Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques

Hsinchun Chen and Andrea L. Houston
Artificial Intelligence Lab, Management Information Systems Department, University of Arizona, Tucson, AZ 85721. E-mail: hchen@bpa.arizona.edu; ahouston@bpaosf.bpa.arizona.edu

Robin R. Sewell
Management Information Systems Department, University of Arizona, Tucson, AZ 85721. E-mail: rrs@ai2.bpa.arizona.edu

Bruce R. Schatz
Digital Library Research Program, University of Illinois at Urbana-Champaign, Urbana, IL 61801. E-mail: schatz@csl.ncsa.uiuc.edu

The Internet provides an exceptional testbed for developing algorithms that can improve browsing and searching large information spaces. Browsing and searching tasks are susceptible to problems of *information overload* and *vocabulary differences.* Much of the current research is aimed at the development and refinement of algorithms to improve browsing and searching by addressing these problems. Our research was focused on discovering whether two of the algorithms our research group has developed, a Kohonen algorithm category map for browsing, and an automatically generated concept space algorithm for searching, can help improve browsing and/or searching the Internet. Our results indicate that a Kohonen self-organizing map (SOM)-based algorithm can successfully categorize a large and eclectic Internet information space (the Entertainment subcategory of Yahoo!) into manageable sub-spaces that users can successfully navigate to locate a homepage of interest to them. The SOM algorithm worked best with browsing tasks that were very broad, and in which subjects skipped around between categories. Subjects especially liked the visual and graphical aspects of the map. Subjects who tried to do a directed search, and those that wanted to use the more familiar mental models (alphabetic or hierarchical organization) for browsing, found that the map did not work well. The results from the concept space experiment were especially encouraging. There were no significant differences among the precision measures for the set of documents identified by subject-suggested terms, thesaurus-suggested terms, and the combination of subject- and thesaurus-suggested terms. The recall measures indicated that the combination of subject- and thesaurus-suggested terms exhibited significantly better recall than subject-suggested terms alone. Furthermore, analysis of the homepages indicated that there was limited overlap between the homepages retrieved by the subject-suggested and thesaurus-suggested terms. Since the retrieved homepages for the most part were different, this suggests that a user can enhance a keyword-based search by using an automatically generated concept space. Subjects especially liked the level of control that they could exert over the search, and the fact that the terms suggested by the thesaurus were "real" (i.e., originating in the homepages) and therefore guaranteed to have retrieval success.

1. Introduction

The Information Age has produced a wealth of information which is supposed to be readily available to anyone who wishes to use it. Indeed with the increased popularity of online services, more people have access to more information than ever before, and that information appears to be growing at an exponential rate. This sheer volume of information is often overwhelming to users, a phenomenon commonly referred to as "information overload" (Blair & Maron, 1985).

Individuals seeking to use this information typically have one of two goals in mind. They either want to explore the information space, to gain familiarity with it and to locate something of interest to them, or they want to search the information space (either for information on a given topic or for a specific piece of information), and retrieve the relevant information. To successfully accomplish this, users need to know two things. First, they must have a working knowledge of the system where the information is stored, in particular how to navigate through the information system (Chen & Dhar, 1990) and of how the information is organized or categorized. Second, they must have a knowledge of the subject of interest, in particular the vocabulary of the subject domain. Users with different levels of subject expertise and system familiarity (Chen, 1994; Furnas, 1982; Furnas, Landauer, Gomez, & Dumais, 1987), combines with the often imprecise nature of human language to create the difficulty of information browsing, searching, and retrieval that is often called the ''vocabulary (differences) problem'' (Chen, 1994; Courteau, 1991; Frenkel, 1991).

The Internet, with its tremendous diversity and volume of information, provides an especially interesting and challenging testbed to investigate solutions to the problems of ''information overload'' and ''vocabulary differences.'' Today's Internet explorers, often called ''surfers,'' have high expectations of Internet services. These range from a simple desire to find something interesting (for no particular reason) to searching for useful information on a specific topic. Internet service providers have responded to such expectations by increasing the size of their indexed World Wide Web (WWW) homepages, and by improving their Internet searching engines (focusing initially on improving user interfaces and query formation). The Lycos server at CMU claims to have indexed 50+ million URLs as of August 1996. Alta Vista (through a strategic alliance with Yahoo!) claims to have indexed 30+ million URLs as of August 1996. HotBot (a recently introduced Internet searching product from Hot Wired) claims to have indexed the largest number of homepages, 54 million URLs as of August 1996.

The University of Arizona has developed several techniques to help improve the organization and categorization of large volumes of information in order to assist users in overcoming both the problems of *information overload* and *vocabulary differences* while browsing and searching large information spaces. One technique, based on an automatically generated thesaurus or concept space, has been successfully used in several environments, including a molecular biology community system (Chen, Martinez, Ng, & Shatz, 1997a; Chen, Schatz, Yim, & Fye, 1995) and the Illinois Interspace Digital Library Initiative Project (Chen et al., 1996a; Schatz & Chen, 1996; Schatz et al., 1996). The second technique, based on the use of a Kohonen self-organizing map (SOM) algorithm, has been used experimentally within the electronic meeting system environment (Orwig, Chen, & Nunamaker, 1997) and for Internet homepage categorization (Chen, Schuffels, & Orwig, 1996b). In this project, we were interested in exploring the usability of the two techniques. In particular, we wanted to compare the performance of these techniques with a widely used Internet browsing service (i.e., the hierarchically structured Yahoo!) and a commonly used Internet searching method (i.e., a keyword-based search).

In Section 2, we review the current status of Internet browsing services and briefly discuss the implementation of the SOM-based method to categorize a set of about 110,000 entertainment-related WWW homepages. In Section 3, we review the current status of Internet searching and the thesaurus or concept space approach to the ''vocabulary differences'' problem, and briefly discuss the implementation of an automatically generated thesaurus on the same set of WWW homepages. Our experimental design is presented in Section 4. Section 5 presents the qualitative results of the browsing task experiment. Section 6 presents both the quantitative results and the qualitative results of the searching task experiment. Conclusions are presented in Section 7.

2. Internet Browsing

2.1. Background

Several researchers have defined browsing behavior in the context of hypertext environments (Carmel, Crawford, & Chen, 1992; Liebscher & Marchionini, 1988; Marchionini, 1987; Marchionini & Shneiderman, 1988). For our research, browsing can best be described by a combination of quotes from Marchionini's work: Browsing is ''an exploratory, information seeking strategy that depends upon serendipity'' and is ''especially appropriate for ill-defined problems and for exploring new task domains'' (Marchionini & Shneiderman, 1988, p. 71). Browsingis ''characterized by the absence of planning'' (Marchionini, 1987, p. 69) and is often used as ''an alternative to the complex Boolean (keyword-based) search strategy (Marchionini, 1987). In essence, browsing explores both the organization or structure of the information space *and* its contents.

Because browsing is frequently used in new or relatively unknown (unexplored) information spaces, users typically rely on pre-existing mental models of information organization as they explore. Mental models are defined as ''cognitive representations of a problem [or information] situation or system'' (Marchionini & Shneiderman, 1988, p. 72). They help a user represent the content, structure, and relationship of information in the information space. This, in turn, helps the user to understand the organization of the space, draw inferences about navigating through the space, and respond to conditions that occur during navigation. For example, when examining a new book or magazine, users typically scan the table of contents or index to get an overall sense of what the book/magazine is about and what topics it covers. When browsing in a bookstore or video store, users typically scan the aisles that contain a favorite genre (e.g., mystery, science fiction, action), or look for the work of a favorite author, actor/actress, or director. Similar behavior occurs when users search online journal indexes for new articles of interest: A favorite subject area or author is entered and the browse is limited to a given time period (i.e., most recent month or year). The most common mental models are either alphabetically-based or based on hierarchical categories (where the categories used are determined by the subject domain and media).

Similarly, users have developed mental browsing models for the Internet which are based on the initial structures developed by Internet information browsing services and searching tools offered by Internet software providers. We have identified two major approaches:

- *Hypertext Browsing.* Hypertext browsing services support Internet browsing by providing links between keywords and topics embedded in the text that can be explored at will by the user. The most common examples of Internet hypertext browsing services are: *NCSA Mosaic, Netscape Browser,* and *Microsoft's Internet Explorer.* Problems arise when a user's mental model of the information space does not conform to that of the author of the hypertext document and/or that of the designer of the information space (Carlson & Ram, 1990). A user with this problem, known as the *embedded digression problem,* can easily become disoriented, lost, and confused. As a result, the user can spend a great deal of time wandering around the information space while learning nothing of interest, a situation known as the *art museum phenomenon* (Foss, 1989).

- *Directories.* One method of improving the efficiency of exploring a large information space is to partition it into distinct subject categories that are meaningful to users. Categorization and subject classification are common practices in library and information sciences (e.g., the INSPEC database for the computer engineering domain, the ERIC database for sociology, etc.). Subject partitioning creates smaller databases which can be more efficiently explored. Furthermore, the subject directory can be used by explorers in ''directory-browsing.'' Directory browsing on the Internet is a user-guided information-seeking behavior exemplified by the previously popular Gopher information servers. Gopher users would connect via Gopher to WWW sites of interest and browse the available directories at that site. If a directory appeared to be of interest, it could be explored in more depth. This is an example of a hierarchically organized directory.

Yahoo! was the first Internet browsing/searching service to offer a directory that partitions the Internet information space by providing meaningful subject categories (e.g., science, entertainment, business, etc.). More recently, Lycos, one of the largest Internet service providers, added a directory service to its traditional keyword-based searching. While partitioning the information space via subject or topic categories can improve exploration by making it more efficient, this method is not without its own set of problems. The most common problems are: 1) The categories are limited in their granularity, and timeliness, and 2) the process of creating the categories and connecting homepages to them is manual, slow, and cumbersome. We believe that this method can be improved by incorporating an intelligent, automatic categorization algorithm, such as the Kohonen self-organizing feature map, as part of the directory creation process.

2.2. A Kohonen SOM Approach to Internet Browsing

Quillian (1968), originally suggested that semantic networks could be used to encode and associate word meanings and therefore could be used to visualize or construct mental models of the information space. Neural network algorithms, in particular, are a natural starting point for organizing large amounts of information in a manner consistent with human mental models. After examining several neural network algorithms in previous research (Lippmann, 1987), our research group concluded that a variant of the Kohonen self-organizing feature map (SOM) appears to be the most promising algorithm for organizing large volumes of information. The algorithm can be used to create an intuitive, graphical display of the important concepts contained in textual information (Chen et al., 1996b; Orwig et al., 1997). This research reports our investigation of the use of a Kohonen SOM as a categorization technique for the Internet information space.

Lin (1991) was the first to adopt the Kohonen SOM for information retrieval. In his prototype, self-organizing clusters of important concepts in a small database of several hundred documents were generated. A scaleable multi-layered, graphical SOM approach to Internet categorization was developed in our previous research (Chen et al., 1996b) and the resulting prototype was tested for usability in this experiment. The prototype was developed using only a portion of the Internet, the Yahoo! Entertainment sub-category (about 110,000 homepages) and hence is called the ET-Map. For more details about the Kohonen SOM algorithm, our modifications and the development of the Internet prototype, the reader is referred to Chen et al. (1996b).

3. Internet Searching

3.1. Background

As Internet services based on the WWW have become more popular, information overload has become an increasingly pressing research problem (Bowman, Danzig, Manber, & Schwartz, 1994). The Internet human–computer interaction paradigm has shifted from a simple hypertext-like *browsing* interaction to a content-based *searching* one. Searching is characterized as a process in which a user describes a request via a query and the system must locate information that matches or satisfies the request. Many researchers and practitioners have considered Internet searching to be one of the most urgent, challenging, and rewarding areas of research for future National Information Infrastructure (NII) applications. For example, Internet searching has been the hottest topic at World Wide Web Conferences (Bowman, 1994; DeBra & Post, 1994; Pinkerton, 1994).

Based on our analysis, there are two basic approaches to searching on the Internet:

• *Keyword Search.* In this approach, users enter a keyword or set of keywords that, in their opinions, best characterizes their information needs. The information system translates this request into a query and searches the information space for appropriate matches, which are returned. Advanced keyword searching allows users to enter more than one keyword and to relate multiple keywords to each other via the use of Boolean operators (''AND,'' ''OR,'' and ''NOT''). Some sophisticated versions even allow users to assign different weights to each of the multiple input keywords. Keyword searching is provided by the following Internet searching services: *Alta Vista* (developed by Digital Equipment Corporation), *Excite*, *Open Text*, and *HotBot* (which uses the Inktomi search engine).

• *Combined Keyword Search and Categorization.* Some searching engines allow the user to further refine a keyword search by restricting it to a given directory or sub-division of the entire database. This is more efficient than searching the entire database but, as a consequence, the user is unable to identify relevant information that may exist outside of the directory chosen. Combined keyword searching and categorization is provided by the following Internet searching services: *Lycos* (at CMU, which includes access to Point and A2Z), *Yahoo!, Infoseek, Magellan*, and *Web Crawler* (America OnLine Inc.). All of these services have an association with a directory service which manually creates categories by subject or topic, and groups homepages under the appropriate category. Searching tasks can be limited to a given category by using the directory, or the entire database of indexed homepages can be searched.

Sophisticated Internet searchers are now requesting subject categories that are more up-to-date or real-time, and more fine-grained. Internet searching services are responding to this demand with features like highlighting new categories and newly obtained homepages, and creating special categories like ''What's Hot,'' and ''What's New.''

Searching the Internet is also vulnerable to the vocabulary differences problem (also referred to as the semantic barrier; Nadis, 1996). This problem has been extensively studied over the years by cognitive psychologists and information scientists (Furnas, 1982; Furnas et al., 1987). Furnas et al., found that in spontaneous word choice tasks for objects in five domains, two people favored the same term with less than 20% probability. This fundamental property of language limits the success of various design methodologies for vocabulary-driven interaction. In information science, the existence of vocabulary differences has been recognized as one of the primary sources of information retrieval problems.

Previous research (Bates, 1986) has shown that different indexers, well-trained in an indexing scheme, often assign different index terms to the same document. It has also been observed that an indexer may use different terms for the same document at different times (possibly because of learning or the cognitive state of mind at indexing). Searchers also tend to use different terms for the same information sought. Because of the indeterminism involved in indexing and searching, an exact match between the searcher's terms and those of the indexer is unlikely (Chen & Dhar, 1987), contributing to poor recall and poor precision in searching tasks.

There are several approaches to helping the user improve searching tasks. They include: Query expansion (Ekmekcioglu, Robertson, & Willett, 1992; Peat & Willett, 1991; Smeaton & van Rijsbergen, 1983); relevance feedback (Green, 1995; Green & Bean, 1995; Heine, 1995); Multidimensional Scaling (McCain, 1995), including Metric Similarity Modeling (Bartell, Cottrell, & Belew, 1995) and Latent Semantic Indexing (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990); and thesauri use (Crouch, 1990; Voorhees, 1993; Wang, Vandendorpe, & Evens, 1985). A more thorough discussion of these approaches can be found in Chen, Zhang, & Houston (1997b). We have focused on thesauri use as it provides one solution to the vocabulary differences problem by providing users with synonyms for the keywords they have chosen or providing memory triggers for other keywords that can better describe a searching task. There have been two approaches to using thesauri to improve keyword-based searches:

• *Incorporating Existing Thesauri.* Many research groups have created vocabulary-based search aids for online information retrieval systems by making use of existing thesauri or dictionaries to provide alternate terms to use in searching. Thesauri, in particular, exhibit a structure similar to human word-association networks. Some recent examples include: The National Library of Medicine's Unified Medical Language System (UMLS) project (Lindberg & Humphreys, 1990; McCray & Hole, 1990); the Vocabulary Switching System (VSS), (Chamis, 1991); Knapp's BRS/TERM vocabulary database (Knapp, 1984); the NTIS thesaurus (Piternick, 1984); the Art and Architecture Thesaurus (AAT), (Petersen, 1983, 1990); the Genentech library (Bellamy & Bickham, 1989); the ''relational thesauri'' (Fox, 1987; Fox, Nutter, Ahlswede, Evens, & Markowitz, 1988); and the ''lexical database'' (Ahlswede & Evens, 1988). Other researchers involved in thesaurus work include: Chaplan (1995), and Niehoff and associates at Battelle Columbus Laboratories (Niehoff, 1976; Niehoff & Kwansy, 1979).

While these tools are able to provide the searcher with alternate terms, they do not overcome the *knowledge acquisition bottleneck* (Hayes-Roth, Waterman, & Lenat, 1983)—the cognitive demand upon humans (indexers or domain experts) to *manually* create thesauri or dictionaries. A reasonable alternative approach therefore is based on *automatic* thesaurus generation.

• *Automatic Thesaurus Generation.* Numerous investigators have developed algorithmic approaches to automatic thesaurus generation. Automatic thesaurus generation techniques provide all of the advantages of thesauri-aided searching and the further advantage of reducing or eliminating the labor-intensive effort of thesaurus creation. Most of these approaches employ techniques that compute coefficients of ''relatedness'' between terms using statistical co-occurrence algorithms (e.g., cosine, Jaccard, Dice similarity functions, EMIM, etc.) (Chen & Lynch, 1992; Crouch, 1990; Rasmussen, 1992; Salton, 1989). Some algorithms perform cluster analysis to further group terms of similar meanings (Everitt, 1980; Rasmussen, 1992). Chen et al., have conducted a series of experiments which included several large-scale, domain-specific, automatically generated thesauri (Chen & Dhar, 1991; Chen et al., 1996a, 1997a; Chen, Lynch, Basu, & Ng, 1993; Chen & Ng, 1995).

Our research group has proposed a concept space approach to Internet information searching and retrieval. By analyzing the co-occurrence probabilities of keywords in homepages of specific subject categories (or sub-directories), a *concept space* which represents the important terms in that sub-directory and their weighted relationships in a graphical structure can be automatically created, akin to an associative man-made thesaurus. In a recent experiment involving an electronic community system and actual molecular biologists, a system-generated (nematode) worm concept space was shown to be an excellent ''memory-jogging'' tool that supported learning and serendipitous browsing (Chen et al., 1997a). Despite some occurrences of obvious noise, the system was useful in suggesting relevant concepts for the researchers' queries and improved concept *recall* (Chen et al., 1995). The usefulness of the concept space or thesaurus approach to accessing the diverse and large-scale Internet information space remains to be tested, hence the interest in the second part of our experiment.

3.2. An Automatically Generated Thesaurus Approach to Internet Searching

The specific steps and algorithms that were adopted to create our Entertainment Thesaurus (concept space) included: *Document collection, automatic indexing, co-occurrence analysis,* and *associative retrieval.* A brief overview of these techniques, in the context of our Internet searching experiment, is presented below. For a more complete discussion of the technique and algorithmic details, readers are referred to Chen et al. (1996a, 1997a).

• *Document Collection.* In any automatic thesaurus building effort, the first task is to identify the collection of documents in specific subject domain(s) that will serve as the basis of the thesaurus. In creating the Entertainment Thesaurus, we collected documents from a single source: The already partitioned Entertainment sub-directory in Yahoo!. The homepage collection was done using a breadth-first search (BFS) spider which began the searching process with all the homepages linked to the Yahoo! Entertainment sub-directory. The entire process of extracting 110,000 entertainment-related homepages lasted about 3 days using our HP 735 workstation (on T1 connection).

• *Automatic Indexing.* The purpose of this step is to automatically identify the content of each homepage collected in the previous step. Based on a revised automatic indexing technique (Salton, 1989), subject descriptors on each homepage are identified, and the number of times that each descriptor appears in the entire collection of homepages is computed. Then, a stop-word list is used to remove non-semantic-bearing words (e.g., ''the,'' ''a,'' ''on,'' ''in''). After removing the stop words, a stemming algorithm is used to identify the word stem for the remaining words. In our process, the stop-word list was also applied to all of the stemmed words. In the case of the Internet Entertainment Thesaurus, the stop-word list contained roughly 3,500 words. Incidental or ''noisy'' descriptors were further removed by setting a term-frequency threshold (typically two or three occurrences of a term in the entire collection) for a homepage.

• *Co-occurrence Analysis.* The importance of each descriptor or term in representing the content of the entire homepage varies. Using term frequency and inverse homepage frequency, the cluster analysis step assigns weights to each term on a homepage to represent the term's level of importance. Term frequency measures how often a particular term occurs in the entire collection. Inverse homepage frequency indicates the specificity of the term and allows terms to acquire different strengths or levels of importance based on their specificity. A term can be a one-, two-, or three-word phrase.

Cluster analysis then converts this raw data (indexes and weights) into a matrix that shows the similarity and dissimilarity of the terms by a distance function. The distance function used in this step is based on the asymmetric ''Cluster Function'' developed by Chen and Lynch (1992) and previously shown to better represent term association better than the popular cosine function. Using it, a net-like concept space of terms and their weighted relationships was created.

• *Associative Retrieval.* Our previous research produced two associative retrieval algorithms, one based on the serial branch-and-bound algorithm and the other based on a parallel Hopfield net algorithm (Chen & Ng, 1995).

The Hopfield algorithm, in particular, has been shown to be ideal for concept-based information retrieval (Chen et al., 1993).

Each term in the network-like thesaurus was treated as a neuron, and the asymmetric weight between any two terms was taken as the unidirectional, weighted connection between neurons. Using user-supplied terms as input patterns, the Hopfield algorithm activated their neighbors (i.e., strongly associated terms), combined weights from all associated neighbors (by adding collective association strengths), and repeated this process until convergence occurred.

4. Experimental Design

4.1. Browsing Experiment

This section describes the experiment we performed to compare our prototype Kohonen SOM-based Internet browsing tool (ET-Map) with an existing Internet browsing tool, Yahoo!. Yahoo! was selected as the basis for comparison because the collection spider used Yahoo!'s entertainment sub-directory to define the boundaries of its collection task. Thirty-four University of Arizona students, recruited from four MIS summer school classes and one fall semester class, and several Library Science students participated in the experiment. The summer school students and the Library Science students volunteered; fall semester students received a nominal amount of extra class credit. After seven of the subjects performed two sets of browsing tasks, the experimenters decided that their experimental sessions took too long and subsequent subjects performed only one set of browsing tasks. Subjects were not given any training in navigating either Yahoo! or the ET-Map, but were basically allowed to explore on their own. Experimenters did answer questions if asked.

Subjects were started in either the Yahoo! Entertainment sub-directory (see Fig. 2) or the ET-Map (see Fig. 1) and were asked to browse for ''something of interest to you.'' The browsing task was described to them as ''window shopping'' and they were asked to start without a specific goal in mind. Subjects were asked to think ''out loud'' and describe the reasoning behind their navigation choices. These verbalizations were recorded by the experimenters and were later analyzed using verbal protocol analysis. Experimenters were particularly interested in capturing the choices (the navigational path), the reasoning behind the choices, and the subject's satisfaction with the process.

The browsing tasks were ended either after 10 minutes (in later experimental sessions this was extended to 20 minutes), or after subjects had successfully located a homepage of interest. At the completion of a browsing task, subjects were asked to attempt to repeat the browse using the other tool. Half the subjects began the browsing task using Yahoo! and the other half began the browsing task using the ET-Map. Three searches had to be dropped from consideration because subjects tried to do a directed search instead of browsing or could not perform the browsing task without resorting to a keyword search.

In an attempt to compare the tools fairly, browsing in Yahoo! was begun in the Entertainment Section of Yahoo! and subjects had to use the links and the hierarchical organization of Yahoo! to navigate. They were not permitted to use the keyword searching feature of Yahoo!. We wanted to compare how successful subjects were in using a hierarchically organized directory-based browsing mechanism versus the SOM-based browsing mechanism, and how satisfied they were with the process. At the end of the experiment, subjects were asked to comment on the ET-Map. In particular, experimenters asked about likes, dislikes, ideas for improvements, if the subjects would like to use the Map again, and when they thought the Map would be most helpful. We analyzed the browsing tasks to try to determine the characteristics of successful and unsuccessful browses using the ET-Map.

4.2. Searching Experiment

This section describes the experiment we performed to compare our concept space or thesaurus-based Internet searching tool with a commonly used Internet searching technique, keyword-based searching. Thirty-eight university students (most of whom were different from the previous browsing subjects), were recruited from four MIS summer school classes, one fall semester MIS class, and from a Library Science class. As in the previous experiment, the summer school subjects and the Library Science students were volunteers, and the fall semester subjects were given a nominal amount of class extra credit for participation. The experimenters briefly demonstrated how to use both the keyword-based searching tool and the thesaurus-based searching tool before the experiment was started. Experimenters also answered subject's questions as the search progressed.

Subjects were asked to think of a topic in the entertainment area (and standard suggestions were given as to what constituted ''entertainment''), verbalize their topic, and try to locate WWW related homepages. Subjects were asked to verbalize the reasoning behind the ''navigation'' choices that they made (usually why they chose the particular words or terms to initiate and refine the search), and their satisfaction or frustration with the searching as it progressed. Once subjects had accessed a list of WWW homepages recommended by the searching process, they were asked to indicate how many of those homepages actually related to their chosen topic. Experimenters recorded the time each search took, the total set of homepages suggested by the searching mechanism, and the number of homepages from that set that subjects indicated were relevant.

Figure 5 is a picture of the user interface used for the searching experiment. If the subject was performing a

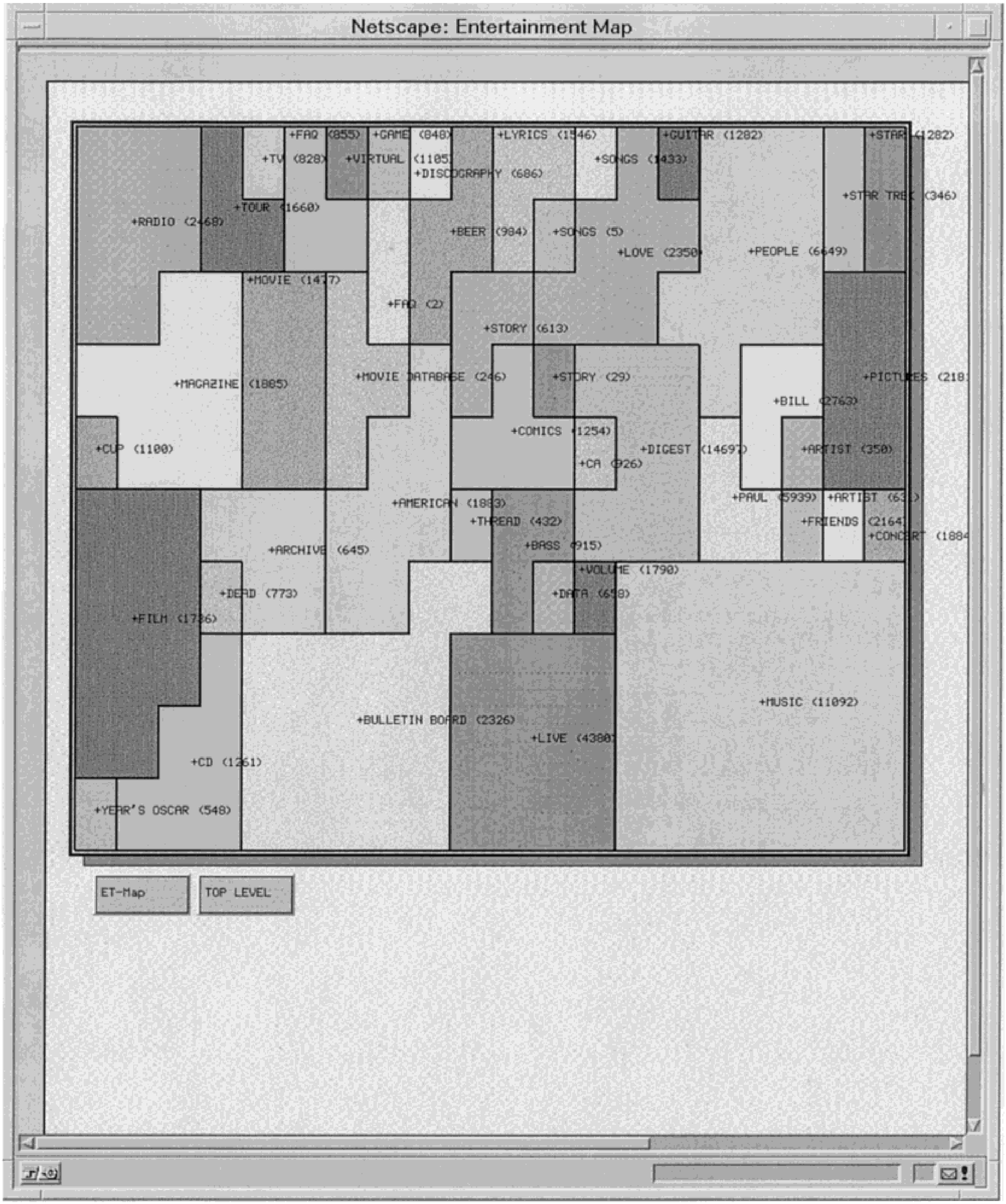

FIG. 1. The ET-Map interface.

keyword-based search, the search went directly to the homepage or document space (the URL index). If the subject was performing a thesaurus-based search, his or her initial input term was used as an entry point into the thesaurus or concept space and the searching tool would return a list of suggested terms that were related to the input term, according to the thesaurus. Figure 6 illustrates terms suggested by the thesaurus when a subject entered

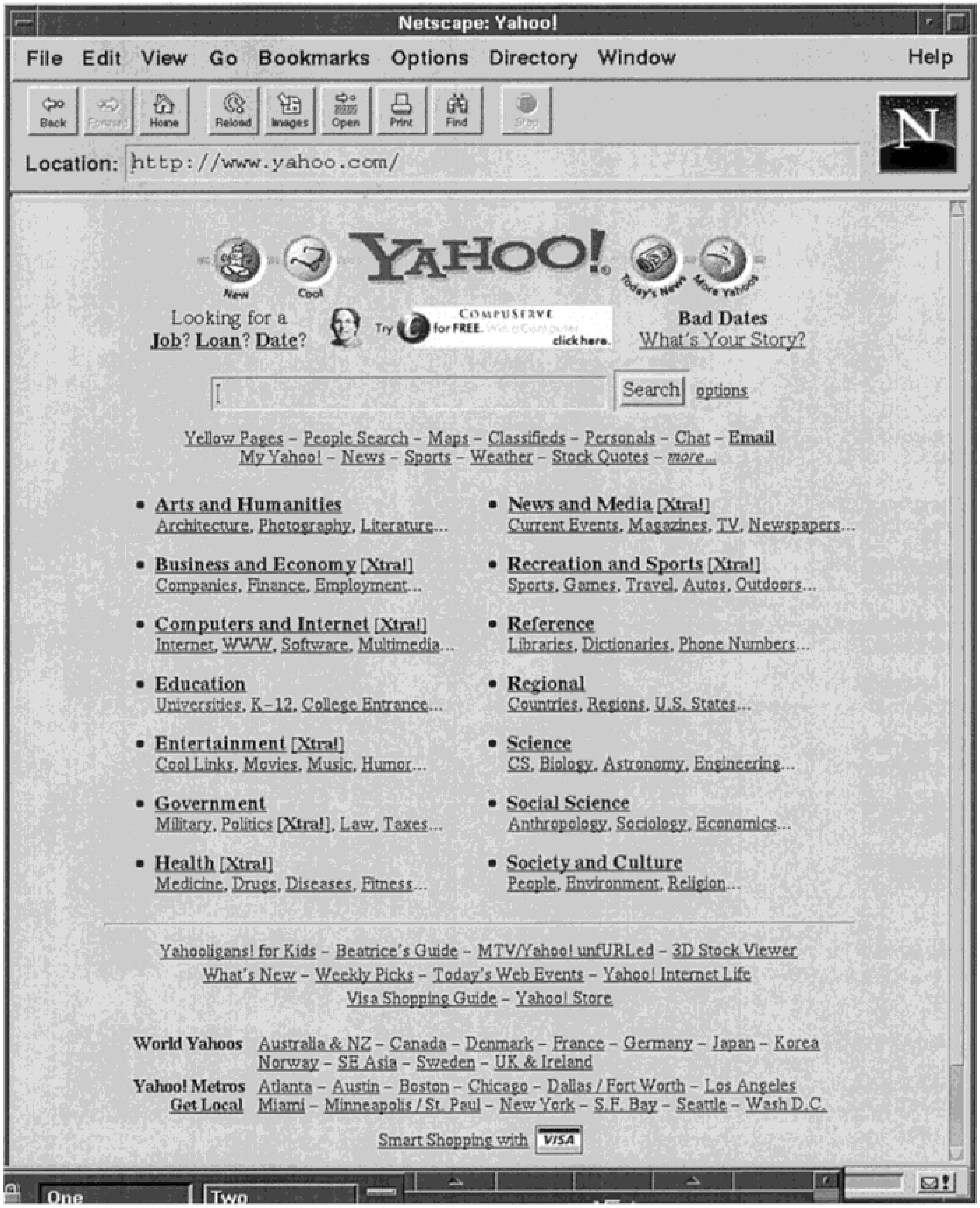

FIG. 2. The Yahoo! Entertainment sub-directory.

the following search request: ''Steel OR Guitar.'' Figure 7 illustrates the list of documents (in this case links to URLs) that the tool determined were relevant to the subject's selected thesaurus term ''Pedal Steel Guitar.''

Half the subjects began their searching task with the keyword-based search and the other half began with the thesaurus. Subjects were recruited until 35 usable experiments were collected, 18 starting with keyword search and 17 starting with the thesaurus. Once a search had been successfully completed using one tool, each subject was asked to try to duplicate the search using the other tool for comparison.

The experimenters then evaluated each of the searching tasks, analyzing the set of final documents (URLs) that each task returned. Using recorded protocol analyses, we identified searching terms as originating from the subject, or from the thesaurus. Next, we analyzed each set of returned documents, identifying relevant (in our opinion) documents and indicating which searching term was responsible for the retrieval, noting whether it was a subject-

- **TRD** - The total number of relevant documents (TRD) was computed by combining all the relevant documents (as identified by the experimenters) suggested by the keyword search, the thesaurus search and the Alta Vista search, removing all duplicates.
- **TRS** - the total of all documents identified in the keyword search and all relevant documents identified in the thesaurus search, after removing duplicates.

For each subject, the results of both the keyword search and the thesaurus search were analyzed and the following measures were calculated:

- **SRK** - the number of relevant documents in the keyword search that were identified based on terms suggested by the subject (without assistance from the thesaurus).
- **SRT** - the number of relevant documents in the thesaurus search that were identified based on terms suggested by the subject (without assistance from the thesaurus). This occurred when subjects kept their initial searching terms or didn't like any of the terms suggested by the thesaurus.
- **STS** - the total number of documents that were identified based on terms suggested by the subject (without assistance from the thesaurus).
- **TRK** - the number of relevant documents in the keyword search that were identified based on terms suggested *solely* by the thesaurus. Note that this means that if the thesaurus also suggested a term that the subject suggested, only the subject got credit for the term.
- **TRT** - the number of relevant documents in the thesaurus search that were identified based *solely* on terms suggested by the thesaurus. Note that this means that if the thesaurus also suggested a term that the subject suggested, only the subject got credit for the term.
- **TTS** - the total number of documents that were identified based on terms suggested *solely* by the thesaurus. Note that this means that if the thesaurus also suggested a term that the subject suggested, only the subject got credit for the term.

FIG. 3. Baseline measures.

suggested term or a thesaurus-suggested term. We also performed a keyword search on Alta Vista, using the same searching terms used by the subjects. Alta Vista was chosen since it has an alliance with Yahoo! but indexes more homepages than Yahoo!.

In all cases, only the top 40 retrieved documents were examined. Pilot studies indicated that subjects had a difficult time evaluating more than 40 documents at a time. We also had discovered from pilot studies that the weighting algorithms used by our searching tools and by Alta Vista usually included the most relevant documents in the top 40 retrieved documents. In our pilot studies, subjects rarely marked any document below document 40 as relevant.

Concept recall and concept relevance were computed for each searching task as follows. First the baseline measures were computed. If a term was suggested by both the subject and the thesaurus, the subject always received credit for the term, and the thesaurus did not. We felt this was reasonable as the thesaurus did not assist the subject in any way, since subjects had already thought of the term on their own. See Figure 3 for information on the baseline measures and how these figures were derived. Using the baseline measures, the next step was to compute the recall measures. See Figure 4 for the details of how these measures were calculated. Recall results are summarized in Figure 10. The following measures of recall were computed:

- *Subject recall,* which reflects the percentage of the total potential relevant set of documents that were located using a term (or terms) suggested by the subject.
- *Thesaurus recall,* which represents the percentage of the total potential relevant set of documents that were located using a term (or terms) suggested by the thesaurus.
- *Combined recall,* which indicates the percentage of the total potential relevant set of documents that were located using terms suggested by the subject in conjunction with terms suggested by the thesaurus.

Finally, the precision measures were calculated, using the baseline measures. See Figure 4 for the details of how these measures were computed. Precision results are summarized in Figure 9. The following measures of precision were computed:

- *Subject precision,* which represents the percentage of relevant documents located by the searching methods we used that were located by a term (or terms) suggested solely by the subject.
- *Thesaurus precision,* which reflects the percentage of

- **Subject Recall** = ((SRK + SRT) - duplicates) / TRD
- **Thesaurus Recall** = ((TRK + TRT) - duplicates) / TRD
- **Combined Recall** = ((SRK + SRT + TRK + TRT) - duplicates) / TRD
- **Subject Precision** = ((SRK + SRT) - duplicates) / STS
- **Thesaurus Precision** = ((TRK + TRT) - duplicates) / TTS
- **Combined Precision** = ((SRK + SRT + TRK + TRT) - duplicates) / TRS

FIG. 4. Recall and precision formulas.

relevant documents located by the searching methods that we used that were located by a term (or terms) suggested by the thesaurus.

- *Combined precision,* which indicates the percentage of the total relevant set of documents retrieved by our searching techniques that were located using terms suggested by the subject in conjunction with terms suggested by the thesaurus.

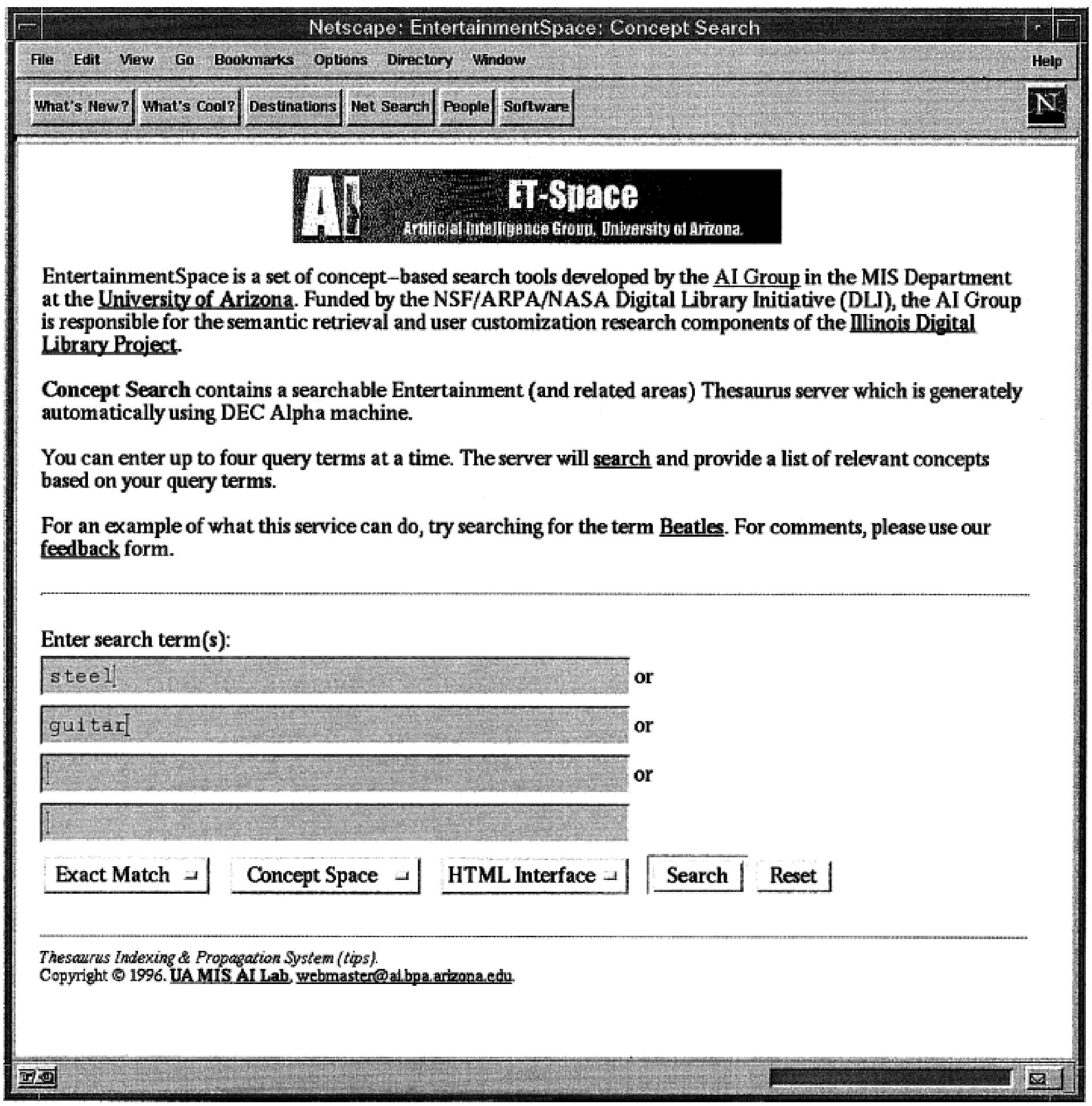

FIG. 5. The searching experiment interface.

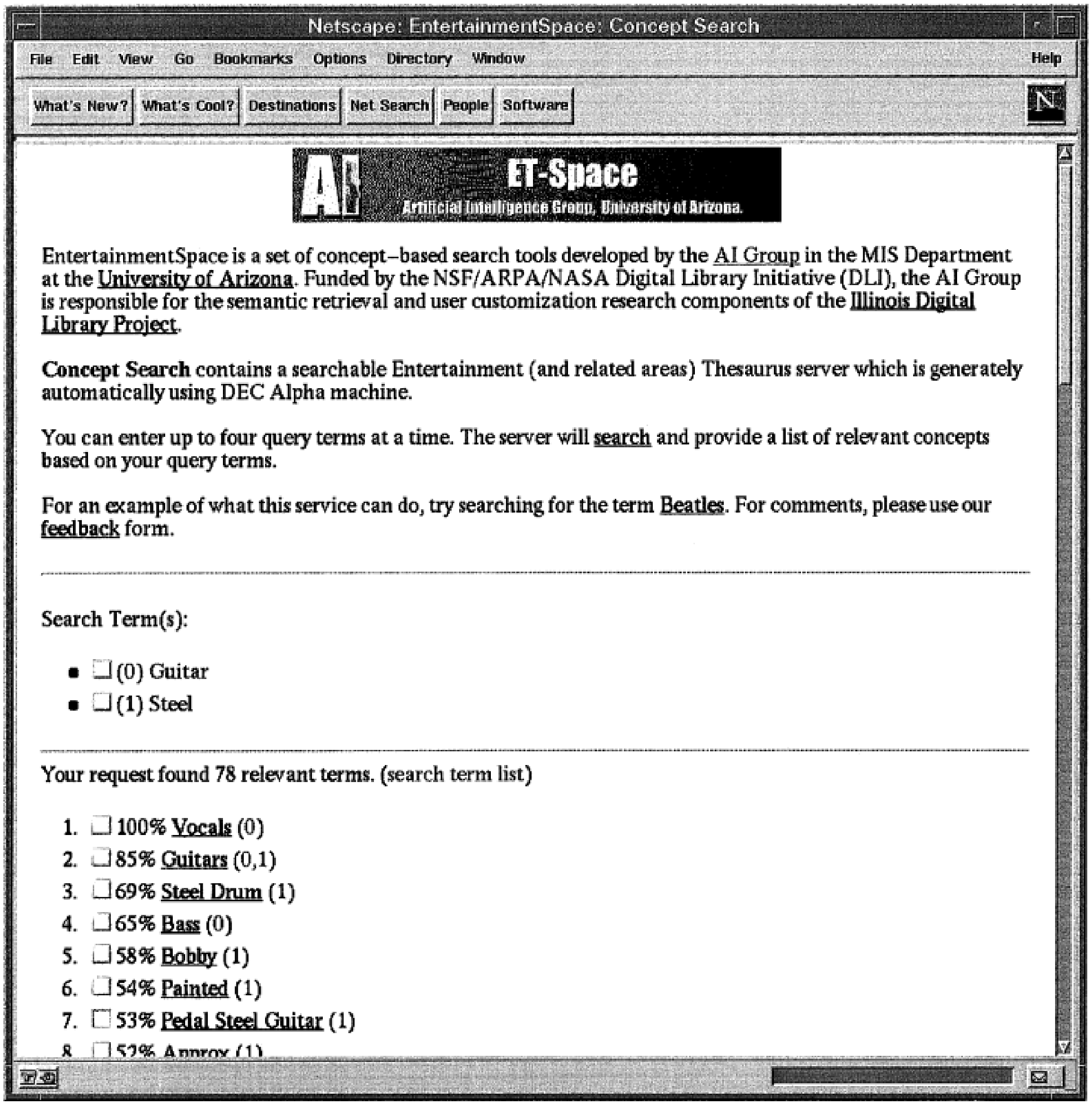

FIG. 6. Terms suggested by the thesaurus.

5. Browsing Experiment Results and Discussion

The results of the browsing task are summarized in Table 1. For the subjects who started the task using the ET-Map, 11 of the 15 usable browses resulted in the subject's locating an interesting homepage (or a set of homepages). The 11 subjects with successful browses were then asked to try to repeat their browsing tasks in Yahoo!. Eight of the 11 subjects were able to do so successfully. For subjects who started the task using Yahoo!, 14 of the 16 usable browses resulted in the subject's locating an interesting homepage (or a set of homepages). However, only two of the 14 successful subjects were able to repeat their browsing task using the ET-Map.

We believe that the experiment confirms our expectation that the Kohonen SOM-based technique (of which our ET-Map is a small working prototype) can be used effectively and scaleably to browse a large information space such as the Internet. Subjects could locate homepages of interest to them approximately 73% of the time. We further believe that our success rate with the Kohonen SOM-based tool could be improved with some modifications to the user interface, or increased user training or user experience with the tool.

In Table 1, the label ''Repeatable Browses'' is somewhat of a misnomer, because the second ''browsing'' task clearly was not a browsing task at all but rather a searching task that did not use keywords. In each instance, the first browsing task was simply an exploration of the information space without a specific or well-defined goal. In the second browsing task, the subjects were asked to try to locate the same homepage they had identified as interesting using the first browsing tool. This second task had a very definite goal. We believe that our results indi-

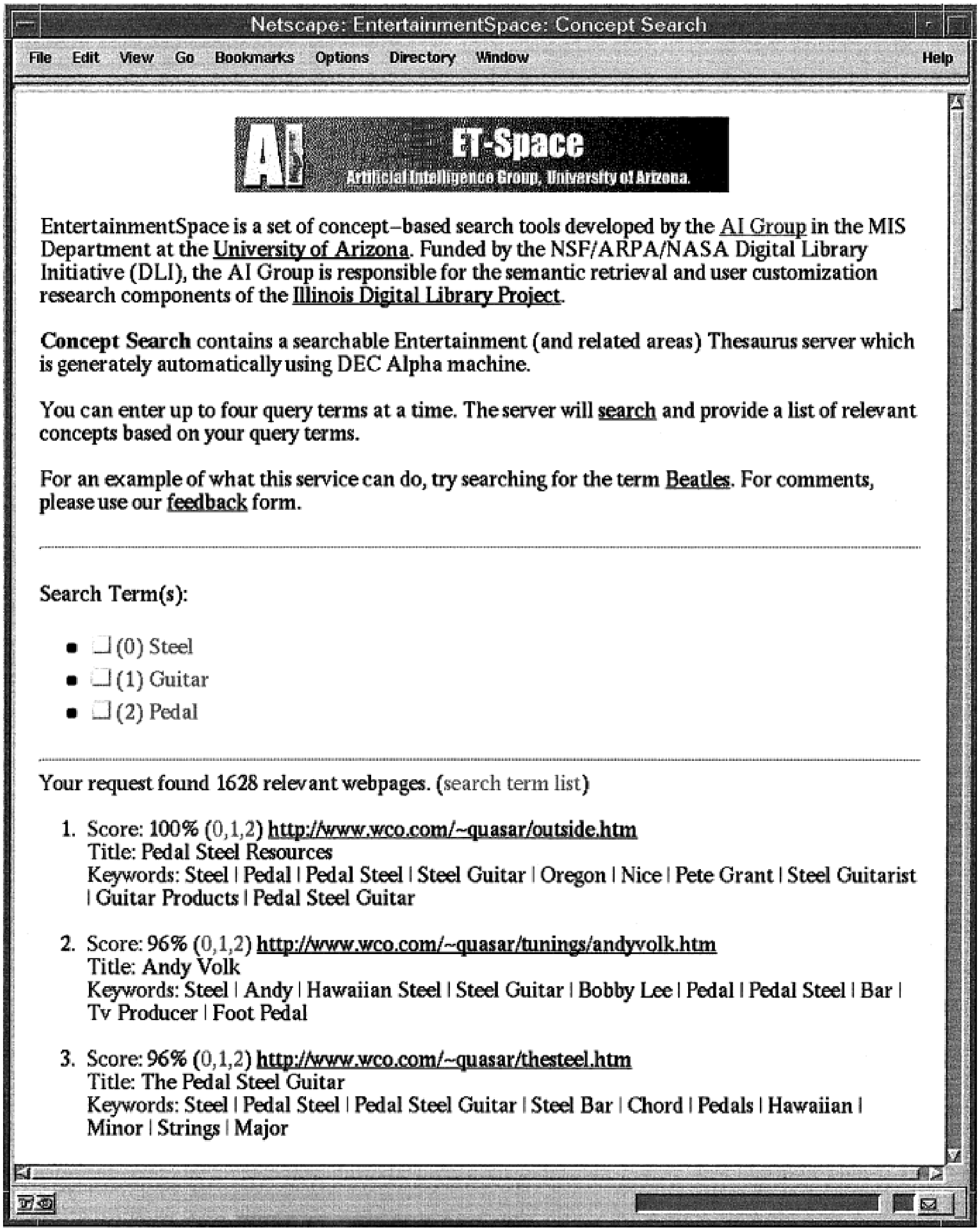

FIG. 7. The document interface.

cate that while the ET-Map can be used as a browsing tool for the Internet, it cannot be used as a searching tool, nor as a tool that supports a browsing task with a specific goal.

Verbal protocol analysis of these sessions reveal some interesting browsing behaviors.

• *Mental Models.* The repeatability of the browses can be partially explained by the mental model phenomenon. Only 14% of the subjects could successfully recreate a Yahoo! browsing session using the ET-Map, compared with 73% who could successfully recreate an ET-Map browsing session in Yahoo!. We believe that, to some extent, this is due to the fact that the structures or organization of the two browsing methods are very different and that subjects formed a mental model of browsing from using the first method that could not be successfully applied to the second. We also discovered that the majority of our subjects were strongly rooted in the conventional (or more familiar) categorization and organization

TABLE 1. Results of browsing experiment

Starting tool	Browses		
	Attempted	Successful	Repeatable
ET-Map	15	11	8
Yahoo!	16	14	2

mental models (i.e., alphabetic and hierarchical). This accounts for the fact that subjects were able to recreate browsing tasks in Yahoo! by abandoning the mental model formed by using the associative ET-Map in favor of the more familiar hierarchical Yahoo!. We also noticed that subjects who started with Yahoo! had a more difficult time adjusting to the associative organization of the ET-Map than subjects who started with the ET-Map. For example, one subject expressed concern that he could not use the ET-Map for a hierarchically structured browse, and then added ''The relationships [in the ET-Map] don't seem to make much sense to me. It doesn't fit with my mental map. On the other hand, I found the Doom page much quicker with the ET-Map!'' Another subject commented, ''I think once I had more experience using the map and understood the organization better, I would like it better.''

• *Non-cohesive Information Source and Application.* We did not get as positive a response to the ET-Map as we have had from other SOM-based applications. We believe that this is in part due to the eclectic and non-cohesive nature of both the information source (WWW homepages) and the application (entertainment). Previous implementations of the SOM technique have been applied to textual documents, frequently articles or abstracts. This kind of document has cohesiveness both in function or purpose and in content, i.e., it was written to convey a specific informational message.

WWW homepages are more eclectic documents. The purpose or function and the content vary tremendously between commercial pages and personal pages. Personal pages, in particular, often contain information about a variety of unrelated topics. What provides cohesion is that all of the topics on the homepage are of interest to its owner. For example, the homepage of a college student may contain a resume, information about a favorite type of music or musical group, information about favorite hobbies or recreational activities (skiing for example), organizations to which the individual belongs, and perhaps even information about the school the individual is attending. In general, these data are unrelated. What causes the relationship or connection is the often unique interests of the individual who created or owns the homepage. One solution to this might be to attempt to ''weed out'' personal pages and concentrate on more cohesive homepages (such as commercial ones). But, at least in the entertainment sub-category, we feel this would cause the loss of relevant information. Many of our subjects felt that some of the most relevant homepages were either personal homepages or pages linked to them.

The one sub-category that was an exception to this phenomenon was the music sub-category. Most of the subjects who were browsing in the music sub-category had a much higher success rate and located information of interest more quickly than subjects browsing other areas. We believe this is because the music sub-category is more cohesive and the subject's mental models for music were more consistent with the ET-Map's conceptual model for music. See Figure 8 for the second level map for the music sub-category. Another possibility is that music was obviously the largest sub-category and the Kohonen SOM had more experience (in essence more training sessions) with that sub-category, which may have resulted in a better representation.

• *Knowledge of English.* We found that native speakers had an easier time using the map than non-native speakers. They seemed to better understand the organization of the map. The subjects who were most frustrated or most confused by the map were non-native speakers, and indeed some of their browsing sessions had to be dropped. However, we did not collect specific information on language comprehension per se.

5.1. Positive Feedback about the ET-Map

There were some general themes in the positive feedback that we received about the ET-Map as a browsing tool. They are as follows:

• *Graphical Aspects: Spatial Factor and Color.* Subjects liked the spatial factor, especially the fact that size of the area was related to number of URLs connected to the area. This is consistent with the well-known phenomenon that graphics are more readily understood than text (''A picture is worth a thousand words''). Users could quickly determine which area of the map had the most referenced URLs. Typical user comments include: ''I like the fact that I can immediately look at it and tell that music is the biggest area (on the top level map).'' See Figure 1 for a picture of the top level of the ET-Map.

Subjects also liked the variety of colors used to differentiate the various areas of the map. ''Different colors are eye appealing,'' was a typical statement. Again, subjects mentioned that the colors helped to define the areas and made area location and definition easy to determine quickly.

Another commonly expressed positive feature of the map was that subjects liked the fact that the map was on one page. Many expressed pleasure that they could get a global picture or the ''big picture'' from the map. However, many subjects then wanted the ability to zoom in or ''blow up'' a region of the map for closer inspection, and expressed disappointment that the map did not yet have that capability. For example, a typical comment

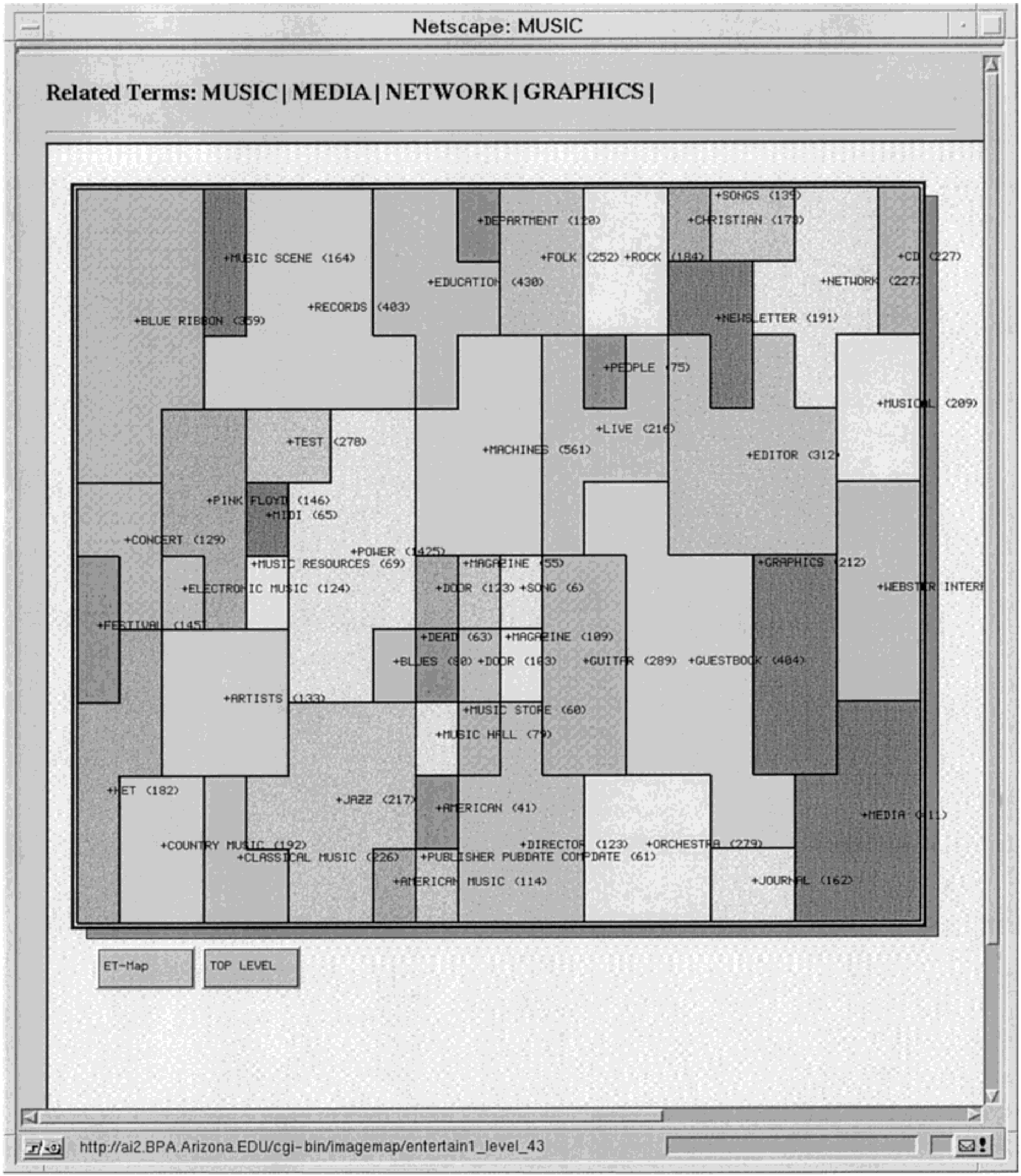

FIG. 8. The music sub-category.

was: ''I like the fact that the entire space is on one page. You get an overall view, but it would be nice to be able to zoom in on an area.''

• *Usage and Navigation.* Several subjects quickly realized that the ET-Map was specifically designed for a browsing task and was not useful in any kind of directed search. These subjects were ones who began their browse as a directed search, despite instructions, and were frustrated in trying to locate their specific interest. Once they changed tactics to a true browse (e.g., ''Oh, let's just look around and see what's here'') and began exploring, they were generally able to find a topic or area of interest to them. This resulted in comments like: ''I would use this if I was only searching for fun. If I wanted information for a report, I would use keywords or the concept space'' or ''This is okay if you are not looking for something specific, but just exploring.''

Another interesting observation was that subjects who really liked the ET-Map tended to do broader exploring than subjects who did not like the map. These subjects quickly discovered that when they got bored with one area, it was easy to jump to a new area of the map. In Yahoo!, due to the hierarchical nature of the directory's organization, subjects had to back out, or start over to change browsing areas. ''It is really easy to jump to another topic if you change your mind.''

• *Diversity and Novelty.* Some of the subjects were interested in the diversity and the novelty of the way that the map organized WWW homepages. This kind of comment was often on the order of: ''This is an interesting way to divide things up.'' One subject particularly liked the diversity of the map: ''The map seems very repetitious, but very diverse. It is interesting. I really like the fact that there are lots of different routes that you can use to get you to where you want to go.''

• *Layers and Labels.* Subjects tended to like the fact that the map had layers or levels. ''I like the fact that there are different levels, it breaks up the database. It is easier to search with sub-categories.'' ''The map is easier to use at lower levels. The initial level had no terminology that I was used to or familiar with, but the labels on the lower levels made sense to me.'' One subject liked the one-term categorization or label. She felt that it allowed her quickly to determine whether or not the area was going to be of interest to her. All of these comments indicate that the ET-Map was helpful in reducing the impact of information overload by creating browsing regions of a manageable size.

5.2. Negative ET-Map Feedback

We found that most of the negative feedback we received had to do with either the unfamiliarity of the word association structure (new mental model compared with the more classical alphabetic or hierarchical organization structures), concern over the user interface or status of the document collection, or attempts to use the tool for searching as opposed to browsing. The most serious criticism from a usability viewpoint is the tool's inability to generalize and to present information at similar levels of abstraction.

• *Hierarchical Organization or Conventional Organization.* Some subjects wanted to have access to more conventional organization strategies. This was particularly true of novice browsing subjects who seemed to be rooted in the conventional (alphabetic or hierarchical) organizational browsing mental models. Typical comments from this group include: ''There is no order, alphabetical or otherwise. The order is difficult to see'' or ''Ordering of the choices is not systematic, or does not seem to be'' or ''It would be better if it was organized in a more meaningful way. For example by popularity. The areas that people chose most often should be brighter, or bigger, or at the top of the map.''

Many subjects asked for an alphabetic list or index. ''I had to read the entire map, just to see if there was anything I was interested in. There was no way to scan it quickly. I would like an index at the side.'' Again, this supports our premise that subjects were more comfortable with conventional mental models of information organization. Interestingly, many subjects wanted the colors used to differentiate areas on the map to have more meaning (for example, related terms should either be in the same color, in colors close on the color wheel, or unique to each area).

• *Word Association.* A common criticism was that subjects, particularly subjects who had trouble navigating the map, wanted the word associations to be clearer or more personally meaningful. Comments included: ''How did the map come up with this set of terms?'' or ''I really like the map, but I wish the labels were more meaningful. They seem kinda useless unless they are the specific term you want.'' For example, one subject chose the area labeled ''BILL'' out of curiosity. He was surprised to see the second level had topics like counting (related to the monetary aspects or definitions of the word). ''I thought that I would find Bill Clinton or a rock group or something.'' This criticism relates back to the mental model issue discussed above.

• *Getting Lost.* Some subjects, in particular novice browsers, tended to get lost or confused. ''How do you determine what level of the map you are on?'' ''It is too easy to lose my place. I forgot where I started from and how I got here.'' Interestingly, many of the same users had similar problems in Yahoo!.

• *Readability.* Most subjects felt that the words or labels were difficult to see. They wanted better readability. For example, there was one map that subjects would abandon as soon as they chose it because there were too many small areas that were too difficult to sort out. Typical comments included: ''Words are too small to read'' or ''It is confusing when the words overlap the spaces. Should I choose the word or the space?'' or ''The words should be centered in the areas to make it easier to use.'' Notice that these comments have more to do with the user interface than the ability of the SOM algorithm per se. That is not to say that they are not issues critical to the usability of the map, but rather, they are not problems inherent to the technique itself.

• *Flexibility.* One subject thought that the map was not flexible enough. This person felt that there should be more ways to use it (for example, a combination of keyword search and map and hierarchy, similar to Yahoo! and Lycos). We believe that this would be an important improvement to future user interfaces for the ET-Map.

• *Collection Process.* In some cases, subjects expressed concern that the collection process was limited to Yahoo!'s definition of entertainment (particularly when it did not seem to coincide with theirs). They felt that this was too limiting. A commonly given example was sports, a large enough category to have its own sub-directory in Yahoo! that is separate from the entertainment subdirectory. Many people argued that sports should be considered entertainment. This represents a failing of the collection process, not of the SOM algorithm.

• *Age of Collection.* Another common criticism was that the collection was not real-time or continuous, but static. While subjects (especially those familiar with the frustration of waiting for Internet searches) were impressed with the rapid response (due to the local nature of the ET-Map), they were concerned when they looked at URLs that had been changed (either deleted generating ''not found'' errors or changed so the keywords and labels chosen by the algorithms to represent the homepage were no longer relevant). Subjects who were looking for current or newly released movie reviews, for example, were especially disappointed that the collection had occurred several weeks before. This is another striking difference between our ''document'' collection and the more classic textual documents. Typically, once documents are created and classified in print, they do not change. Textual documents do not tend to be as dynamic as WWW homepages. In the textual world, changes to a document become a new document (for example, a newer edition of a book). This situation implies that, in future prototypes, it will be important to regenerate the ET-Map more frequently (perhaps daily).

• *Use of Map for Searching.* The ET-Map is not a useful tool for directed searching. It was designed as a browsing mechanism. Nevertheless, several subjects initially tried to use it as a directed searching tool, and those subjects who were directed to repeat Yahoo! browses using the ET-Map clearly were attempting to do a browse that had a definite goal. In both cases, most subjects were frustrated by their lack of ability to get directly to a homepage (''cut to the chase''). Many of these subjects tended to be self-described experienced Internet keyword searchers. Typical comments from this group included: ''How do you get to the cream?'' or ''What if I just wanted to go directly to the URLs?''

• *Inability to Generalize Different Levels of Abstraction.* The ET-Map suffered from the same limitation as other neural-network-based categorization techniques: Namely the lack of an ability to generalize or to present topics at the same level of abstraction. Comments included: ''The initial topics are either too specific or lack meaning'' or ''The word terminology and association seems rather vague'' or ''Star Trek is very specific and music is general, but they are on the same level.'' ''Terms should become more specific as you go down through the levels but they don't. Many of the terms are still very general.''

6. Searching Experiment Results and Discussion

We were especially pleased with the results of the searching experiment. We found no statistically significant difference in document precision between terms suggested by the subject, and terms suggested by the thesaurus (refer to Fig. 9). This means that the terms suggested by the thesaurus were no worse from a precision point of view than terms suggested by the subject. In fact, based on our qualitative analysis, we believe that the thesaurus would have performed better in precision if we had changed the user interface to allow subjects to use the Boolean operator ''AND.'' We discovered that, when using the multiple input the thesaurus permitted, most subjects incorrectly assumed that either an ''AND'' operator was linking the terms, or that the thesaurus would use the terms they chose to further refine a retrieval based on their initial input term.

Even more encouraging were the results from the recall calculations (refer to Fig. 10). In this case, there was no statistically significant difference between the recall ability of the thesaurus compared to the subject. This means that the terms suggested by the thesaurus were no worse in identifying relevant homepages from a pool of potentially relevant homepages. What was especially interesting was that there was a statistically significance difference between the recall ability of the subject's terms and the combined terms, and between the thesaurus's terms and the combined terms. This means that the thesaurus was identifying relevant homepages that the subject's terms were not locating. This result indicates that a subject's hit rate, or ability to identify relevant homepages, can be significantly improved by adding terms suggested by the thesaurus to a directed search.

We noticed some general themes or trends when analyzing the qualitative information from the searching tasks. This information is summarized below.

• *Searching Experience.* We noticed that, in general, subjects who described themselves as experienced Internet searchers took to the concept space better than the inexperienced subjects. This group seemed to understand many of the problems with keyword searching (in some instances relating negative experiences with other search engines). Experienced searchers tended to quickly grasp the idea that the thesaurus could be used to refine their

```
ONE-WAY ANALYSIS OF VARIANCE

ANALYSIS OF VARIANCE
Source      DF        SS        MS         F        p
Factor       2    0.0450    0.0225      0.25    0.779
Error       97    8.7016    0.0897
Total       99    8.7466
                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
 Level       N      Mean     StDev  -------+---------+---------+---------
Subject     32    0.2689    0.2860  (--------------*--------------)
Thesaurus   33    0.3187    0.3469         (--------------*-------------)
Combined    35    0.3085    0.2610        (-------------*-------------)
                                    -------+---------+---------+---------
Pooled StDev =    0.2995                 0.210     0.280     0.350
```

FIG. 9. Precision comparison by term source.

search (either narrow or broaden it) and why this was an improvement over strictly keyword-based searching.

Inexperienced searchers appeared to have a more difficult time understanding the value of the thesaurus. This group tended to want simply to enter words in different combinations in a keyword manner. Some of them even commented that they did not understand why a searcher would want to spend extra time looking at the thesaurus when they could just ''type in a bunch of keywords and get directly to the URLs.''

• *Narrow vs. Broad Searches.* Another general observation is that subjects who started out with searches that were too narrow (very specific) had more trouble than subjects who started out with searches that were too broad. Subjects starting out too broadly (for example, beginning with a simple term like ''movie''), immediately saw how the thesaurus could help them refine or narrow their requests by suggesting related terms that were more specific. A typical comment was: ''I like the fact that I can use the thesaurus to narrow or expand my search.'' The thesaurus tended to suggest fewer related terms, and those terms tended to be more general for searches that were very narrowly defined from the beginning. These subjects tended to know exactly what they wanted and felt that their choice of keywords was just as good, and in some cases better, than what the thesaurus recommended.

• *Spelling Errors and Typos.* Spelling errors and typos were particularly frustrating to some subjects. In more than one case, sophisticated users wanted to be able to ''wild card'' a term or request a spell checker to recommend the correct spelling when no hits were found due to a spelling error (particularly true of proper names—places, groups, or individuals).

• *Information Source.* The information source itself was frequently another interesting source of frustration to some subjects. The Yahoo! Entertainment sub-category is heavily influenced by the homepages of young people (especially high school- and college-age people). As a consequence, there tend to be more homepages that contain entertainment information of interest to this age group (for example, information about music and movies that appeal to this age group). Information specifically about children's entertainment or entertainment for older people was sparser. Subjects with these searching interests had a more difficult time and expressed greater frustration than subjects who were interested in topics popular with a younger group.

• *Knowledge of English.* Interestingly, contrary to our experience with the ET-Map, non-native speakers did not have a noticeably harder time with the concept space than native speakers. Experienced searchers who were not native speakers seemed to like the fact that they could get valid searching term recommendations from the thesaurus.

6.1. Positive Thesaurus/Concept Space Comments

We were especially encouraged by the amount of positive feedback that we received about the thesaurus. Again, there were some noticeable trends or themes around which we have organized the comments.

• *Thesaurus Organization.* Subjects tended to like the organization of the thesaurus. Specifically, subjects liked the idea that they could look in two places, an ''index'' or directory place (the thesaurus) and a ''document'' place (the URLs). ''I like the fact that I can get a 'total' listing.'' One novice subject recommended that

```
ONE-WAY ANALYSIS OF VARIANCE

ANALYSIS OF VARIANCE
Source     DF        SS        MS        F        p
Factor      2    0.4314    0.2157     5.30    0.007
Error      97    3.9476    0.0407
Total      99    4.3789
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
 Level      N      Mean     StDev  -------+---------+---------+---------
Subject    32    0.2168    0.2312       (--------*--------)
Thesaurus  33    0.1754    0.1672  (--------*--------)
Combined   35    0.3290    0.2026                    (-------*--------)
                                   -------+---------+---------+---------
Pooled StDev =   0.2017                 0.160     0.240     0.320


TWO SAMPLE T-TEST AND CONFIDENCE INTERVAL

Two sample T for Subject vs Combined
           N     Mean     StDev   SE Mean
Subject   32     0.217    0.231     0.041
Combined  35     0.329    0.203     0.034

95% C.I. for mu Subject - mu Combined: ( -0.219,  -0.005)
T-Test mu Subject = mu Combined(vs not =): T= -2.10  P=0.040  DF=  61


Two sample T for Thesaurus vs Combined
           N     Mean     StDev   SE Mean
Thesaurus 33     0.175    0.167     0.029
Combined  35     0.329    0.203     0.034

95% C.I. for mu Thesaurus - mu Combined: ( -0.243,  -0.064)
T-Test mu Thesaurus = mu Combined(vs not =): T= -3.42  P=0.0011  DF=  64
```

FIG. 10. Recall comparison by term source.

more clarification on how many URLs a suggested term found would be useful. She felt that she could not really tell if the thesaurus was going to help limit the search (by reducing the number of URLs returned) until she got into the document area.

• *Multiple Terms.* Many subjects really liked the fact that multiple terms could be entered. ''I like the fact that you can search all (many terms) at the same time. This gives more options than the traditional stuff.'' Experienced searchers particularly liked this feature. One caveat related to this is that most subjects who took advantage of the multiple term feature also requested that future releases of the interface allow them to input Boolean operators (''AND,'' ''OR,'' and ''BUT NOT'').

• *Ranking, Term Indicator Information, and Totals.* Subjects liked the ranking and the term indicator information indicating which input term caused a recommended term or term set to be included by the thesaurus in its list. They also liked the fact that this information was included when the URLs were displayed. The URLs were ranked and followed by a term indicator which identified which searching term or terms caused that URL to be included on the list. A typical comment is: ''I like being able to choose from the thesaurus, and the 0, 1, 2 (term indicators) that apply to the categories is helpful.''

Subjects also liked the fact that when they reached the URL listing level, they were told how many URLs were returned in total. This quickly allowed them to decide if the search was too narrow or too broad, before they started looking closely at the URL list. One subject would not bother to review a list of more than 30 URLs. Several subjects expressed the notion that a search was too broad if it returned more than 20–30 URLs. A similar criticism was made about the ET-Map. Subjects did not want to review a list of more than 40 URLs. This is consistent with what we know about information overload, in general, and with our experience with information overload in electronic brainstorming categorization, in particular.

• *User Control.* Subjects liked the fact that they were in control of the search. They generally appeared to feel that the thesaurus gave them more control over a search than a simple keyword-based search. Some subjects emphasized that they felt the thesaurus gave them an intelligent alternative because suggested terms actually derived from the documents themselves were guaranteed to have matches. ''I can choose the search but with intelligence because the thesaurus tells me what the sub-categories are.''

Typical comments included: ''I like the fact that I can control if the match is a partial match or an exact match'' or ''I like the choices'' or ''I like the fact that I can control whether I search the concept space or the URLs'' or ''It is nice to see the computer limit the choices instead of the way I structure the query.''

However, one subject thought that the control was bad. This subject commented: ''Random things that come up can be fun too.''

• *Search Refinement.* Many of the subjects, especially the more experienced Web searchers, liked the fact that the thesaurus could be used to refine their searches. Most quickly understood the concept of narrowing the search space with more specific terms and liked the idea that the thesaurus could suggest ''real'' or valid search terms. Typical comments included: ''I like the fact that the narrow searching terms actually exist in the database. When I use Sabio (our university's online library searching service) and try to narrow the search, I often end up with too narrow a search because I don't know what additional keywords will work'' or ''It helps to hopefully zero in on your topic.'' Several subjects liked to enter just one keyword and then let the thesaurus help them narrow the search, as opposed to entering multiple keywords.

What fewer subjects grasped was the ability of the thesaurus to broaden a too narrowly defined search. The ones that did understand were very much impressed. They tended to say things like: ''It helps you get into other things that your own organization (mental model or structure) isn't into.''

6.2. Negative Thesaurus/Concept Space Comments

In general, the negative feedback that we received about the thesaurus-based tool could be attributed to problems with mental models of the subjects being more attuned to alphabetic or hierarchical structures as opposed to term associations, or to criticisms with the user interface and limited functionality (no Boolean operators). In both cases, these criticisms are not criticisms of the technique per se but rather of the implementation of the technique.

• *Term Relationships.* Some subjects had trouble understanding how the terms suggested by the thesaurus were related to each other and to the input term(s). They said things like: ''If I click on soft (subject entered soft rock music as input), will it give me marshmallows?'' or ''Sometimes the words in the concept space didn't make sense to me. I didn't understand how they were related.''

• *User Interface Problems and Search Experience.* Several subjects felt that the tool needed clearer instructions. Some expressed concern that they would not have been able to use the thesaurus without some instruction or guidance from the experimenters. Their comments were: ''. . . difference between concept and document choices is not clear'' or ''instructions for exact and partial matches is not clear nor intuitively set up'' or ''It was somewhat confusing switching between the thesaurus and the URL part, sometimes I got confused.'' In general, these type of comments came from the more inexperienced searchers.

Sophisticated searchers were unhappy that the thesaurus was limited to ''OR'' searches; they wanted to be able to use other Boolean operators (for example, ''AND'' and ''NOT'') and other search-limiting abilities similar to the advanced Alta Vista searching capabilities. These searchers saw the value and the power of the thesaurus but wanted more functionality.

The less sophisticated searchers had a harder time with the thesaurus. In general, these subjects had very specific searches in mind and did not see the value of the thesaurus over a keyword search. A typical comment was: ''Since I already know what I want to see, why would I waste time looking in the thesaurus? Why wouldn't I go directly to the homepages I wanted?'' Another related comment was: ''I much prefer search engines. The amount of information on the Web is in general overwhelming. I like to find a few key sites and visit them regularly.''

One subject thought that the thesaurus would be best used as a browsing tool and that keywords were better for searching. He thought that the related terms could be used to see what other topics were related to the topic of interest.

7. Conclusions

7.1. Browsing Experiment

We believe that the results from the browsing experiment confirm our belief that a Kohonen SOM-based browsing mechanism can successfully categorize a large amount of information from a huge information space (in this case, the Internet). We further believe that this can be done in such a way that users are able to successfully locate information of interest. The technique is capable of organizing Internet homepage information into browsing areas that are of a manageable size. User feedback also clearly indicates that users like the graphical nature of the information organization, which may eventually be preferable to textually-based organization schemes.

The prototype ran into problems in three major areas: Mental models, cohesive source information, and user interface design issues. The reduced ability of subjects to recreate their browsing task in Yahoo!, after successfully locating an interesting homepage with the ET-Map, can mostly be explained by mental-model inconsistency. The conceptual model used by the SOM algorithm was unfamiliar to the subjects and did not match more traditional models of information organization (i.e., alphabetic and hierarchical). As several subjects pointed out, this problem can be partially overcome with more experience using the ET-Map and/or with some initial training in the use of the ET-Map. Based on our experiences with other implementations of the SOM algorithm, we feel that this technique is sensitive to the cohesion or lack of cohesion of the documents in the information space. The area that subjects had the easiest time browsing, music, clearly had a conceptual structure that was more consistent with subjects' mental models, and appeared to have a larger number of more cohesive homepages in it. For example, there were many homepages dedicated either to music in general, or musicians or groups in particular.

It is also clear that several user interface improvements need to be made to enhance the usability of the prototype. Potential future improvements requested by subjects include: A zoom-in/pan-out feature, running the collection program on a more frequent basis, improvements to the readability of the labels on the map, allowing the colors of the regions to assume significance, and perhaps providing an alternative to reading all the individual labels, such as an alphabetized index of the map's labels off to the side. Our research group is currently experimenting with several visualization techniques (fractals and fisheye views) to respond to the request for a way to enlarge or enhance a given region on the map.

7.2. Searching Experiment

We were extremely encouraged by the results of the searching experiment. The vast majority of the subjects were enthusiastic about the use of the thesaurus. The precision results indicated that the addition of thesaurus terms to subject terms did not significantly reduce the precision of the task. This means that the thesaurus terms added little to no superfluous noise (seen as additional irrelevant returned homepages) to the retrieval set. The recall results indicated that a combination of thesaurus terms and subject terms returned a more complete set of relevant homepages than either could on its own. This supports a conclusion that the thesaurus did not recommend terms that resulted in irrelevant homepages being retrieved, and that the homepages it was responsible for retrieving were, for the most part, different from the ones that the subject terms were responsible for retrieving. These findings support our supposition that an automatically generated thesaurus or concept space could be successfully used to supplement or improve a keyword-based search of the Internet.

Feedback from subjects indicated that the thesaurus was most useful in further refining a search that was initially too broad. Subjects particularly liked the fact that the terms suggested by the thesaurus came directly from the homepages themselves. This gave the subjects confidence that a more narrow search would actually retrieve relevant homepages. A common comparison was that when subjects tried to devise narrower keywords on their own, they frequently got no hits or hits that were not what they were interested in. Subjects also requested that the interface be changed so that a searcher could keep an initial term that was too broad as a subject category, and then choose more narrow terms from the thesaurus, but limit the search to within the category defined by the initial term. For example, if a user entered the term ''guitar,'' and then chose ''steel'' from the thesaurus, the search would be limited to the category of guitar (initially returned homepages that related to the term guitar) and then, within that set, find all homepages that were also related to the term steel). A few subjects also noticed that the thesaurus could also be used to widen a search that was initially too narrow.

Further feedback indicated other areas of improvement for the user interface. It needs either a set of help screens or instructions on how to use the thesaurus, or it needs to be more intuitive in its design (or both). Advanced searchers wanted more advanced query capabilities (e.g., more Boolean operators—''AND,'' ''OR,'' and ''NOT,'' hierarchical or category limitation, and wild carding ability).

We believe the first experiment indicates that an SOM-based Internet browsing prototype compares favorably with the hierarchical, hypertext browsing mechanism used by Yahoo! insofar as it enhances a subject's ability to locate a homepage of interest. We further believe that the second experiment indicates that an automatically generated thesaurus or concept space-based Internet searching mechanism compares extremely favorably with a simple keyword-based search and, in fact, can be used to improve it.

Anyone interested in experimenting with either prototype, the Kohonen SOM-based browsing tool (ET-Map) may contact the following Internet address: http://ai2.BPA.Arizona.EDU/ent/entertain1/

The automatically generated thesaurus or concept-based searching tool (Entertainment Thesaurus) is available at the following Internet address: http://ai.bpa.arizona.edu/cgi-bin/tng/ETSpace

8. Acknowledgments

This project was mainly supported by two grants: (1) A Digital Library Initiative grant awarded by NSF/ARPA/NASA (''Building the Interspace: Digital Library Infrastructure for a University Engineering Community,'' PIs: B. Schatz, H. Chen et al., 1994–1998, IRI-9411318), and (2) an NSF/CISE grant (''Concept-Based Categorization and Search on Internet: A Machine Learning, Parallel Computing Approach,'' PI: H. Chen, 1995–1998, IRI-9525790). We thank the following instructors for permitting us to recruit subjects from their classes: Akhilesh Bajaj (MIS 331), Divakaran Liginlal (MIS 111), and Randy Johnson (MIS 441). We also thank all of the University of Arizona students from these classes, and from the School of Library Science, who so generously donated their time, ideas, and suggestions.

References

Ahlswede, T., & Evens, M. (1988). Generating a relational lexicon from a machine-readable dictionary. *International Journal of Lexicography, 1(3)*, 214–237.

Bartell, B. T., Cottrell, G. W., & Belew, R. (1995). Representing documents using an explicit model of their similarities. *Journal of the American Society for Information Science, 46,* 254–271.

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

Bellamy, L. M., & Bickham, L. (1989). Thesaurus development for subject cataloging. *Special Libraries, 80(1)*, pp. 9–15.

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM, 28(3)*, 289–299.

Bowman, C. M. (1994). The Harvest information discovery and access system. *Computer Networks and ISDN Systems, 28(1–2)*, Dec. 1995, 119–125.

Bowman, C. M., Danzig, P. B., Manber, U., & Schwartz, F. (1994). Scalable internet resource discovery: Research problems and approaches. *Communications of the ACM, 37(8)*, 98–107.

Carlson, A., & Ram, S. (1990). HyperIntelligence: The next frontier. *Communications of the ACM, 33(3)*, 311–321.

Carmel, E., Crawford, S., & Chen, H. (1992). Browsing in hypertext: A cognitive study. *IEEE Transactions on Systems, Man and Cybernetics, 22(5)*, 865–884.

Chamis, A. Y. (1991). *Vocabulary control and search strategies in online searching.* New York: Greenwood Press.

Chaplan, M. A. (1995). Mapping Laborline thesaurus terms to Library of Congress subject headings: Implications for vocabulary switching. *Library Quarterly, 65(1)*, 39–61.

Chen, H. (1994). Collaborative systems: Solving the vocabulary problem. *IEEE Computer, 27(5)* [Special issue on computer-supported cooperative work (CSCW)], 58–66.

Chen, H., & Dhar, V. (1987). Reducing indeterminism in consultation: A cognitive model of user/librarian interaction. In H. Shrobe & K. Forbus (Eds.), *Proceedings of the 6th National Conference on Artificial Intelligence (AAAI-87),* Seattle, WA. (pp. 285–289). Menlo Park, CA: AAAI.

Chen, H., & Dhar, V. (1990). User misconceptions of online information retrieval systems. *International Journal of Man-Machine Studies, 32(6)*, 673–692.

Chen, H., & Dhar, V. (1991). Cognitive process as a basis for intelligent retrieval systems design. *Information Processing and Management, 27(5)*, 405–432.

Chen, H., & Lynch, K. J. (1992). Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics, 22(5)*, 885–902.

Chen, H., Lynch, K. J., Basu, K., & Ng, D. T. (1993). Generating, integrating, and activating thesauri for concept-based document retrieval. *IEEE Expert* [Special series on artificial intelligence in text-based information systems], *8(2)*, 25–34.

Chen, H., Martinez, J., Ng, D. T., & Schatz, B. R. (1997a). A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the Worm Community System. *Journal of the American Society for Information Science, 48,* 157–170.

Chen, H., & Ng, D. T. (1995). An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic branch-and-bound vs. connectionist Hopfield net activation. *Journal of the American Society for Information Science, 46,* 348–369.

Chen, H., Schatz, B. R., Ng, T. D., Martinez, J. P., Kirchhoff, A. J., & Lin, C. (1996a). 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, 18(8)*, 771–782.

Chen, H., Schatz, B. R., Yim, T., & Fye, D. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science, 46,* 175–193.

Chen, H., Schuffels, C., & Orwig, R. (1996b). Internet categorization and search: A machine learning approach. *Journal of Visual Communications and Image Representation, 7(1)*, 88–102.

Chen, H., Zhang, Y., & Houston, A. L. (1997b). Semantic indexing and searching using a Hopfield net (Tech. Rep.). Tucson, AZ: Department of MIS, College of Business and Public Administration, University of Arizona. Working paper.

Courteau, J. (1991). Genome databases. *Science, 254,* 201–207.

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

DeBra, P., & Post, R. (1994). Information retrieval in the World-Wide Web: Making client-based searching feasible. *Computer Networks and ISDN Systems, 27(2)*, Nov. 1995, 183–192.

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.

Ekmekcioglu, F. C., Robertson, A. M., & Willett, P. (1992). Effectiveness of query expansion in ranked-output document retrieval systems. *Journal of Information Science, 18,* 139–147.

Everitt, B. (1980). *Cluster analysis* (2nd ed.). London: Heinemann Educational Books.

Foss, C. L. (1989). Tools for reading and browsing hypertext. *Information Processing and Management, 25,* 407–418.

Fox, E. A. (1987). Development of the CODER system: A testbed for artificial intelligence methods in information retrieval. *Information Processing and Management, 23(4)*, 341–366.

Fox, E. A., Nutter, J. T., Ahlswede, T., Evens, M., & Markowitz, J. (1988). Building a large thesaurus for information retrieval. In B. Ballard (Ed.), *2nd Conference on Applied Natural Language Processing, Association for Computational Linguistics* (pp. 101–108). Morristown, NJ: Bell Communications Research.

Frenkel, K. A. (1991). The human genome project and informatics. *Communications of the ACM, 34(11)*, 41–51.

Furnas, G. W. (1982). Statistical semantics: How can a computer use what people name things to guess what things people mean when they name things? In *Proceedings of the Human Factors in Computer Systems Conference* (pp. 251–253). Gaithersburg, MD: Association for Computing Machinery.

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

Green, R. (1995). Topical relevance relationships. I. Why topic matching fails. *Journal of the American Society for Information Science, 46,* 646–653.

Green, R., & Bean, C. A. (1995). Topical relevance relationships. II. An exploratory study and preliminary typology. *Journal of the American Society for Information Science, 46,* 654–662.

Hayes-Roth, F., Waterman, D. A., & Lenat, D. (1983). *Building expert systems.* Reading, MA: Addison-Wesley.

Heine, M. H. (1995). An investigation of the relation influences of database informativeness, query size and query term specificity on the effectiveness of MEDLINE searching. *Journal of the American Society for Information Science, 21,* 173–185.

Knapp, S. D. (1984). Creating BRS/TERM, a vocabulary database for searchers. *Database, 7(4),* 70–75.

Liebscher, P., & Marchionini, G. (1988). Browse and analytical search strategies in a full-text cd-rom encyclopedia. *School Library Media Quarterly, 16,* 223–233.

Lin, X., Soergel, D., & Marchionini, G. (1991). A self-organizing semantic map for information retrieval. In A. Bookstein, Y. Chiaramella, G. Salton, & V. V. Raghavan (Eds.), *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval,* Chicago, IL (pp. 262–269). New York: ACM Press.

Lindberg, D. A., & Humphreys, B. L. (1990). The UMLS knowledge sources: Tools for building better user interface. In R. Miller (Ed.), *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care* (pp. 121–125). Los Alamitos, CA: Institute of Electrical and Electronics Engineers.

Lippmann, R. P. (1987). An introduction to computing with neural networks. *IEEE Acoustics Speech and Signal Processing Magazine, 4(2),* 4–22.

Marchionini, G. (1987). An invitation to browse: Designing full text systems for novice users. *Canadian Journal of Information Science, 12(3),* 69–79.

Marchionini, G., & Shneiderman, B. (1988). Finding facts vs. browsing knowledge in hypertext systems. *IEEE Computer, 21(3),* 70–79.

McCain, K. W. (1995). Biotechnology in context: A database-filtering approach to identifying core and productive non-core journals supporting multidisciplinary R and D. *Journal of the American Society for Information Science, 46,* 306–319.

McCray, A. T., & Hole, W. T. (1990). The scope and structure of the first version of the UMLS semantic network. In R. Miller (Ed.), *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care* (pp. 126–130). Los Alamitos, CA: Institute of Electrical and Electronics Engineers.

Nadis, S. (1996). Computation cracks semantic barrier between databases. *Science, 272,* 1419.

Niehoff, R. T. (1976). Development of an integrated energy vocabulary and the possibilities for on-line subject switching. *Journal of the American Society for Information Science, 27,* 3–17.

Niehoff, R. T., & Kwansy, S. (1979). The role of automated subject switching in a distributed information network. *Online Review, 3(2),* 181–194.

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.

Peat, H. J., & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science, 42,* 378–383.

Petersen, T. (1983). The AAT: A model for the restructuring of LCSH. *Journal of Academic Librarianship, 9(4),* 207–210.

Petersen, T. (1990). Developing a new thesaurus for art and architecture. *Library Trends, 38(4),* 644–658.

Pinkerton, B. (1994). Finding what people want: Experiences with the WebCrawler. In *Proceedings of the Second International World Wide Web Conference '94,* Chicago, IL. Urbana-Champaign, IL: NCSA.

Piternick, A. B. (1984). Searching vocabularies: A developing category of online search tools. *Online Review, 8(5),* 441–449.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing* (pp. 227–270). Cambridge, MA: The MIT Press.

Rasmussen, E. (1992). Clustering algorithms. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms* (pp. 419–442). Englewood Cliffs, NJ: Prentice Hall.

Salton, G. (1989). *Automatic text processing.* Reading, MA: Addison-Wesley.

Schatz, B. R., & Chen, H. (1996). Building large-scale digital libraries. *IEEE Computer, 29(5),* 22–26.

Schatz, B. R., Mischo, B., Cole, T., Hardin, J., Bishop, A., & Chen, H. (1996). Federating repositories of scientific literature: The Illinois digital library project. *IEEE Computer, 29(5),* 28–36.

Smeaton, A. F., & van Rijsbergen, C. J. (1983). The effects of query expansion on a feedback document retrieval system. *Computer Journal, 26,* 239–246.

Voorhees, E. (1993). Using WordNet to disambiguate word senses for text retrieval. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), *Proceedings of the ACM SIGIR,* Pittsburgh, PA (pp. 171–180). New York: ACM Press.

Wang, Y. -C., Vandendorpe, J., & Evens, M. (1985). Relational thesauri in information retrieval. *Journal of the American Society for Information Science, 36,* 15–27.