

Ying Ding, Gobinda Chowdhury, Schubert Foo
Division of Information Studies, School of Applied Science
Nanyang Technological University, Singapore 639798

ORGANISING KEYWORDS IN A WEB SEARCH ENVIRONMENT: A METHODOLOGY BASED ON CO-WORD ANALYSIS

Organising keywords in a web search environment: a methodology based on co-word analysis

Abstract

The rapid development of the Internet and World Wide Web has caused some critical problem for information retrieval. Researchers have made several attempts to solve these problems. Thesauri and subject heading lists as traditional information retrieval tools have been criticised for their efficiency to tackle these newly emerging problems. This paper proposes an information retrieval tool generated by cocitation analysis, comprising keyword clusters with relationships based on the co-occurrences of keywords in the literature. Such a tool can play the role of an associative thesaurus that can provide information about the keywords in a domain that might be useful for information searching and query expansion.

1. Introduction

The rapid development of the Internet and World Wide Web has given a tremendous boost to the growth and availability of electronic information resources and has also changed the way people search information. However, alongside this, new problems associated with the searching and retrieval of the required information have emerged. Researchers have made several attempts to solve these problems.

Thesauri and subject heading lists have been used in the library and information world for a long time to solve the problems of inconsistencies in indexing, and also for providing support for users in query formulation, query expansion, and so on. However, existing thesauri often represent a general subject area, and therefore they usually need significant enhancement to be tailored to a specific application. The structure of thesauri, in particular the relationships among descriptors, is also questioned by IR researchers (Harter & Cheng, 1996). With the rapid development of various specialized domains, more and more new concepts, methods, theories or new sub-domains are emerging making the thesaurus dated. Building up or amending thesauri is extremely time-consuming and labor-intensive. Chen & Lynch (1992) pointed out that among the major reasons to cause the difficulty of information retrieval are the lack of explicit semantic clustering of or linkages between relevant information and the limits of conventional keyword-driven search techniques. A research group in the Artificial Intelligence Lab, University of Arizona has conducted research on automatic thesauri (Chen, Yim, Fye & Schatz, 1995). This kind of automatically generated thesaurus component plays an important role in solving searchers' vocabulary problems during information retrieval. The specific algorithms adopted in such research include term filtering, automatic indexing, and cluster analysis, which are complicated and time and resource-intensive process.

Here we propose an alternative way to developing a tool comprising keyword clusters with relationships based on the co-occurrences of keywords in the literature. Such a tool can play the role of an associative thesaurus that can provide information about the keywords in a domain that might be useful for information searching and query expansion. Our approach is based on the bibliometric co-word analysis method, which is commonly used to analyze papers in order to identify keywords that describe their research content and linking papers by the degree of co-occurrence of these keywords to produce a 'map index' of a specialty (King, 1987). Here we have applied this technique to identify the relationships among words and to create keyword maps that may be useful for information retrieval purposes.

In this study, we chose Information Retrieval (IR) as the domain. The first part of this paper briefly discusses the co-word analysis in the field of Information Retrieval. Then, the result of the co-word analysis has been compared with traditional thesauri to identify the difference. The last part reports Bibliometric Information Retrieval System that organizes and displays keyword clusters for information searching in the web environment.

2. Methodology

The IR papers were retrieved from SCI (Science Citation Index) and SSCI (Social Science Citation Index) covering the period of 1987-1997. A number of retrieved articles that did not have any abstracts, or were book reviews, editorial, meeting abstracts, newsletters or notes were excluded. Finally 2,012 articles were selected as the co-word analysis sample. From each of these papers, we have not only accepted all the keywords added by the SCI and SSCI database indexers but have also extracted important keywords from titles and abstracts manually. All these keywords were standardized using the LISA thesaurus, LCSH (Library Congress Subject Heading) and Thesaurus of Information Technology Terms (TITT) in order to make them consistent (singular/plural), unified (synonyms), and as far as possible unambiguous (homonyms).

A total of 3,227 unique keywords were collected from the chosen 2,012 articles. In these literature, some related concepts are represented by different words or phrases. Such words or phrases were standardized by selecting an appropriate heading from the vocabulary tools that would represent them, such as words from LISA thesaurus, LCSH, and TITT. The following examples illustrate the process:

- Synonyms: citations + citation analysis = citation analysis; linguistics + linguistic analysis = linguistic analysis; navigating + browsing = browsing; inquiries + searching = searching; relevance searching + relevance feedback = relevance feedback; digital library concept + electronic library = digital libraries;
- Antonyms: Boolean strategies + Non-Boolean strategies = Boolean strategies; and so on.
- Ambiguity: strategies + search strategies = searching; CD-ROMs + CD-ROM databases = CD-ROMs; user aids + user guides = user training; and so on.
- Broad term/Narrow term: retrieval performance measures + performance measures = performance measures; end users + users = users; automatic indexing + indexing = indexing; research students + foreign students = students; education activities + education = education; school children + children = children; optical discs + CD-ROMs = CD-ROMs; and so on.
- See or See Also term: information work + reference work = information work; terms + keywords = keywords; and so on.
- Use or Use for term: undergraduate students + students = students; and so on.
- Others: retrieval evaluation + performance measures = performance measures; user groups + users = users; user needs + user satisfaction = user needs; and so on.

- General terms were excluded, such as: knowledge, theories, tests, influence, projects, criteria, development, errors, applications, production, competition, status, implementation, definition, annotations, and so on.

Words with a frequency of one or two were merged with their BROAD terms. Words with frequency of one or two, which did not have any BROAD or similar term in our list were ignored. Finally, 240 keywords with frequency of more than two were chosen as the research sample for the co-word analysis. In order to compare the dynamic features of these word clusters based on the co-word analysis, we divided the whole 11-year period into two consecutive parts: the first five-year period (1987-1991) and the second six-year period (1992-1997).

Specifically built Foxpro programs were used to calculate the number of times two keywords appeared together in the same publication. Thus, we have formed a co-occurrence matrix of 240*240 keywords. In the cell of keyword X and keyword Y we put the co-occurrence frequency of X and Y. The diagonal values of the matrix were treated as missing data (McCain, 1990). The matrix was transformed into a correlation matrix by using the Pearson's correlation coefficient indicating the similarity and dissimilarity of each keyword pair. So each keyword has its own relevant keywords. But unlike the traditional thesaurus, which is built up by domain experts, this is more like automatic thesaurus mentioned before because both are based on term co-occurrence and can be built up automatically (Chen, Yim, Fye and Schatz, 1995).

Peat and Willett (1991) found that similar terms identified by symmetric co-occurrence functions tended to occur very frequently in the database being searched and thus did little or nothing to improve the discriminatory power of the original query. They also concluded that this can help explain Sparck Jones's (1971) findings that the best retrieval results were obtained if only the less frequently occurring terms were clustered and if the more frequently occurring terms were left unclustered. This also happened in this research. In order to alleviate this negative effect, we recalculated the co-occurrence frequency with the Salton Index which can avoid high frequency words to be linked with almost every other keyword in the research sample (Noyons & van Raan, 1998).

For each keyword in the research sample during each period of study (1987-1997, 1987-1991 and 1992-1997), we chose the top 20 words (20s) with high Salton Index with this keyword to compare with its corresponding semantic descriptors for the three selected traditional thesauri (TT). As some words' semantic descriptors in each single thesaurus is very little or even empty, we combined the semantic descriptors from these three traditional thesauri (LISA thesaurus, LCSH, and TITT) for each word in the research sample and reduced the semantic duplications to form one whole semantic set for comparison. Through the comparison of these two sets of data, we wanted to find the difference between the co-word analysis and traditional thesauri.

3. Results

Comparison co-word analysis with traditional thesauri

First, we compared the co-word analysis with traditional thesauri to observe the difference between them (Table 1). For each period, around 50% of the sample keywords have similarity in its 20s and TT, but the average similarity per sample keyword is very low. This means that the associations of words identified by co-word analysis were different from those obtained from traditional thesauri. One important conclusion coming out from this comparison is that there exists the difference between co-word analysis and traditional thesaurus. The conclusion is consistent with

Chen's result (Chen, Ng, Martinez, & Schatz, 1997). It indicates that co-word analysis can become a significant tool to support traditional thesaurus to generate search varieties.

Table 1. Comparison of co-word analysis with traditional thesauri

Period	Sample keywords	Keywords with similarity		Keywords with lowest similarity		Keywords with highest similarity		Average similarity
		No.	%	No.	Similarity	No.	Similarity	
1987-1997	216	102	47.2%	60	5%	2	25%	7.9%
1987-1991	176	75	42.6%	20	5%	1	100%	12.4%
1992-1997	216	92	42.6%	52	5%	1	33%	7.9%

Changes of co-word analysis over time

Second, we compared the co-word analysis in different periods to perceive the dynamic changes among them (Table 2). Both the separate periods (five years and six years) have high similarities with the entire period (11 years). But the results for the two separate periods are less similar. Thus, this comparison captures the changes of co-word analysis. In other words, co-word analysis can catch the changes of its domain area to provide better and timely information guide for users.

Table 2. Dynamic changes of co-word analysis during three different periods

Periods	Sample keywords	Keywords with similarity		Keywords with lowest similarity		Keywords with highest similarity		Average similarity
		No.	%	No.	Similarity	No.	Similarity	
1987-1997 vs. 1987-1991	193	192	99.5%	1	<=10%	4	>90% and <=100%	52.2%
1987-1997 vs. 1992-1997	239	239	100%	1	>30% and <=40%	61	>90% and <=100%	83.6%
1987-1991 vs. 1992-1997	192	168	87.5%	14	<=10%	1	>60% and <=70%	26.1%

This research shows that the results of co-word analysis can be used for organizing knowledge through keyword maps and they may be quite useful to compliment the traditional vocabulary tools.

Bibliometric Information Retrieval System (BIRS)

Based on the above result of co-word analysis, BIRS was designed to help end-users formulate and expand queries for searching information on a number of media ranging from OPAC to online database and World Wide Web (WWW). BIRS is implemented and maintained in an environment running Microsoft Windows 98/NT operating system with Microsoft Access 97 as BIRS database and ODBC server as the connection between web application and database.

Information organization features of BIRS

- Information organization and visualization feature: The maps show a visual organization of the knowledge or concepts in the domain. For example, in Figure 1 and 2, users can obtain general information about the IR field via the overview map. Once they go deeper to the selected cluster, they will be provided with detailed information about the sub-domain, such as the intellectual location of specific subject, the relationships of different subjects, relevance of different subjects, and so on.

- Multilevel information organization feature: Multilevel information organization and browsing is incorporated into BIRS to support layering so that users can slice and dice to get different levels of information about interesting topics. For example, three levels of details are available for the keyword map as shown in Figure 3. The top level (Level A) offers an overview of the IR field. Clicking on a cluster results in a more detailed map of the specific cluster (Level B). Clicking on an appropriate keyword results in the 20 most relevant associated keywords (Level C).

System evaluation

A preliminary system evaluation was conducted involving 35 users. Among them, six were from IT-related companies and the remaining 29 were postgraduate students in the MscIS programme at Nanyang Technological University (Singapore). The results show that: 28 (80%) users got a good or very good understanding of the IR area with the help of the information organization of BIRS; 27 (77%) users agreed that BIRS can greatly help them form and expand their queries; 25 (71%) subjects indicated good satisfaction with the multi-level information organization and browsing system; 28 (80%) users gave good or very good comments on the helpfulness of the information organization and visualization feature of BIRS. Twenty-two (63%) users added new keywords to expand and refine their queries, while these subjects had experienced problems to form their queries before using the BIRS system.

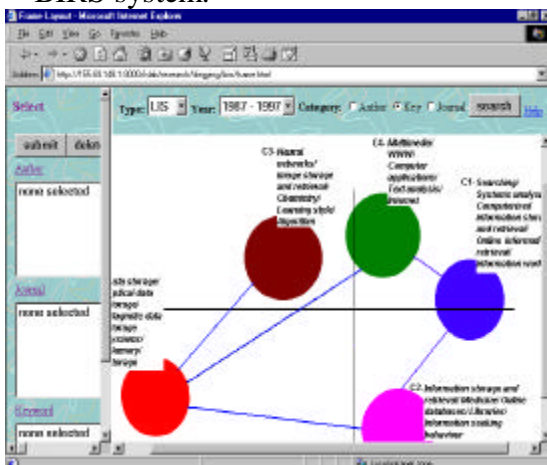


Figure 1. Overview keyword map

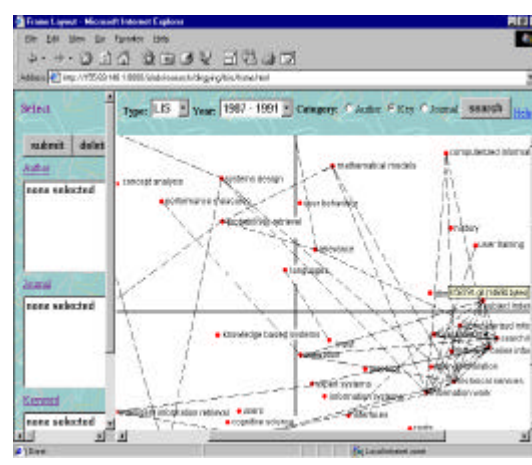


Figure 2. Detail keyword map

In summary, this evaluation indicated that BIRS was found useful in assisting query formation and expansion, and provided a useful means to acquire background information about the domain area in one integrated system. The information visualization, multilevel browsing and common user interface are also deemed as the novel characteristics of BIRS.

4. Discussion

This research has proposed an alternative method for knowledge organization by using the result of a co-word analysis. Based on the results of co-word analysis, two comparisons were conducted in an attempt to reveal the performance of co-word analysis and its novel characteristics in comparison with the conventional thesauri, and also to reveal its dynamic changes in different periods. The associations of words identified by co-word analysis were found to be different from those obtained from

traditional thesauri. It indicates that co-word analysis can become a significant information organization tool to support traditional thesaurus to generate search variety and to re-organize information. Furthermore, co-word analysis can catch the dynamic changes of a domain area and thus can provide better information guide for users.

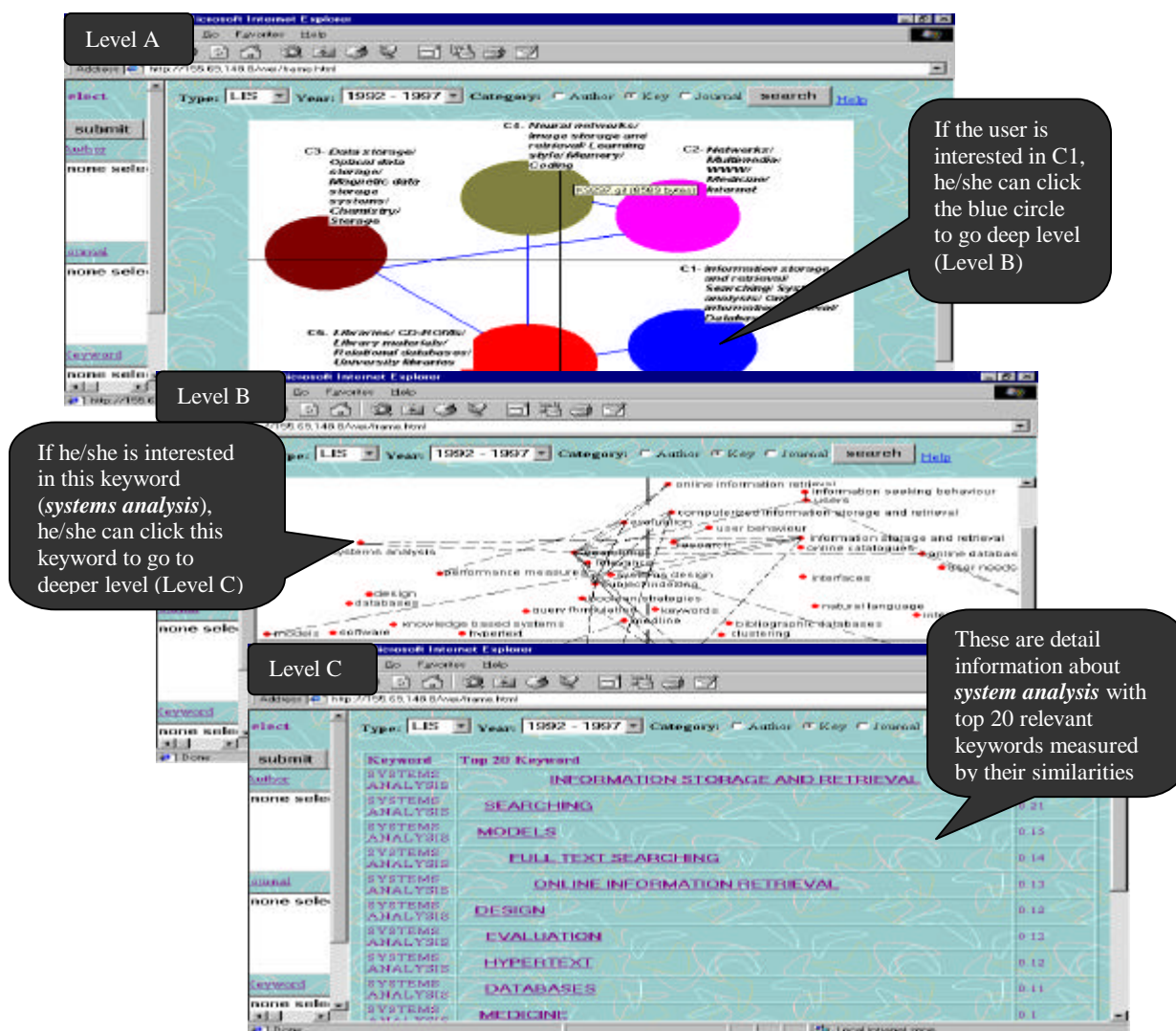


Figure 3. Multilevel browsing feature of the BIRS

Users of information retrieval often face the critical problem to form and expand their queries (Bates, 1986; Peat & Willett, 1991; Chowdhury, 1999; Chowdhury and Chowdhury, 1999; and Voorbij, 1999). We have attempted to incorporate and integrate the results of the above co-word analysis to form BIRS to help users in (1) query formulation and expansion, (2) acquiring new understanding in a particular subject domain, and (3) showing an alternative approach to organize information. The results of the user evaluation of the BIRS confirm that this system can efficiently help user form and expand their queries as well as aid users to better understand the information retrieval domain area. User feedback also clearly indicates that users like the graphical nature of information organization and multi-level information organization and browsing system.

As this is the first version of BIRS, many areas need further refinement, enhancing and development. It is also undeniable that BIRS needs to be extended to cover larger subject domains to make it more useful to a wider community of users. It is possible for BIRS not only to incorporate new maps or other forms of data representation, but also to incorporate additional or new forms of search engines, thereby providing a useful one-stop tool for information retrieval sessions. All these provide much scope for future on using the results of co-word analysis in various subject for organizing information and using such organized information to help users in the actual information retrieval operations using BIRS.

References

- Bates, M. J. (1986). Subject access in online catalogs: A design model. *Journal of the American Society for Information Science*, 37, 357-376.
- Bradley, P. (1995). Towards a common user interface. *Aslib proceedings*, 47(7/8), 179-184.
- Chen, H. & Lynch, K. J. (1992). Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(5), 885-902.
- Chen, H., Ng, T. D., Martinez, J. & Schatz, B. R. (1997). 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), 17-31.
- Chen, H., Yim, T., Fye, D. & Schatz, B. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science*, 46(3), 175-193.
- Chowdhury, G. G. (1999). The Internet and information retrieval research: A brief review. *Journal of Documentation*, 55(2), 209-225.
- Chowdhury, G.G. & Chowdhury, S. (1999). Digital library research: major issues and trends. *Journal of Documentation*, 55(4), 409-448.
- Harter, S.P. & Cheng, Y. R. (1996). Colinked descriptors: Improving vocabulary selection for end-user searching. *Journal of the American Society for Information Science*, 47(4), 311-325.
- King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, 13, 261-276.
- McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6), 433-443.
- Noyons, E.C. M. & van Raan, A.F. J. (1998). Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research. *Journal of the American Society for Information Science*, 49(1), 68-81.
- Peat, H. J. & Willett, P. (1991). The limitation of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5), 378-383.
- Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. London: Butterworths.
- Voorbij, H. J. (1999). Searching scientific information on the Internet: A Dutch academic user survey. *Journal of the American Society for Information Science*, 50(7), 598-615.