Int J Digit Libr (1997) 1: 132–152

INTERNATIONAL JOURNAL ON
Digital Libraries

A geographic knowledge representation system for multimedia geospatial retrieval and analysis

Hsinchun Chen[1], Terence R. Smith[2], Mary L. Larsgaard[3], Linda L. Hill[4], Marshall Ramsey[5]

[1] Associate Professor, MIS Department, Karl Eller Graduate School of Management, University of Arizona, McClelland Hall 430Z, Tucson, AZ 85721, USA, Visiting Senior Research Scientist, NCSA, hchen@bpa.arizona.edu, (520) 621-4153.
[2] Director, Alexandria Digital Library Project, University of California at Santa Barbara, Santa Barbara, CA 93106, USA, smithtr@cs.ucsb.edu
[3] Assistant Director, Map and Imagery Laboratory, Davidson Library, University of California at Santa Barbara, Santa Barbara, CA 93106, USA, mary@sdc.ucsb.edu
[4] Senior Research Scientist, University of Maryland, College of Library and Information Services, Universities Space Research Association, Center of Excellence in Space Data and Information Sciences, Goddard Space Flight Center, Code 930.5, Greenbelt, MD 20771, USA, lhill@alexandria.sdc.ucsb.edu
[5] MIS Department, Karl Eller Graduate School of Management, University of Arizona, McClelland Hall 430W, Tucson, AZ 85721, USA, mramsey@bpa.arizona.edu, (520) 621-2748

Received: 15 October 1996 / Accepted: 14 January 1997

Abstract. Digital libraries serving multimedia information that may be accessed in terms of geographic content and relationships are creating special challenges and opportunities for networked information systems. An especially challenging research issue concerning collections of geo-referenced information relates to the development of techniques supporting *geographic information retrieval* (GIR) that is both fuzzy and concept-based. Viewing the meta-information environment of a digital library as a heterogeneous set of services that support users in terms of GIR, we define a *geographic knowledge representation system* (GKRS) in terms of a core set of services of the meta-information environment that is required in supporting concept-based access to collections of geospatial information. In this paper, we describe an architecture for a GKRS and its implementation in terms of a prototype system. Our GKRS architecture loosely couples a variety of multimedia knowledge sources that are in part represented in terms of the semantic network and neural network representations developed in artificial intelligence research. Both textual analysis and image processing techniques are employed in creating these textual and iconic geographical knowledge structures. The GKRS also employs spreading activation algorithms in support of concept-based knowledge retrieval. The paper describes implementational details of several of the components of the GKRS as well as discussing both the lessons learned from, and future directions of, our research.

Key words: Geographic information retrieval – Geographic knowledge representation – Geospatially referenced information – Multimedia geospatial queries – Geospace

1 Introduction

Digital library technology is beginning to support major increases in both the availability and usefulness of geospatial information [27]. In accordance with the results of user surveys of traditional map libraries [13] [19], digital libraries that provide access to geospatially referenced material have become increasingly popular in supporting users of various levels, ranging from scientists to general users, for such diverse purposes as research, coursework, business, and recreation. These uses require many of the classes of multimedia documents that are contained in the collections of such digital libraries including: traditional, text-based geographic literature; technical reports and surveys; maps, aerial photos, satellite images, and digital elevation models; and a large variety of scientific datasets. In relation to such materials, the networked information systems and multimedia technologies popularized by the Internet and the Web have opened a floodgate of expectations among users for concept-based, geospatial access to such information and for the analysis of such information.

There are many different classes of users of geospatially referenced materials, ranging from researchers and college students to home owners and school children, and many different applications in which such information can be used. Geoscience researchers, for example, have begun to use various satellite-sensing data and images to study global climatic changes and their environmental impacts; home owners and developers can use digital elevation models, census data, and aerial photos to evaluate construction projects and perform feasibility studies; students at various educational levels are increasingly able to use Internet browsers in accessing various maps

and metadata for class assignments or personal recreation.

Based on several user studies, Gluck [19] suggested that conventional map libraries, which are currently the main providers of geospatial information, are weak in supporting the analysis and interpretation of geo-referenced materials in these various applications. Furthermore, the recent emergence of Internet-accessible geographic information systems (GIS), that provide support for such analysis and interpretation [22] [27] [44], are of limited value in such environments because they lack support for fuzzy, concept-based queries that may be applied to multimedia, georeferenced databases.

Gluck [19] concluded that multimedia information types are essential for supporting geospatial queries. In particular, maps seem to provide important survey knowledge and to expose spatial patterns, while text supports answering factual questions and clarifying causal relationships. Such complementary roles for text and maps have also been confirmed by Hill [22]. Gluck's experiments suggest that both traditional and digital libraries should provide more value-added services and products for geospatial analysis and interpretation [19]. From a system development perspective, these results suggest the importance of creating more pro-active and "intelligent" digital libraries to support complex geospatial retrieval and analysis.

In relation to the various classes of users and uses, the emergence of digital libraries with geospatially referenced, multimedia content has created special challenges and opportunities. A particularly challenging research issue relating to geospatial collections is to develop technologies that support *geographic information retrieval* (GIR) in concept-based and fuzzy terms [27]. GIR is particularly complex owing to the diversity of the information media and the fuzziness of geospatial queries. The two primary classes of geospatial queries, "What's there?" and "Where's that?" involve the description of geographic locations ("Where") using either precise terms (e.g., coordinates) or fuzzy terms, such as place names or features (e.g., river, Santa Barbara county).

Traditional information retrieval techniques are clearly inadequate since, for example, classical database management systems feature deterministic retrieval of complete items using structured query languages and exact matching methods [27]. In particular, such systems cannot address the various problems associated with text-based geo-referencing, such as the lack of uniqueness, spatial boundary changes, name changes, spatial and naming variation, spelling variation, and neologisms [49]. In addition, describing geographic phenomena ("What") using subject terms is a classical difficulty in information retrieval which suffers from the vocabulary difference problem [4] [26].

Since it is clearly important that retrieval support probabilistic searches and perform partial matches of relevant multimedia documents using fuzzy, natural language queries, various researchers have recently begun to address the system design issues relating to multimedia, concept-based GIR. Among various ongoing projects, we mention in particular

- The GIPSY Project [49], which developed an automatic geo-referencing system. The system employs the words and phrases of textual documents containing geographic place names and geographic feature references to extract probabilistic functions that encode elementary spatial reasoning and approximate coordinates of the location being referenced in the text [49] [27].
- The Alexandria Digital Library (ADL) project [6] [44], which is one of six projects funded by the NSF/DARPA/NASA-supported Digital Library Initiative (DLI), is developing scalable technologies to assist users in accessing and analyzing multimedia collections for which geospatial access is of central importance. In particular, the project has developed texture extraction and image segmentation techniques [29] for indexing aerial photos with the goal of creating a conceptual, iconic thesaurus that can assist users in image-based, concept-based browsing [29].
- The University of Illinois' DLI project [8], which encompasses all fields of engineering, is investigating the feasibility of creating textual geographic thesauri (called concept spaces) on the basis of textual analysis and clustering techniques. The resulting geographic concept spaces may be used to suggest subject descriptors and place names for GIR.

The systems developed by these projects, although experimental and preliminary in nature, are paving the way for the design and implementation of "intelligent", user-friendly geographic retrieval and analysis systems that support multimedia collections in terms of concept-based access.

In this paper, we discuss preliminary results from a collaborative research effort in the area of concept-based geospatial access between the Alexandria and Illinois DLI projects. In particular, we discuss the application of artificial intelligence (AI) based approaches that have led to our formulation of the concept of a *Geospatial Knowledge Representation System* (GKRS). The goal of the GKRS architecture is to provide a major set of services that support the *meta-information environment* of a digital library by integrating various multimedia knowledge sources in order to support concept-based GIR. Based on semantic network and neural network representations, GKRS loosely couples different knowledge sources and adopts spreading activation algorithms for concept-based reasoning.

The paper is structured as follows. We first define the meta-information environment of a digital library in terms of a set of services that provide users with appropriate access to the information in the collections of the library. We then define a GKRS as a core component of the meta-information environment of a digital library

[6] The Alexandria Project is centered at the University of California at Santa Barbara

supporting geospatial information, and describe the various components of an architecture for a GKRS. Following this, we describe a GKRS testbed that is currently under construction as part of the Alexandria-Illinois DLI projects, first in terms of a set of component *knowledge representation systems* (KRS) and second in terms of a set of procedures that operate on KRS. Finally, we discuss some preliminary results arising from our research that relate to concept spaces, category maps, and visual thesauri. We also indicate future directions for our research.

2 Digital libraries and geographic knowledge representation systems

We now discuss the meta-information environment of a digital library, which is a component enabling users to access and manipulate geospatially referenced information. We first discuss such an environment in general terms and then focus on a specific example of a meta-information environment that we have designed and investigated as part of the Alexandria and Illinois DLI project research agendas.

2.1 Digital libraries as heterogeneous sets of distributed services

We may view a digital library as a heterogeneous set of distributed services that a user may access and integrate in various ways in order to locate information of interest from the set of *information bearing objects* (IBOs) in the collections of the library. Such services may be implemented, for example, within a distributed object framework which may be based upon standards such as CORBA [37].

A particularly important set of services relates to the meta-information environment of a digital library [45], which we define to be

> *the set of* **all** *information services accessible to users of the library, together with all available means for coordinating the use of these services, that enable users to access, evaluate, and use any information that may be extracted from the total information resources of the library.*

Figure 1 illustrates a high-level design for a meta-information environment for DLs. The design is intended to be extensible and views the meta-information environment of a DL as a set of high-level services that provide the essential functionality of a library.

In order to characterize further the manner in which the sets of services shown in Fig. 1 provide support for user access to information, it is useful to introduce the concept of a *knowledge representation system* (KRS). A KRS may be defined as a system for representing and reasoning about the knowledge in some domain of discourse, and is generally comprised of: (1) an underlying knowledge representation language (KRL), whose expressions are intended to represent knowledge about some domain of discourse; (2) a semantics for the language; (3) a set of reasoning rules for drawing inferences from expressions in the language; and, most importantly, (4) a body of knowledge about the domain of discourse, expressed in terms of the KRL.

We argue that an important component of the functionality of the six sets of meta-information services in digital libraries is provided by a diverse set of KRSs. In

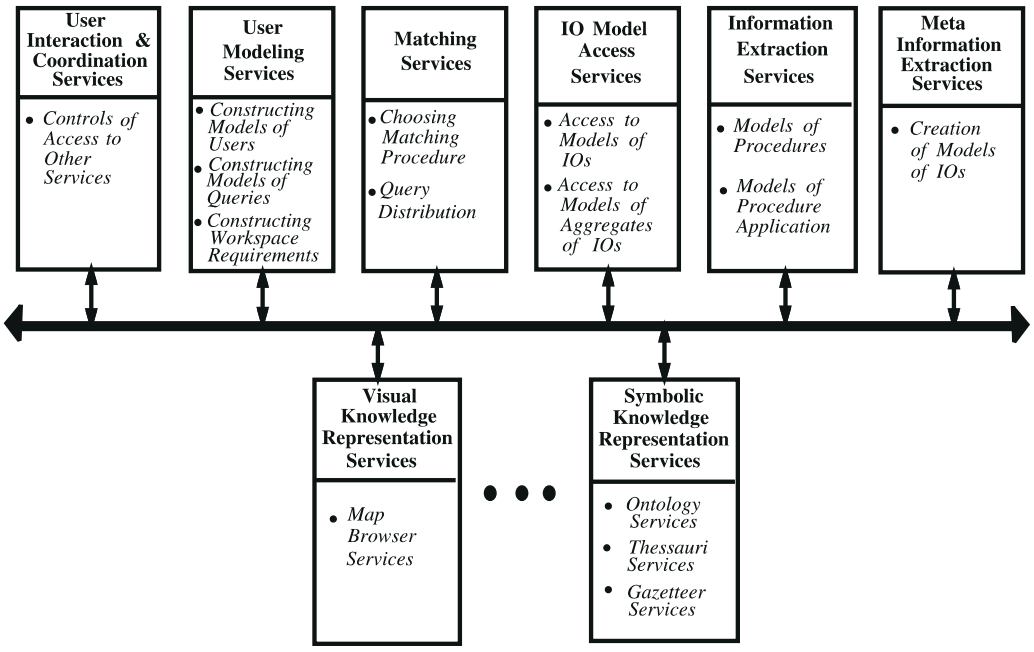

Fig. 1. A high-level design for the meta-information environment of a DL

particular, we may view the KRSs of a library as providing a diverse set of services that are of particular value in modeling both the IBOs of the library's collection and the user queries that are applied to the IBOs. For example, KR are of particular significance in supporting the modeling of IBOs in terms of their *content*, since, in principle, the content of library materials may refer to *any* representable aspect of our knowledge. This knowledge-based system view of digital libraries appears to be consistent with the experiences of AI and expert systems researchers [17] [20] [21].

We briefly summarize the main clusters of services represented in Fig. 1.

- A first set of services provides support for the coordination of interactions between the user and the meta-information environment.
- A second set of services is employed in modeling the user, the queries presented by the user, and the workspace requirements of the user. The services are intended to represent some of the functionality of a librarian in relation to similar services in traditional libraries. In modeling a user, for example, there may be a service for determining a user's area of expertise and, on this basis, choosing a KRL in which the user's query may be expressed.
- A third set of services supports storage of, and access to, models of the IBOs available in the collections of the library, as well as other corresponding libraries. In general, the models of IBOs may be interpreted in terms of various *relations between some symbolic or iconic representation of the IBO itself and representations of the characteristics of the IBO*. In particular, the representations of the IBO itself may be provided in terms of *access paths* and the representations of the characteristics of the IBO expressed in some KRL. As in the catalog of a traditional library, these services support direct access to IBOs on the basis of the characteristics of the IBOs. These services may be generalized to provide models of *aggregates* of IBOs and even of whole libraries. Such aggregate representations are of value for realizing the efficiencies associated with hierarchical search [15].
- A fourth set of services supports choosing and applying appropriate matching procedures between models of user queries and models of IBOs. The goal of these services is to return appropriate IBOs to the user. The matching services may involve, for example, query translation (since the models of IBOs may be represented in languages that are different from the languages in which the user's query is represented) or branching by search type (such as hierarchical search or iterative search.) Matching may employ different matching services depending on the nature of the query using, for example, standard information retrieval procedures for text information or a browsing-type search for images based on a relevance feedback algorithm. The process may proceed iteratively and hierarchically, by returning to the user information that allows the user to have input into the search process. For example, the system may present generalized information about the content of various subcollections in order to obtain information on the most appropriate subcollections to search. There may also be services that support the distribution of queries that cannot be satisfied to other library services.
- A fifth set of services that supports access to, and application of, procedures that may be applied to retrieved IBOs in order to extract useful information. Such services may, include the modeling of procedures and the modeling of the results of applying procedures to IBOs.
- A sixth set of services that provides support for librarians in creating models of IBOs. These services may also support, the automated creation of aggregate representations of collections of IBOs and of whole libraries.

This list of sets of services is not intended to be exhaustive.

2.2 Complex geospatial queries

The importance of supporting the analysis and interpretation of geographically referenced materials, as suggested by Gluck [19], for example, calls for the development of meta-information environments supporting access to geospatially referenced material. Such environments are required for supporting many classes of complex, multimedia, geospatial queries, which typically involve complex concepts that must be represented in the KRS of the meta-information environment. Typical examples of such queries, with their associated concepts shown in italics, include:

- "Find me information, in the form of texts, maps, or images, about *orchards* along the *Santa Cruz River* in *Arizona*."
- "What are the *major valleys* along the *California* and *Arizona borders* and where are their *highest points*?"
- "Could you find me an *up-scale residential area* in *Santa Barbara County* which was *not flooded* in *1994*?"
- "I am planning a field trip. Could you find me textual or map information about a *lake* or *creek* that has a lot of *shade* surrounding it and is *close* to *Highway 101* and *Highway 5* in *Ventura County*?"
- "I am planning on moving my operation to *Los Angeles County*. Please find me images of a site which is close to *major highways*, but with *a lot of green*. It should have some existing *parking lots* in the area and be close to *city and federal buildings*. Hopefully, the site will not be too far from *major residential areas* and *schools*."

In order to answer such complex geographic queries, multiple knowledge sources must be consulted, correlated, and analyzed and other advanced image processing and natural language processing techniques must be employed.

We now discuss the development of a meta-information environment that involves various large-scale *geographic* knowledge representation systems (GKRS),

which we define to be KRSs supporting geospatial retrieval and analysis. We believe that such a GKRS architecture defines an important approach to building intelligent and pro-active systems that are capable of assisting users in retrieving, analyzing, and interpreting geospatially indexed, multimedia information on the basis of a large set of domain-specific concepts.

2.3 Geographic knowledge representation systems

Based on previous research in scientific information retrieval and sharing [7] [43], we have developed a GKRS-based meta-information environment that integrates multiple, multimedia geographic knowledge sources. The underlying knowledge representations are mainly based on symbolic semantic networks [46] and probabilistic neural networks [28]. The semantics are provided, for example, by textual subject descriptors and iconic visual images, which represent textual and visual representations of such concepts as parking lots (images), housing developments (images), power stations (text), and gravity dams (text). The representations rely heavily on the use of spreading activation-type reasoning methods [7] [11]. The body of geographic knowledge to be represented includes existing human-created geographic subject headings and thesauri as well as automatically generated textual and visual concept spaces. The subject headings and thesauri are generated with the use of selected inductive machine learning and neural network techniques, while the concept spaces are probabilistic networks of concepts represented by textual subject descriptors and visual images.

We show in Fig. 2 a schematic diagram of the GKRS architecture. The figure illustrates a top-down, ontological view of knowledge structure development as well as a bottom-up, inductive approach to extracting knowledge from textual and image databases. In the diagram, knowledge sources or structures are depicted by database icons; processes and techniques are represented by rectangular boxes; resulting bodies of integrated geographic knowledge are shown in ovals as loosely coupled networks of concepts, with alphabetic symbols representing textual concepts and square icons representing visual images.

We now discuss the various components of this architecture and their role in the creation of the GKRS component of the meta-information environment of a digital library.

2.3.1 Existing knowledge sources: geographic subject headings and thesauri

A variety of knowledge sources that require little processing in order to convert them into the form of a KRS currently exist in the domain of geospatially referenced information. For example, various thesauri and collections of geographic subject headings, represented on the right side of Fig. 2, have been developed manually by information specialists and domain experts for representing bodies of general geographic knowledge. For example, the GeoRef thesaurus contains about 27000 geographic terms and place names.

Such existing knowledge sources resemble the symbolic semantic network representation developed by AI researchers [32]. Conceptual subject descriptors of place names and geographic phenomena can be represented as nodes in a network and their symbolic relationships (often in terms of hierarchical *broader term*, *narrower term*, and *related term* relationships) are represented as directed links between nodes. Specialized geographic knowledge sources may contain representations of geographic relationships such as "is near" or "is contained in." Relatively simple knowledge networks are capable of capturing a significant body of geographic knowledge in terms of their nodes and links and they are especially suited for fuzzy geographic access involving imprecise concepts such as "Where" and "What."

Human-generated subject headings and thesauri, however, suffer from a lack of fine-grained subject coverage, especially for new and emerging geographic concepts. Like expert system development, knowledge elicitation and update is often a slow and painstaking process.

2.3.2 Specialized geo-referenced materials

There also exist various specialized geo-referenced collections that may be employed in creating specialized geographic knowledge for KRS using domain-specific data analysis techniques, as shown on the right side of Fig. 2. For example, the US Geological Survey's (USGS) Geographic Names Information System (GNIS) is a large gazetteer containing 1.8 million place names, organized hierarchically into 15 geographic feature classes, with a point location for each feature; Digital Elevation Models (DEMs) contain coordinate and elevation information of various geographic regions; while Advanced Very High Resolution Radiometer (AVHRR) data contains information about vegetation cover, land surface temperature, and soil temperature. DEM data may be used to answer queries relating to the geographic elevation of such features as valleys, rivers, and peaks while AVHRR data may be used in answering queries about the vegetation cover or temperature of a particular area.

2.3.3 Textual and image databases

A variety of existing and "lower-level" information sources may be employed in creating KRSs with the application of inductive knowledge discovery techniques, as shown at the bottom of Fig 2. Over the past several years, various databases relating, for example, to drug side effects, retail shopping patterns, tax and welfare frauds, and frequent flyer patterns have been employed in

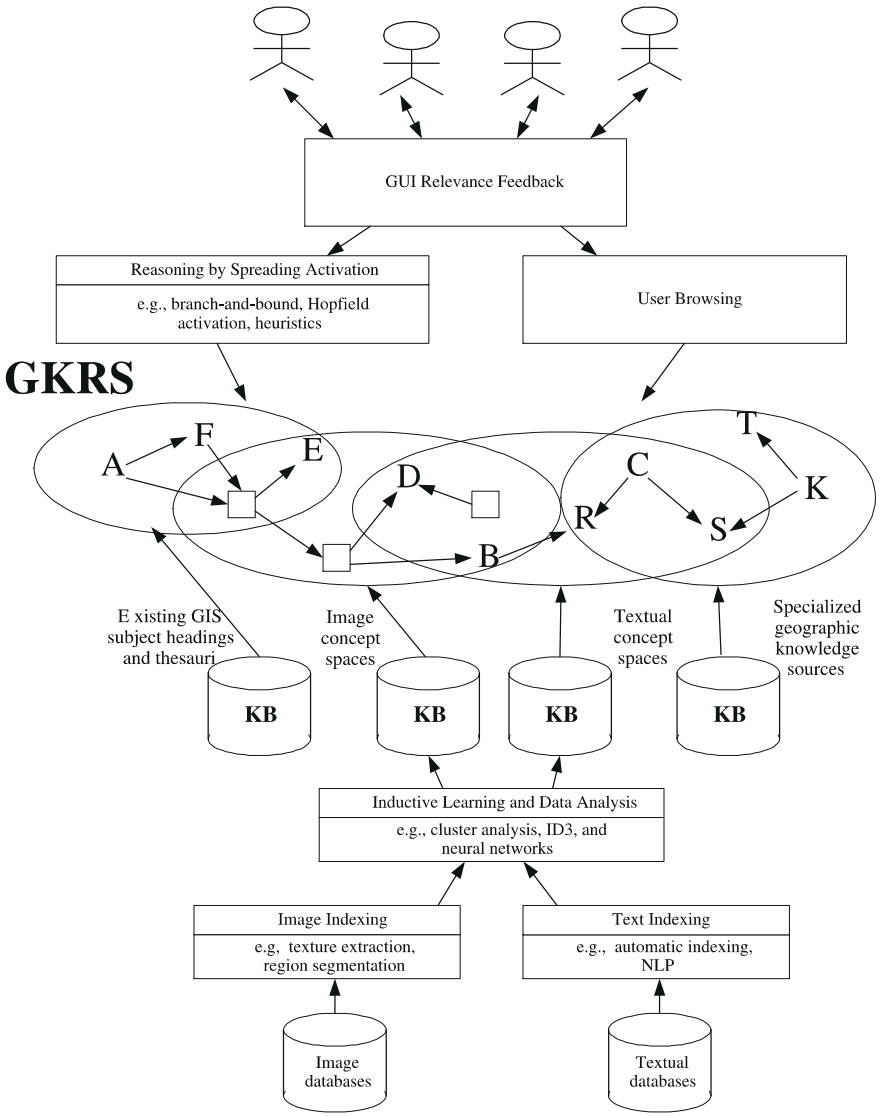

Fig. 2. The geographic knowledge representation system (GKRS) architecture

identifying collection-specific knowledge [18]. While the size of many real-life databases prohibits their analysis by human beings, the availability of unused computing cycles within many organizations has led to the use of computers for knowledge discovery [38] [18]. Massively parallel computers, and even supercomputers, have also been employed in the analysis of large business or scientific databases [43].

In the realm of geospatially-referenced information, many of the existing textual geographic databases can be used as knowledge sources for creating KRSs. Good examples of such databases are the Compendex database (covering all scientific and engineering fields including geography), the GeoRef database, and Petroleum Abstracts. In addition to these conventional textual sources, digitized aerial photos and satellite images have become increasingly important for supporting geospatial queries. With scalable image extraction and segmentation techniques, such images can be automatically segmented, extracted, and indexed, a process similar to textual indexing. Many image processing techniques are now sufficiently mature and scalable to support indexing and analysis of large-scale geo-referenced aerial photos and images [29] [44].

2.3.4 Textual and image indexing

For textual knowledge sources, domain-independent automatic indexing techniques are often adopted to extract potentially meaningful subject descriptors [42]. More recently, syntactic natural language parsing techniques have been adopted for more fine-grained textual analysis [31]. For images, texture extraction, region segmentation, and color and shape detection are techniques frequently used to segment and index image collections [29] [47] [48]. Texture extraction based on edge and orientation and region segmentation based on edge

flow and texture appear to be promising techniques for indexing aerial photos and satellite images.

2.3.5 Inductive learning and data analysis

Various inductive learning and data analysis techniques have been developed over the past few decades by statisticians, information scientists, and AI researchers. Statistical algorithms [38] are typically applied to quantitative data in order to (1) cluster descriptors in terms of a relatively small set of characteristics, as in the case of factor analysis, principal components analysis [34], and cluster analysis [16]; (2) test hypotheses concerning differences among populations, as in the case of t-tests and analysis of variance (ANOVA) [33]; (3) perform trend analysis, as in the case of time series analysis [34] [35]; and to construct correlations among sets of variables, as in the case of computing correlation coefficients and multiple regression analysis [33].

Recently, "classical" (or symbolic) AI learning algorithms, such as ID3 [40] and AQ [30], as well as a new generation of neural net learning algorithms that include feedforward networks [41] and Kohonen self-organizing maps [25], have provided new perspectives on knowledge discovery. These techniques allow effective analysis of both qualitative and quantitative data. Unlike the statistical approaches, which are typically based on some underlying model involving strong assumptions and/or stringent conditions, many AI-based techniques are more flexible, more powerful, and easier to use as well as producing output that is more meaningful to users (see [2] [14] [18] [23] for overviews of AI-based learning techniques).

Several recent large-scale digital library experiments have adopted cluster analysis for creating domain-specific concept spaces consisting of networks of terms and their weighted relationships [8] [9]. The Kohonen self-organizing map algorithm has been extended to create textual graphical category maps [10] and visual geographic thesauri [29]. In our proposed GKRS architecture, we suggest the use of concept spaces and category map techniques in order to create compatible network-based knowledge representation systems. Textual and visual concept spaces and category maps can be extracted from the underlying textual and image databases respectively to create multimedia geographic knowledge structures.

2.3.6 Geographic knowledge representation systems

Existing geographic subject headings and thesauri, together with automatically generated textual and visual concept spaces and category maps, may be viewed in terms of loosely coupled network-based KRSs that can be selected, mixed, and matched in response to fuzzy, multimedia geographic queries. For example, by selecting the GeoRef thesaurus and a visual concept space generated from selected aerial photos as the search aid, a user can potentially query a geographic database using a combination of subject descriptors (e.g., dams and rivers) and image icons (e.g., orchard pattern). The integrated Geographic Knowledge Representation System can thus be thought of as a knowledgeable digital map librarian, ready to assist in the analysis and interpretation of complex geo-referenced materials. We suggest that the potential usefulness of such a large body of geographic knowledge cannot be over stated in this era of networked information systems and large-scale digital collections.

2.3.7 Reasoning by Spreading Activation

As shown in the upper portion of Fig. 2, users are free to use the GKRS in terms of either a user-controlled browsing mode or a system-supported spreading activation mode. If users employ hypertext browsing to follow manually the links of network-based knowledge structures as subject headings, thesauri, textual concept spaces and category maps, and visual concept spaces and category maps, they can only traverse a small portion of a large knowledge space. In general, user-controlled, relevance-feedback search processes have been found to be productive, but at times cumbersome [7] [42].

In order to manage and utilize the potentially rich and complex nodes and links in GKRS, system-aided reasoning methods may be needed to intelligently suggest a small portion of the relevant knowledge to the user. For semantic network and neural network representations, spreading activation, which is a well-known memory recall process [1], appears to be well suited for the reasoning task [32]. Several information systems researchers have developed symbolic, heuristic-based spreading activation methods for information retrieval systems and digital libraries [3] [11]. More recently, we have adopted the serial branch-and-bound search algorithm and the parallel Hopfield network activation method to traverse multiple knowledge networks [7]. Hopfield network activation, in particular, appears to be a robust and scalable technique for term suggestion and vocabulary switching across different knowledge domains.

2.3.8 GUI relevance feedback

Due to the multimedia nature of GKRS, a graphical user interface (GUI) is needed for viewing and activating multiple data types and the underlying geographic IBOs such as maps and aerial photos. The Internet, Web protocols, and Java-like languages now make it possible to implement adequate GUIs for GKRS. In particular, the spreading activation reasoning process of a GKRS can be "visualized" using advanced graphical and visualization techniques. We believe that making the system's reasoning process transparent to users is essential for the success of GIR. Such a user "buy-in

factor'' has been recognized in expert systems research [21].

Despite such system-aided reasoning and visualization, however, users still need to control the relevance feedback process by selecting relevant and interesting items and determining directions for search. We believe that such a user-system collaborative process is crucial to the success of fuzzy, concept-based geographic information retrieval.

3. A GKRS testbed

As part of a joint project between the Illinois DLI Interspace project [43] and UCSB DLI Alexandria project [44], we are in the process of creating a large testbed for implementing the proposed GKRS. The testbed includes substantial coverage of textual and image databases and existing knowledge structures. For specialized geographic collections such as DEM data, AVHRR data, aerial photos, and satellite images, our initial focus is on having a complete coverage of Southern California.

Our current GKRS testbed involves approximately 100 GB of storage on the National Center for Supercomputing Applications (NCSA) Unitree file storage system. Several experiments are underway to generate individual textual and visual thesauri, concept spaces, category maps, and specialized knowledge sources. We are also in the process of developing different spreading activation reasoning techniques for integrating these multiple, multimedia geographic knowledge sources.

3.1 Textual databases and knowledge sources

We now describe the various components shown in Fig. 1 that have been implemented in our testbed.

3.1.1 Compendex geographic category, database, concept space, and category map

The Compendex database consists of 3 major categories and 43 sub-categories which are related to geoscience. Such categories are presented as a hierarchy. In addition, through the Illinois DLI project, about 50,000 Compendex geoscience-related abstracts (70 MB, 1991–1995) have been extracted. The title, abstract, and author fields of the Compendex records are used to generate Compendex-specific concept spaces and category maps. In the following section, we illustrate the resulting sample Compendex concept space and category map.

3.1.2 GeoRef thesaurus, database, concept space, and category map

The GeoRef thesaurus was recently incorporated into our testbed. The thesaurus contains more than 27,000 terms, with several standard symbolic relationships, such as *broader term*, *narrower term*, *related term*, and *use for*. It also includes usage notes, dates of addition, and coordinates for selected place names.

In addition, we have also obtained two large collections of GeoRef records: 300,000 recent GeoRef records (most without abstracts, 1990–1995, 300 MB), and 20,000 GeoRef records with abstracts (1981–1995, 70 MB) from the American Geological Institute. Like the Compendex collection, the GeoRef-specific concept spaces and category maps will be generated using the title, abstract, and author fields of the records. Because many GeoRef records contain both place names and coordinates, we will also correlate this information.

3.1.3 Petroleum abstracts thesaurus, database, concept space, and category map

While GeoRef covers many of the concepts of geography and geology, Petroleum Abstracts (PA) covers concepts for petroleum engineering and petroleum exploration. The abstracts overlap with GeoRef only in relation to earth science concepts. We have obtained the 1985–1995 collection of the Petroleum Abstracts (about 800,000 abstracts, 600 MB). Titles, abstracts, and author names are being used to create the PA-specific concept spaces and category maps. The PA thesaurus, similar to the GeoRef thesaurus in structure, has been made available through the University of Tulsa.

3.1.4 GNIS gazetteer

Gazetteers are useful knowledge sources for identifying relationships between precise coordinates and fuzzy place names. Since in our current experiment we are developing an integrated GKRS to support queries specific to Southern California, we have extracted only the Southern Californian portion of the USGS GNIS gazetteer, consisting of about 56,000 place names, 7 feature classes (e.g., hydrology), and 60 feature types (e.g., canal), organized in a hierarchy similar to that of a standard thesaurus.

3.2 Image databases

We have been unable to identify image knowledge sources, such as human-created visual thesauri, for georeferenced collections. Through the Map and Imagery Laboratory (MIL) at UCSB, however, we have been able to develop a substantial collection of aerial photos and satellite images covering Southern California.

3.2.1 Aerial photos

Frames from three historically important and popular flights have been scanned. These flights involve 761, 1216,

and 355 frames, respectively, and cover Santa Barbara, Ventura, and Los Angeles counties. Each black-and-white frame occupies approximately 30 MB of memory, and together they constitute a database of approximately 75 GB. Identifiable visual patterns on such flights include housing developments, parking lots, schools, parks, vegetation, highways, airports, factories, etc. Based on the texture extraction and region segmentation techniques developed earlier in the Alexandria project [29], we are in the process of automatically indexing these images to create visual concept spaces and category maps for Southern California. We plan to index and analyze these items in order to reveal meaningful, geographic relationships between man-made and natural artifacts, such as parks next to residential areas, vegetation next to rivers, and factories next to highways.

3.2.2 Satellite images

Unlike aerial photos, satellite images reveal higher-level, coarse-grained geographic features such as mountains, populated areas, and lakes. We believe that the image indexing techniques developed for aerial photos can be extended to satellite images. With the help of UCSB's MIL we have obtained a collection of nearly 200 SPOT satellite images of California for our experiment.

It is important to note that each aerial photo and satellite image also contains coordinate information which may be correlated with other textual knowledge sources. For example, the GNIS gazetteer and the GeoRef thesaurus contain information about place names and coordinates that may be used to label selected image regions automatically.

3.3 Specialized geographic knowledge sources

While the above textual and image knowledge sources have many common characteristics, other specialized geographic collections need to be processed in a more ad hoc fashion. Most of the techniques considered for such collections are still experimental.

3.3.1 Digital elevation models

DEMs consist of sampled arrays of elevations for ground positions that are usually, but not always, spaced at regular intervals. A California DEM data set at 1:250,000 scale was extracted from the USGS Web site. This data set consists of about 93 MB (70 files) of coordinate and elevation information covering all of California. We are experimenting with several heuristics-based computational models to obtain conceptual descriptors (e.g., valley, canyon, hill, mountain, etc.) relating to different geographic regions. This information is being correlated with other coordinate-specific knowledge sources including the GNIS gazetteer and the GeoRef thesaurus.

3.3.2 Digital line graphs

Also available from the USGS, Digital Line Graphs contain linear map information in digital form and include information on planimetric-based categories, such as transportation, hydrography, and boundaries. A California DLG data set at a scale of 1:100,000 was extracted from the USGS Web site. This data set consists of about 159 MB (21 files) of coordinate and hydrography and transportation information covering all of California. As in the case of the DEMs, we are experimenting with several heuristics-based computational models to match the hydrography and transportation information with other knowledge sources such as the GNIS gazetteer, aerial photos, and satellite images.

3.3.3 Advanced Very High-Resolution Radiometer data

Lastly, a testbed of AVHRR data, collected from NOAA's Polar Orbiting Environmental Satellites (POES), was extracted to include vegetation cover, land surface temperature, and soil temperature information for California. For experimental purposes, our testbed contains only monthly (8km, 5 MB) and weekly (1km,

The combined weight of term j in document i, d_{ij} is computed, based on the product of "term frequency" and "inverse document frequency" as follows:

$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j} \times w_j\right)$$

where N represents the total number of documents in the collection, tf_{ij} represents the number of occurrences of term j on document i, w_j represents the number of words in descriptor T_j, and df_j, represents the number of documents in a collection of n documents in which term j occurs. Multiple-word terms are assigned heavier weights than single-word terms because multiple-word terms usually convey more precise semantic meaning than single-word terms.

Fig. 3. Frequency computations

600 MB) information collected in 1995. We are also experimenting with several heuristics-based computational models to correlate this vegetation and temperature information with conceptual descriptors (e.g., warm, hot, green, desert, etc.) and place names (through coordinates).

4 Algorithms: cluster analysis and self-organizing maps

In this section we summarize the nature and use of the algorithms employed in our current prototyping efforts.

4.1 Cluster analysis

The specific steps and algorithms that were adopted to create our textual concept space include: *automatic indexing*, *co-occurrence analysis*, and *associative retrieval*. A brief overview of these techniques, in the context of our experiment, is presented below. For further details of the algorithms, see [8].

4.1.1 Automatic indexing

The purpose of this step is to automatically identify the content of each textual document. Based on a revised automatic indexing technique [42], subject descriptors on each document are identified, and the number of times that each descriptor appears in the entire collection of documents is computed. A "stop-word" list is used to remove non-semantic bearing words such as "the", "a", "on", and "in", after which a stemming algorithm is applied to identify the word stem for the remaining words. In our processes, the stop-word list was also applied to all of the stemmed words.

4.1.2 Co-occurrence analysis

The importance of each descriptor or term in representing the content of the entire document varies. Using term frequency (tf) and inverse document frequency (idf), the cluster analysis step assigns weights to each term in a document to represent the term's level of importance. Term frequency measures how often a particular term

$$ClusterWeight(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \times WeightingFactor(T_k)$$

$$ClusterWeight(T_k, T_j) = \frac{\sum_{i=1}^{n} d_{ikj}}{\sum_{i=1}^{n} d_{ik}} \times WeightingFactor(T_j)$$

These two equations indicate the similarity weights from term T_j to term T_k (the first equation) and from term T_k to term T_j (the second equation). d_{ij} and d_{ik} are calculated based on the equation in the previous step. d_{ijk} represents the combined weight of both descriptors T_j and T_k in document i. d_{ijk} is similarly defined:

$$d_{ijk} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_j\right)$$

In order to *penalize* general terms (terms which appeared in many places) in the co-occurrence analysis, the following weighting schemes, which are similar to the *inverse document frequency* function, were developed:

$$WeightingFactor(T_k) = \frac{\log \frac{N}{df_k}}{\log N}$$

$$WeightingFactor(T_j) = \frac{\log \frac{N}{df_j}}{\log N}$$

Imposing this penalty factor, allows the thesaurus to make more precise and specific suggestions.

Fig. 4. Cluster analysis computations

occurs in the entire collection. Inverse document frequency indicates the specificity of the term and allows terms to acquire different strengths or levels of importance based on their specificity. A term can be a one-, two-, or three-word phrase. We describe in Fig. 3 how such numbers are calculated.

Cluster analysis is then used to convert these raw data of indices and weights into a matrix showing the similarity and dissimilarity of the terms using a distance function. The distance function used in this step is based on the asymmetric "cluster function" developed by Chen and Lynch [5]. We show in Fig. 4 a more detailed de-

The Hopfield net algorithm relies on an activating and iterative process, where

$$\mu_j(t+1) = f_s[\sum_{i=0}^{n-1} t_{ij}\mu_i(t)],\ 0 \le j \le n-1$$

$\mu_j(t+1)$ is the activation value of neuron (term) j at iteration $t+1$, t_{ij} is the co-occurrence weight from neuron i to neuron j, and f_s is the continuous SIGMOID transformation function, which normalizes any given value to a value between 0 and 1 [?] [?]. This formula shows the *parallel relaxation* property of the Hopfield net. (Readers are referred to [?] for algorithmic detail.)

Fig. 5. Hopfield net algorithm

1. **Initialize input nodes, output nodes, and connection weights:**

   Represent each document (or image) as an input vector of N keywords (or image features) and create a two-dimensional map (grid) of M output nodes (e.g., 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 or image in order:**

   Represent each document or image by a vector of N features 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 [?] for the algorithmic detail of neighborhood adjustment.)

5. **Label regions in map:**

   After the network is trained through repeated presentation of all inputs, submit unit input vectors of single terms to the trained network and assign the winning node the name of input feature. Neighboring nodes which contain the same feature then form a concept or topic region. The resulting map thus represents regions of important terms or image patterns (the more important a concept, the larger a region) and the assignment of similar documents or images to each region.

6. **Apply the above steps recursively for large regions:**

   For each map region which contains more than k (e.g., 100) documents or images, conduct a recursive procedure of generating another self-organizing map until each region contains no more than k documents or images.

Fig. 6. The SOM algorithm used in textual category map and visual thesaurus generation

scription of this function. We have shown elsewhere that this asymmetric similarity function represents term associations better than the cosine function. Using the function, a net-like concept space of terms and their weighted relationships can be created.

4.1.3 Associative Retrieval

In previous research, we developed two associative retrieval algorithms, one based on the serial branch-and-bound algorithm and the other based on a parallel Hopfield net algorithm [7] which, in particular, has been shown to be ideal for concept-based information retrieval [6].

Each term in the network-like thesaurus was treated as a neuron and the asymmetric weight between any two terms was taken as the unidirectional, weighted connection between neurons. Using user-supplied terms as input patterns, the Hopfield algorithm activated neighboring (i.e., strongly associated) terms, combined weights from all associated neighbors (by adding collective association strengths), and repeated this process until convergence occurred. We outline this algorithm in Fig. 5.

4.2 The self-organizing map algorithm

Quillian [39] suggested that semantic networks could be used to encode and associate word meanings and that such networks could then be used to visualize or construct mental models of an information space. Neural network algorithms, in particular, appear to be a natural starting point for organizing large amounts of information in a manner consistent with human mental models. After investigating several neural network algorithms

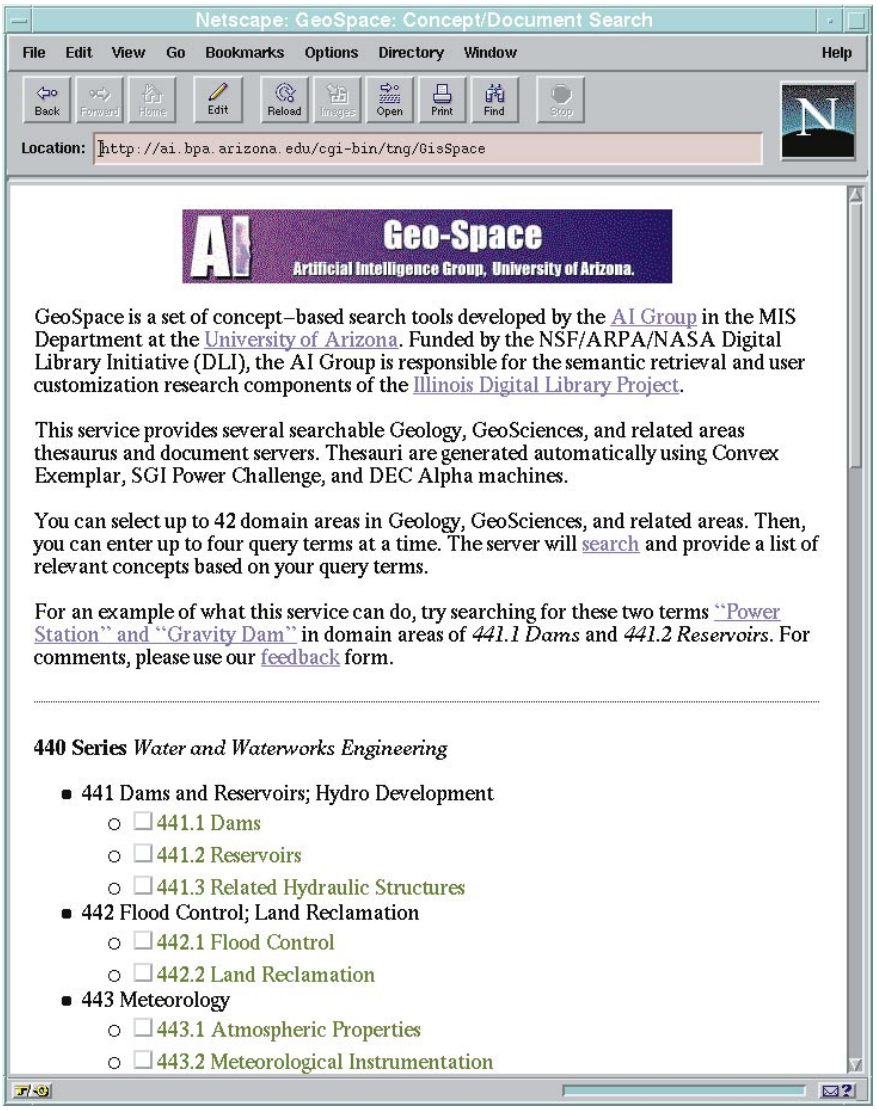

Fig. 7. Geo-Space provides 42 searchable sub-domain concept spaces in geoscience based on the Compendex classification code, e.g., 441.1 Dams. Searchers can select multiple concept spaces to identify other relevant geoscience terms

[28], we have concluded that a variant of the Kohonen self-organizing feature maps (SOM) appears to be the most promising algorithm for organizing large volumes of information. The algorithm can be used to create an intuitive, graphical display of the important concepts contained in textual information [10] [36].

In the algorithm's basic form, continuous-valued vectors (of document keywords or image features) are presented sequentially without specifying the desired output. After a sufficient number of input vectors have been presented, network connection weights specify cluster or vector centers that sample the input space in such a way 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 such that topologically close nodes are sensitive to inputs that are similar. Figure 6 describes how the Kohonen SOM algorithm was modified to create textual category maps and image-based visual thesaurus.

5 Preliminary results: concept space, category map, and visual thesaurus

Three GKRS components based on the Compendex geoscience collections have been developed recently to illustrate the concept space, category map, and spreading activation techniques. A total of approximately 50,000 Compendex records (70 MB) in Water and Waterworks Engineering, Engineering Geology, and Petroleum Engineering were used in this experiment.

Fig. 8. Terms related to "Gravity Dam" and "Power Station" in the Dams and Reservoirs concept spaces are displayed in ranked order, e.g., "Concrete Gravity Dam," "Hydro-electric Power Station," etc. A searcher can select any system-suggested terms for document searches

5.1 The Compendex concept space

In the Illinois DLI project, a scalable concept space generation technique was developed to create textual thesauri automatically. Using a week of dedicated computer time on the HP Convex Exemplar at NCSA, concept spaces were generated for 10,000,000 journal abstracts across 1000 subject areas covering all of engineering and science [8]. Based on subject-independent automatic indexing and cluster analysis techniques, the resulting concept space represents a network of terms (subject descriptors) and their weighted relationships, akin to a single-layered probabilistic neural network [8] [9].

In this experiment, we extracted 3 geoscience domains from the Compendex collection, covering 42 geoscience subdomains, e.g., Dams, Reservoirs, etc. By using each subdomain collection, we generated 42 small geoscience concept spaces (each consisting of about 5000 terms and 200,000 weighted relationships), which were placed as online thesauri on our Web site, as shown in Fig. 7. Users may click on the dialog boxes to select concept spaces for consultation. As shown in Fig. 8, by clicking on the Dams and Reservoirs concept spaces using "Gravity Dam" and "Power Station" as the initial search terms, the system suggests "Concrete Gravity Dam," "Hydro-electric Power Station," "Finite Element Method," etc, as relevant concepts (the probabilistic weights indicate the strengths of relevance).

Using HTML/CGI programs, the GIS space supports iteration of concepts when the user clicks on an individual term or uses a combination of search terms. After deciding on appropriate search terms, a user can then employ the server in locating relevant documents. Our previous research has shown the usefulness of such a system-aided term suggestion tool, especially in im-

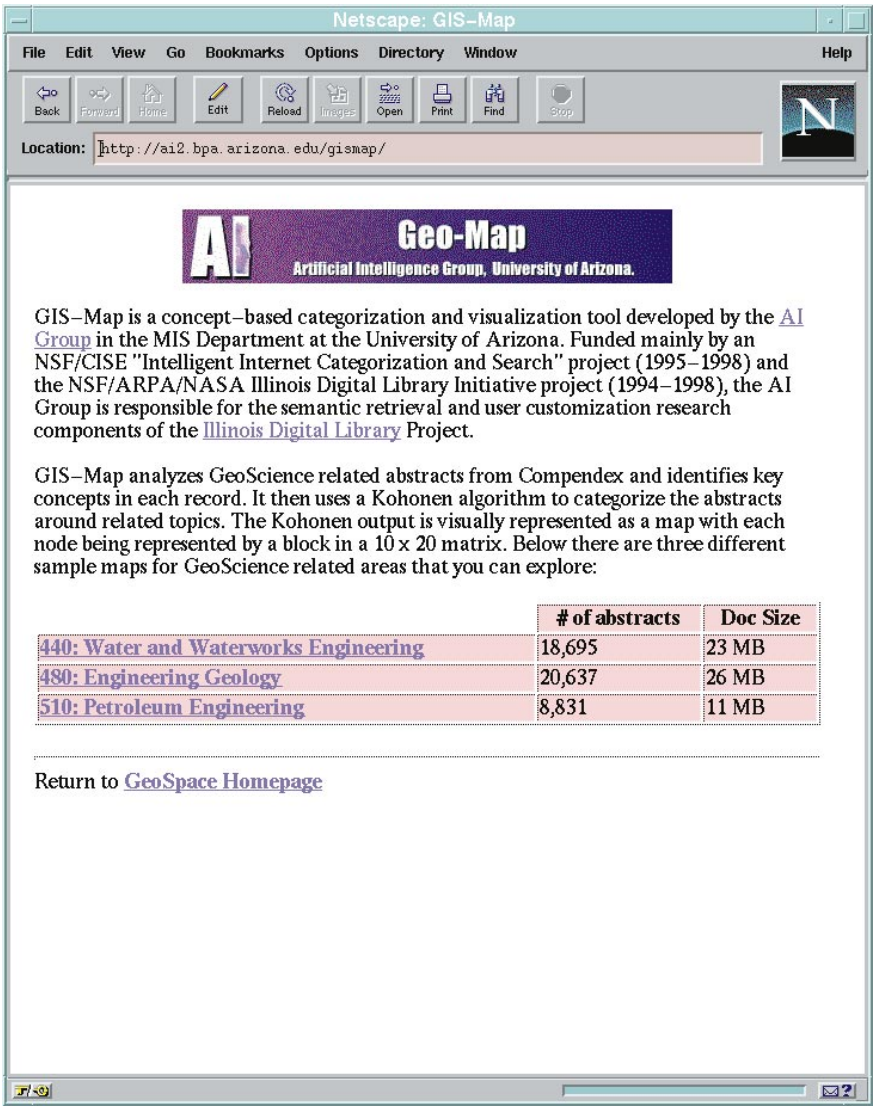

Fig. 9. Three Compendex geoscience collections have been analyzed to create category maps in Water and Waterworks Engineering, Engineering Geology, and Petroleum Engineering, respectively. Searchers can select any of these category maps to perform a concept-based browsing of documents

proving search recall [9]. We are currently designing a large-scale experiment to examine the performance of the GIS space in supporting text-based geographic queries.

5.2 The Compendex category map

By partitioning the Compendex geoscience collection into three main areas of Water and Waterworks Engineering, Engineering Geology, and Petroleum Engineering, we generated three large category maps automatically. The multi-layered self-organizing map (SOM) techniques adopted have been shown to be an efficient and robust method for creating directories of categories graphically [10] [36].

A GIS-map server has been created to support concept-based geospatial browsing (see Fig. 9). By clicking on an individual domain (e.g., Water and Waterworks Engineering), the server brings up a graphical category map as shown in Fig. 10.

Category maps, displayed as jigsaw puzzles, could graphically represent the importance of a subject category in terms of size (i.e., the larger a region, the more important a category, e.g., the "Water" region on the right side of Fig. 10 has 2956 abstracts assigned to it). In addition, categories which are similar conceptually are often clustered in a neighborhood, e.g., "Treatment," "Water Quality," and "Wastewater Treatment" topics in

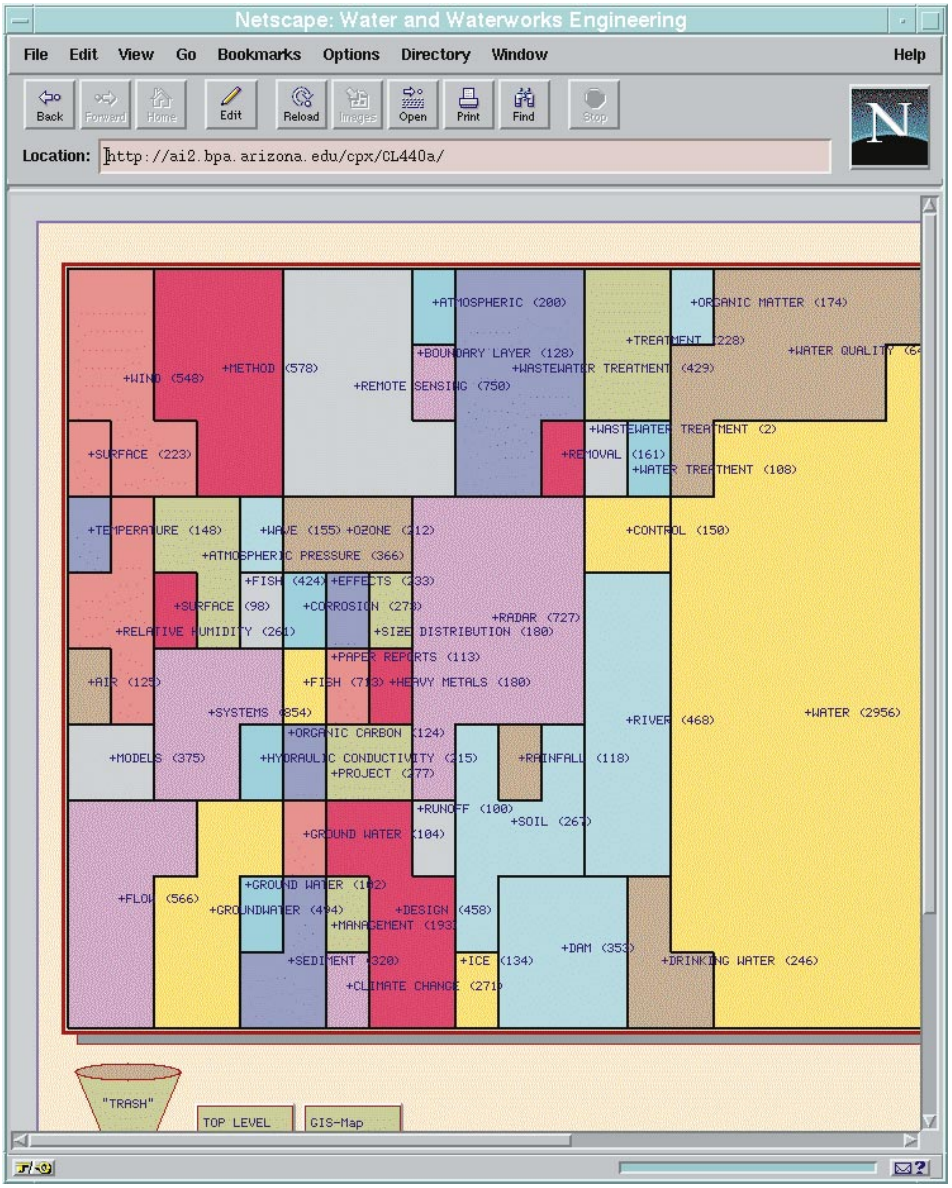

Fig. 10. Top-level SOM category map related to "Water and Waterworks Engineering." Different topics appear to be grouped in different regions, e.g., "Water," "River," and "Drinking Water" at the bottom-right-hand corner. "Wind," "Atmospheric Pressure," and "Temperature" are seen at the top-left-hand corner

the upper-right-hand corner of Fig. 10, and "Wind," Surface," "Atmospheric Pressure," and "Ozone" topics in the upper-left-hand corner. Such graphical display has been shown to be meaningful and user-friendly for browsing [36].

By utilizing a multi-layered category map, a user can drill down a region level by level until he/she sees the desired documents. For example, a user clicks on the "Wind (548)" region in Fig. 10 and retrieves the sub-category map for the Wind-related topics only, as shown in Fig. 11. "Wind Speed," "Wind Field," "Wind Velocity," "Wind Tunnel," etc, are categorized in the same "Wind" region. Clicking on the "Wind Speeds (30)" region of Fig. 11 brings up 30 documents which are related to Wind Speeds, as shown in Fig. 12. In the same experiment planned for the Geo-Space, we will also evaluate the performance and usefulness of the GIS-Map as a browsing tool.

5.3 A SOM-based visual thesaurus for aerial photos

Using an aerial photo testbed made available through the Map and Imagery Lab at UCSB, we have developed a prototype system to assist in image-based visual thesaurus browsing. Approximately 950 $\times$ photos were scanned and each frame was represented as a 5000 $\times$ 5000 pixel image file (approximately 50 MB per image). A 1994 aerial survey flight had covered all of Santa Barbara County. Each image was then partitioned into about 1600 (40×40) small image blocks and indexed using Gabor filters (represented by 60 image features) [29]. Figure 13 shows one partitioned image, processed and displayed by a Java-based prototype system. Using the Gabor filters and a simple Euclidean distance similarity function, we were able to perform an image-based similarity search. After clicking on any small image block, the system displays the best matches, in different

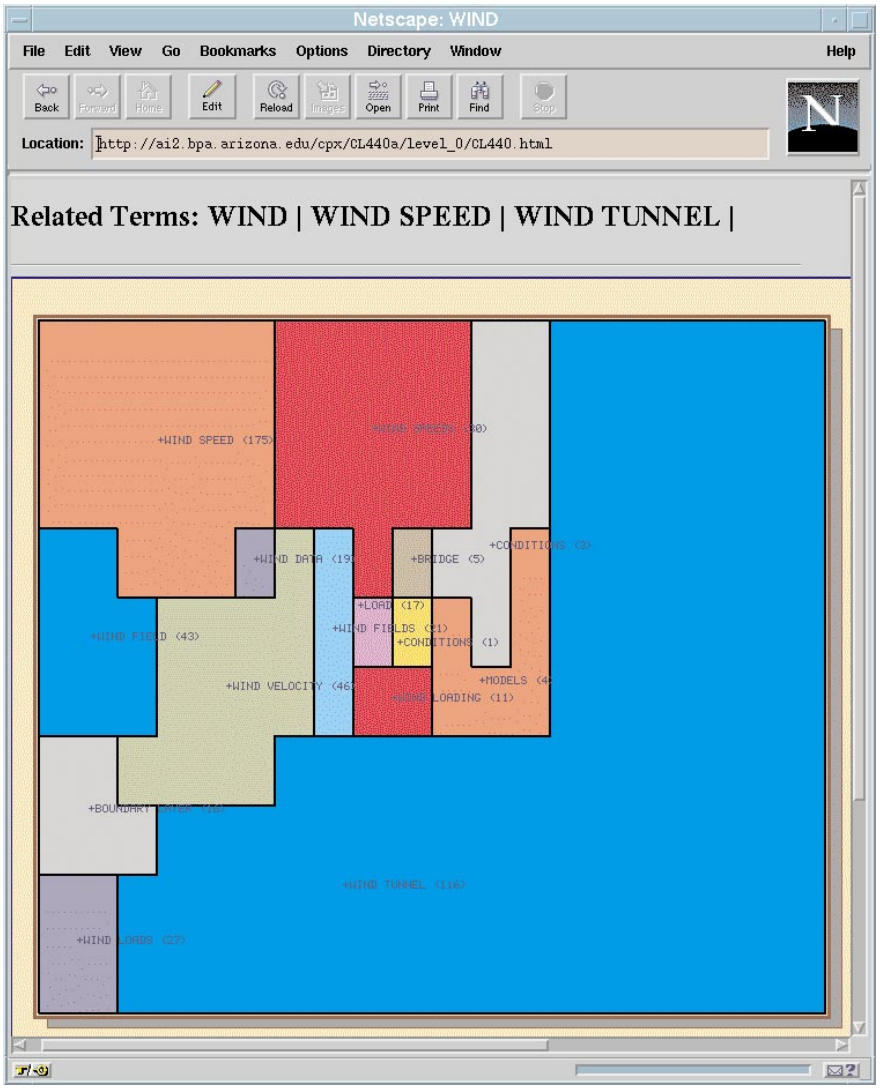

Fig. 11. Second-level SOM category map related to "Wind." Secondary wind-related topics were identified (e.g., "Wind Speeds," "Wind Field," "Wind Tunnel," etc.)

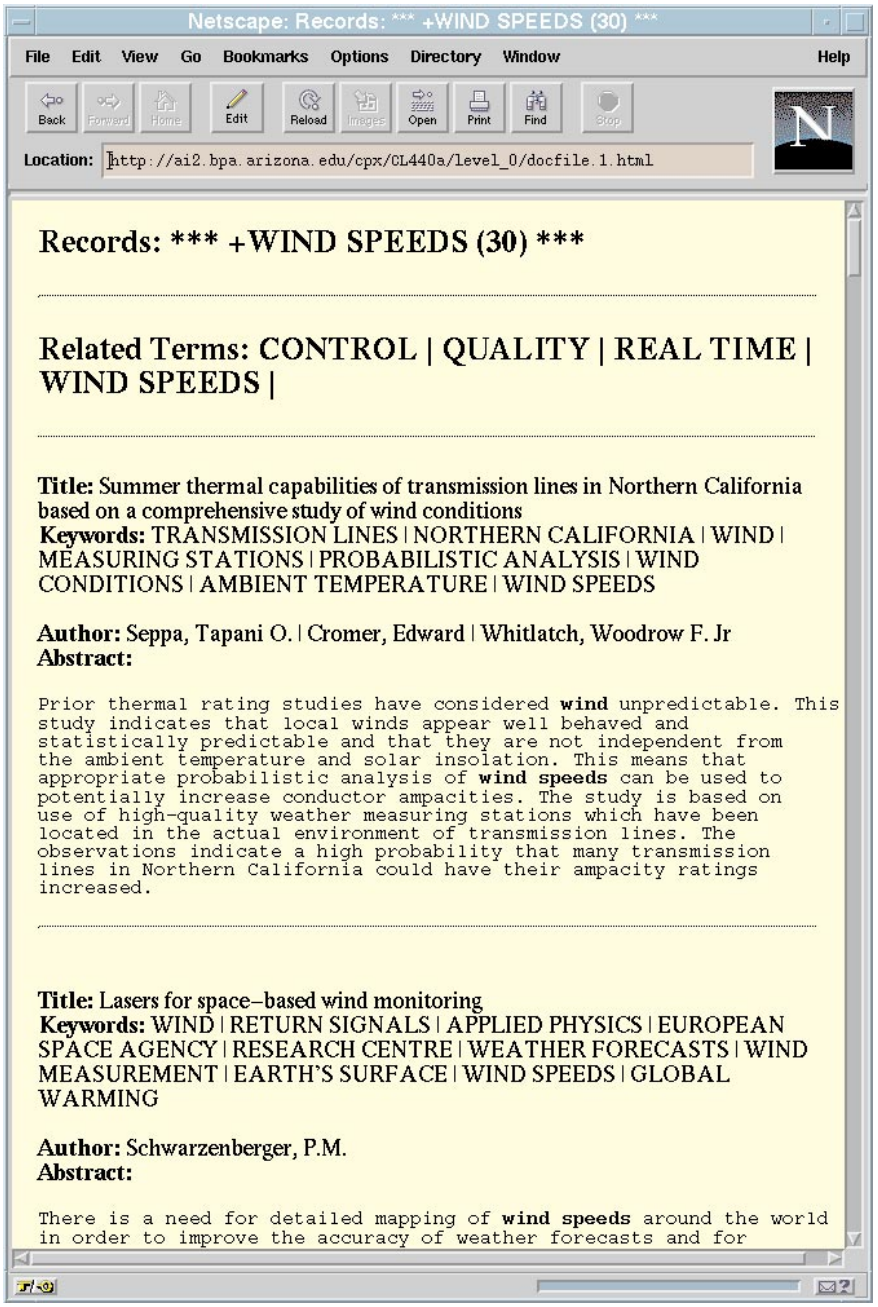

Fig. 12. Documents classified under the "Wind Speeds" region. Documents are ranked and displayed in order and topic terms ("Wind Speeds") in abstracts are highlighted

aerial photos or different parts of the same photo, to selected image patterns in the first image displayed. Figure 13, for example, illustrates image blocks representing residential areas. Our prototype system currently supports image-based similarity search of aerial photos.

A second prototype system was later developed to support image-based visual thesaurus browsing. Using the same 60 image features as the input vector and the SOM algorithm described earlier, we clustered similar image patterns (e.g., residential areas, vegetation, parking lots, farm lands, etc.) in a graphical two-dimensional display. As shown in Fig. 14, similar image patterns were grouped together in different regions of the (semantic map) display, e.g., vegetation patterns at the bottom-right-hand corner. Clicking on each representative image brought out another window that displayed other image patterns that were classified as similar (see Fig. 15). By mapping each image's coordinate to the GNIS gazetteer, we are able to suggest place information matching the geographic location of each image, as shown in Fig. 13.

6 Conclusion and discussion

Developing scalable techniques to support fuzzy, concept-based retrieval from multimedia geo-referenced information (GIR) is a pressing research issue for digital libraries. We are investigating technical and research issues relating to GIR on the basis of an integrated and scalable AI approach.

In this paper, we have proposed a Geospatial Knowledge Representation System (GKRS) architecture which comprises the major component of the meta-information environment of a digital library serving geospatial information. The GKRS integrates multiple knowledge sources concerning such multimedia items as images and text in response to concept-based, geographic queries. Based on semantic network and neural network representations, GKRS loosely couples different knowledge sources and adopts spreading activation algorithms for concept-based knowledge reasoning. Our preliminary experiments using the Compendex collection have yielded promising results concerning geographic concept spaces, graphical category maps, and a Java-based visual thesaurus for airphotos. Our extensive multimedia testbed of textual, image, and specialized geographic collections will allow us to continue to expand on our techniques and gradually evolve toward an intelligent and complete Geographic Knowledge Representation System.

Our ongoing research efforts mainly involve:

- conducting an experiment on the performance and usefulness of the geographic concept spaces, category maps, and vocabulary switching system;
- performing concept space and category map generation for other textual collections, e.g., GeoRef and Petroleum Abstracts;
- experimenting with automatic region segmentation techniques for aerial photos and satellite images;
- developing special-purpose computational models for specialized geographic collections, e.g., DEM, DLG, and AVHRR data sets;
- designing a fully functional Java-based system to integrate all GKRS components.

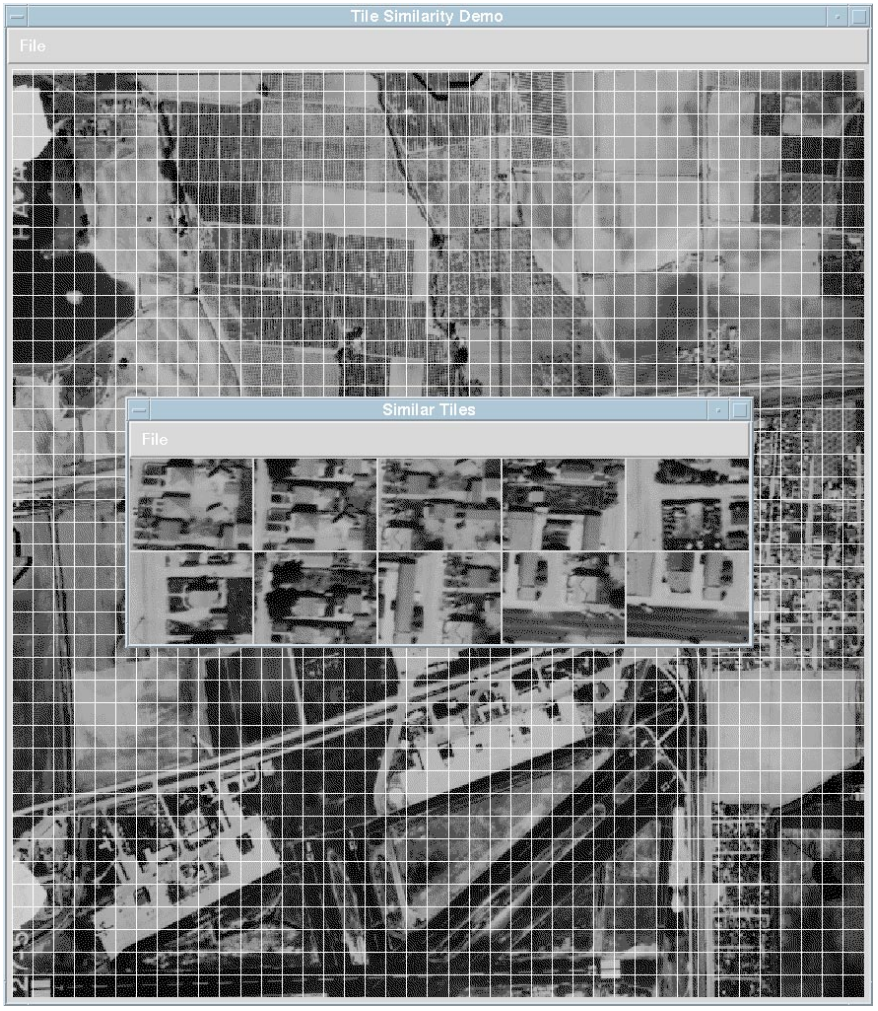

Fig. 13. Each aerial photo is partitioned into 40×40 small image blocks and indexed by Gabor filters. After the user clicks on an image block, the Java-based system pops out a separate window to display the image block (the top-leftmost image in the popped-out window) and its similar blocks (seven similar image blocks displayed in ranked order from left to right and top to bottom)

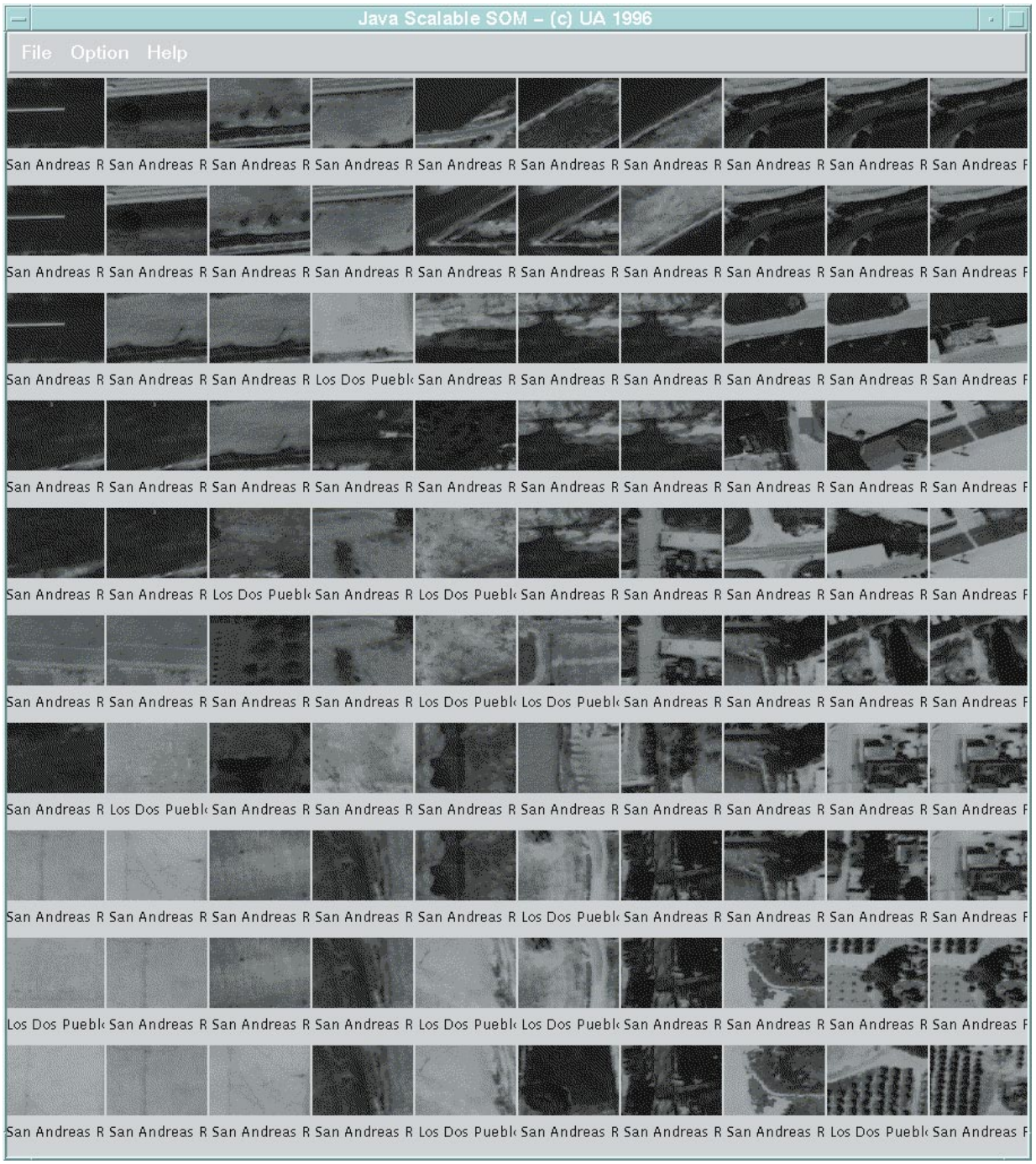

Fig. 14. The SOM algorithm has clustered similar image blocks, e.g., vegetation (bottom-right-hand corner), farm lands (top right-hand corner), etc., and presented them as a "visual thesaurus." Each representative image also has been labeled by the corresponding place name obtained from the coordinate information in the GNIS gazetteer, e.g., "San Andreas Rift." Clicking on each representative image would display all other similar images in different aerial photos

Acknowledgement. We would like to thank Bob Cowen of Engineering Information, Inc., for providing the Compendex collection, SPOT Image Corporation for providing the California SPOT images, John Mulvihill of the American Geological Institute for providing the GeoRef collection, and Rafael Ubico of the University of Tulsa for providing the Petroleum Abstracts collection. Many thanks to Bruce Schatz, Kevin Powell, and Charles Herring of the University of Illinois and Larry Carver, B. Manjunath, Wei Ma, Greg Hajic, Randy Kemp, Jason Simpson, Alex Wells, Jim Frew, and Qi Zheng of UCSB for their kind assistance and insightful comments.

This project was supported in part by the following grants:

- NSF/ARPA/NASA Digital Library Initiative, IRI94-113301, 1994–1998 (T. Smith, M. Goodchild, et al., "The Alexandria Project: Towards a distributed digital library with comprehensive services for images and spatially-referenced information")

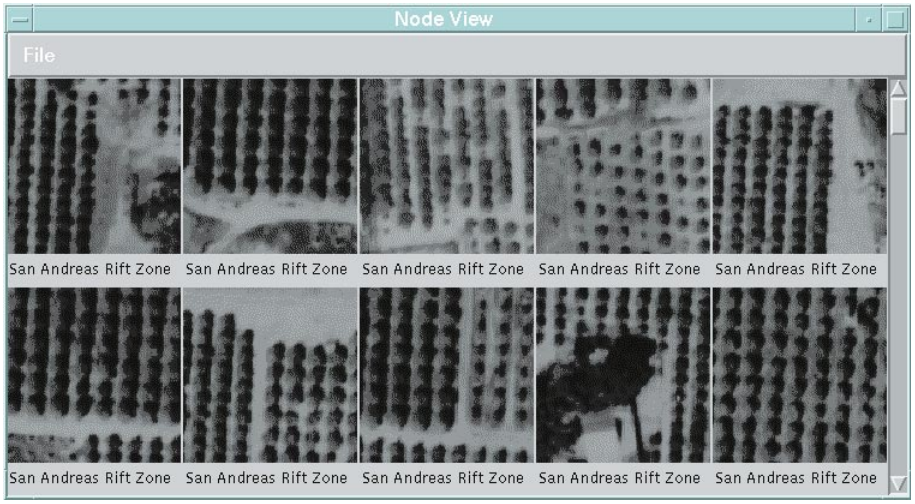

Fig. 15. Clicking on a representative image, e.g., a vegetation pattern, on the visual thesaurus results in a display of similar aerial photo patterns that are labeled by their corresponding place names

- NSF/ARPA/NASA Digital Library Initiative, 1996–1998 (H. Chen and T. Smith, "Supplement to Alexandria DLI Project: A semantic interoperability experiment for spatially-oriented multimedia data")
- NSF/ARPA/NASA Digital Library Initiative, IRI-9411318, 1994–1998 (B. Schatz, H. Chen, et al., "Building the Interspace: Digital library infrastructure for a university engineering community")
- NSF CISE, IRI-9525790, 1995–1998 (H. Chen, "Concept-based categorization and search on internet: A machine learning, parallel computing approach")
- National Center for Supercomputing Applications (NCSA), High-Performance Computing Resources Grants, 1994-1996 (H. Chen)

References

1. J. R. Anderson. *Cognitive Psychology and Its Implications, 2nd edn*. W. H. Freeman and Co., New York, NY, 1985
2. J. G. Carbonell, R. S. Michalski, T. M. Mitchell. An overview of machine learning. In *Machine Learning, An Artificial Intelligence Approach*, Pages 3–23, Michalski, R. S., Carbonell, J. G., and Mitchell, T. M.(eds.), Tioga Publishing Co., Palo Alto, CA, 1983
3. H. Chen. *An Artifical Intelligence Approach to the Design of Online Information Retrieval Systems*. Information Systems Department, New York University, Unpublished Ph.D. Thesis, 1989
4. H. Chen. Collaborative systems: solving the vocabulary problem. *IEEE Computer*, 27(5):58–66, Special Issue on Computer Supported Cooperative Work (CSCW), May 1994
5. H. Chen, K. J. Lynch. Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics*, 22(5):885–902, September/October 1992
6. H. Chen, K. J. Lynch, K. Basu, D. T. Ng. Generating, integrating, and activating thesauri for concept-based document retrieval. *IEEE Expert, Special Series on Artifical Intelligence in Text-based Information Systems*, 8(2):25–34, April 1993
7. H. Chen, 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. *Journal of the American society for Information Science*, 46(5):348–369, June 1995
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*, 18(8):771–782, August 1996
9. H. Chen, B. R. Schatz, T. Yim, D. Fye. Automatic thesaurus generation for an electronic community system. *Journal of the American for Society Information Science*, 46(3):175–193, April 1995
10. H. Chen, C. Schuffels, R. Orwig. Internet categorization and search: a machine learning approach. *Journal of Visual Communications and Image Representation*, 7(1):88–102, March 1996
11. P. R. Cohen, R. Kjeldsen. Information retrieval by constrained spreading activation in semantic networks. *Information Processing and Management*, 23(4):225–268, 1987
12. J. Dalton, A. Deshmane. Artificial neural networks. *IEEE Potentials*, 10(2):33–36, April 1991
13. D. Deckelbaum. A user survey conducted in the Henry J. Bruman map library. *Wester Association of Map Libraries Information Bulletin*, 20(3),170–197, June 1989
14. T. G. Dietterich, R. S. Michalski. A comparative review of selected methods for learning from examples. In *Machine Learning, An Artificial Intelligence Approach*, Pages 41–81, Michalski, R. S. Carbonell, J. G., Mitchell, T. M. (eds.), Tioga Publishing Co., Palo Alto, CA, 1983
15. K. Doan, C. Plaisant, B. Schneiderman. Query previews in networked information systems. In *Proceedings of the Third Forum ON Research and Technology Advances in Digital Libraries*, pages 120–129. IEEE Computer Society, 1996
16. B. Everitt. *Cluster Analysis*. Heinemann Educationai Books, London, UK, 1980
17. E. A. Feigenbaum. The art of artificial intelligence: themes and case studies in knowledge engineering. In *International Joint Conference of Artificial Intelligence*, pages 1014–1029, 1977
18. W. J. Frawley, G. Pietetsky-Shapiro, C. J. Matheus. Knowledge discovery in databases: an overview. In *Knowledge Discovery in Databases*, pages 1–30, G. Piatetsky-Shapiro, W. J. Frawley. (eds.), MIT Press, Cambridge, MA, 1991
19. M. Gluck. Geospatial information needs of the general public: text, maps, and users' tasks. In *Geographic Information Systems and Libraries*: *Patrons, Maps, and Spatial Information*, pages 151–172, (eds.), L. C. Smith, M. Gluck. University of Illinois, Champaign, IL, 1996

20. F. Hayes-Roth, N. Jacobstein. The state of knowledge-based systems. *Communications of the ACM*, 37(3):27–39, March 1994
21. F. Hayes-Roth, D. A. Waterman, D. Lenat. *Building Expert Systems*. Addison- Wesley, Reading, MA, 1983
22. L. L. Hill. Spatial access to, and display of, global change data: avenues for libraries. In *Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information*, pages 125–150, (eds.) L. C. Smith, and M. Gluck. University of Illinois, Champaign, IL, 1996
23. K. Knight, Connectionist ideas and algorithms. *Communications of the ACM*, 33(11):59–74, November 1990
24. T. Kohonen. *Self-Organization and Associative Memory*. (3rd edn.) Springer-Verlag, Berlin Heidelberg New York, 1989
25. T. Kohonen. *Self-Organization Maps*. Springer-Verlag, Berlin Heidelberg New York, 1995
26. F. W. Lancaster. *Information Retrieval Systems*. John Wiley &Sons, New York, NY, 1979
27. R. R. Larson. Geographic information retrieval and spatial browsing. In *Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information*, pages 81–124, (eds.), L. C. Smith, M. Gluck. University of Illinois, Champaign, IL, 1996
28. R. P. Lippmann. An introduction to computing with neural networks. *IEEE Acoustics Speech and Signal Processing Magazine*, 4(2):4–22, April 1987
29. B. S. Manjunath, 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. R. S. Michalski, J. B. Larson. Selection of most representative training examples and incremental generation of VL1 hypotheses: the underlying methodology and the description of programs ESEL and AQ11. In *Department of Computer Science*, University of Illinois, 1978
31. G. A. Miller. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41, November 1995
32. M. Minsky. *Semantic Information Processing*. MIT Press, Cambridge, MA, 1968
33. D. D. Montgomery. *Design and Analysis of Experiments*. John Wiley &Sons, New York, NY, 1976
34. D. F. Morrison. *Multivariate Statistical Methods*. McGraw-Hill Book Co., New York, NY, 1976
35. C. R. Nelson, *Applied Time Series Analysis of Managerial Forecasting*. Holden-Day, San Francisco, CA, 1973
36. R. Orwig, H. Chen, 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
37. A. Paepcke and et al. Using distributed objects for digital library interoperability. *Computer*, 29:61–68, 1996
38. K. Parsaye, M. Chignell, S. Khoshafian, H. Wong. *Intelligent Databases*. John Wiley & Sons, New York, NY, 1989
39. M. R. Quillian, Semantic memory. In *Semantic Information Processing*, M. Minsky, Editor, MIT Press, Cambridge, MA, 1968
40. J. R. Quinlan. Learning efficient classification procedures and their application to chess end games. In *Machine Learning, An Artificial Intelligence Approach*, Pages 463–482, Michalski, R. S. Carbonell, J. G., Mitchell, T. M. (eds.), Tioga Publishing Co., Palo Alto, CA, 1983
41. D. E. Rumelhart, G. E. Hinton, R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing*, Pages 318–362, D. E. Rumelhart, J. L. McClelland, PDP Research Group. (eds.), MIT Press, Cambridge, MA, 1986
42. G. Salton, *Automatic Text Processing*. Addison-Wesley, Reading, MA, 1989
43. B. R. Schatz, B, Mischo, T. Cole, J. Hardin, A. Bishop, H. Chen. Federating repositories of scientific literature. *IEEE Computer*, 29(5):28–36, May 1996
44. T. R. Smith. A digital library for geographically referenced materials. *IEEE Computer*, 29(5):54–60, May 1996
45. T. R. Smith. The meta-information environment of digital libraries. *D-Lib Magazine*, July 1996
46. J. F. Sowa, *Principles of Semantic Networks*. Morgan Kaufmann, Mateo, CA, 1991
47. H. D. Wactlar, T. Kanade, M. A. Smith, S. M. Stevens. Intelligent access to digital video: Informedia project. *IEEE Computer*, 29(5):46–53, May 1996
48. R. Wilensky. Toward work-centered digital information services. *IEEE Computer*, 29(5):37–45, May 1996
49. A. G. Woodruff, C. Plaunt. GIPSY: Automated geographic indexing of text documents. *Journal of the American Society for Information Science*, 45(9):645–655, October 1994