Journal of Information Science, 26(6), 429-452

Incorporating the results of co-word analyses to increase search variety for information retrieval

Ying Ding, Gobinda G. Chowdhury, Schubert Foo

Division of Information Studies, School of Applied Science
Nanyang Technological University, Nanyang Avenue, Singapore 639798
{p143387632, asggchowdhury, assfoo}@ntu.edu.sg

Abstract

This research aims to incorporate the results of co-word analysis into information retrieval as a means to increase search variety for end users in the domain of information retrieval. Relevant data were first collected from the Science Citation Index and Social Science Citation Index for the period of 1987-1997. The results of co-word analysis on the data were compared with similar data obtained from three thesauri, namely, the LISA thesaurus, LCSH (Library Congress Subject Heading) and the Thesaurus of Information Technology Terms. The differences detected between them indicate that the search variety may be increased by combining co-word analysis with the use of traditional thesauri. Subsequently, the results of co-word analysis were compared with each other for two different periods (1987-1991 and 1992-1997). The changes among them were identified implying co-word analysis may be used to directly identify dynamic changes in its chosen domain area, thereby providing better up-to-date information to aid the information search process.

Introduction

The large amounts of information available through online databases, Internet and other networks are changing the way we gather, process, and retrieve information. However, gaining access to such information is often difficult as a result of inconsistency involved in the processing of information and the way queries is expressed by searchers.

Bates (1986 & 1998) pointed out the gap between the end user and the indexer which guarantees the mismatches between user search terms and indexing terms on the same records. Although there are some ways to fill up the gap, like maintaining the consistency by enhancing the standards between indexers and the users, developing user-friendly interface to link both the indexer and users together, problems still exist on full-text searching, natural language searching, and so on. Researches also demonstrated that end users like to use a very wide range of different terms and none of those terms will occur very frequently (Saracevic and Kantor, 1988). Experienced online database searchers also know that if they do a thorough search, they need to use as many different terms and term variants as possible and they will scan various thesauri from the subject area and enter all the relevant terms they could find (Bates, 1986 & 1998). So increasing search variety for the end users is an important aspect to succeed in information retrieval (Bates, 1986 & 1998). A number of researches have been conducted or are developing to increase the search variety for the end users (Gomez, Lochbaum and Landauer, 1990; Peat and Willett, 1991; Chen & Ng, 1995; Byrne and McCracken, 1999).

The commonality between bibliometrics and information retrieval is not apparent. Bibliometrics deals with research products generated by researchers and scholars, while information retrieval is principally concerned with information storage, search and retrieval. However, more holistic approaches to research in bibliometrics and information retrieval would be likely to add to our understanding of research problems in both the fields (Harter, 1992). Harter & Cheng (1996) applied the co-citation concept in bibliometrics to information retrieval and generated a new method or concept: colinked descriptor. Shalini (1993) has utilized citation profiles to improve relevance in a two-stage retrieval system. Quoniam et al. (1998) have employed Zipf's law into information retrieval to get a first impression of documents data set by querying without any keyword.

This paper proposes that the results of a co-word analysis may be used to generate search variety for the end-users. The keyword sets generated through a co-word analysis in the domain of information retrieval, are compared with the corresponding keyword sets obtained from three thesauri, namely, the LISA thesaurus, LCSH (Library Congress Subject Heading) and the Thesaurus of Information Technology Terms. The keyword clusters in the same subject domain generated during two different time periods have been compared to identify the changes. Based on the findings, it is apparent that the results of a co-word analysis can produce keyword sets that are different from those that are obtained from traditional thesauri, and that the results of co-word analyses in the same subject domain produced different results over different time periods. From this, it becomes evident that the results of co-word analyses may be used to yield better search variety in an information retrieval environment.

Background

Search variety

In an information retrieval system, the users usually use keywords or controlled terms (index terms) or terms from full-text articles or natural language to formulate their queries and fulfill their information needs. If inappropriate, incorrect, or an insufficient variety of words are used to form the queries or index the records in the system, the users may not be able to find the objects they desire (Aitchison & Gilchrist, 1997).

The primary techniques available to identify good names for stored objects are manual indexing and automatic text analysis (Bates, 1986 & 1998). The basic assumption underlying is that "If an index term is good at discriminating relevant from non-relevant documents then any closely associated index term is also likely to be good at this" (Van Rijsbergen, 1979). But many studies have found disappointingly low agreement in the assignment of indexing terms and the mismatches between the terms users employed and the terms the indexer adopted (Bates, 1986 &1998; Saracevic and Kantor, 1988). Bates pointed out that "in order to succeed in information retrieval, the searcher need to generate as much variety in a search formulation as there is variety in the indexing of the topic of interest" (Bates, 1986). However, most current systems do little to help the searcher generate the search variety (Bates, 1998).

In order to generate the search variety for the end-users, traditional thesauri, automatic thesauri and term co-occurrence researches have become the research focus. While traditional thesauri have been used to generate search varieties in information retrieval systems, there have been a number of serious shortcomings:

- Different indexers might assign index terms for a given document differently (Bates 1986).

- Searchers tend to use different terms for the same information needs (Chen & Dhar, 1991).
- It is difficult to let searchers, who are not familiar with the specific subject area and/or terminology of a database, articulate their information needs accurately so that it can be translated into relevant queries. This has always been a classical and pressing question in information science research (Chen & Dhar, 1991; Quoniam, et al., 1998);
- The structure of the thesauri, in particular the relationships among descriptors, is also questioned by IR researchers (Harter & Cheng, 1996)
- As more and more new concepts, methods, theories or new sub-domains continue to emerge in most domains, building up or amending thesauri to make it up-to-date and relevant is an extremely time-consuming and labor-intensive task..

Many research groups have created automatically-generated thesaurus components, akin to a manually created thesaurus, that have played an important role in solving searchers' vocabulary problems during information retrieval (For example, Chen & Dhar, 1991; Chen & Ng, 1995; Chen & Lynch, 1992; Chen, Ng, Martinez & Schatz, 1997; Chen, Martinez, Kirchhoff, Ng & Schatz, 1998; Chen, Yim, Fye & Schatz, 1995).

Virtually all techniques for automatic thesaurus generation are based on the statistical co-occurrence of word types in text (Chen & Lynch, 1992; Crouch, 1990; Salton, 1989). For example, the specific algorithms in Chen's research include: term filtering, automatic indexing and cluster analysis. Based on Everitt's (1980) cluster analysis, Salton's Vector Space Model (Salton, Wong & Yang, 1975) has been adopted in Chen's automatic thesaurus generation techniques. In these techniques, the most commonly used algorithms compute probabilities of terms co-occurring in all documents of a database.

Extending co-word analysis to Information Retrieval

Usually, researchers use co-word techniques to analyze papers in order to identify keywords that describe their research content and link papers by the degree of co-occurrence of these keywords to produce a 'map index' of a specialty (King, 1987). The traditional co-word analysis techniques have been applied in a number of studies, e.g. (1)

using the occurrence of particular concepts such as “information theory” in materials, as tracers of the influence of one researcher or group of researchers on other researchers (King, 1987); (2) using high-frequency concepts to profile the concerns of a field of research (Gregory, 1983); (3) using longitudinal shifts in concept clusters to characterize the succession of theoretical paradigms in fields of research (Chu, 1992 and Lau, 1995); (4) using article information content to re-evaluate scientific productivity (Seglen, 1996; Noyons, Moed & Luwel, 1999); (5) studying disciplinary formation processes and disciplinary functions (Borgman, 1990; Coulter, Monarch and Konda, 1998); (6) mapping the structure aspects and dynamic aspects of scientific research on the level of research specialties and tracing the history of specific domain (Braam, Moed and van Raan, 1991; Law and Whittaker, 1991; Cambrosio, et al., 1993); (7) Stimulating knowledge growth and development according to a local positive feed-back rule within small sets of word associations (Courtial, Cahlik and Callon, 1994; Courtial, 1994). All these co-word applications just follow the traditional co-word research to analyze disciplinary development or knowledge growth. They do not touch the idea of extending co-word analysis into information retrieval field to generate helpful tools for user to retrieve information.

Recently, some researchers have conducted research and tried to apply co-word analysis techniques to information retrieval field. Peters, Braam and van Raan (1995) measured word-profile similarities between citing and cited publications and found that publications with a citation relationship are significantly more content-related than other publications. De Looze and Lemarie (1997) analyzed different corpuses by means of co-word analysis in plant proteins. These co-word studies show the trend of applying co-word analysis into information retrieval field.

Method

In this study, we will compare the word clusters generated by co-word analysis with corresponding word blocks obtained from the traditional thesauri. The research domain is Information Retrieval (IR) itself. One important characteristic of our study is that we will include the ‘time’ dimension of the documents and concepts in our research. This is so since almost all automatic thesaurus researches so far have not considered this ‘time’ dimension in their studies. Chen, Yim, Fye and Schatz (1995) believe that by time-tagging each concept and analyzing the activities associated with it (i.e. when it first appeared, when it was most actively used, etc.), a more fluid and time-precise thesaurus can be created and this will add valuable information to automatic thesaurus research and also improve information retrieval results. In this research, we have tried to build up the term clusters or blocks (as in a thesaurus) based on co-word analysis, while also considering the ‘time’ dimension.

Data collection

A DIALOG search was conducted on the SCI (Science Citation Index) and SSCI (Social Science Citation Index) to retrieve literature on information retrieval. The search statement was created carefully to retrieve documents on all the different aspects of information retrieval. A total of 3,325 items were retrieved covering the period of 1987-1997. A number of these articles without abstracts, 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.

Keyword Standardization

A total of 3,227 unique keywords were collected from the chosen 2,012 articles 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. One of the major problems was that the keywords were not standard. Three thesauri, namely, the LISA thesaurus, LCSH (Library Congress Subject Heading) and Thesaurus of Information Technology Terms were used in combination in an attempt to make the keywords consistent (singular/plural), unified (synonyms), and as far as possible, unambiguous (homonyms). Three thesauri were used since a single thesaurus was insufficient to cater to all the identified keywords. Thus, the various keywords and phrases were standardized by selecting an appropriate heading from the vocabulary tools. The following examples illustrate how the keywords and phrases were standardised.

- 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.

This process also helped to reduce the number of keywords and phrases significantly. Words with a word frequency of one or two were merged with those terms that either broader or similar to them. Words with frequency of one or two, which did not have any broader 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. In order to find out whether the features of these word clusters change over time, 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).

Foxpro programs were written to calculate the number of times two keywords appear together in the same publication. Thus, a co-occurrence matrix of 240*240 keywords was formed. The cell of keyword X and keyword Y stores 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 Pearson's correlation coefficient indicating the similarity and dissimilarity of each keyword pair. We recalculated the co-occurrence frequency with the Salton Index (Hamers, et. al., 1989) that has a value of more than 0.2. 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. In other words, this is used to eliminate high frequency words that can be linked to almost every other keyword in the research sample (Noyons, 1998, 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 20 co-occurring words (20s) with high Salton Index. These 20 keywords were then compared (i.e. checked whether they were present or absent) with the set of terms that appear in the three selected traditional thesauri (TT). We combined all the terms that appear with a given keyword in all the three thesauri. We did this because if we had taken the terms appearing only in one thesaurus, then the number of keywords appearing with a given search term would be very few. In other words, combining the keywords that appear along with a given keyword in all the three thesauri will yield as long a list as possible. We compared the two sets of keywords – the one obtained through the co-word analysis versus the one obtained from the thesauri, by presenting them side by side in order to check how similar or dissimilar they were. Our proposition is that the keyword set generated by the co-word analysis shows the co-occurrences of keywords in the literature, and therefore they may be help the end users select appropriate search terms in a given search session. Furthermore, if many of these keywords do not appear in the list of keywords obtained from the three thesauri, then it would be definitely useful to use the results of such co-word analysis to add search varieties in an information retrieval session.

Co-word analysis and traditional thesauri

For 240 keywords in the research sample, 24 keywords did not have any semantic descriptors from the three traditional thesauri. Thus, 216 (90%) keywords in the research sample have semantic descriptors from three traditional thesauri. So these 216 keywords in the research sample were used for comparison.

Results of co-word analysis for 1987-1997 compared with traditional thesauri

During this period, out of these 216 keywords, only 102 (47.2%) keywords were found to have common words in their corresponding 20s and TTs, and the average similarity was 7.9% (Figure 1). In those 102 keywords, 60 have 5% common words, 5 have 20% common words, and 2 have 25% common words. The *TTs* and *Keywords' 20s* of two keywords viz. *Expert systems* and *information storage and retrieval* have 25% common words. The keywords with 20% common words in their 20s and TT are: *cataloguing, artificial intelligence, multiprocessor systems, information services* and *performance measures.* All of these are high frequency keywords in this study.

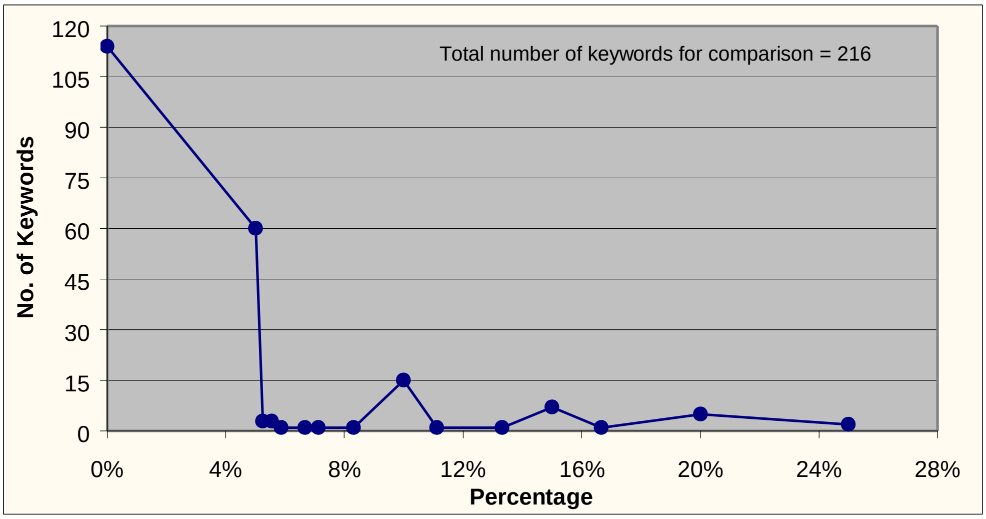

Figure 1. Comparison of co-word analysis and traditional thesaurus in 1987-1997

Figures 2 to 8 show the keywords' 20s and TT for each of the seven keywords that have more than or equal to 20% common words in its 20s and TT. In these figures, * represents the thesaurus entry from TITT; ~ represents the thesaurus entry from LISA and ' represents thesaurus entry from LCSH. Keywords shown shaded in the figures are common in both the co-word analysis results and traditional thesauri.

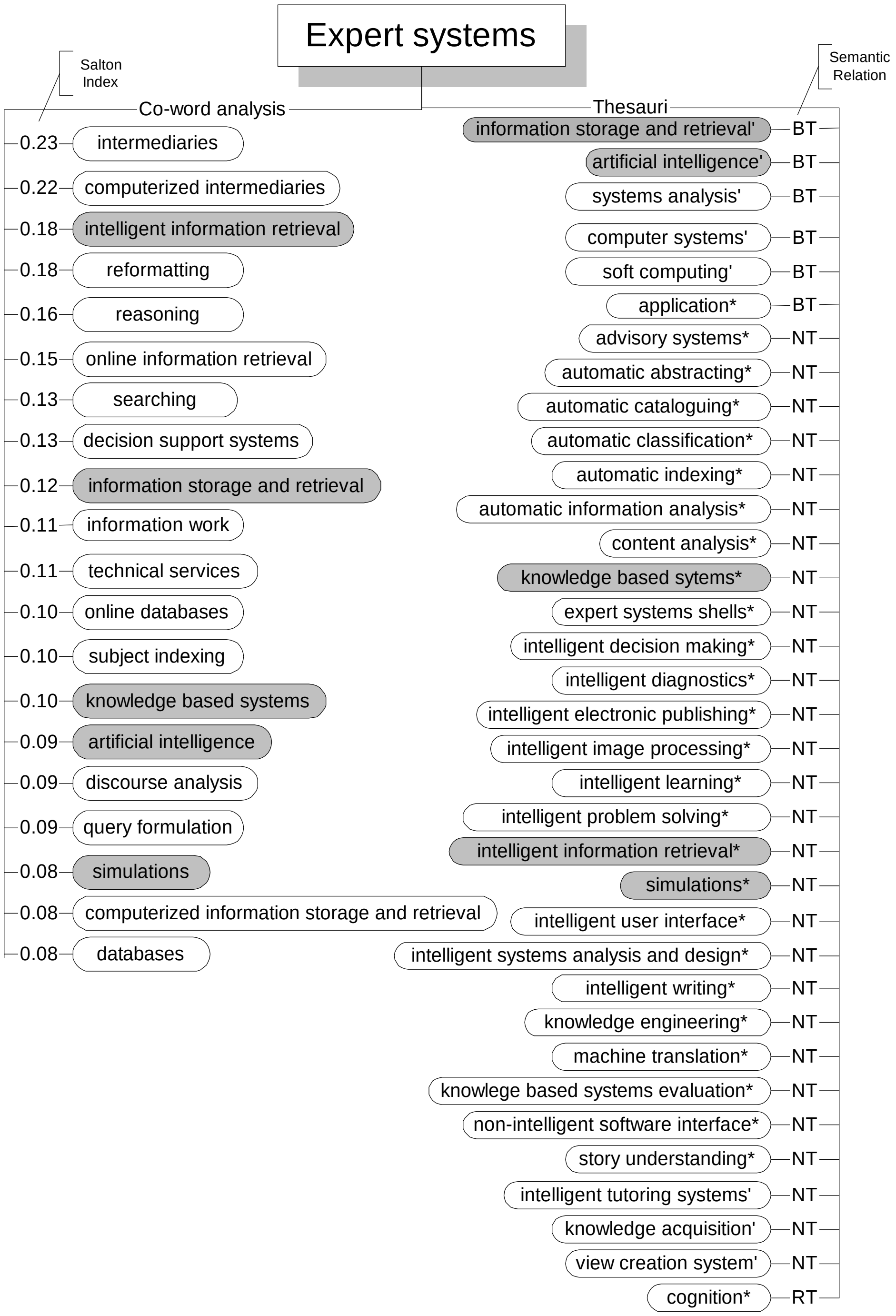
Expert systems
Salton Index
Co-word analysis
Thesauri
Semantic Relation
0.23 intermediaries
0.22 computerized intermediaries
0.18 intelligent information retrieval
0.18 reformatting
0.16 reasoning
0.15 online information retrieval
0.13 searching
0.13 decision support systems
0.12 information storage and retrieval
0.11 information work
0.11 technical services
0.10 online databases
0.10 subject indexing
0.10 knowledge based systems
0.09 artificial intelligence
0.09 discourse analysis
0.09 query formulation
0.08 simulations
0.08 computerized information storage and retrieval
0.08 databases
information storage and retrieval' BT
artificial intelligence' BT
systems analysis' BT
computer systems' BT
soft computing' BT
application* BT
advisory systems* NT
automatic abstracting* NT
automatic cataloguing* NT
automatic classification* NT
automatic indexing* NT
automatic information analysis* NT
content analysis* NT
knowledge based sytems* NT
expert systems shells* NT
intelligent decision making* NT
intelligent diagnostics* NT
intelligent electronic publishing* NT
intelligent image processing* NT
intelligent learning* NT
intelligent problem solving* NT
intelligent information retrieval* NT
simulations* NT
intelligent user interface* NT
intelligent systems analysis and design* NT
intelligent writing* NT
knowledge engineering* NT
machine translation* NT
knowlege based systems evaluation* NT
non-intelligent software interface* NT
story understanding* NT
intelligent tutoring systems' NT
knowledge acquisition' NT
view creation system' NT
cognition* RT

Figure 2. Comparison of expert systems' 20s and TT in 1987-1997

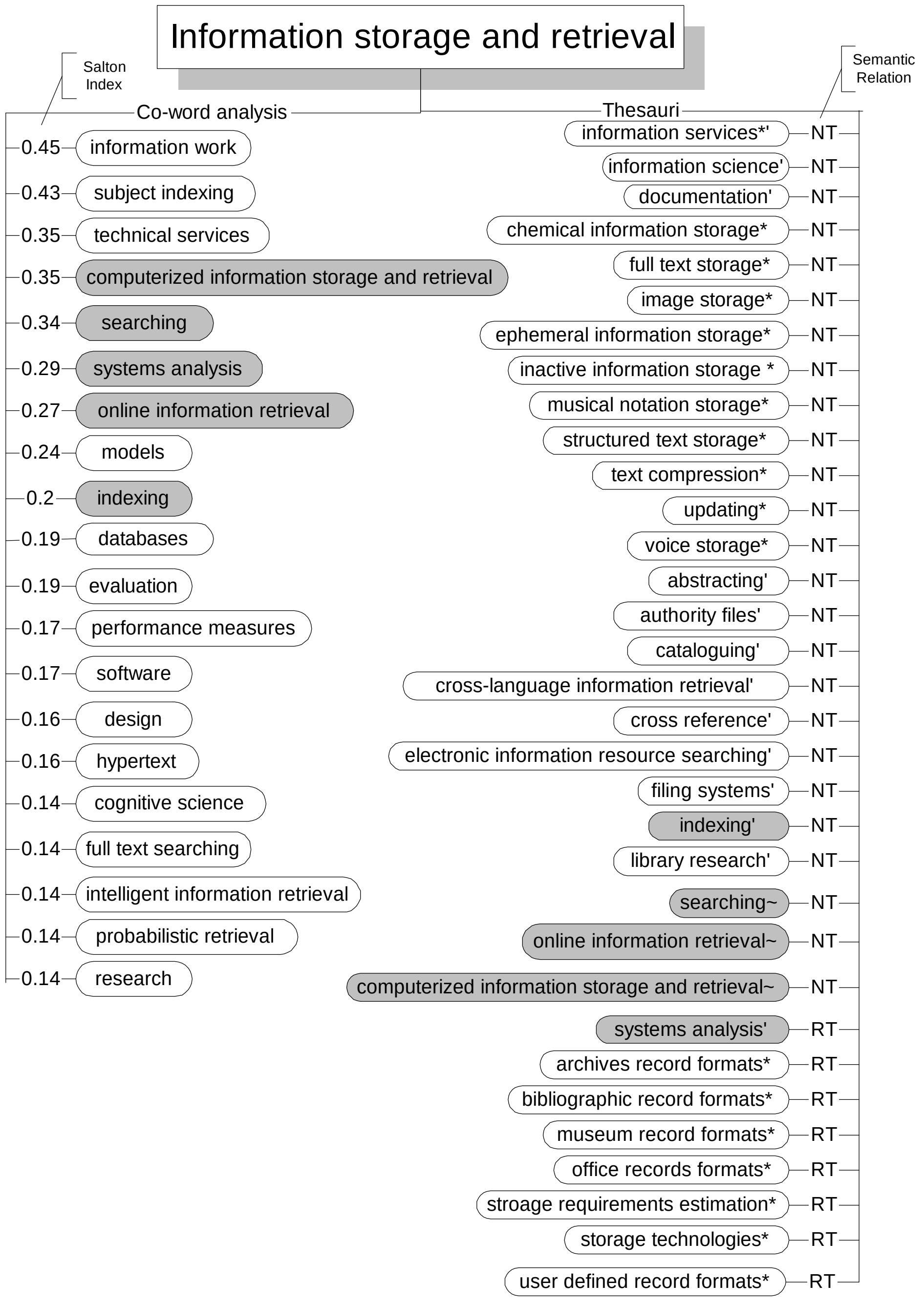
Information storage and retrieval
Salton Index
Semantic Relation
Co-word analysis
Thesauri
0.45 information work
0.43 subject indexing
0.35 technical services
0.35 computerized information storage and retrieval
0.34 searching
0.29 systems analysis
0.27 online information retrieval
0.24 models
0.2 indexing
0.19 databases
0.19 evaluation
0.17 performance measures
0.17 software
0.16 design
0.16 hypertext
0.14 cognitive science
0.14 full text searching
0.14 intelligent information retrieval
0.14 probabilistic retrieval
0.14 research
information services*' NT
information science' NT
documentation' NT
chemical information storage* NT
full text storage* NT
image storage* NT
ephemeral information storage* NT
inactive information storage * NT
musical notation storage* NT
structured text storage* NT
text compression* NT
updating* NT
voice storage* NT
abstracting' NT
authority files' NT
cataloguing' NT
cross-language information retrieval' NT
cross reference' NT
electronic information resource searching' NT
filing systems' NT
indexing' NT
library research' NT
searching~ NT
online information retrieval~ NT
computerized information storage and retrieval~ NT
systems analysis' RT
archives record formats* RT
bibliographic record formats* RT
museum record formats* RT
office records formats* RT
stroage requirements estimation* RT
storage technologies* RT
user defined record formats* RT

Figure 3. Comparison of information storage and retrieval' 20s and TT in 1987-1997

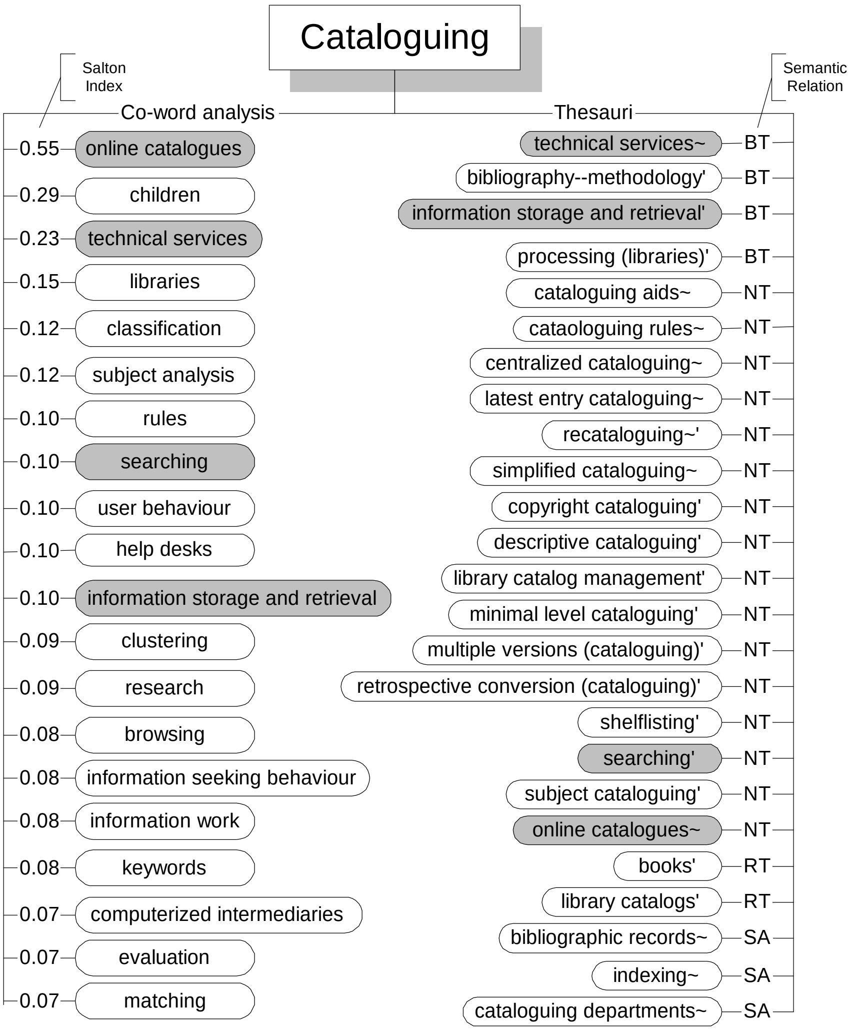

Figure 4. Comparison of cataloguing's 20s and TT in 1987-1997

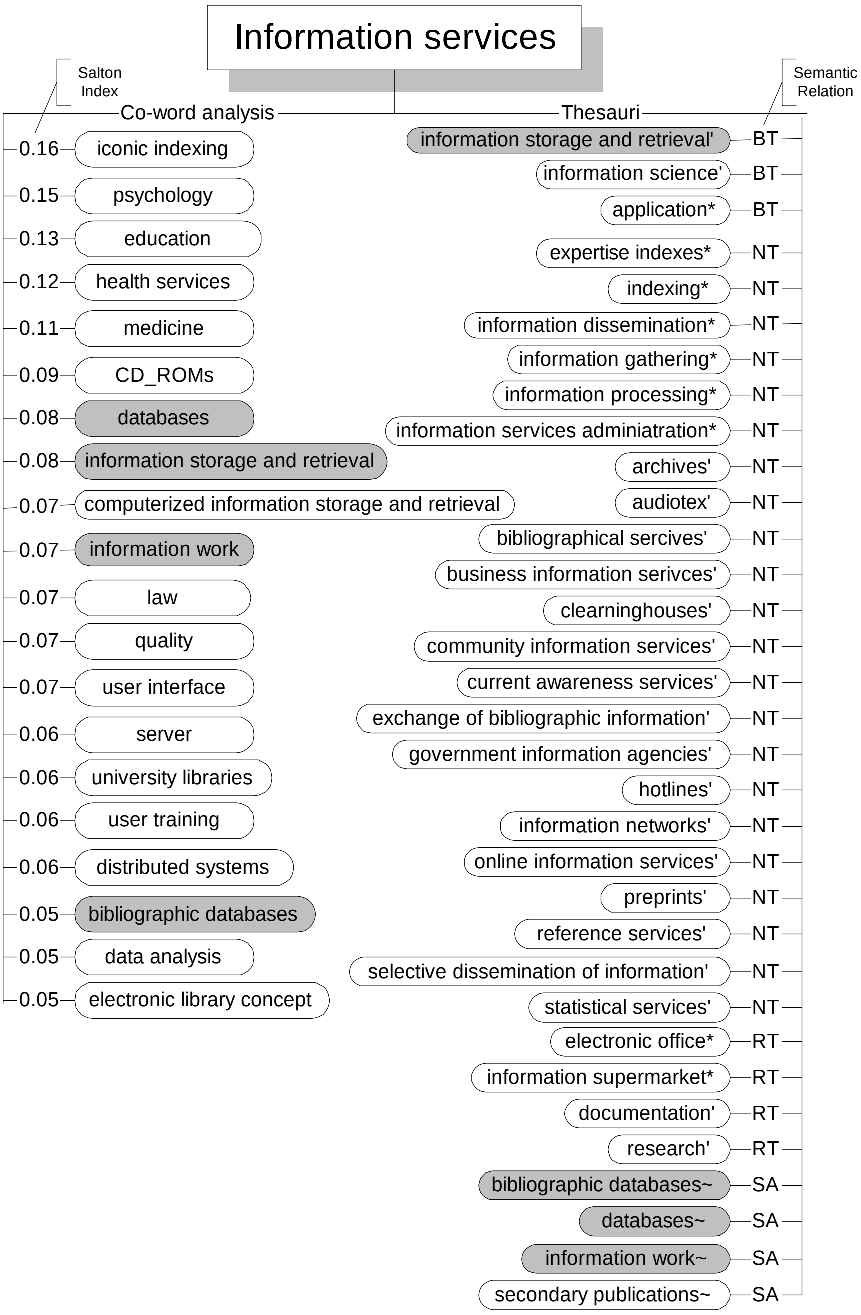
Information services
Salton Index
Co-word analysis
Thesauri
Semantic Relation
0.16 iconic indexing
0.15 psychology
0.13 education
0.12 health services
0.11 medicine
0.09 CD_ROMs
0.08 databases
0.08 information storage and retrieval
0.07 computerized information storage and retrieval
0.07 information work
0.07 law
0.07 quality
0.07 user interface
0.06 server
0.06 university libraries
0.06 user training
0.06 distributed systems
0.05 bibliographic databases
0.05 data analysis
0.05 electronic library concept
information storage and retrieval' BT
information science' BT
application* BT
expertise indexes* NT
indexing* NT
information dissemination* NT
information gathering* NT
information processing* NT
information services adminiatration* NT
archives' NT
audiotex' NT
bibliographical sercives' NT
business information serivces' NT
clearninghouses' NT
community information services' NT
current awareness services' NT
exchange of bibliographic information' NT
government information agencies' NT
hotlines' NT
information networks' NT
online information services' NT
preprints' NT
reference services' NT
selective dissemination of information' NT
statistical services' NT
electronic office* RT
information supermarket* RT
documentation' RT
research' RT
bibliographic databases~ SA
databases~ SA
information work~ SA
secondary publications~ SA

Figure 5. Comparison of information services' 20s and TT in 1987-1997

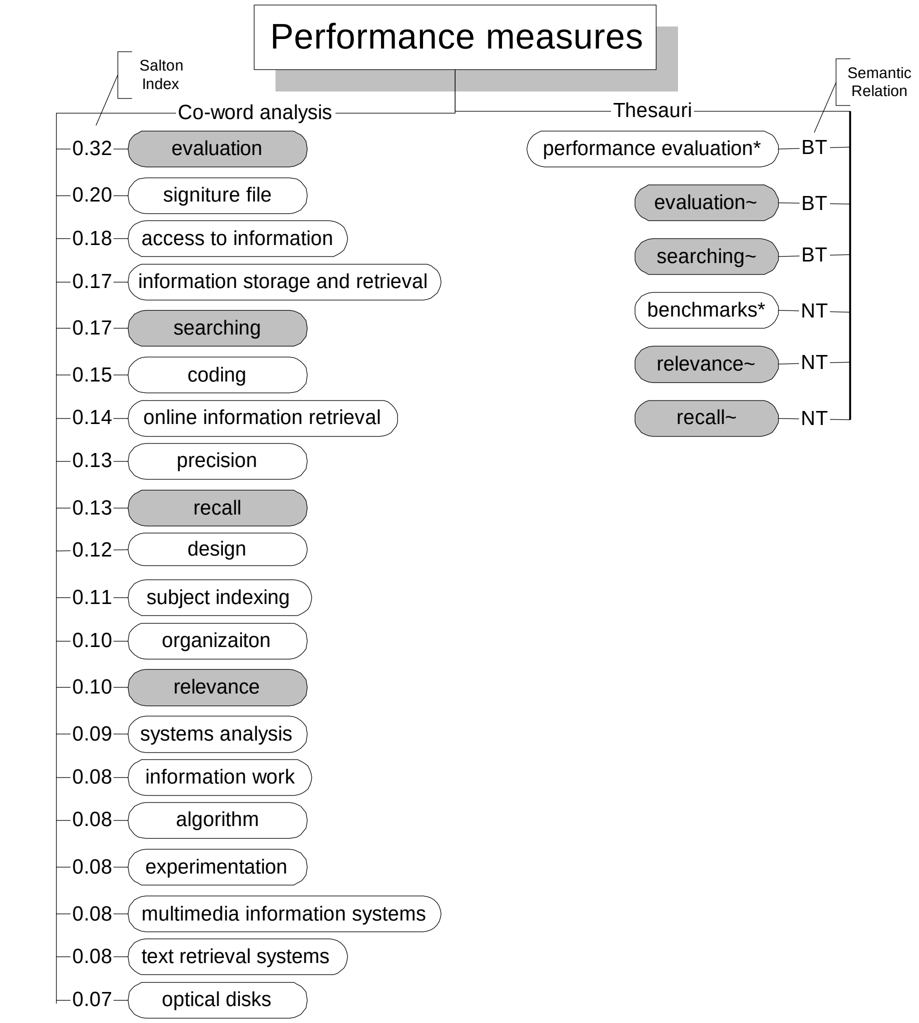

Figure 6. Comparison of performance measures' 20s and TT in 1987-1997

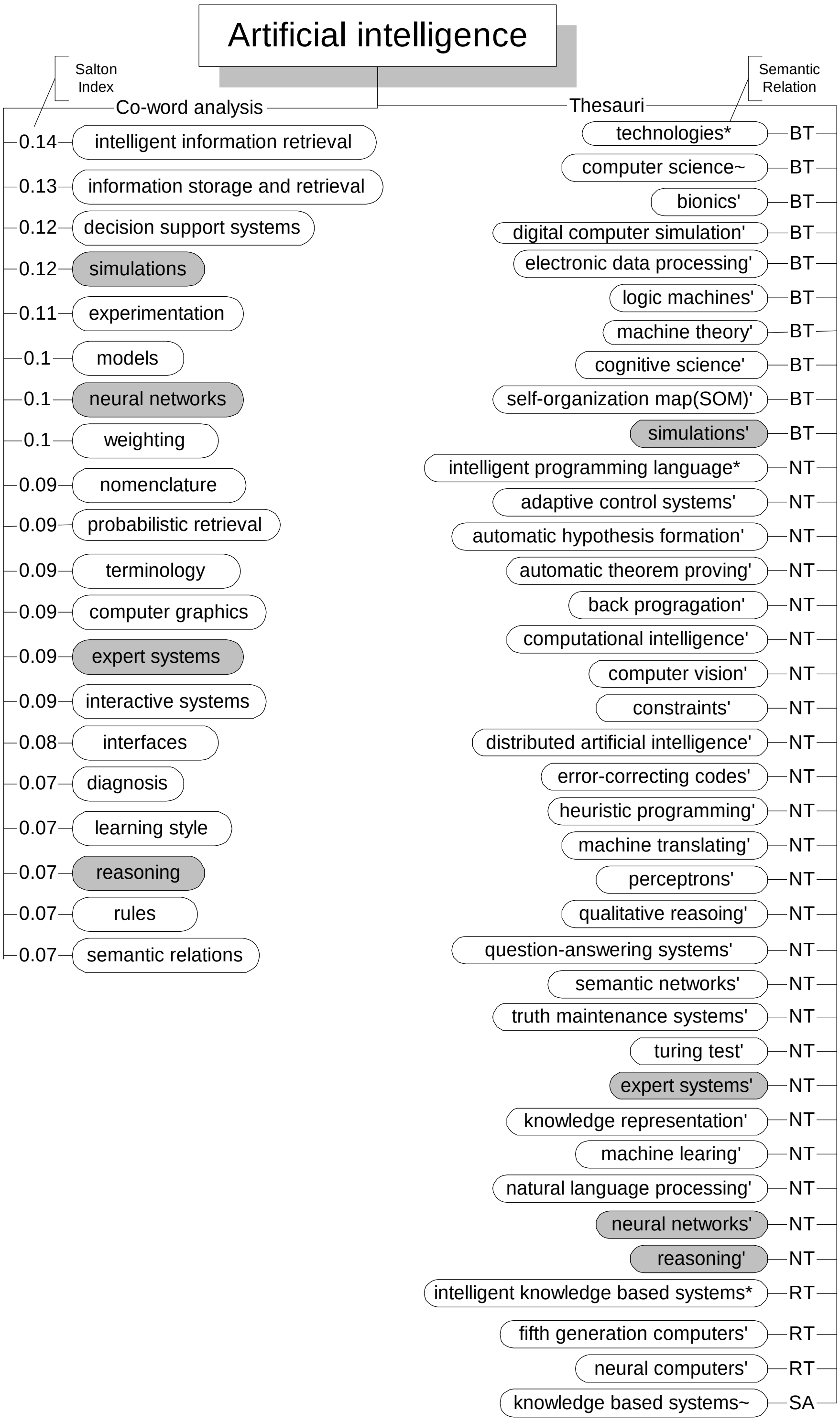

Figure 7. Comparison of artificial intelligent' 20s and TT in 1987-1997

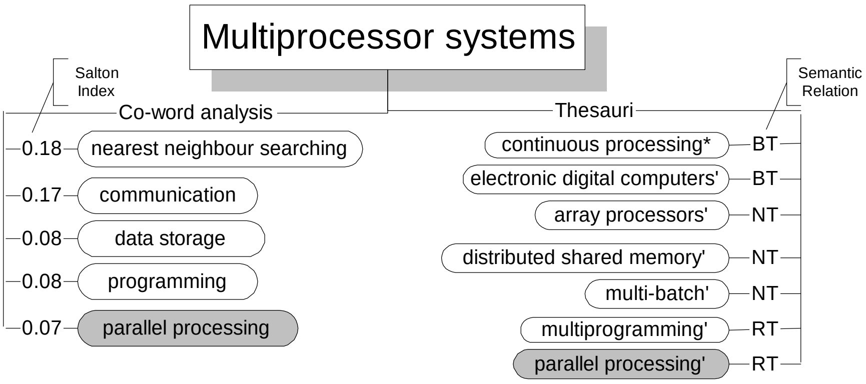

Figure 8. Comparison of multiprocessor systems' 20s and TT in 1987-1997

From Figures 2, 3, 4, 5, 7 and 8, we can see that in most cases the combination of thesauri provide more information than co-word analysis (though this might not be the case if the Keywords 20s were compared with the corresponding term blocks (TTs) in only one thesaurus). Co-word analysis provides more variety for the end-users by identifying additional and different terms that are not found in the traditional thesauri. Thus, co-word analysis can play an important role to assist traditional thesauri to provide more search varieties to the end users. However, it is acknowledged that co-word analysis cannot supply semantic relations between words. Nevertheless, these two systems can be used to supplement one another.

Figure 1 shows a descending trend implying that few keywords (20s and TT) share common results. This illustrates the difference between the results of co-word analysis and traditional thesauri. The similarity of these two methods is very low because only 3.2% keywords have more than or equal to 20% common words in their 20s and TT. Through the comparison the top seven keywords sharing up to 20% common words in each 20s and TT, the difference of co-word analysis and traditional thesauri was confirmed again.

Results of co-word analysis for 1987-1991 compared with traditional thesauri

During this period, out of the 216 keywords sample, 40 keywords did not co-occur with any other keyword in the research sample during 1987-1991. Thus, a total of 176 keywords were used for comparison.

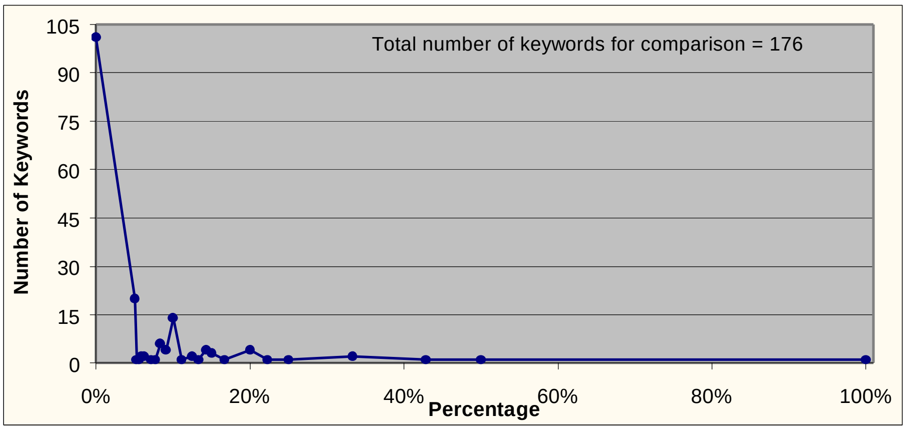

Figure 9. Comparison of co-word thesaurus and traditional thesaurus in 1987-1991

Out of these 176 keywords, only 75 (42.6%) keywords were found to have common words in their corresponding 20s and TT and the average similarity is 12.4% (Figure 9). In those 75 keywords, 20 had 5% common words in their corresponding 20s and TTs. One keyword, *Dialog*, had 100% common words in its 20s and TTs (*Dialog* only co-occurred with other two keywords during this period, so there are only two words in its 20s).

Figures 10 to 20 show the keywords with more than or equal to 20% common words in their 20s and TTs. These keywords include *dialog*, *controlled vocabulary*, *text compression*, *multiprocessor systems*, *multimedia information systems*, *information storage and retrieval*, *human-computer interaction*, *data compression*, *cataloguing*, *information services*, and *expert systems.*

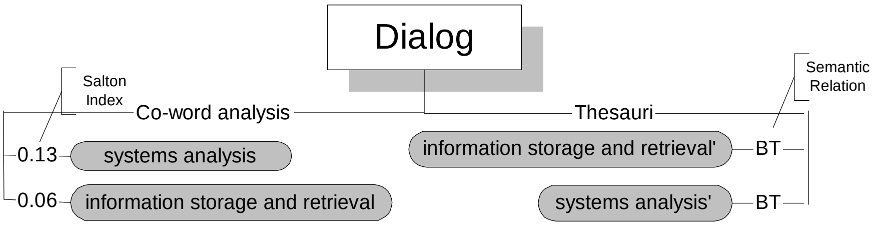

Figure 10. Comparison of dialog's 20s and TT in 1987-1991

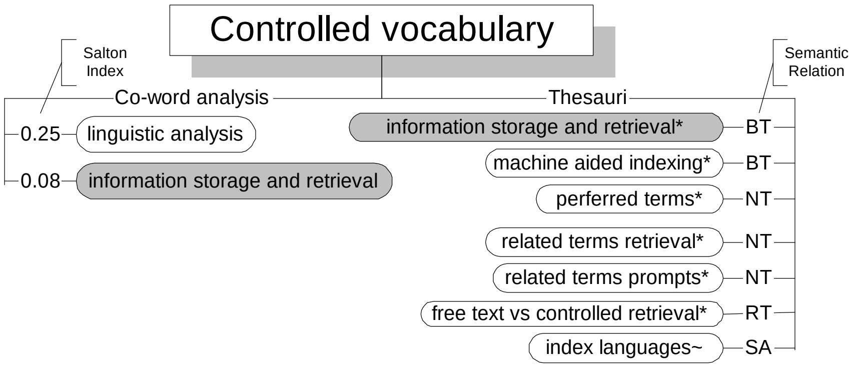
Controlled vocabulary
Salton Index
Semantic Relation
Co-word analysis
Thesauri
0.25 linguistic analysis
0.08 information storage and retrieval
information storage and retrieval* BT
machine aided indexing* BT
perferred terms* NT
related terms retrieval* NT
related terms prompts* NT
free text vs controlled retrieval* RT
index languages~ SA

Figure 11. Comparison of controlled vocabulary' 20s and TT in 1987-1991

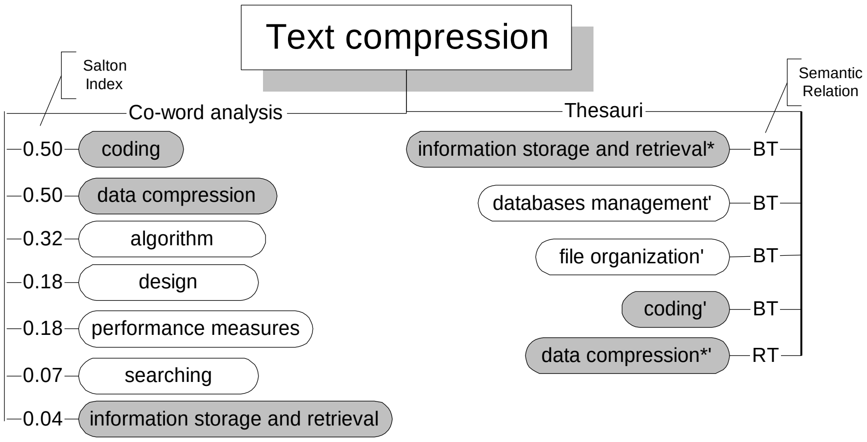

Figure 12. Comparison of text compression' 20s and TT in 1987-1991

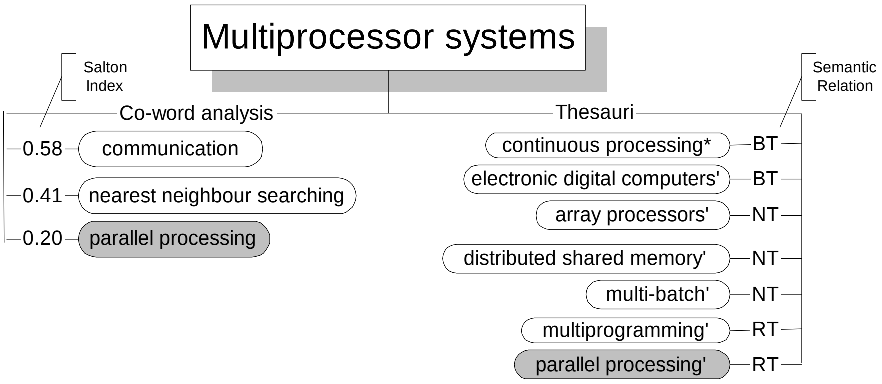

Figure 13. Comparison of multiprocessor systems' 20s and TT in 1987-1991

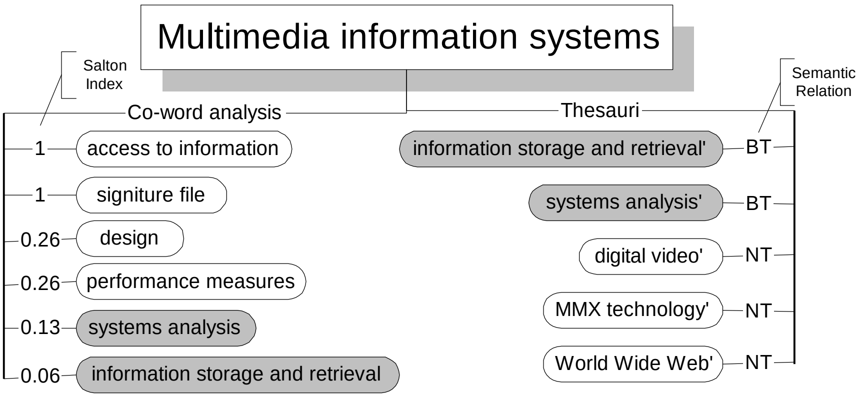

Figure 14. Comparison of multimedia information systems' 20s and TT in 1987-1991

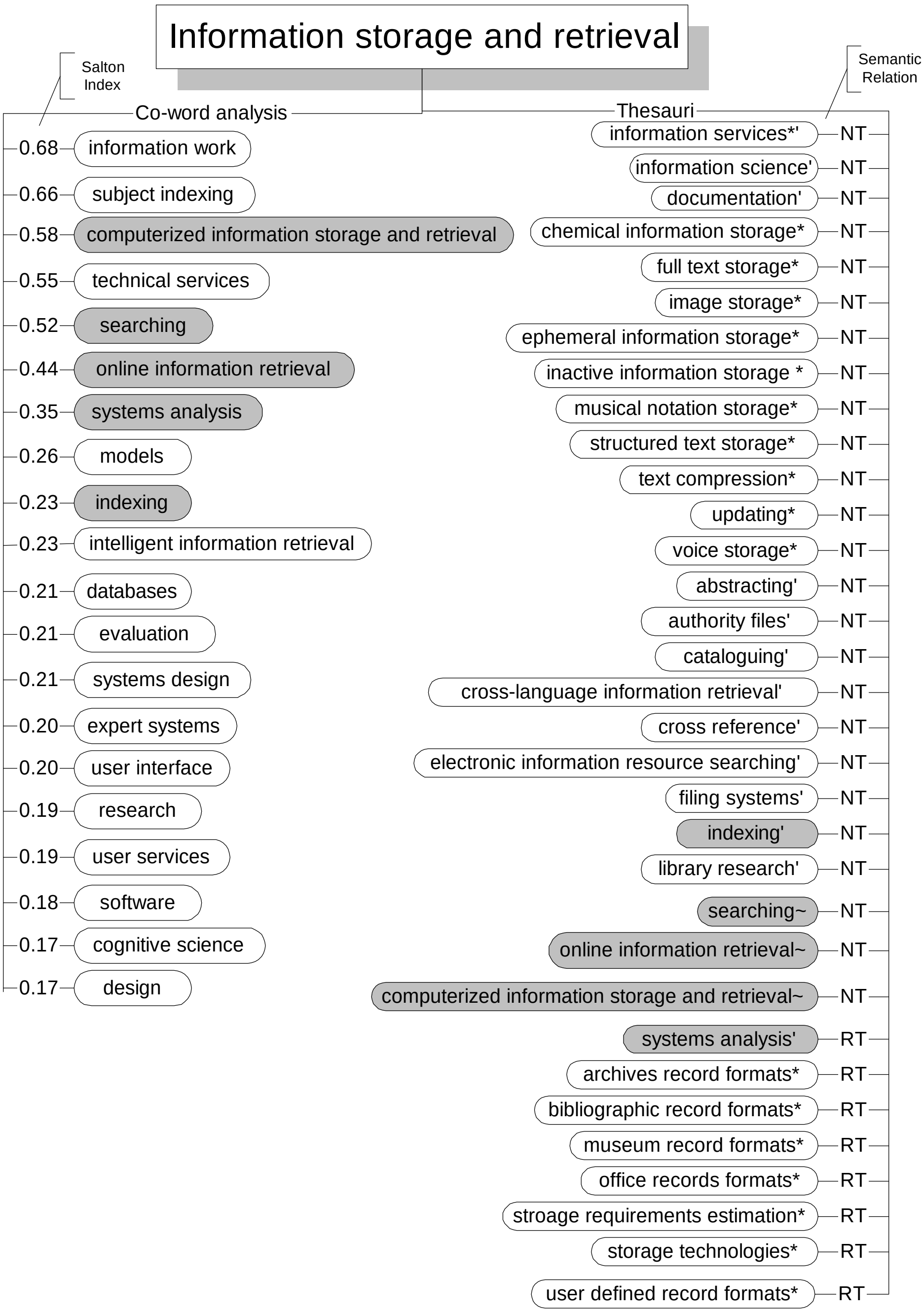
Information storage and retrieval
Salton Index
Semantic Relation
Co-word analysis
Thesauri
0.68 information work
0.66 subject indexing
0.58 computerized information storage and retrieval
0.55 technical services
0.52 searching
0.44 online information retrieval
0.35 systems analysis
0.26 models
0.23 indexing
0.23 intelligent information retrieval
0.21 databases
0.21 evaluation
0.21 systems design
0.20 expert systems
0.20 user interface
0.19 research
0.19 user services
0.18 software
0.17 cognitive science
0.17 design
information services*' NT
information science' NT
documentation' NT
chemical information storage* NT
full text storage* NT
image storage* NT
ephemeral information storage* NT
inactive information storage * NT
musical notation storage* NT
structured text storage* NT
text compression* NT
updating* NT
voice storage* NT
abstracting' NT
authority files' NT
cataloguing' NT
cross-language information retrieval' NT
cross reference' NT
electronic information resource searching' NT
filing systems' NT
indexing' NT
library research' NT
searching~ NT
online information retrieval~ NT
computerized information storage and retrieval~ NT
systems analysis' RT
archives record formats* RT
bibliographic record formats* RT
museum record formats* RT
office records formats* RT
stroage requirements estimation* RT
storage technologies* RT
user defined record formats* RT

Figure 15. Comparison of information storage and retrieval' 20s and TT in 1987-1991

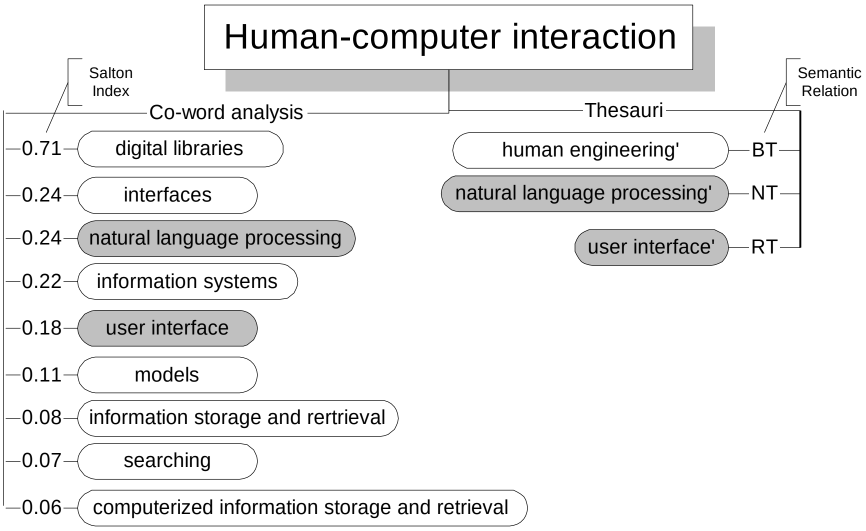

Figure 16. Comparison of human-computer interaction's 20s and TT in 1987-1991

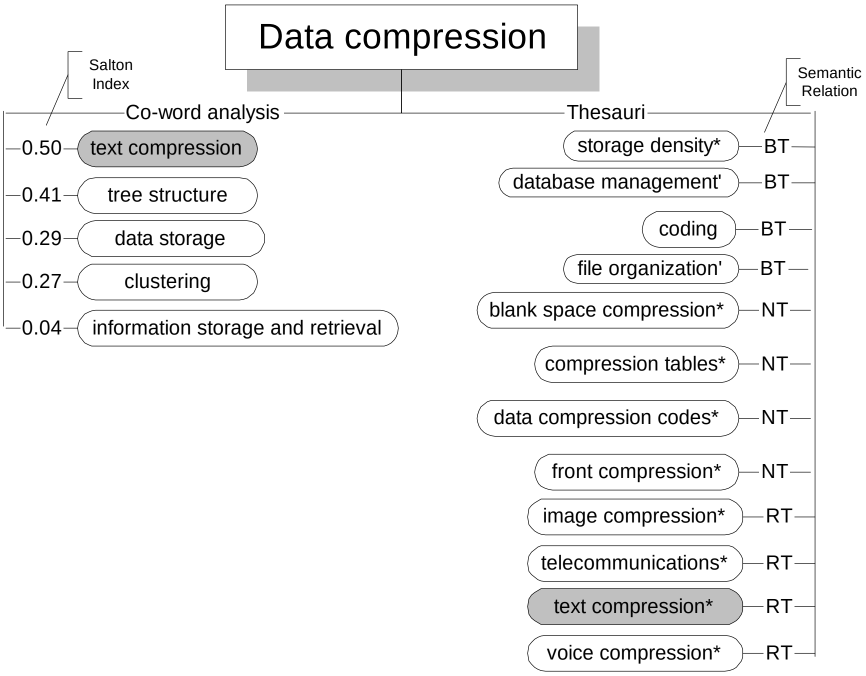

Figure 17. Comparison of data compression's 20s and TT in 1987-1991

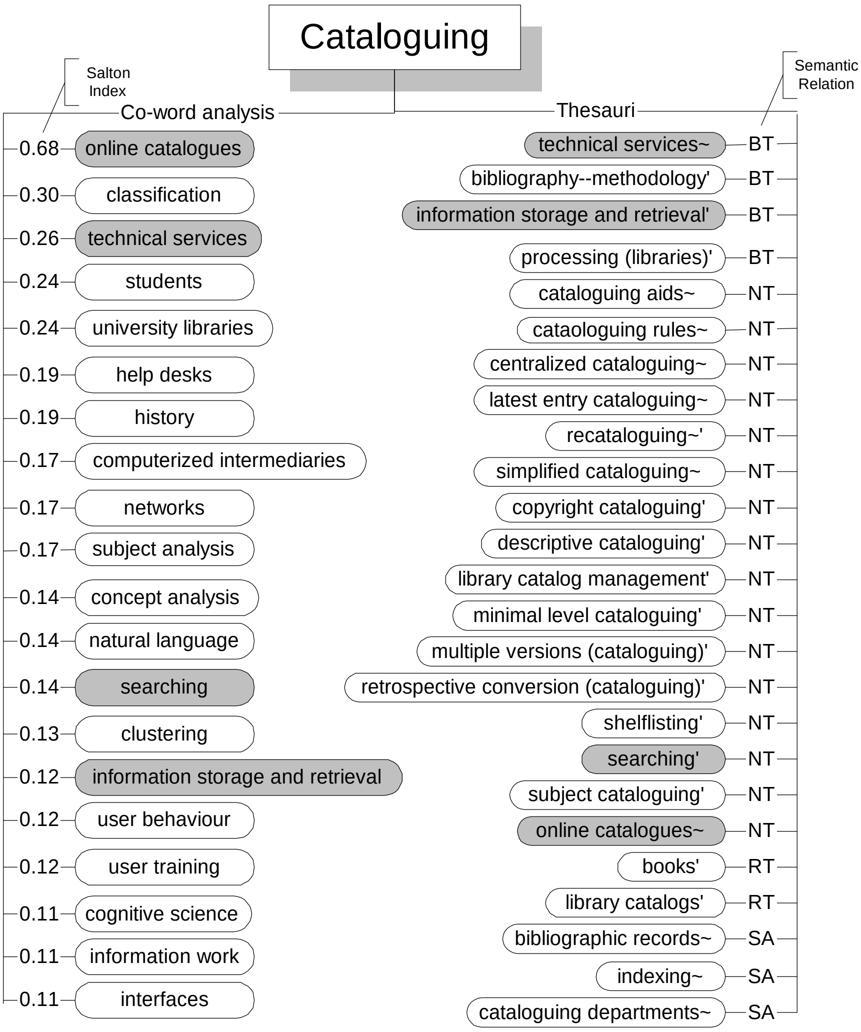

Figure 18. Comparison of cataloguing's 20s and TT in 1987-1991

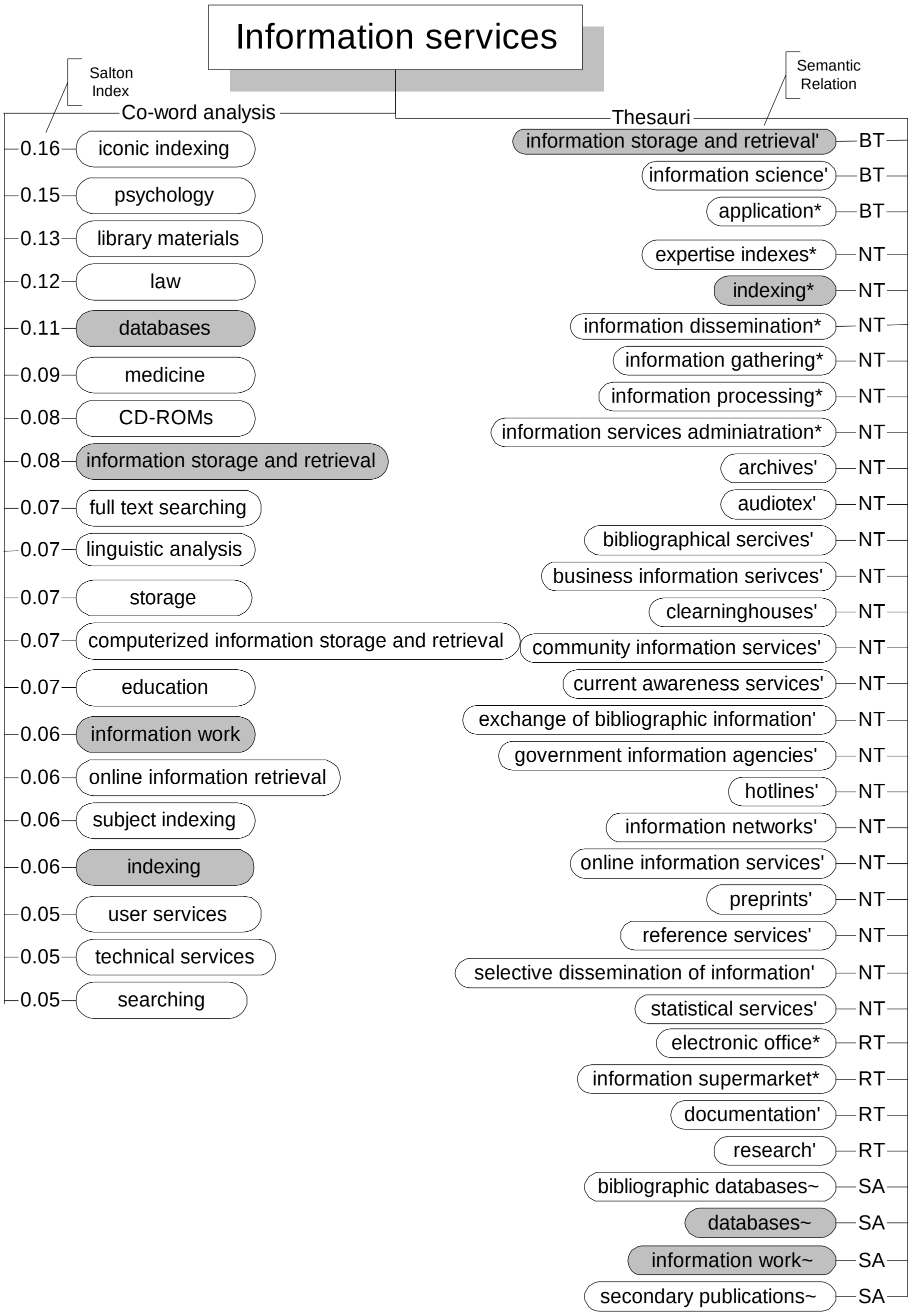

Figure 19. Comparison of information services' 20s and TT in 1987-1991

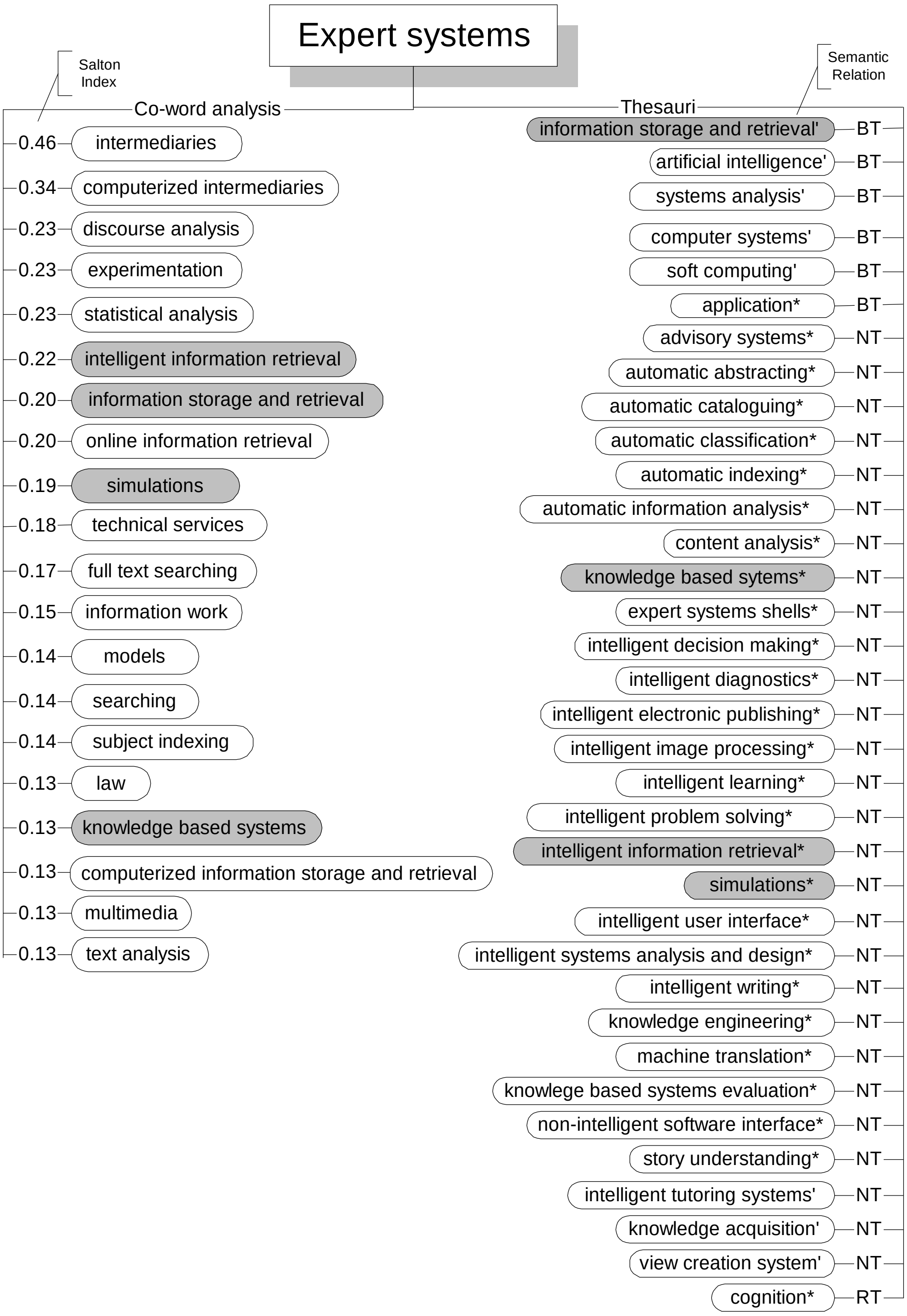
Expert systems
Salton Index
Co-word analysis
Semantic Relation
Thesauri
0.46 intermediaries
0.34 computerized intermediaries
0.23 discourse analysis
0.23 experimentation
0.23 statistical analysis
0.22 intelligent information retrieval
0.20 information storage and retrieval
0.20 online information retrieval
0.19 simulations
0.18 technical services
0.17 full text searching
0.15 information work
0.14 models
0.14 searching
0.14 subject indexing
0.13 law
0.13 knowledge based systems
0.13 computerized information storage and retrieval
0.13 multimedia
0.13 text analysis
information storage and retrieval' BT
artificial intelligence' BT
systems analysis' BT
computer systems' BT
soft computing' BT
application* BT
advisory systems* NT
automatic abstracting* NT
automatic cataloguing* NT
automatic classification* NT
automatic indexing* NT
automatic information analysis* NT
content analysis* NT
knowledge based sytems* NT
expert systems shells* NT
intelligent decision making* NT
intelligent diagnostics* NT
intelligent electronic publishing* NT
intelligent image processing* NT
intelligent learning* NT
intelligent problem solving* NT
intelligent information retrieval* NT
simulations* NT
intelligent user interface* NT
intelligent systems analysis and design* NT
intelligent writing* NT
knowledge engineering* NT
machine translation* NT
knowlege based systems evaluation* NT
non-intelligent software interface* NT
story understanding* NT
intelligent tutoring systems' NT
knowledge acquisition' NT
view creation system' NT
cognition* RT

Figure 20. Comparison of expert systems' 20s and TT in 1987-1991

Comparing with the period of 1987-1997, the 20s lists of *multiprocessor systems*, *information storage and retrieval*, *cataloguing*, *expert systems* and *information services* are noted to be different. Some of them are changing according to the time. This gives an evidence of the 'time' characteristic of co-word analysis. It indicates that co-word analysis can reflect the changing relations among keywords and can thus be used to supplement traditional thesauri.

Results of co-word analysis for 1992-1997 compared with traditional thesauri

During this period, out of the 216 keywords, only 92 (42.6%) have common words in their 20s and TT and the average similarity is 7.9%. In those 92 keywords, 52 have 5% common words in their 20s and TTs. One keyword, *text compression*, has 33% common words in its 20s and TTs. The same declining tendency was confirmed during this period (see Figure 21). Figures 22 to 24 show the keywords with more than or equal to 20% common words in its 20s and TT, namely, *text compression*, *cataloguing*, and *performance measures*. After comparing these three keywords with their corresponding figures in the other two periods (1987-1997 and 1987-1991), we noted that the common keywords they share in their 20s and TTs remain the same, but the lists of co-word analysis for these three keywords have slightly changed. These are also good examples to manifest the 'time' characteristic of co-word analysis. For example, in the term *text compression, algorithm* remains in both time periods, but terms such as *electronic library concept*, *text retrieval systems*, *text analysis*, *hypertext*, and *indexing* have now emerged in the latter 1992-1997 period

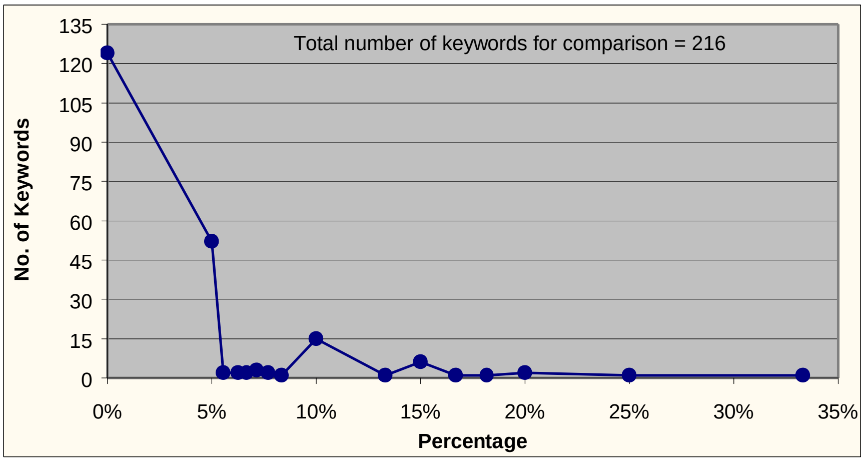

Figure 21. Comparison of co-word thesaurus and traditional thesaurus in 1992-1997

.

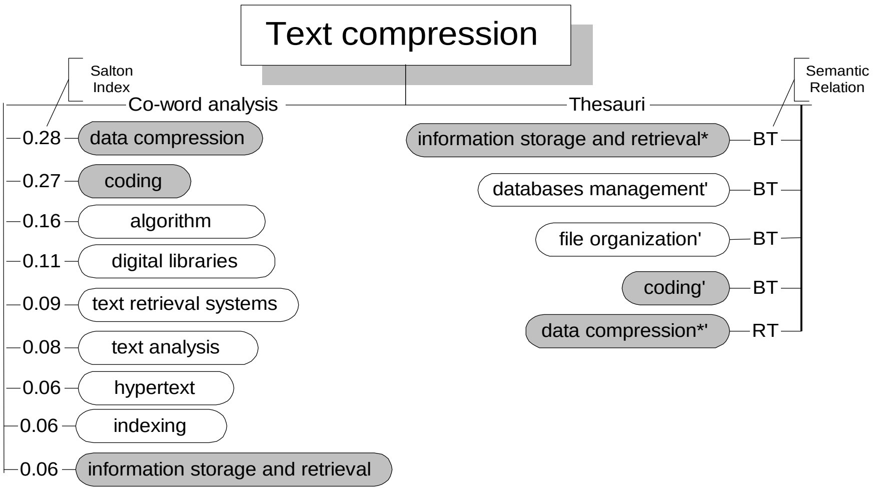

Figure 22. Comparison of text compression's 20s and TT in 1992-1997

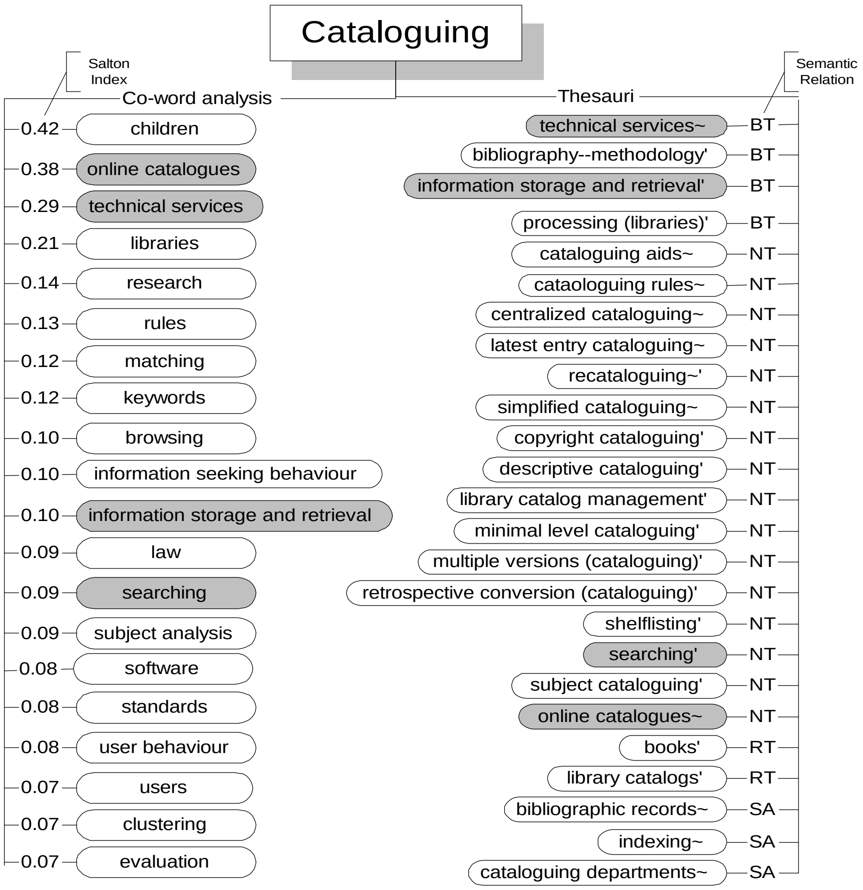

Figure 23. Comparison of cataloguing's 20s and TT in 1992-1997

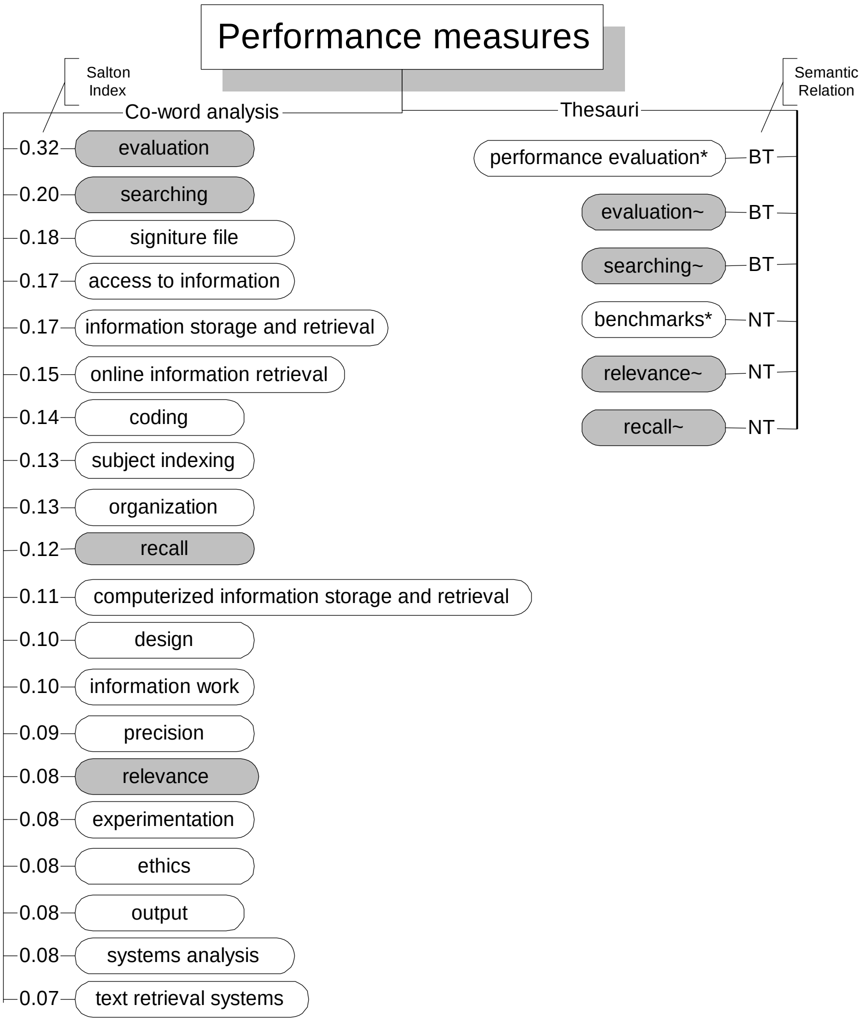

Figure 24. Comparison of performance measures' 20s and TT in 1992-1997

Co-word analysis in 1987-1991 vs. in 1992-1997

Out of the 240 keywords, only 192 keywords appeared together with other keywords in the sample during the period of 1987-1991 and 1992-1997. So only these 192 keywords could be used as the research sample to compare the dynamic changes of co-word thesaurus in 1987-1991 and 1992-1997. Among these 192 keywords, 168 keywords at least have one common word in their 1987-1991’s 20s and 1992-1997’s 20s. There are 24 keywords which do not share any common word in their 1987-1997’s 20s and 1987-1991’s 20s. For each of these 168 keywords, the average similarity percentage is 26.1%. Most of keywords in the sample have 10% to 30% similarities. The similarity of these periods is not high (Figure 25). The dynamic changes of co-word thesaurus in 1987-1991 and 1992-1997 were captured through this comparison.

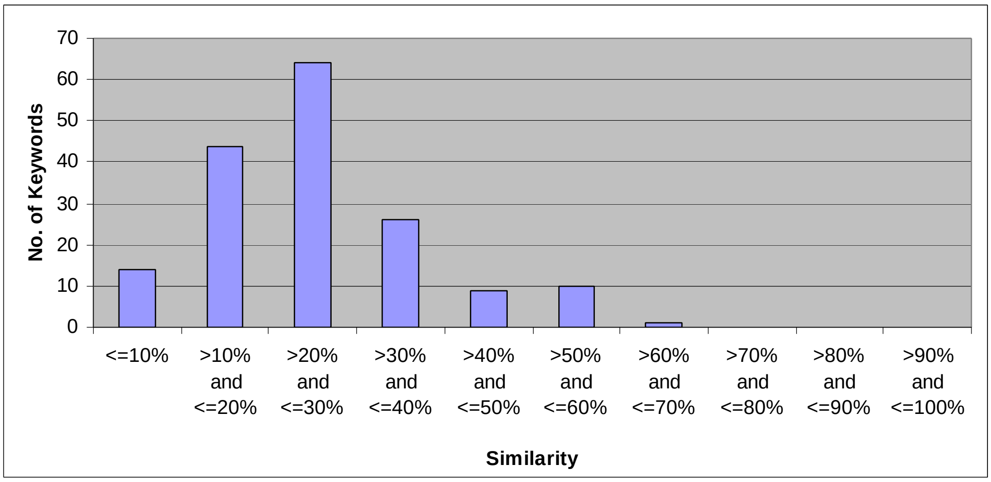

Figure 25 Changes of co-word analysis during 1987-1991 and 1992-1997.

Discussion and Conclusion

Table 1 summarizes the results of comparison among between the keywords sets obtained through the co-word analysis and the three combined thesauri. For each period, around 50% of 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. So 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, et al., 1997).

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%

The list of terms generated by traditional thesauri reflects the organization of term based on human intelligence. The list of terms in the co-word analysis is produced according to the frequency of their co-occurrence. Thus, users are given more keywords, that have already co-occurred in the literature to expand and refine their queries. The results of comparisons among the keywords sets generated during the two different time periods (1987—1991 Vs. 1992—1997) indicate that there are some changes in the keywords sets. This implies that co-word analysis has the ability to capture the dynamic changes in the domain area and to provide additional information to end users seeking information in this domain area. However, this needs to be explored further.

In conclusion, we believe this research has provided insights concerning the application of co-word analysis in the information retrieval area. In other words, this research has shown that side by side with the traditional thesauri, we should also consider the application of co-word analysis to create an automatic thesaurus that can be subsequently integrated to improve concept-based information retrieval by providing search varieties for end users.

Acknowledgement:

We wish to thank Dr. Alan Gilchrist for his valuable suggestions and comments on the LIS/IS thesauri during the early stage of this research. We also acknowledge the valuable comments obtained from the referees that helped us improve the paper significantly.

References:

Aitchison, J. & Gilchrist, A. (1997). Thesaurus construction : a practical manual (3rd). London: Aslib.

Bates, M. J. (1986). Subject access in online catalogs: A design model. *Journal of the American Society for Information Science*, 37, 357-376.

Bates, M. J. (1998). Indexing and access for digital libraries and the Internet: Human, database, and domain factors. *Journal of the American Society for Information Science*, 49(13), 1185-1205.

Borgman, C. L. (Ed.). (1990). *Scholarly communication and bibliometrics*. Newbury Park, CA: Sage.

Braam, R. T. , Moed, H. F. & van Raan, A. F. J. (1991). Mapping of science by combined cocitation and word analysis. II: Dynamical aspects. *Journal of the American Society for Information Science*, 42(4), 252-266.

Byrne, C. C. & McCracken, S.A. (1999). An adaptive thesaurus employing semantic distance, relational inheritance and nominal compound interpretation for linguistic support of information retrieval. *Journal of Information Science*, 25(2), 113-131.

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.

Chen, H. & Dhar, V. (1991). Cognitive process as a basis for intelligent retrieval systems design. *Information Processing and Management*, 27(5), 405-432.

Chen, H. & Ng, T. (1995). An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic branch-and-bound search vs. connectionist hopfield net activation. *Journal of the American Society for Information Science*, 46(5), 348-369.

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., Martinez, J., Kirchhoff, A., Ng, T. D. & Schatz, B. R. (1998). Alleviating search uncertainty through concept associations: automatic indexing, co-occurrence analysis, and parallel computing. *Journal of the American Society for Information Science*, 49(3), 206-216.

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.

Chu, H. (1992). Communication between Chinese and non-Chinese scientists in discovery of high-Tc superconductor: I. The formal perspective. *Scientometrics*, 25(2), 229-252.

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 keyword simulation of knowledge growth. *Scientometrics*, 31(2), 173-192.

Crouch, C. J. (1990). An approach to the automatic construction of global thesauri. *Information Processing and Management*, 26, 629-640.

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

Everitt, B. (1980). *Cluster analysis* (2nd ed.). London: Heinemann.

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

Gregory, J. G. (1983). Citation study of peripheral theories in an expanding research front. *Journal of Information Science*, 7, 73-80.

Hamers, L. et. al. (1989). Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. Information Processing & Management, 25(3), 315-318.

Harter, S.P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Sience*, 43, 602-615.

Harter, S.P. & Cheng, Y. R. (1996). Colinked descriptors: Imporving vocabulary selection for end-user searching. *Journal of the American Society for Information Sience*, 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.

Lau, T. Y. (1995). Chinese communication studies: A citation analysis of Chinese communication research in English-language journals. *Scientometrics*, 33(1), 65-91.

Law, J. & Whittaker, J. (1991). Mapping acidification research: A test of the co-word method. *Scientometrics*, 23(3), 417-461.

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. (1998). Personal communication.

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.

Noyons, E.C.M.; Moed, H.F. & Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the American Society for Information Science*, 50(2), 115-131.

Peters, H. P. F., Braam, R. R. & van Raan, A. F. J. (1995). Cognitive resemblance and citation relations in chemical engineering publications. *Journal of the American Society for Information Science*, 46(1), 9-21.

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.

Quoniam, L.; Balme, F.; Rostaing, H.; Giraud, E. & Dou, J.M. (1998). Bibliometric law used for information retrieval. *Scientometrics*, 41(1-2), 83-91

Salton, G. (1989). *Automatic text processing*. Reading, MA: Addison Wesley.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18, 613-620.

Saracevic, T. & Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches, and overlap. *Journal of the American Society for Information Science*, 39, 197-216.

Seglen, P. O. (1996). Quantification of scientific article contents. *Scientometrics*, 35(3), 335-366.

Shalini, R. (1993). 'Citation profiles' to improve relevance in a two-stage retrieval system: a proposal. *Information Processing & Management*, 29(4), 463-470.

Van Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths.