Concept-based searching and browsing: a geoscience experiment

Roslin V. Hauck

University of Arizona, USA

Robin R. Sewell

University of Arizona, USA

Tobun D. Ng

Carnegie Mellon University, Pennsylvania, USA

Hsinchun Chen

University of Arizona, USA

Received 11 January 2001
Revised 30 March 2001

Abstract.

In the recent literature, we have seen the expansion of information retrieval techniques to include a variety of different collections of information. Collections can have certain characteristics that can lead to different results for the various classification techniques. In addition, the ways and reasons that users explore each collection can affect the success of the information retrieval technique. The focus of this research was to extend the application of our statistical and neural network techniques to the domain of geological science information retrieval. For this study, a test bed of 22,636 geoscience abstracts was obtained through the NSF/DARPA/NASA funded Alexandria Digital Library Initiative project at the University of California at Santa Barbara. This collection was analyzed using algorithms previously developed by our research group: concept space algorithm for searching and a Kohonen self-organizing map (SOM) algorithm for browsing. Included in this paper are discussions of our techniques, user evaluations and lessons learned.

Correspondence to: R.V. Hauck, Department of Management Information Systems, University of Arizona, Tucson, Arizona 85721, USA. Tel: +1 520 621 2748. E-mail: rrv@bpa.arizona.edu

1. Introduction

As part of the University of California at Santa Barbara Alexandria Digital Library Initiative research on geological science and geographic information systems [1], this project explores the realm of textual information retrieval and analysis by experts in the geoscience discipline. As one study from an ongoing series of studies, this paper will summarize two algorithms developed earlier for the purpose of searching and browsing geoscience literature. There is then presented in detail how results from these two algorithms can be applied to the geoscience domain.

Geoscientists have information-seeking behaviours that are very similar to those of members of other scientific disciplines. While colleagues and contacts are the first source of information, they often use journal articles as their primary source of print information. Articles are usually located by referral from a colleague or contact, by browsing personal or library journal collections or by using large online databases [2].

Geoscientists encounter many of the same difficulties expressed by other database users when attempting to access information online and one of the biggest obstacles to information retrieval is the 'vocabulary problem' [3, 4]. Geoscientists are often frustrated in their attempts to find information, due to unfamiliarity with the indexing vocabulary used by the database or by an indexing terminology that is too broad to be useful. In addition, many geoscientists are unfamiliar with the print and online thesauri that can be used to facilitate their searches [2].

The University of Arizona Artificial Intelligence Lab has developed several methods and tools to improve access to information in large databases. These tools include concept space generation for searching and self-organizing maps for browsing. Other scientific communities have experienced an improvement in information retrieval with the use of computer-generated concept spaces [5]. The purpose of this study was to determine how the use of a geoscience concept space affects document recall and precision for a database of 22,636 geoscience documents obtained from the Alexandria Digital Library collection. In addition, the usefulness of a geoscience self-organizing map as a browsing tool for geoscience information was explored.

2. Literature review

Human and automatic indexing are two basic approaches to support textual geological science information retrieval. Automatic indexing can be further utilized for concept space and automatic thesaurus generation. In addition to indexing, this research focuses on the categorization of documents resulting in an interactive visual environment for browsing. This section presents a brief review of both searching and browsing techniques used in this study.

2.1. Indexing and searching

Human indexing and thesaurus

Indexing is the attachment of descriptive keyword terms to a unit of information to assist in its retrieval from a data collection [6]. This task is usually performed by a human subject specialist and can be both time-consuming and costly, making it unsuitable for very large collections. Because human indexers use terms selected from a subject-specific controlled vocabulary or thesaurus, attempting to select terms that adequately represent or describe the object can be difficult. Problems may be encountered at two levels of the indexing process: the human level and at the level of the thesaurus or controlled vocabulary.

As with any human activity, indexing can suffer from human inconsistency. Studies have found that a human indexer will rarely index an article the same way twice. Keyword selection can be influenced by changes in the indexer's subject knowledge or level of indexing expertise. In addition, it has been found that no two individuals, indexers or users, will use the same set of terms to describe a given object. For any two indexers, this represents a problem in indexing consistency. When the two are a user and an indexer, there will be an additional impact on information recall. The indexer may not see aspects of an object that a user may feel are important or significant enough to warrant representation in the indexing [6, 7, 8, 9]. Such a lack of descriptive compatibility can be compounded further by a user's unfamiliarity with the subject domain, leading to a vocabulary problem.

As a hierarchically arranged controlled subject vocabulary which attempts to identify terms accepted for use within a discipline, a thesaurus can even contribute to the difficulty of users attempting to access information. Often, the index terms that are identified by a vocabulary as the 'best terms' do not reflect the terminology commonly employed by users, especially novice users. The observation that many geoscientists do not take advantage of the thesauri available, and that those who do have found it frustrating, may be attributed to their having found that available thesauri tend to favour broad terminology and lack the fine granularity they desire in their searches. In addition, many searchers do not keep abreast of changes in indexing terminology and policy that can compound information retrieval difficulties [2].

Automatic indexing and automatic thesaurus

Automatic indexing is an attempt to address the time-consuming and labour-intensive process of human indexing of large databases of information, as well as the problem of human indexing inconsistency by using computer analysis of document term frequencies to select indexing terms. It relies on the assumption that all the terms necessary to describe a document are contained within the document itself. Terms can be selected and weighted based on their appearance within individual documents and within a collection as a whole. It has been found that terms with very high or very low frequencies in a document collection are ineffective as indexing terms. High-frequency terms have very low discrimination value and their use results in excessive document recall, while low-frequency terms appear too infrequently to be useful. The best terms seem to be those appearing with medium frequency [10, 11, 12].

Automatic analysis of co-occurring patterns on all pairs of terms in a document collection makes it possible to create a concept space or automatic thesaurus that can be used to address the vocabulary problem. Unlike a traditional human-generated thesaurus that establishes a term hierarchy, a concept space relies on the statistical analysis of the terms within documents and within a collection to generate a list of co-occurring terms. The undesirable side-effects and poor retrieval encountered with symmetric co-occurrence analysis [13] have been avoided by Chen *et al.*, by the application of an asymmetric co-occurrence analysis function and an algorithm that utilizes term filtering, automatic indexing and cluster analysis [14]. The two algorithms presented in this research are examples of techniques that create an automatic thesaurus or subject hierarchy.

2.2. Categorization and visualization

In addition to analysis of textual information, another problem with dealing with a large collection of documents is the issue of displaying the output in an unambiguous manner. One way in which developers are addressing this issue is through the use of categorization and visualization tools.

One such algorithm that combines the aspects of information categorization with visualization is Kohonen's self-organizing map (SOM). The name of this algorithm suggests a powerful property: because the map is self-organizing, pre-existing categorization structuring of the collection set is not needed. This results in an algorithm that is both robust and graphical in nature.

Many studies have applied the SOM to a wide variety of applications (e.g. image recognition, signal processing) [15]. One area where the SOM has been utilized is in that of textual analysis. Our research group has adapted an SOM algorithm that supports its applicability to the classification of a large-scale collection of geoscience abstracts.

2.3. Summary of techniques

Concept space generation

In order to create the concept space, the test bed of abstracts was first processed to remove journal information but retain the abstract document, title and the authors' names. The documents were then additionally processed in three steps: automatic indexing, co-occurrence analysis and associative retrieval [5].

Automatic indexing analyzes the frequency of term occurrence both within a document and within a document set. Stop words (e.g. 'a', 'the', 'in', 'on') were identified using a stopword list, and a stemming algorithm was applied to remove any variations of other stop words, such as plural forms. After identifying the stopwords, an algorithm was used to create one-, two- and three-word phrases from the remaining words. These phrases were used to create both a concept space and a SOM.

Co-occurrence analysis was used to generate what was called the GeoAbs Concept Space. After the term phrases were created using automatic indexing, an analysis of term frequency (frequency within a document) and inverse document frequency (frequency within the collection) was conducted. The document set was filtered to remove low-frequency words by eliminating any terms that occurred fewer than two times within a document and terms that appeared in only two documents. Co-occurrence analysis then allowed the determination of which terms co-occurred with one another both within a document and within the document set. A frequency computation algorithm was used to determine a co-occurrence weight, a number between 0 and 1, with the term(s) having weights closest to 1 given the highest ranking. A file was generated in which each term was given an assigned identification number followed by the list of ranked co-occurring terms.

Self-organizing map generation

The multi-layered self-organizing map has been explained in reports of previous research and utilizes a variation of the Kohonen SOM algorithm [16]. This allows concepts from abstracts to be organized based on co-occurrence within a two-dimensional space. The result is the grouping of similar documents in similar regions of the map.

In the SOM algorithm, continuous-value vectors between input and output nodes are presented without specification of the desired output. The underlying network connections of weights are then used to map input nodes to output nodes so that their vectors have the smallest Euclidean distance. All neighbouring nodes of the mapped node are then adjusted proportionally so that clusters of vector centres are created. These clusters are then displayed such that topologically close nodes represent similar input patterns (documents).

The basic SOM algorithm that was used for textual analysis in this study is similar to that described in [15] and [16].

The variant of the basic SOM algorithm has been used in our categorization of the geoscience documents. The scalable, multi-layered, graphical SOM that resulted was tested for usability in this experiment.

3. Experimental design

One goal of this research stream focuses on the distinguishing qualities of the geoscience collection from the collection of entertainment homepages discussed previously [18]. In the Internet study, a collection of 110,000 entertainment-related homepages was extracted from Yahoo!'s 'Entertainment' subdirectory for analysis by both the concept space and SOM techniques. Subjects, consisting of students attending summer school at the Management Information Systems and Library Science departments, were asked to evaluate both the concept space and SOM applications created from the Internet pages. The study concluded that, although the evaluations of both tools were positive, the applications did suffer in performance due to the eclectic nature of the Internet collection. We feel that focusing on domain-specific characteristics of the geoscience collection *will* lead to improvements in this current study.

The underlying characteristics of the geoscience collection differ from that of the collection used in the Internet experiment in ways that should result in the increased effectiveness of the searching as well as the browsing tools. The geoscience collection consists of a somewhat uniform structure, where each abstract contains certain elements (e.g. keywords, title, authors, abstract, etc). Compared to the collection of Internet homepages, the geoscience collection contains less noise and is a much cleaner dataset, which should result in improved algorithm performance. In addition, the users of the geoscience collection are subject-specific, in that those involved in evaluating the collection have a basic knowledge of geological science. With better algorithmic performance and a more domain-specific subject pool, it could be expected that both the concept space and the SOM applications should prove to be more effective with the geoscience collection than the Internet collection.

Utilizing the 22,636 geoscience abstracts obtained from the Alexandria Digital Library collection, a concept space and SOM were created and evaluated. For the evaluation of the concept space searching tool, recall and precision served as the dependent measures for the quantitative analysis. In addition, qualitative data in the form of user feedback was also collected. The SOM was evaluated specifically as a browsing tool and user feedback was collected.

3.1. Concept space searching experiment

The purpose of the study was to discover whether the use of a concept space tool could improve the recall and precision of geoscience information retrieval over using only the searcher's keywords. Keyword searching refers to the approach in which participants enter a keyword or set of keywords that they feel best characterize their information needs. Although research indicates that the use of keyword searching alone leads to satisfactory results [17], the goal here is to improve upon keyword searching by complementing it with concept space searching.

Twelve subjects with geoscience backgrounds participated in the study for a small monetary reward. Of the subjects, two were from the geoscience specialties of hydrology and environmental engineering, seven were geology students (Master's and doctoral levels) and three were professionals with the US Geological Survey. Each subject performed four searches on two different queries.

For each query, participants were asked to think of a topic of interest in the geoscience domain, verbalize their topic and then try to locate relevant abstracts. Each query was used to perform both a keyword and a concept space-based search. For example, if a search used the keywords 'clastic sediments' and 'fossils' to access the abstracts, the query would be repeated using the same keywords to enter the concept space and allow the subject to select terms to broaden or narrow the search. The order of searching method was alternated so that results were not confounded by method order; if the first search started with a keyword, the second search was begun with a concept space search and vice versa. In addition, the searching order was alternated between subjects, seven starting the first search with a keyword search and five starting with a concept space search.

The system is able to accept up to four keywords or keyword phrases to search a document or concept space. Figs. 1, 2, 3 and 4 show the interface and output. The interface contains radio buttons, which allow the user to access either the documents directly (*Geoscience Abstracts*) or the concept space. When the concept space is selected, a terms list is returned with numerical identifiers to show with which keyword term(s) the concept space term is associated in the test bed of abstracts. A subject may select more than one concept term by clicking on the boxes located to the left of the term. The subject may then choose to access the documents or may see another selection of concept terms related to the term chosen. If one term only is desired, the subject may choose between the document or concept space and click on the hyperlinked term to access the space selected. When multiple terms are selected for document display, documents retrieved by the highest number of terms (i.e. AND logic) are returned first, followed by documents retrieved by only one term (i.e. OR logic). The document space displays the abstracts with numbers identifying the terms responsible for the abstract's retrieval.

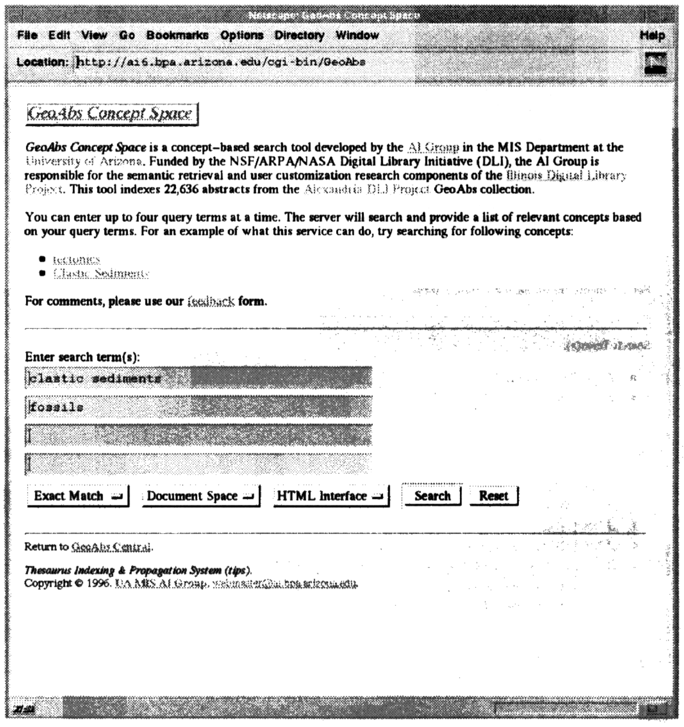
Subjects were encouraged to think aloud and give feedback on their search goals and the performance of the system. At the end of a concept space or keyword search, the abstracts were retrieved for examination. The system defaults to display the top 40 documents, though more may have matched the query. Due to time constraints, only the top 20 abstracts were examined. The subjects were asked to examine the abstracts, determine each document's relevance and briefly comment on why the document was or was not considered relevant to the particular query.

The document display (refer to Fig. 2) provided a number associated with the retrieval term, allowing us to determine whether the document was retrieved based on:

- keywords alone;
- concept space terms alone; or
- both keywords and concept space terms.

In calculation of keyword-retrieved documents, documents retrieved only by participant formulated keywords were counted. For calculation of number of documents retrieved only by concept space terms, credit was given only to the concept space terms for documents that would not have been retrieved by the subject's keyword(s). Thus, if a document was retrieved from both keyword term(s) and concept space term(s), the keyword always received credit for the term. It was felt that this was reasonable, since the participant would have retrieved the document without the help of the concept space. Finally, since the goal of the study was to see if the concept space terms could be used to improve users searches, the combined keyword and concept space retrieved documents was used in the analysis.

Fig. 5 details the formulas used to calculate recall and precision for the keyword searches, concept space searches and the combination searches. The calculation of recall was based on comparison of the number of documents the subject found relevant to the number of relevant documents existing in the entire collection, as determined by a subject expert. The unit of analysis used for this design was each search task. The subject chosen as the expert was the scientist-in-chief of the locally-based US Geological Survey office. The geology subject expert was asked to recreate each search by using modified search terms to obtain the set of total relevant documents within the entire testbed collection. Using the same procedure for assigning relevant documents to either the keyword or concept space systems, precision was calculated by comparing the number of relevant documents retrieved with the total document set found by the subject.


Netscape: GeoAbs Concept Space
File Edit View Go Bookmarks Options Directory Window Help
Location:

http://ai6.bpa.arizona.edu/cgi-bin/GeoAbs

GeoAbs Concept Space

GeoAbs Concept Space is a concept–based search tool developed by the AI Group in the MIS Department at the University of Arizona. Funded by the NSF/ARPA/NASA Digital Library Initiative (DLI), the AI Group is responsible for the semantic retrieval and user customization research components of the Illinois Digital Library Project. This tool indexes 22,636 abstracts from the Alexandria DLI Project GeoAbs collection.

You can enter up to four query terms at a time. The server will search and provide a list of relevant concepts based on your query terms. For an example of what this service can do, try searching for following concepts:

- tectonics
- Clastic Sediments

For comments, please use our feedback form.

Enter search term(s):

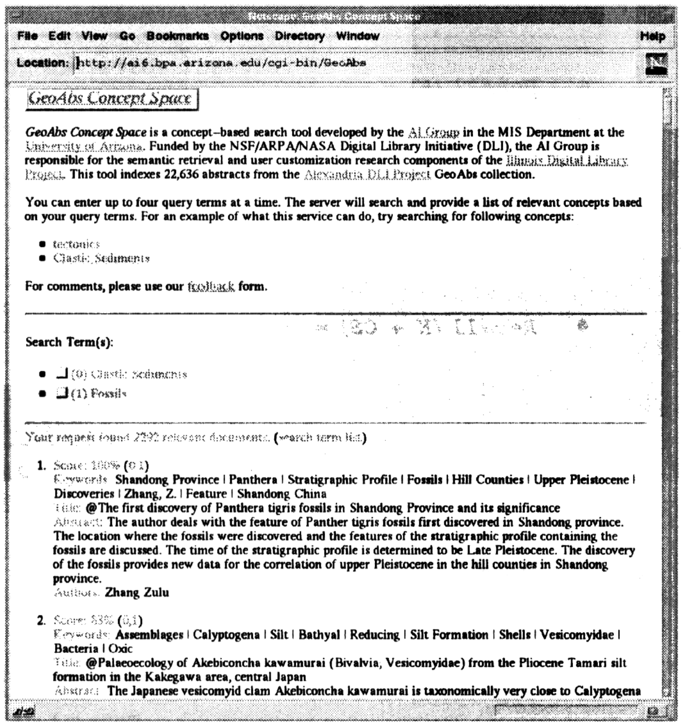
clastic sediments

fossils

Return to GeoAbs Central.

Thesaurus Indexing & Propagation System (tips).
Copyright © 1996. UA MIS AI Group, webmaster@ai.bpa.arizona.edu.

Fig. 1. A keyword search using the terms 'clastic sediments' and 'fossils'.


Netscape: GeoAbs Concept Space
File Edit View Go Bookmarks Options Directory Window Help
Location:

http://ai6.bpa.arizona.edu/cgi-bin/GeoAbs

GeoAbs Concept Space

GeoAbs Concept Space is a concept–based search tool developed by the AI Group in the MIS Department at the University of Arizona. Funded by the NSF/ARPA/NASA Digital Library Initiative (DLI), the AI Group is responsible for the semantic retrieval and user customization research components of the Illinois Digital Library Project. This tool indexes 22,636 abstracts from the Alexandria DLI Project GeoAbs collection.

You can enter up to four query terms at a time. The server will search and provide a list of relevant concepts based on your query terms. For an example of what this service can do, try searching for following concepts:

- tectonics
- Clastic Sediments

For comments, please use our feedback form.

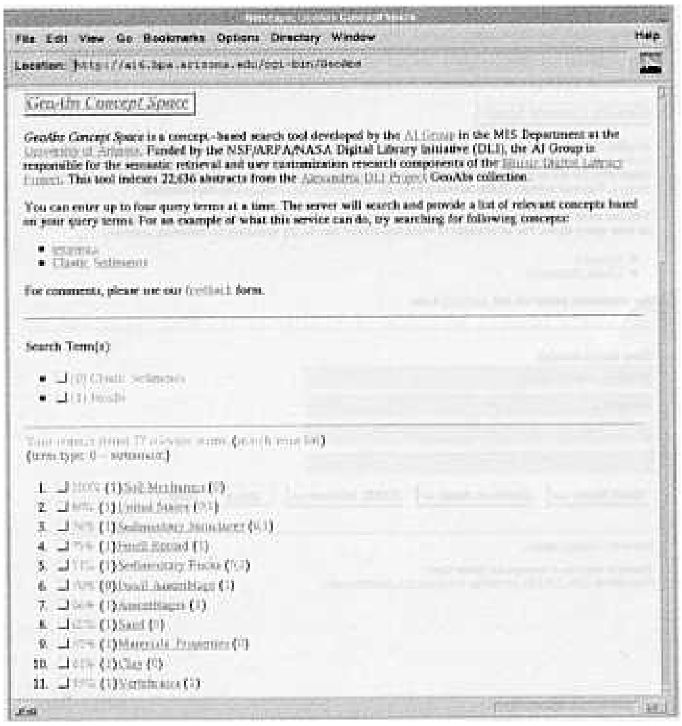
Search Term(s):

- □ (0) Clastic Sediments
- □ (1) Fossils

Your request found 2292 relevant documents. (search term list)

1. Score: 100% (0,1)
   Keywords: Shandong Province | Panthera | Stratigraphic Profile | Fossils | Hill Counties | Upper Pleistocene | Discoveries | Zhang, Z. | Feature | Shandong China
   Title: @The first discovery of Panthera tigris fossils in Shandong Province and its significance
   Abstract: The author deals with the feature of Panther tigris fossils first discovered in Shandong province. The location where the fossils were discovered and the features of the stratigraphic profile containing the fossils are discussed. The time of the stratigraphic profile is determined to be Late Pleistocene. The discovery of the fossils provides new data for the correlation of upper Pleistocene in the hill counties in Shandong province.
   Authors: Zhang Zulu
2. Score: 83% (0,1)
   Keywords: Assemblages | Calyptogena | Silt | Bathyal | Reducing | Silt Formation | Shells | Vesicomyidae | Bacteria | Oxic
   Title: @Palaeoecology of Akebiconcha kawamurai (Bivalvia, Vesicomyidae) from the Pliocene Tamari silt formation in the Kakegawa area, central Japan
   Abstract: The Japanese vesicomyid clam Akebiconcha kawamurai is taxonomically very close to Calyptogena

Fig. 2. Abstracts retrieved by the keyword search. *Note* the numerical reference for each keyword, which corresponds to the keyword responsible for the retrieval.



Location: http://ai6.bpa.arizona.edu/cgi-bin/GeoAbs

GeoAbs Concept Space

GeoAbs Concept Space is a concept-based search tool developed by the AI Group in the MIS Department at the University of Arizona. Funded by the NSF/ARPA/NASA Digital Library Initiative (DLI), the AI Group is responsible for the semantic retrieval and user customization research components of the Illinois Digital Library Project. This tool indexes 22,636 abstracts from the Alexandria DLI Project GeoAbs collection.

You can enter up to four query terms at a time. The server will search and provide a list of relevant concepts based on your query terms. For an example of what this service can do, try searching for following concepts:

- [illegible]
- Clastic Sediments

For comments, please use our feedback form.

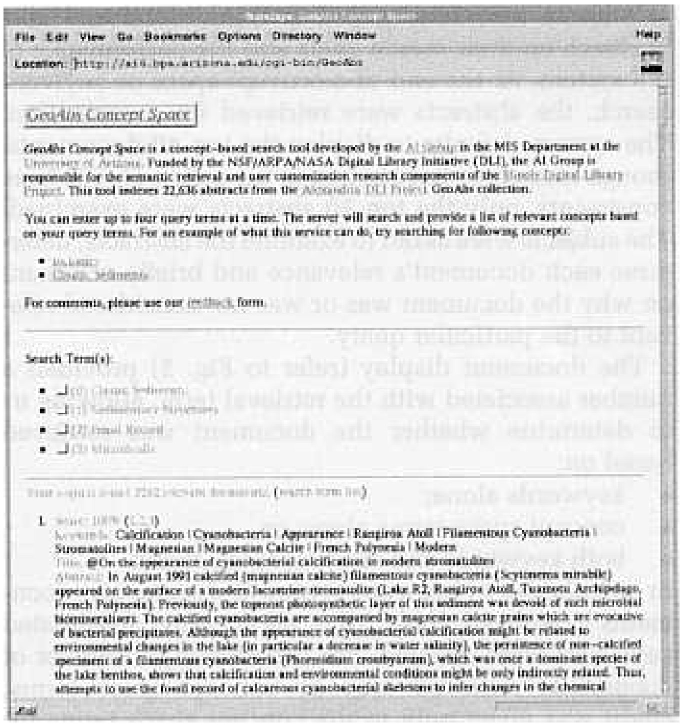
Search Term(s):

- ☐ (0) Clastic Sediments
- ☐ (1) Fossils

[illegible] (search term list)
(term type: 0 – [illegible])

1. ☐ 100% (1)Soil Mechanics (0)
2. ☐ [illegible] (1)United States (0,1)
3. ☐ [illegible] (1)Sedimentary Structures (0,1)
4. ☐ [illegible] (1)Fossil Record (1)
5. ☐ [illegible] (1)Sedimentary Facies (0,1)
6. ☐ [illegible] (0)Fossil Assemblages (1)
7. ☐ [illegible] (1)Assemblages (1)
8. ☐ [illegible] (1)Sand (0)
9. ☐ [illegible] (1)Materials Properties (0)
10. ☐ [illegible] (1)Clay (0)
11. ☐ [illegible] (1)Vertebrates (1)

Fig. 3. Concept space searching. The same keywords have been used to access the concept space, rather than the documents.



Location: http://ai6.bpa.arizona.edu/cgi-bin/GeoAbs

GeoAbs Concept Space

GeoAbs Concept Space is a concept-based search tool developed by the AI Group in the MIS Department at the University of Arizona. Funded by the NSF/ARPA/NASA Digital Library Initiative (DLI), the AI Group is responsible for the semantic retrieval and user customization research components of the Illinois Digital Library Project. This tool indexes 22,636 abstracts from the Alexandria DLI Project GeoAbs collection.

You can enter up to four query terms at a time. The server will search and provide a list of relevant concepts based on your query terms. For an example of what this service can do, try searching for following concepts:

- [illegible]
- Clastic Sediments

For comments, please use our feedback form.

Search Term(s):

- ☐ (0) Clastic Sediments
- ☐ (1) Sedimentary Structures
- ☐ (2) Fossil Record
- ☐ (3) [illegible]

[illegible] (search term list)

1. Score: 100% ([illegible])
Keywords: Calcification | Cyanobacteria | Appearance | Rangiroa Atoll | Filamentous Cyanobacteria | Stromatolites | Magnesian | Magnesian Calcite | French Polynesia | Modern
Title: @On the appearance of cyanobacterial calcification in modern stromatolites
Abstract: In August 1991 calcified (magnesian calcite) filamentous cyanobacteria (Scytonema mirabile) appeared on the surface of a modern lacustrine stromatolite (Lake R2, Rangiroa Atoll, Tuamotu Archipelago, French Polynesia). Previously, the topmost photosynthetic layer of this sediment was devoid of such microbial biomineralizers. The calcified cyanobacteria are accompanied by magnesian calcite grains which are evocative of bacterial precipitates. Although the appearance of cyanobacterial calcification might be related to environmental changes in the lake (in particular a decrease in water salinity), the persistence of non-calcified specimens of a filamentous cyanobacteria (Phormidium crosbyanum), which was once a dominant species of the lake benthos, shows that calcification and environmental conditions might be only indirectly related. Thus, attempts to use the fossil record of calcareous cyanobacterial skeletons to infer changes in the chemical

Fig. 4. Abstracts retrieved by the concept space terms.

- $$\text{Recall (K)} = \frac{\text{(\# of relevant abstracts found from keywords alone)}}{\text{Total \# of Relevant documents in data set}}$$
- $$\text{Recall (CS)} = \frac{\text{(\# of relevant abstracts found from consept space alone)}}{\text{Total \# of Relevant documents in data set}}$$
- $$\text{Recall (K + CS)} = \frac{\text{(\# of relevant abstracts found from both keywords and concept spaces)}}{\text{Total \# of Relevant documents in data set}}$$
- $$\text{Precision (K)} = \frac{\text{(\# of relevant abstracts found from keywords alone)}}{\text{Total \# of Relevant documents retrieved by User}}$$
- $$\text{Precision (CS)} = \frac{\text{(\# of relevant abstracts found from concept space alone)}}{\text{Total \# of Relevant documents retrieved by User}}$$
- $$\text{Precision (K +CS)} = \frac{\text{(\# of relevant abstracts found from both keyword and concept space)}}{\text{Total \# of Relevant documents retrieved by User}}$$

Fig. 5. Recall and precision formulas.

3.2. Self-organizing map experiment

The self-organizing map was presented to subjects as a browsing tool. As discussed before, previous research demonstrated subjects' inability to use it for searching [18]. At the end of the concept space searching experiment, each subject was asked to use the SOM for browsing and give verbal feedback as to its organization, effectiveness at categorization and general usefulness. Because this was a browsing process and not a search, precision and recall could not be calculated. Figs. 6, 7 and 8 show examples of browsing the map from the top layer down to the abstracts.

The top layer of the multi-layered organization of the SOM map provides a high-level picture of the geoscience collection (see Fig. 6). Each space or region in the map is categorized by a single topic. All regions in a higher layer can be broken up into multiple categories by entering its sub-layer (Fig. 7). This process continues until the number of abstracts contained within a space falls below the threshold of 200. Once this sub-layer is reached, the actual abstracts are returned (Fig. 8).

The individual regions in a map are shown in different colours, to indicate the space boundaries. Additionally, the relative size of the region specifies the number of abstracts contained by the space. The number to the right of the topic name indicates the exact number of abstracts contained in that region. Adjacent regions in the map indicate clusters of abstracts that axe categorized to be similar. For example, in Fig. 6, the region titled 'Water' contains abstracts that are categorized as similar to those abstracts in the adjacent regions 'Ground Water', 'Hydrology', 'Paleozoic' and 'Sediments'. *Note* that there are two regions of 'Hydrology'. This indicates that the different regions contain different abstracts that are found to be similar to other different regions.

4. Searching experiment results and discussion

The results of the quantitative data analysis on recall and precision are quite supportive of concept space searching.

The initial analyses dealt only with precision and recall of keyword terms alone and concept space terms alone. This allows us to understand the performance of the two sets of terms, as they would perform independently. The second analysis is the primary one of interest: the comparison between keyword term alone and the combination of keyword and concept space terms. The first set of recall and precision calculations, which compares keywords versus concept space terms, leads to an interesting observation. The data show that keyword searching alone results in significantly higher recall than concept space searching alone: F (1,46) = 10.30, p<0.01. In precision analysis, keywords searching alone is also higher than concept space searching alone: F (1,46) = 15.85, p<0.001.

Since the intended use of concept space is to complement the search efficiency of keyword searching, the inclusion of the concept space terms with keyword terms compared to searches via keyword alone was then analyzed. Results of the recall analysis show that there was a significant difference when the recall ability of the concept space with keywords was compared to keyword searching alone: F (1,46) = 3.77, p<0.10. The interpretation of this result is that inclusion of terms suggested by the concept space significantly improved identification of relevant abstracts from a pool of potentially relevant abstracts.

No significant difference in precision was found between keyword searching and the addition of concept searching: F (1,46) = 0.07, *ns*. This means that inclusion of terms suggested by the concept space plus keywords chosen by the subject performed no more

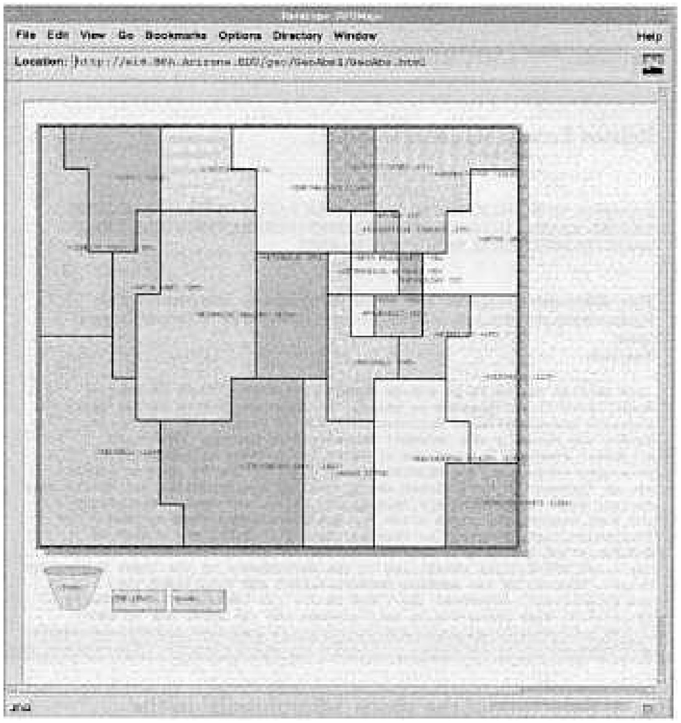
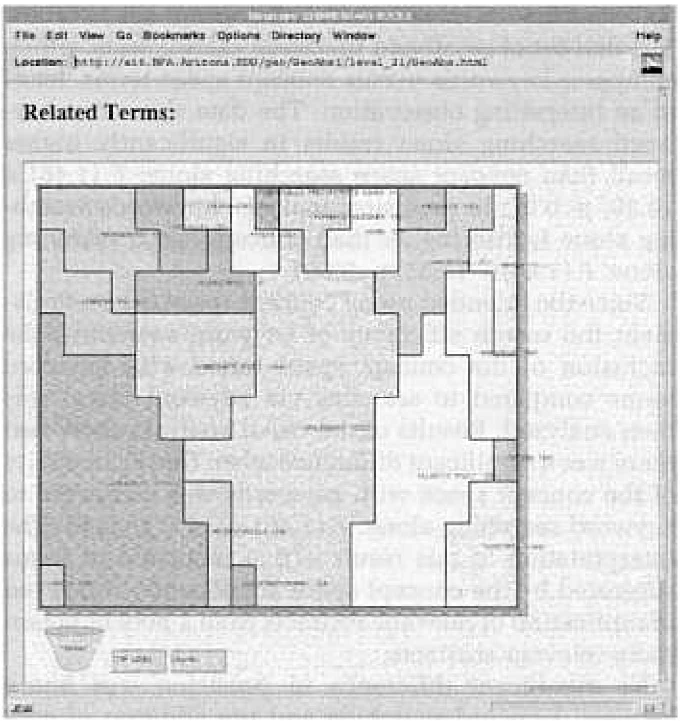
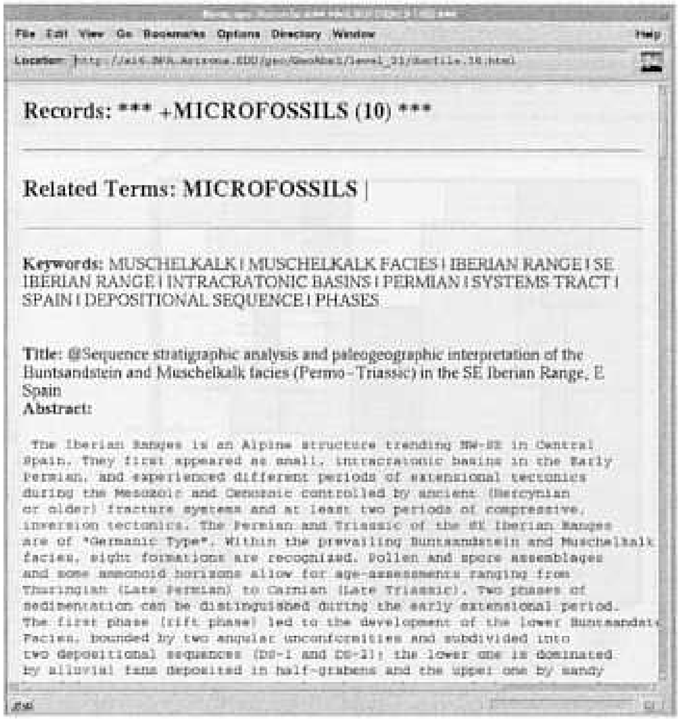


Fig. 6. The top level of the GeoAbs map.



Related Terms:

Fig. 7. Selection of the space identified as 'Sedimentary rocks' takes the user to the second level of the map.



Records: *** +MICROFOSSILS (10) ***

Related Terms: MICROFOSSILS |

Keywords: MUSCHELKALK | MUSCHELKALK FACIES | IBERIAN RANGE | SE IBERIAN RANGE | INTRACRATONIC BASINS | PERMIAN | SYSTEMS TRACT | SPAIN | DEPOSITIONAL SEQUENCE | PHASES

Title: @Sequence stratigraphic analysis and paleogeographic interpretation of the Buntsandstein and Muschelkalk facies (Permo-Triassic) in the SE Iberian Range, E Spain
Abstract:

```
 The Iberian Ranges is an Alpine structure trending NW-SE in Central
Spain. They first appeared as small, intracratonic basins in the Early
Permian, and experienced different periods of extensional tectonics
during the Mesozoic and Cenozoic controlled by ancient (Hercynian
or older) fracture systems and at least two periods of compressive,
inversion tectonics. The Permian and Triassic of the SE Iberian Ranges
are of "Germanic Type". Within the prevailing Buntsandstein and Muschelkalk
facies, eight formations are recognized. Pollen and spore assemblages
and some ammonoid horizons allow for age-assessments ranging from
Thuringian (Late Permian) to Carnian (Late Triassic). Two phases of
sedimentation can be distinguished during the early extensional period.
The first phase (rift phase) led to the development of the lower Buntsandst
Facies, bounded by two angular unconformities and subdivided into
two depositional sequences (DS-1 and DS-2); the lower one is dominated
by alluvial fans deposited in half-grabens and the upper one by sandy
```

Fig. 8. Selection of the space 'Microfossils' in the 'Sedimentary rocks' map takes the user to ten abstracts related to the selected term.

poorly in terms of precision than searching by keywords alone.

In addition to the results of the quantitative analysis, the analysis of the qualitative information obtained from subject feedback and researcher observation led to the emergence of a number of general themes.

4.1. Searching experience

A majority of the subjects reported liking the concept space search tool better than the keyword search tool. Both technically experienced and less-experienced subjects were able to demonstrate that the concept space search tool provided a helpful method of finding specific information.

Number of chosen concept space terms

It was observed that subjects who chose fewer terms tended to find abstracts that were reported to be more useful. Subjects who chose a large number of concept space search terms seemed to generate more diluted results.

When presented with the generated concept space, subjects were often delighted by the novelty of the tool. It appeared that this encouraged them to choose a large number of concept space search terms. Therefore, it is posited that the dilution of results stemming from choosing too many concept space terms is an artefact of the subject's unfamiliarity with how the tool should be used.

Spelling errors and different forms of words

Spelling errors and use of singular versus plural forms of words resulted in frustration for some subjects. Although a subject might have used the correct term in the search field, the search tool may have failed to recognize the term because the system contained the plural rather than the singular form. Subjects would return to the previous screen and change the form of the term, in the hope that a different form of the word would be recognized. Subjects often asked if a 'wild card' term could be used in the search.

4.2. Positive concept space comments

During the course of this study, a number of recurring positive comments surfaced. These comments can be generalized in the following categories.

Search refinement

Many of the subjects reported that the ability to refine their search by using concept space terms was very useful. For example, as one subject put it, the concept space tool provides 'options you wouldn't think of on your own'. This type of search method would save time and thus allow users to be more efficient. In general, the documents found by concept space searching were reported to be more relevant to the subjects' search criteria than those found by keyword searches.

Ranking, term indicator information and totals

Many subjects indicated that they liked the score ranking and the indication of terms from the initial input that retrieved the document (refer to Fig. 4). Subjects used this information to quickly assess whether or not an abstract contained the search term specified in the concept search.

4.3. Negative concept space comments

In general, negative comments addressed a number of issues that could be attributed to the subjects' unfamiliarity with this type of search tool. Many subjects voiced uncertainty about how the concept terms were generated. Questions regarding the user interface and limited functionality are explored in further depth.

Term relationship

Some subjects had trouble understanding how the terms suggested by the thesaurus were related to each other and to the input term(s). One subject stated that she did not 'understand [the] relationship between keywords and concepts generated'. The underlying issue most likely stems from the subjects' lack of familiarity with this type of term association model, rather than the more familiar hierarchical mental structure.

User interface problems

In the discipline of geological science, many subjects focused on specific geographical locations or temporal periods for their research. This issue brought out interesting feedback on the user interface and the capabilities of the concept space tool. For example, a subject interested in a number of general terms (e.g. tectonics and transcurrent fault) and one specific term (e.g. the North Slope) may have felt that the documents generated did not address the specific term so they were not useful.

A number of subjects who had this problem suggested having a method of allocating weights to the concept terms to specify the importance of each term in the document retrieval process. Being able to 'demand inclusion of [some] terms' would force the more relevant documents to be ranked higher in the document retrieval. This notion would also allow users to be more in control of a search by indicating which terms are most important.

5. Browsing experiment results and discussion

Subjects were asked to explore GeoMap, specifically addressing the comprehension of the relationship between terms as well as the usefulness and logic of this Kohonen SOM-based technique of browsing. Verbal protocols during browsing revealed a number of interesting issues.

5.1. Positive feedback about the GeoMap

Some general themes were apparent in the positive feedback that was received about the GeoMap as a browsing tool.

Graphical aspects: spatial factor and colour

Subjects repeatedly commented on the use of visualization as a technique for browsing documents. They liked the visual distribution of the information created by the spatial arrangement that allowed them to survey the information quickly. In addition to providing topic identification, the use of colour seemed to make the map more visually interesting to the subjects.

SOM usage

Several subjects commented that this tool would be most beneficial to users with little prior knowledge of geology. One subject commented that using the map was similar to 'looking through an encyclopedia'. Because the SOM is a browsing rather than a searching tool, it seemed less appropriate for use by subjects specializing in the field.

Novelty

Many subjects were intrigued by the novelty of the categorization of the terms in the graphical representation. The notion that the maps were computer-generated held much intrigue for subjects. The grouping of topics and sub-layers of topics was described as 'interesting' and a 'great idea'.

5.2. Negative feedback about the GeoMap

Most negative feedback had to do with the unfamiliarity of the word association structure (new mental model compared with more classical alphabetic or hierarchical organization structures), issues dealing with the user interface or attempts to utilize the tool for searching rather than for browsing.

Use of GeoMap for searching

Although it was presented as a browsing tool, many subjects still directed their efforts toward using the tool for searching, which often resulted in frustration from being unable to specify search terms. One subject suggested including a search tool in conjunction with the maps to further limit the search.

Hierarchical organization or conventional organization

Many subjects voiced a need for a more systematic organization of terms in the map. Some suggested mandatory inclusion of terms not currently included in the map topics. One subject elaborated on this idea by suggesting 'some coaching to delineate sub-categories would help make the grouping of the articles more sensible'. Another subject argued that the categories found on the top level of the map were not parallel. For example, although the top map consisted of a mixture of types of fields, types of rock and time periods, these areas were not complete (e.g. one time period was missing).

Inability to generalize

One recurring criticism of the GeoMap was a lack of ability to generalize or to present topics at the same level of abstraction. Some comments include 'categories are broad', 'big pieces are ambiguous', 'it is hard to see the relationship' between terms and 'terms should be connected easily'. One subject commented that if she could not find a key term on the top map, she was forced to explore a number of sub-maps in order to locate the particular topic.

6. Recommendations

In light of the previous discussion, a number of recommendations can be posited for future iterations of this and similar studies, as well as future generations of the technology itself.

6.1. Concept space search

After reviewing the subjects' feedback on this tool, a number of areas are targeted for further development.

Weight allocation for search terms

It appears that including a function in the tool's design that allows subjects to specify levels of importance or weight to each search term would serve to better refine and limit a search. For example, the user would be able to specify that the returned documents would always include a particular term. For the geological discipline, this type of ability should prove quite valuable.

Experimental protocol

Due to subjects' lack of familiarity with this type of search tool, an experimenter should be able to guide the subjects by highlighting useful tips for using the concept spare tool. Limiting the number of concept terms selected in order to prevent the dilution of the relevant documents might prove helpful.

6.2. GeoMap browsing

In reviewing subjects' comments and researchers' observations, a number of recommendations for the GeoMap evaluation are suggested.

User characteristics

A majority of the subjects reported that the GeoMap was not useful for searching purposes. It is also generally true that a user with a high level of familiarity with the extant literature and topics will most likely choose not to utilize a tool that browses rather than searches for specific topics. For this reason, it could be concluded that the subject's characteristics constitute a mediating variable in determining the usefulness of this browsing tool.

By exploring how users with a lower level of familiarity with the discipline use this tool, its effectiveness can be re-evaluated in terms of serving the purpose of browsing in order to orient users with the material. It is possible that by focusing on user characteristics as a variable contributing to usability, the extent to which this tool can be most effective in a specific context can be determined.

Organization of GeoMap

It is recommended that the categorization and sub-categorization of the topics in the map be explored further. Because the top-level map is not mutually exclusive, users who wish to explore some topic of interest may not be able to locate that topic in the first few levels of the map. This is especially crucial for general categories in the traditional partitioning of the information. Being able to specify particular categories that must be included in at least the top-level map would allow users to determine the location of particular topics more accurately.

7. Conclusions

In conclusion, the authors feel encouraged by the study results for both the GeoAbs Concept Space and the GeoMap SOM. This research study indicates that characteristics of the overall collection, as well as characteristics of the user group, can determine the performance of the application.

The statistical results for precision and recall indicate that the inclusion of concept space terms equals (precision) or surpasses (recall) the performance of keywords searching. Positive comments about the concept search include its search refinement ability, as well as the ranking and term indicators information included in the Concept Space interface. Although subjects did have some difficulty resulting from unfamiliarity with this type of search tool, the comments offered present useful suggestions for future directions of concept space.

The GeoMap also received many positive comments from the subjects. The use of graphical aspects, such as spatial arrangements and colour, added visualization value for the subjects. The authors also learned that user characteristics should also be considered when determining the most efficient way for this tool to be deployed. Users who have an in-depth knowledge of the specific domain would find the GeoMap browsing tool to be less effective than would those having a limited knowledge of the geoscience literature. Issues of topic organization in the SOM surfaced many questions concerning the possible need for a more systematic organization of the map categories.

Although these results are preliminary, they are encouraging to ongoing effort. It is planned to continue to address and further explore these and other issues for various other information retrieval and analysis applications.

Acknowledgements

This project was funded primarily by a grant from the NSF/DARPA/NASA Digital Library Initiative, 'Supplement to Alexandria DLI Project: A Semantic Interoperability Experiment for Spatially-Oriented Multimedia Data' (1996–1998, IRI-9411330) and the NSF/DARPA/NASA Digital Library Initiative, 'Building the Interspace: Digital Library Infrastructure for a University Engineering Community' (1994–1998, HU-9411318). In addition, this work was partially supported by National Computational Science Alliance under HU970002N and utilized the NCSA SGI/CRAY Origin2OOO.

We would like to thank all of the subjects from the Geology, Hydrology and Environmental Engineering departments and the US Geological Survey who participated in this study. Special thanks go to Andrew Moosmann, for his help with editing, and the anonymous reviewers for their insightful comments and feedback.

References

[1] T.R. Smith, A digital library for geographically referenced materials, *IEEE COMPUTER* 29(5) (1996) 54–60.

[2] J. Bichteler and D. Ward, Information-seeking behavior of geoscientists, *Special Libraries* 80(3) (1989).

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

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

[5] H. Chen, J. Martinez, D.T. 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* 48(1) (1997) 17–31.

[6] 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) (1987) 964–971.

[7] M.J. Bates, Subject access in online catalogs: a design model, *Journal of the American Society for Information Science* 37(6) (1986) 357–376.

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

[9] L.M. Gomez, C.C. Lochbaum and T.K. Landauer, All the right words: finding what you want as a function of the richness of indexing vocabulary, *Journal of the American Society for Information Science* 41(8) (1990) 547–559.

[10] G. Salton, Another look at automatic text-retrieval systems, *Communications of the ACM* 29(7) (1986) 648–656.

[11] G. Salton and M. McGill, *Introduction to Modern Information Retrieval* (McGraw-Hill, New York, 1983).

[12] G. Salton, A. Wong and C.S. Yang, A vector space model for automatic indexing, *Communications of the ACM* 18(11) (1975) 613–620.

[13] H.J. Peat and P. Willett, The limitations of term co-occurrence data for query expansion in document retrieval systems, *Journal of the American Society for Information Science* 42(5) (1991) 378–383.

[14] B. Everitt, *Cluster Analysis* 2nd ed. (Heinemann Educational Books, London, 1980).

[15] T. Kohonen, *Self-organizing Maps* (Springer-Verlag, Berlin, 1995).

[16] H. Chen, C. Schuffels and R. Orwig, Internet categorization and search: a machine learning, *Journal of Visual Communications and Image Representation* 7(l) (1996) 88–102.

[17] J. Tillotson, Is keyword searching the answer? *College and Research Libraries* 56 (1995) 199–206.

[18] H. Chen, A. Houston, R. Sewell and B. Schatz, Internet browsing and searching: user evaluations of category map and concept space techniques, *Journal of the American Society for Information Science* 49(7) (1998) 582–603.