

Internet Categorization and Search: A Self-Organizing Approach

Hsinchun Chen,¹ Chris Schuffels,² and Richard Orwig³

Management Information Systems Department, University of Arizona, Tucson, Arizona 85721

Received July 6, 1995; accepted December 5, 1995

The problems of information overload and vocabulary differences have become more pressing with the emergence of increasingly popular Internet services. The main information retrieval mechanisms provided by the prevailing Internet WWW software are based on either keyword search (e.g., the Lycos server at CMU, the Yahoo server at Stanford) or hypertext browsing (e.g., Mosaic and Netscape). This research aims to provide an alternative concept-based categorization and search capability for WWW servers based on selected machine learning algorithms. Our proposed approach, which is grounded on automatic textual analysis of Internet documents (homepages), attempts to address the Internet search problem by first *categorizing* the content of Internet documents. We report results of our recent testing of a multilayered neural network clustering algorithm employing the Kohonen self-organizing feature map to categorize (classify) Internet homepages according to their content. The category hierarchies created could serve to partition the vast Internet services into subject-specific categories and databases and improve Internet keyword searching and/or browsing. © 1996 Academic Press, Inc.

1. INTRODUCTION

Despite the usefulness of database technologies, users of online information systems are often overwhelmed by the amount of current information, the subject and system knowledge required to access this information, and the constant influx of new information [11]. The result is termed “information overload” [3]. A second difficulty associated with information retrieval and information sharing is the classic vocabulary problem, which is a consequence of diversity of expertise and backgrounds of system users [29, 30, 9]. The fluidity of concepts and vocabularies in various domains further complicates the retrieval issue [9, 26, 18]. A concept may be perceived differently by different searchers and it may also convey different meanings at different times. To address the information overload and the vocabulary problem in a large information space

that is used by searchers of varying backgrounds a more intelligent and proactive search aid is needed.

The problems of *information overload* and *vocabulary differences* have become more pressing with the emergence of increasingly popular Internet services [47, 24]. Although Internet protocols such as WWW/http support significantly easier importation and fetching of online information sources, their use is accompanied by the problem of users not being able to explore and find what they want in an enormous information space [2, 6, 55]. While the Internet services are popular and appealing to many online users, difficulties with search on Internet, we believe, will worsen as the amount of online information increases. We consider that devising a scalable approach to Internet search is critical to the success of Internet services and other current and future national information infrastructure applications.

The main information retrieval mechanisms provided by the prevailing Internet WWW-based software are based on either keyword search (e.g., the Lycos server at CMU and the Yahoo server at Stanford) or hypertext browsing (e.g., NCSA Mosaic and Netscape browser). Keyword search often results in relatively low precision and/or poor recall, as well as slow response time due to the limitations of the indexing and communication methods (bandwidth), controlled language based interfaces (the vocabulary problem), and the inability of searchers themselves to fully articulate their needs. Furthermore, browsing allows users to explore only a very small portion of the large Internet information space. An extensive information space accessed through hypertext-like browsing can also potentially confuse and disorient its user, resulting in the embedded digression problem, and can cause the user to spend a great deal of time while learning nothing specific, the art museum phenomenon [8, 25].

Internet “surfers” also have begun to raise their expectations of the Internet services—from a simple desire to find something fun (for no particular reason) to hoping to find something that might be useful (to their work or personal interests). For example, the Lycos server at CMU has become one of the hottest and most popular servers on the Internet due to its comprehensive listing and indexing of

¹ E-mail: hchen@bpa.arizona.edu.

² E-mail: cschuffels@bpa.arizona.edu.

³ E-mail: rorwig@bpa.arizona.edu.

Internet homepages (10+ million URLs in October 1995 and growing) and (keyword) search capability. However, the Lycos server has been hampered severely by the information overload and communication bandwidth problems discussed above.

Our proposed approach, which is grounded on automatic textual analysis of Internet documents (homepages), aims to address the Internet search problem by first automatically *categorizing* the content of Internet documents and subsequently providing category-specific search capabilities. As the first step to intelligent categorization and search for Internet, we proposed a multilayered neural network clustering algorithm employing a Kohonen self-organizing feature map to categorize (classify) the Internet homepages according to their content. The category hierarchies could serve to partition the vast Internet services into subject-specific categories and databases. After individual subject categories had been created, subject-specific searches or browsing could be performed.

In Section 2, we first present an overview of machine learning techniques for information retrieval. We then review the current status of Internet categorization and searching and associated problems. In Section 3, we present our framework for addressing these problems. Sections 4 and 5 discuss the specific algorithm and results from our ongoing Internet categorization research. Conclusions and discussion are provided in Section 6.

2. INTERNET CATEGORIZATION AND SEARCH: TECHNIQUES AND PROBLEMS

2.1. Machine Learning for Information Retrieval

Searching is a concept frequently discussed in the context of information retrieval research. In this section, we provide a brief summary of the emerging machine learning approach to searching. For a complete review of other techniques, readers are referred to [10].

Inductive machine learning techniques have drawn attention from researchers in computer and information sciences in recent years. In particular, Doszkoecs *et al.* [22] have provided an excellent review of connectionist models for information retrieval and Lewis [37] and Chen [10] have surveyed and experimented with various machine learning algorithms in information retrieval and discussed promising areas for future research at the intersection of these two fields.

Neural Networks and IR. Neural networks computing, in particular, seems to fit well with conventional retrieval models such as the vector space model [54] and the probabilistic model [43]. The work of Belew is probably the earliest connectionist model adopted in IR. In AIR [1], he developed a three-layer neural network of authors, index terms, and documents. The system used relevance feed-

back from its users to change its representation of authors, index terms, and documents over time. Kwok [36] also developed a similar three-layer network of queries, index terms, and documents. A modified Hebbian learning rule was used to reformulate probabilistic information retrieval. Wilkinson and Hingston [59, 60] incorporated the vector space model in a neural network for document retrieval. Their network also consisted of three layers: queries, terms, and documents.

While the above systems represent information retrieval applications in terms of their main components of documents, queries, index terms, authors, etc., other researchers have used different neural networks for more specific tasks. Lin [39] adopted a Kohonen network for information retrieval. Kohonen's self-organizing feature map (SOM), which produced a two-dimensional grid representation for N -dimensional features, was applied to construct a self-organizing (unsupervised learning) visual representation of the semantic relationships between input documents. The Kohonen SOM approach was further extended by Orwig and Chen [48] to analyze electronic meeting comments and present graphical representation of group consensus. In [41], a neural algorithm developed by MacLeod was used for document clustering. The algorithm compared favorably with conventional hierarchical clustering algorithms. Chen *et al.* [13–15] reported a series of experiments and system developments which generated an automatically created weighted network of keywords from large textual databases and integrated it with several existing man-made thesauri (e.g., LCSH). Instead of using a three-layer design, Chen's systems developed a single-layer, interconnected, weighted/labeled network of keywords (concepts) for "concept-based" information retrieval. A blackboard-based design which supported browsing and automatic concept exploration using the Hopfield neural network's parallel relaxation method was adopted to facilitate the use of several thesauri [14]. In [15] the performance of a branch-and-bound serial search algorithm was compared with that of the parallel Hopfield network activation in a hybrid neural-semantic network. Both methods achieved similar performance, but the Hopfield activation method appeared to activate concepts from different networks more evenly.

Symbolic Learning and IR. Despite the popularity of using neural networks for information retrieval, we see only limited use of symbolic learning techniques for IR. In [4], the researchers used discriminant analysis and a simple symbolic learning technique for automatic text classification. Their symbolic learning process represented the numerical classification results in terms of IF-THEN rules. Text classification involves the task of classifying documents with respect to a set of two or more predefined classes [38]. A number of systems have been built based

on human categorization rules (a knowledge-based system approach) [52]. However, a range of statistical techniques including probabilistic models, factor analysis, regression, and nearest neighbor methods also have been adopted [38, 44, 4]. Fuhr *et al.* [27] adopted regression methods and ID3 for their feature-based automatic indexing technique. Crawford, Fung, and their co-workers [28, 19, 20] have developed a probabilistic induction technique called CONSTRUCTOR and have compared it with the popular CART algorithm [7]. Their experiment showed that CONSTRUCTOR's output is more interpretable than that produced by CART, but CART can be applied to more situations (e.g., real-valued training sets). In [17], Chen and She adopted ID3 and the incremental ID5R algorithm for information retrieval. Both algorithms were able to use user-supplied samples of desired documents to construct decision trees of important keywords which could represent the users' queries.

Genetic Algorithms and IR. Our literature search revealed several implementations of genetic algorithms in information retrieval. In [31], Gordon presented a genetic algorithms based approach to document indexing in which competing document descriptions (keywords) are associated with a document and altered over time by using genetic mutation and crossover operators. In his design, a keyword represents a gene (a bit pattern), a document's list of keywords represents individuals (a bit string), and a collection of documents initially judged relevant by a user represents the initial population. Based on a Jaccard matching function (fitness measure), the initial population evolved through generations and eventually converged to an optimal (improved) population—a set of keywords which best described the documents. In [32], Gordon adopted a similar approach to document clustering. Raghuvan and Agarwal [50] have also studied genetic algorithms in connection with document clustering. In [49], Petry *et al.* applied genetic programming to a weighted information retrieval system. In their research, a weighted Boolean query was modified in order to improve recall and precision. They found that the form of the fitness function has a significant effect upon performance. Yang and his co-workers [61, 62] have developed adaptive retrieval methods based on genetic algorithms and the vector space model using relevance feedback. They reported the effects of adopting genetic algorithms in large databases, the impact of genetic operators, and GA's parallel searching capability. In [12], a GA-NN hybrid system, called GANNET, was developed by Chen and Kim for IR. The system performed *concept optimization* for user-selected documents using genetic algorithms. It then used the optimized concepts to perform *concept exploration* in a large network of related concepts through the Hopfield net parallel relaxation procedure. A Jaccard's score was also adopted to compute the "fitness" of subject descriptions for information retrieval.

Based on our experience in this area, we believe that the new and emerging machine learning algorithms that analyze the common characteristics of documents and retrieval patterns of searchers are promising and may provide a viable solution to the complex and large-scale Internet categorization and search problem.

2.2. Internet Categorization and Search: An Overview

In its roots as the ARPANET, the Internet was conceived primarily as a means for remote login and experimentation with telecommunication [6]. However, the predominant usage quickly became e-mail communication. This trend continues into the present form of the Internet, but with increasingly diverse support for collaborative data sharing and distributed, multimedia information access, especially using the World-Wide Web (WWW). Many people consider the Internet and the WWW the backbone of the Information Superhighway and the window to Cyberspace.

The WWW was developed initially to support physicists and engineers at CERN, the European Particle Physics Laboratory in Geneva [2]. In 1993, when several browser programs (most noticeably the NCSA Mosaic) became available for distributed, multimedia, hypertext-like information fetching, Internet became the preview of a rich and colorful information cyberspace [55]. However, as Internet services based on WWW have become more popular, information overload has become a pressing research problem [6]. The user interactions paradigm on Internet has been shifted from simple hypertext-like *browsing* (human-guided activity exploring the organization and contents of an information space) to content-based *searching* (a process in which the user describes a query and a system locates information that matches the description). Many researchers and practitioners have considered Internet searching to be one of the more pressing and rewarding areas of research for future NII applications.

Internet searching has been the hottest topic at recent World-Wide Web Conferences. Two major approaches have been developed and experimented with: one is the client-based search spider (agent) and the other is online database indexing and searching. However, many systems contain components of both approaches.

Client-Based Search Spiders (Agents). Several software programs based on the concept of spiders, agents, or softbots (software robots) have been developed. tueMosaic and the WebCrawler are two prominent examples.

DeBra and Post [21] reported tueMosaic v2.42, modified at the Eindhoven University of Technology (TUE) using the Fish Search algorithm, at the First WWW Conference in Geneva. Using tueMosaic, users can enter keywords, specify the depth and width of search for links contained in the homepage currently displayed, and request the spider

agent to fetch homepages connected to the current homepage. However, potentially relevant homepages that do not connect with the current homepage cannot be retrieved and the search space becomes enormous when the depth and breadth of search become large (an exponential search). The inefficiency and local search characteristics of the BFS/DFS-based spiders and the communication bandwidth bottleneck on Internet severely constrained the usefulness of such an agent approach. At the Second WWW Conference, Pinkerton reported a more efficient spider (crawler). The WebCrawler extends the tueMosaic's concept to initiate the search using its index and to follow links in an intelligent order. Despite its refinement, the local search and communication bottleneck problems persist. A more efficient and global Internet search algorithm is needed to improve client-based searching agents.

Online Database Indexing and Searching. An alternative approach to Internet resource discovery is based on the database concept of indexing and keyword searching.

The World Wide Web Worm (WWWW) [46], developed by McBryan, uses crawlers to collect Internet homepages. After homepages are collected, the system indexes their subject headings and anchor texts and provides UNIX grep-like routines for database keyword searching. However, WWWW does not create a large collection of documents in the database, due to the limitation of its crawlers and its grep-like keyword searching. AliWeb [35], developed by Koster, adopts an owner-registration approach to collecting Internet homepages. Homepage owners write descriptions of their services in a standard format to register with AliWeb, which regularly retrieves all homepages included in its registration database. AliWeb alleviates the spider traversal overhead. However, the manual registration approach places a special burden on homepage owners and thus is unpopular. The AliWeb database is small and too incomplete for realistic searches.

The Harvest information discovery and access system [5], developed by Bowman *et al.*, presents a big leap forward in Internet searching technology. Harvest includes a Gatherer designed to run at the service provider's site, saving a great deal of server load and network traffic. Similar to AliWeb, it also allows service owners to determine which documents are worth indexing. Harvest also includes Agrep and Glimpse for fuzzy pattern matching and significantly cuts down the index size.

One of the most comprehensive and useful searchable Internet databases is Lycos at CMU [45]. Lycos uses a combination of spider fetching and simple owner registration. Internet servers can access the Lycos server and complete registration in a few simple steps. No extensive registration form like that requested for AliWeb is required. In addition, Lycos uses spiders based on the connections to the registered homepages to identify other unregistered

homepages. With this suite of techniques, Lycos has acquired an impressive list of URLs on the Internet (10+ million URLs in October, 1995). Its homepage growth rate has been exponential over the past 6–10 months. Lycos adopted a heuristics-based indexing approach for the homepages it acquired (only the first 20 lines of text in a homepage are indexed). However, Lycos's success also shows the vulnerability of the approach and the daunting task of creating "intelligent" and efficient Internet search engines. Its popularity has caused a severe degradation of information access performance, due to the communication bottleneck and the task of finding selected documents in an all-in-one database of Internet homepages. An automatic and robust method of partitioning the Lycos database based on its subject content is needed to assist in more efficient and fruitful Internet searching.

The Yahoo server developed at the Stanford University represents one attempt to partition the Internet information space and provide meaningful subject categories (e.g., science, entertainment, engineering, etc.). However, the subject categories are limited in their granularity and the process of creating such categories is a manual effort. The demand to create up-to-date and fine-grained subject categories and the requirement that an owner place a homepage under a proper subject category has significantly hampered Yahoo's success and popularity. Only about 100,000 servers have been registered in November 1995 and placed in the Yahoo subject directory. We believe an automatic, machine learning approach based on content analysis of large-scale Internet documents could help alleviate some of these Internet categorization and classification problems and provide a scalable solution to efficient and fruitful concept-based Internet searching.

3. A FRAMEWORK FOR INTERNET CATEGORIZATION AND SEARCH

In this section we present our overall framework and design for Internet categorization and search and the specific machine learning algorithms to use. Relevant findings and preliminary testing results from our own research will also be presented. Status of our current research (i.e., Stage 1. Internet categorization) will be reported in the next section. Our proposed design consists of three stages, to be executed consecutively.

1. Stage 1. Internet Categorization Using a Multilayered Kohonen Self-Organizing Feature Map. In order to improve the efficiency of searching on Internet, the first task is to partition the Internet information space into distinct subject categories meaningful to Internet users. Categorization and subject classification are common practices in library and information sciences (e.g., the INSPEC database for the computer engineering domain and the ERIC database for sociology). Subject partitioning creates

smaller databases, which are more efficient for searching. In addition, a subject directory created as a result of a categorization or classification can also aid searchers' "directory-browsing," a searcher-guided information seeking behavior frequently seen in the previously popular Gopher information servers. Many searchers of the Yahoo database adopted a combination of directory browsing and keyword searching within the specific subcategory of homepages.

After examining several clustering algorithms in the areas of computer science (e.g., hierarchical and nonhierarchical methods) and neural network algorithms in our previous research (to be discussed in detail below), we concluded that a variant of the Kohonen self-organizing feature maps (SOM) appears promising. The algorithm has been shown to be robust in numerous image processing and pattern recognition applications [56]. It also creates an intuitive, graphical display of important concepts contained in textual information [39, 48].

A multilayered graphical SOM approach to Internet categorization have been adopted for this research. By analyzing keywords/descriptors in Internet homepages and their probabilities of co-occurrence, we should be able to represent the most important Internet subject categories (e.g., science, engineering, business, politics, entertainment, etc.) in different regions of a map. For each large region, a recursive process of analyzing homepages in the region and creating submaps could then be undertaken. Because each map might contain 30–50 categories, 5–6 layers of maps could then easily represent a number of homepages on the order of ten to hundred millions (e.g., 30^5 – 50^6). After subject categories had been created, searchers would be able to browse the subject directory to locate the appropriate partition in which to launch their keyword searching.

2. *Stage 2. Concept-Based Search Based on Cluster Analysis and Hopfield Net Associative Retrieval.* In addition to keyword searching in a subject category, we propose a concept space approach to information retrieval. By analyzing the co-occurrence probabilities of keywords in homepages of specific subject categories, we could create a *concept space* for each subject category. Such a concept space would represent the important terms and their weighted relationships in a graph structure, akin to an associative manmade thesaurus. A system-created concept space has been shown to be an effective tool to suggest alternative terms for searching and to articulate and reformulate precise queries during information retrieval. 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. Despite some occurrences of obvious noise, the system was useful in suggesting relevant

concepts for the researchers' queries and it helped improve concept *recall* [16].

The success of the concept space approach has been shown in various domain-specific applications such as Russian computing [13] and molecular biology [16]. However, the usefulness of such an approach in accessing the diverse and large-scale Internet servers remains to be tested. After concept spaces (graphs) have been created for each subject category, we plan to incorporate into Internet searches several graph traversal algorithms previously tested in other applications, e.g., branch-and-bound and Hopfield net association [15]. We believe the thesaurus-like concept spaces created automatically for each subject category will serve as an excellent memory-jogging and term suggestion aid for searchers on Internet.

3. *Stage 3. Intelligent Spider (Agent) Using Genetic Algorithm.* In addition to enhancing Internet categorization and concept-based IR capabilities, the subject categories created during the categorization process could also be used to develop "intelligent" global-search spiders (agents) for more efficient and optimal client-based search of relevant Internet information.

Based on our experience with various serial and parallel search algorithms and the analysis of the characteristics of the Internet structure, a genetic algorithm-based spider is proposed. By following homepages linked to starting homepages (a form of crossover) and performing sampling on a category-specific list of all other potentially relevant homepages (a form of mutation), a stochastic process of global evolution toward the "fittest" (the most similar) homepages can be achieved. We believe that this GA-based search algorithm is efficient and that using it will help us obtain optimal global search results on Internet that are based on users' preferences (i.e., identify a list of homepages most relevant to the user-supplied starting homepages). Such a genetic algorithms-based approach has been adopted successfully in recent "intelligent agent" research for human-computer interaction design [42] and for inductive query by examples [12].

4. RESEARCH DESIGN AND FINDINGS FOR INTERNET CATEGORIZATION AND SEARCH

The specific system design and research findings adopted for Stage 1 of our research are reported below.

4.1. Research Design: Multilayered Self-Organizing Feature Maps (M-SOM)

Categorization and classification are processes which involve clustering/grouping items of similar nature. Tagging similar items with meaningful labels (names) results in subject categories. When pairwise similarities are obtained between items, a hierarchical agglomerative cluster gener-

ation process can be adopted (a process often used in information science [54]). Several serial clustering algorithms exist, e.g., *single-link clustering* and *complete-link clustering* (based on popular minimal-spanning tree algorithms such as Prim's and the Kruskal's) [58, 53, 51]. While these methods have demonstrated their usefulness in clustering documents, a somewhat newer and more promising approach based on the connectionist paradigm, or neural network computing, has attracted a resurgence of interest [33, 57]. There are several reasons for this, including the appearance of faster digital computers on which to simulate large networks, interest in building massively parallel computers, and, most important, the development of more powerful neural network architectures and algorithms.

Kohonen's self-organizing feature maps [40, 34], in particular, have drawn significant attention in various engineering and scientific domains. In the basic form, continuous-valued vectors are presented sequentially in time without specifying the desired output. After enough input vectors have been presented, network connection weights will specify cluster or vector centers that sample the input space so that the point density function of the vector centers tends to approximate the probability density function of the input vectors. In addition, the connection weights will be organized so that topologically close nodes are sensitive to inputs that are physically similar. Lin [39] first adopted the Kohonen SOM for information retrieval. In his prototype, he generated self-organizing clusters of important concepts in a small database of several hundred documents.

In order to organize the large number of homepages (10M+) on Internet, we proposed a multilayered SOM algorithm, which permitted unlimited layers of Kohonen maps (we refer to it as M-SOM). A sketch of our proposed M-SOM algorithm is presented below:

1. *Initialize Input Nodes, Output Nodes, and Connection Weights.* Use the top (most frequently occurring) N terms (say 1000) from all homepages as the input vector and create a two-dimensional map (grid) of M output nodes (say a 20-by-10 map of 200 nodes). Initialize weights from N input nodes to M output nodes to small random values.

2. *Present Each Document (Homepage) in Order.* Represent each document (homepage) by a vector of N terms and present to the system.

3. *Compute Distances to All Nodes.* Compute distance d_j between the input and each output node j using

$$d_j = \sum_{i=0}^{N-1} (x_i(t) - w_{ij}(t))^2$$

where $x_i(t)$ is the input to node i at time t and $w_{ij}(t)$ is the weight from input node i to output node j at time t .

4. *Select Winning Node j^* and Update Weights to Node j^* and Neighbors.* Select winning node j^* as that output node with minimum d_j . Update weights for node j^* and its neighbors to reduce their distances (between input nodes and output nodes). (See [34, 40] for the algorithmic detail of neighborhood adjustment.)

5. *Label Regions in Map.* After the network is trained through repeated presentation of all homepages (each homepage is presented at least five times), submit unit input vectors of single terms to the trained network and assign the winning node the name of input term. Neighboring nodes which contain the same name/term then form a concept/topic region (group). Similarly, submit each homepage as input to the trained network again and assign it to a particular node in the map. The resulting map thus represents regions of important terms/concepts (the more important a concept, the larger a region) and the assignment of homepages to each region. Concept regions that are similar (conceptually) will also appear in the same neighborhood.

6. *Apply the Above Steps Recursively for Large Regions.* For each map region which contains more than k (say 100) homepages, conduct a recursive procedure of generating another self-organizing map until each region contains no more than k homepages.

We believe that, with 5–6 layers of self-organizing maps and a simple subject category browsing interface, we can partition Internet resources into meaningful and manageable sizes, ready for hypertext browsing and/or category-specific searching.

4.2. Research Findings

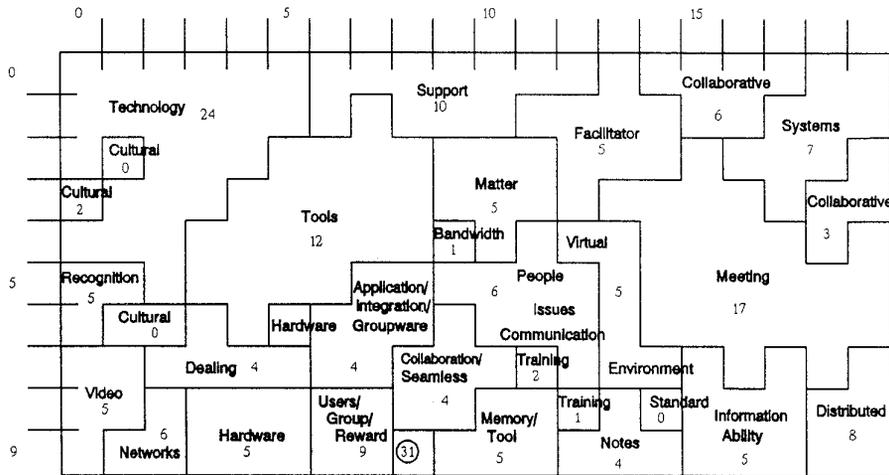
We have adopted the proposed algorithm in various applications, which varied in sizes: electronic brainstorming comments (10+ KBs, several hundred comments) and Internet entertainment-related homepages (3+ MBs, 10,000+ homepages). In this section, we report the results and status of our system implementation and evaluation.

4.2.1. Categorizing Electronic Brainstorming Comments

In [48], Orwig describes research in the application and evaluation of a Kohonen SOM algorithm to the problem of categorization of brainstorming output using electronic meeting systems.

A major advantage of the meeting software is its ability to let meeting participants brainstorm ideas in a parallel mode. Brainstormers can sit around a table and "talk" at the same time, using their keyboards. Often as many as

Kohonen Self-Organizing Map EBS Output



EBS Question: What are the most important information technology problems with respect to Collaborative Systems to be solved over the next five years?

FIG. 1. SOM-generated list of topics.

several hundred comments can be generated by a group of 10–20 meeting participants during a typical 1-h electronic brainstorming (EBS) session. While meeting software has been shown to be extremely useful for *idea generation*, a *divergent task*, the process of categorizing crucial ideas embedded in meeting comments and generating a consensus list of important topics (*idea categorization*), a *convergent task*, is more difficult.

Because of the relatively small number of data required of the categorization process (often on the order of 10–30 KBs or 50–300 comments), a single-layered SOM algorithm was developed and tested. Figure 1 shows the SOM output of actual brainstorming comments in an electronic meeting session. Twenty group participants, who included managers and users of GroupSystems (an electronic meeting system developed at the University of Arizona) from various companies and government agencies, were asked to use the electronic brainstorming tool to respond to the following question: *What are the most important information technology problems with respect to Collaborative Systems to be solved over the next five years?* The group generated 201 comments over a period of 30 min. While the participants were responding to the question, an expert group facilitator utilized an existing Categorizer tool to produce a list of the major topics addressed by the respondents (by manually browsing the participants' comments). When he recognized that a new concept was appearing frequently in the comments, it was added to his list, and relevant comments were attached to the topic. The expert spent

the entire 30 min during the EBS session, plus more time during break, to arrive at a list of 20 items.

In a recent experiment [48], we compared a facilitator-generated list of topics with one generated by SOM. Eight facilitator subjects participated in the experiment. Subjects were given the text output of the brainstorming session results and the two lists. Subjects read through the actual comments first and then corrected each list by deleting inappropriate topics and adding topics that they thought were missing. The resulting lists of topics were then used to compute the recall and precision levels of the two lists. Statistical results for hypothesis testing were also obtained.

On an average (with sample size $N = 8$), the facilitator list obtained an 81% precision level and the SOM list obtained a 55% precision level. The difference was statistically significant (at 5% significance level). Compared with human facilitators, the SOM algorithm was less precise in generating topics. In recall, the facilitator list reached an 88.5% level and the SOM list reached an 81% level. The difference was statistically insignificant. However, the SOM algorithm took significantly less time to produce a list of topics (45 min for the facilitator and 4 min for the system). Considering the cognitive demand for generating topics manually and the prospect of using the SOM output as an information visualization and decision aid (i.e., using the SOM output as a straw-man list for further user refinement), we believe the results from this experiment were encouraging. It suggests an efficient, algorithmic alternative for information searchers or sys-

tem users. Based on this initial research, we proceeded to test the SOM approach in several other larger Internet applications (where human categorization becomes even more difficult).

4.2.2. *Categorizing Internet Entertainment Homepages*

In order to examine the scalability of the SOM approach to Internet categorization, we created a testbed of about 10,000 Internet homepages related to entertainment, using the Yahoo server. (We developed a spider/softbot that traversed and fetched the homepages at the top three layers of the entertainment section of the Yahoo directory.) The resulting testbed was about 3 MB in size. The experiment aimed to use the M-SOM approach to classify the 10,000+ homepages into meaningful, multilayered categories.

The first-layer SOM process took 1 h and 37 min on a DEC Alpha 3000/600 (200 MHz, 128 MB RAM) and produced about 50 regions (groups) on the map. Using 100 homepages as the threshold for further SOM categorization, the second-layer SOM process took about 1 h and 21 min. The SOM categorization process for the 10,000+ entertainment homepages ended after four levels.

The computational characteristics and initial output produced for the entertainment homepage analysis are interesting. We observed that many of the larger concept regions appeared to be meaningful and to relate to each other (e.g., SAN FRANCISCO and LOS ANGELES form neighboring regions). Initial browsing of the homepages in a concept region showed relevant homepages; e.g., we found Star Trek fan homepages under the SCIENCE FICTION concept region.

A sample concept browsing WWW server using SOM was developed recently. The resulting *ET-Map* server contains about 50 concept regions at the first layer and is available at <http://ai.bpa.arizona.edu/ent/et-map.html>. The large concept regions (using 100 URLs as the threshold) can be clicked on to produce sub-regions. For example, Fig. 2 shows the top-level map for all entertainment homepages. By clicking on the STAR TREK (503 URLs) concept region of the top-level map, the system displayed a sub-map which contained SCIENCE FICTION (23 URLs), as shown in Fig. 3. Clicking on the SCIENCE FICTION region resulted in a ranked list of URLs summarized with titles and top keywords (Fig. 4). Each URL is “live” and can be clicked on the fetch an actual homepage. We felt that this experimental server was interesting enough for some initial user evaluation.

5. USER EVALUATION: AN EXPERIMENT ON THE ENTERTAINMENT MAP

In order to assess the usefulness of the ET-Map for Web browsing, we designed a qualitative experiment based

upon protocol analysis [23]. Our research goal was to understand the characteristics of the SOM output as demonstrated in the ET-Map server and its potential for becoming an alternative for concept browsing and searching for WWW services, using the Yahoo entertainment directory as the benchmark for comparison.

5.1. Experimental Design

The experiment involved 10 subjects: 5 graduate students from the Library Science Department at the University of Arizona, 4 graduate students from the MIS Department, and one System Administrator, also from the MIS Department. The subjects compared the ET-Map (<http://ai.bpa.arizona.edu/ent/et-map.html>) created by the Kohonen SOM algorithm to the manually catalogued Entertainment hierarchy of Yahoo (<http://www.yahoo.com/Entertainment/>). The subjects were asked to perform three searches twice, once by searching Yahoo’s Entertainment hierarchy and the other by searching the ET-Map. The subjects verbalized their thought processes and comments while searching and the experimenters collected the protocols for analysis.

The searches were all performed using Netscape 1.1 or higher on either an X-terminal or a Macintosh. The only “search” mechanism which subjects could use was the Netscape Find function, which searches the text on the loaded page for the input word. No instructions were given to the subjects as to whether the searches should be broad (any home pages on a specific subject) or narrow (a particular page). Due to the open endedness of the experiment and the nature of browsing the WWW, some of the searches became extensive and unsuccessful (most subjects would abandon an unsuccessful search after approximately 10 min). At the other end of the spectrum, a few of the searches remained relatively short and successful in under a minute, with the majority of the search times falling between the two extremes. All subjects completed the experiment within 1 hour and 15 min.

5.2. Experimental Results

5.2.1. *Patterns of Problems Common to Both Tools*

Subjects Became “Lost” in Both the ET-Map and the Yahoo Entertainment Hierarchy. The Yahoo Entertainment hierarchy was problematic because subjects would become involved with browsing and they would pay little attention to where they were in the hierarchy. Yahoo lists the entire hierarchical division at the top of every page, but subjects would become so engrossed in their searches that they would not think to look up. For example, a subject had spent some time searching for comedy shows, and was located at ENTERTAINMENT:TELEVISION:COMEDIES in the Yahoo Enter-

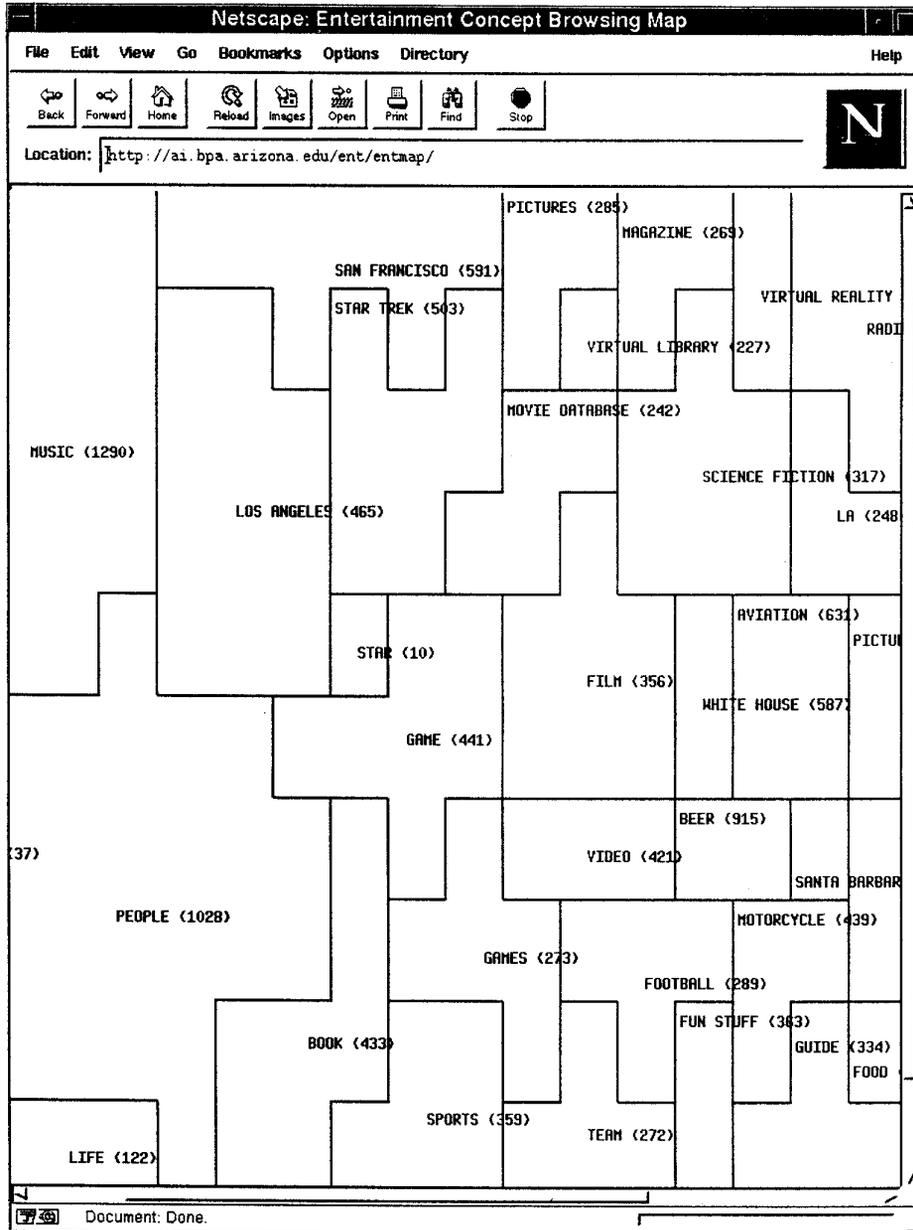


FIG. 2. ET-Map, top-most layer.

tainment hierarchy. The subject spent at least 10 min looking at different comedy shows, before turning her attention to MUSIC. However, she did not realize that she was still in COMEDIES and questioned why there were no listings for Pop or Classical. At that point the subject looked up and realized where she was located, commenting, "Oh, I need to go back."

Although subjects became lost and were unsure of where they were in the Yahoo hierarchy and the ET-Map, this disorientation caused more havoc for subjects when they searched the ET-Map. Unlike Yahoo, the ET-Map does

not provide map titles to identify its various levels. Although the areas of the maps are labeled, the maps themselves are not. Some of the map location problems also occur because the area labels are longer than the area. For example, from the first level several subjects chose the area labeled SPACE, but thought they were choosing the area labeled ENTERTAINMENT.

Subjects Did Not Understand the Logic and Organization behind Either Tool. Although it is apparent that Yahoo presents a hierarchy, there is no explanation

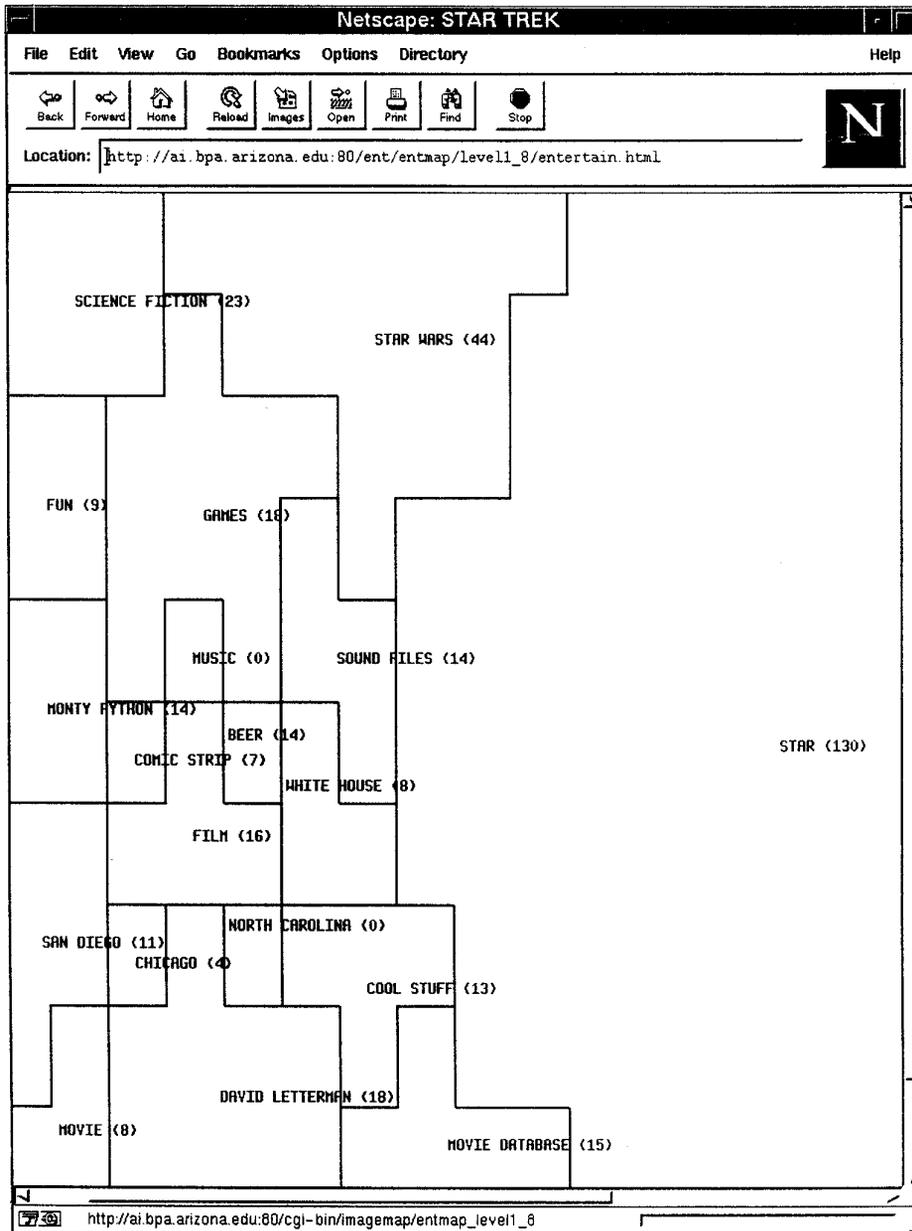


FIG. 3. Subregions for STAR TREK region.

present in the tool for a variety of its gadgets. On the other hand, most of the subjects were still thinking hierarchically and linearly when they went to search the ET-Map, but the map is not hierarchical or linear, and this caused problems. Only one subject could intuitively grasp the organization of the ET-Map from the very beginning. This subject is very artistically inclined and has worked as a cartographer. The subject decided that she was interested in film clips. While other subjects also searched for films, they all began in the areas labeled MOVIE or MOVIE DATABASE. Although this subject

started with MOVIE, when she was unable to locate that in which she was interested, she returned to the first level and choose to browse LA because "that's where the movie business is." Most of the subjects did not think in this manner and when a search for films failed in the FILM area of the map, they would give up the search.

Subjects' Lack of Knowledge (of English Words, of WWW Browsing, of Netscape, of the Names of the People for Whom They Were Searching, etc.) Created Problems.

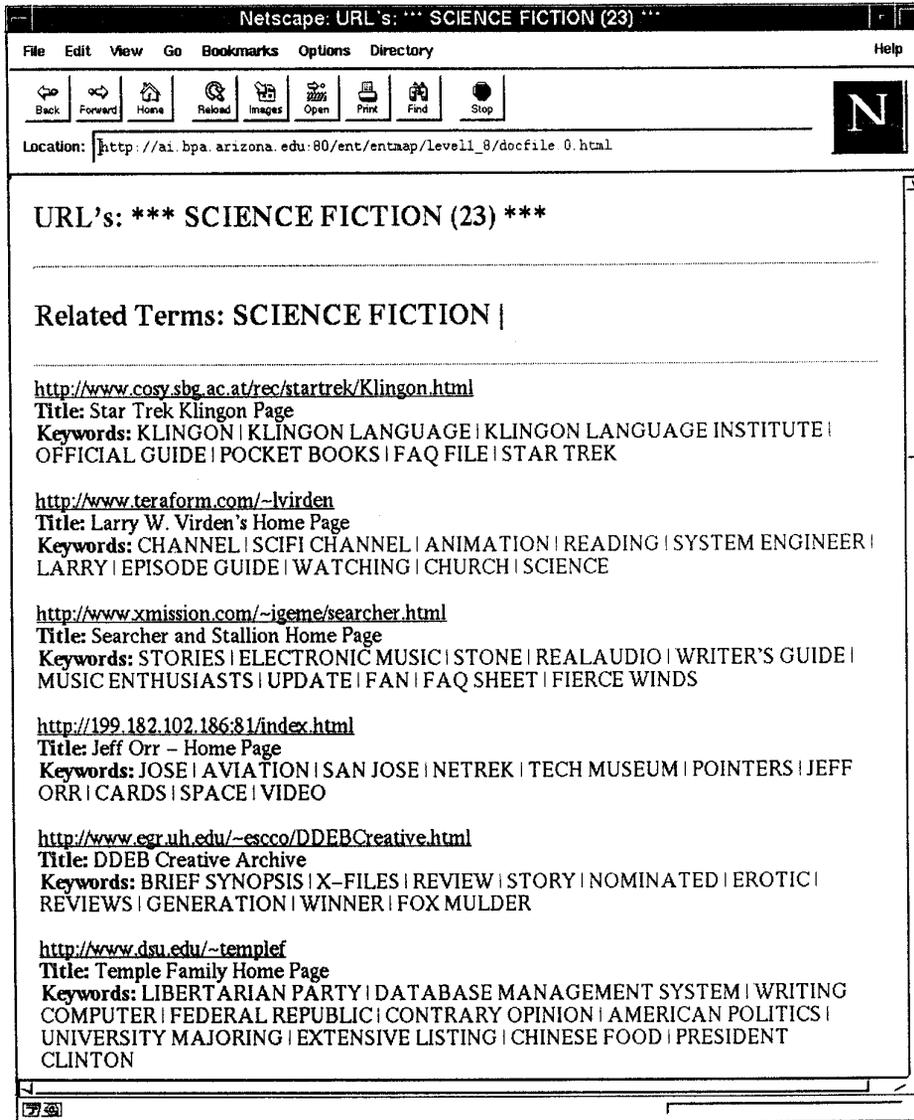


FIG. 4. URLs in SCIENCE FICTION subregion of the STAR TREK region.

Those subjects more familiar with Web browsing could speed through searches on Yahoo in less than a minute, whereas those with less experience in searching or Yahoo could take much longer. However, even those subjects experienced in searching found it difficult to isolate items in which they were interested on the ET-Map. Knowledge of the English language also played a part, especially in searching the Yahoo hierarchy, which is dependent upon the meanings and definitions of the subdivisions of its hierarchies. If you are in search of information on the latest James Bond movie, you must be aware that 007 movies are "genre" films, as that is how they are classified in the Yahoo hierarchy.

As well, several subjects were interested in finding information on the actors and actresses of their favorite TV

shows. However, the subjects did not know the names of the actors and actresses. Pages about actors and actresses are cataloged into the ACTORS AND ACTRESSES subdivision of the Yahoo hierarchy, so without their names, searching becomes very difficult. The best the subjects could hope for was to find a link to a page devoted to the TV show which would provide further links to information about the actors and actresses on the show.

Familiarity with the Web and searching also affected subjects' opinions about the size of the ET-Map. The size of the map did not bother those with more experience on the Web. However, subjects with less experience in searching found the size daunting and had difficulty navigating across the width and breadth of it.

Subjects Were Unsure of What Items Were Considered

“*Entertainment*” by *Either Tool*. Several subjects were interested in sports or cars, which are not part of the ET-Map or the Yahoo Entertainment hierarchy, although the subjects considered them “entertainment”. As well, a subject in search of circus information was unsure whether it had not been considered entertainment and therefore had been excluded from the ET-Map and the Yahoo hierarchy or whether the information simply did not exist on the Web.

5.2.2. *The ET-Map Strengths and Weaknesses*

Through analysis of the search patterns and verbal protocols of the subjects, it became apparent that the ET-Map has certain weaknesses and certain strengths.

Weaknesses of the ET-Map. The nature of the map is not obvious and often the placement of documents into areas of the maps seemed arbitrary and random to the subjects. An often reiterated comment among the subjects was that the areas of the map “need more of an explanation.” For example, several subjects became very frustrated when they chose WHITE HOUSE from the first level of the map, but were unable to find the interactive tour of the White House.

Although the listings of titles, URLs, and keywords presented in the map are ranked, subjects frequently requested an “alphabetical ordering” to this list. Subjects felt the ranking did little to assist their search and that alphabetical order would be better. Also, subjects commented on the length of the ultimate URL lists. “I don’t really have the patience to go through all of them,” was a statement repeated by several of the subjects.

Some large “entertainment” areas did not appear on the first level of the map. A TELEVISION area may be found in FILM from the first level. Many subjects were interested in television, but became discouraged when neither the term Television nor the term TV appeared as area labels on the first level of the ET-Map. The subjects made frequent comments about the organization of the map as “random” or the “mapping is sorta ... there’s no pattern in it.”

Strengths of the ET-Map. By the end of the experiment, at least three subjects were becoming more accustomed to searching the ET-Map and its associative regions. For example, one subject was interested in the television show, ER. When he did not find a heading for television on the first level of the map, he chose DAVID LETTERMAN because “he’s on TV.” This search was successful. A Netscape Find search on the word TV inside the DAVID LETTERMAN listing of titles returned numerous hits, and a search on NBC returned two very helpful hits (once again, though, a subject’s knowledge of the field assisted the search—the subject needed to know that ER is shown on

NBC). Following the links with NBC in the keywords led to the NBC Home Page and from there the subject was able to find ER. Another subject who really grasped the associative nature of the map was the former cartographer. Although the map was designed to reflect human, associative thought patterns, it may be that people are too well trained in hierarchical searches to intuitively grasp the associative map.

The listing of keywords was popular among the subjects. Although most felt the lists of titles were too long, they found the keywords to be of great assistance in describing the pages. Most wished that the map areas could be as well described as the documents within them. The keywords may have been so well liked because home page titles often are vague and the keywords can be used to determine the subject matter in a page where the title is of no help.

Subjects liked the “idea” of the map; the idea of being able to visually browse the Web (or the entertainment sections of the Web) at a glance. Most subjects enjoyed browsing the map for ideas. Through the map, subjects learned which topics existed on the Web. Most people started off their searches with a comment along the lines of “just gonna look and see what is on the map.” Most of the subjects did not enter the experiment with search items fixed in their minds. They would browse the map and see what looked interesting and then follow those interesting links further, i.e., they would “surf” the map. Those subjects who were simply browsing found the map much more helpful than those who tried to search for specific items using the map. The difference between serendipitous browsing and goal-driven searching has been well documented in prior hypertext studies [8].

It also appears that some of the problems associated with the map are due to the nature of the documents on the Web. A large percentage of the “entertainment” home pages are personal home pages, which are difficult to classify. On most personal home pages, the owners may discuss their Professional Experience (e.g., Ikos Systems, Ready Systems, Link Flight Simulation, Dymac, Simpact Associates), their Education (e.g., M.A. Applied Mathematics, UCSD), and their Personal Interests (e.g., astronomy, Bay area, comics, computer languages, family, fantasy and science fiction, games, history, maps, mathematics, sports). The only appropriate classification for these pages is probably as a personal home page region. If the personal home pages could be weeded out of the collection and placed into their own category, the map would be much cleaner and more clearly attain its promise.

CONCLUSION

This research aimed to address current and future Internet searching problems by developing and testing prom-

using neural network categorization techniques. Based on a general research framework for Internet categorization and searching, the first stage of our research involves an automatic, multilayered, self-organizing approach to categorizing Internet homepages based on their contents (terms). Results of this categorization process can then be used at the second stage to create category-specific concept spaces for assisting in concept-based, associative information retrieval.

The multilayered SOM (M-SOM) algorithm has been tested in several applications including electronic brainstorming comments and Internet entertainment-related homepages. The initial testing results were interesting. The techniques appeared to produce meaningful results for small-scale applications (e.g., EBS categorization) and potentially useful concept maps for serendipitous browsing for large-scale applications (e.g., Internet homepage categorization). However, more systematic refinement and user evaluation for large-scale Internet applications and parallelization for selected algorithms are needed (and are under way).

ACKNOWLEDGMENTS

This project is supported by a Research Initiation Award grant awarded by the Division of Information, Robotics, and Intelligent Systems, NSF ("Building a Concept Space for an Electronic Community System," PI: H. Chen, 1992–1994, IR19211418), a National Collaboratory grant awarded by NSF ("Systems Technology for Building a National Collaboratory", PI: B. Schatz, 1990–1994), 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, IR19411318), and an NSF/CISE grant ("Concept-based Categorization and Search on Internet: A Machine Learning, Parallel Computing Approach," PI: H. Chen, 1995–1995, IR19525790).

REFERENCES

1. R. K. Belew, Adaptive information retrieval, in *Proceedings of the Twelfth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Cambridge, MA, June 25–28, 1989*, pp. 11–20.
2. T. Berners-Lee, R. Cailliau, A. Luotonen, H. F. Nielsen, and A. Secret, The World-Wide Web, *Comm. ACM* **37**(8), 1994, 76–82.
3. D. C. Blair and M. E. Maron, An evaluation of retrieval effectiveness for a full-text document-retrieval system, *Comm. ACM* **28**(3), 1985, 289–299.
4. M. J. Blosseville, G. Hebrail, M. G. Monteil, and N. Penot, Automatic document classification: Natural language processing, statistical analysis, and expert system techniques used together, In *Proceedings of the Fifteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, June 21–24 1992*, pp. 51–57.
5. C. M. Bowman, The Harvest information discovery and access system, in *Proceedings of the Second International World Wide Web Conference '94, Chicago, October 17–20, 1994*.
6. C. M. Bowman, P. B. Danzig, U. Manber, and F. Schwartz, Scalable internet resource discovery: Research problems and approaches, *Comm. ACM* **37**(8), 1994, 98–107.
7. L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Tree*, Wadsworth, Monterey, CA, 1984.
8. E. Carmel, S. Crawford, and H. Chen, Browsing in hypertext: A cognitive study, *IEEE Trans. Systems Man Cybernet* **22**(5), 1992, 865–884.
9. H. Chen, Collaborative systems: Solving the vocabulary problem, in Special Issue on Computer—Supported Cooperative Work (CSCW), *IEEE Computer* **27**(5), 1994, 58–66.
10. H. Chen, Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms, *J. Am. Soc. Inform. Sci.* **46**(3), 1995, 194–216.
11. H. Chen and V. Dhar, User misconceptions of online information retrieval systems, *Internat. J. Man-Machine Stud.* **32**(6), 1990, 673–692.
12. H. Chen and J. Kim, GANNET: A machine learning approach to document retrieval, *J. Management Inform. Systems* **11**(3), 1994–1995, 7–41.
13. H. Chen and K. J. Lynch, Automatic construction of networks of concepts characterizing document databases, *IEEE Trans. Systems Man Cybernet.* **22**(5), 1992, 885–902 1992.
14. H. Chen, K. J. Lynch, K. Basu, and D. T. Ng, Generating, integrating, and activating thesauri for concept-based document retrieval, *Special Series on Artificial Intelligence in Text-based Information Systems IEEE Expert* **8**(2), 1993, 25–34.
15. H. Chen and D. T. Ng, An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): symbolic branch-and-bound vs. connectionist Hopfield net activation, *J. Am. Soc. Inform. Sci.* **46**(5), 1995, 348–369.
16. H. Chen, B. R. Schatz, T. Yim, and D. Fye, Automatic thesaurus generation for an electronic community system, *J. Am. Soc. Inform. Sci.* **46**(3), 1995, 175–193.
17. H. Chen and L. She, Inductive query by examples (IQBE): A machine learning approach, in *Proceedings of the 27th Annual Hawaii International Conference on System Sciences (HICSS-27), Information Sharing and Knowledge Discovery Track, Maui, HI, January 4–7, 1994*.
18. J. Courteau, Genome databases, *Science* **254**, October 11, 1991, 201–207.
19. S. L. Crawford, R. Fung, L. A. Appelbaum, and R. M. Tong, Classification trees for information retrieval, in *Proceedings of the 8th International Workshop on Machine Learning*, pp. 245–249, Morgan Kaufmann, 1991.
20. S. L. Crawford and R. M. Fung, An analysis of two probabilistic model induction techniques, *Statist. Comput.* **2**(2), 1992, 83–90.
21. P. DeBra and R. Post, Information retrieval in the World-Wide Web: Making client-based searching feasible, in *Proceedings of the First International World Wide Web Conference '94, Geneva, Switzerland, 1994*.
22. T. E. Doszko, J. Reggia, and X. Lin, Connectionist models and information retrieval, *Annu. Rev. Information Sci. Technol.* **25**, 1990, 209–260.
23. K. A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data*, MIT Press, Cambridge, MA, 1993.
24. O. Etzioni and D. Weld, A softbot-based interface to the Internet, *Comm. ACM* **37**(7), 1994, 72–79.
25. C. L. Foss, Tools for reading and browsing hypertext, *Inform. Process. Management* **25**(4), 1989, 407–418.

26. K. A. Frenkel, The human genome project and informatics, *Comm. ACM* **34**(11), 1991, 41–51.
27. N. Fuhr, S. Hartmann, G. Knorz, G. Lustig, M. Schwantner, and K. Tzeras, AIR/X—A rule-based multistage indexing system for large subject fields, in *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90), Boston, MA, July 29–August 3, 1990*, pp. 789–795.
28. R. Fung and S. L. Crawford, Constructor: a system for the induction of probabilistic models, in *Proceedings of the 8th National Conference on Artificial Intelligence (AAAI-90), Boston, MA, July 29–August 3, 1990*, pp. 762–769.
29. G. W. Furnas, 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, Gaithersburg, MD, March 1982*, pp. 251–253.
30. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, The vocabulary problem in human-system communication, *Comm. ACM* **30**(11), November 1987, 964–971.
31. M. Gordon, Probabilistic and genetic algorithms for document retrieval, *Comm. ACM* **31**(10), 1988, 1208–1218.
32. M. D. Gordon, User-based document clustering by redescribing subject descriptions with a genetic algorithm, *J. Am. Soc. Inform. Sci.* **42**(5), 1991, 311–322.
33. K. Knight, Connectionist ideas and algorithms, *Comm. ACM* **33**(11), 1990, 59–74.
34. T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed. Springer-Verlag, Berlin/Heidelberg, 1989.
35. M. Koster, ALIWEB: Archie-like indexing in the web, in *Proceedings of the First International World Wide Web Conference '94, Geneva, Switzerland, 1994*.
36. K. L. Kwok, A neural network for probabilistic information retrieval, in *Proceedings of the Twelfth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Cambridge, MA, June 25–28, 1989*, pp. 21–30.
37. D. D. Lewis, Learning in intelligent information retrieval, in *Proceedings of the 8th International Workshop on Machine Learning*, pp. 235–239, Morgan Kaufmann, San Mateo, CA, 1991.
38. D. D. Lewis, An evaluation of phrasal and clustered representations on a text categorization task, in *Proceedings of the Fifteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, June 21–24 1992*, pp. 37–50.
39. X. Lin, D. Soergel, and G. Marchionini, A self-organizing semantic map for information retrieval, in *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Chicago, October 13–16, 1991*, pp. 262–269.
40. R. P. Lippmann, An introduction to computing with neural networks, *IEEE Acoust. Speech Signal Process.* **4**(2), 1987, 4–22.
41. K. J. MacLeod and W. Robertson, A neural algorithm for document clustering, *Inform. Process. Management* **27**(4), 1991, 337–346.
42. P. Maes, Agents that reduce work and information overload, *Comm. ACM* **37**(7), 1994, 30–40.
43. M. E. Maron and J. L. Kuhns, On relevance, probabilistic indexing and information retrieval, *J. Assoc. Comput. Mach.* **7**(3), 1960, 216–243.
44. B. Masand, L. Gordon, and D. Waltz, Classifying news stories using memory-based reasoning, in *Proceedings of the Fifteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 59–65, Copenhagen, June 21–24 1992.
45. Mauldin and Leavitt, Web-agent related research at the CMT, in *Proceedings of the ACM Special Interest Group on Networked Information Discovery and Retrieval (SIGNIDR-94), August 1994*.
46. O. McBryan, GENVL and WWW: Tools for taming the web, in *Proceedings of the First International World Wide Web Conference '94, Geneva, Switzerland, 1994*.
47. K. Obraczka, P. B. Danzig, and S. Li, Internet resource discovery services, *IEEE Computer* **26**(9), 1993, 8–24.
48. R. Orwig, H. Chen, and J. F. Nunamaker, A graphical, self-organizing approach to classifying electronic meeting output, *J. Am. Soc. Inform. Sci.* in press.
49. F. Petry, B. Buckles, D. Prabhu, and D. Kraft, Fuzzy information retrieval using genetic algorithms and relevance feedback, in *Proceedings of the ASIS Annual Meeting, 1993*, pp. 122–125.
50. V. V. Raghavan and B. Agarwal, Optimal determination of user-oriented clusters: An application for the reproductive plan, in *Proceedings of the Second International Conference on Genetic Algorithms and Their Applications, Cambridge, MA, July 1987*, pp. 241–246.
51. E. Rasmussen, Clustering algorithms, in *Information Retrieval: Data Structures and Algorithms* (W. B. Frakes and R. Baeza-Yates, Eds.), Prentice-Hall, Engelwood Cliffs, NJ, 1992.
52. L. F. Rau and P. S. Jacobs, Creating segmented databases from free text for text retrieval, in *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Chicago, October 13–16, 1991*, pp. 337–346.
53. G. Salton, Generation and search of clustered files, *ACM Trans. Database Systems* **3**(4), 1978, 321–346.
54. G. Salton, *Automatic Text Processing*, Addison-Wesley, Reading, MA, 1989.
55. B. R. Schatz and J. B. Hardin, NSCA Mosaic and the World Wide Web: Global hypermedia protocols for the Internet, *Science* **265**, 12 August 1994, 895–901.
56. P. K. Simpson, *Artificial Neural Systems: Foundations, Paradigms, Applications, and Implementations*, McGraw-Hill, New York, 1990.
57. P. K. Simpson, Fuzzy min-max neural networks. 2. Clustering, *IEEE Trans. Fuzzy Systems* **1**(1), 1993, 32–45.
58. K. Sparck Jones, *Automatic Keyword Classification for Information Retrieval*, Butterworths, London, 1971.
59. R. Wilkinson and P. Hingston, Using the cosine measure in neural networks for document retrieval, in *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Chicago, October 13–16, 1991*, pp. 202–210.
60. R. Wilkinson, P. Hingston, and T. Osborn, Incorporating the vector space model in a neural network used for document retrieval, *Library Hi Tech* **10**(12), 1992, 69–75.
61. J. Yang and R. R. Korfhage, Effects of query term weights modification in document retrieval: A study based on a genetic algorithm, in *Proceedings of the Second Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, April 26–28, 1993*, pp. 271–285.
62. J. Yang, R. R. Korfhage, and E. Rasmussen, Query improvement in information retrieval using genetic algorithms: A report on the experiments of the TREC project, in *Text Retrieval Conference (TREC-1), Gaithersburg, MD, November 4–6 1993*, pp. 31–58.



HSINCHUN CHEN received the Ph.D. degree in information systems from New York University in 1989. He is an associate professor of management information systems at the University of Arizona and director of the Artificial Intelligence Group. He received an NSF Research Initiation Award in 1992 and the Hawaii International Conference on System Sciences (HICSS) Best Paper Award and an AT&T Foundation Award in Science and Engineering in 1994. He was recently awarded a major Digital Library Initiative grant by NSF/NASA/ARPA for a joint project with the University of Illinois, 1994–1998, and an “intelligent” Internet categorization and search project from NSF/CISE, 1995–1998. Dr. Chen has published more than 30 articles in publications such as *Communications of the ACM*, *IEEE Computer*, *Journal of the American Society for Information Science*, *IEEE Transactions on Systems, Man, and Cybernetics*, *IEEE Expert*, and *Advances in Computers*.



CHRISTOPHER SCHUFFELS received the B.S. degree in management information systems from the University of Arizona in 1995. He is currently a software engineer at the Ventana Corporation in Tucson, Arizona. He is also a research scientist in the University of Arizona’s Artificial Intelligence Group. His interests are in group systems, neural network algorithms, and software engineering.



RICHARD ORWIG is a research scientist with the Center for the Management of Information in the Management Information Systems Department at the University of Arizona. He received his Ph.D. from the University of Arizona in 1995. He has published in *Communications of the ACM*, *Journal of the Management Information Systems*, and *Group Decision and Negotiation*.