ELSEVIER SCIENCE Ltd. [DTD 4.1.0]
JOURNAL IPM ARTICLE No. 517
PAGES 1-26 DISPATCH 4 November 2000
IPM 517
PROD. TYPE: FROM DISK
A

PERGAMON
Information Processing and Management 000 (2000) 000–000
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Bibliography of information retrieval research by using co-word analysis

Ying Ding *, Gobinda G. Chowdhury, Schubert Foo

Division of Information Studies, School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798

Abstract

The aim of this study is to map the intellectual structure of the field of Information Retrieval (IR) during the period of 1987–1997. Co-word analysis was employed to reveal patterns and trends in the IR field by measuring the association strengths of terms representative of relevant publications or other texts produced in IR field. Data were collected from Science Citation Index (SCI) and Social Science Citation Index (SSCI) for the period of 1987–1997. In addition to the keywords added by the SCI and SSCI databases, other important keywords were extracted from titles and abstracts manually. These keywords were further standardized using vocabulary control tools. In order to trace the dynamic changes of the IR field, the whole 11-year period was further separated into two consecutive periods: 1987–1991 and 1992–1997. The results show that the IR field has some established research themes and it also changes rapidly to embrace new themes.

Keywords: Co-word analysis; Information retrieval research; Research trends; Science citation index; Social science citation index

1. Introduction

It is important nowadays, for both intellectual and policy reasons, to be able to map the relationship between concepts, ideas and problems in science and social sciences. There are several ways in which such mapping may be attempted. The traditional way, both in science studies and science policy, is to seek the views of relatively small number of experts (peer review) (Law &

* Corresponding author.
E-mail addresses: ying@cs.vu.nl (Y. Ding), asggchowdhury@ntu.edu.sg (G.G. Chowdhury), assfoo@ntu.edu.sg (S. Foo).

0306-4573/00/$ - see front matter
PII: S0306-4573(00)00051-0

Whittaker, 1992). Bibliometric research is another way to achieve this task from quantitative perspective.

Bibliometric research is devoted to quantitative studies of literature. It encompasses a number of empirical methods, viz., citation and co-citation analyses. Co-citation analysis is an important subset of bibliometrics. Since Small (1973) introduced the concept and defined it as "the frequency with which two items of earlier literature are cited together by the later literature", co-citation analyses have been successfully applied to examine the intellectual structure of many disciplines. The criteria generally involve counting the number of times certain markers occur or co-occur, giving rise to information on such author co-citation, journal co-citation, keyword co-citation, and so on. In particular, they can be applied to the formal record of scholarly communication from different points, such as authors, journals and textual content. Ding, Chowdhury, and Foo (1999a,b, 2000) have recently shown how bibliometric studies can be used to trace the development of subject mapping/cartography using author co-citation, and journal co-citation analysis in the field of information retrieval. Bibliometric studies also also aim at the advancement of knowledge on the development of science and technology in elation to social and to policy questions (van Raan, 1997). Traditional bibliometric techniques such as author and journal co-citation analyses are based on the analysis of the citations contained in scientific papers. While this kind of analysis leads to interesting results, it does not provide an immediate picture of the actual content of the research topics dealt with in the literature. Co-word analysis, that counts and analyzes the co-occurrences of keywords in the publications on a given subject, on the other hand, has the potential to address precisely this kind of analytic problem (Callon, Courtial, & Laville, 1991).

Co-word analysis reduces and projects the data into a specific visual representation with the maintenance of essential information containing in the data. It is based on the nature of words, which are the important carrier of scientific concepts, idea and knowledge (van Raan & Tijssen, 1993). Many researchers have used co-word analysis as an important method to explore the concept network in different fields, for instance, software engineering (Coulter, Monarch, & Konda, 1998), polymer chemistry (Callon et al., 1991), scientometrics (Courtial, 1994), neural network research (Noyons & van Raan, 1998a; van Raan & Tijssen, 1993), biological safety (Cambrosio, Limoges, Courtial, & Laville, 1993), acidification research (Law & Whittaker, 1992), patents (Courtial, Callon, & Sigogneau, 1993), optomechatronics (Noyons & van Raan, 1994), bioelectronics (Hinze, 1994), medicine (Rikken, Kiers, & Vos, 1995), biology (Rip & Courtial, 1984; Looze & Lemarie, 1997), condensed matter physics (Bhattacharya & Basu, 1998), and so on.

With the rapid development of online databases, Internet and World Wide Web, information retrieval has experienced drastic changes over the past few years. In the present study, we have applied co-word analysis to trace these changes, in particular, the concept or idea derivation in the field of Information Retrieval (IR) during the period of 1987–1997.

2. Methodology

Co-word analysis draws upon the assumption that a paper's keywords constitute an adequate description of its content or, the links the paper established between problems. Two keywords co-occurring within the same paper are an indication of a link between the topics to which they refer

(Cambrosio et al., 1993). The presence of many co-occurrences around the same word or pair of words points to a locus of strategic alliance within papers that may correspond to a research theme. Co-word analysis reveals patterns and trends in a specific discipline by measuring the association strengths of terms representative of relevant publications produced in this area. The main feature of co-word analysis is that it visualizes the intellectual structure of one specific discipline into maps of the conceptual space of this field, and that a time-series of such maps produces a trace of the changes in this conceptual space.

2.1. Steps of co-word analysis

2.1.1. Data collection

Words are the most important research elements in co-word analysis. There are two ways to extract words from journal articles, conference papers, reports or even chapters of books. One way is to extract keywords from keyword lists, title, abstract, and sometimes even including classification codes. Many journals, abstracting services and databases already provide such keywords. Cambrosio et al. (1993) chose keywords added by indexers and title words as the research data because of the poor quality of indexing in their specific database. The resulting lists of descriptors were standardized to eliminate different spellings and variants of the same terms. Coulter et al. (1998) selected descriptors chosen by professional indexers. They believed that it is useful to study a fixed system that imposes a common nomenclature. Professional indexers' experiences assure standard application of that taxonomy. Looze and Lemarie (1997) conducted co-word study based on the keywords proposed by the experts. Some researchers downloaded keywords from online database, which are added by database indexers and authors (Courtial, 1994; Law & Whittaker, 1992; Courtial, Cahlik, & Callon, 1994). Noyons and van Raan (1998b) mapped the coarse overall structure in the field of neural networks by using the co-occurrence of classification codes.

One of the most significant reservations about this data collection from controlled vocabulary is the possibility of an "indexer effect". The fear is that indexing might reflect the prejudices and points of view developed by indexers during the course of their training (Law & Whittaker, 1992) and the probable inconsistencies in keyword selection by professional indexers working for different databases (King, 1987). This effect was eliminated by the good results of interviews (Law & Whittaker, 1992; Cambrosio et al., 1993; Tijssen, 1993; Courtial, 1994). Law and Whittaker (1992) mentioned that "after an analysis of 83 interviews we are able to report that anxieties about the quality of the indexing in the PASCAL database are substantially unfounded. The match between the keywords chosen by indexers and those chosen by respondents is reasonable and even more significantly."

Another method of data collection involves extracting words directly from full-text documents by using some software, such as *NPtools* (Voutilainen, 1993). The words or phrases with proper frequency are chosen as the subject of co-word analysis to represent the core topics of the specific field. This method was chosen to avoid the negative effort of indexer and time problem of thesauri and classification systems, such as the lengthy time involved in constructing the thesauri or classification systems, difficulty to maintain and keep abreast of new development in the corresponding fields and so on.

This study has used a combination of the two methods of data collection discussed above. We have chosen the keywords added by the ISI database indexers, and also have extracted keywords from the titles and abstracts of the corresponding documents/articles.

2.1.2. Data standardizing

In co-word analysis, once a research subject is selected, a matrix based on the word co-occurrence is built. The value of the cell of two words is decided by the times these two words both appear in the same document. The higher co-occurrence frequency of the two words means the closer relationship between them. The matrix is then transformed into a correlation matrix by using specific correlation coefficient.

2.1.3. Data mapping

There are several approaches to mapping the data. The most commonly used methods are multidimensional scaling and clustering techniques. Other methods include the use of specific software: LEXIMAPPE program for co-word mapping, developed as a science policy tool and has already been used to analyze publications from various research fields (Looze & Lemarie, 1997; Law & Whittaker, 1992; Cambrosio et al., 1993; Courtial, 1994, and so on); Content Analysis and Information Retrieval (CAIR) developed by Software Engineering Institute in Caregie Mellon University, was employed by co-word analysis researchers (Coulter et al., 1998); Bibliometric Technology Monitoring (BibTechMon) developed by Austrian Research Centers is another software used for co-word analysis (Kopcsa & Schiebel, 1998; Widhalm, 1999).

Another approach is to use Kohonen's neural network algorithms to map the data. Polanco, Francois, and Keim (1998) have applied artificial neural network technology, that includes the Adaptive Resonance Theory (ART), a Multilayer Perceptron (MLP) and an associative network with unsupervised learning (KOHONRN), to assess and map the research area of Science and Technology Information. WEBSOM research group is one of such example. They built up their web-based document map interface to map the words from large collections of articles by using a Self-Organizing Map (Kohonen, 1995). WEBSOM performs a completely automatic and unsupervised full-text analysis of the document set using Self-Organizing Maps. The results of the analysis, an ordered map of the document space, display directly the similarity relations of the subject matters of the documents. They are reflected as distance relations on the document map. The density of documents in different parts of the document space can be illustrated with shades of color on the document map display (Honkela, Kaski, & Kohonen, 1996).

2.2. Method used in this study

2.2.1. Data collection

A total of 3325 IR papers were retrieved from the Science Citation Index (SCI) and Social Science Citation Index (SSCI) covering the period of 1987–1997. A total of 1313 documents were excluded because they were not articles, but abstracts, book reviews, editorials, meeting abstracts, news, letters, or notes. Finally 2012 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 added by indexers or chosen from titles or abstracts are standardized using the

LISA thesaurus, LCSH and Thesaurus of Information Technology Terms, in order to make them consistent (singular/plural), unified (synonyms), and unambiguous (homonyms). The average number of keywords per article is found to be 5.09. The range of keywords for each article varies from one to ten. Around 5.4% articles have 10 keywords while 93.4% of articles have more than one keyword.

2.2.2. Vocabulary standardization

A total of 3227 unique keywords were collected from the chosen 2012 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 Thesaurus of Information Technology Terms. The following examples will 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 word 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 more than two were chosen as the research sample for co-word analysis.

Specifically built Foxpro programs were used to calculate the number of times two keywords appear 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. The matrix was transformed into a correlation matrix by using Pearson's correlation coefficient indicating the similarity and dissimilarity of each keyword pair.

2.2.3. Data mapping

In order to have a clearer picture of the IR field, we have used hierarchical clustering techniques with Ward's method to divide these 240 keywords into five clusters (for detailed discussion of this method see, Arabie, Carroll, & DeSarbo, 1987; Norusis, 1997). Subsequently, keywords with high frequencies in each cluster were chosen to represent the cluster, so that the coarse general overview map was achieved by using MDS (multidimensional scaling) techniques to represent the overall positions of these five clusters in the IR field.

In order to generate refined MDS maps for each cluster, all the keywords in one cluster were chosen as the group of variables and MDS was applied to them to yield a two dimensional map. Thus, the five refined MDS maps for each of the five clusters can display the specific relationship of the keywords within the clusters concerned. This technique is also known as 'multi-level mapping' (Noyons & van Raan, 1998b). The multi-level mapping concerns maps of a field with more than one level. We first generated a coarse structure of the field (the general overview map). Refined maps for each cluster is then used to present the detailed structure of the specific cluster. By using multi-level maps, one can zoom into a sub-domain to get more detailed information about particular areas of interests.

The raw data 240 × 240 matrix was recalculated (Pearson correlation coefficient) in order to find proximity on the basis of the 240-vector. In other words, the similarity between two words was calculated on the basis of all co-occurrence frequency that these two words have with all the other items in the same matrix. So the words with high Pearson correlation coefficient are located

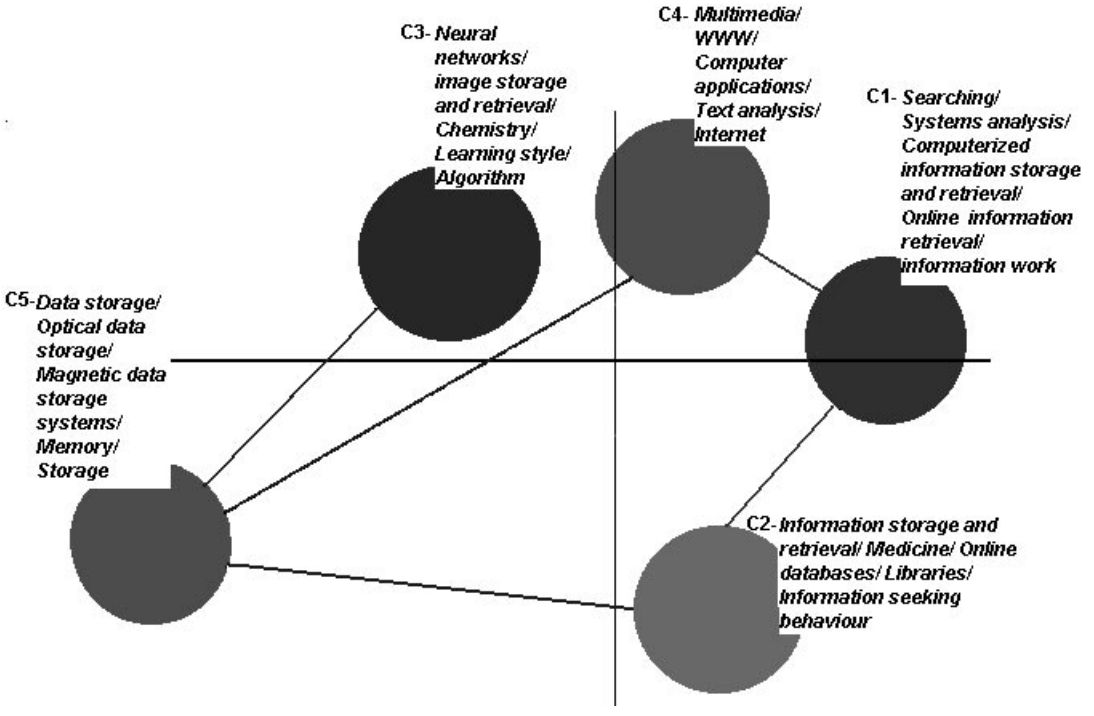

Fig. 1. General overview map of IR field in 1987–1997 (lines between clusters indicate strongest linkage according to the Salton Index).

together in the map, and those words located together in the map have high similarity in terms of co-occurrence profile within the whole matrix.

The dotted line between two keywords in the maps (Figs. 2–6, 8–12, 14–18) indicates the high correlation of them with the Salton Index (Hamers et al., 1989) that has a value of more than 0.2. The Salton Index is one of the important indices that can screen the negative effect of keywords with high occurrence frequency, and at the same time, reflects the direct similarity of two individual words in terms of co-occurrence frequency (Noyons, 1998, Peters & van Raan, 1993a,b). These links are most interesting because the information about the correlation (in terms of 'co-occurrence profile') is already captured by the positioning (Noyons, 1998).

3. Results

In order to grasp the overall co-word analysis in the whole period (1987–1997), we analyzed keywords based on the whole period. Then we divided the whole period into two parts, 1987–1991 and 1992–1997, so that we can identify the dynamic changes during these two periods.

3.1. Co-word analysis of 1987–1997

The top ten keywords with high frequency in each cluster were chosen to represent these five clusters because of the limitation of MDS in SPSS (as mentioned previously). A general overview map of the IR field in 1987–1997 is generated by MDS based on these fifty representative key-

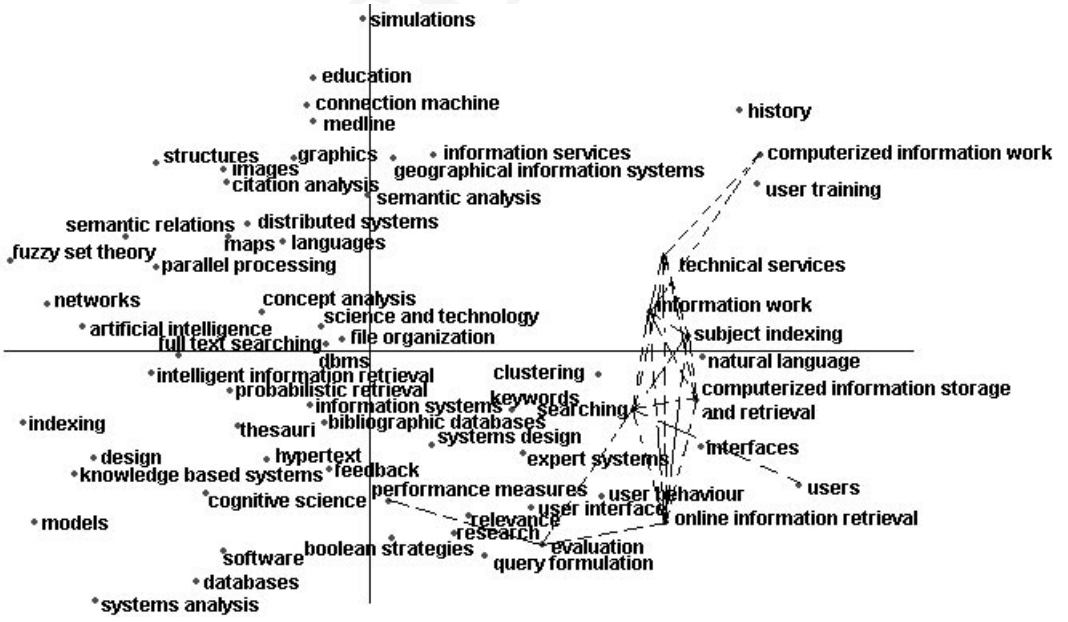

Fig. 2. The fine structure of Cluster 1 (C1) in 1987–1997 (the dotted line represents the link between two keywords with The Salton Index (>0.2)).

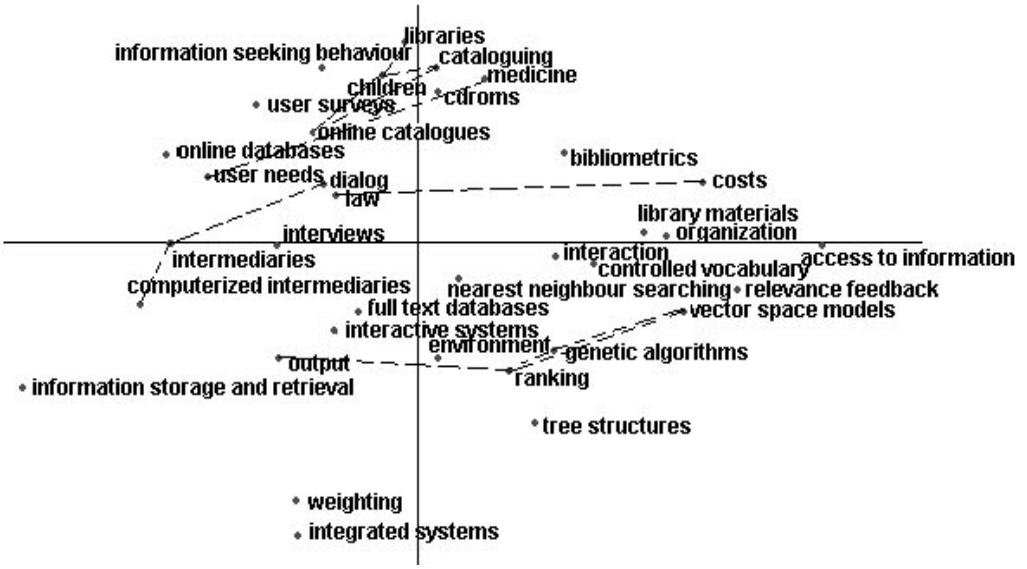

Fig. 3. The fine structure of Cluster 2 (C2) in 1987–1997 (the dotted line represents the link between two keywords with theSalton Index (>0.2)).

words. The rough position of each cluster is decided by its ten representative keywords (Fig. 1). Each cluster (sub-domain) is labeled according to the most frequent keywords appearing in the cluster.

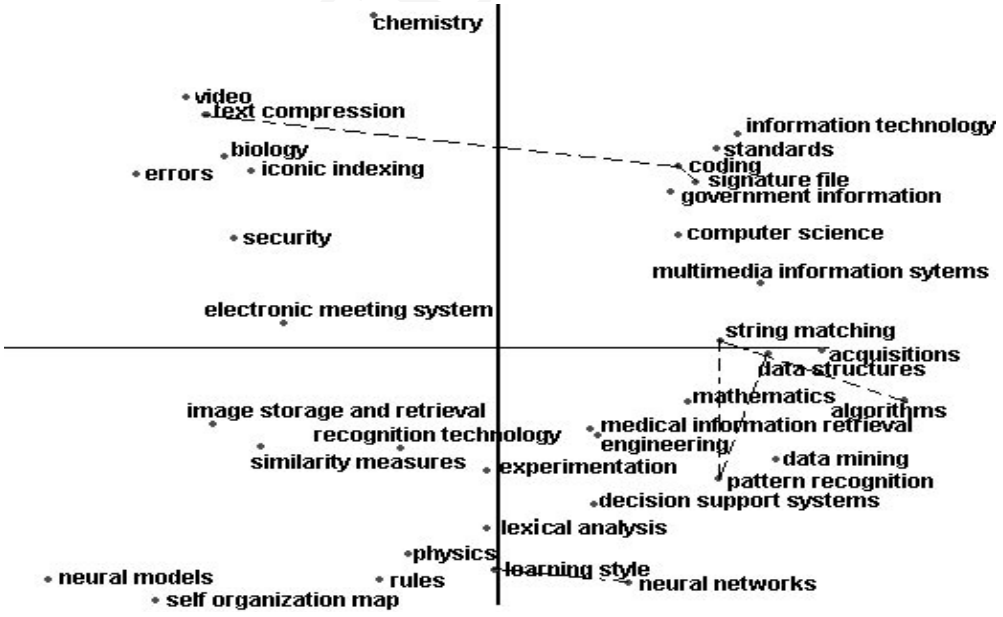

Fig. 4. The fine structure of Cluster 3 (C3) in 1987–1997 (the dotted line represents the link between two keywords with the Salton Index (>0.2)).

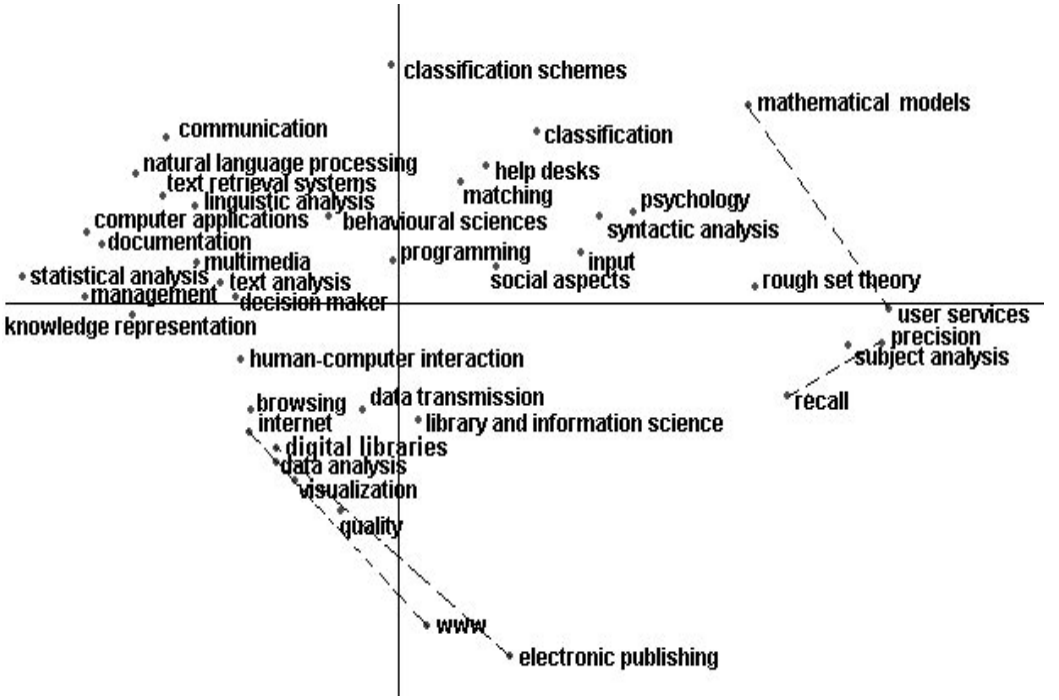

Fig. 5. The fine structure of Cluster 4 (C4) in 1987–1997 (the dotted line represents the link between two keywords with the Salton Index (>0.2)).

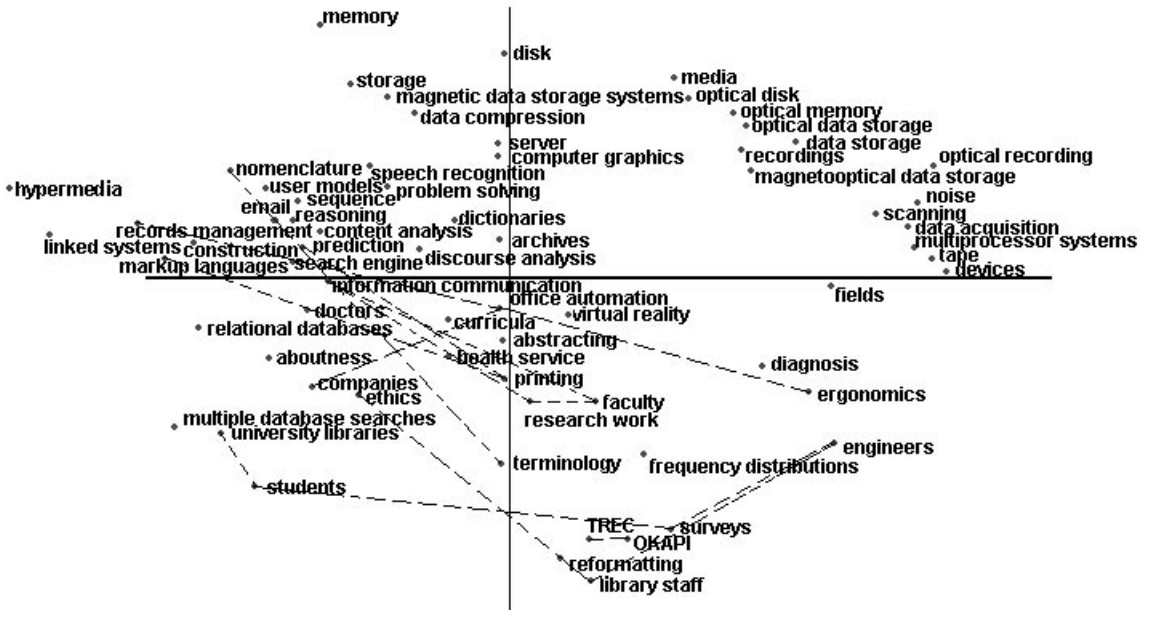

Fig. 6. The fine structure of Cluster 5 (C5) in 1987–1997 (the dotted line represents the link between two keywords with the Salton Index (>0.2)).

Each cluster contains around 50 keywords. In order to construct 'fine-structure' or detailed maps, each cluster was chosen as the input variable to map the sub-domain based on 240 × 240 correlation matrix. Thus, five detailed sub-domain maps (Figs. 2–6) were generated to reflect specific characters of each of the five sub-domains in the IR field.

Cluster 1 includes the research topics relating to searching, systems analysis, computerized information storage and retrieval, online information retrieval, information work, and so on. Words located together indicate that they often appear together in the same publications. Cluster 2 includes research topics relating to information storage and retrieval, medicine, online catalogues, libraries, information seeking behavior, and so on. Cluster 3 includes research topics relating to neural networks, image storage and retrieval, chemistry, learning style, algorithms, and so on. Cluster 4 includes topics on multimedia, WWW, computer applications, text analysis, Internet, and so on. Finally, Cluster 5 includes topics on data storage, optical data storage, magnetic data storage, memory, storage, and so on.

A cluster can be defined in two different ways. First, it can be seen as a point in a general network, one which is characterized by its position, that is to say by the bundle of links uniting it to other clusters/points in the general network. Secondly, it can be seen as a cluster made up of words linked with each other – it itself defines a more or less dense network, one which is more or less coherent and robust (Callon et al., 1991). The inter-relationships among these five clusters are demonstrated by the links of keywords in different clusters. Although we use cluster techniques to separate the whole keyword sample into five clusters, unavoidably, some keywords with links (the Salton Index (>0.2)) are divided into different clusters (Fig. 1).

Table 1 shows a comparison of five identified clusters during the period of 1987–1997. In the table, the outer link refers to the number of links between words in the cluster and words in the rest of other clusters. The inner link refers to the number of words within the cluster. The outer and inner link percentages refer to the value of outer and inner link in comparison to the total sum of these links respectively. The outer link key refers to the number of keywords in the cluster that have links with other keywords in other clusters, while the total key refers to the total keywords in the cluster. Centrality is defined as the mean of the outer link (sum of Salton-index of the outer

Table 1
Comparison of five clusters during the period of 1987–1997[a]

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Average
Outer link	31	33	11	21	38	26.8
Inner link	40	24	12	8	28	22.4
Total link	71	57	23	29	66	49.2
Outer link %	44%	58%	48%	72%	58%	56%
Inner link %	56%	42%	52%	28%	42%	44%
Outer link key	19	18	9	14	27	17.4
Total key	63	35	36	39	67	48
Outer link key %	30%	51%	25%	36%	40%	36%
Average link/key	1.13	1.63	0.64	0.74	0.99	1.03
Centrality	0.276	0.260	0.250	0.249	0.247	–
Density	0.413	0.311	0.285	0.385	0.286	–

[a] *Note*: Here link denotes the link between keywords with the Salton Index (>0.2).

links/outer link) and density is defined as the mean of the inner link (sum of Salton-index of the inner links/inner link).

From Table 1, it can be seen that for each cluster, the outer links among keywords are slightly more than its inner links. Around 36% of keywords in each cluster have outer links with other keywords based on the Salton Index. These indicate that the links among keywords based on the Salton Index do aggregate within the cluster to some degree, but not intensely. Rather, around 50% of such kind of links are located among the inter-relationships of different clusters. These links not only illustrate the substantial relationships among clusters, but also show a stable internal composition in each cluster. Meanwhile, on the average, each keyword has around one link (outer link or inner link). It is clearly advisable to show links between keyword of different clusters.

Centrality measures for a given cluster the intensity of its links with other clusters. The more numerous and stronger are these links, the more this cluster designates a set of research problems considered crucial by the scientific or technological community. It occupies a strategic position. The centrality of a given cluster could be measured by calculating the mean of the links with other clusters (Callon et al., 1991). Density characterizes the strength of the links that tie the words making up the cluster together. The stronger these links are, the more the research problems

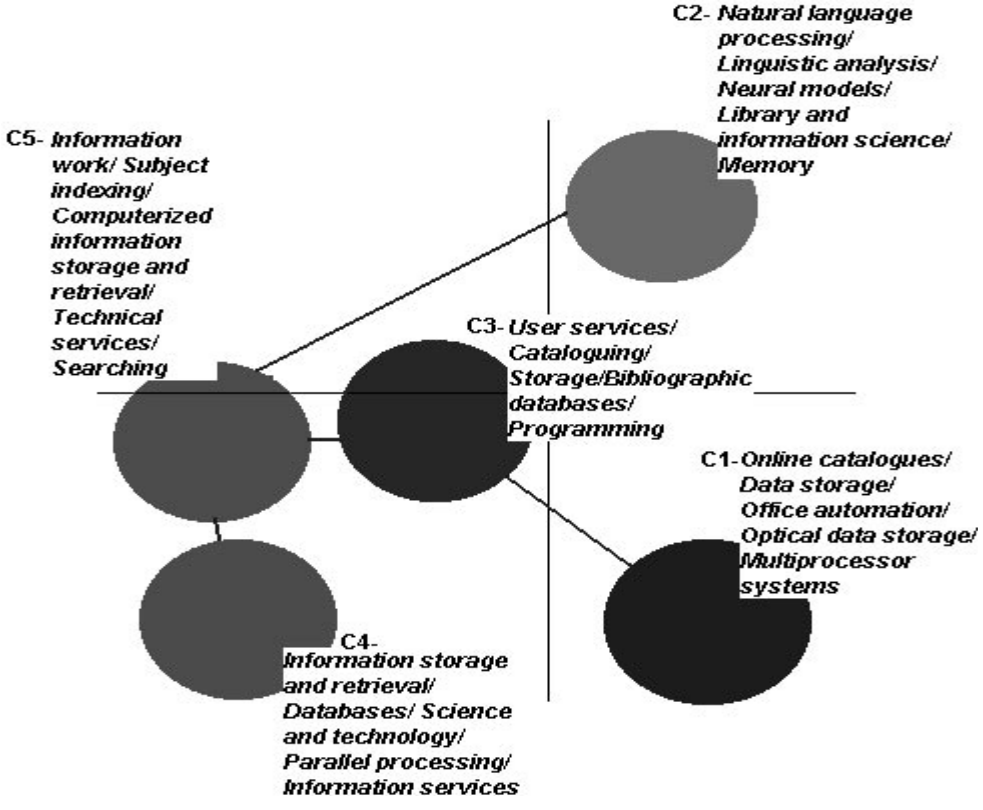

Fig. 7. General overview map of IR field in 1987–1991 (lines between clusters indicate strongest linkage according to the Salton Index).

corresponding to the cluster constitute a coherent and integrated whole. It could be said that density provides a good representation of the cluster's capacity to maintain itself and to develop over the course of time in the field under consideration. The value of the density of a given cluster can be measured by simply calculating for each cluster the mean value of its internal links (Callon et al., 1991). Cluster 1 with the highest centrality and density is both central to the general network (it is strongly connected to other clusters) and has intense internal links (it displays a high degree of development). Cluster 1 in some sense constitutes the file's core. Its position is strategic, and it is probably dealt with systematically and over a long period by a well-defined group of researchers. From the words in this cluster, the above results can be easily understood.

These five fine structures of sub-domains in IR research explain the research status of the IR field during the whole period of 1987–1997. The keyword's position in the map indicates not only its location in IR research field and but also its relations with other keywords in the sub-domain research fields. Next, we will discuss the dynamic changes of these keywords' positions in the two separate periods (1987–1991 and 1992–1997).

3.2. Co-word analysis of 1987–1991

Among the 240 keywords, 47 keywords (Appendix A), that did not appear during this period, were excluded from this period of research. Thus, the remaining 193 keywords were chosen as the keyword research sample for this period. The same method was employed to generate the general overview map of the IR field in 1987–1991 by MDS (Fig. 7) and each cluster (sub-domain) was labelled by the most frequent keywords within the cluster as before.

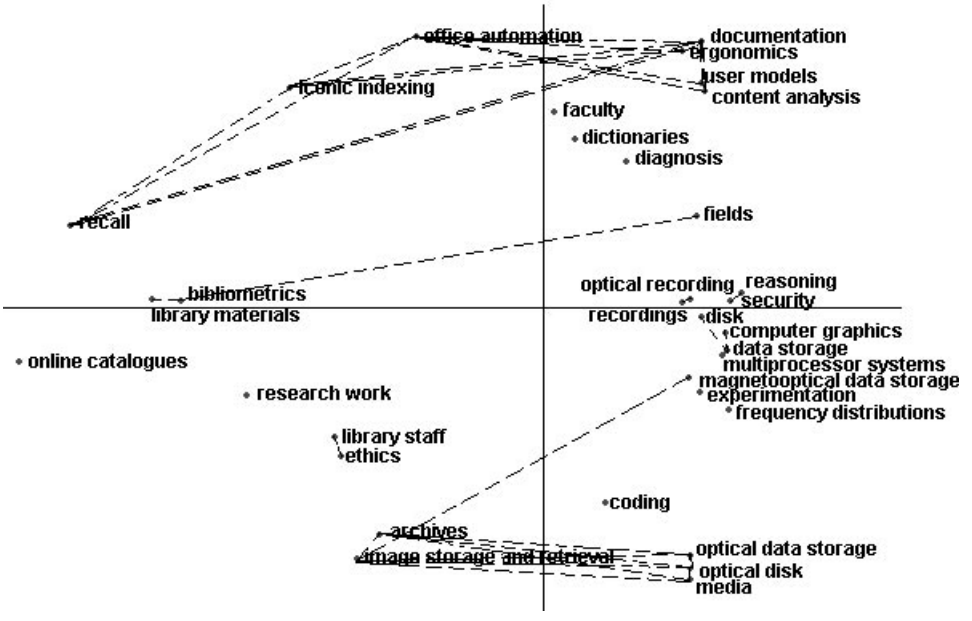

Fig. 8. The fine structure of Cluster 1 (C1) in 1987–1991 (the dotted line represents the link between two keywords with the Salton Index (>0.2)).

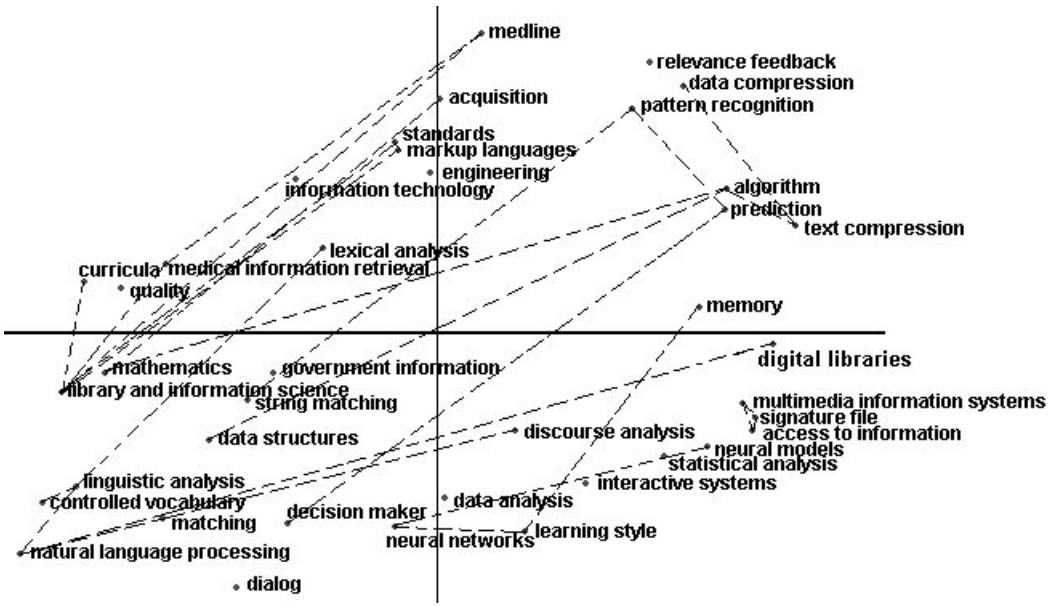

Fig. 9. The fine structure of Cluster 2 (C2) in 1987–1991 (the dotted line represents the link between two keywords with the Salton Index (>0.2)).

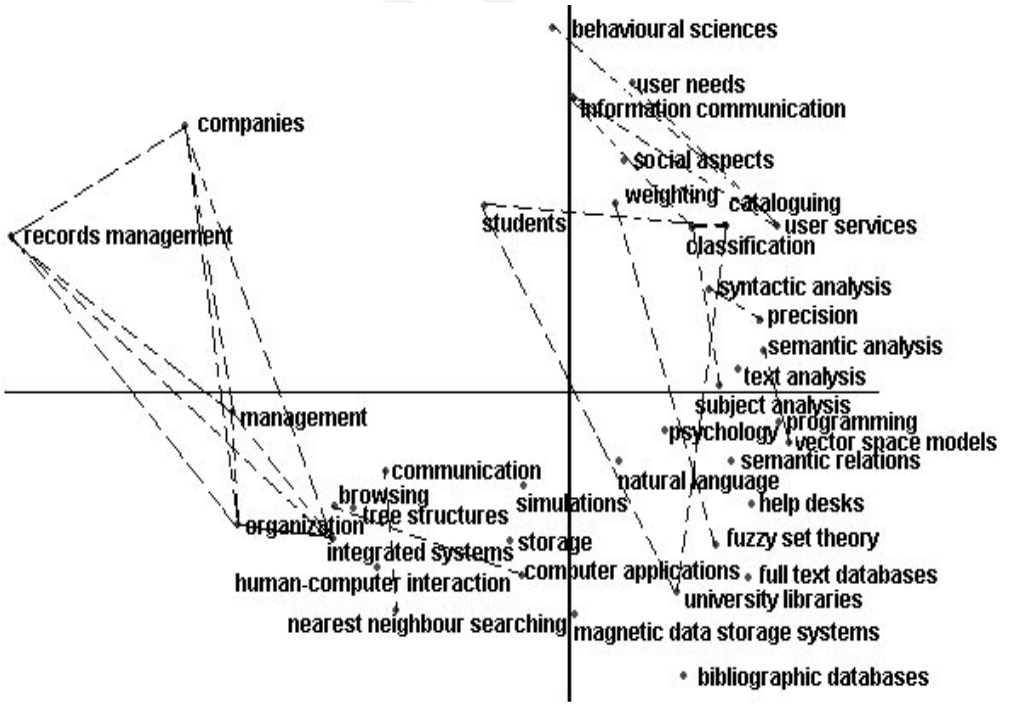

Fig. 10. The fine structure of Cluster 3 (C3) in 1987–1991 (the dotted line represents the link between two keywords with the Salton Index (>0.2)).

Each cluster contains around 50 keywords. In order to construct 'fine-structure' or detailed maps, each cluster was chosen as the input variable to map the sub-domain based on 193 × 193 correlation matrix. Thus, five detailed sub-domain maps (Figs. 8–12) were achieved to reflect the specific characters of each sub-domain in IR field.

Cluster 1 during this period (1987–1991) describes research topics relating to online catalogues, data storage, office automation, optical data storage and multiprocessor systems, and so on. Cluster 2 describes research topics relating to natural language processing, linguistic analysis, neural models, library and information science, memory, and so on. Cluster 3 describes topics on user services, cataloguing, storage, bibliographic databases, programming, and so on. Cluster 4 focuses on information storage and retrieval, databases, science and technology, parallel processing, information services, and so on. Finally, Cluster 5 appears to focus on information work, subject indexing, computerized information storage and retrieval, technical services, searching, and so on.

As shown in Table 2, during this period, for each cluster, the outer links among keywords are much more than its inner links. Around 88% of keywords in each cluster have outer links with other keywords based on the Salton Index. These indicate that the links among keywords based on the Salton Index do not aggregate within the cluster. In other words, around 62% of such kind of links are located among different clusters. So, these links not only reflect the abundant relationships among clusters, but also show a loosely internal composition in each cluster. But, the

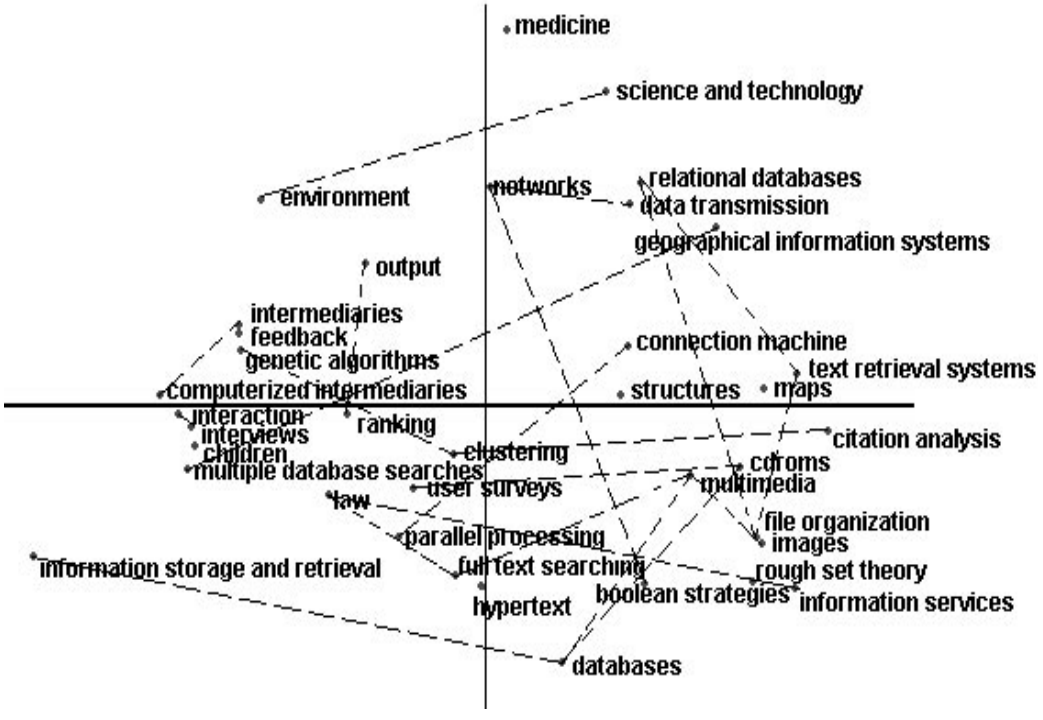

Fig. 11. The fine structure of Cluster 4 (C4) in 1987–1991 (the dotted line represents the link between two keywords with the Salton Index (>0.2)).

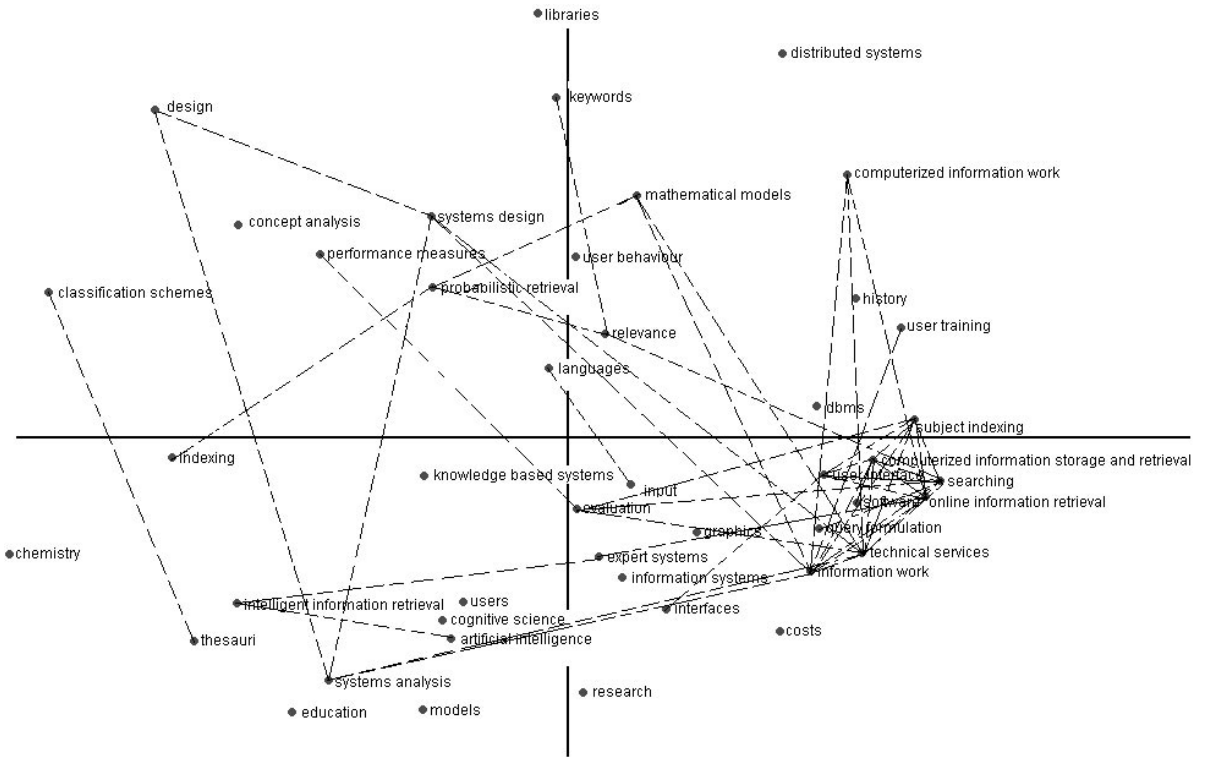

Fig. 12. The fine structure of Cluster 5 (C5) in 1987–1991 (the dotted line represents the link between two keywords with the Salton Index (>0.2)).

average link per keyword is very high during this period. It strongly suggests the evidence of the abundant links among keywords in these five clusters. During this period, Cluster 1 (not to be

Table 2
Comparison of five clusters during the period of 1987–1991[a]

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Average
Outer link	83	73	113	99	138	101.2
Inner link	66	52	48	42	98	61.2
Total link	149	125	161	141	236	162.4
Outer link %	56%	58%	70%	70%	58%	62%
Inner link %	44%	42%	30%	30%	42%	38%
Outer link key	28	33	38	31	40	34
Total key	34	39	38	37	45	38.6
Outer link key %	82%	85%	100%	84%	89%	88%
Average link/key	4.38	3.21	4.24	3.81	5.24	4.18
Centrality	0.340	0.300	0.311	0.301	0.276	–
Density	0.585	0.442	0.373	0.363	0.374	–

[a] *Note*: Here link denotes the link between keywords with the Salton Index (>0.2).

confused with Cluster 1 in 1987–1997) is the one with both highest centrality and density which indicates Cluster 1 is the important sub-area considered crucial by the IR community and it is able to maintain itself to develop over the course of time in IR field.

3.3. Co-word analysis of 1992–1997

During this period, among 240 keywords, only one keyword (i.e., neural models) did not appear so that this keyword was excluded from this period of research. Thus, 239 keywords were chosen as the keyword research sample in this period. The same method was used to generate the general overview map of the IR field in 1992–1997 by MDS (Fig. 13) and each cluster (sub-domain) was labeled by the most frequent keywords within the cluster as before.

Each cluster contains around 50 keywords. In order to construct 'fine-structure' or detailed maps as well, each cluster was chosen as the input variable to map the sub-domain based on 239×239 correlation matrix. Five detailed sub-domain maps (Figs. 14–18) were generated to reflect specific characters of each sub-domain in IR field.

Cluster 1 describes research topics relating to information storage and retrieval, searching, systems analysis, online information retrieval, database, and so on. Cluster 2 during this period describes research topics relating to networks, multimedia, WWW, medicine, Internet, and so on. Cluster 3 mainly describes topics on data storage, optical data storage, magnetic data storage, chemistry, storage, and so on. Keywords were located discretely in Cluster 4 that focuses on

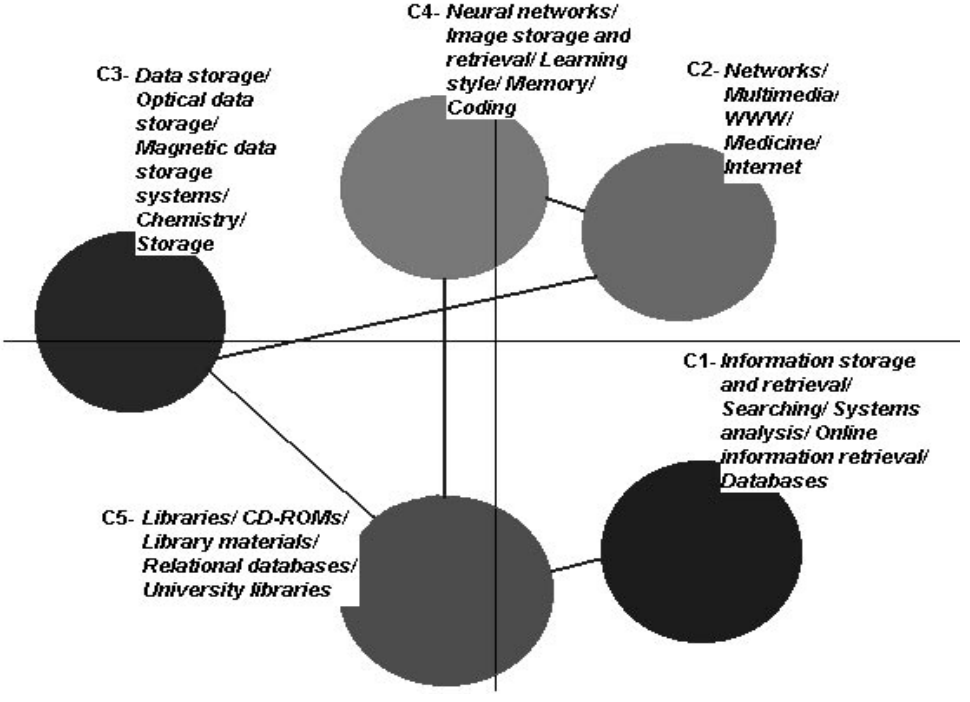

Fig. 13. General overview map of IR field in 1992–1997 (lines between clusters indicate strongest linkage according to the Salton Index).

neural networks, image storage and retrieval, learning style, memory, coding, and so on. Finally, Cluster 5 focuses on libraries, CD-ROM, library materials, relational databases, university libraries, and so on.

As shown in Table 3, during this period of 1992–1997, for each cluster, there are many more links among keywords than inner links. Around 46% of keywords in each cluster have outer links with other keywords based on. This indicates that the links among keywords based on Salton Index do not aggregate within the cluster. In other words, around 68% of such kind of links are located among different clusters. So, as in the period of 1987–1991, these links not only reflect the relationships among clusters, but also show a loosely internal composition in each cluster. During this period, Cluster 4 is the one with highest centrality indicating its strong linkage with other clusters. Topics in Cluster 4 are more related to inter-disciplinary areas such as neural networking, image storage and retrieval and so on. This is one of the reasons for Cluster 4 to appear with high centrality. Cluster 5 is the one with highest density. It means that topics in Cluster 5 have already form their own sub-fields with strong internal composition.

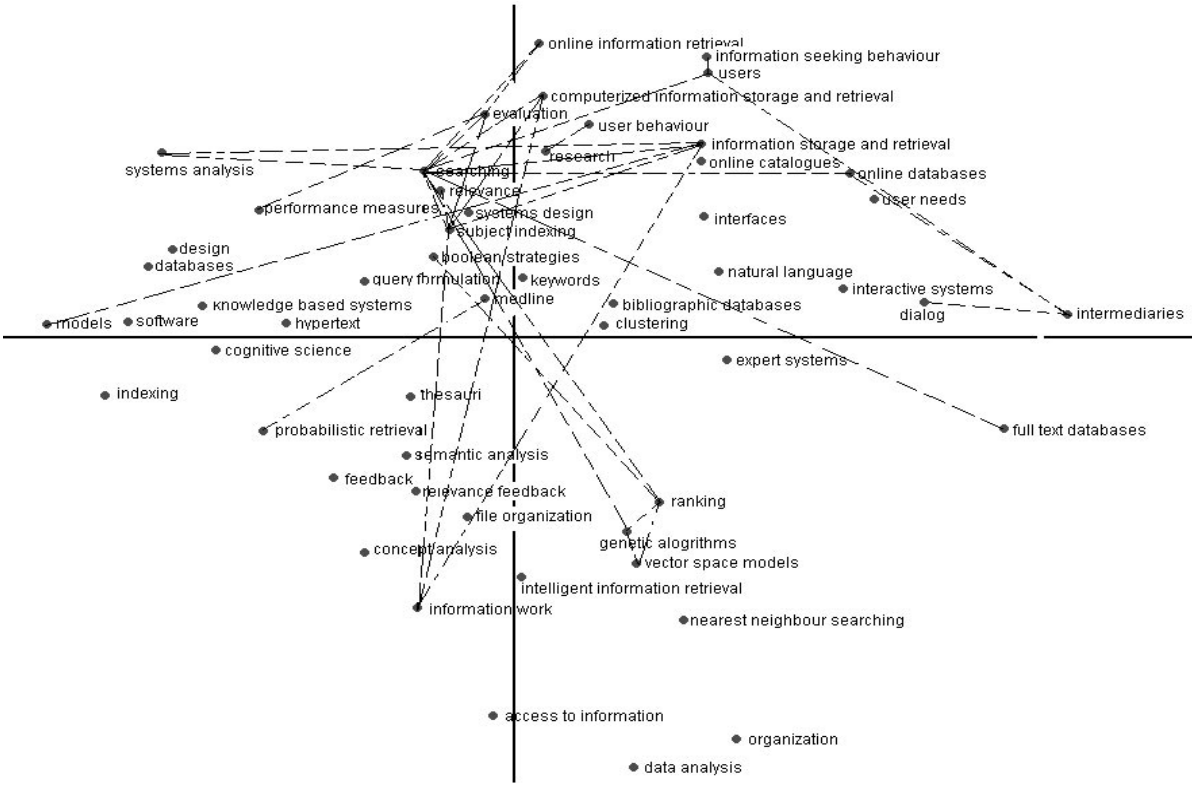

Fig. 14. The fine structure of Cluster 1 (C1) in 1992–1997 (the dotted line represents the link between two keywords with the Salton Index (>0.2)).

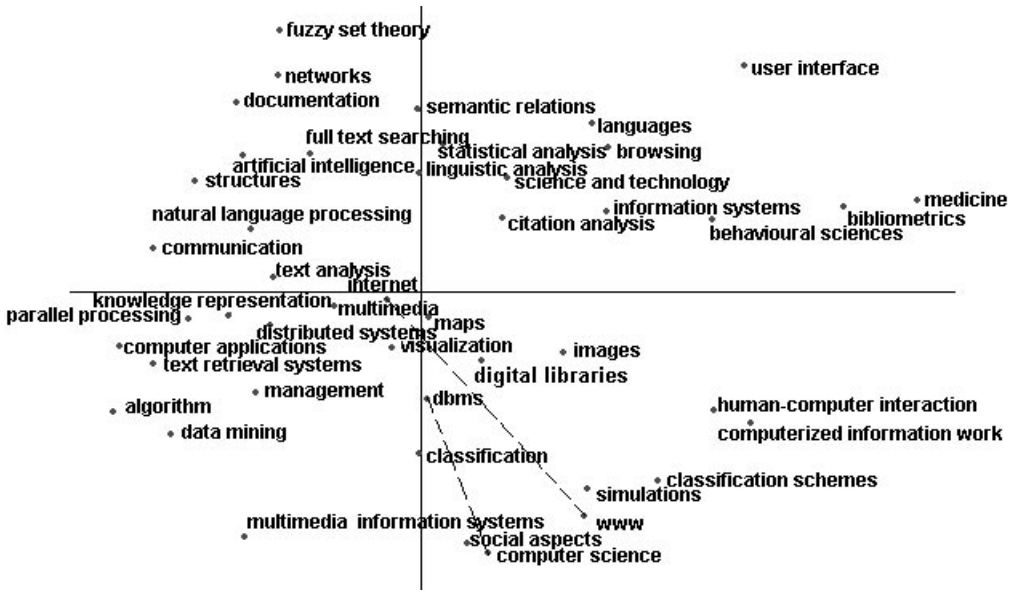

Fig. 15. The fine structure of Cluster 2 (C2) in 1992–1997 (the dotted line represents the link between two keywords with The Salton Index (>0.2)).

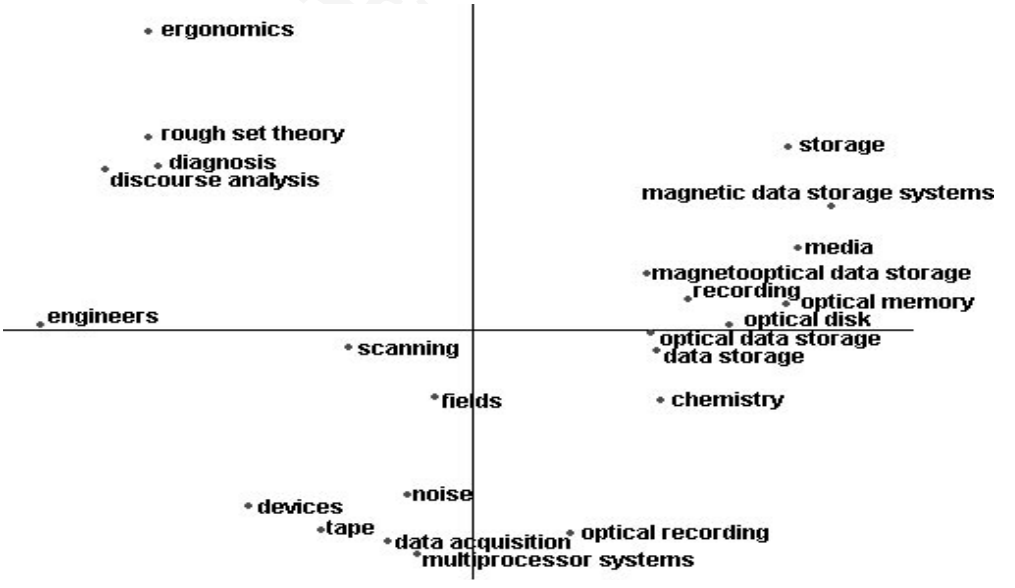

Fig. 16. The fine structure of Cluster 3 (C3) in 1992–1997 (the dotted line represents the link between two keywords with The Salton Index (>0.2)).

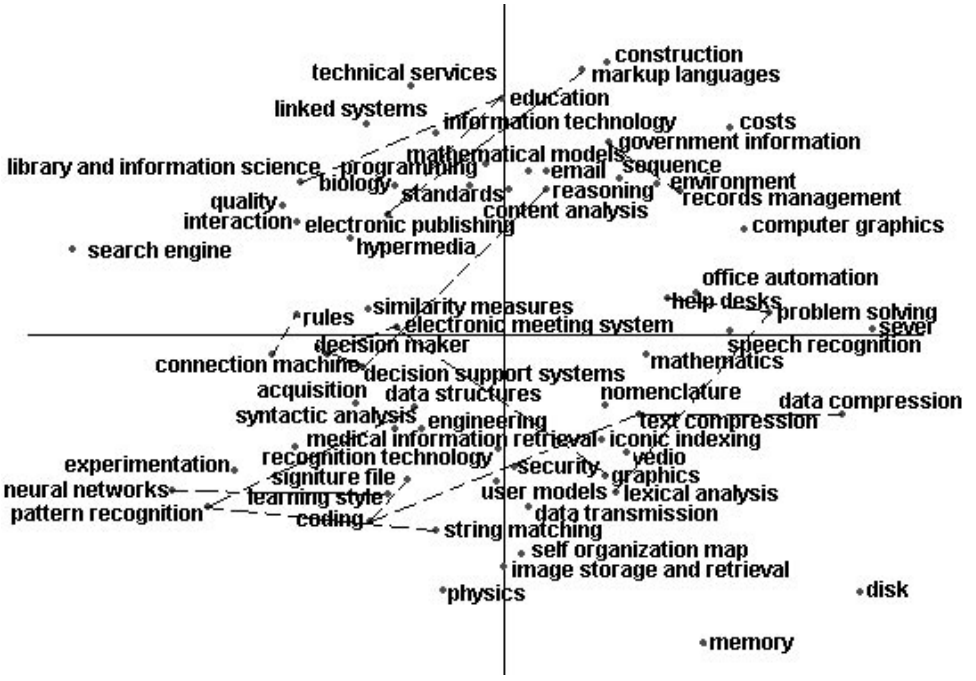

Fig. 17. The fine structure of Cluster 4 (C4) in 1992–1997 (the dotted line represents the link between two keywords with The Salton Index (>0.2)).

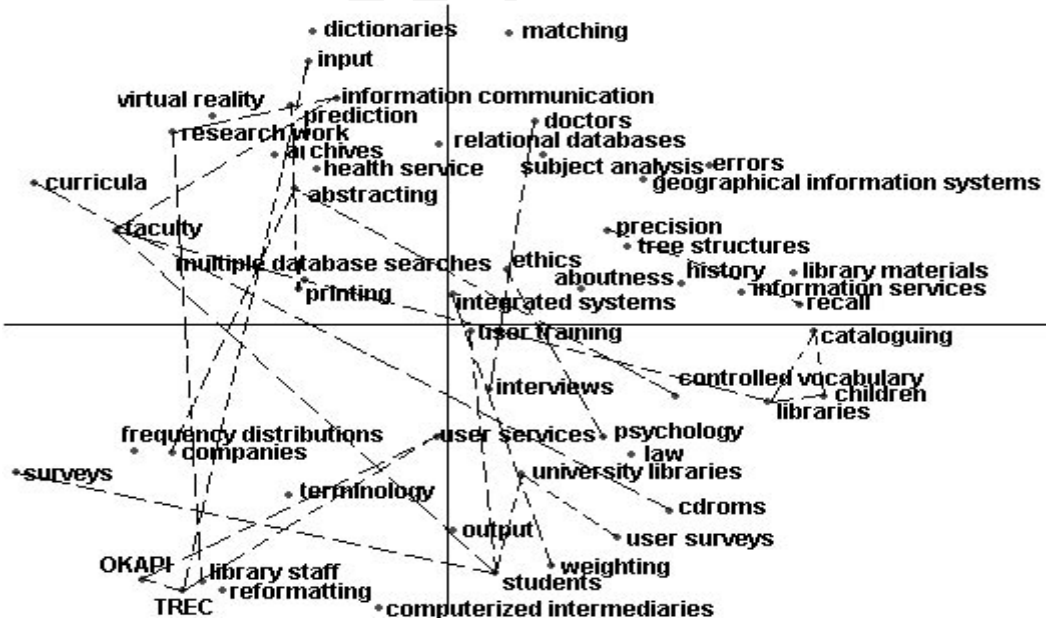

Fig. 18. The fine structure of Cluster 5 in (C5) 1992–1997 (the dotted line represents the link between two keywords with The Salton Index (>0.2)).

4. Discussion

Co-word analysis enables the structuring of data at various levels of analysis: as networks of links and nodes; as distributions of interacting networks; and as transformation of networks over time periods. These structures and changing relationships provide a basis for tracing IR researches. Co-word analysis reduces a large space of related descriptors to multiple related smaller spaces that are easier to comprehend but are also indicative of actual partitions of interrelated concepts in the literature under consideration.

What can we conclude about the state of IR field based on the co-word study of IR publications? First, the field is rapidly evolving, as is demonstrated by the increasing number of keywords in Internet, digital library, library networks and online database. The analysis of the 1987–1991 data shows a research trend focusing on traditional library science, library education, user theory, and information storage and retrieval. Consistent themes are evident over the second time period, but the focuses are moving towards data storage techniques, user needs, digital library, multimedia, networks, hypertext and so on. At the same time, some topics are growing dimmer, such as user services, technical services, information work, subject indexing and even computerized information storage. Some new areas have emerged during the second period, such as World Wide Web, Internet, information seeking behavior, online database, hypermedia, electronic publishing, artificial intelligence, knowledge representation, neural networks, information visualization, data mining, search engine, and so on (see Table 4).

While some IR themes seem to have well-defined genealogies, such as user theory, others appear to emanate from multiple preceding themes, such as intelligent information retrieval borrowing some knowledge from artificial intelligence, neural networks and so on; still others emerge quickly with little evidence of ancestry, such as World Wide Web, Internet, search engine and so on. So, the field has some established research themes, but it also changes rapidly to embrace new themes.

This study demonstrates the feasibility of co-word analysis as a viable approach for extracting patterns from, and identifying trends, in large corpora where the texts collected are from the same domain or sub-domain and are divided into roughly equivalent quantities for different time periods. Hence, this examination, of IR's emergence, points the way for further research into IR

Table 3
Comparison of five clusters during the period of 1992–1997[a]

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Average
Outer link	39	35	11	45	60	38
Inner link	64	4	0	34	46	29.6
Total Link	103	39	11	79	106	67.6
Outer link %	38%	90%	100%	57%	57%	68%
Inner link %	62%	10%	0%	43%	43%	32%
Outer link key	19	21	7	33	35	23
Total key	54	45	23	65	52	47.8
Outer link key %	35%	47%	30%	51%	67%	46%
Average link/key	1.91	0.87	0.48	1.22	2.04	1.3
Centrality	0.250	0.239	0.248	0.273	0.260	–
Density	0.288	0.305	0	0.257	0.319	–

[a] *Note*: Here link denotes the link between keywords with the Salton Index (>0.2).

Table 4
Comparison co-word analysis during the period 1987–1991 and 1992–1997[a]

Top topics in 1987–1991		Top topics in 1992–1997		Top Topics with positive increase from 1987–1991 to 1992–1997		Top Topics with negative increase from 1987–1991 to 1992–1997		Top Newly emerged topics from 1987–1991 to 1992–1997	
Topics	Freq	Topics	Freq	Topics	%	Topics	%	Topics	Freq
Information storage and retrieval	294	Information storage and retrieval	497	Optical data storage (4; 102)	2450	User services (21; 2)	−90	World Wide Web	38
Information work	151	Searching	229	Media (1; 25)	2400	Technical services (107; 14)	−87	Internet	35
Subject indexing	130	Systems analysis	155	Information technology (1; 23)	2200	Mathematical models (10; 2)	−80	Information seeking behaviour	26
Computerized information storage and retrieval	120	Online information retrieval	128	Magnetooptical data storage (1; 22)	2100	Information work (151; 39)	−74	Online databases	24
Technical services	107	Databases	109	Disk (1; 22)	2100	Companies (3; 1)	−67	Hypermedia	23
Searching	92	Data storage	105	User needs (1; 19)	1800	Help desks (3; 1)	−67	Biology	15
Online information retrieval	62	Models	104	Medline (1; 19)	1800	Subject indexing (130; 51)	−61	Electronic publishing	15
Systems analysis	56	Optical data storage	102	Genetic algorithms (1; 18)	1700	Psychology (7; 3)	−57	Server	14
Models	41	Computerized information storage and retrieval	72	Access to information (1; 18)	1700	Computerized information work (11; 5)	−55	Decision support systems	13
Databases	31	Performance measures	66	Digital libraries (1; 18)	1700	Intelligent information retrieval (26; 12)	−54	Knowledge representation	12
Intelligent information retrieval	26	Hypertext	62	Data storage (6; 105)	1650	Office automation (4; 2)	−50	Visualization	12

IPM 517

Table 4 (*continued*)

Top topics in 1987–1991		Top topics in 1992–1997		Top Topics with positive increase from 1987–1991 to 1992–1997		Top Topics with negative increase from 1987–1991 to 1992–1997		Top Newly emerged topics from 1987–1991 to 1992–1997	
Topics	Freq	Topics	Freq	Topics	%	Topics	%	Topics	Freq
User services	21	Evaluation	58	Multimedia information systems (1; 17)	1600	Rough set theory (2;1)	−50	Errors	11
Expert systems	19	Indexing	55	Image storage and retreival (3; 43)	1333	Input (2;1)	−50	Self-organizing map (SOM)	11
Evaluation	19	Knowledge based systems	53	Magnetic data storage systems (3; 43)	1333	Computerized information storage and retrieval (120; 72)	−40	Optical memory	10
Indexing	18	Users	52	Statistical analysis (1; 14)	1300			Recognition technology	10
Artificial intelligence	18	Design	51	Multimedia (3; 39)	1200			Video	10
Systems design	17	Subject indexing	51	Browsing (2; 26)	1200			Computer science	9
User interface	16	Networks	48	User models (1; 13)	1200			Linked systems	9
Design	15	Neural networks	48	Networks (4; 48)	1100			Sequence	9
Performance measures	15	Software	46	Coding (2; 24)	1100			Abstracting	8
Education	14	Magnetic data storage systems	43	Feedback (1; 12)	1100			Physics	8
Chemistry	14	Image storage and retreival	43	Recordings (1; 12)	1100			Similarity measures	7
Software	13	Information work	39	Medical information retrieval (2; 23)	1050			Construction	6
Probabilistic retrieval	13	Relevance	39	Acquisitions (1; 11)	1000			Data mining	6
Knowledge based systems	13	Multimedia	39	Security (1; 11)	1000			Search engine	6

Table 4 (*continued*)

Top topics in 1987–1991		Top topics in 1992–1997		Top Topics with positive increase from 1987–1991 to 1992–1997		Top Topics with negative increase from 1987–1991 to 1992–1997		Top Newly emerged topics from 1987–1991 to 1992–1997	
Topics	Freq	Topics	Freq	Topics	%	Topics	%	Topics	Freq
Research	13	Query formulation	40	Documentation (2; 23)	1050			Surveys	7
Query formulation	12	Learning style	38	Content analysis (1; 11)	1000			Trec	6
Computerized information work	11	World Wide Web	38	Relevance feedback (1; 11)	1000			Virtual reality	6
Relevance	10	Memory	38	Interactive systems (1; 11)	1000			Doctors	5
Mathematical models	10	Internet	35	Text analysis (3; 32)	967			Problem solving	5
Information systems	10	Medicine	35	Standards (1; 10)	900			Devices	4
Users	10	Systems design	34	Signature file (1; 10)	900			Electronic mail	4
Cognitive science	9	Information systems	33	Behavioural sciences (1; 10)	900			Health service	4
Cataloguing	9	Probabilistic retrieval	32	Quality (1; 10)	900			Rules	4
Science and technology	9	Chemistry	32	Neural networks (5; 48)	860				
Natural language processing	9	Text analysis	32	Learning style (4; 38)	850				
Interfaces	9	Computer applications	32	Library materials (2; 18)	800				
		Online catalogues	31	Optical recording (1; 9)	800				
		Research	31	Hypertext (7; 62)	786				

[a] *Notes*: Two numbers in the bracket, the first one represents the occurrence frequency of that word or phrase during 1987–1991; the second one presents the occurrence frequency of that word or phrase during 1992–1997.

IPM 517

field and well-restricted sub-domains. On the other hand, co-word analysis successfully visualizes the inter-relations of the keywords and sub-fields of IR, while the importance of visualizing methods in the convincing presentation of results has not been sufficiently understood in the past. Co-word analysis opens a new opportunity for cartography of science and information visualization. The co-word results have produced a great deal more than statistical artifact. We aimed to exploit the visualization effect of the co-word maps to the aid of searchers in the IR domain, and the results are quite encouraging. A separate paper on this endeavor is currently under preparation and will appear soon.

Overall, this study has led us to an increased confidence in the co-word analysis. As Law and Whittaker (1992) pointed out "looked at in the light of co-word analysis thus makes a modest claim: it notes that it is indeed dependent on its context but, by virtue of this fact, claims a degree of sensitivity to the nuances of scientific context."

Acknowledgements

The authors wish to thank Professor Ronald Rousseau (KHBO, Department of Industrial Sciences and Technology, Belgium) for his valuable comments on an earlier draft of this paper. We are also grateful for useful suggestions from Dr. Ed Noyons (CWTS, Leiden University, The Netherlands).

Appendix A

Forty-seven keywords which did not appear together with any other keyword in the research sample during the first period (1987–1991) are the following:

Aboutness	Abstracting	Biology
Computer science	Construction	Data acquisition
Data mining	Decision Support systems	Devices
Doctors	Electronic meeting system	
Electronic publishing	Electronic mail	Engineers
Errors	Health service	Hypermedia
Information seeking behaviour		Internet
Linked systems	Noise	Nomenclature
Okapi	Online databases	Optical memory
Physics	Printing	Problem solving
Recognition technology	Reformatting	Rules
Scanning	Search engine	Sequence
Similarity measures	Self-organizing map (SOM)	
Speech recognition	Surveys	Tape
Terminology	Trec	Video
Virtual reality	Visualization	World Wide Web

References

Arabie, P., Carroll, J. D., & DeSarbo, W. S. (1987). *Three-way scaling and clustering*. Newbury Park: Sage Publications.

Bhattacharya, S., & Basu, R. K. (1998). Mapping a research area at the micro level using co-word analysis. *Scientometrics*, *43*(3), 359–372.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, *22*(1), 153–205.

Cambrosio, A., Limoges, C., Courtial, J. P., & Laville, F. (1993). Historical scientometrics? Mapping over 70 years of biological safety research with co-word analysis. *Scientometrics*, *27*(2), 119–143.

Coulter, N., Monarch, I., & Konda, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science*, *49*(13), 1206–1223.

Courtial, J. P. (1994). A coword analysis of scientometrics. *Scientometrics*, *31*(3), 251–260.

Courtial, J. P., Cahlik, T., & Callon, M. (1994). A model for social interaction between cognition and action through a key-word simulation of knowledge growth. *Scientometrics*, *31*(2), 173–192.

Courtial, J. P, Callon, M., & Sigogneau, A. (1993). The use of patent titles for identifying the topics of invention and forecasting trends. *Scientometrics*, *26*(2), 231–242.

Ding, Y., Chowdhury, G., & Foo, S. (1999a). Mapping intellectual structure of information retrieval: An author cocitation analysis, 1987–1997. *Journal of Information Science*, *25*(1), 67–78.

Ding, Y., Chowdhury, G., & Foo, S. (1999b). Mapping the development in information retrieval specialty: A bibliometric analysis via journals. In. Macias-Chapula, C.A. (Ed). *Seventh Conference of the International Society for Scientometrics and Informetrics-Proceedings* (pp. 139–149). Mexico, 5–9 July 1999.

Ding, Y., Chowdhury, G.G., & Foo, S. (2000). Journal as markers of intellectual space: Journal co-citation analysis of information retrieval area, 1987–1997. *Scientometrics*, *47*(1).

Hinze, S. (1994). Bibliographical cartography of an emerging interdisciplinary discipline: The case of bioelectronics. *Scientometrics*, *29*(3), 353–376.

Honkela, T., Kaski, S. Lagus, K., & Kohonen, T. (1996). *Newsgroup exploration with WEBSOM method and browsing interface*. Report A32, Laboratory of Computer and Information Science, Helsinki University of Technology, Finland.

King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, *13*, 261–276.

Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer.

Kopcsa, A., & Schiebel, E. (1998). Science and technology mapping: A new iteration model for representing multidimensional relationships. *Journal of the American Society for Information Sicnece*, *49*(1), 7–17.

Law, J., & Whittaker, J. (1992). Mapping acidification research: A test of the co-word method. *Scientometrics*, *23*(3), 417–461.

Looze, M. D., & Lemarie, J. (1997). Corpus relevance through co-word analysis: An application to plant proteins. *Scientometrics*, *39*(3), 267–280.

Norusis, M.J. (1997). *SPSS 7.5 guide to data analysis*. Upper Saddle River, NJ: Prentice-Hall.

Noyons, E.C.M. (1998). Personal communication.

Noyons, E. C. M., & van Raan, A. F. J. (1994). Bibliometric cartography of scientific and technological development of an R & D field. *Scientometrics*, *30*(1), 157–173.

Noyons, E. C. M., & van Raan, A. F. J. (1998a). 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.

Noyons, E. C. M., & van Raan, A. F. J. (1998b). Advanced mapping of science and technology. *Scientometrics*, *41*(1–2), 61–67.

Peters, H. P. F., & van Raan, A. F. J. (1993a). Co-word based science maps of chemical engineering, Part I: Representations by direct multidimensional scaling. *Research Policy*, *22*, 23–45.

Peters, H. P. F., & van Raan, A. F. J. (1993b). Co-word based science maps of chemical engineering. Part II: Combined clustering and multidimensional scaling. *Research Policy*, *22*, 47–71.

Polanco, X., Francois, C., & Keim, J. P. (1998). Artificial neural network technology for the classification and cartography of science and technology information. *Scientometrics*, *41*(1–2), 69–82.

Rikken, P., Kiers, H. A. L., & Vos, R. (1995). Mapping the dynamics of adverse drug reactions in subsequent time periods using INDSCAL. *Scientometrics*, *33*(3), 367–380.

Rip, A., & Courtial, J. P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, *6*, 381–400.

Small, H. (1973). Cocitation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*, 265–269.

Tijssen, R. J. W. (1993). A scientometric cognitive study of neural network research: expert mental maps versus bibliometric maps. *Scientometrics*, *28*(1), 111–136.

van Raan, A. F. J. (1997). Scientometrics: State-of-the-art. *Scientometrics*, *38*(1), 205–218.

van Raan, A. F. J., & Tijssen, R. J. W. (1993). The neural net of neural network research. *Scientometrics*, *26*(1), 169–192.

Voutilainen, A. (1993). NPtool. A detector of English noun phrases. In: *Proceedings of the workshop on very large corpora Columbus*. Ohio: Ohio State University, 22 June 1993.

Widhalm, C. (1999). Personal Communication, November 1999.