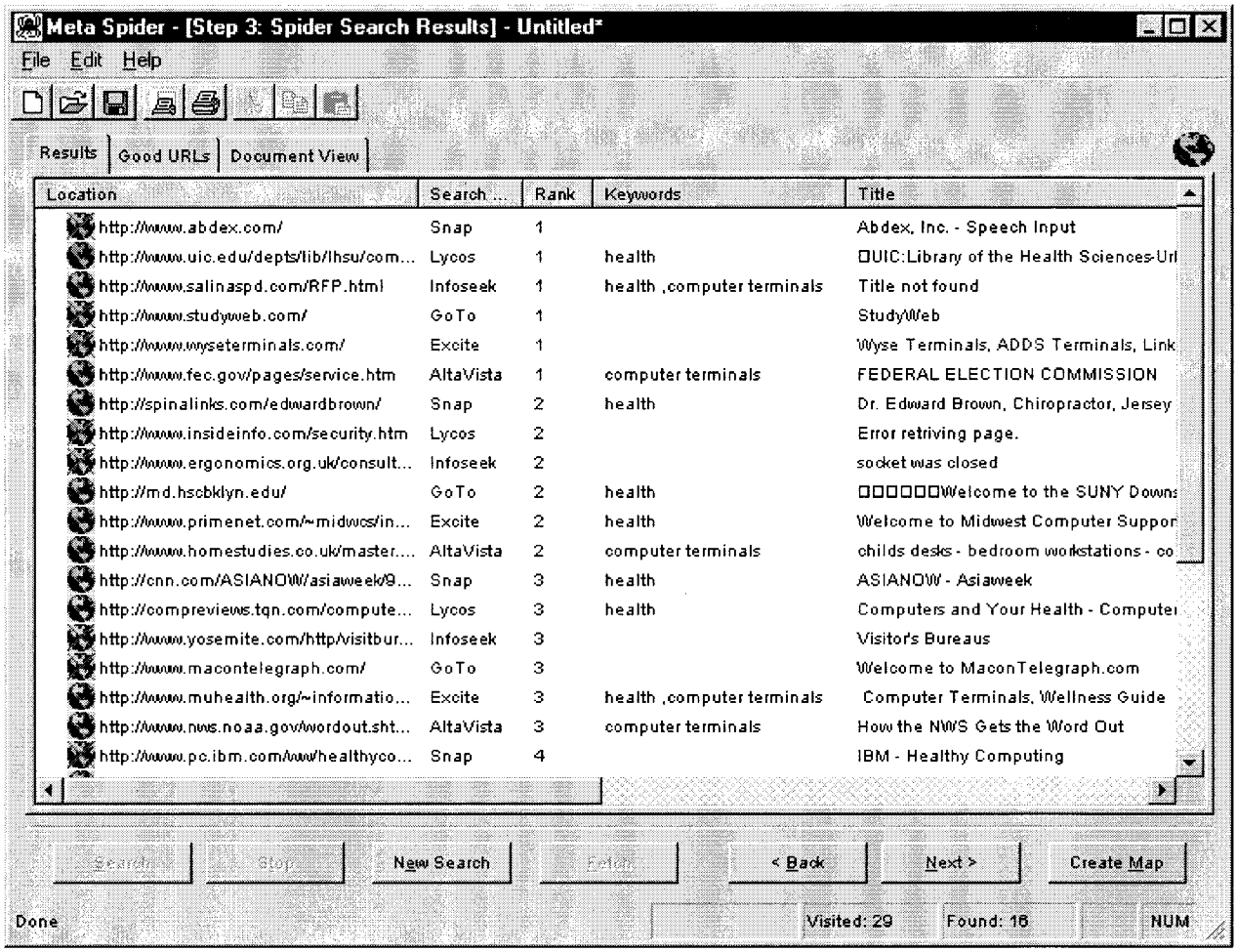

FIG. 4. MetaSpider: Fetching.

concept most representative of the cluster of Web pages in that region (e.g., "computer eye strain").

The SOM algorithm creates an intuitive, graphic display of important concepts contained in textual information (Lin et al., 1991; Orwig et al., 1997). In the context of Internet searching and browsing, the size of the region color block indicates the relative significance of the phrase to the documents collected. The relative proximity reveals the distance between the two concepts presented by the respective phrases.

Sophisticated users can tailor the AZNP by deselecting some of the trivial phrases to make the most sense of the map. For example, on the initial map created for "health and computer terminals," "health" and "computer terminals" take up the whole area. To allow other words or concepts to be seen, users can go back to the AZNP page and deselect such frequently appearing terms as "computer terminals" to permit inclusion of terms such as "repetitive stress injury," "eye strain," and "ergonomics," as shown in Figure 7. Self-organizing map provides very convenient browsing. It is especially helpful when the number of documents is large (Chen et al., 1996). The user can click on any of the labeled regions to go to the list of Web pages that contain the corresponding phrases.

4. Evaluation Methodology

4.1 Experimental Tasks

We conducted a user study to evaluate the proposed approach, implemented in the MetaSpider system. The main research question explored was, "Is MetaSpider effective and efficient in helping users locate useful information on the Web and understand the retrieved document set as a whole?"

Because MetaSpider has been designed to facilitate and integrate both document retrieval and automated categorization, traditional evaluation methodologies that treat document retrieval and categorization completely separately are not directly applicable. We have developed a new evaluation framework based on theme identification. Within this framework we have designed experimental tasks to permit evaluation of the extent to which combined document retrieval and categorization facilitate users' identification of major themes related to a certain topic in a given search session. This evaluation and the related evaluation methodology are among the intended research contributions of the research reported in this article.

We first report on how experimental tasks were generated. Six of the 50 topics created by the National Institute of Stan-

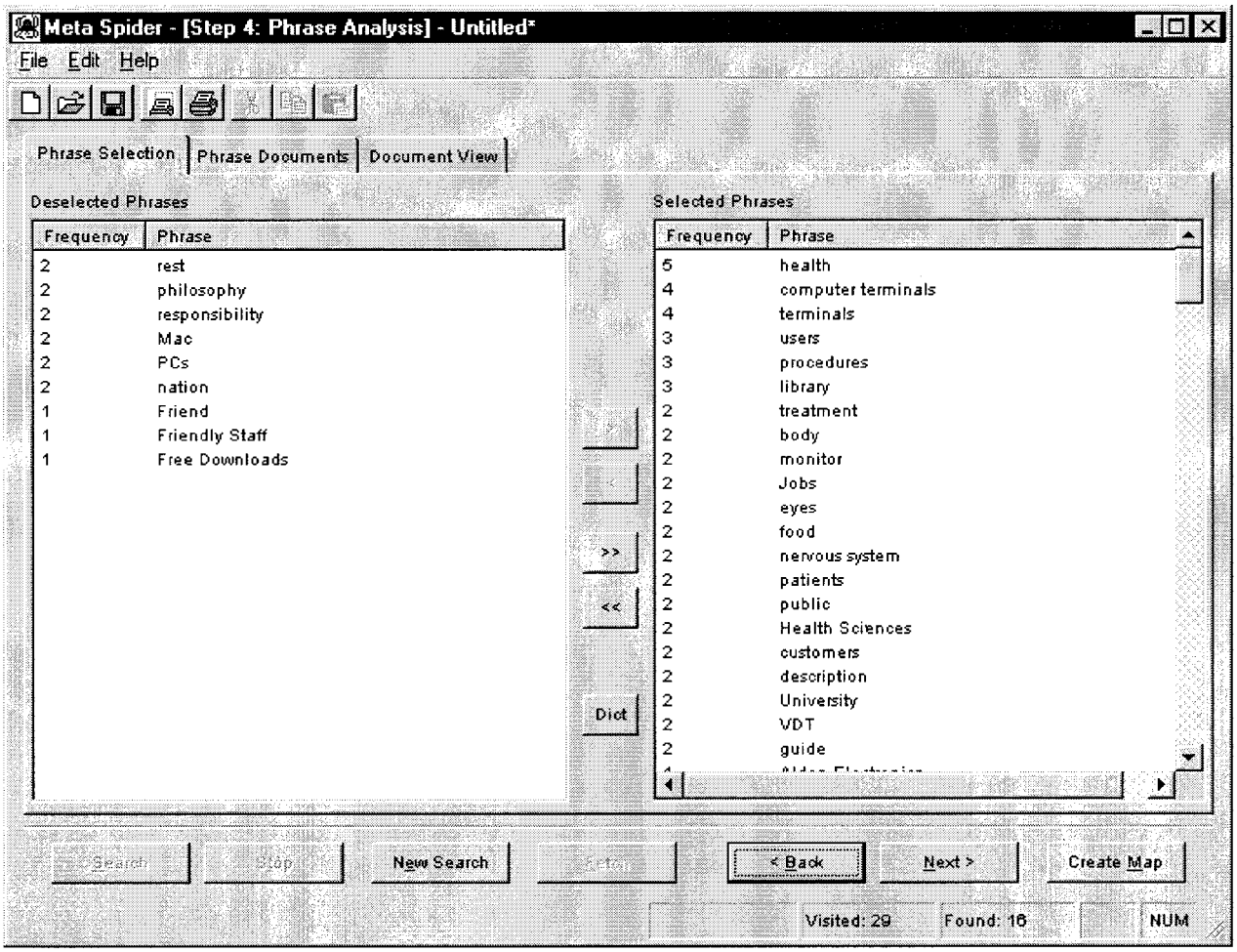

FIG. 5. MetaSpider: Phrase analysis.

dards and Technology (NIST) for the TREC-6 ad hoc task were selected and modified for use in the context of Web searching. The TREC (Text Retrieval Conference) series is sponsored by the NIST and the Defense Advanced Research Projects Agency (DARPA) to encourage research in IR from large text collections. We based our experimental tasks on six TREC topics for the following reasons. First, TREC strives to provide a common task evaluation that allows cross-system comparisons (Voorhees & Harman, 1998), which is consistent with our user study. Second, TREC-6 tasks have been well studied and many evaluation results can be found in the literature (Cormack, Palmer, & Clarke, 1998; Singhal, 1998), providing a solid foundation and reference framework for our research. Third, TREC-6 topics are amenable to iterative query construction methods, permitting users to look at individual documents retrieved by the ad hoc queries and then reformulate the queries based on the documents retrieved.

The six topics we used in our experiments were:

- Hubble telescope achievement
- Implant dentistry
- Radio waves and brain cancer
- Undersea fiber optic cable
- New fuel sources
- Health and computer terminals

Each of the topics defines an information need accompanied by a short description regarding the task, the domain involved, and the related questions. Here is an example:

Topic Title: *Health and computer terminals*

Description: *Is it hazardous to the health of individuals to work with computer terminals on a daily basis? What are the potential problems?*

Given such a search task, subjects are expected to summarize the findings of their Web searching or browsing (facilitated by the IR system being evaluated) as a number of themes. In our experiments, a theme was defined as "a short phrase, which describes a certain topic." Phrases like "repetitive stress injury" and "suppression of immune system" are examples of themes in our experiments. By examining the themes that subjects came up with using different search tools, we were able to evaluate how effectively and efficiently each IR system helped a user locate a collection

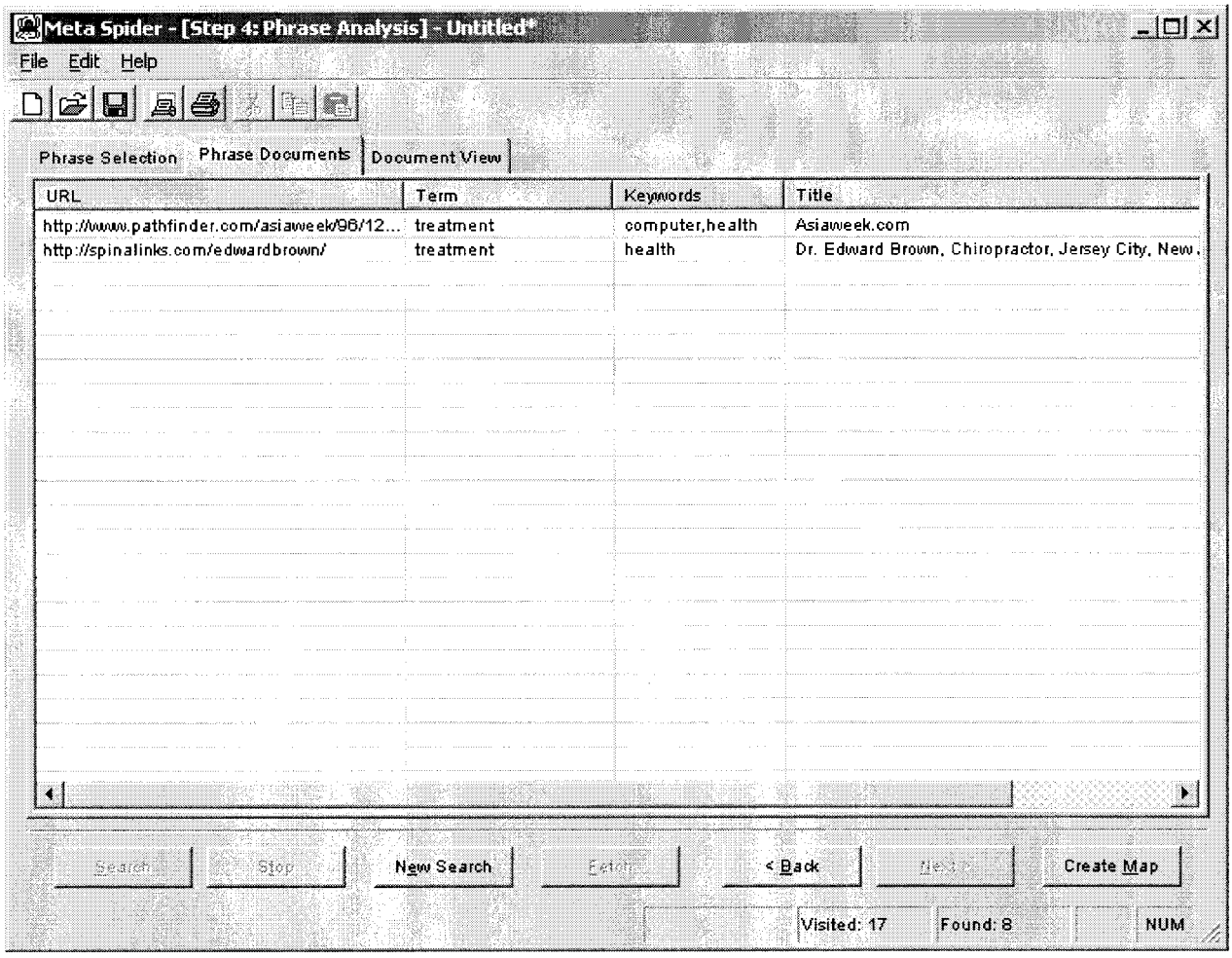

FIG. 6. Web pages grouped by key phrases.

of relevant documents from the Web and gain a general understanding of the returned documents.

4.2 Experimental Design and Hypotheses

In our experiments, MetaSpider was compared with two closely related IR systems: MetaCrawler and NorthernLight. MetaCrawler, developed at the University of Washington, is a widely used meta-search engine. It provides features such as the analysis of relevance rankings from source search engines, and the elimination of duplicates (Selberg & Etzioni, 1997). Recently, MetaCrawler has added the query expansion feature to assist users in formulating queries and narrowing the search scope by displaying related search topics. NorthernLight is a commercial search engine that organizes its retrieved documents in what are known as "custom search folders." Those folders are based on information type (e.g., press release, current news, special collection), subject, language, or source (e.g., government site, educational site). NorthernLight does not reveal the method used to create these folders.

Our study addressed the research question concerning whether MetaSpider effectively and efficiently supports the user's ability to locate useful information and to understand a retrieved document as a whole. We used precision and recall for theme identification as the primary measures of effectiveness as follows:

$$\textit{precision} = \frac{\textit{number of correct themes identified by the subject}}{\textit{number of all themes identified by the subject}}$$

$$\textit{recall} = \frac{\textit{number of correct themes identified by the subject}}{\textit{number of correct themes identified by expert judges}}$$

A theme is considered correct if the expert judges considered that it matched with one of the themes generated earlier by the same expert judges.

Because MetaSpider needs to fetch all the located Web pages from the Internet, it takes significantly more time than the other two systems in terms of the CPU time. However, the CPU time does not directly reflect the time needed by a

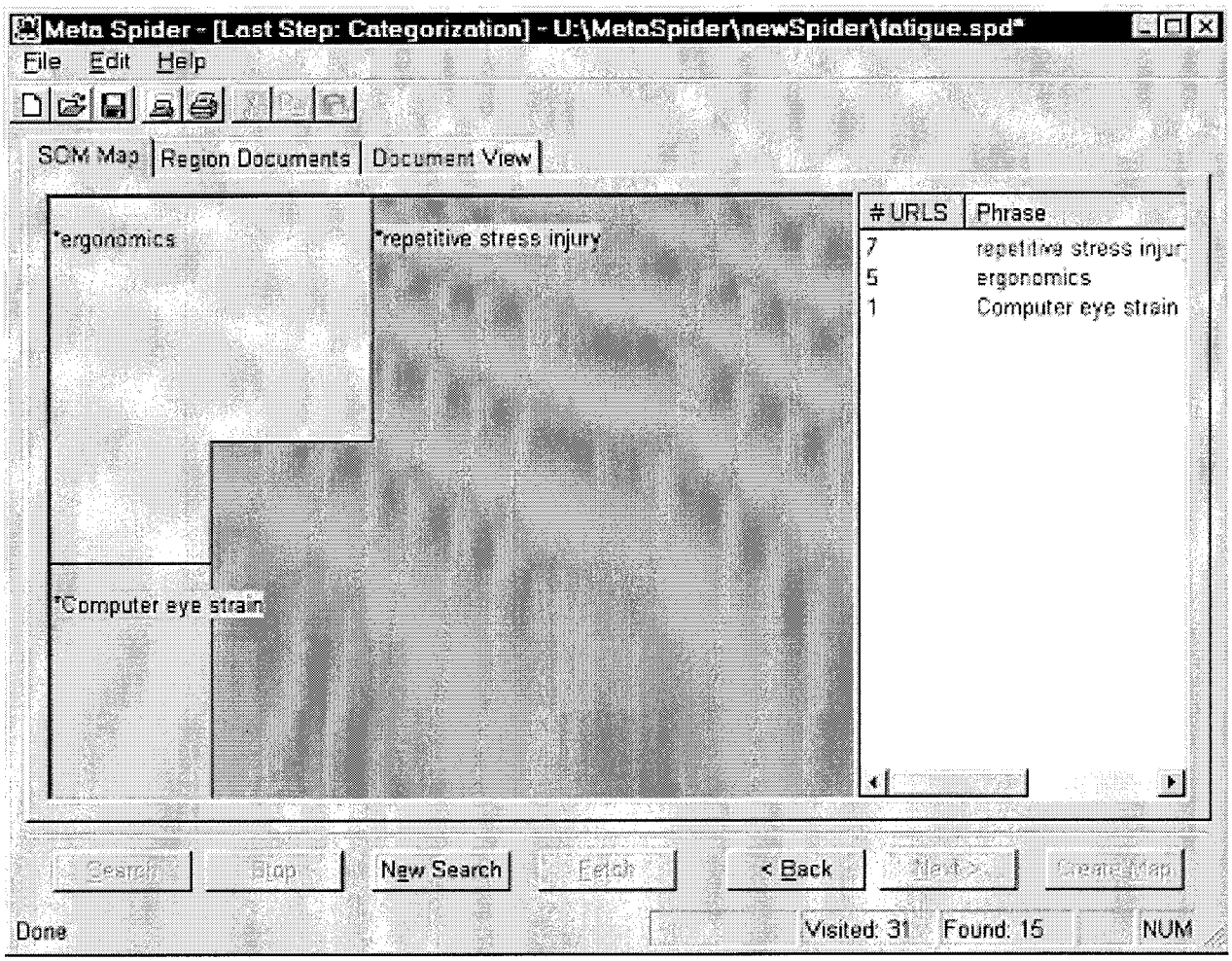

FIG. 7. Documents clustered into different categories in SOM.

user to understand the topic being searched, which is the purpose of this experiment. Therefore, the length of time of a complete user search session was used to measure efficiency. In addition, we recorded and analyzed several secondary measures, including the number of documents browsed and the number of switchings between actual Web documents and ranked document lists. An effective and efficient IR system should require the browsing of fewer documents and less-frequent switching.

The specific hypotheses examined in our user study were:

- *H1*: MetaSpider helps users achieve a higher precision for theme identification than MetaCrawler and NorthernLight.
- *H2*: MetaSpider helps users achieve a higher recall for theme identification than MetaCrawler and NorthernLight.
- *H3*: MetaSpider requires less user time for searching and understanding of Web documents than MetaCrawler and NorthernLight.
- *H4*: MetaSpider users require less manual browsing effort than MetaCrawler and NorthernLight users.

4.3 Subjects and Performance Measurement

Thirty undergraduate students from a junior-level computer programming class at the University of Arizona were recruited to participate in our experiments. Although these subjects are all computer literate, some of them have barely done any focused Web searches. In fact, we observed a variety of search strategies adopted by the subjects.

Subjects were assigned information search tasks (described in section 4.1) and required to jot down the themes they had identified after searching on a given IR system. To avoid possible previous influence, each subject was given three different search tasks and was instructed to perform each of them using a different IR system. To avoid a potential fatigue effect, we rotated the order in which each IR system was evaluated. In total, six search tasks were used. Although subjects were not given a specific time frame within which to perform the searches, they were encouraged to stop after 20 minutes (most subjects took less than 20 minutes to finish the tasks).

Two graduate students majoring in library science were recruited as experts to perform all six searches on all the three IR systems under investigation. After spending approximately 2 hours on each search query, these two experts compared notes and agreed upon a list of relevant themes. This list in turn was used as the reference set to compute precision and recall for theme identification of all information search sessions

TABLE 1. Experimental results.

	MetaSpider	MetaCrawler	NorthernLight
Precision			
Mean	0.816	0.697	0.561
Variance	0.281	0.315	0.402
Recall			
Mean	0.308	0.331	0.203
Variance	0.246	0.291	0.181
Time (min)			
Mean	10.93	11.13	11.00
Variance	4.04	4.72	5.30

performed by the subjects. The expert judges worked together to look at every theme generated by all the subjects and decided whether it matched with the reference set.

We collected and examined both quantitative and qualitative data. Quantitative data included precision and recall rates, which were used to measure effectiveness, and time spent on each search, which was used to measure efficiency. During the experiments, subjects were encouraged to think aloud, producing qualitative data that were then recorded and studied for major comments and observations. To permit further comparison of the three different systems, subjects also filled out questionnaires at the end of their search sessions.

5. Experimental Results and Analysis

5.1 Experimental Results

The main quantitative findings of our experiments are summarized in Table 1. Precision and recall rates were computed according to the definitions presented in section 4.2. Time was recorded as the total duration of the search task, including both the response time of the system and the browsing time of the subject. In this section, we discuss in detail these three main measures and present our analysis of two secondary measures: number of documents browsed and number of switching between Web pages and ranked document lists.

5.1.1 Precision

A pairwise *t*-test was applied to compare the three systems as to their precision rate performance. The mean precision rate of MetaSpider (0.816) ranked highest among the three. This confirms our first hypothesis (*H1)* that MetaSpider would achieve better precision than the other two IR systems. As shown in Table 2, the difference in precision between MetaSpider and NorthernLight was statistically significant (0.013), while the difference between MetaCrawler and MetaSpider was not.

A key contributory factor in the high precision achieved by MetaSpider was that MetaSpider performs real-time indexing and analysis and ensures that every page shown to the user contains the queried keywords. NorthernLight apparently performs no post-retrieval analysis but simply uses pre-defined categories based on off-line indexing. A careful review of the collected experimental data reveals supporting evidence in addition to statistics that suggests the superiority of MetaSpider in terms of precision. For instance, we found that among all 90 search sessions, Meta Spider had two zero-scores (the subject failed to produce any correct themes), MetaCrawler had two, and NorthernLight had seven.

Although MetaSpider had higher mean precision than MetaCrawler and delivered more consistent performance (lower variance), the difference was not statistically significant. We suspect two major factors contributed to this statistical insignificance. First, MetaCrawler has several features designed to assist the user in searching that MetaSpider does not offer. These features include summarization or description of each snippet and query refinement. Snippet descriptions enable the user to glimpse the general content of a document quickly. Query refinement features expand the original query to include other related topics, a very useful refinement for user who has no background knowledge about the search topic. For example, should a subject get stuck when trying to define and search for "new fuel source", related topics such as "alternative fuel" would be listed to provide critical hints for a better, alternative search query. Subjects were quoted as saying "those descriptions are very helpful" and "those suggested names helped me out." These desirable features helped MetaCrawler maintain a competitive high precision rate. Second, the MetaSpider indexing and categorization tools, i.e., Noun Phraser and SOM, were seldom used by subjects not familiar with them. Sixteen out of 30 subjects indicated that they rarely utilized Noun Phraser and SOM in their searches. Some users essentially stopped at the fetching phase and did not utilize the clustering tool at all.

5.1.2 Recall

A similar pairwise *t*-test was conducted to compare the performances of the three IR systems in terms of recall (Table 2). MetaCrawler ranked highest among the three. However, since none of the pairwise *t* tests showed statistical significance, no definite conclusions can be drawn on our hypothesis *H2*. We suspect that part of reason is the setting of certain experimental parameters. In theory, the SOM clustering approach works best with large document sets. However, concerned about possible time delay at the fetching phase, in our experiments we set the number of

TABLE 2. Pairwise *t*-test comparison.

	MetaSpider vs. MetaCrawler	MetaCrawler vs. NorthernLight	MetaSpider vs. NorthernLight
Precision	0.540	0.360	*0.013
Recall	1.000	0.139	0.304
Time	1.000	1.000	1.000

* The mean difference is significant at the 0.05 level.

documents returned from each search engine at five. This guaranteed timely responsiveness of the system but severely limited input to the post-retrieval clustering. As a result, the recall performance of MetaSpider may have been adversely affected.

5.1.3 Time

The *t*-test results show that the three search methods did not differ significantly in time requirements (Table 2). MetaSpider requires the least search time among the three systems, but *H3* is not confirmed as the differences are not statistically significant. As defined in the previous section, the time used for comparison is the total searching and browsing time. Real-time indexing and fetching, which usually takes more than 3 minutes, were also included in the total time for MetaSpider. In other words, we anticipate that MetaSpider requires less user time and effort in the whole search process, because the user only needs to browse the post-processed and categorized results.

5.1.4 Manual Browsing Effort

One of the assumptions of our experiments was that fewer documents browsed and a smaller number of times the user switched between actual Web pages and the document list indicate a need for less mental effort. These two measures have been collected and analyzed, as shown in Table 3. The number of documents browsed using either MetaCrawler or NorthernLight is greater than that using MetaSpider. Use of MetaCrawler or NorthernLight also required more switching than using MetaSpider. These figures to some extent support (although not statistically significantly) our hypothesis (*H4*) that MetaSpider provides a higher level of automation and requires less user browsing effort.

5.2 Strengths and Weaknesses of MetaSpider

Based on subjects' spontaneous reactions during searching and their general comments, we performed a verbal protocol analysis that revealed three main areas of user feedback: interface, searching, and clustering.

5.2.1 User interface

Subjects indicated that they liked the clean, simple design of MetaSpider. "The interface is clear-cut, looks very professional." Subjects also appreciated some of the design details, considering them convenient for browsing and intuitive to use. "I like the globe symbol with the red cross on it for bad pages. It is very straightforward."

However, the interface also received many suggestions from the subjects for improvement. For instance, the *back* and *next* buttons on MetaSpider received much criticism. The MetaSpider interface consists of five panels progressing from searching to generating the SOM. The *back* and *next* buttons are used for navigating through different panels, but not for switching windows in the same panel. Tabs such as *Good URL* and *Document view* accomplish within-panel window switching. Subjects who are accustomed to simple Web browser navigation found the distinction between within-panel and interpanel switching confusing.

TABLE 3. Comaprison of manual efforts.

	MetaSpider	MetaCrawler	NorthernLight
Documents browsed	5.36	7.98	5.86
Times of switching	3.23	4.20	4.16

Another useful suggestion relates to the display of words. Many subjects recommended keywords be highlighted with a different color and a "find keyword" function be implemented such that the user can quickly spot the keywords and surrounding text in the displayed document.

5.2.2 Searching

Subjects found it easier to locate useful information using MetaSpider than using other systems, especially when they were searching for multiple phrases. For example, subject 2 was not able to find any useful information using MetaCrawler after trying different query combinations for search task 3 (radio waves and brain cancer). His MetaCrawler queries included "brain cancer and radio wave," "brain cancer," "radio wave," "waves cancer," and "radio cancer." Subject 7's comments about MetaCrawler include "the result related to one word only, not the whole-phrase. . . [I] look up every single URL and guess." MetaSpider, on the contrary, provides better search support for multiple-phrase queries because MetaSpider looks for exact-phrase matching of the same word sequence. NorthernLight's "Custom Search Folder" seemed to offer somewhat similar support for phrase-based queries. However, as one of the subjects quickly pointed out, "it is not as effective as MetaSpider in grouping because folders limit choices; topics are not related to the keywords."

5.2.3 Clustering

Fourteen out of 30 subjects to varying degrees expressed preferences for using either the Noun Phraser or SOM. Ten subjects clicked on the terms on Noun Phraser to browse the Web pages represented by those terms. Other subjects commented that they liked the interactive analysis feature of SOM. Subject 16 was quoted as saying, "I like the fact that I can click on a map to go to relevant URLs." Subject 25 said, "It (MetaSpider) gives the best analysis of the search results." Subjects were particularly interested in the 2-D (map) display of search results. Among all search sessions, the two shortest ones were completed in 5 minutes and 8 minutes, respectively. In both these sessions, the subjects utilized either SOM or Noun Phraser, or both. Sometimes subjects obtained the themes directly from SOM or Noun Phraser.

Although the subjects who used SOM and Noun Phraser to analyze search results reported that the clustering tool had

improved their information search and theme formation, they were disappointed that SOM failed to give them complete visualization and comprehensive understanding of the documents collected. This can be partially attributed to the small number of documents fetched because the default number of documents returned by each search engine had been set to five. There were only a few documents for the clustering tool to work with after filtering, and we believe that the value and usefulness of clustering could be better demonstrated by working with large document collections. In general, subjects' overall opinion on MetaSpider tended to be positive. Although some subjects expressed confusion about the interface and complained about the time lag, many others commented positively about their experience with MetaSpider. For instance, some subjects claimed that "[MetaSpider] is so much easier" and that "I should have been allowed to do all the searches [using MetaSpider]."

5.3 Discussion

MetaSpider distinguished itself in our experiments for offering precise information. However, should this high precision be attributed to the two-tier filtering mechanism, to the high clustering quality, or to the combination? Seeking to answer this question, we compared the results of two searching strategies, namely, manual browsing and automatic categorizing.

In post-searching questionnaires, 14 out of the 30 subjects indicated that they found the analysis tools (including Noun Phraser and SOM) helpful. Sixteen considered being given a list of all valid pages (pages that contained the keywords) and actual Web pages to be more helpful. Two subjects thought that both the Web pages and analysis tools were helpful. This could imply that the subjects were more comfortable with traditional manual browsing than with automatic categorizing. Based on the original data collected, we performed some further analysis. We divided the subjects into two groups, those who preferred manual browsing and those who preferred using the categorizing tools provided by MetaSpider. We then compared the mean precision and mean recall achieved by each group, as shown in Table 4. Both the precision and recall levels of the group that preferred using the categorizing tools surpassed those of subjects who preferred manual browsing. In addition, a comparison based on the number of documents browsed also demonstrated that clustering using SOM and Noun Phraser required less user effort than manual browsing.

From our experiments, we observed that individual user's sophistication in using IR systems appeared to influence their degree of comfort in using the clustering tool. Experienced users who knew the frustrations of the so-called "art museum phenomenon" (browsing, but with little specific results) tended to show more interest and appreciation of the facilitation provided by the categorizing tool. Inexperienced users were more comfortable using the traditional ranked list display and browsing documents manually; they also were more conservative in exploring new system features.

TABLE 4. Comparison of manual browsing and automatic categorizing.

	Manual Browsing	Automatic Categorizing
Average precision	0.75	0.91
Average recall	0.28	0.34
Average no. of documents browsed	6.2	3.9

6. Conclusion and Future Directions

The research reported in this article is part of an ongoing effort to address Internet searching problems by integrating meta-search engines with textual clustering tools. As a meta-search engine, MetaSpider is designed to offer high precision by collating and further processing documents returned from primary search engines. Post-retrieval processing includes validation, indexing, and categorizing. After verifying the content of the returned Web pages, the Noun Phraser extracts all noun phrases from each document based on part-of-speech tagging and linguistic rules. The SOM automatically and in real-time clusters Web pages into different regions on a 2-D map to give the user a graphical overview of the whole document set. In addition, MetaSpider permits the user to fine-tune the categorization results in an iterative manner to gain different perspectives of the search results. MetaSpider can be downloaded from http://ai.bpa.arizona.edu/go/download/metaspider/index.html. As a client-side stand alone application, MetaSpider can be easily installed and run, and contains useful features such as saving of user search sessions and caching of past search results.

We present in this article our user evaluation of MetaSpider. We have developed an evaluation framework based on themes of search topics. The initial evaluation results are promising. They have shown that MetaSpider performs better in precision when compared with several widely used meta-search systems. Because of its built-in automatic indexing and categorizing components, MetaSpider greatly reduces the manual effort required of the user for Web searching and browsing.

For ongoing and future research, we are in the process of extending MetaSpider vertically so that it provides in-depth information support for specific domains. One such system currently under development is a MedSpider specialized in the medical domain and capable of querying and aggregating authoritative medical databases. We are also actively developing multi-agent, collaborative MetaSpider to be used in group settings. Another research direction that we are pursuing is related to developing a multi-lingual MetaSpider through replacing the current English Noun Phraser with a multi-lingual indexer.

Acknowledgments

The MetaSpider project was mainly supported by:

- NSF Digital Library Initiative-2, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management," IIS-9817473, April 1999–March 2002.
- NSF/CISE/CSS, "An Intelligent CSCW Workbench: Personalized Analysis and Visualization," IIS-9800696, June 1998–June 2001.

We would like to thank all the subjects at the University of Arizona who participated in the user study. We would also like to thank the members of the University of Arizona Artificial Intelligence Lab, particularly Wojciech Wyzga, Hadi Bunnalim, Ye Fang, Ming Yin, and Bill Oliver, for their insightful suggestions and programming support.

References

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

Chen, H., Schufels, C., & Orwig, R. (1996). Internet categorization and search: A self-organizing approach. Journal of Visual Communication and Image Representation, 7, 88–102.

Chignell, M.H., Gwizdka, J., & Bodner, R.C. (1999). Discriminating meta-search: A framework for evaluation. Information Processing and Management, 35, 337–362.

Cormack, G.V., Palmer, C., & Clarke, C. (1998). Efficient construction of large test collections. In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98). New York: ACM Press.

Garman, N. (1999). Meta search engines, ONLINE. [On-line]. Available at http://www.onlineinc.com/onlinemag/OL1999/garman5.html

Hearst, M. (1995). TileBars: Visualization of term distribution information in full text information access. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '95) (pp. 59–66). New York: ACM Press.

Hearst, M., & Pedersen, J.O. (1996). Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96) (pp. 76–84). New York: ACM Press.

Kohonen, T. (1995). Self-organizing maps. Berlin: Springer-Verlag.

Kohonen, T. (1997). Exploration of very large databases by self-organizing maps. In Proceedings of the IEEE International Conference on Neural Networks, 1 (pp. PL1–PL6). IEEE.

Lawrence, S., & Giles, C.L. (1999). Accessibility of information on the Web. Nature, 400, 107–109.

Leighton, H.V., & Srivastava, J. (1999). First 20 precision among World Wide Web search services (search engines). Journal of the American Society for Information Science, 50, 870–881.

Leuski, A. (1998). Evaluating a visual presentation of retrieved documents. CIIR Technical Report [On-line]. Available at: http://ciir.cs.umass.edu/

Leuski, A., & Allan, J. (1999). The best of both worlds: Combining ranked list and clustering. CIIR Technical Report [On-line]. Available at: http://ciir.cs.umass.edu/

Lin, X. (1997). Map displays for information retrieval. Journal of the American Society for Information Science, 48, 40–54.

Lin, X., Soergel, D., & Marchionini, G. (1991). A self-organizing semantic map for information retrieval. In Proceedings of the 14th International ACM SIGIR Conference on Research and Development in Information Retrieval Retrieval (SIRIR '91) (pp. 262–269). New York: ACM Press.

Morgan, L. (1999, May). Make Web searches more powerful. InternetWeek, 766. [Online]. Available at http://www.internetwk.com/reviews/rev052499-3.htm

Orwig, R., Chen, H., & Nunamaker, J.F. (1997). A graphical self-organizing approach to classifying electronic meeting output. Journal of the American Society for Information Science, 48, 157–170.

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

Selberg, E., & Etzioni, O. (1995). Multi-service search and comparison using the MetaCrawler. In Proceedings of the 4th World Wide Web Conference, Boston, Mass, USA, December 1995.

Selberg, E., & Etzioni, O. (1997). The MetaCrawler architecture for resource aggregation on the Web. IEEE Expert. IEEE.

Singhal, A. (1998). AT&T at TREC-6. In Proceedings of the Sixth Text Retrieval Conference (TREC-6) (pp. 215–226). Gaithersburg, Maryland: National Institute of Standards and Technology.

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

Van Rijsbergen, C.J. (1979). Information retrieval (2nd ed.). London: Butterworths.

Veerasamy, A., & Belkin, N.J. (1996). Evaluation of a Tool for Visualization of Information Retrieval Results. In Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96) (pp. 85–92). New York: ACM Press.

Voorhees, E., & Harman, D. (1998). Overview of the Sixth Text Retrieval Conference (TREC-6). In Proceedings of the Sixth Text Retrieval Conference (TREC-6) (pp. 1–24). Gaithersburg, Maryland: National Institute of Standards and Technology.

Zamir, O. (1998). Visualization of search results in document retrieval systems. Unpublished General Examinations Paper, University of Washington, Seattle.

Zamir, O., & Etzioni, O. (1999). Grouper: A Dynamic Clustering Interface to Web Search Results. In Proceedings of the Eighth World Wide Web Conference, Toronto, May 1999.