Available online at www.sciencedirect.com

SCIENCE DIRECT®

Information Processing and Management 42 (2006) 1614–1642

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Towards mapping library and information science

Frizo Janssens [a,*], Jacqueline Leta [b,c], Wolfgang Glänzel [b,d], Bart De Moor [a]

[a] Katholieke Universiteit Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[b] Katholieke Universiteit Leuven, Steunpunt O&O Statistieken, Dekenstraat 2, B-3000 Leuven, Belgium
[c] Instituto de Bioquímica Médica, Centro de Ciências da Saúde, Cidade Universitária, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
[d] Hungarian Academy of Sciences, Institute for Research Policy Studies, Nádor u. 18, H-1051 Budapest, Hungary

Received 16 March 2006; accepted 16 March 2006
Available online 18 May 2006

Abstract

In an earlier study by the authors, full-text analysis and traditional bibliometric methods were combined to map research papers published in the journal *Scientometrics*. The main objective was to develop appropriate techniques of full-text analysis and to improve the efficiency of the individual methods in the mapping of science. The number of papers was, however, rather limited. In the present study, we extend the quantitative linguistic part of the previous studies to a set of five journals representing the field of Library and Information Science (LIS). Almost 1000 articles and notes published in the period 2002–2004 have been selected for this exercise. The optimum solution for clustering LIS is found for six clusters. The combination of different mapping techniques, applied to the full text of scientific publications, results in a characteristic tripod pattern. Besides two clusters in bibliometrics, one cluster in information retrieval and one containing general issues, webometrics and patent studies are identified as small but emerging clusters within LIS. The study is concluded with the analysis of cluster representations by the selected journals.
© 2006 Elsevier Ltd. All rights reserved.

Keywords: Full-text analysis; Text-based clustering; Mapping of science; Library and information science

1. Introduction

1.1. A concise overview of the application of quantitative linguistics in informetrics and bibliometrics

Quantitative linguistics dates back to at least the middle of the 19th century (see Grzybek & Kelih, 2004). However, the classical theoretical work by Zipf (1949) is considered pioneering in quantitative linguistic (or text) analysis. Since the 1970s, a remarkable increase in interest has been observed for this topic of information science. As for its application on scientific literature, Wyllys' (1975) study is among the first ones.

[*] Corresponding author.
E-mail addresses: frizo.janssens@esat.kuleuven.be (F. Janssens), Jacqueline.Leta@econ.kuleuven.be (J. Leta), Wolfgang.Glanzel@econ.kuleuven.be (W. Glänzel), bart.demoor@esat.kuleuven.be (B. De Moor).

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.03.025

Along the decades, studies on quantitative linguistics have changed a lot in terms of focus and methodology. At present, the most frequent techniques, co-word, co-heading and co-author clustering, are based on the analysis of co-occurring keywords, terms extracted from titles, abstracts and/or full text, subject headings or cited authors. The method was developed by Callon, Courtial, Turner, and Brain (1983), more than two decades ago, for purposes of evaluating research. The methodological foundation of co-word analysis is the idea that the co-occurrence of words describes the contents of documents. By measuring the relative intensity of these co-occurrences, simplified representations of a field's concept networks can be illustrated (Callon, Courtial, & Laville, 1991).

Van Raan and Tijssen (1993) have discussed the "epistemological" potentials of bibliometric mapping based on co-word analysis. Leydesdorff (1997) analysed 18 full-text articles and sectional differences therein, and considered that the subsumption of similar words under keywords assumes stability in the meanings, but that words can change both in terms of frequencies of relations with other words, and in terms of positional meaning from one text to another. This fluidity was expected to destabilize representations of developments of the sciences on the basis of co-occurrences and co-absences of words. However, Courtial (1998) replied that words, in co-word analysis, are not used as linguistic items to mean something, but as indicators of links between texts.

Many researchers have used this methodology to investigate concept networks in different fields, among others, de Looze and Lemarie (1997) in plant biology, Bhattacharya and Basu (1998) in condensed matter physics, Peters and van Raan (1993) in chemical engineering, Ding, Chowdhury, and Foo (2001) in information retrieval (IR) and Onyancha and Ocholla (2005) in medicine. Co-heading analysis was introduced by Todorov and Winterhager (1990). Polanco, Grivel, and Royauté (1995) used partial parsing of titles and abstracts and hypothesized that two informetric–linguistic indexes, terminological variation but also non-variation (stabilization), can be used as science watch indicators. Co-word analysis resulted in term networks including both non-variant and variant terms, which would remain undetected without using an electronical dictionary of inflected word forms.

The extension of co-word analysis towards the full texts of large sets of publications was possible as early as large textual databases became available in electronic form. The descriptive power of controlled terms or of the vocabulary used by authors to summarise their work in title and abstract, makes it possible to use text mining and co-word analysis as sophisticated tools both in structural (Tijssen & van Raan, 1989) and dynamic bibliometrics (e.g., Zitt, 1991; Zitt & Bassecoulard, 1994). Nonetheless, the added value of full text with respect to title and abstract information can be high; Glenisson, Glänzel, and Persson (2005) and Glenisson, Glänzel, Janssens, and De Moor (2005) have found that the use of full text included more relevant phrases for interpretation.

The idea of studying the full text of scientific literature by means of mathematical statistics, and combining these tools with bibliometrics, was already present in the work of Mullins, Snizek, and Oehler (Mullins, Snizek, & Oehler, 1988; Snizek, Oehler, & Mullins, 1991). The integration of full-text based techniques, above all of text mining into bibliometric standard methodology, has also persistently been advanced by Ronald Kostoff (Kostoff, Toothman, Eberhart, & Humenik, 2001, or more recently Kostoff, Buchtel, Andrews, & Pfeil, 2005).

Co-word analysis has recently also become the preferred tool for the mapping of science at CWTS (Leiden, the Netherlands), where bibliometric mapping is used within a science policy and research management context (Noyons, 2001). The reason why the emphasis has shifted from co-citation analysis to co-word techniques is twofold. The first reason is a practical one; co-word analysis allows application to non-citation indexes as well. The second relates to methodology; co-citation analysis complicates the combined analysis of field dynamics and trends in the actors' activity (Noyons & van Raan, 1998).

The statistical analysis of natural language obviously has a long history. Manning and Schütze (2000) have provided a comprehensive introduction. For science and technology research, Leopold, May, and Paaß (2004) have given an overview of data and text mining fundamentals. Porter and Newman (2004) coined the term "tech mining" for text mining of collections of patents on a specific topic, in support of technology management.

There is also a wealth of documentation available on the use of clustering techniques in text and data mining, i.e., the unsupervised grouping of objects based on pairwise similarities or distances. In this paper, we opted for agglomerative hierarchical text-based clustering using Ward's method (Jain & Dubes, 1988), but we also report on experiments with the *k*-means algorithm. Useful surveys on clustering can be found in Jain, Murty, and Flynn (1999), Berkhin (2002), and Xu and Wunsch (2005). These works treat other linkage methods and different clustering approaches like, among other, divisive or partitional clustering, nearest neighbour, density-based, grid-based, fuzzy, and model-based clustering. Halkidi, Batistakis, and Vazirgiannis (2001) have given an overview of quality assessment of clustering results and cluster validation measures.

Although bibliometrics has early been applied to study its own field and the field of Library and Information Science (e.g., Stefaniak, 1985), relatively few studies have been devoted to general aspects or concept networks of this field. Bonnevie (2003) has used primary bibliometric indicators to analyse the *Journal of Information Science*, while He and Spink (2002) compared the distribution of foreign authors in *Journal of Documentation* and *Journal of the American Society for Information Science and Technology*. Bibliometric trends of the journal *Scientometrics*, another important journal of the field, have been examined by Schubert and Maczelka (1993), Wouters and Leydesdorff (1994), Schoepflin and Glänzel (2001), Schubert (2002), Dutt, Garg, and Bali (2003). The main journals of the field were also analysed in terms of journal co-citation and keyword analyses (Marshakova, 2003; Marshakova-Shaikevich, 2005). The co-citation network of highly cited authors active in the field of IR was studied by Ding, Chowdhury, and Foo (1999). Finally, Persson (2000, 2001) analysed author co-citation networks on basis of documents published in the journal *Scientometrics*.

Courtial (1994) has studied the dynamics of the field by analysing the co-occurrence of words in titles and abstracts. Courtial described scientometrics as a hybrid field consisting of invisible colleges, conditioned by demands on the part of scientific research and end-users. Although this situation might have somewhat changed during the last decade, this conclusion illustrates how heterogeneous the much broader field of LIS – comprising subdisciplines such as traditional library science, IR, scientometrics, informetrics, patent analyses and most recently the emerging specialty of webometrics – nowadays is.

1.2. Main objectives of the present paper

In recent papers, Glenisson, Glänzel, and Persson (2005), Glenisson, Glänzel, Janssens, and De Moor (2005), Janssens, Glenisson, Glänzel, and De Moor (2005) have applied full-text based structural analysis in combination with “traditional” bibliometric methods to bibliometrics and its subdisciplines. In these studies, the approach was expanded from a limited set of papers, published in a special issue of the journal *Scientometrics*, to the complete 2003 volume of this journal. The full-text analysis consisted of text extraction, preprocessing, multidimensional scaling, and Ward’s hierarchical clustering (Jain & Dubes, 1988). Latent semantic indexing was used to reduce the dimensionality of the vector space and to map terms on a much lower number of statistically derived factors, based on the common context in which the terms or phrases generally appear (Berry, Dumais, & O’Brien, 1995; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). An optimal number of clusters was determined using the stability-based method of Ben-Hur, Elisseeff, and Guyon (2002). Each cluster was profiled using term networks, and the clustering results were compared with those of a clustering based on bibliometric indicators. It was clear that the clusters found through application of text mining provided additional information that could be used to extend, improve and explain structures found on basis of bibliometric methods, and vice-versa. The third study by Janssens et al. (2005) aimed at advancing the methodology for integrating lexical and bibliographic information.

The present paper is actually not aiming at providing a theoretical contribution to computational linguistics, but at applying and extending our methodological approach to a broader, more heterogeneous set of documents. In particular, we will study a broader field within Library and Information Science, which will be represented by a set of selected journals devoted to methodological, theoretical studies and quantitative approaches. The challenge is not the growing number of articles (the first study was based on 19 papers, the second one on 85 articles and the present study will use almost 1000 full-text articles), but the heterogeneity of this hybrid field and the variety of terms and concepts used by scientists in our field.

With scalability and complexity issues in mind, and given the fact that statistical methods can provide surprisingly good results, we did not make use of advanced or “deep” natural language processing (NLP) methods. Although the linguistic structure of sentences is thus neglected in the adopted “bag-of-words” representation, we did however make use of shallow parsing techniques as a means to filter important terms and phrases.

According to the observations by Glänzel and Schoepflin (1994), new topics emerged very early in the field and sub-disciplines began drifting apart. In order to monitor the situation in the field of library and information science about one decade later, we will conduct our research along the following questions.

1. Can the assumed heterogeneity be characterised by means of quantitative linguistics?
2. What are the main topics in current research in information science?
3. Have new, emerging topics already developed their own "terminology"?
4. Can the cognitive structure be visualised and represented using multivariate techniques?
5. How are topics and sub-disciplines represented by important journals of the field?

In order to be able to answer these questions, we will elaborate vocabularies for subdisciplines within LIS, and compare different methods of clustering and mapping in order to reach the optimum presentation of the cognitive structure of our field.

2. Materials

Since this study is aiming at the extension of earlier papers on structural analysis of scientometrics and informetrics, we have selected a set of journals with strong focus on both fields and related specialties. The document set used for our study consists of 938 full-text articles or notes, published between 2002 and 2004 in one of five journals. In particular, Table 1 shows the distribution of the 938 documents over the selected journals.

3. Methods

An overview of the text-based analysis is presented in Fig. 1. The different steps will be discussed in the corresponding subsections below.

3.1. Text representation

A considerable part of the text mining analyses performed in this study is comparable to those used in another paper by Glenisson, Glänzel, Janssens, and De Moor (2005). In short, the textual information is encoded in the vector space model using the TF-IDF weighting scheme, and similarities are calculated as the cosine of the angle between the vector representations of two items (see Salton & McGill, 1986; Baeza-Yates & Ribeiro-Neto, 1999). The term-by-document matrix A is again transformed into a latent semantic index A_k (LSI), an approximation of A, but with rank k much lower than the term or document dimension of A. A latent semantic analysis is advisable, especially when dealing with full-text documents in which a lot of noise is observed. One advantage of LSI is the fact that synonyms or different term combinations describing the same concept are mapped on the same factor, based on the common context in which they generally appear (Berry et al., 1995; Deerwester et al., 1990). There is, however, no straightforward rule for determining the number of factors. When choosing too many, there might still be a lot of noise, while too few factors might result in loss of important information. The decay of the singular values can give a clue about

Table 1
The distribution of the 938 articles or notes over the 5 selected journals

Journal	Number of papers	%
Information Processing & Management (IPM)	143	15.3
Journal of the American Society for Information Science and Technology (JASIST)	309	32.9
Journal of Documentation (JDoc)	85	9.1
Journal of Information Science (JIS)	137	14.6
Scientometrics (SciMetr)	264	28.1
Total	938	100

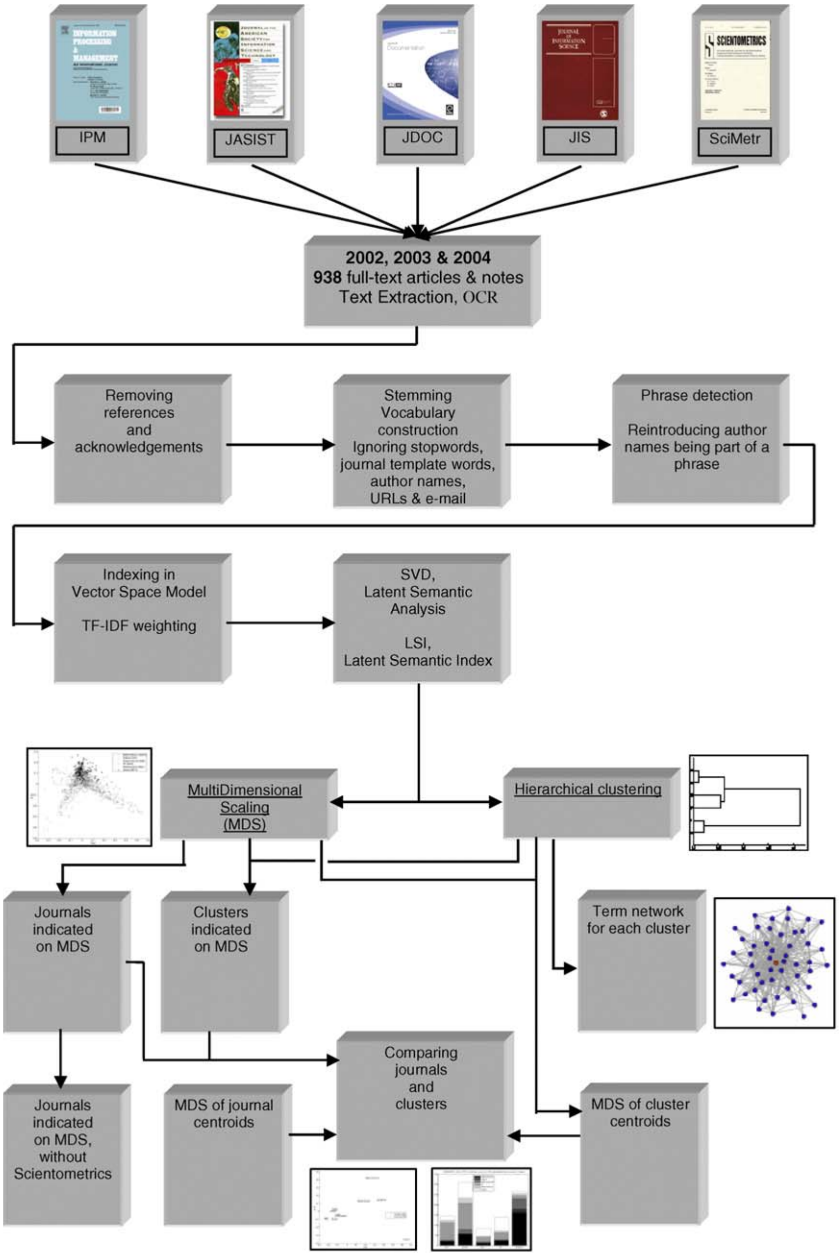
IPM
JASIST
JDOC
JIS
SciMetr
2002, 2003 & 2004
938 full-text articles & notes
Text Extraction, OCR
Removing references and acknowledgements
Stemming
Vocabulary construction
Ignoring stopwords, journal template words, author names, URLs & e-mail
Phrase detection
Reintroducing author names being part of a phrase
Indexing in Vector Space Model
TF-IDF weighting
SVD, Latent Semantic Analysis
LSI, Latent Semantic Index
MultiDimensional Scaling (MDS)
Hierarchical clustering
Journals indicated on MDS
Clusters indicated on MDS
Term network for each cluster
Comparing journals and clusters
Journals indicated on MDS, without Scientometrics
MDS of journal centroids
MDS of cluster centroids

Fig. 1. Overall framework of the analysis.

a good choice (Fig. 2). Based on this plot, we set k to 150 factors. Except for a comparison between clustering results with or without LSI in Section 4.4, all subsequent analyses will make use of this latent semantic index A_{k150}.

3.2. Text preprocessing

Mining full texts instead of titles and abstracts introduces extra difficulties and noise, but is to be preferred if the full text is available, as observed in previous research of Glenisson, Glänzel, Janssens, and De Moor (2005). The mining process we adopted comprises the following different steps:

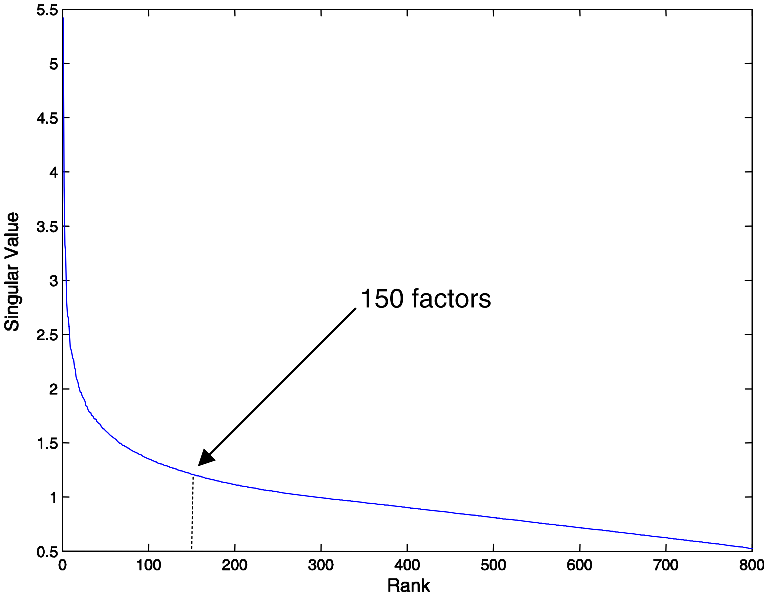

Fig. 2. Plot of the singular values of the term-by-document matrix A.

- *Text extraction*

  The first necessary step to constructing a mathematical representation of the textual information contained in documents, is the extraction of the text. The full-text papers used in our study were mainly saved in the Portable Document Format (PDF) or Microsoft Word format (DOC). We used the StarOffice Software Development Kit[1] to extract the content from the DOC files. The PDF files were extracted by making use of Xpdf PDF text extractor,[2] licensed under the GNU General Public License[3] (GPL). Unfortunately, text extraction from PDF files was not always possible as some publication years, especially the earlier, only contained graphical scanned images of the papers. For these documents, we used the Optical Character Recognition (OCR) techniques from Scansoft's commercial package *Omnipage 14*.
- *Automatically separating acknowledgements and references from article content*

  The aim was to analyse only the "pure" scientific content that is written in the body of a paper and to exclude all bibliographic or other components. Acknowledgements introduce a lot of extra terms relating to institutions, funding agencies, persons, etc. These paragraphs were omitted in order to prevent that the similarity between papers could be influenced by, for example, common acknowledged research funding. Article notes and appendices were considered not problematic and no special effort was done for removing them. In practice, they were removed only when occurring after the reference lists.
- *Neglecting of author names, stop- and template words, URLs and e-mail addresses*

  Each of an article's references of course has an anchor somewhere in the full text. In order not to let cited author names influence the text mining and, above all, the clustering results, they were removed. This was accomplished by an indexing run only considering the reference lists of each article. Combined with pattern matching, we were then able to construct a vocabulary of author and institution names that were ignored in the final indexing run of the research articles and notes. A considerable manual component was necessary here for checking the final list. Sometimes an author's name may indeed correspond to a content-bearing word that should not be withheld from the vocabulary. Besides ignoring author names, also plain stopwords (words with little or no semantic value), month and day names, terms with one or two characters and all terms containing non-alphabetical characters were neglected. Note that a lot of files were extracted using text extraction or OCR techniques that can introduce errors, special characters and sometimes even

[1] http://www.sun.com/software/star/staroffice/sdk/, visited in January 2006.

[2] http://www.foolabs.com/xpdf/, visited in January 2006.

[3] http://www.gnu.org/licenses/gpl.txt, visited in January 2006.

strange, very long strings. Journal-specific information, i.e., terms frequently printed on each page of an article, was also neglected as it could severely influence results. Note that the TF-IDF weighting is not powerful enough to downweight these journal-specific template terms. Finally, pattern matching was performed to match and ignore all e-mail addresses and URLs throughout the indexing process. These addresses introduce a lot of specific tokens that are not necessarily included in the list of words to ignore constructed from the references. However, those tokens, if included, might also heavily deteriorate our results (e.g., clustering based on institution or domain names, etc.). On all remaining terms we applied the popular stemmer by Porter (1980). Stemming involves the removal of a word affix such as plurals, verb tenses and deflections, and the replacement by the canonized equivalent. The Porter Stemmer uses a simple rule-based scheme to process the most common English words. An advantage of stemming is the equation of different forms of the same word, resulting in a reduced dimensionality of the vector space and thus lessening computational costs and the "curse of dimensionality" for a clustering task. A disadvantage is the possible loss of morphological information necessary for discerning between different meanings of two similar words.

– *Phrase and synonym detection, reintroducing author names being part of a phrase*
A lot of time was devoted to the detection of phrases. Since the best phrase candidates can be found in noun phrases, the programs LT POS and LT CHUNK[4] have first been applied to detect all noun phrases in the complete document collection. LT POS is a part-of-speech tagger that uses a lexicon and a hidden markov model disambiguation strategy. LT CHUNK is a syntactic chunker or partial parser. It uses the part-of-speech information provided by LT POS and employs mildly context-sensitive grammars to detect boundaries of syntactic groups. For the scoring of the phrase candidates, Ted Dunning's log-likelihood method for detection of bigrams was followed to detect composite terms within those noun phrases (see Dunning, 1993; Manning & Schütze, 2000). The likelihood ratio tests the hypothesis that terms occur independently in a corpus. When rejected, the words presumably are correlated. It is a parametric statistical text analysis based on the binomial or multinomial distribution and may lead to more accurate results than other text analyses that, often unjustifiably, assume normality, what limits the ability to analyse rare events. Memory-based language processing techniques, which are more recent and advanced methods for, among other purposes, part-of-speech tagging, are described by Daelemans and van den Bosch (2005). Often an author's name has become eponymic and thus part of a phrase that has much power in describing the content of an article. For example, phrases describing a law (e.g., *Lotka's law*, *Bradford's law*, etc.), disease (e.g., *Alzheimer disease*), model, index (e.g., *Price Index*), indicator or method, etc. Such bi-grams were extracted from the texts and, if not yet contained in it, added to the phrase list. Finally, we manually defined a list of synonyms.

For the indexing of the documents we used the Jakarta Lucene[5] indexing platform, which is a high-performance, open source, full-featured text search engine library written entirely in Java. Most of the subsequent analyses were performed in Matlab[6] and Java.[7]

3.3. Multidimensional scaling

In order to get a view of the field, first of all we applied multidimensional scaling (MDS). MDS represents all high-dimensional points (documents) in a two- or three-dimensional space in a way that the pairwise distances between points approximate the original high-dimensional distances as precisely as possible (see Mardia, Kent, & Bibby, 1979).

3.4. Clustering

The agglomerative hierarchical cluster algorithm using Ward's method (see Jain & Dubes, 1988) was chosen to subdivide the documents into clusters. Like any algorithm, it has its advantages and weaknesses and is

[4] http://www.ltg.ed.ac.uk/software/pos/index.html, visited in January 2006.
[5] http://lucene.apache.org/, visited in January 2006.
[6] http://www.mathworks.com/products/matlab/, visited in January 2006.
[7] http://java.sun.com/, visited in January 2006.

certainly not perfect. One of the disadvantages of agglomerative hierarchical clustering is that wrong choices (merges) that are made by the algorithm in an early stage can never be repaired (Kaufman & Rousseeuw, 1990). What we sometimes observe when using hierarchical clustering is the forming of one very big cluster and a few small very specific clusters. We have also experimented with the k-means cluster algorithm on this dataset, but as expected, the hierarchical algorithm seemed to outperform it. Another point why the clustering results could not be expected to be perfect is that, in the present study, we only make use of text and neglect all other information. In further research, we will asses the performance of different schemes for integrating textual and bibliometric information. Note that journal information is of course never used for clustering; the cluster algorithm was just run on the A_{k150} matrix described in Section 3.1.

To determine a statistically optimal number of clusters, we used the stability-based method as proposed by Ben-Hur et al. (2002) and also used in the paper by Glenisson, Glänzel, Janssens, and De Moor (2005). In this method, the optimal number of clusters k is determined by inspection of a stability diagram as in Fig. 6. For higher k, the distances between the curves decrease and form a band. For the optimal number of clusters the largest k is chosen such that partitions into k clusters are still stable. This comes down to looking for a transition curve to the band of wide distributions. A second opinion was offered by observing the plot of mean Silhouette values for 2 up to 25 clusters (as in Fig. 7). The Silhouette value for a document ranges from -1 to $+1$ and measures how similar it is to documents in its own cluster vs. documents in other clusters (Rousseeuw, 1987). The mean Silhouette value for all documents assesses the overall quality of a clustering solution. We also observed the dendrogram resulting from hierarchically clustering of the documents. On a dendrogram, see Fig. 8 for an example, the horizontal lines connect documents (or clusters of documents) in a hierarchical tree. The length of a line represents the distance between the two objects being connected. A vertical line from top to bottom may illustrate a cut-off point.

For ease of interpretation we also made a table containing summary statistics, and a term network for each cluster. A term network is mainly intended to provide a qualitative rather than quantitative way of cluster evaluation. As already described by Glenisson, Glänzel, Janssens, and De Moor (2005), a term network shows the 50 best TF-IDF terms with an edge between two terms meaning that both co-occur in a document of the corresponding cluster, but within a given distance, set to one in this case (ignoring stopwords). We used Biolayout Java by Enright and Ouzounis (2001) for visualising the term networks. Next to the automatically determined best TF-IDF term for a cluster, we will also propose a name for each cluster.

4. Results

4.1. Indexing

The list of detected phrases contained 261 instances and we wrote 58 synonym rules, most of them for mappings like, e.g., coauthor onto "co(-)author", but also for dealing with acronyms, e.g., wif onto "web impact factor". While an initial indexing phase resulted in a vocabulary of 65019 terms, after preprocessing the final index to be used for subsequent analyses only contained 11151 stemmed terms or phrases. In this index also Zipf's curve was cut off to neglect all terms occurring in only one document or in more than 50% of all documents. The final term-by-document matrix A was thus of size 11151×938, transformed by LSI to A_{k150} (150×938). Appendix 1 contains a table with the most important words for each journal and for the whole dataset (terms with highest mean TF-IDF score).

4.2. Visualising library and information science

Figs. 3 and 4 show the multidimensional scaling (MDS) map of the 938 articles or notes in two and three dimensions. Each of the five journals considered is indicated with a different symbol and colour.[8] Fig. 4 is the

[8] For color version of figures, reader is referred to the web version of this article.

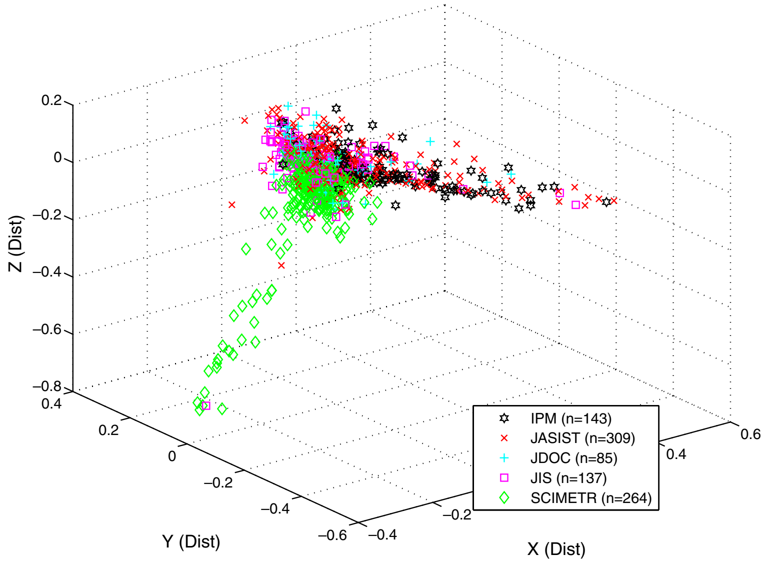

Fig. 3. 3D multidimensional scaling plot of the 938 LIS articles or notes. Each of five journals is indicated with a different symbol and colour.

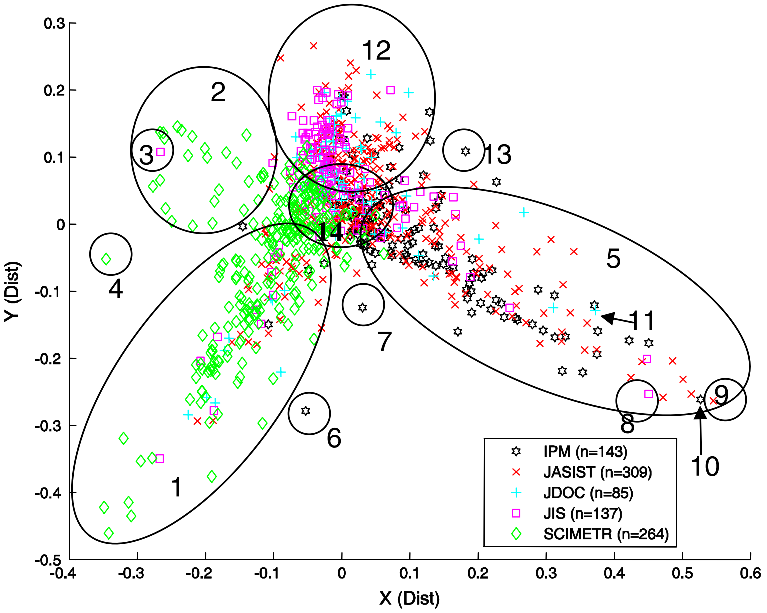

Fig. 4. 2D multidimensional scaling plot of the 938 LIS papers. This figure is the projection of Fig. 3 on the *X*–*Y* plane.

projection of Fig. 3 on the *X*–*Y* plane. The journal *Scientometrics* can be largely separated from the other journals (which is also confirmed by the different term profile in the table of Appendix 1), and exhibits two different foci (best visible in Fig. 4).

In what follows, the different subsets of papers indicated in Fig. 4 will be analysed in more detail. The first "leg", indicated by the ellipse with number 1 and by and large containing the first focus of the journal *Scientometrics*, clearly contains papers in bibliometrics. The 10 best TF-IDF terms for "leg" #1 are: *citat*, *cite*, *impact factor*, *self citat*, *co citat*, *scienc citat index*, *citat rate*, *isi*, *countri* and *bibliometr*. The second "leg of *Scientometrics*", indicated by number 2, is characterised by the best terms *patent*, *industri*, *biotechnolog*, *inventor*, *invent*, *compani*, *firm*, *thin film*, *brazilian* and *citat*. The *JIS* paper (#3) embedded in this patent "leg" might be considered an outlier for that journal, but it was put in the right place since it is concerned with "The many applications of patent analysis" (Appendix 2: Breitzman & Mogee, 2002). One *Scientometrics* paper (#4) seems not to belong to either focus. Indeed, it is about "Patents cited in the scientific literature: An exploratory study of reverse citation relations" (Appendix 2: Glänzel & Meyer, 2003). An important focus of LIS is indicated by ellipse #5 and can be profiled as "Information Retrieval" (IR) when looking at the highest scoring terms: *queri*, *search engin*, *web*, *node*, *music*, *imag*, *xml*, *vector* and *weight*. "Interdisciplinary" *IPM* papers (#6 and #7), between ellipses #1 and #5, are the following: "Mining a Web citation database for author co-citation analysis" (Appendix 2: He & Hui, 2002) and "Real-time author co-citation mapping for online searching" (Appendix 2: Lin et al., 2003). While *Scientometrics* seems to have never published any paper specifically about IR, all other considered journals have (e.g., papers #8, #9, #10 and #11 in Fig. 4). The fourth distinguishable subpart of LIS (#12) is about *digit*, *internet*, *servic*, *seek*, *behaviour*, *health*, *knowledg manag*, *organiz*, *social* and *respond*; so encompassing the more social aspects. The *IPM* paper bridging the gap between IR and more social oriented research (#13) is entitled "The SST method: a tool for analysing Web information search processes" (Appendix 2: Pharo & Jarvelin, 2004). The goal of that paper was "to analyse the searching behaviour, the process, in detail and connect both the searchers' actions (captured in a log) and his/her intentions and goals, which log analysis never captures". The remaining large subpart is somewhat the central part (#14). It consists of papers leading to a mean profile containing the terms *web*, *web site*, *classif*, *domain*, *web page*, *languag*, *scientist*, *region*, *catalog*, and *web impact factor*.

In order to zoom in on the differences between the other journals, Fig. 5 depicts the two-dimensional scaling plot without the journal *Scientometrics*. The above-observed "tripod shape" resolved into a more homogeneous scatter plot when papers published in *Scientometrics* were removed. At the right-hand side, we find information retrieval, while the left-hand side represents rather bibliometric issues such as citation, collaboration and web related issues. Only a few outliers can be found at the bottom of the diagram, one published in *IPM*, one in *JIS* and all others published in *JASIST*. All of them are concerned with music information retrieval. *JASIST* 55 (12) 2004 was a special issue devoted to this topic.

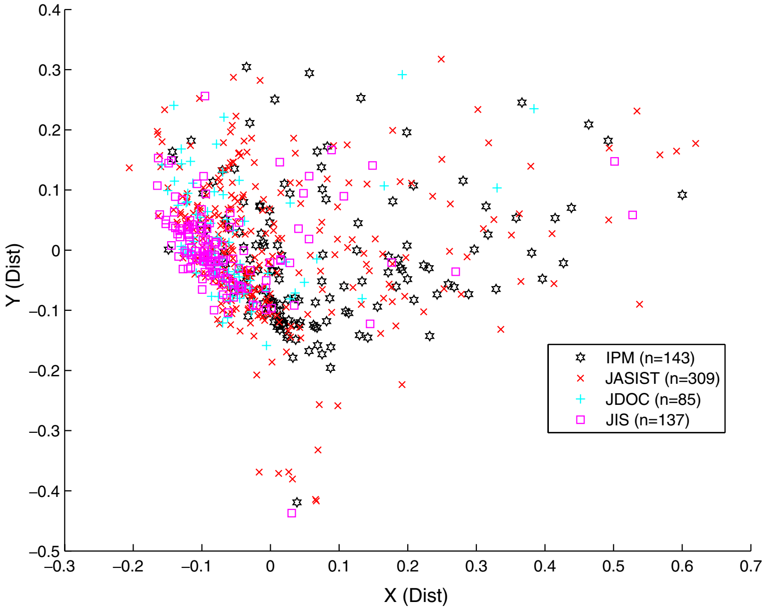

Fig. 5. 2D multidimensional scaling plot (recalculated) of all articles and notes (674) not published in *Scientometrics*.

4.3. Clustering full-text articles to map LIS

We have experimented with the *k*-means cluster algorithm, but as expected, the hierarchical algorithm seemed to outperform it, even when using a more intelligent version of *k*-means in which the means were initialised by a preliminary clustering on a 10% subsample and by selecting the best out of ten runs (*k*-means can get stuck in local minima). At first sight, *k*-means did however seem to do well, the cluster Silhouettes (see further on) even looking a bit nicer than for hierarchical clustering. But upon closer investigation, some undesirable effects showed up due to the nature of the algorithm. For instance, one of the clusters was a combination of papers about patents and papers about music information retrieval (MIR). This was definitely a spurious merge of clusters relatively far from other clusters. The reason why they ended up in one cluster is probably due to the averaging effect of *k*-means. In every step of iteration, each document is assigned to the cluster with closest mean and each mean is updated to become the average document profile in its cluster. The MDS diagram of the documents in that cluster indeed showed two very different orientations. The clustering results and MDS diagrams in this section will corroborate that patent papers are closer to bibliometrics papers and the MIR papers closer to information retrieval, what complies better with our intuition (e.g., Figs. 15, 18 and 19).

Fig. 6 shows, for agglomerative hierarchical clustering using Ward's method, the stability diagram for detection of the optimal number of clusters *k*. The plot shows, for 2 up to 30 clusters, the cumulative distribution of the pairwise similarities between 1000 pairs of clustering solutions for random subsamples of the data set, each comprising 703 documents (sampling ratio of 75%).

For higher *k* the distances between the curves decrease and form a band. Admittedly, the stability diagram for this dataset does not show a clear transition from a score distribution that is concentrated near 1 to a wider distribution. One might conclude that the structure in the data is not well defined or even lacking. But, intuitively, there must be some discernible clusters in Library and Information science. The not so clear-cut stability is due to the high dimensionality of the vector space involved. *Ben-Hur* also stated that partitions obtained on a large set of variables, like thousands of genes from DNA microarrays or thousands of terms in the LIS vocabulary, are usually unstable. On the other hand, the nature of natural language usage might be of influence. When dealing with full-text documents in comparable subject areas, the amount of overlap-

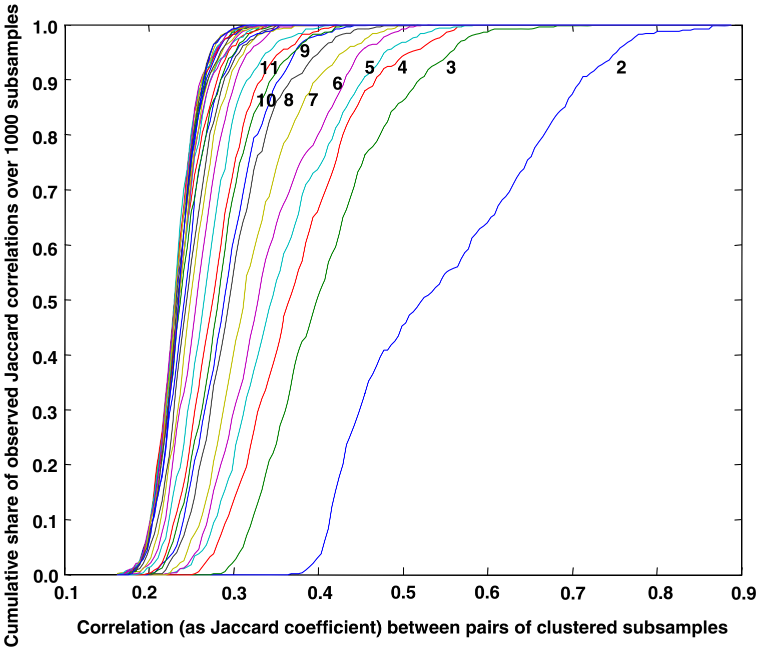

Fig. 6. Stability diagram for determining the number of clusters *k* (according to Ben-Hur et al., 2002).

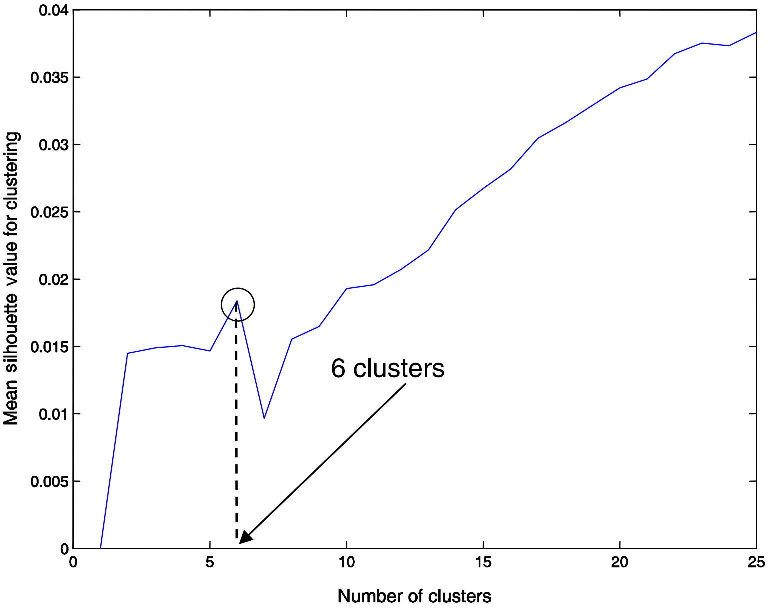

Fig. 7. Mean silhouette coefficient for solutions of 2 up to 25 clusters, with local maximum at 6 clusters.

ping words between different papers is, of course, considerable. Another reason is the "hard" nature of the cluster algorithm involved, forcing each document in one (and only one) cluster. For intertwined subjects this is not always straightforward, resulting in documents shifting between clusters when different subsamples are drawn.

Although the stability plot exhibited no obvious optimal number of clusters, a few observations could be made. Partitioning the documents into two groups, resulted in the most stable solution. But we were looking for a somewhat finer grained clustering solution. Other relatively stable options were 3, 4, 5, 6 and maybe even 7 clusters, but not more.

A second opinion was offered by observing the plot of mean Silhouette values, assessing the overall quality of a clustering solution, for 2 up to 25 clusters (see Fig. 7). It is clear that a local maximum was present at six clusters.

After the sharp drop at 7 clusters, the mean Silhouette value increases again and from 10 clusters onwards it is larger than the value for 6 clusters, but according to the stability diagram those clustering solutions are less stable. So, we chose 6 as the optimal number of clusters. However, the overall mean Silhouette values each seem low, again hinting at no clearly separable groups of documents in the dataset. But a standard *t*-test for the difference in means revealed that the difference between mean Silhouette values for 6 and 7 clusters was statistically significant at the 1% significance level (*p*-value of 2.25×10^{-7}).

Fig. 8 depicts the dendrogram resulting from hierarchically clustering of the documents, cut-off at 25 clusters. The vertical line illustrates the cut-off point for our optimal 6 clusters with best terms *countri*, *patent*, *citat*, *queri*, *web site* and *seek* for c1 to c6, respectively. For each of 25 clusters, the best mean TF-IDF term is shown. Figs. 9–14 show the term network for each of the 6 clusters (c1–c6).

The 6 clusters formed two groups according to their size, particularly three large clusters with more than 200 papers each and three small ones with less than 100 articles each. The large clusters are Cluster 1, manually labelled as "Bibliometrics1", Cluster 4, labelled as "IR", and Cluster 6, labelled as "Social". Cluster names have been chosen on basis of the terms representing these clusters. The term network of Cluster 1 allowed the conclusion that the papers belonging to this cluster are concerned with domain studies, studies of collaboration in science, citation analyses, national research performance and similar issues. Indeed, the analysis of the papers close to the medoid, representing about 20% of all papers in the cluster, confirmed this assumption. The medoid is a paper by Persson et al. on "Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies" (Appendix 2: Persson et al., 2004). This is a

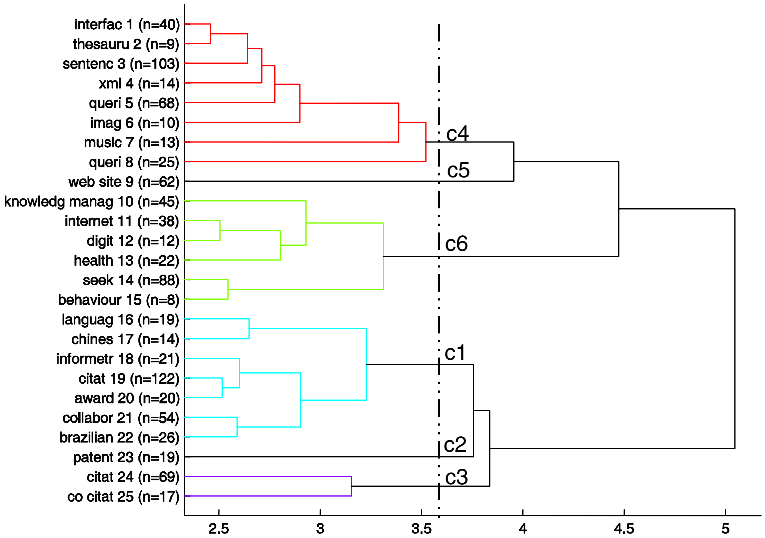

Fig. 8. Dendrogram, cut-off at 25 clusters at the left-hand side and at 6 clusters (c1–c6) at the vertical line, for hierarchical clustering of the 938 LIS papers. For each of 25 clusters, the best mean TF-IDF term is shown.

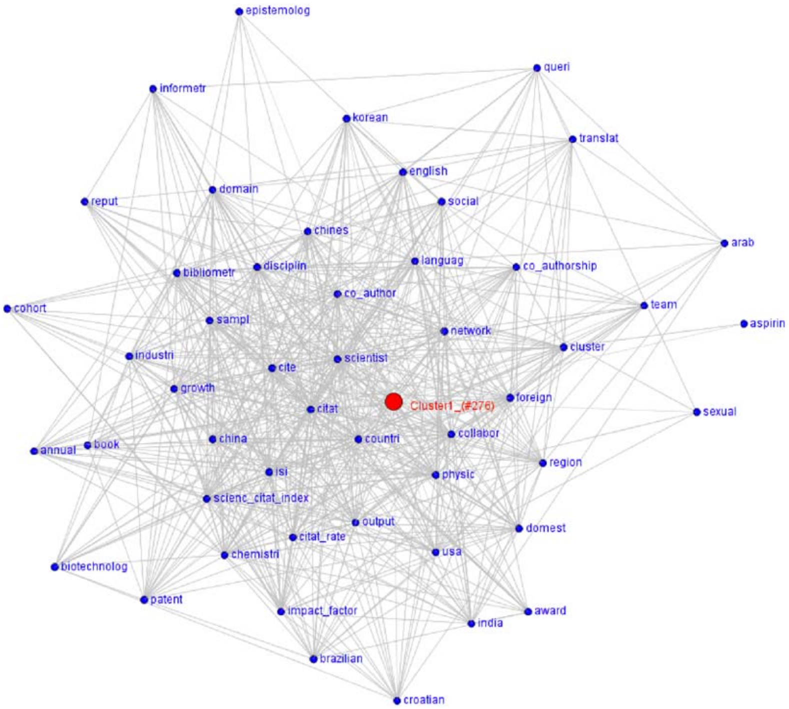

Fig. 9. Term network for Cluster 1 (276 documents), labelled as "Bibliometrics1".

methodological paper with strong implications for research evaluation, combining research collaboration with citation analysis and construction of national science indicators. Other papers "in the neighbourhood" are concerned with research evaluation, such as the assessment of research performance of countries, institutes and research groups, with journal impact analysis, with domain studies and interdisciplinarity in science, with

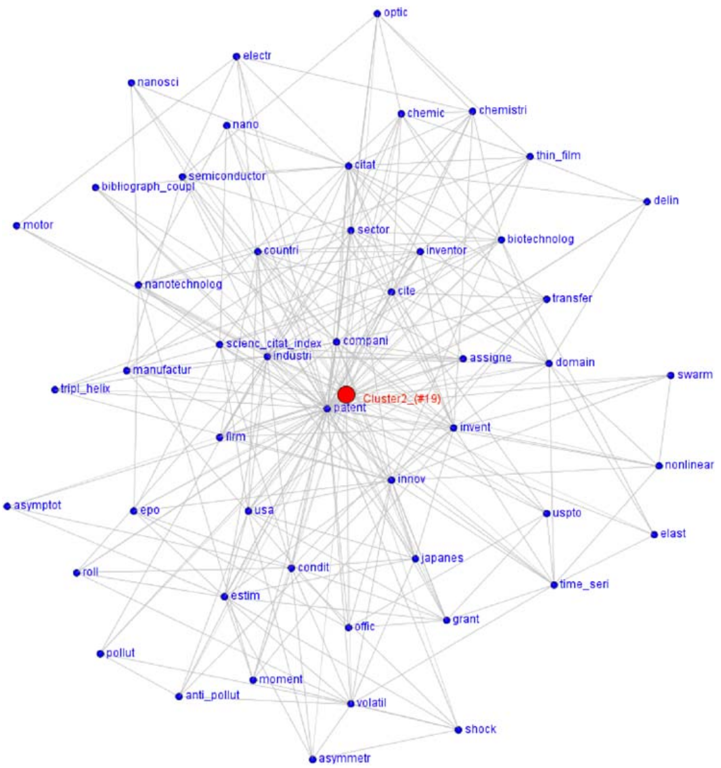

Fig. 10. Term network for Cluster 2 (19 documents), labelled as "Patent".

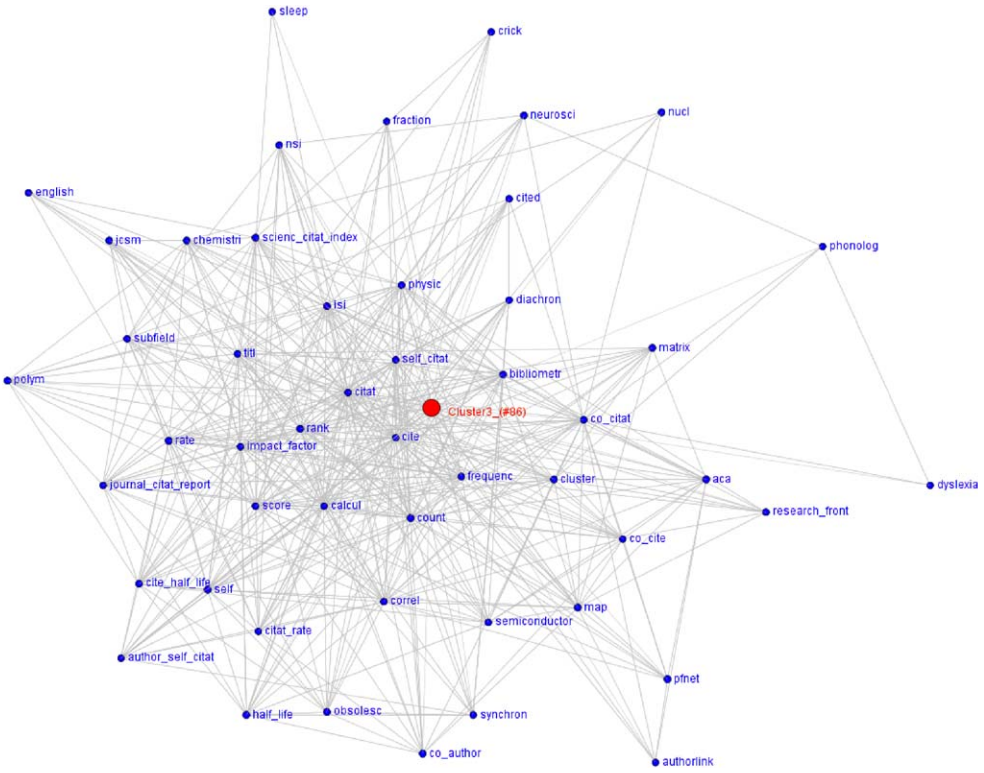

Fig. 11. Term network for Cluster 3 (86 documents), labelled as "Bibliometrics2".

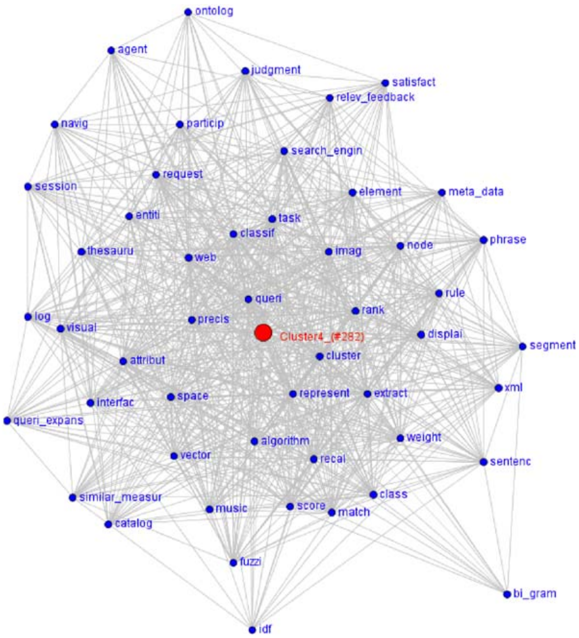

Fig. 12. Term network for Cluster 4 (282 documents), labelled as "IR" (Information Retrieval).

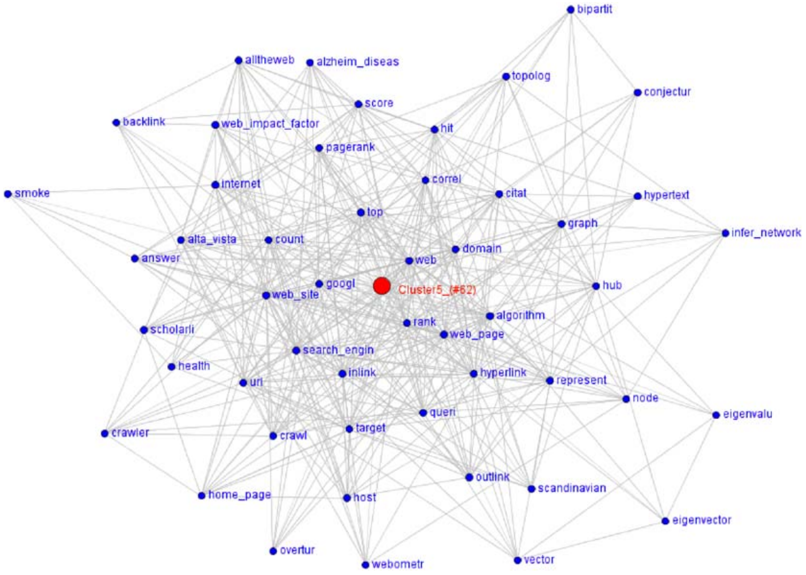

Fig. 13. Term network for Cluster 5 (62 documents), labelled as "Webometrics".

collaboration, co-publication networks and networks in the web environment, etc. Besides application, this cluster also comprises the sociological approach, technical questions in the context of bibliometrics and IR, as well as database-related aspects.

The smaller bibliometrics cluster (Cluster 3: manually labelled as "Bibliometrics2") is of more methodological/theoretical nature. The medoid is the state-of-the-art report "Journal impact measures in

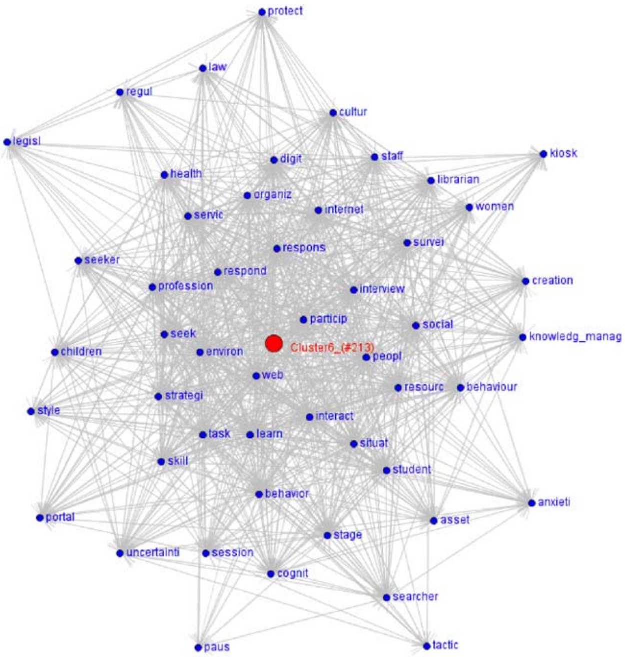

Fig. 14. Term network for Cluster 6 (213 documents), labelled as “Social”.

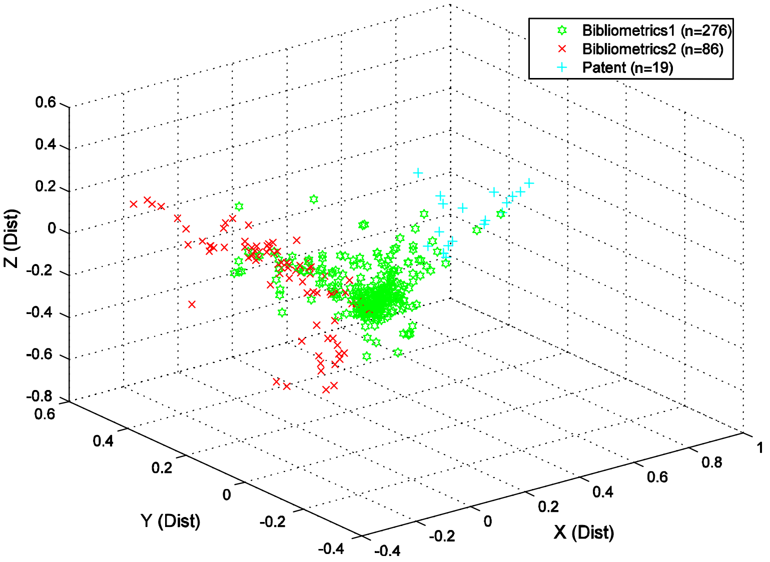

Fig. 15. MDS plot only considering the two Bibliometrics and the Patent clusters.

bibliometric research” (Appendix 2: Glänzel & Moed, 2002). Papers close to this article are dealing with methodological questions of citation analysis such as problems of delayed recognition, long-term citation analysis, highly cited publications, models for and properties of citation distributions and processes, the role of self-citations, the relationship between journal productivity and citations, author co-citation, etc.

Table 2
Share of documents and terms in each cluster and share of the 5% best terms in common with other clusters

Cluster number and name	Share of documents (%)	Number of terms (%)	Share (%) of the 5% best TF-IDF terms in common with cluster					
			1	2	3	4	5	6
1. Bibliometrics1	29.4	71.4	–	14	29	27	14	32
2. Patent	2.0	21.0	46	–	25	17	12	22
3. Bibliometrics2	9.2	44.5	46	12	–	27	12	17
4. IR	30	70	27	5	17	–	13	29
5. Webometrics	6.7	27.1	38	9	21	34	–	25
6. Social	22.7	72.2	32	6	10	29	9	–
Total	938 documents	11151 distinct terms or phrases						

The term networks for the two bibliometrics clusters just described contain a few overlapping terms (*bibliometr*, *chemistri*, *citat*, *citat rate*, *cite*, *cluster*, *countri*, *impact factor*, *isi*, *physic*, *rank* and *scienc citat index*). The MDS plot of Fig. 15 confirms that there is no clear border between Bibliometrics1 and Bibliometrics2, but that there is a gradual transition. One of the papers was clearly an IR paper about collaborative filtering and should even have been put in another cluster. But by applying the Porter stemmer (Porter, 1980), the stem for "collaborative" (collabor) is the same as for "collaboration", which was the second most important term for the Bibliometrics1 cluster. This might just serve as an example for a case when incorporating bibliographic coupling information in the clustering process might prevent the spurious association with the Bilbiometrics1 cluster.

The almost tiny Cluster 2 (19 papers, Fig. 10) represents patent analysis. A paper on "Methods for using patents in cross-country comparisons" forms the medoid of this cluster (Appendix 2: Archambault, 2002). This cluster proved to be homogeneous; all papers are concerned with technology studies, linkage between science and technology, and are at least partially relying on patent statistics. On the MDS plot of Fig. 15, the Patent cluster is much closer to Bibliometrics1 than to Bibliometrics2. The dendrogram of Fig. 8 reveals that Bibliometrics1 ("c1") even is combined first with Patent ("c2") before being combined with Bibliometrics2 ("c3").

Cluster 4, with 282 papers, is the largest one. We have labelled it "Information Retrieval". The medoid paper is entitled "Querying and ranking XML documents" (Appendix 2: Schlieder & Meuss, 2002). The full spectrum of IR related issues can be found here. Both theoretical and applied topics are represented. Although "traditional" information retrieval is covered too, Web search related issues are in the foreground. Music retrieval is also covered by this cluster; among others, all papers of the special issue on music information retrieval (*JASIST* 55 (12), 2004) can be found here.

Cluster 5, with 62 papers, belongs to the small clusters. Both terms and papers close to the medoid characterise this cluster as "Webometrics". The medoid paper is entitled "Motivations for academic web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication" (Appendix 2: Wilkinson et al., 2003).

Cluster 6 (213 papers) proved to be the most heterogeneous cluster. We have labelled it "Social", however, we could also have called it "General & miscellaneous issues". "Approaches to user-based studies in information seeking and retrieval: a Sheffield perspective" is the title of the medoid paper (Appendix 2: Beaulieu, 2003). Knowledge management, social information, evaluation of digital libraries, user feedback, user requirements for information systems, special aspects of IR such as contexts of information seeking, gender issues, use of internet facilities, etc. are among the topics covered by this cluster.

Table 2 shows the share of documents in each cluster and the share of terms or phrases from the complete vocabulary that have been used in one or more of the included papers. Next, the percentages of the terms which are among the 5% best TF-IDF terms for a cluster, and which are also present in the list of 5% best terms of another cluster, are indicated. The most frequently common terms are *citat*, *cluster*, *web*, *countri*, *domain*, *scientist*, *search engin*, *chemistri*, *queri*, *score*, *map*, *compani*, *industri*, *internet*, *task*, *bibliometr*, *collabor* and *china*.

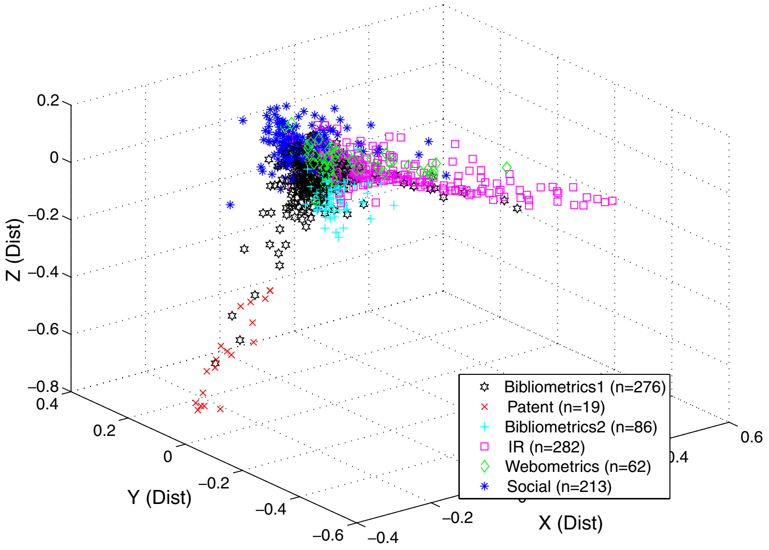

Fig. 16. 3D multidimensional scaling plot of the 938 LIS articles or notes.

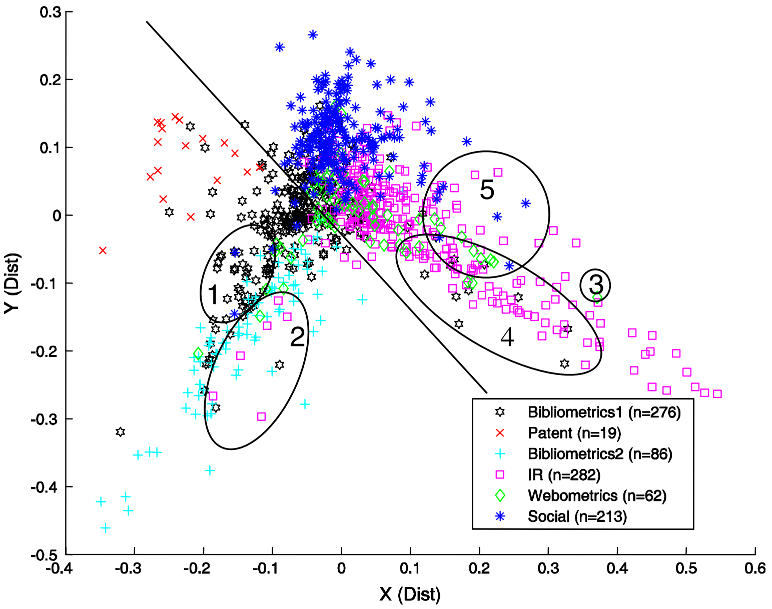

Fig. 17. 2D multidimensional scaling plot of the 938 LIS articles or notes. This figure is the projection of Fig. 16 on the X–Y plane.

Figs. 16 and 17 show the same MDS maps as in Figs. 3 and 4, but now the clusters instead of the journals are indicated. Note that there is no correspondence between journals and clusters with the same symbol or colour.

The Patent cluster can be clearly separated from the rest of LIS. The subspace under the line is almost completely occupied by Bilbiometrics1, Bibliometrics2 and Patent. Papers put by the cluster algorithm in a seemingly suspicious group will be verified in the following. First of all, there are Social papers located in the middle of the bibliometrics clusters (#1), were they rightfully added to the Social cluster? Remember that no "fuzzy clustering" was performed, so a paper could be attributed to only one cluster. Two of the papers are from *Grant Lewison* (both in *Scientometrics*) and one is by the hand of *Irene Wormell* (*JIS*). Analysing the associated titles already gives a clue about the social scope of the papers:

- "From biomedical research to health improvement" (Appendix 2: Lewison, 2002a),
- "Researchers' and users' perceptions of the relative standing of biomedical papers in different journals" (Appendix 2: Lewison, 2002b), and
- "Bibliometric navigation tools for users of subject portals" (Appendix 2: Wormell, 2003).

Next, the six IR papers that are as well embedded in the bibliometrics space (#2).

- "Algorithmic procedure for finding semantically related journals" (Appendix 2: Pudovkin & Garfield, 2002),
- "Identifying core documents with a multiple evidence relevance filter" (Appendix 2: Christoffersen, 2004),
- "Co-citation, bibliographic coupling and a characterisation of lattice citation networks" (Appendix 2: Egghe & Rousseau, 2002),
- "Citation analysis using scientific publications on the Web as data source: A case study in the XML research area" (Appendix 2: Zhao & Logan, 2002),
- "Exploiting citation overlaps for information retrieval: Generating a boomerang effect from the network of scientific papers (Appendix 2: Larsen, 2002),
- "Introduction to bibliometrics for construction and maintenance of thesauri. Methodical considerations" (Appendix 2: Schneider & Borlund, 2004).

Based on the titles, it can be concluded that at least 4 out of 6 are indeed closely related to IR and thus correct members of that cluster.

The paper that is most deeply infiltrated in the IR space but still belonging to Webometrics (#3) is "Automatic performance evaluation of Web search engines", by *Can* et al. (Appendix 2: Can et al., 2004). Again a straightforward choice. However, for most of the 11 Bibliometrics1 papers grafted onto the IR "leg" (#4) it is not at all clear why they are put in the Bibliometrics1 instead of the IR Cluster (except maybe for the second one):

- "An entropy-based interpretation of retrieval status value-based retrieval, and its application to the computation of term and query discrimination value" (Appendix 2: Dominich et al., 2004),
- "Multidimensional data model and query language for informetrics" (Appendix 2: Niemi & Hirvonen, 2003),
- "Query expansion and query translation as logical inference" (Appendix 2: Nie, 2003),
- "On bidirectional English-Arabic search" (Appendix 2: Aljlayl et al., 2002),
- "Implicit ambiguity resolution using incremental clustering in cross-language information retrieval" (Appendix 2: Lee et al., 2004),
- "Web searching for sexual information: an exploratory study" (Appendix 2: Spink et al., 2004),
- "Transitive dictionary translation challenges direct dictionary translation in CLIR" (Appendix 2: Lehtokangas et al., 2004),
- "Applying query structuring in cross-language retrieval" (Appendix 2: Pirkola et al., 2003),
- "Connectionist interaction information retrieval" (Appendix 2: Dominich, 2003),
- "Analysis of large data logs: an application of Poisson sampling on excite web queries" (Appendix 2: Ozmutlu & Spink, 2002),
- "The effectiveness of query-specific hierarchic clustering in information retrieval" (Appendix 2: Tombros et al., 2002).

The usage of common words among IR and Bibliometrics1 papers (e.g., *cluster*, *english*, *arab*, *entropi*, *sample*, *poisson*) might have contributed to the high similarity of the above papers to the Bibliometrics1 cluster.

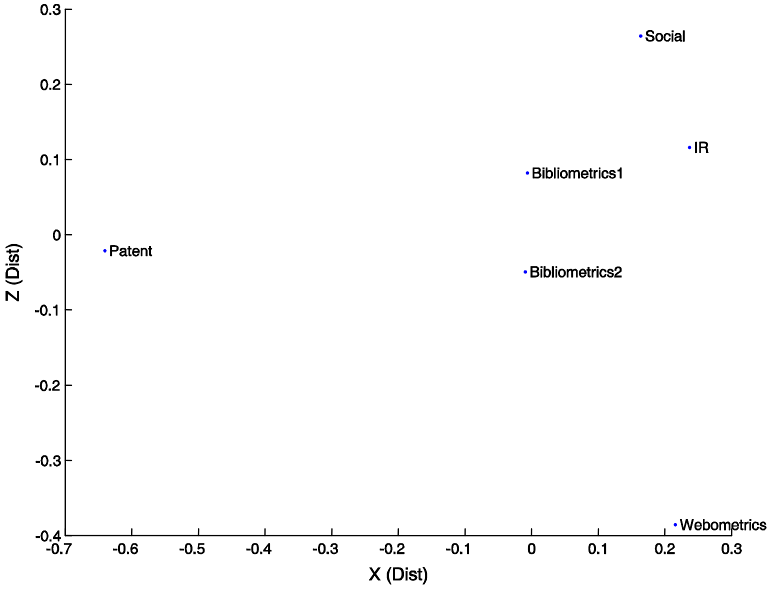

Fig. 18. MDS plot showing the distances between the centres (centroids) of the six clusters.

However, by observing the plot of individual Silhouette values for each paper in the Bibliometrics1 cluster (not shown), it was apparent that the Bibliometrics1 cluster, besides being the second largest cluster, contained the highest share of negative Silhouette values. This means that the corresponding papers should rather have been put in another cluster. The worst score, even as low as −0.4, meant that this paper was definitely put in a wrong cluster. This is an illustration of the weaknesses of the chosen agglomerative hierarchical cluster algorithm as described earlier. Since the negative Silhouette values can be detected, the corresponding documents and interpretations can be handled with care.

To conclude, the mixed character of about half of the 7 papers of the Social cluster that are also most connected to the IR field (#5), is obvious based on the titles:

- "Topic modeling for mediated access to very large document collections" (Appendix 2: Muresan et al., 2004),
- "Students' conceptual structure, search process, and outcome while preparing a research proposal: A longitudinal case study" (Appendix 2: Pennanen & Vakkari, 2003),
- "Strategic help in user interfaces for information retrieval" (Appendix 2: Brajnik et al., 2002),
- "Web search strategies and retrieval effectiveness: An empirical study" (Appendix 2: Ford et al., 2002),
- "Changes of search terms and tactics while writing a research proposal. A longitudinal case study" (Appendix 2: Vakkari et al., 2003),
- "Combining evidence for automatic Web session identification" (Appendix 2: He et al., 2002),
- "The challenge of commercial document retrieval, Part II: A strategy for document searching based on identifiable document partitions" (Appendix 2: Blair, 2002).

The centroid of a cluster is defined as the linear combination of all documents in it and is thus a vector in the same vector space. For each cluster, the centroid was calculated and the MDS of the pairwise distances between all centroids is shown in Fig. 18. As expected, the Patent cluster is the most separated one, and closest to the bibliometrics clusters. The more applied Bibliometrics1 cluster is closer to IR and Social than Bibliometrics2 is. Webometrics is, however, somewhat closer to the more methodological Bibliometrics2 cluster. Again, it should be stressed that only a two-dimensional approximation of the very high-dimensional distances is provided and that conclusions must be handled with care.

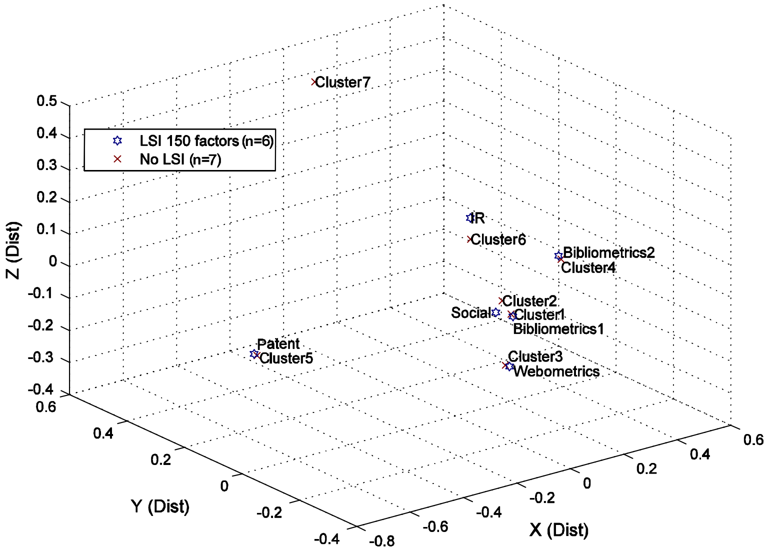

Fig. 19. MDS plot comparing the cluster centres (centroids) of the six clusters found in the LSI-transformed concept-by-document matrix (150 factors, clustering "*A*"), with the seven cluster centres when not using LSI (clustering "*B*").

4.4. Clustering without LSI

As already mentioned in Section 3.1, the number of factors for the latent semantic index is difficult to account for. In order to assess the effect of LSI on the clustering results, Fig. 19 compares the cluster centres found as described at the end of Section 4.3 (using 150 LSI factors, clustering "*A*" further on) to those of a clustering not using LSI but on the plain term-by-document matrix ("clustering *B*"). In the latter case, there is one extra cluster ("Cluster 7") because the plot of the mean Silhouette Coefficients (as in Fig. 7, but not shown here) revealed a local maximum for 7 clusters. The LSI transformation seems not to have that much influence as most of the *A* clusters correspond nicely to one *B* cluster, except for the new cluster. When analysing its contents, we observed that Cluster 7 is a dense cluster containing 14 documents about music information retrieval (MIR) with the largest mean Silhouette coefficient of all seven clusters. The terms with highest mean TF-IDF score are: *music*, *audio*, *pitch*, *mir*, *melodi*, *song*, etc. Cluster 7 contains the complete special issue of *JASIST* about "Sound Music Information Retrieval" (*JASIST* 55 (12), 2004), another *JASIST* paper and one paper from *IPM*, *SciMetr* and *JIS*. The *JIS* paper, being the medoid or the closest paper to the centroid and thus the most characteristic for the cluster, is a paper of *Aura Lippincott* about "Issues in content-based music information retrieval" (Appendix 2: Lippincott, 2002). Cluster 7 is closest to Cluster 6 and on the dendrogram (not shown), Cluster 7 is first combined with that Cluster 6, which is very close to the IR cluster from clustering *A*. Moreover, that IR cluster contains all 14 papers of Cluster 7 (MIR) of clustering *B*.

Now, why was the optimal number of clusters higher when LSI was not used? Why was MIR then considered a separate cluster? A plausible explanation is that it is an illustration of the power of latent semantic indexing to identify the general concept of information retrieval and the fact that music information retrieval is included as a part of it. Indeed, the most important terms in the MIR cluster are all very specifically about music, but because of the (possibly higher-order) co-occurrences with a lot of general information retrieval terms, they are mapped on the same LSI factors (each is a linear combination of terms). Looking at the dendrogram of Fig. 8, in the case of LSI, the MIR cluster is only split off when asking for 8 or more clusters (in this case it only contains 13 papers, the paper that was most distant from the centroid here belonging to another cluster). As we were trying to understand the field of LIS and looking for overall patterns, we preferred the solution in which a highly specific and, in this dataset at least, temporary cluster like MIR was considered part of the more general concept of IR. Thus, we deem it an advantage of LSI, next to its general noise reduction capabilities.

5. Comparing journals and clusters

The two-dimensional projection provides interesting insight in the journal presentation of LIS. IR and *IPM* almost collide in this 2D projection (Fig. 20). This means that Cluster 4 (“IR”) is very close to the scope of this journal. The “Social” cluster with general and miscellaneous topics as well as “Webometrics” are close to *JIS*, *JDoc* and *JASIST*, too. Moreover, the “Social” cluster is almost equidistant to all traditional journals in Information Science. Although this is a 2D-projection, we can nevertheless conclude that those three clusters are mainly represented by the above-mentioned four journals. The remaining three clusters, namely Bibliometrics1, Bibliometrics2 and Patent, form a triangle in the centre of which the journal *Scientometrics* is located. The relatively large distances among these clusters and between each cluster and the journal, strongly indicate that a quite large spectrum of bibliometric, technometric and informetric research using different vocabularies is covered by the journal *Scientometrics*. This observation is in line with the findings by Schoepflin and Glänzel (2001) that scientometrics consists of several subdisciplines such as informetric theory, empirical studies, indicator engineering, methodological studies, sociological approach and science policy; and that case studies and methodology became dominant by the late 1990s. At the end of the 1990s, also technology related studies based on patent statistics became an emerging subdiscipline of the field. This trend was also confirmed by the size of the bibliometric/technometric clusters (see Section 4.3). The patent cluster, still the smallest one, has the largest distance from all other clusters.

Fig. 21 visualises the share of each journal’s papers in the different clusters. Indeed, Bibliometrics1, Bibliometrics2 and Patent are predominant in *Scientometrics*. *JASIST* has a complementary profile with dominant papers about General LIS, IR and Webometrics. The reverse way of presenting, namely the share of the clusters’ papers published in the five journals, can be found in Fig. 22. Here, we have to take the journal size into account. The “triangle clusters” have their “best” representation in *Scientometrics*, while papers of the IR cluster are mostly published in *JASIST* and *IPM*. The small cluster of webometrics is almost uniquely distributed over all journals, however, there is a certain bias in favour of *JASIST*.

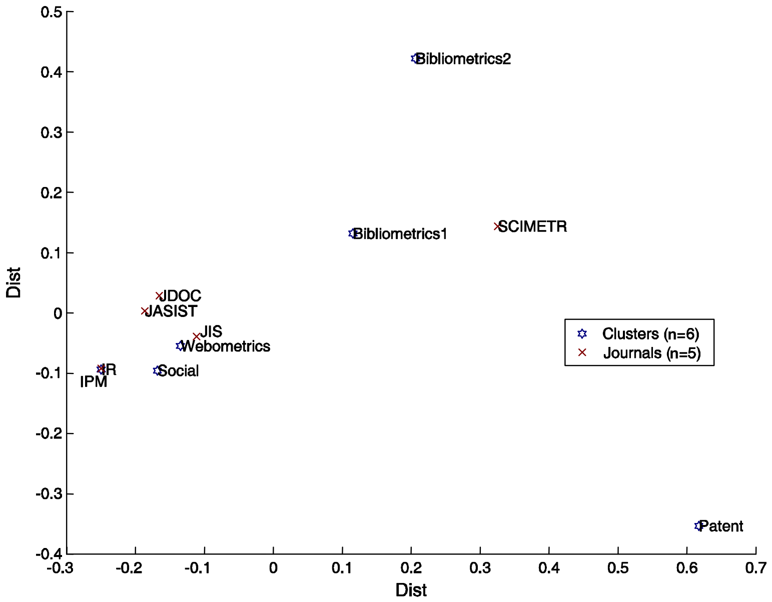

Fig. 20. MDS plot with the six cluster centres (centroids) and the five journal centroids.

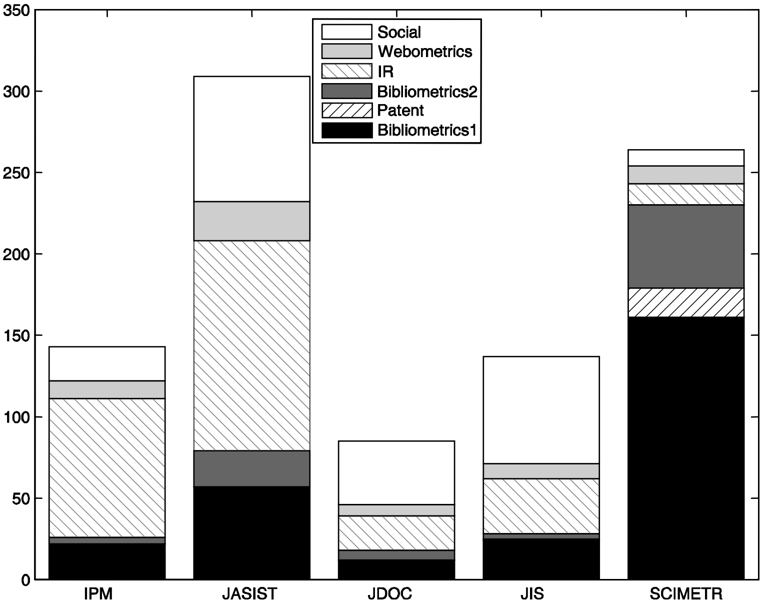

Fig. 21. Share of each journal's papers in the different clusters.

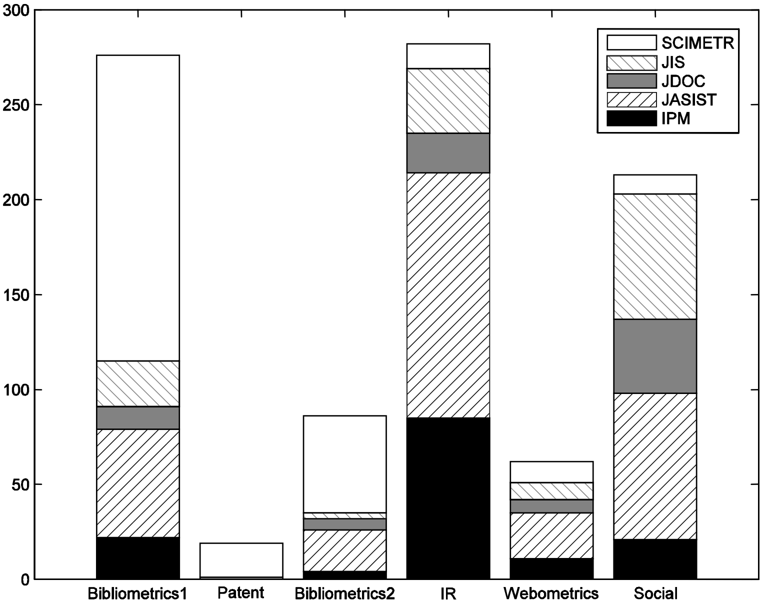

Fig. 22. For each cluster the share of papers belonging to the five journals.

6. Discussion and conclusions

In the present paper, we have analysed the concept structure of five journals representing a broad spectrum of topics in the field of Library and Information Science (LIS). We have focussed on the analysis of the "pure" text corpus, excluding any bibliographic or bibliometric components which might influence or even distort the quantitative linguistic analysis of the scientific text. We have excluded author names (except for eponyms), addresses, cited references, journal information and acknowledgements, which might otherwise already have provided cognitive links to other relevant literature. We have applied different techniques of clustering and visualising the structure of the field and of its journals.

Cluster-stability analysis according to *Ben-Hur* and the mean Silhouette value resulted in an optimum of six clusters for the selected journals and for the period 2002–2004. We have found two clusters in bibliometrics, of which a big one in applied bibliometrics/research evaluation and a smaller one in methodological/theoretical issues; also we have found two large clusters in information retrieval and general and miscellaneous issues and, finally, two small emerging clusters in webometrics and patent and technology studies. Within the IR cluster, we have found a small subcluster on music retrieval, which might be a temporary phenomenon since the journal *JASIST* has published a special issue on this topic.

The combination of cluster analysis, MDS and journal assignment has revealed interesting details about cognitive journal structure and cluster representation by journals. The about 1000 LIS papers form a characteristic "tripod" in the 3D multidimensional scaling plot. According to the expectation, IR, General issues and Webometrics were represented by four of the five journals, namely *JIS*, *IPM*, *JASIST* and *JDoc*, while the two bibliometrics and the patent clusters were the domain of the journal *Scientometrics*. The papers published in *Scientometrics* were arranged in two of the three legs forming the tripod. The "two legs" were formed by Bibliometrics1 and Patent on the one hand, and Bibliometrics1 and Bibliometrics2 on the other hand. The border between the two bibliometrics clusters is fuzzy; there is a gradual transition between methodology and application. From the viewpoint of concept structure, patent analysis can be considered an extension of evaluative bibliometrics. Moreover, the cluster dendrogram has shown that Bibliometrics1 is combined first with Patent, before being combined with Bibliometrics2. A similar polarised structure has already been observed by Schoepflin and Glänzel (2001) who have based their study on the combination of an intuitive cognitive classification and the bibliometric analysis of reference literature.

Our findings are also in line with the results of an earlier study of the authors (Glenisson, Glänzel, Janssens, & De Moor, 2005). In that study, the authors have applied a combination of full-text and bibliometric information in mapping the 85 papers published in volume 2003 of *Scientometrics*. They have studied a fine-grained structure using six clusters for the journal and indicators of cited references and they have also found a similar polarisation of scientometrics literature. The question arises whether we can validate the results of the present study by bibliometric methods and whether we can find further bibliometric characteristics of the clusters we have identified. Indicators of cited references, bibliographic coupling and cross-citations among subclusters might be appropriate tools to extend the methodology for combining full-text analysis with bibliometric indicators as suggested by the authors. However, this will be the task of a future publication.

Acknowledgement

The authors would like to thank Dr. Xavier Polanco and an anonymous reviewer for fruitful comments, and acknowledge support from:

- The Flemish Government: Steunpunt O&O Statistieken; FWO PhD/postdoc grants, projects G.0115.01 (microarrays/oncology), G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalisation), research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope, Silicos;
- Research Council KUL (GOA AMBioRICS, IDO, several PhD/postdoc & fellow grants);
- Belgian Federal Science Policy Office: IUAP P5/22 (Dynamical Systems and Control: Computation, Identification and Modelling, 2002–2006);
- EU-RTD: FP5-CAGE (Compendium of Arabidopsis Gene Expression); ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain.

Appendix 1

The 50 most important stems or stemmed phrases according to mean TF-IDF score, for each journal and for the complete dataset (938 full-texts articles or notes)

Term	IPM	JASIST	JDoc	JIS	SciMetr	All journals
queri	X	X	X	X		X
imag	X			X		X
node	X	X				X
cluster	X				X	X
vector	X	X				X
algorithm	X					X
fuzzi	X					
weight	X					X
similar measur	X					
paus	X					
session	X	X				
dierent	X					
segment	X					
web	X	X				X
sentenc	X					
bi gram	X					
web page	X					X
precis	X					
represent	X					
speciyc	X					
task		X	X	X		X
particip		X				X
student		X				X
children		X				
cognit		X				X
seek		X		X		X
music		X				X
behavio(u)r		X				X
catalog		X				
co citat		X			X	X
interact		X				X
scienc technolog		X				
score		X				X
digit		X	X	X		X
search engin		X				X
book			X			
organis			X			
thesauru			X			
frbr			X			
kiosk			X			
servic			X	X		X
loan			X			
jcsm			X			
epistemolog			X			
film			X			
entiti			X			
serendip			X			
women			X			
fiction			X			
borrow			X			
health			X	X		X
alzheim diseas				X		
meta data				X		
asset				X		
law				X		
knowledg manag				X		
respond				X		

Appendix 1 (*continued*)

Term	IPM	JASIST	JDoc	JIS	SciMetr	All journals
internet				X		X
topic map				X		
creation				X		
organiz				X		
web site				X		X
regul				X		
sim				X		
legisl				X		
preserv				X		
citat					X	X
patent					X	X
cite					X	X
countri					X	X
collabor					X	X
impact factor					X	X
scienc citat index					X	
scientist					X	X
korean					X	
self citat					X	
chines					X	X
brazilian					X	
physic					X	X
chemistri					X	
citat rate					X	
co author					X	
isi					X	
co authorship					X	
network						X
rank						X
domain						X
languag						X
social						X
electron						X
china						X
classif						X
disciplin						X
resourc						X
item						X
industri						X
interfac						X
map						X
titl						X

Shaded cells in a column mean that the corresponding terms are sorted by decreasing weight for that journal. When a term is also present in the list of a journal more to the left in the table, it is marked in a non-shaded cell on the same row, meaning that it is taken out of the ordered list for that journal. Note the last important term for IPM, ‘‘speciyc’’, which might be an illustration of errors that can occur when using OCR or text extraction techniques.

Appendix 2

Bibliographic sources of papers published in *IPM*, *JASIST*, *JDoc*, *JIS* or *Scientometrics* between 2002 and 2004, and referred to in the text as subject of analysis (in alphabetical order of the first authors):

Aljlayl et al. (2002). *JASIST, 53*(13), 1139.
Archambault (2002). *Scientometrics, 54*(1), 15.
Beaulieu (2003). *Journal of Information Science, 29*(4), 239.

Blair (2002). *Information Processing & Management, 38*(2), 293.
Brajnik et al. (2002). *JASIST, 53*(5), 343.
Breitzman & Mogee (2002). *Journal of Information Science, 28*(3), 187.
Can et al. (2004). *Information Processing & Management, 40*(3), 495.
Christoffersen (2004). *Scientometrics, 61*(3), 385.
Dominich (2003). *Information Processing & Management, 39*(2), 167.
Dominich et al. (2004). *JASIST, 55*(7), 613.
Egghe & Rousseau (2002). *Scientometrics, 55*(3), 349.
Ford et al. (2002). *Journal of Documentation, 58*(1), 30.
Glänzel & Meyer (2003). *Scientometrics, 58*(2), 415.
Glänzel & Moed (2002). *Scientometrics, 53*(2), 171.
He et al. (2002). *Information Processing & Management, 38*(5), 727.
He & Hui (2002). *Information Processing & Management, 38*(4), 491.
Larsen (2002). *Scientometrics, 54*(2), 155.
Lee et al. (2004). *Information Processing & Management, 40*(1), 145.
Lehtokangas et al. (2004). *Information Processing & Management, 40*(6), 973.
Lewison (2002a). *Scientometrics, 54*(2), 179.
Lewison (2002b). *Scientometrics, 53*(2), 229.
Lin et al. (2003). *Information Processing & Management, 39*(5), 689.
Lippincott (2002). *Journal of Information Science, 28*(2), 137.
Muresan et al. (2004). *JASIST, 55*(10), 892.
Nie (2003). *JASIST, 54*(4) 335.
Niemi & Hirvonen (2003). *JASIST, 54*(10), 939.
Ozmutlu & Spink (2002). *Information Processing & Management, 38*(4), 473.
Pennanen & Vakkari (2003). *JASIST, 54*(8), 759.
Persson et al. (2004). *Scientometrics, 60*(3), 421.
Pharo & Jarvelin (2004). *Information Processing & Management, 40*(4), 633.
Pirkola et al. (2003). *Information Processing & Management, 39*(3), 391.
Pudovkin & Garfield (2002), *JASIST, 53*(13), 1113.
Schlieder & Meuss (2002). *JASIST, 53*(06), 489.
Schneider & Borlund (2004). *Journal of Documentation, 60*(5), 524.
Spink et al. (2004). *Information Processing & Management, 40*(1), 113.
Tombros et al. (2002). *Information Processing & Management, 38*(4), 559.
Vakkari et al. (2003). *Information Processing & Management, 39*(3), 445.
Wilkinson et al. (2003). *Journal of Information Science, 29*(1), 49.
Wormell (2003). *Journal of Information Science, 29*(3), 193.
Zhao & Logan (2002). *Scientometrics, 54*(3), 449.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Cambridge: Addison-Wesley.

Ben-Hur, A., Elisseeff, A., & Guyon, I. (2002). A stability based method for discovering structure in clustered data. In *Pacific symposium on biocomputing* (vol. 7, pp. 6–17). Available from http://helix-web.stanford.edu/psb02/benhur.pdf.

Berkhin, P. (2002). Survey of clustering data mining techniques. Technical report (Accrue Software). Available from http://www.accrue.com/products/rp_cluster_review.pdf.

Berry, M., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review, 37*(4), 573–595.

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

Bonnevie, E. (2003). A multifaceted portrait of a library and information science journal: The case of the Journal of Information Science. *Journal of Information Science, 29*(1), 11–23.

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.

Callon, M., Courtial, J. P., Turner, W., & Brain, S. (1983). From translations to problematic networks. An introduction to co-word analysis. *Social Science Information, 22*(2), 191–235.

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

Courtial, J. P. (1998). Comments on Leydesdorff's article. *Journal of the American Society for Information Science, 49*(1), 98.

Daelemans, W., & van den Bosch, A. (2005). *Memory-based language processing*. Cambridge: Cambridge University Press.

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

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

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

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management, 37*(6), 817–842.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61–74.

Dutt, B., Garg, K. C., & Bali, A. (2003). Scientometrics of the international journal Scientometrics. *Scientometrics, 56*(1), 81–93.

Enright, A. J., & Ouzounis, C. A. (2001). BioLayout JAVA. *Bioinformatics, 17*, 853–854.

Glänzel, W., & Schoepflin, U. (1994). Little scientometrics – big scientometrics ... and beyond. *Scientometrics, 30*(2–3), 375–384.

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full-text and bibliometric information in mapping scientific disciplines. *Information Processing & Management, 41*(6), 1548–1572.

Glenisson, P., Glänzel, W., & Persson, O. (2005). Combining full-text analysis and bibliometric indicators. A pilot study. *Scientometrics, 63*(1), 163–180.

Grzybek, P., & Kelih, E. (2004). Anton S. Budilovič (1846–1908) – A forerunner of quantitative linguistics in Russia? *Glottometrics, 7*, 94–97.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems, 17*(2–3), 107–145.

He, S. Y., & Spink, A. (2002). A comparison of foreign authorship distribution in JASIST and the Journal of Documentation. *Journal of the American Society for Information Science and Technology, 53*(11), 953–959.

Jain, A., & Dubes, R. (1988). *Algorithms for clustering data*. New Jersey: Prentice Hall.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Janssens, F., Glenisson, P., Glänzel, W., & De Moor, B. (2005). Co-clustering approaches to integrate lexical and bibliographical information. In P. Ingwersen & B. Larsen (Eds.). *Proceedings of the 10th international conference of the International Society for Scientometrics and Informetrics* (1, pp. 284–289). Stockholm: Karolinska University Press.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: John Wiley and Sons Inc.

Kostoff, R. N., Buchtel, H. A., Andrews, J., & Pfeil, K. M. (2005). The hidden structure of neuropsychology: Text mining of the journal Cortex, 1991–2001. *Cortex, 41*(2), 103–115.

Kostoff, R. N., Toothman, D. R., Eberhart, H. J., & Humenik, J. A. (2001). Text mining using database tomography and bibliometrics: A review. *Technological Forecasting and Social Change, 68*(3), 223–253.

Leopold, E., May, M., & Paaß, G. (2004). Data mining and text mining for science & technology research. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems* (pp. 187–213). Dordrecht: Kluwer Academic Publishers.

Leydesdorff, L. (1997). Why words and co-words cannot map the development of the sciences. *Journal of the American Society for Information Science, 48*(5), 418–427.

Manning, C. D., & Schütze, H. (2000). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. New York: Academic Press, Harcourt Brace & Co.

Marshakova, I. V. (2003). Journal co-citation analysis in the field of information science and library science. In P. Nowak & M. Gorny (Eds.), *Language, information and communication studies* (pp. 87–96). Poznan: Adam Mieckiewicz University.

Marshakova-Shaikevich, I. (2005). Bibliometric maps of field of science. *Information Processing & Management, 41*(6), 1534–1547.

Mullins, N., Snizek, W., & Oehler, K. (1988). The structural analysis of a scientific paper. In A. F. J. vanRaan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 81–105). New York: Elsevier Science.

Noyons, E. (2001). Bibliometric mapping of science in a science policy context. *Scientometrics, 50*(1), 83–98.

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*, 68–81.

Onyancha, O. B., & Ocholla, D. N. (2005). An informetric investigation of the relatedness of opportunistic infections to HIV/AIDS. *Information Processing & Management, 41*(6), 1573–1588.

Persson, O. (2000). A tribute to Eugene Garfield – Discovering the intellectual base of his discipline. *Current Science, 79*(5), 590–591.

Persson, O. (2001). All author citations versus first author citations. *Scientometrics, 50*(2), 339–344.

Peters, H. P. F., & van Raan, A. F. J. (1993). Co-word-based science maps of chemical-engineering. 1. Representations by direct multidimensional-scaling. *Research Policy, 22*(1), 23–45.

Polanco, X., Grivel, L., & Royauté, J. (1995). How to do things with terms in informetrics: Terminological variation and stabilization as science watch indicators. In M. E. D. Koening & A. Bookstein (Eds.), *Proceedings of the fifth international conference of the International Society for Scientometrics and Informetrics* (pp. 435–444). Medford, NJ: Learned Information Inc.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Porter, A. L., & Newman, N. C. (2004). Patent profiling for competitive advantage. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems* (pp. 587–612). Dordrecht: Kluwer Academic Publishers.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York: McGraw-Hill, Inc.

Schoepflin, U., & Glänzel, W. (2001). Two decades of "Scientometrics" – An interdisciplinary field represented by its leading journal. *Scientometrics, 50*(2), 301–312.

Schubert, A. (2002). The Web of Scientometrics – A statistical overview of the first 50 volumes of the journal. *Scientometrics, 53*(1), 3–20.

Schubert, A., & Maczelka, H. (1993). Cognitive changes in Scientometrics during the 1980's, as reflected by the reference patterns of its core journal. *Social Studies of Science, 23*(3), 571–581.

Snizek, W., Oehler, K., & Mullins, N. (1991). Textual and nontextual characteristics of scientific papers: Neglected science indicators. *Scientometrics, 20*(1), 25–35.

Stefaniak, B. (1985). Periodical literature of information science as reflected in Referativnyj Zhurnal, section 59, informatika. *Scientometrics, 7*(3–6), 177–194.

Tijssen, R. J. W., & van Raan, A. F. J. (1989). Mapping co-word structures – A comparison of multidimensional-scaling and Leximappe. *Scientometrics, 15*(3–4), 283–295.

Todorov, R., & Winterhager, M. (1990). Mapping Australian geophysics – A co-heading analysis. *Scientometrics*(1–2), 35–56.

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

Wouters, P., & Leydesdorff, L. (1994). Has Price's dream come true? Is scientometrics a hard science? *Scientometrics, 31*(2), 193–222.

Wyllys, R. E. (1975). Measuring scientific prose with rank-frequency ("Zipf") curves: A new use for an old phenomenon. *Proceedings of the American Society for Information Science, 12*, 30–31.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645–678.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge: Addison-Wesley.

Zitt, M. (1991). A simple method for dynamic scientometrics using lexical analysis. *Scientometrics, 22*(1), 229–252.

Zitt, M., & Bassecoulard, E. (1994). Development of a method for detection and trend analysis of research fronts built by lexical or co-citation analysis. *Scientometrics, 30*(1), 333–351.