Using Text Analysis to Inform Clients of the Subject of a Document

Offer Drori
Hebrew University of Jerusalem and SHAAM – Information Systems, Jerusalem, Israel

offerd@cs.huji.ac.il

Abstract

Contemporary information databases contain many millions of electronic documents. Locating information on the Internet today is problematic, due to the enormous number of documents it contains. Several other studies have found that associating documents with a subject or list of topics can improve locatability of information on the Internet (Drori, 2000a 2000b 2000c). Effective cataloguing of information is performed manually, requiring extensive resources. Consequently, most information is currently not catalogued. This paper aims to present a software tool that automatically locates the subject of a document and to show the results of a test performed, using the software tool, *TextAnalysis*, specially developed for this purpose. The main purpose of this study is to inform clients of the subject of the corpus of texts it obtains from search engines as a search results list.

Keywords: Frequently occurring words, Web documents classification, Search results list, Identify topics of documents

Introduction

SearchEngineWatch reports that there are over one billion pages of information on the Web, excluding the information contained in hosted databases. Because of the sheer quantity of information, and the huge resources required for cataloguing, it can hardly be expected that any significant proportion of this information base will be catalogued. In a world where the quantity of information is growing at a rapidly increasing pace, automatic cataloguing of information and documents is a necessity, and vital for locating information.

Cataloging information comprises associating information with a list of predefined subjects or concepts. Librarians pioneered the cataloguing of human knowledge, developing methods to associate a book collection (and subsequently other media) by subject. The most widely used and long established cataloguing systems are the Dewey classification system (which divides subjects into categories by decimal numbers) and LCSH (Library of Congress Subject Headings, which is based on a list of terms and is the method used to catalog the U.S. Library of the 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 (Yahoo!), which manually catalogs large numbers of pages on the Internet. Because the cataloguing is manual, however, the coverage offered by the Yahoo! search engine is very limited.

Several studies have been undertaken on the automatic organizing of information. Most of them deal with the Internet and the display of information it contains. In 1995, a prototype was developed to display search results using the Dewey method (Allen, 1995). In the Superbook project (1993), text paragraphs were arranged in a table of hierarchical content, similar to a table of contents (Landauer, 1993). A table of contents, resembling searches in the Library of Congress, was also created by Marchionini (Marchionini et al., 1998). In the WebCutter system (Maarek et al., 1997), the user search map is category-based.

Clustering is another means of automatically organizing a group of documents. In clustering, groups of documents are arranged on the basis of similarities rather than on a predefined set of categories. There are a number of projects underway in this area, but since most barely touch on the issue of cataloguing, which is the crux of this article, only a few references are mentioned here: (Zamir & Etzioni, 1998) (Zamir & Etzioni, 1999), (Hearst & Pedersen, 1996), (Sahami et al., 1998).

Classification is a third way to organize documents into groups. This method 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-catalogued. This method is less relevant to the contents of this article and so here, too, only a few sources are mentioned: (Chekuri et al., 1997), (Mladenic, 1998), and (Chen & Dumais, 2000).

The fundamental problem of computerized textual information management is automatic natural processing. There are three basic approaches for textual documents processing (Korfhage, 1997): lexical, syntactic, and semantic analysis. The syntactic approach seeks to enable the computer to understand a natural language sentence structure, while the semantic analysis attempts to identify the semantic structure of a document and thus to discover its meaning. The most popular and effective so far is the lexical text analysis method. The aim of this method is usually to determine the most important terms, the "keywords," which characterize each document, documents group, or topic, using different statistical techniques based on how frequently the terms appear in a text. Once the keywords list for existing documents in search results list has been generated, the documents related to its topics may be found and displayed.

Locating the Subject of the Article

A series of experiments to determine the most effective means 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 (Drori, 2000a 2000b 2000c). 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 actually having to read all the documents in the list. Displaying the subject of each document in the search results list enables the user to focus only on the documents that meet the defined subject of interest. Other conclusions from the study (Drori, 2000c) where:

1. The addition of keywords to the information displayed in the list of the search results will reduce the search time, as opposed to information displayed without keywords.
2. The addition of keywords to the information displayed in the list of search results will improve the user's feeling of ease, as opposed to the same information displayed without keywords.
3. The addition of keywords to the information displayed in the list of search results will improve the user's feeling of satisfaction, as opposed to the same information displayed without keywords.

The subject of a document may be a category from a pre-defined list of categories, or some words that list the topics of the document or list of keywords. In our study, we chose to focus on key words as part of previous studies and for several other reasons that will be addressed later.

There are several ways of locating the subject of a document. The most accepted method is based on manual characterization of the document according to different categories. In the case of scientific documents, the document 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). On the Internet, the directory staff adds the category with which the document is associated (see Yahoo! for example) to computerized directories. In addition to these manual, relatively accurate but resource-intensive methods, a computerized system of characterization is required to catalog the extensive body of documents from different sources, including documents that have not been indexed or catalogued.

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 realtime, automatically retrieves, parses, and organizes these documents into categories (Sahami et al. 1998).

On the Internet as we know it today and in the absence of specific sample queries and user relevance assessments, little is known *a priori* about what renders an item relevant or irrelevant. Instead of concentrating on the relevance or nonrelevance of particular documents, it seems reasonable first to consider the 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 of a collection. Moreover, most function words are characterized by high occurrence frequencies in ordinary texts. On the other hand, nonfunctional words that may actually relate to document content tend to occur with greatly varying frequencies in the different texts of a collection. Furthermore, the frequency of occurrence of nonfunctional words may actually be used to indicate term importance for content representation (Salton, 1989).

To conduct the tests, a fully automated software tool was developed to analyze the texts of the required documents. The output of the tool is a set of significant words from the document, which effectively constitute the subject of the document by presenting its topics. The software tool “reads” the document text and uses statistical analysis to determine the most frequently-occurring words in the document. The text is analyzed after words that are meaningless for the subject are eliminated using a Stop List, and after limited processing of linguistic elements. The software tool can process both English and Hebrew documents, as well as documents containing both languages. It is able to process the statistical elements in any language, but is less successful in processing the linguistic elements.

The limited linguistic processing recognizes the inflections in the language, 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 software as a tool for setting the subject of a document by using its most relevant words. Our research examined whether those words can give an accurate definition of the document.

The Study

To determine the extent to which document analysis by the software can reliably define the document subject, four sets of scientific documents were chosen in which the subject was directly available from the document keywords and document title. The research assumption was that a title and keywords of a document (defined by the author), could give a good impression of its content. The success of the software tool would be measured by comparing the document subject as defined by the software tool with the document subject as defined by the author using his title and keywords. The test included full text analysis of 300 scientific documents from four areas of research 1) General management 2) Industrial management, 3) Geography, 4) Family study. For a list of the journals from which the articles were culled, see Appendix 1.

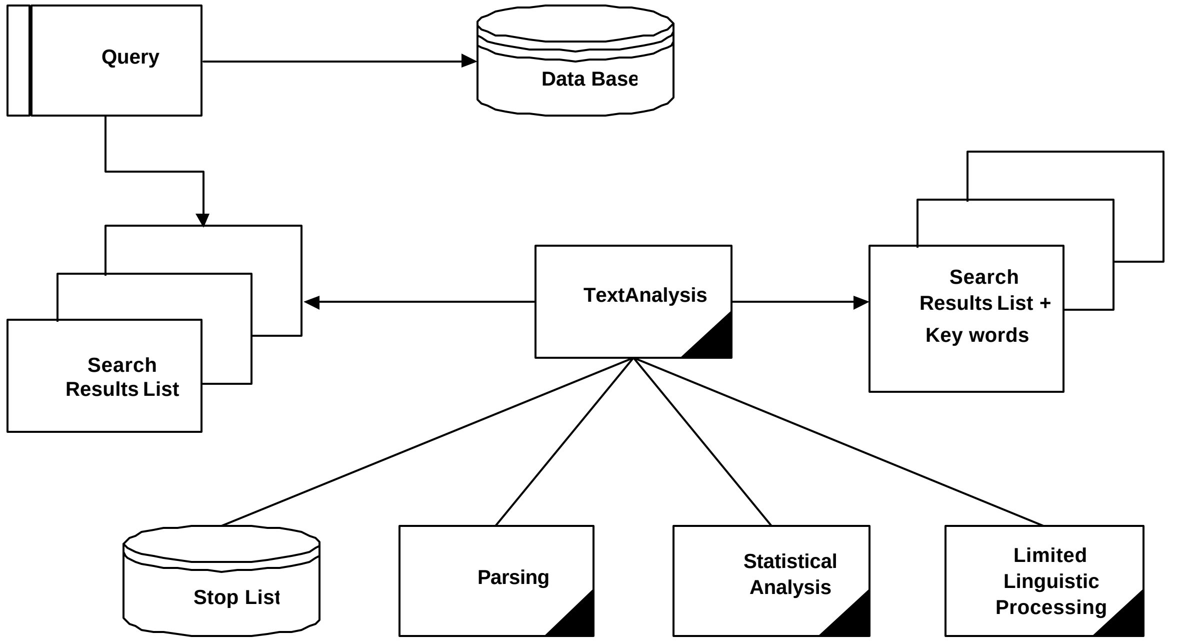

Figure 1: Processing documents using TextAnalysis

A comparison was made between the significant words identified by the software tool and those in the two existing sets of words, namely keywords defined by the document's author and words appearing in the document title (also intended to more explicitly characterize the document). Successful prediction of the document subject by the software tool would be reflected in a high degree of conformance between the terms selected by the software tool and the terms contained in the document (keywords or words in the document title). Sample results on automatically generated topics can be seen at table 7.

To validate the data in another way, four people with skills relevant to the content of the articles were assigned to read them. Based on the significant words selected by the software tool, they then had to assess to what level they were able to determine the document's subject after reading the actual document.

Words rated significant by the software tool were words that appeared more than 20 times in an article (average number of words per article: 6000). In most cases (93.7%) the top ten words where words occurring more than 20 times, in other cases (6.3%) we used the top ten words even though they occurred less than 20 times. For the purposes of the test, the study examined three word groupings: the top three words in the list, the top five words in the list (the top three, plus another two words), and the top ten words in the list.

Findings

Each scientific sphere was analyzed separately, to determine whether texts from different spheres behave differently (see Table 1).

In the case of General Management documents, 50.4% of the keywords assigned to the articles were contained in the top ten words preferred by the software tool. 49.7% of the words from the titles of the articles were contained in the top ten words culled by the software tool. When keywords and titles were combined, and duplicate words filtered out, the identification rate was 43.9%. The findings for the top 5 frequent words were a 56.4% identification rate, and for the top three frequent words, 70.6% identification rate comparing to the combined words of the titles and keywords.

In the case of Industrial Management documents, 46% of the keywords assigned to the articles were contained in the top ten words preferred by the software tool. 46.7% of the words from the titles of the articles were contained in the top ten words culled by the software tool. When keywords and titles were

	General Management	Industrial Management	Geography	Family studies
% Keywords identified (SD)	50.4 (0.27)	46.0 (0.23)	53.9 (0.20)	52.4 (0.24)
% Title words identified	49.7 (0.21)	46.7 (0.21)	55.6 (0.20)	47.2 (0.18)
% Index terms words identified	-	-	45.1 (0.21)	44.8 (0.21)
% Rate by top 5 key-words identified	56.4 (0.22)	55.9 (0.22)	46.2 (0.26)	42.8 (0.23)
% Rate by top 5 title words identified	55.2 (0.22)	54.8 (0.22)	45.5 (0.19)	40.6 (0.22)
% Rate by top 5 index terms words identified	-	-	37.7 (0.22)	40.0 (0.19)

Table 1: Summary of the identification rates by the TextAnalysis software.

combined, and duplicate words filtered out, the identification rate was 40.8%. The findings for the top 5 frequent words were a 55.9% identification rate, and for the top three frequent words, 64.1% identification rate comparing to the combined words of the titles and keywords.

When comparing frequently occurring words, there was a significant difference between the identification rates by the top 10 frequent words compared with the top 5 frequent words (P<0.0001), and between the top 10 frequent words compared with the top 3 frequent words (P<0.0001). There was also a significant difference between the identification rates by the top 5 versus the top 3 frequently occurring words (P<0.0001).

	% Subject identification by evaluator using 5 top words	% Subject identification by evaluator using 3 top words
General Management	68.6	52.9
Industrial Management	74.0	40.0
Geography	78.0	53.0
Family studies	91.0	88.0
Overall	78	58

Table 2: Summary of identification rates for the article subject based on the TextAnalysis tool's top 3 and top 5 words, respectively

Since the software tool is intended to identify the document subject automatically by analyzing the text, the software results were validated by two persons with the skills appropriate to the content of the articles. They were asked whether they felt that the subject of the article could be understood on the basis of the software tool's recommendation. After reading the articles the referees gave their opinion by providing a yes/no answer for each article ('yes' – for understanding the subject of the article based on the tool's recommendation). Table 2 illustrates that the subject of the document was identified in more than 70% of cases using the 5 top words (74% in industrial management, 69% in general management). The identification rate declined when attempted on the basis of only 3 words.

There was a significant difference between identification based on 5 words and identification based on 3 words (P<0.0003), see Table 2.

The study also investigated whether the size of the documents influenced identification. In the case of General Management documents, the shortest article contained 1,537 words and the longest ran to 16,001 words.

The average word count was 6,324. As Table 3 demonstrates, the greater the number of words in the text, the higher the rate of identification of the keywords (57% vs. 45%), and of combined words (keywords + title words) by the top 5 and top 3 words, respectively, culled by the TextAnalysis tool. It is also evident that the identification rates decline for title words (50.9% vs. 48.3%) and when using the top 10 words (48% vs. 42.3%) culled by the software tool. The decline in the identification of the title words and the top 10 words can be explained by the title not being sufficiently informative (for example, one of the articles was entitled “Can Elephants Fly?”).

In the case of Industrial Management documents, the shortest article contained 2,294 words, and the longest ran to 9,734 words. The average word count was 4,915. As Table 4 demonstrates, the greater the number of words in the text, the higher the identification rates of the key words (47.1% vs. 45.2%), of words in the title (48.3% vs. 45.5%), and of the top 5 and top 3 words (56.4% vs. 55.5% and 68.2% vs. 60.7%) proposed by TextAnalysis.

The shortest geography article contained 598 words, and the longest ran to 14,792 words. The average word count was 5,686. Table 5 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 ran to 19,354 words. The average word count was 9,537. Table 6 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 <6,324	NWD >6,324
% Keywords identified (SD)	44.7 (0.26)	57.4 (0.29)
% Title words identified	50.9 (0.21)	48.3 (0.20)
Combined rate (keywords + title words)	40.9 (0.18)	41.4 (0.18)
% Of top 10 words identified	48 (0.19)	42.3 (0.17)
% Of top 5 words identified	55.4 (0.22)	57.7 (0.23)
% Of top 3 words identified	70.7 (0.28)	71 (0.28)

Table 3: Summary of identification rates based on article size (General Management).

Number of words in document (NWD)	NWD <4,915	NWD >4,915
% Keywords identified (SD)	45.2 (0.23)	47.1 (0.23)
% Title words identified	45.5 (0.20)	48.3 (0.23)
Combined rate (keywords + title words)	48.8 (0.19)	38.7 (0.18)
% Of top 10 words identified	44.7 (0.15)	43.6 (0.15)
% Of top 5 words identified	55.5 (0.22)	56.4 (0.22)
% Of top 3 words identified	60.7 (0.24)	68.2 (0.28)

Table 4: Summary of identification rates based on article sizes (Industrial Management).

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 5: 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 6: Summary of identification rates based on article size (family studies).

Automatically Generated Key words	Sample Document Titles	**Sample Document keywords**
Purchasing International Factor Study Results Table	International purchasing practices of US and Indian managers: a comparative analysis	India International Trade Purchasing USA
Quality TVS Management Team Firm	Maintaining quality through evolving strategy: the TVS Partnership	Quality Awards Australia Professional Services
Quality Program Company Management Project Process	Practical experience with quality improvement in small companies	Kaizen Problem Solving Process Management Quality
Product Selling Performance Salespeople Related Market	Determinants of new product selling performance: an empirical examination in The Netherlands	New Product Launch Netherlands Performance Sales Management
Strategic Transformation Organization Marks Spencer Change	Developing skills in strategic transformation	Strategic Planning Organizational Change Management Techniques
Fire Safety Management Passenger Risk Terminals	Fire safety management at passenger terminals	Fire Public Safety Transport
Construction ISO Environmental EMS Sustainable Development	A framework for implementing ISO 14000 in construction	Construction Industry Environmental Impact Sustainable Development
Consulting Management Strategic Capabilities Success	Strategic capabilities which lead to management consulting success in Australia	Management Consultants Success Australia

Table 7: Sample results on automatically generated topics

Conclusions

Several conclusions can be drawn from the findings of the study.

1. The TextAnalysis tool can be used to identify the subject of a document in up to 91% of cases (according to expert opinion), using the tool's five top rated words. We used keywords to identify the subject of a document
2. The greater the number of words in the document, the higher the rate of keyword identification.
3. The identification rate for keywords and title words combined is lower than for either category independently.
4. The rate for keyword identification is slightly higher than for title words. Practically speaking, there are many documents that do not have keywords registered for them, whereas most documents do have titles. Since there is no significant difference in the identification rates for keywords and title words, respectively, title words can be used in the absence of keywords.
5. Identification rates differ by type of material. For most of the subjects examined in our case, identification rates were lower for Industrial Management and General Management than Geography and Family Studies.
6. It is recommended that the administrators of full text databases without categories or key words locate and register categories/keywords using automated tools (such as TextAnalysis) for the benefit of search engine users in the search process.

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 with its title in a search results list. In this study, 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 subject.

The use of articles from data bases was employed to examine the effectiveness of the tool, based upon additional data from the data base such as keywords, titles, etc. which were compared to the automatic analysis by the tool. The analysis tool will be used for free texts found on the Internet, for the purpose of effective display of search results.

Three hundred scientific articles were examined in four specific areas, and the software tool's identification rate was calculated by the degree of conformance to the keywords and title words of the respective article.

Document length is one parameter that might influence identification rates.

The current study challenges software developers to increase identification rates. Several comments were also noted for improvements to the software, including the processing of expressions, acronyms, meaningless words that were weighted, etc. An attempt will also be made to improve the algorithm in order to improve software performance. Also warranting further research is the ability to define the subject of a document based on only two to three words from the five most significant words, taking into account specific professional terminology. The collection studied was homogeneous for two subjects. Further research can be done on larger collections on different subjects.

Acknowledgment

I would like to thank Shirley Sharvit, Rachel Moffie, Michal Katz and Ruth Chelouche for their help in the research.

References

ACM Digital library - http://www.acm.org/dl/

Allen, R. (1995). Two digital library interfaces that exploit hierarchical structure. *Proceedings of DAGS95: Electronic Publishing and the Information Superhighway*, (Boston, May-June 1995).

Chekuri, C. et al. (1997). Web search using automated classification. *Sixth International World Wide Web Conference*, (Poster no. POS725), Santa-Clara, CA.

Chen, H., Dumais, S. (2000). Bringing order to the web: automatically categorizing search results. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'00)*, ACM, 145-152.

Drori, O. (2000a). 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, 17-22. Available at http://shum.huji.ac.il/~offerd/papers/drori052000.pdf

Drori, O. (2000b). 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, Jerusalem. Available at http://shum.huji.ac.il/~offerd/papers/drori082000.pdf

Drori, O. (2000c). The Benefits of Displaying Additional internal Document Information on Textual database Search Results Lists, *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries – ECDL2000* (Lisbon, Portugal, September 2000), LNCS 1923, Springer, 69-82. Available at http://shum.huji.ac.il/~offerd/papers/drori092000.pdf

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

Korfhage, R. (1997). Information Storage and Retrieval, N.Y.: John Wiley.

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

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

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

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

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

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

Searchenginewatch - http://www.searchenginewatch.com

Yahoo! - http://www.yahoo.com

Zamir, O., Etzioni, O. (1998). Web document clustering: a feasibility demonstration. Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR98), 46-54.

Zamir, O., Etzioni, O. (1999). Grouper: A dynamic clustering interface to web search results. *Proceedings of WWW8*, Toronto, Canada.

Biography

Offer Drori, Ph.D. in computer science. Homepage: http://shum.huji.ac.il/~offerd/, E-mail: mailto:offerd@cc.huji.ac.il

Lecturer at the School of Business Administration, School of Engineering and Computer Science of The Hebrew University of Jerusalem.

Head of information database systems, SHAAM Information Systems.

Chairman, Special Interest Group on Text Retrieval Systems (SIGTRS) - http://sigtrs.huji.ac.il

Area of professional engagement and research: Online information systems, system analysis and design, large information databases, information retrieval, human-computer interaction.

Membership: ACM (Association for Computing Machinery), Israel Association of System Analysts, IPA (Information Processing Association of Israel)

Appendix 1 - List of Journals from which Articles were taken for this Study

General management

Innovation Management European Journal of Management Decision
European Business Review
Accounting Auditing & Accountability Journal
Disaster Prevention and Management
International Journal of Entrepreneurial Behaviour & Research
Environmental Management and Health
Journal of Management History

Industrial management

Industrial Management & Data Systems
Integrated Manufacturing Systems
Logistics Information Management
International Journal of Operations & Production Management
International Journal of Agile Management Systems
International Journal of Service Industry Management
International Journal of Manpower

Geography

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

Family studies

Family Relations
Journal of Marriage and the Family