

# Identifying the Subject of Documents in Digital Libraries Automatically Using Frequently-Occurring Words - Study and Findings

*[Offer Drori](#)*

The Hebrew University of Jerusalem

E-mail: [offerd@cs.huji.ac.il](mailto:offerd@cs.huji.ac.il)

## **ABSTRACT**

Contemporary information databases contain millions of electronic documents. The immense number of documents makes it difficult to conduct efficient searches on the Internet. Several studies have found that associating documents with a subject or list of topics can make them easier to locate online [5] [6] [7]. Effective cataloging of information is performed manually, requiring extensive resources. Consequently, at present most information is not cataloged. This paper will present the findings of a study based on a software tool (TextAnalysis) that automatically identifies the subject of a document. We tested documents in two subject categories: geography and family studies. The present study follows an earlier one that examined the subject categories of industrial management and general management.

**KEYWORDS:** frequently-occurring words, Web documents classification, search results list, identify topics of documents, catalog, cataloging

## **INTRODUCTION**

SearchEngineWatch (June 2000) reports that the World Wide Web contains over one billion pages of information, not including the information contained in hosted databases. Because of the sheer quantity of data, and the huge resources required for cataloging, it is unreasonable to expect that any significant part of this information base will be cataloged. In a world where the quantity of information is growing at a rapidly increasing pace, automatic cataloging of information and documents is a necessity. It is vital for locating information.

Cataloging information comprises associating information with a list of predefined subjects or concepts. Librarians pioneered the cataloging of human knowledge, developing methods to associate a book collection (and subsequently other media) by subject. The most widely-used and long-established cataloging systems are the Dewey classification system (which uses decimal numbers to divide subjects into categories) and Library of Congress Subject Headings, or LCSH (which uses a list of terms to catalog the library of the United States Congress).

Alongside these traditional methods, additional tools have been developed to associate existing texts with a given set of categories. One of the most widely-used tools is the Yahoo! web site directory [17], which manually catalogs large numbers of pages on the Internet. Because the cataloging is manual, however, Yahoo!'s search engine offers only very limited coverage.

Several studies have been conducted on the automatic organizing of information. Most address the Internet and the manner in which information is displayed online. Allen [2] developed a prototype to display search results using the Dewey method. Another initiative, the Superbook project (1993), arranged text paragraphs in a table of hierarchical contents, similar to a table of contents [13]. Marchionini et al. [11] developed a table of contents that resembles searches in the Library of Congress. The WebCutter system, developed by Maarek et al., features a category-based user search map [10].

Clustering is another way to automatically organize a group of documents. With clustering, groups of documents are arranged on the basis of similarities rather than on a predefined set of categories. Several clustering-based projects are currently underway, but since most barely touch on the issue of cataloging, which is the crux of this article, we will cite only a few: Zamir & Etzioni [18] [19], Hearst and Pedersen [8], and Sahami et al. [14].

Classification is a third means of organizing documents into groups. It applies statistical techniques to documents for which a category has been defined by other means. The system learns the behavior of the documents with respect to the defined categories, and enables the creation of a similar category for documents that were not pre-cataloged. This method is less relevant to the contents of this article, and so, here too, we will cite only a few sources: Chekuri et al. [3], Mladenic [12], and Chen and Dumais [4].

Automatic natural processing is the fundamental challenge of computerized textual information management. Korfhage [9] defined three basic approaches for processing text documents: lexical, syntactic, and semantic analysis. The syntactic approach attempts to enable the computer to understand a natural language sentence structure, and semantic analysis attempts to identify the semantic structure of a document and thus disclose its meaning. To date, lexical text analysis has, however, proven the most popular and effective method. It uses different statistical techniques based on terms' frequency of appearance to pinpoint the most important terms – the “keywords” – that characterize each document, documents group, or topic. Once the keywords list for existing documents in a search results list is generated, the documents may be found and displayed in relation to its topic.

### **LOCATING THE SUBJECT OF THE ARTICLE**

A series of experiments to determine the most effective way of presenting information in a list of search engine results found that presenting the subject of the document, or the category with which it is associated, offered users several benefits [5] [6] [7]. The principal advantage was that the user was able to find the required document by viewing the list of results from the search query, without having to actually read all the documents in the list. Displaying the subject of each document in the search results list allowed the user to focus only on documents in the subject category of interest. The study's conclusions [7] were:

1. Adding keywords to the information displayed in the search results list cuts down on search times, compared with a display of information that does not include keywords.
2. Adding keywords to the information displayed in the search results list will improve the user's sense that the defined search terms will yield the correct answer, compared with a display of information that does not include keywords.
3. Adding keywords to the information displayed in the list of search results will improve the user's feeling of satisfaction, compared with a display of the same information that does not include keywords.

The listed subject of a document may be a category derived from a pre-defined list, or several words that are the list of topics of the document or list of keywords. In our study, we choose to focus on keywords, because this was part of the subject of earlier studies we conducted and for a number of other reasons that will be addressed hereunder. The user interface used for the study is presented in Appendix 2.

Several methods can be applied to identify the subject of a document. The most popular is based on manual characterization of the document according to different categories. In the case of scientific documents, the author specifies the keywords.

In digital libraries, the database team adds a list of terms relevant to the article (for an example, see the Index Terms in the ACM Digital Library [1]). On the Internet, the directory's staff adds the category with which the document is associated (see Yahoo! [17]) to computerized directories. In addition to these manual methods, which are relatively accurate but resource-intensive, a computerized system of characterization is required to catalog an extensive body of documents from different sources, including documents that have not been indexed or cataloged.

The SONIA system (Service for Organizing Networked Information Autonomously) employs a combination of technologies that takes the results of queries to networked information sources and, in real-time, automatically retrieves, parses, and organizes these documents into categories [14].

On the Internet in its present state, and absent specific sample queries and user relevance

assessments, we know little *a priori* about what renders an item relevant or irrelevant. Rather than focusing on the relevance or irrelevance of particular documents, it seems reasonable to first consider the frequencies of occurrence properties of the terms in complete document collections. In the composition of written texts, grammatical function words such as “and,” “of,” and “or” exhibit approximately equal frequencies of occurrence in all the documents in a collection. Moreover, most function words are characterized by high occurrence frequencies in ordinary texts. Notwithstanding, non-function words that may in fact be relevant to the document content tend to occur with greatly varying frequencies in the different texts of a collection. Furthermore, the frequency of occurrence of non-function words may actually be used to indicate the term’s importance as representative of the document’s content. [15].

To conduct the tests, we developed a fully automated software tool, TextAnalysis, to analyze the texts of the documents chosen for the study. The program’s output is a set of significant words from the document which effectively constitute the subject of the document by presenting its topics. The program “reads” the document text and uses statistical analysis to determine the most frequently-occurring words. The text is analyzed to eliminate (using a Stop List and after limited processing of linguistic elements) from the results list words that are meaningless in terms of representing the document’s subject. TextAnalysis can process English, Hebrew, and bilingual documents in both languages. It can process the statistical elements in any language, but is less successful at processing the linguistic elements.

The limited linguistic processing recognizes language inflections as well as the insignificant words that must be filtered out by means of a special Stop List specific to each language (see figure 1). In our study, we used the program to specify the subject of a document by using its most relevant words. Our study examined whether these words offer an accurate definition of the document’s subject.

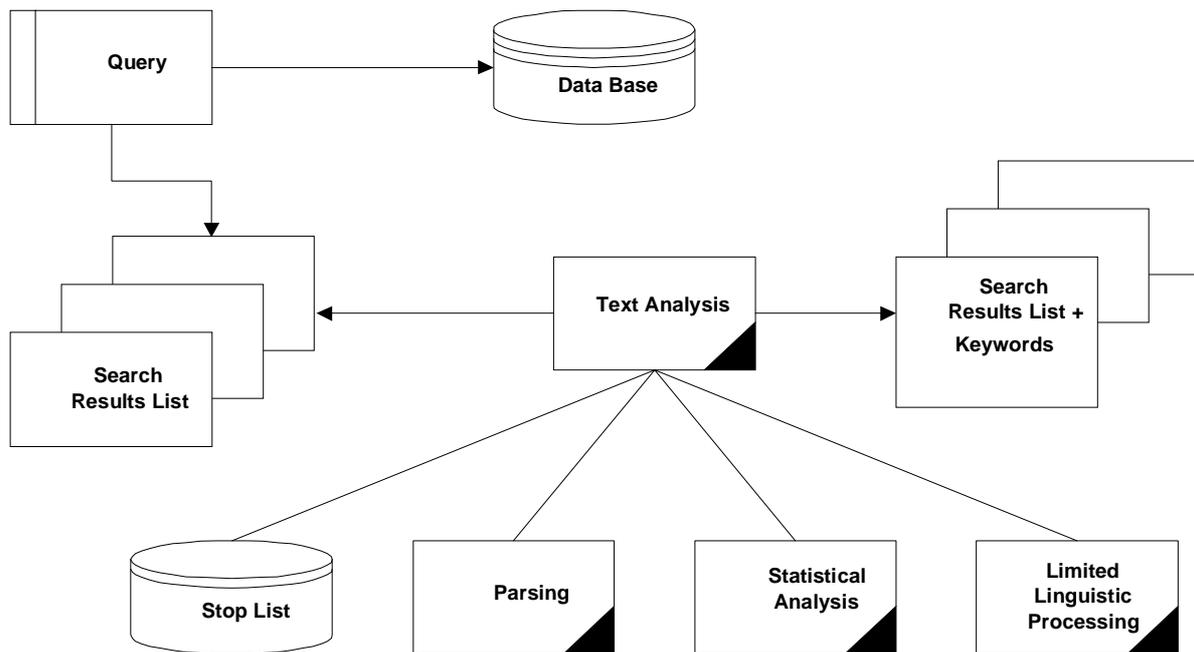


Figure 1: Processing documents using text analysis

Table 6 presents sample results of automatically-generated keywords.

### THE STUDY

To determine the extent to which document analysis performed by TextAnalysis can reliably define the document’s subject, we chose two sets of scientific documents in which the

subject could be derived directly from the keywords and title. Our study assumed that the words in the document's title and its keywords (defined by the author) offered an accurate account of its content. TextAnalysis' efficiency would be measured by comparing the document subject it defined with the document subject as defined by the author in the title and keywords. The test included a full-text analysis of 200 scientific documents in two subject categories: geography and family studies. See Appendix 1 for a list of the journals from which the articles were culled.

An alternate validation of the data was obtained by assigning two persons with expert knowledge to read the articles. These validators were then asked to assess how helpful the significant words designated by TextAnalysis were in reliably determining the document's subject.

TextAnalysis rated words that appeared more than 20 times in an article (average number of words per article: 7,611) as significant. In most cases (93.7%), the 10 most frequently-occurring words were words that appeared more than 20 times. In other cases (6.3%), however, some of the words ranked in the list of 10 most-frequently appearing words appeared less than 20 times. For the purposes of the test, the study examined three word groupings: the 3 most frequently-occurring words in the list, the 5 most frequently-occurring in the list, and the 10 most frequently-occurring words in the list.

## FINDINGS

Each subject category was analyzed separately to determine whether TextAnalysis performed differently on texts in different subject categories (see Table 1).

For documents in the subject category of geography, 53.9% of the previously-designated keywords and 55.6% of the words that appeared in the document titles were among the 10 most frequently-occurring words identified by TextAnalysis. In addition, 45.1% of the index terms (based on words added by the publisher) were included in TextAnalysis' list of the 10 most frequently-occurring words. The findings for the 5 most frequently-occurring words showed that the software identified 46.2% of the keywords, 45.5% of the words appearing in the document title, and 37.7% of the index terms.

	Geography (SD)	Family studies (SD)
% Keywords identified	53.9 (0.20)	52.4 (0.24)
% Title words identified	55.6 (0.20)	47.2 (0.18)
% Index terms identified	45.1 (0.21)	44.8 (0.21)
% Keywords by 5 most frequently-occurring words identified	46.2 (0.26)	42.8 (0.23)
% Title words by 5 words most frequently-occurring identified	45.5 (0.19)	40.6 (0.22)
% Index terms by 5 most frequently-occurring words identified	37.7 (0.22)	40.0 (0.19)

Table 1: Summary of TextAnalysis' identification rates (analyzing the text of the articles) reflects the keywords and titles of the articles

For documents in the subject category of family studies, 52.4% of the previously-designated keywords, 47.2% of the words appearing in the document title, and 44.8% of the index terms

(based on words added by the publisher) were included in TextAnalysis' list of the ten most frequently-occurring words. The findings for the 5 most frequently-occurring words showed that TextAnalysis identified 42.8% of the keywords, 40.6% of the words appearing in the document title, and 40.0% of the index terms.

A comparison of the most frequently-occurring words indicated that there was no significant difference between the rate of identification of the 10 most frequently-occurring words and the 5 most frequently-occurring words ( $P < 0.0001$ ). There was, however, a significant difference in the rate of identification of the 10 and the 5 most frequently-occurring words compared with the 3 most frequently-occurring words ( $P < 0.0001$  for each).

	% subject identification by evaluator using 3 most frequently-occurring words (SD)	% subject identification by evaluator using 5 most frequently-occurring words (SD)	% subject identification by evaluator using 10 most frequently-occurring words (SD)
Geography	53 (0.50)	78 (0.42)	83 (0.38)
Family studies	70 (0.46)	88 (0.32)	91 (0.29)
Overall	61.5	83	87

Table 2: Summary of the rate of effective identification of the article's subject based on TextAnalysis' list of the 3, 5, and 10 most frequently-occurring words, respectively

Since TextAnalysis is designed to automatically identify the document's subject by analyzing the text, its results were validated by two persons with expert knowledge. After reading the articles, these validators were asked to determine, by indicating "yes" or "no," whether the terms identified by TextAnalysis reliably identified the subject of the article. Table 2 demonstrates that TextAnalysis correctly identified the subject in more than 75% of cases using the 5 most frequently-occurring words (78% for geography articles, 88% for family studies articles). The identification rate declined when TextAnalysis attempted to define the subject based on only the 3 most frequently-occurring words.

There was a significant difference between identification based on 5 or 10 words and identification based on 3 words ( $P < 0.0001$ ) (see Table 2).

The study also investigated whether the overall length of the documents influenced the degree of effective identification. The findings indicate that for both subject categories, geography and family studies, the longer the text, the lower the rate of effective identification of the keywords, words in the document title, and index terms.

The shortest geography article contained 598 words, and the longest contained 14,792 words. The average word count was 5,686. Table 3 demonstrates how the length of the article affects the rate of identification of keywords, title words, and index terms.

The shortest family studies article contained 5,488 words, and the longest contained 19,354 words. The average word count was 9,537. Table 4 demonstrates how the length of the article affects the rate of identification of keywords, title words, and index terms.

Number of words in document (NWD)	NWD <5,686	NWD >5,686
% keywords identified (10 most frequently-occurring words)	57.3	50.1
% title words identified (10 most frequently-occurring words)	55.6	55.6
% index terms words identified (10 most frequently-occurring words)	48.9	40.8
% keywords identified (5 most frequently-occurring words)	49.9	41.8
% title words identified (5 most frequently-occurring words)	46.1	37.1
% index terms words identified (5 most frequently-occurring words)	42.3	32.3

Table 3: Summary of identification rates based on article size (geography)

Number of words in document (NWD)	NWD <9,537	NWD >9,537
% keywords identified (10 most frequently-occurring words)	54.7	48.7
% title words identified (10 most frequently-occurring words)	50.2	42.2
% index terms words identified (10 most frequently-occurring words)	47.7	40.0
% keywords identified (5 most frequently-occurring words)	44.1	39.8
% title words identified (5 most frequently-occurring words)	44.2	34.3
% index terms words identified (5 most frequently-occurring words)	41.3	37.7

Table 4: Summary of identification rates based on article size (family studies).

As we found in our earlier study [7a], there is no significant difference in the rate of identification of the subject of the articles based on the 10 most frequently-occurring words or the 5 most frequently-occurring words. Inasmuch as our study sought to identify the subject of documents using the minimal number of words, Table 5 demonstrates the rate of identification of the subject based on the 5 most frequently-occurring words and the 3 most frequently-occurring words in articles in 4 subject categories.

	% subject identification by evaluator using 5 most frequently-occurring words	% subject identification by evaluator using 3 most frequently-occurring words
General Management	69	53
Industrial Management	74	40
Geography	78	53
Family studies	88	70
Overall	77	54

Table 5: Summary of identification rates of the subject of articles based on TextAnalysis' designation of the 5 most frequently-occurring and 3 most frequently-occurring words, respectively

In Table 6, we can see sample results of automatically-generated topics. The first column lists keywords automatically generated by TextAnalysis, and the following columns list the original document titles and keywords. The first five rows are for articles in the subject category of geography, and the next five rows are for articles in the subject category of family studies.

Automatically-Generated Keywords	Sample Document Titles	Sample Document keywords
Geography Australian Volume Journal Papers Issue published	Development of a journal: Some brief comments relating to the part index for the Australian Geographer, 1928-98	Geography Australia Journals publishing
Unemployment Labour Party Political Contextual Economic Zealand Minor Voters electoral	Social democracy and contextual unemployment: New Zealand labour in 1990	New Zealand Context Unemployment Labour politics social democracy electoral reform
Grey Exploration Bushmanship Experience Australia Competence Barrow land-based John	Bushmanship: The explorers' silent partner	Bushmanship Exploration George Grey Edward John Eyre Australian History
Tourism Meanings Visitors Research Daintree Tribulation Cape travel	The social construction of tourist places	Tourism place meanings place promotion social constructions quantitative assessment Daintree and Cape Tribulation area Far North Queensland
Fisheries Resources Shetland Management Lau Coastal sustainable	Contrasting approaches to the management of common property resources: An institutional analysis of fisheries development strategies in Shetland and the Solomon Islands	Scotland Shetland Islands Solomon Islands Malaita Lau community fisheries policy resource management sustainable resource bases

Parent Training Behavior Child Program Evaluation Family Children	Guidelines for evaluating parent training programs	parent training program evaluation CIPP model
Family Ethics Scientists Education Principles Guidelines Students Professional NCFR	The development and teaching of the ethical principles and guidelines for family scientists	ethics education ethical principles and guidelines National Council on Family Relations professional ethics
child adoption parents mothers self- reflectiveness scores	Adjustment among adopted children: The role of maternal self-reflectiveness	Adjustment Adoption externalizing symptoms self-reflectiveness
Family Community Model Public Knowledge Professional Parent Life researchers	Family science and family citizenship: Toward a model of community partnership with families	family and community marriage education parenting parent education
Family Sibling Fairness Farm Conflict Transfer Agreement relationships	Sibling relationships, fairness, and conflict over transfer of the farm	Fairness family conflict family farm farm transfer intergenerational relations sibling relationships

Table 6: Sample results of automatically-generated topics

## CONCLUSIONS

The study's findings can lead to several conclusions.

1. The 5 most frequently-occurring words identified by the TextAnalysis software tool reliably reflect the subject of a document up to 88% of the time (according to the expert opinions of the validators and according to the subject of the document).
2. The longer the document, the lower the rate of TextAnalysis' effective identification of its subject.
3. The rate for keyword identification is slightly higher, although not significantly so, than the rate for identification of words appearing in the document title. On a practical level, many documents lack designated keywords, whereas most documents do have titles. Since there is no significant difference in the respective identification rates for keywords and words

appearing in the document title, the latter can be used in cases where there are no keywords.

4. Identification rates differ by type of material. For the subjects examined in our study, the rate of identification of the subject was lower for articles in the subject category of geography articles than for articles in the subject category of family studies.
5. It would be beneficial for search engine users if administrators of full-text databases lacking categories or keywords would locate and register both using automated tools such as TextAnalysis.

### **SUMMARY AND RECOMMENDATIONS FOR FURTHER RESEARCH**

The purpose of this study was to examine whether automated software tools can be used to identify the subject of a document, to be displayed along the document title in a search results list. A software tool, TextAnalysis, was developed to perform statistical analyses on words in a given text and generate a list of significant words meant to represent the document's subject.

Two hundred scientific articles were examined in two specific subject categories. TextAnalysis' rates of effective identification were calculated by the degree to which the most frequently-occurring words it identified conformed with previously-designated keywords and words contained in the document's title.

The length of a document may influence the rate of effective identification of the document's subject.

The current study challenges software developers to increase the rates of successful identification of a document's subject. We also noted several comments for improvements to the software. These include the processing of expressions, acronyms, meaningless words that were weighted in the calculation of the results, etc. We will also attempt to improve the algorithm in order to refine TextAnalysis' performance. Another area that warrants further study is the ability to define the subject of a document based on only two to three words out of the 5 most frequently-occurring words, taking into account specific professional terminology. The collection of articles used for this study was homogeneous for two subject categories. Further research can be done on larger collections in more subject categories.

### **ACKNOWLEDGEMENT**

I would like to thank to Michal Katz and Ruth Chelouche for their help in the study.

### **REFERENCES**

1. ACM Digital library - <http://www.acm.org/dl/>
2. Allen, R., Two digital library interfaces that exploit hierarchical structure, *Proceedings of DAGS95: Electronic Publishing and the Information Superhighway*, (Boston, May-June 1995).
3. Chekuri, C. et al., Web search using automated classification. *Sixth International World Wide Web Conference*, (Poster no. POS725), Santa-Clara, CA, 1997.
4. Chen, H., Dumais, S., Bringing order to the web: automatically categorizing search results. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'00)*, ACM, 2000, 145-152.
5. Drori, O., Using Text Elements by Context to Display Search Results in Information Retrieval Systems. *Information Doors - Where Information Search and Hypertext Link (a workshop proceedings held in conjunction with the ACM Hypertext 2000 and ACM Digital Libraries 2000 conferences)*, San Antonio, Texas, USA, 2000, 17-22.

<http://shum.huji.ac.il/~offerd/papers/drori052000.pdf>

6. Drori, O., Improving Display of Search Results in Information Retrieval Systems - User's Study. *Technical Report of the Leibniz Center for Research in Computer Science*, No. 2000-34, 2000, Jerusalem.

<http://shum.huji.ac.il/~offerd/papers/drori082000.pdf>

7. Drori, O., The Benefits of Displaying Additional internal Document Information on Textual database Search Results Lists, *Proceedings of the 4<sup>th</sup> European Conference on Research and Advanced Technology for Digital Libraries – ECDL2000* (Lisbon, Portugal, September 2000), LNCS 1923, Springer, 2000, 69-82.

<http://shum.huji.ac.il/~offerd/papers/drori092000.pdf>

- 7a. Drori, O. (b). "Using Frequently Occurring Words to Identify the Subject of a Document", *Technical Report No. 2001-7 of the Leibniz Center for Research in Computer Science*, School of Computer Science and Engineering, Hebrew University of Jerusalem, May 2001.

<http://shum.huji.ac.il/~offerd/papers/drori052001.pdf>

8. Hearst, M., Pedersen, J., Reexamining the cluster hypothesis: Scatter/Gather on retrieval results, *Proceedings of 19<sup>th</sup> annual international ACM/SIGIR conference (Zurich, Switzerland, August 1996)*, ACM Press, 1996, 76-84.

9. Korfhage, R., *Information Storage and Retrieval*, N.Y.: John Wiley, 1997.

10. Maarek, Y. et al., WebCutter: a system for dynamic and tailorable site mapping. *Proceedings of the 6<sup>th</sup> International World Wide Web Conference*, Santa-Clara CA, 1997.

11. Marchionini, G. et al., Interfaces and tools for the Library of Congress national digital library program. *Information Processing and Management*, 34, 1998, 535-555.

12. Mladenic, D., Turning Yahoo into an automatic web page classifier, *Proceedings of the 13<sup>th</sup> European Conference on Artificial Intelligence (ECAI'98)*, Brighton, UK: ECCAI Press, 1998, 473-474.

13. Landauer, T. et. al., enhancing the usability of text through computer delivery and formative evaluation: the SuperBook project. *Hypertext - A Psychological Perspective*, New York: Ellis Horwood, 1993, 71-136.

14. Sahami, M., Yusufali, S., Baldonado, M., SONIA: A Service for Organizing Networked Information Autonomously, *Proceedings of ACM Digital Libraries '98* (Pittsburgh, PA, USA), ACM Press, 1998, 200-209.

15. Salton, G., *Automatic text processing: the transformation, analysis, and retrieval of information by computer*, Reading, Massachusetts: Addison-Wesley, 1989, 279-281.

16. Searchenginewatch - <http://www.searchenginewatch.com>

17. Yahoo! - <http://www.yahoo.com>
18. Zamir, O., Etzioni, O., Web document clustering: a feasibility demonstration. Proceedings of the 19<sup>th</sup> International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR98), 1998, 46-54.
19. Zamir, O., Etzioni, O., Grouper: A dynamic clustering interface to web search results. *Proceedings of WWW8*, Toronto, Canada, 1999.

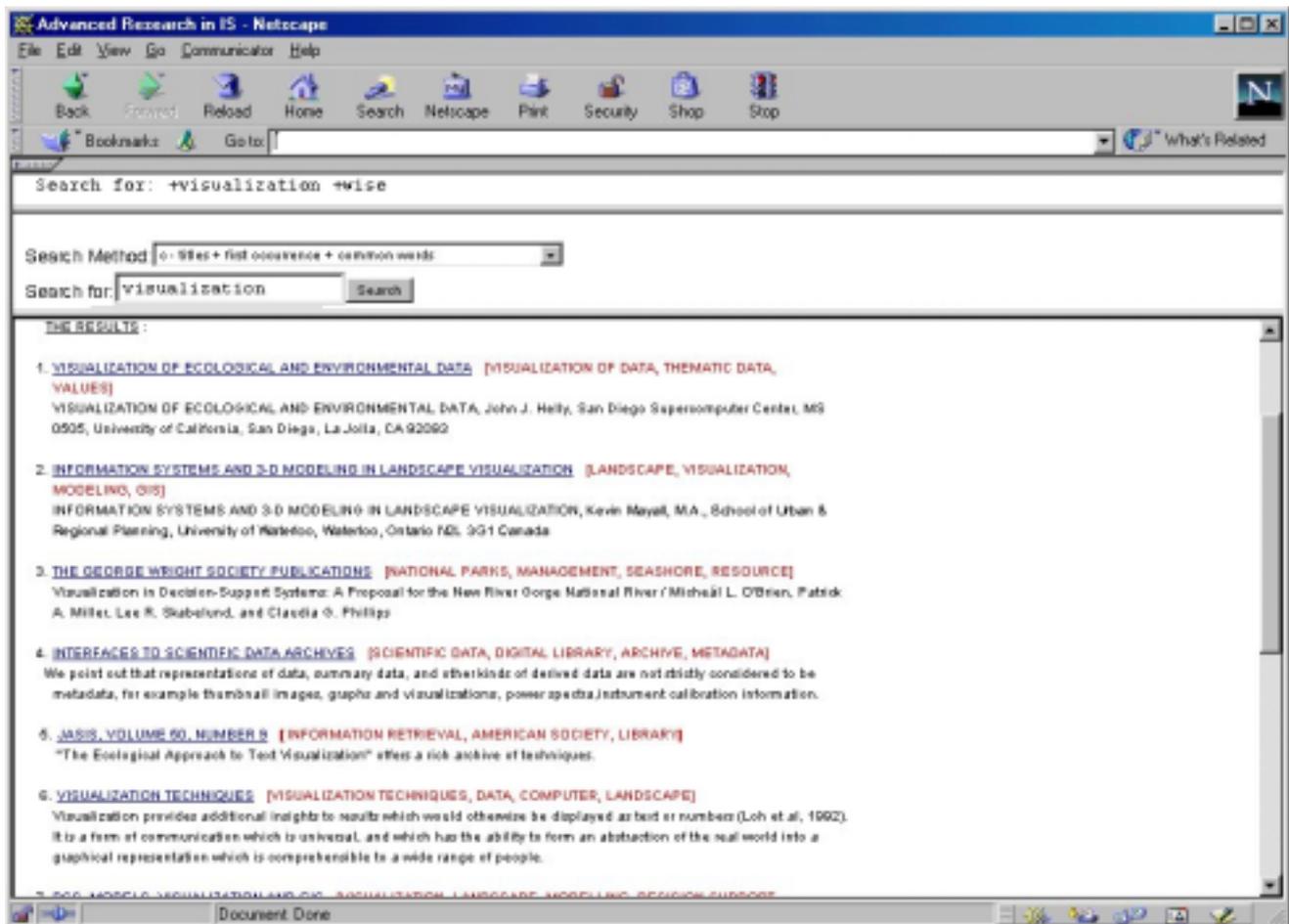
#### **Appendix 1 - List of journals from which articles were taken for this study**

##### **Geography**

Economic Geography  
Journal of Geography in Higher Education  
Canadian Geographer  
Australian Geographer

##### **Family studies**

Family Relations  
Journal of Marriage and the Family



Appendix 2 – The study's user interface