High-performance Digital Library Classification Systems:
From Information Retrieval to Knowledge Management

Hsinchun Chen, Ph.D.
Professor, MIS Department/Director, Artificial Intelligence Lab
McClelland Hall 430Z, University of Arizona
Tucson, AZ 85721, hchen@bpa.arizona.edu, (520) 621-4153
Robin R. Sewell, Doctor of Veterinary Medicine/MA in Library Science
Research Scientist, Artificial Intelligence Lab, University of Arizona
Tucson, AZ 85721, rsewell@bpa.arizona.edu, (520) 621-6219

Contents

High-performance Digital Library Classification Systems: From Information Retrieval to Knowledge Management

Principal Investigator (PI): Hsinchun Chen, Ph.D., Professor, Management Information Systems Department, Director, Artificial Intelligence Lab, McClelland Hall 430Z, University of Arizona, Tucson, AZ 85721, (520) 621-4153, hchen@bpa.arizona.edu. **Co-PI:** Robin R. Sewell, Doctor of Veterinary Medicine/MA in Library Science, Research Scientist, Artificial Intelligencce Lab, University of Arizona, Tucson, AZ 85721, (520) 621-6219, rsewell@bpa.arizona.edu.

Partnerships/Supports: (1) Computing: SiliconGraphics Computer Systems, National Center for Supercomputing Applications; **(2) Collections:** National Library of Medicine, National Cancer Institute, GeoRef Information Services, Petroleum Abstracts; **(3) User Evaluation:** Arizona Cancer Center, Arizona Health Sciences Library, UA Main Library and Science and Engineering Library

Introduction: In this era of the Internet and distributed, multimedia computing, new and emerging digital library applications have swept into the lives of office workers and everyday people. As the digital library applications become more overwhelming, pressing, and diverse, several well-known information retrieval (IR) problems have become even more urgent. Conventional approaches to addressing information overload and interoperability problems are manual in nature, requiring human experts as information intermediaries to create knowledge structures and/or classification systems (e.g., the National Library of Medicine's Unified Medical Language System, UMLS) to bridge the gap resulting from vocabulary differences. As information content and collections become even larger and more dynamic (thus rendering manual knowledge structures more difficult to create), we believe a system-aided, algorithmic, bottom-up approach to creating large-scale digital library classification systems is needed.

Research Questions: (1) Can various clustering algorithms produce classification results comparable to those of classification systems generated by human beings? Which algorithm produces the best result and under what condition? (2) Are clustering algorithms to create classification systems based on large-scale domain-specific digital library collections computationally feasible? What optimization and parallelization techniques are needed to achieve the required scalability?

Research Plan: The proposed research aims to develop an architecture and the associated techniques needed to automatically generate classification systems from large domain-specific textual collections and to unify them with manually created classification systems to assist in effective digital library retrieval and analysis. Both algorithmic developments and user evaluation in several sample domains will be conducted. Scalable automatic clustering methods including Ward's clustering, multi-dimensional scaling, latent semantic indexing, and the self-organizing map will be developed and compared. Most of these algorithms, which are computationally intensive, will be optimized based on the sparseness of common keywords in textual document representations. Using parallel, high-performance platforms as a time machine for simulation, we plan to parallellize and benchmark these clustering algorithms for large-scale collections (on the order of millions of documents) in several domains. Results of automatic classification systems will be represented using several novel hierarchical display methods.

The testbed of research will include three application domains that incorporate both large-scale collections and existing classification systems: (1) medicine: CancerLit (700,000 cancer abstracts) and the NLM's UMLS (500,000 medical concepts), (2) geoscience: GeoRef and Petroleum Abstracts (800,000 abstracts) and Georef thesaurus (26,000 geoscience terms), and (3) Web application: a WWW collection (1.5M web pages) and the Yahoo! classification (20,000 categories). Medical professionals, geo scientists, and WWW search engine users will be used in our evaluation plan.

Program Announcement: DIGITAL LIBRARY - NSF 98-63

Introduction

In this era of the Internet and distributed, multimedia computing, new and emerging classes of information systems applications have swept into the lives of office workers and everyday people. New applications ranging from digital libraries, multimedia systems, geographic information systems, and collaborative computing to electronic commerce, virtual reality, and electronic video arts and games have created tremendous opportunities for information and computer science researchers and practitioners.

As applications become more overwhelming, pressing, and diverse, several well-known information retrieval (IR) problems have become even more urgent. *Information overload* resulting from easy creation and transmittal of information via Internet and WWW, has become prominent in people's lives (e.g., even stockbrokers and elementary school students, heavily exposed to various WWW search engines, are versed in such IR terminology as recall and precision). Significant variations of database formats and structures, the richness of information media (text, audio, and video), and an abundance of multilingual information content also have created severe *interoperability* problems.

The conventional approaches to addressing information overload and interoperability problems are manual in nature, requiring human experts as information intermediaries to create knowledge structures and/or classification systems (e.g., the National Library of Medicine's Unified Medical Language System project, UMLS) [30]. Such manually-created classification systems, which represent subject vocabularies and their relationships, are often used to index and/or organize collections or to suggest search terms during retrieval processes. But as information content and collections become even larger and more dynamic, thus rendering manual knowledge creation to become more difficult, if not infeasible, we believe a complementary system-aided, algorithmic, bottom-up approach to creating large-scale digital library classification systems is needed.

Digital Libraries, Semantic Interoperability, and Knowledge Networking

The Information Infrastructure Technology and Applications (IITA) Working Group, the highest level of the country's National Information Infrastructure (NII) technical committee, held an invited workshop in May 1995 to define a research agenda for digital libraries.

The shared vision that emerged is an entire Net of distributed repositories, where objects of any type can be searched within and across different indexed collections [45]. In the short term, technologies must be developed to search across these repositories transparently, handling any variations in protocols and formats (i.e., addressing structural interoperability [35]). In the long term, technologies must be developed to handle the variations in content and meanings (knowledge) transparently as well. Meeting these requirements constitutes steps along the way toward matching the concepts requested by users with objects indexed in collections [44].

The ultimate goal, as described in the IITA report, is the Grand Challenge of Digital Libraries:

> deep semantic interoperability - the ability of a user to access, consistently and coherently, similar (though autonomously defined and managed) classes of digital objects and services, distributed across heterogeneous repositories, with federating or mediating software compensating for site-by-site variations...Achieving this will require breakthroughs in description as well as retrieval, object interchange and object retrieval protocols. Issues here include the definition and use of metadata and its capture or computation from objects (both textual and multimedia), the use of computed descriptions of objects, federation and integration of heterogeneous repositories with disparate semantics, clustering and automatic hierarchical organization of information, and algorithms for automatic rating, ranking, and evaluation of information quality, genre, and other properties.

A focus on to semantic interoperability has prompted several of the NSF/DARPA/NASA funded large-scale digital library initiative (DLI) projects to explore various statistical, and pattern recognition techniques, e.g., concept spaces and category maps in the Illinois project [46] [5], textile and word sense dis-ambiguiation in the Berkeley project [53], voice recognition in the CMU project [50], and image segmentation and clustering in the UCSB project [29]. ``Definition and use of metadata'' and ``clustering and automatic hierarchical organization of information,'' which require significant future research, are the key components needed to build classification systems for digital libraries automatically.

In the Santa Fe Workshop on Distributed Knowledge Work Environments: Digital Libraries held in March, 1997, the panel of digital library researchers and practitioners suggested three areas of research for the planned Digital Library Initiative-2 (DLI-2): system-centered issues, collection-centered issues, and user-centered issues. *Scalability, interoperability, adaptability and durability*, and *support for collaboration* are the four key research directions relevant to system-centered issues. System interoperability, syntactic (structural) interoperability, linguistic interoperability, temporal interoperability, and semantic interoperability are recognized by researchers as the most challenging and rewarding research areas.

Knowledge Management in Digital Libraries

In a new NSF Knowledge Networking (KN) initiative, a group of domain scientists and information systems researchers were invited to a Workshop on Distributed Heterogeneous Knowledge Networks at Boulder, Colorado, in May, 1997. They considered that scalable techniques to improve semantic bandwidth and knowledge bandwidth are among the priority research areas, as described in the KN report:

> The Knowledge Networking (KN) initiative focuses on the integration of knowledge from different sources and domains across space and time... KN research aims to move beyond connectivity to achieve new levels of interactivity, increasing the semantic bandwidth, knowledge bandwidth, activity bandwidth, and cultural bandwidth among people, organizations, and communities.

``Knowledge networking'' or, using a more general term, ``knowledge management'' (KM), has attracted significant attention from academic researchers and even executives in Fortune 500 companies [52] [51]. Daniel O'Leary provides the following definition for KM [33]:

> Enterprise knowledge management entails formally managing knowledge resources in order to facilitate access and reuse of knowledge, typically by using advanced technology. KM is formal in that knowledge is classified and categorized according to a pre-specified - but evolving - ontology into structured and semi-structured data and knowledge bases.

Many commercial software systems are considered enabling technologies for KM, including Internet/Intranet search engines, groupware, electronic document management systems, full-text retrieval systems, database management systems, and electronic mail [52]. Major consulting firms, on the other hand, are promoting KM practices such as organizational development, strategic planning, performances metrics, and methodology. Despite of obvious differences in technologists and consultants views of what KM should be, their calls for advanced classification and categorization technologies to help analyze, organize, and present mission-critical information and knowledge are the same - turning information overload into knowledge assets.

In ``Practical Digital Libraries'' [26], Lesk describes knowledge representation methods by which people have tried to organize and arrange knowledge, with the idea of making searching simple. As Lesk commented:

> In practice, it seems unlikely that any single knowledge representation scheme will serve all purposes. The more detailed such a scheme is, the less likely it is that two different people will come up with the same place in it for the same document. And the less detailed it is, the less resolving power it has and the less use it is.

Trained librarians, who are versed in classification scheme and domain knowledge have crated library classification systems and subject-specific thesauri such as the Library of Congress classification, Dewey classification, or the NLM's Unified Medical Language Systems (UMLS), which are significant human efforts to label knowledge consistently [17] [3]. Library classification systems and thesauri often capture nouns or noun phrases and represent only limited relationships (e.g., broader terms, narrower term, etc.). The representations are often coarse, but precise, and they do support their practical goal of supporting indexing and searching, but significant human efforts are needed to create and maintain such large-scale classification systems. (In this research we will examine mostly hierarchical classification systems. If a manual classification system is unavailable, we will resort to a subject thesaurus with hierarchical relationships.)

Artificial intelligence representations such as semantic networks, expert systems, or ontologies represent another approach to capturing knowledge, e.g., Lenat's CYC common sense knowledge base [24] [25] [13]. Their representations are often richer and more fine-grained and the goal of capturing human intelligence is ambitious and difficult. Due to the granularity required of such representations, knowledge creation is slow and painstaking. Only experimental prototypes in small, limited domains have been created. Their usefulness in large-scale digital library applications remains suspect.

A Scalable Bottom-Up Approach to Supporting Knowledge Management in Digital Libraries

The traditional approach to creating classification systems and knowledge sources in library science and classical AI is often considered top-down since knowledge representations and formats are pre-defined by human experts or trained librarians and the process of generating knowledge is structured and well-defined. A complementary bottom-up approach to knowledge creation has been suggested by researchers in machine learning, statistical analysis, and neural networks.

Based on actual databases or collections, researchers develop programs which systematically segment and index documents in various databases (text, image, and video) and identify patterns within such databases. Analyzing databases which contain structured and numeric data (e.g., credit card usage, a frequent flyer program) is often referred to as data mining or knowledge discovery [37] [28]. Generating knowledge algorithmically from multimedia databases (especially text, e.g., customer complaint email, machinery repair reports, brainstorming outputs) is considered the core of knowledge management [33].

Two stages of such a bottom-up knowledge management approach are often required:

- **Object Recognition, Segmentation, and Indexing:**

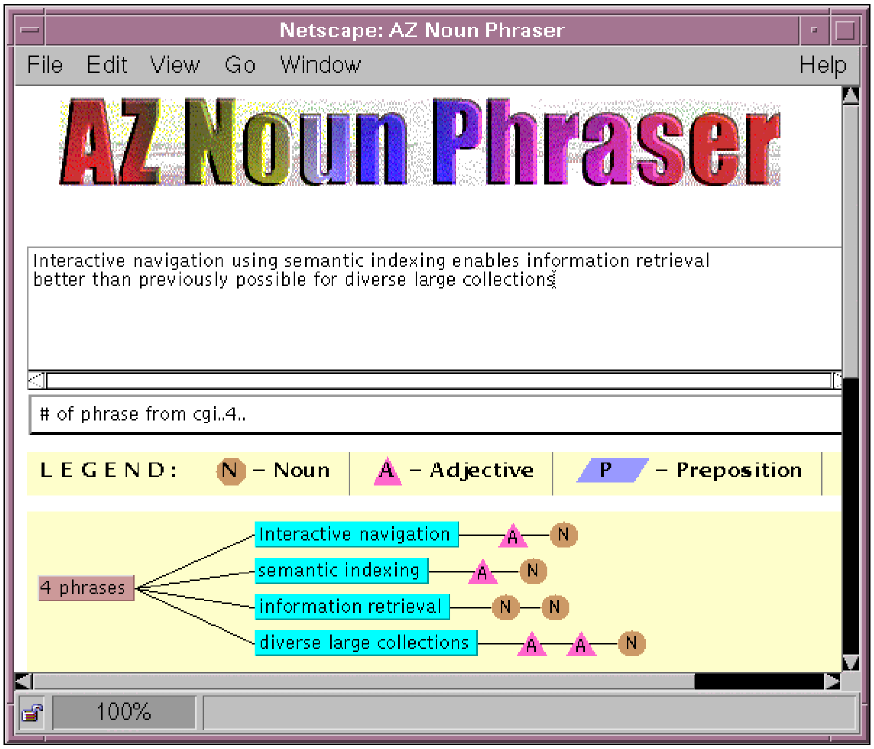
  The most fundamental techniques in IR involve identifying key features in objects. For example, automatic indexing and natural language processing (e.g., noun phrase extraction or object type tagging) are frequently used to extract meaningful keywords or phrases from texts automatically [43] [5]. Texture, color, or shape-based indexing and segmentation techniques are often used to identify images [29]. For audio and video applications, voice recognition, speech recognition, and scene segmentation techniques can be used to identify meaningful descriptors in audio or video streams [50]. Many of these techniques have been developed previously in the six large-scale DLI-1 projects [45]. (The numeric data mining approach does not require such a process since most data are structured and well organized in the first place.)

- **Analysis and Classification:**

  Several classes of techniques have been used for semantic analysis of texts or multimedia objects. Symbolic machine learning (e.g., ID3, version space), graph-based clustering and classification (e.g., Ward's hierarchical clustering), statistics-based multivariate analyses (e.g., latent semantic indexing, multi-dimensional scaling, regressions), artificial neural network-based computing (e.g., backpropagation networks, Kohonen self-organizing maps), and evolution-based programming (e.g., genetic algorithms) are among the popular techniques [2]. Most of these are computationally intensive and are particularly suitable for creating classification systems or knowledge structures from unstructured textual collections. Large-scale image (and video) based knowledge extraction is still at an early stage of development due to the difficulty of designing image-invariant content-based segmentation and indexing algorithms [7] [29].

An example of adoption of noun phrasing and use of a neural network based clustering algorithm in analyzing project reports (project summaries submitted to the DARPA/ITO program) is shown below. The techniques were developed by The University of Arizona Artificial Intelligence Lab (headed by Dr. Hsinchun Chen) for the Illinois DLI project. For detailed technical discussions, readers are referred to [5] [4].

- **Noun Phrase Indexing:** Noun phrase indexing aims to identify concepts (grammatically correct noun phrases) from a collection for term indexing. It begins with a text tokenization process to separate punctuation and symbols, followed by part-of-speech-tagging (POST) using variations of the Brill tagger and 30-plus grammatic noun phrasing rules. Figure 1 shows an example of tagged noun phrases for a simple sentence. (The system is referred to as AZ Noun Phraser.) For example, ``interactive navigation'' is a noun phrase that consists of an adjective (A) and a noun (N).

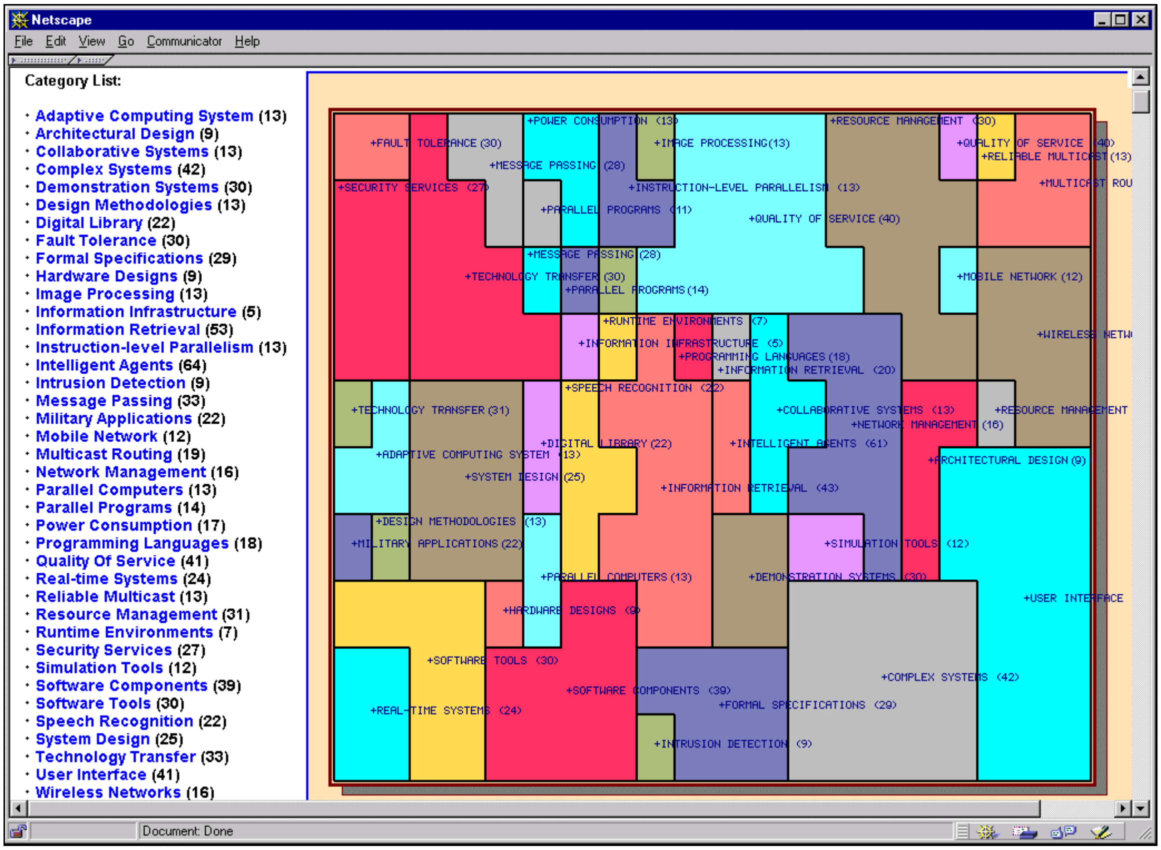


Interactive navigation using semantic indexing enables information retrieval better than previously possible for diverse large collections

# of phrase from cgi..4..

Figure 1 - Tagged Noun Phrases

- **Automatic SOM Classification:** A category map is the result of performing a neural network-based clustering (self-organizing map, SOM) of similar documents and automatic category labeling. Documents that are similar (in noun phrase terms) to each other are grouped together in a neighborhood on a two-dimensional display. As shown in the colored jigsaw-puzzle display in Figure 2, each colored region represents a unique topic that contains similar documents. Topics that are more important often occupy larger regions. By clicking on each region, a searcher can browse documents grouped in that region. An alphabetical list that is a summary of the 2D result is also displayed on the left-hand-side of Figure 2, e.g., Adaptive Computing System (13 documents), Architectural Design (9 documents), etc. The SOM algorithm can also create multi-layered category maps, resulting a hierarchical structure [6].



Category List:

- Adaptive Computing System (13)
- Architectural Design (9)
- Collaborative Systems (13)
- Complex Systems (42)
- Demonstration Systems (30)
- Design Methodologies (13)
- Digital Library (22)
- Fault Tolerance (30)
- Formal Specifications (29)
- Hardware Designs (9)
- Image Processing (13)
- Information Infrastructure (5)
- Information Retrieval (53)
- Instruction-level Parallelism (13)
- Intelligent Agents (64)
- Intrusion Detection (9)
- Message Passing (33)
- Military Applications (22)
- Mobile Network (12)
- Multicast Routing (19)
- Network Management (16)
- Parallel Computers (13)
- Parallel Programs (14)
- Power Consumption (17)
- Programming Languages (18)
- Quality Of Service (41)
- Real-time Systems (24)
- Reliable Multicast (13)
- Resource Management (31)
- Runtime Environments (7)
- Security Services (27)
- Simulation Tools (12)
- Software Components (39)
- Software Tools (30)
- Speech Recognition (22)
- System Design (25)
- Technology Transfer (33)
- User Interface (41)
- Wireless Networks (16)

Figure 2 - Category Map

Research Questions

General-purpose clustering algorithms have been in existence since the 1960s. Hierarchical and non-hierarchical clustering algorithms have been developed mostly for numeric analysis purposes [19]. Some of these general-purpose algorithms have then been adopted in information retrieval applications. (Rasmussen provides a good review of clustering algorithms in information retrieval in [39].) Several factor analysis based techniques such as latent semantic indexing (LSI) [9] and multi-dimensional scaling (MDS) [23] have also been adopted in textual analysis. More recently, the SOM-based textual classification systems have been reported by Chen et al. in several large-scale digital library applications [4] [34]. As evident in the comments made in the digital library and knowledge networking initiatives, clustering and classification techniques for real-life, large-scale collections are critically needed for developing knowledge structures for the next-generation digital libraries.

Several challenging problems have been reported to be associated with these clustering techniques:

- **Evaluating Clusters:**

  No consistent methodology for evaluating clusters has been adopted. Some research has compared clustering results with human clusters. Other experiment have report only cluster topology and simulation results. There is surprisingly little literature on evaluating partitions and other hierarchies, especially for textual documents.

Most statistics-based clustering methods were developed in the 1960s or 1970s. Neural network based clustering (unsupervised learning), on the other hand, has largely been developed in the 1980s and 1990s (although having historical roots that could be traced back to selected statistical algorithms developed in then 1960s). We have not seen a consistent and systematic evaluation of results generated by these diverse algorithms. (Most comparisons are of the statistics-based algorithms.) In light of significant interest in adopting clustering algorithms in digital libraries and knowledge management, there is a pressing need to develop a methodology for comparing these techniques systematically.

- **Optimization and Parallelization:**

  Most of the more robust clustering algorithms (e.g., Voorhees method, Ward's method) are computationally intensive, are often are $O(N^2)$ or $O(N^3)$ in complexity, where N is the number of objects to be clustered [39]. This is part of the reason that clustering algorithms rarely have been used in large-scale, real-life applications. Using thousands of objects is often the upper bound in clustering experiments.

  In our recent experiment of adopting the ``keyword sparseness'' characteristic (each document contains only a few non-zero indices) in textual classification, we were able to turn an $O(N^2)$ neural network algorithm into an $O(N)$ algorithm [42], which was computationally efficient in generating clusters for hundreds of thousands of objects (web pages) [6]. In addition, hardware advancement in recent years has made large-scale, parallel analysis a promising approach. Deep Blue's brute force computing approach to chess playing is a good example. Large-scale commercial data mining applications using shared memory multi-processors such as the SGI Origin2000 or IBM's SP2 are also becoming common practice in Fortune 500 companies [47]. In [5], we also report our experiment in using parallel supercomputers to analyze millions of engineering abstracts and automatically generate engineering concept spaces (thesauri). Algorithms developed in the 1960s and 1970s may be ready for prime time simply due to the hardware advances.

In light of these challenges, our project aims to address the following research questions:

1. **Can various clustering algorithms produce classification results comparable to classification systems generated by human beings? Which algorithm produces the best result and under what condition?**

2. **Are these clustering algorithms computationally feasible to create classification systems based on large-scale (millions) domain-specific digital library collections? What optimization and parallelization techniques are needed to achieve such scalability?**

The following 3 sections consist of our technical plan, testbed plan, and user evaluation plan, respectively, for the proposed research.

Technical Plan: Automatic Classification and High-Performance Computing

After five decades of active research in syntactic and semantic analysis for machine translation, linguistic analysis, speech recognition, and natural language processing, it has become clear that syntactic, linguistic analysis of domain-independent texts is computationally feasible [14] [36]. However, for detailed semantic analysis (i.e., understanding who did what to whom), most techniques still rely on domain-dependent lexicons or heuristic parsing rules and are not scalable across different subject areas or applications [18].

In the DARPA-funded decade-long multi-million dollar TIPSTER projects [48], less ambitious, yet scalable techniques have been developed. Most of these techniques rely on noun phrase extraction (e.g., proper names, place names, company names) and mathematical analysis of a large corpus. In the Message Understanding Workshop (MUC) and Multilingual Entity Tasks (MET), also under TIPSTER, lexicons, part-of-speech-taggers (POSTs), and linguistic parsing rules were adopted to extract key phrases in unstructured text - an approach adopted by successful systems developed at SRA, SRI, and BBN [48]. The state of the art in scalable, domain-independent text parsing seems to point conclusively to use of NLP phrase extraction.

Several well-known POSTs that previously had been adopted in MUC and MET were identified. Among them, Eric Brill's Brill Part of Speech Tagger, the MIT Media Lab Machine Understanding Group's Chopper, and LingSoft's NPTool are considered the most promising. The Brill tagger is rule-based and trainable, relying on transformation-based error-driven learning to improve its accuracy. Its accuracy has been reported to be as high as 97.2% when used on the Penn Treebank Wall Street Journal Corpus. Chopper is a natural language analysis engine (based on the Princeton's WordNet [31]) developed by the Machine Understanding Group at the MIT Media Laboratory under the direction of Dr. Ken Haase. NPTool is a commercially available noun phrase detector originally developed by Dr. Atro Voutilainen at the Department of General Linguistics at Helsinki University. LingSoft, a Finnish company currently distributes NPTool, which relies on hand-coded linguistic instead of statistical/stochastic methods of prediction rules to determine parts of speech [49].

Based on the Brill's tagger and the LingSoft's noun phrasing rule, we have developed the Arizona Noun Phraser (ANP) in the context of the Illinois DLI project [16]. Our experiments have shown that ANP is comparable to LingSoft in phrase recall and precision, and it significantly better than Chopper. The ANP will be adopted in our proposed research for text indexing.

A. Automatic Classification: Clustering Algorithms

After representing each textual document as a feature vector of noun phrases (using vector space rpresentation), it is possible to ``learn'' from the content of the entire collection. Conventional categorization techniques, including hierarchical clustering and multivariate statistical analysis such as multi-dimensional scaling and latent semantic indexing, and neural network computing techniques such as the self-organizing map are prime candidates for analyzing large-scale textual collections. During the course of this project, we will also continue to explore other new and promising clustering techniques.

- **Hierarchical Clustering: Ward's Algorithm**

  In a review of conventional clustering algorithms for textual analysis [39], Rasmussen suggested the multi-link based Ward's algorithm as the defacto standard for hierarchical clustering. Ward's clustering was proposed by statistician J. Ward in 1963. The algorithm uses a reciprocal nearest neighbor approach to identifying closest neighbors, an incremental, non-backtracking process reported by Murtagh in 1984 [32]. A hierarchy is created with each branch connecting two different clusters. Despite its robustness, the algorithm is serial and computationally expensive (an $O(N^2)$ algorithm, where N is the the total number of objects to be clustered). There is little comparison of hierarchical clustering methods with other newer techniques.

- **Metric Similarity Modeling: MDS and LSI**

An attempt to perform more extensive ``semantic analysis'' has been made by Metric Similarity Modeling (MSM). MSM uses a multi-dimensional semantic space where vectors are determined for documents. The nature of the multidimensional aspect of the semantic space no longer requires use of common keywords as the only measurement for document similarity (the basis of similarity computation in Ward's algorithm). ``MSM allows semantic associations to directly determine the interpretation of terms and the representations of text in the multidimensional space'' [1].

Multi-dimensional scaling (MDS) is a popular technique within MSM in which ``objects are represented as points in a multi-dimensional space; points are chosen so that the inter-point similarities meet a set of externally imposed constraints on the similarities'' [1]. MDS for textual analysis allows documents to be placed in spatial proximity limited by their similarity constraints. In addition to MDS, Latent Semantic Indexing (LSI) is an optimal method of MSM in which statistical techniques are used to uncover the underlying latent semantic structure in the data [9]. Despite being rooted in statistical analysis developed in 1960s and 1970s, MDS and LSI have become promising candidates for computational analysis, due in part to significant improvements in computing power [9] [1]. Both techniques were theoretically sound, but had previously been computationally expensive for large-scale analysis of collections.

- **Neural Networking Clustering: SOM**

  In 1980s and 1990s, a new class of categorization techniques based on artificial neural networks was developed. Among them, the Kohonen's self-organizing feature map (SOM) algorithm was successfully adopted in various engineering and scientific applications which involve numeric data (e.g., image recognition, signal processing) [21] and more recently for large-scale textual analysis [6] [20].

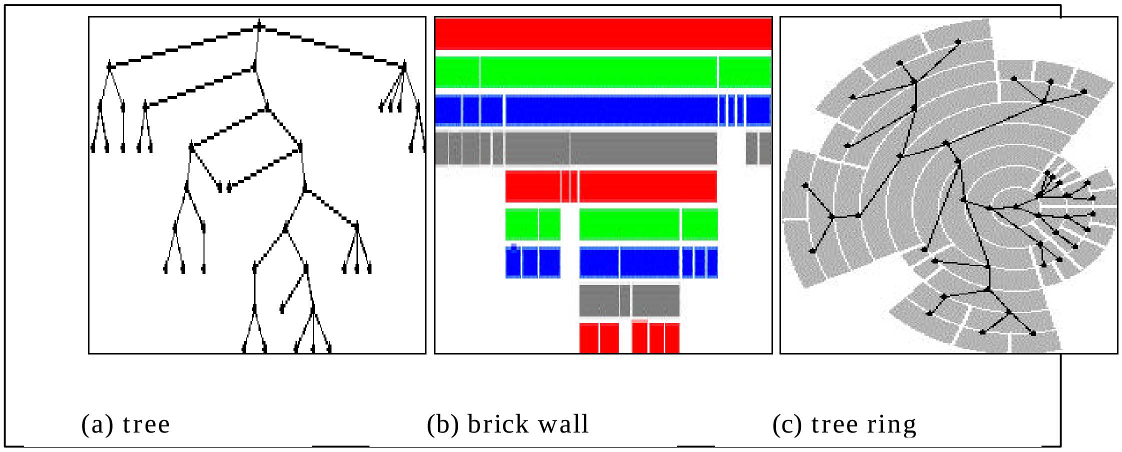
  In the algorithm's basic form, continuous-valued vectors are presented sequentially without specifying the desired output. After enough input vectors have been presented, network connection weights will specify cluster or vector centers that sample the input space such that the point density function of the vector centers tends to approximate the probability density function of the input vectors. In addition, the connection weights are organized such that topologically close nodes are sensitive to inputs that are physically similar.

  Lin [27] was the first to adopt the Kohonen SOM for textual analysis. In his prototype, self-organizing clusters of important concepts in a small database of several hundred documents were generated. A scalable multi-layered, graphical SOM approach to Internet categorization was developed in our previous research [6]. The prototype was developed using only a portion of the Internet, the Yahoo! Entertainment sub-category (about 110,000 homepages). The SOM techniques, which often represent the results in terms of a graphical map, have been considered to exhibit computational properties similar to those of MDS and LSI. In [34], Orwig and Chen reported a prototype system that adopted the graphical SOM approach to organizing electronic brainstorming comments. The system compared favorably with human (expert) categorization results in concept recall, but its system precision levels was significantly worse than human results.

One of the major problems with large-scale classification systems has been the lack of support for effective visualization. Most of the clustering results discussed above can be represented as hierarchies. However, displaying large hierarchies (hundreds of thousands of nodes) is a challenging research problem.

One example of a visualization system supporting hierarchies is the Cone Tree structure developed at Xerox PARC [40]. Cone Trees have a three-dimensional hierarchical configuration. Each node of a tree is the apex of a cone. The 3D representation of the Cone Tree hierarchy maximizes the effective use of available screen space. However, portability difficulty of the 3D interface and the problem associated with displaying large hierarchies are still unresolved issues. Another example of a hierarchical structure used for representing information is a multitree, developed by Furnas and Zacks at Bellcore [11]. A multitree is a directed acyclic graph that has large easily identifiable substructures that are trees. The descendants of any node in a multitree form a tree, and the ancestors of any node form an inverted tree. Multitrees can share both leaves and complete subtrees.

Several novel hierarchy visualization methods have been proposed for displaying large hierarchies in the context of the Icon graphics programming language project [12]. In one variation, a tree is converted to a horizontal brick wall display. In another example, a tree is represented as a tree ring. (See Figure 3, Displaying Hierarchies.) The appealing 2D or even 3D display of large hierarchies (e.g., displaying a large hierarchy like the ``Great Wall of China") makes these methods a promising candidate for our research. (The methods were developed by Dr. Griswold at the Computer Science department of The University of Arizona. Significant local expertise is available.) In the user evaluation phase of our project, the large-scale automatic and manual classification systems produced will be represented using these novel hierarchy visualization methods.



(a) tree (b) brick wall (c) tree ring

Figure 3 - Visualizing Hierarchies

B. High-Performance Computing: Optimization and Parallelization

Although the performance of clustering algorithms, especially the hierarchical methods, has been shown in several small-scale applications, the computational complexity of such algorithms has caused severe implementation problems, especially for mid-to-large-scale applications. For a mid-scale application using the SOM algorithm such as the Internet homepage categorization project reported in [6] (about 100,000 homepages), it took about 10 hours on a DEC Alpha 3000/600 workstation (200 MHz, 128 MBs RAM).

Many researchers have attempted to optimize clustering algorithms, especially for sparse textual analysis applications. Rodriguez and Almeida [41] suggested improving SOM algorithm by starting with a small grid and adding nodes to a grid as the net begins to converge. The locations of the added nodes were interpolated from the locations of the old nodes. The improvement observed varied from marginal for small applications to 10-fold improvement for large networks. Koikkalainen et al. [22] suggested a way to improve the process of finding the winning node in maps which are almost converged. They replaced the exhaustive search method in SOM with an heuristic search technique for finding the winning node. Dmitri and Chen [42] developed a scalable SOM algorithm that took advantage of the sparsity of coordinates in the document input vectors and reduced the SOM computational complexity by several order of magnitude. The resulting complexity of the algorithm is proportional to the average number of non-zero coordinates in an input vector, instead of the total number of input vector coordinates. We believe the same ``keyword sparsity'' optimization principle observed in textual applications could be applied in the optimization of other conventional clustering algorithms as well. Our research will experiment with optimization on Ward's MDS, LSI, and SOM, respectively. Simulation and benchmarking on large-scale collections will be conducted to observe the scalability of these algorithms.

Other researchers have attempted to improve clustering through parallelization. Demian and Mignot [10] optimized SOM on parallel computers. Both SIMD and MIND architectures were tested. They assigned blocks of neurons (nodes) to each processor. The reported performance improvement was about 10 times for 128 processors. Chen and Yang [8] also parallelized SOM on the shared-memory multiprocessor (SMP) Convex Exemplar supercomputers. Multiple processors were used to find the winning node and to update weights of a winning neighborhood. A 10-fold improvement was also noted for SOM implemented on a 24-processor Exemplar. Using the high-performance parallel computing platforms made available to us from NCSA and SGI, we plan to parallellize ward's, MDS, LSI, and SOM for large-scale applications. Similar simulation and benchmarking experiments will be performed.

Testbed Plan: High-Performance Computing, Collections, and Classification Systems

A. High-Performance Computing:

The project PI, Dr. Hsinchun Chen, has worked extensively with the National Center for Supercomputing Applications (NCSA) and the SiliconGraphics Computer Systems (SGI) over the past several years and has received strong support from both organizations. Having served as a visiting senior research scientist with NCSA since 1996, Chen has received several NCSA High-Performance Computing Resources Grants through the competitive Peer Review Board (PRB) process, 1995-present. His recent project entitled: ``Parallel Computation for a Semantic Interoperability Environment'' was awarded 8,000 SUs (processing units) on the SGI Power Challenge and 4,000 on the CRAY SGI Origin2000. An NCSA grant support letter from Dr. Radha Nandkumar (NCSA Allocations Coordinator) is enclosed.

NCSA and its director, Dr. Larry Smarr, have been strong champions in promoting new high-bandwidth data-intensive scientific and knowledge management applications for the new millennium. High-performance digital library analysis is one of the critical areas supported by NCSA. Chen, serving as the representative of the digital library community, is also a member of the NCSA User Advisory Council. (See the attached NCSA User Advisory Council letter to Dr. Hsinchun Chen from Dr. Larry Smarr.) A continuous high-performance computing support from NCSA is expected for the duration of this project.

In addition to the NCSA support, a strong commitment has been made by SGI in support of this project. Last year SGI made a donation to Dr. Chen's AI lab in acquiring an 8-node SGI Origin2000 supercomputer. The equipment, which is in operation in the AI Lab, will serve as the main computing platform for the proposed project (and be upward-compatible with the NCSA's 512-node SGI Origin2000). A new commitment has been made by SGI to provide support for an additional 8-node Origin2000. Using the SGI's ``lego'' approach, we will be able to combine the two hypernodes into a significantly more powerful 16-node supercomputer. (A support letter from the SGI Account Manager, Matt Coover, is attched.) The proposed technical research in optimization and parallelization of clustering algorithms using large-scale collections will be performed using these high-performance SGI platforms.

B. Collections and Manual Classification Systems:

In order to fully benchmark the performances of the various clustering techniques and to explore their scalability across different domains, we propose to test these methods in three different application testbeds, all consisting of significant collections (millions of documents) and existing classification systems or thesauri.

- **Medicine:** A CancerLit collection has been made available to us through an ongoing project with the National Cancer Institute. (``Information Analysis and Visualization for Cancer Literature,'' PI: H. Chen and B. Schatz, National Cancer Institute, National Institutes of Health (NIH), July 1996-July 1999.) The CancerLit collection, which covers cancer abstracts from January 1992 to June 1998, consists of 714,537 documents from about 200 medical journals. Running clustering algorithms for such a large collection will be a significant challenge in optimization and parallelization.

  The NLM's Unified Medical Language System (UMLS), arguably the most comprehensive and fine-grained medical classification systems created by trained librarians, has also been available for our research through a collaborative agreement with NLM. (See the attached NLM agreement letter from Dr. Betsy Humphreys.) The UMLS Metathesaurus consists of 476,313 concepts and 1,051,901 different concept names from more than 40 different medical vocabularies. Mostly hierarchical, UMLS also represents other semantic relationships via its Semantic Network. Using the 700,000+ cancer documents, we plan to compare the resulting automatic classifications with the cancer-portion of the UMLS Metathesaurus.

- **Geoscience:** Two geoscience collections have been made available to us through a previous UCSB Alexandria Digital Library project [7]. (``Supplement to Alexandria DLI Project: A Semantic Interoperability Experiment for Spatially-Oriented Multimedia Data,'' PI: H. Chen and T. Smith, June 1996-May 1998, KK7022, IRI9411330). The 300,000-record Georef database (1990-1995), provide by the American Geological Institute, is one of the largest collections in geoscience. While GeoRef covers most of geography and geology, Petroleum Abstracts (PA) cover petroleum engineering and petroleum exploration. They overlap with GeoRef only on the earth science side. We have obtained the 1985-1995 collection of the Petroleum Abstracts (about 500,000 documents) from Petroleum Abstracts Service of The University of Tulsa.

  A manually-created Georef thesaurus, consisting of more than 27,000 terms with standard hierarchical relationships, has also been made available to us from the American Geological Institute. Automatic geoscience classification systems to be generated using the Georef and Petroleum Abstract collections will be compared with the Georef thesaurus.

- **Web pages:** Lastly, to further validate the robustness of the proposed clustering algorithms, our final testbed will consist of web pages collected by Internet spiders. A testbed of about 1.5M web pages has been created recently in a previous NSF-funded project. (``Concept-based Categorization and Search on Internet: A Machine Learning, Parallel Computing Approach,'' PI: H. Chen, September 1995-August 1998, IRI9525790.) The noise and diversity in web pages pose a significant challenge to clustering techniques.

  We plan to compare our automatic classification systems with the Yahoo! manual classification (14 top category headings and about 20,000 category headings). Having the ability to create robust, efficient, automatic classification systems will make a significant contribution to the organization and discovery of web resources.

User Evaluation Plan: Medicine, Geoscience, and Web

Different quantitative and qualitative evaluation measurements will be developed to compare the performances of the automatic and manual classification systems.

There is surprisingly little literature on evaluating clusters and other hierarchies. At present, we are unaware of any research involving methodological evaluation of clustering of textual documents. In this research we propose to borrow an evaluation procedure adopted in experimental computational linguistics [15]. Hatzvassiloglou and McKeown used human experts and quantitative measures to evaluate partitions of adjectives. They based their measurement on whether or not a pair of objectives was put into the same class by the human expert and by the system. We refer to this measure as *contingency error* and is defined as the number of incorrect associations divided by the the number of pairs of documents.

We also plan to adopt *cluster recall* and *cluster precision*, similar to the traditional IR *recall* and *precision* measures. Rather than examining the number of relevant documents, we count the number of associations between documents. An association is a pair of documents belonging to the same cluster. The *correct associations* are those that are created by human experts.

Three institutions within The University of Arizona will assist in providing expert subjects for user evaluation:

- Arizona Cancer Center: With about 500 researchers and staffs in this comprehensive cancer center, we plan to recruit physicians, researchers, and graduate students as expert subjects to evaluate the cancer-related classification systems.
- Arizona Health Science Library: Serving the entire medical community on the campus, staffs and patrons in the library will be solicited to participate in the evaluation of the proposed cancer-related classification systems.
- UA Main Library and Science and Engineering Library: Serving the entire campus, staffs and patrons of the two UA libraries will be solicited to participate in the evaluation of the proposed geoscience and web-related classification systems.

All three institutions have been involved in Chen's previous research projects and are strongly supportive of the proposed research activities. The following support letters are attached: (1) Carla Stoffle, Dean of Libraries, University Library, The University of Arizona, (2) Rachael Anderson, Director, Arizona Health Science Library, and (3) Garth Powis, Director of Basic Science, Arizona Cancer Center.

User experiments will contrast and evaluate four automatic classification systems and the manual classification (control) using a Latin Square Design. Sampling will be performed on the classification systems to allow subjects to evaluate only a manageable sueset of the classification systems, i.e., selected terms/documents and their relationships. This design allows the use of multiple independent variables for each subject by varying the order in which the independent variables are presented. Subjects of different expertise will be assigned to experiments that require their domain knowledge.

In addition to the above quantitative measures, immediately after the experiment session subjects will be asked to evaluate the effectiveness of the five conditions (four techniques and one control) across the dependent variables. Measurements of these dependent variables will include verbal protocol analysis (during each session), questionnaires (after each session), and interviews consisting of both structured and unstructured questions (after all sessions). Data analysis will consist of content analysis of verbal protocols, various multivariate statistical analyses of questionnaire data, and content analysis of interviews. We will develop an overall index of effectiveness which we will use to evaluate the ``best'' overall technique. However, we will also assess individual strengths and weaknesses of each technique across all dependent variables and explore methods to integrate selected manual and automatic classification systems.

Project Management Plan

Chen, who is experienced in textual analysis and clustering techniques for various digital library applications, will serve as PI of the project and will supervise three graduate research assistants in algorithm development and the prototyping effort. Sewell, who is trained in medicine (10 years), library science, and user study (3 years) will serve as Co-PI and will supervise one graduate research assistant in testbed management and user evaluation.

The proposed research will last three years and will be conducted roughly in four phases (although exploratory work may be conducted prior to the scheduled phases). Phases 1 and 3, each of which will last for one year will mainly concern individual algorithm development (Phase 1) and large-scale testbed analysis (Phase 2), to be supervised by Chen. Phases 2 and 4, which will immediately follow the system development efforts, will focus on user evaluation, each lasting 6 months.

Chen primarily will employ system benchmarking and information systems evaluation measures (e.g., recall/precision, contingency error). Sewell will conduct a series of controlled laboratory experiments to evaluate the effectiveness of the various classification systems.

Results from Prior NSF Support

The proposed project will significantly expand research in three previous NSF-funded projects in scientific collaboration (1992-1994), digital libraries (1994-1998), and Internet computing (1994-1998).

A. Scientific Collaboration:

Hsinchun Chen, Principal investigator (PI), Research Initiation Award, National Science Foundation, ``Building a Concept Space for an Electronic Community System,'' $63,804, June 1992-November 1994 (IRI9211418).

Chen was Co-PI for an NSF-funded National Collaboratory project (``Worm Community System,'' PI: Schatz) which built a community system in molecular biology referenced as a national model for future science information systems [38]. Chen's contribution was in developing information analysis techniques to support scientific information retrieval and information sharing. A (nematode) worm concept space and a fly (Drosophila) concept space have been developed based on automatic indexing and cluster analysis techniques and can be used to assist in cross-domain concept exploration and term suggestion during scientific collaboration. They are in use by worm biologists. Several important journal publications have been generated from this work:

1. H. Chen, K. J. Lynch, K. Basu, and T. Ng, ``Generating, Integrating, and Activating Thesauri for Concept-Based Document Retrieval," *IEEE Expert*, Special Series on Artificial Intelligence in Text-Based Information Systems, Volume 8, Number 2, Pages 25-34, April, 1993.
2. H. Chen, B. Schatz, T. Yim, and D. Fye, ``Automatic Thesaurus Generation for an Electronic Community System," *Journal of the American Society for Information Science*, Volume 46, Number 3, Pages 175-193, April 1995.

B. Digital Libraries:

Hsinchun Chen, Co-PI (PI: B. Schatz, University of Illinois), Digital Library Initiative, NSF/ARPA/NASA, ``Building the Interspace: Digital Library Infrastructure for a University Engineering Community," $4,000,000, September 1994-August 1998 (IRI9411318, UA subcontract: $500,000).

Chen is a Co-PI in the NSF/NASA/ARPA-funded Illinois ``Digital Library Initiative" project and is responsible for developing semantic (concept-based) retrieval, semantic federation, and vocabulary switching capabilities for a large testbed of SGML scientific literature. Chen and Schatz also served as guest editors for the May 1996 and February 1999 issues of *IEEE Computer* on ``Digital Libraries." The project is ending in August 1998 and selected techniques, have started to appear in several major journal publications.

1. B. Schatz and H. Chen, ``Building Large-Scale Digital Libraries," *IEEE Computer,* Special Issue on ``Building Large-scale Digital Libraries," Volume 29, Number 5, Pages 22-27, May, 1996.
2. B. Schatz, B. Mischo, T. Cole, J. Hardin, A. Bishop, and H. Chen, ``Federating Diverse Collections of Scientific Literature," *IEEE Computer,* Special Issue on ``Building Large-scale Digital Libraries," Volume 29, Number 5, Pages 28-36, May, 1996.
3. H. Chen, B. R. Schatz, T. D. Ng, J. P. Martinez, A. J. Kirchhoff, C. Lin, ``A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Special Section on Digital Libraries: Representation and Retrieval, Volume 18, Number 8, Pages 771-782, August, 1996.
4. H. Chen, J. Martinez, A. Kirchhoff, T. D. Ng, and B. R. Schatz, ``Alleviating Search Uncertainty Through Concept Associations: Automatic Indexing, Co-occurrence Analysis, and Parallel Computing," Special Issue on ``Management of Imprecision and Uncertainty in Information Retrieval and Database Management Systems," Volume 49, Number 3, Pages 206-216, 1998.

C. Internet Computing:

Hsinchun Chen, Principal investigator (PI), National Science Foundation, CISE, IRIS, ``Concept-based Categorization and Search on Internet: A Machine Learning, Parallel Computing Approach," $200,755, September 1995-August 1998 (IRI9525790).

Chen is PI of an NSF/CISE-funded ``Internet Categorization and Search" project. The research attempts to address the Internet search problem by first *categorizing* the content of Internet documents (using the SOM algorithm) and subsequently providing semantic search capabilities based on a *concept space* and a genetic algorithm spider (agent). Sample agent-based prototype systems have been developed and have appeared in several journals:

1. H. Chen, C. Schuffels, and R. Orwig, ``Internet Categorization and Search: A Machine Learning Approach," *Journal of Visual Communication and Image Representation,* Special Issue on Digital Libraries, Volume 7, Number 1, Pages 88-102, 1996.
2. H. Chen, A. L. Houston, R. R. Sewell, and B. R. Schatz, ``Internet Browsing and Searching: User Evaluation of Category Map and Concept Space Techniques," *Journal of the American Society for Information Science,* Special Issue on AI Techniques for the Emerging Information Systems Applications, Volume 49, Number 7, Pages 582-603, 1998.
3. H. Chen, Y. Chung, M. Ramsey, and C. Yang, ``A Smart Itsy Bitsy Spider for the Web," *Journal of the American Society for Information Science,* Special Issue on AI Techniques for the Emerging Information Systems Applications, Volume 49, Number 7, Pages 604-618, 1998.

References

1. B. T. Bartell, G. W. Cottrell, and R.K. Belew.
   Representing documents using an explicit model of their similarities.
   Journal of the American Society for Information Science, 46(4):254-271, 1995.

2. H. Chen.
   Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms.
   Journal of the American Society for Information Science, 46(3):194-216, April 1995.

3. H. Chen and V. Dhar.
   User misconceptions of online information retrieval systems.
   International Journal of Man-Machine Studies, 32(6):673-692, June 1990.

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

5. H. Chen, B. R. Schatz, T. D. Ng, J. P. Martinez, A. J. Kirchhoff, and C. Lin.
   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, August 1996.

6. H. Chen, C. Schuffels, and R. Orwig.
   Internet categorization and search: a machine learning approach.
   Journal of Visual Communications and Image Representation, 7(1):88-102, March 1996.

7. H. Chen, T. R. Smith, M. L. Larsgaard, L. L. Hill, and M. Ramsey.
   A geographic knowledge representation system (GKRS) for multimedia geospatial retrieval and analysis.
   International Journal of Digital Library, 1(2):132-152, September 1997.

8. H. Chen and M. Yang.
   Self-organizing map optimization using Exemplar supercomputers.
   In *Center for Management of Information, University of Arizona, Working Paper, CMI-WPS 96-15*, 1996.

9. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman.
   Indexing by latent semantic analysis.
   Journal of the American Society for Information Science, 41(6):391-407, September 1990.

10. V. Demian and J. C. Mignot.
    Implementation of the self-organizing feature map on parallel computers.
    In L. Bouge, M. Cosnard, Y. Robert, and D. Trystram, editors, *Proceedings of the Second Joint International Conference on Vector and Parallel Processing*, pages 775-776, Berlin, Heidelberg, 1992. Springer.

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

12. R. E. Griswold and M. T. Griswold.
The Icon Programming Language.
Prentice Hall, Englewood Cliffs, NJ, 1990.

13. N. Guarino.
The role of formal ontology in the information technology.
International Journal of Human-Computer Studies, 43(5/6):623-624, 1995.

14. M. D. Harris.
Introduction to natural Language Processing.
Reston Publishing Company, Inc., Reston, Virginia, 1985.

15. T. Hatzivassiloglou and R. McKeown.
Towards tth automatic identification of adjectival scales: clustering adjectives according to meaning.
In *Proceedings of the 31st Annual meting of the Association for Computational Linguistics*, pages 172-182, 1993.

16. A. L. Houston, H. Chen, B. R. Schatz, R. R. Sewell, K. M. Tolle, T. E. Doszkocs, S. M. Hubbard, and D. T. Ng.
Exploring the use of concept space, category map techniques, and natural language parsers to improve medical information retrieval.
Decision Support Systems, page forthcoming, 1998.

17. B. L. Humphreys and D. A. Lindberg.
Building the unified medical language system.
In *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care*, pages 475-480, Washington, DC: IEEE Computer Society Press, November, 5-8 1989.

18. P. S. Jacobs and L. F. Rau.
Innovations in text interpretation.
Artificial Intelligence, 63:143-193, 1993.

19. A. K. Jain and R. C. Dubes.
Algorithms for Clustering Data.
Prentice-Hall, Inc., Englewood Cliffs, NJ, 1988.

20. S. Kaski, T. Honkela, K. Lagus, and T. Kohonen.
Creating an order in digital libraries with self-organizing maps.
In *Submitted to WCNN'96, World Congress on Neural Networks, San-Diego, CA*, September 1996.

21. T. Kohonen.
Self-Organization Maps.
Springer-Verlag, Berlin Heidelberg, 1995.

22. P. Koikkalainen.
Fast deterministic self-organizing maps.
In F. Fogelman-Soulié and P. Gallinari, editors, *Proceedings of the International Conference on Artificial Neural Networks*, pages 63-68, Nanterre, France, 1995.

23. J. B. Kruskal.
Multidimensional Scaling.
Sage university papers series. Quantitative applications in the social sciences, Beverly Hills, CA, 1978.

24. D. B. Lenat, A. Borning, D. McDonald, C. Taylor, and S.i Weyer.
Knoesphere: Building expert systems with encyclopedic knowledge.
In *International Joint Conference of Artificial Intelligence*, 1983.

25. D. B. Lenat, R. Guha, K. Pittman, D. Pratt, and M. Shepherd.
CYC: Toward programs with common sense.
Communications of the ACM, 33(8):30-49, August 1990.

26. M. Lesk.
Practical Digital Libraries.
Morgan Kauffmann, Los Altos, CA, 1997.

27. 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*, pages 262-269, Chicago, IL, October 13-16 1991.

28. K. J. Lynch and H. Chen.
Knowledge discovery from historical data: an algorithmic approach.
In *Proceedings of the 25th Annual Hawaii International Conference on System Sciences (HICSS-25), Decision Support and Knowledge Based Systems Track*, pages 70-79, Kaui, HI, January 7-10 1992.

29. B. S. Manjunath and W. Y. Ma.
Texture features for browsing and retrieval of image data.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(8):837-841, August 1996.

30. A. T. McCray and W. T. Hole.
The scope and structure of the first version of the UMLS semantic network.
In *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care*, pages 126-130, Los Alamitos, CA: Institute of Electrical and Electronics Engineers, November, 4-7 1990.

31. G. Miller.
Special issue, wordnet: An on-line lexical database.
International Journal of Lexicography, 3(4), 1990.

32. F. Murtagh.
Complexities of hierarchical clustering algorithms: state of the art.
Computational Statistics Quarterly, 1:101-113, 1984.

33. D. E. O'Leary.
Enterprise knowledge management.
IEEE Computer, 31(3):54-61, March 1998.

34. R. Orwig, H. Chen, and J. F. Nunamaker.
A graphical, self-organizing approach to classifying electronic meeting output.
Journal of the American Society for Information Science, 48(2):157-170, February 1997.

35. A. Paepcke, S. B. Cousins, H. Garcia-Molino, S. W. Hasson, S. P. Ketcxhpel, M. Roscheisen, and T. Winograd.
Using distributed objects for digital library interoperability.
IEEE COMPUTER, 29(5):61-69, May 1996.

36. F. C. Pereira and B. J. Grosz.
Natural Language Processing.
The MIT Press, Cambridge, MA, 1994.

37. G. Piatetsky-Shapiro.
Workshop on knowledge discovery in real databases.
In *International Joint Conference of Artificial Intelligence*, 1989.

38. R. Pool.
Beyond database and e-mail.
Science, 261:841-843, 13 August 1993.

39. E. Rasmussen.
Clustering algorithms.
In *Information Retrieval: Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates, Editors, Prentice Hall, Englewood Cliffs, NJ, 1992.

40. G. Robertson, S.K. Card, and J. Mackinlay.
Information visualization using 3D interactive animation.
Communications of the ACM, pages 57-71, April 1993.

41. J. S. Rodrigues and L. B. Almeida.
Improving the learning speed in topological maps of patterns.
In *Proceedngs of International Conference on Neural Networks*, pages 813-816, Dordrecht, Netherlands, 1990. Kluwer Academic Publishers.

42. D. Roussinov and H. Chen.
A scalable self-organizing map algorithm for textual classification: A neural network approach to automatic thesaurus generation.
Communication and Cognition in Artificial Intelligence Journal, page forthcoming, 1998.

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

44. B. R. Schatz.
Information retrieval in digital libraries: bring search to the net.
Science, 275:327-334, January 17 1997.

45. B. R. Schatz and H. Chen.
Building large-scale digital libraries.
IEEE COMPUTER, 29(5):22-27, May 1996.

46. B. R. Schatz, B. Mischo, T. Cole, J. Hardin, A. Bishop, and H. Chen.
Federating repositories of scientific literature.
IEEE COMPUTER, 29(5):28-36, May 1996.

47. L. Smarr.
Commercial applications program at NCSA.
NCSA Commercial Application Workshop, Urbana-Champaign, IL, April 24-25 1995.

48. Tipster.
TIPSTER Text Phase II.
24-month Workshop, Tysons Center, VA, May, 1996.

49. A. Voutilainen.
A Short Introduction to NPTool, 1997.
http://www.lingsoft.fi/doc/nptool/intro/.

50. H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens.
Intelligent access to digital video: Informedia project.
IEEE COMPUTER, 29(5):46-53, May 1996.

51. E. Field.
Knowledge management.
CIO Magazine, 11(17):41-48, June 15 1998.

52. E. Wakin.
Knowledge management: Tapping intellectual capital.
Beyond Computing, 7(4):1-8, may 1998.

53. R. Wilensky.
Toward work-centered digital information services.
IEEE COMPUTER, 29(5):37-45, May 1996.

Hsinchun Chen

3665 N. Longwood Pl.
Tucson, Arizona 85715
Home: (602) 722-6808
Office: (602) 621-4153

MIS Department
Karl Eller Graduate
School of Management
University of Arizona
Tucson, Arizona 85721

Education

Ph.D.	Information Systems	New York University	1989
Master of Philosophy	Information Systems	New York University	1987
MBA	Management Information Systems	State University of New York	
	Mgmt Science, Finance	at Buffalo	1985
B.S.	Management Science	National Chiao-Tung	
		University, Taiwan	1981

Ph.D. Dissertation: ``An Artificial Intelligence Approach to the Design of Online Information Retrieval Systems,'' October 1989, directed by Professor V. Dhar.

Academic and Professional Experience

University of Arizona, Department of MIS, Professor (with tenure).

Head, University of Arizona, MIS Artificial Intelligence Group/Laboratory, 1992-present.

University of Illinois at Urbana-Champaign, National Center for Supercomputing Applications (NCSA), Visiting Senior Research Scientist, 1996-present.

Guest editor, *IEEE Computer* May 1996 and February 1999 special issues on ``Digital Libraries'' (with B. R. Schatz).

Guest editor, *Journal of the American Society for Information Science* special issue on ``Artificial Intelligence Techniques for Emerging Information Systems Applications,'' May 1998.

Recognition and Awards

``Information Retrieval in Digital Libraries: Bring Search to the Net,'' Featured in Volume 275 of *Science*, January 17, 1997 (cover article).

``Digital Libraries Computation Cracks Semantic Barriers Between Databases,'' Featured in Volume 272 of *Science*, June 7, 1996.

AT&T Foundation Award in Science and Engineering, 1994-1996.

Best Paper Award, ``A Machine Learning Approach to Document Retrieval: An Overview and an Experiment,'' in the *27th Annual Hawaii International Conference on System Sciences (HICSS-27)*, Maui, Hawaii, January 4-7, 1994.

NSF Research Initiation Award, Division of Information, Robotics, and Intelligent Systems, Directorate for Computer and Information Sciences and Engineering, 1992-1994.

Grants

Co-PI of the NSF/NASA/ARPA-funded Illinois ``Digital Library Initiative" project (PI: B. Schatz).

Principal investigator (PI), National Science Foundation, CISE, IRIS, ITO, ``Concept-based Categorization and Search on Internet: A Machine Learning, Parallel Computing Approach," September 1995-August 1998 (IRI9525790).

Principal investigator (PI), Research Initiation Award, National Science Foundation, ``Building a Concept Space for an Electronic Community System," June, 1992- November, 1994 (IRI9211418).

Refereed Journal Publications

1. H. Chen, K. J. Lynch, K. Basu, and T. Ng, ``Generating, Integrating, and Activating Thesauri for Concept-Based Document Retrieval," *IEEE Expert*, Special Series on Artificial Intelligence in Text-Based Information Systems, Volume 8, Number 2, Pages 25-34, April, 1993.

2. H. Chen, ``Collaborative Systems: Solving the Vocabulary Problem," *IEEE Computer*, Special Issue on Computer-Supported Cooperative Work, Volume 27, Number 5, Pages 58-66, May, 1994.

3. H. Chen, P. Hsu, R. Orwig, L. Hoopes, and J. Nunamaker, ``Automatic Concept Classification of Text from Electronic Meetings," *Communications of the ACM*, Volume 37, Number 10, Pages 56-73, October 1994.

4. H. Chen, ``Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms," *Journal of the American Society for Information Science*, Volume 46, Number 3, Pages 194-216, April 1995.

5. H. Chen, C. Schuffels, and R. Orwig, ``Internet Categorization and Search: A Machine Learning Approach," *Journal of Visual Communication and Image Representation*, Special Issue on Digital Libraries, Volume 7, Number 1, Pages 88-102, 1996.

6. B. Schatz and H. Chen, ``Building Large-Scale Digital Libraries," *IEEE Computer*, Special Issue on ``Building Large-scale Digital Libraries," Volume 29, Number 5, Pages 22-27, May, 1996.

7. H. Chen, A. Houston, J. Yen, and J. F. Nunamaker, ``Toward Intelligent Meeting Agents," *IEEE Computer*, Volume 29, Number 8, Pages 62-70, August, 1996.

8. H. Chen, B. R. Schatz, T. D. Ng, J. P. Martinez, A. J. Kirchhoff, C. Lin, ``A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Special Section on Digital Libraries: Representation and Retrieval, Volume 18, Number 8, Pages 771-782, August, 1996.

9. H. Chen, J. Martinez, T. D. Ng, and B. R. Schatz, ``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*, Volume 48, Number 1, Pages 17-31, January, 1997.

10. H. Chen, A. L. Houston, R. R. Sewell, and B. R. Schatz, ``Internet Browsing and Searching: User Evaluation of Category Map and Concept Space Techniques," *Journal of the American Society for Information Science*, Special Issue on AI Techniques for Emerging Information Systems Applications, Volume 49, Number 7, Pages 582-603, 1998.

Collaborators

Bruce Schatz, Mary Larsgaard, Terry Smith, Rich Orwig, B. S. Manjunath, Jay Nunamaker

Robin R. Sewell

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

Education

1996	M.L.A.	University of Arizona	School of Information Resources and Library Science
1986	D.V.M.	Washington State University	Veterinary Medicine
1984	B.S.	Washington State University	Veterinary Science

Academic and Professional Experience

1997-present Research Specialist and Program Coordinator, Artificial Intelligence Laboratory, Management Information Systems Dept, University of Arizona

1996-97 Graduate Research Assistant, Artificial Intelligence Laboratory, Management Information Systems Dept, University of Arizona

1986-96 Doctor of Veterinary Medicine, Small Animal Practice, Phoenix, Arizona

Honors and Awards

President, Student Chapter of the Special Libraries Association, University of Arizona School of Information Resources and Library Science, Fall 1996

Publications Related to the Proposed Research

Chen, H., Houston, A., Sewell, R., and Schatz, B., ``Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques." JASIS, vol.49, no.7 p582-603

Houston, A., Chen, H., Schatz, B., Sewell, R., Tolle, K., Doskocs, T., Hubbard, S., and Ng, T. ``Exploring the Use of Concept Space, Category Map Techniques, and Natural Language Parsers to Improve Medical Information Retrieval." Decision Support Systems, Special Issue on Decision Support for Health Care in a New Information Age, 1998, forthcoming.

Houston, A., Chen, H., Hubbard, S., Schatz, B., Ng, T., Sewell, R., and Tolle, K. ``Health Care Information Infrastructures: A Critical Component of the NII," Journal of the American Society for Information Science, 1998.

Collaborators

Hsinchun Chen, Andrea Houston, Bruce Schatz.

References Cited

54. B. T. Bartell, G. W. Cottrell, and R.K. Belew. Representing documents using an explicit model of their similarities. *Journal of the American Society for Information Science*, 46(4):254-271, 1995.

55. H. Chen. Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms. *Journal of the American Society for Information Science*, 46(3):194-216, April 1995.

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

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

58. H. Chen, B. R. Schatz, T. D. Ng, J. P. Martinez, A. J. Kirchhoff, and C. Lin. 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, August 1996.

59. H. Chen, C. Schuffels, and R. Orwig. Internet categorization and search: a machine learning approach. *Journal of Visual Communications and Image Representation*, 7(1):88-102, March 1996.

60. H. Chen, T. R. Smith, M. L. Larsgaard, L. L. Hill, and M. Ramsey. A geographic knowledge representation system (GKRS) for multimedia geospatial retrieval and analysis. *International Journal of Digital Library*, 1(2):132-152, September 1997.

61. H. Chen and M. Yang. Self-organizing map optimization using Exemplar supercomputers. In *Center for Management of Information, University of Arizona, Working Paper, CMI-WPS 96-15*, 1996.

62. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391-407, September 1990.

63. V. Demian and J. C. Mignot. Implementation of the self-organizing feature map on parallel computers. In L. Bouge, M. Cosnard, Y. Robert, and D. Trystram, editors, *Proceedings of the Second Joint International Conference on Vector and Parallel Processing*, pages 775-776, Berlin, Heidelberg, 1992. Springer.

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

65. R. E. Griswold and M. T. Griswold. *The Icon Programming Language*. Prentice Hall, Englewood Cliffs, NJ, 1990.

66. N. Guarino. The role of formal ontology in the information technology. *International Journal of Human-Computer Studies*, 43(5/6):623-624, 1995.

67. M. D. Harris. *Introduction to natural Language Processing*. Reston Publishing Company, Inc., Reston, Virginia, 1985.

68. T. Hatzivassiloglou and R. McKeown. Towards the automatic identification of adjectival scales: clustering adjectives according to meaning. In *Proceedings of the 31st Annual meting of the Association for Computational Linguistics*, pages 172-182, 1993.

69. A. L. Houston, H. Chen, B. R. Schatz, R. R. Sewell, K. M. Tolle, T. E. Doszkocs, S. M. Hubbard, and D. T. Ng. Exploring the use of concept space, category map techniques, and natural language parsers to improve medical information retrieval. *Decision Support Systems*, page forthcoming, 1998.

70. B. L. Humphreys and D. A. Lindberg. Building the unified medical language system. In *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care*, pages 475-480, Washington, DC: IEEE Computer Society Press, November, 5-8 1989.

71. P. S. Jacobs and L. F. Rau. Innovations in text interpretation. *Artificial Intelligence*, 63:143-193, 1993.

72. A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1988.

73. S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. Creating an order in digital libraries with self-organizing maps. In *Submitted to WCNN'96, World Congress on Neural Networks, San-Diego, CA*, September 1996.

74. T. Kohonen. *Self-Organization Maps*. Springer-Verlag, Berlin Heidelberg, 1995.

75. P. Koikkalainen. Fast deterministic self-organizing maps. In F. Fogelman-Soulié and P. Gallinari, editors, *Proceedings of the International Conference on Artificial Neural Networks*, pages 63-68, Nanterre, France, 1995.

76. J. B. Kruskal. *Multidimensional Scaling.* Sage university papers series. Quantitative applications in the social sciences, Beverly Hills, CA, 1978.

77. D. B. Lenat, A. Borning, D. McDonald, C. Taylor, and S.i Weyer. Knoesphere: Building expert systems with encyclopedic knowledge. In *International Joint Conference of Artificial Intelligence*, 1983.

78. D. B. Lenat, R. Guha, K. Pittman, D. Pratt, and M. Shepherd. CYC: Toward programs with common sense.
Communications of the ACM, 33(8):30-49, August 1990.

79. M. Lesk. *Practical Digital Libraries*. Morgan Kauffmann, Los Altos, CA, 1997.

80. 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*, pages 262-269, Chicago, IL, October 13-16 1991.

81. K. J. Lynch and H. Chen. Knowledge discovery from historical data: an algorithmic approach. In *Proceedings of the 25th Annual Hawaii International Conference on System Sciences (HICSS-25), Decision Support and Knowledge Based Systems Track*, pages 70-79, Kaui, HI, January 7-10 1992.

82. B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837-841, August 1996.

83. A. T. McCray and W. T. Hole. The scope and structure of the first version of the UMLS semantic network.
In *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care*, pages 126-130, Los Alamitos, CA: Institute of Electrical and Electronics Engineers, November, 4-7 1990.

84. G. Miller. Special issue, wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 1990.

85. F. Murtagh. Complexities of hierarchical clustering algorithms: state of the art. *Computational Statistics Quarterly*, 1:101-113, 1984.

86. D. E. O'Leary. Enterprise knowledge management. *IEEE Computer*, 31(3):54-61, March 1998.

87. R. Orwig, H. Chen, and J. F. Nunamaker. A graphical, self-organizing approach to classifying electronic meeting output. *Journal of the American Society for Information Science*, 48(2):157-170, February 1997.

88. A. Paepcke, S. B. Cousins, H. Garcia-Molino, S. W. Hasson, S. P. Ketcxhpel, M. Roscheisen, and T. Winograd.
Using distributed objects for digital library interoperability. *IEEE COMPUTER*, 29(5):61-69, May 1996.

89. F. C. Pereira and B. J. Grosz. *Natural Language Processing*. The MIT Press, Cambridge, MA, 1994.

90. G. Piatetsky-Shapiro. Workshop on knowledge discovery in real databases. In *International Joint Conference of Artificial Intelligence,* 1989.

91. R. Pool. Beyond database and e-mail. *Science,* 261:841-843, 13 August 1993.

92. E. Rasmussen. Clustering algorithms. In *Information Retrieval: Data Structures and Algorithms,* W. B. Frakes and R. Baeza-Yates, Editors, Prentice Hall, Englewood Cliffs, NJ, 1992.

93. G. Robertson, S.K. Card, and J. Mackinlay. Information visualization using 3D interactive animation. *Communications of the ACM*, pages 57-71, April 1993.

94. J. S. Rodrigues and L. B. Almeida. Improving the learning speed in topological maps of patterns. In *Proceedings of International Conference on Neural Networks*, pages 813-816, Dordrecht, Netherlands, 1990. Kluwer Academic Publishers.

95. D. Roussinov and H. Chen. A scalable self-organizing map algorithm for textual classification: A neural network approach to automatic thesaurus generation. *Communication and Cognition in Artificial Intelligence Journal*, page forthcoming, 1998.

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

97. B. R. Schatz. Information retrieval in digital libraries: bring search to the net. *Science*, 275:327-334, January 17 1997.

98. B. R. Schatz and H. Chen. Building large-scale digital libraries. *IEEE COMPUTER*, 29(5):22-27, May 1996.

99. B. R. Schatz, B. Mischo, T. Cole, J. Hardin, A. Bishop, and H. Chen. Federating repositories of scientific literature. *IEEE COMPUTER*, 29(5):28-36, May 1996.

100. L. Smarr. *Commercial applications program at NCSA*. NCSA Commercial Application Workshop, Urbana-Champaign, IL, April 24-25 1995.

101. Tipster. *TIPSTER Text Phase II*. 24-month Workshop, Tysons Center, VA, May, 1996.

102. A. Voutilainen. *A Short Introduction to NPTool*, 1997.
http://www.lingsoft.fi/doc/nptool/intro/.

103. H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens. Intelligent access to digital video: Informedia project. *IEEE COMPUTER*, 29(5):46-53, May 1996.

104. E. Field. Knowledge management. *CIO Magazine*, 11(17):41-48, June 15 1998.

105. E. Wakin. Knowledge management: Tapping intellectual capital. *Beyond Computing*, 7(4):1-8, may 1998.

106. R. Wilensky. Toward work-centered digital information services. *IEEE COMPUTER*, 29(5):37-45, May 1996.