Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 41 (2005) 1548–1572
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Combining full text and bibliometric information in mapping scientific disciplines

Patrick Glenisson [a,c,*,1], Wolfgang Glänzel [a,b], Frizo Janssens [c], Bart De Moor [c]

[a] *Katholieke Universiteit Leuven, Steunpunt O&O Statistieken, Dekenstraat 2, B-3000 Leuven, Belgium*
[b] *Hungarian Academy of Sciences, Institute for Research Organisation, Nádor u. 18, H-1051 Budapest, Hungary*
[c] *Katholieke Universiteit Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium*

Received 27 September 2004; accepted 3 March 2005
Available online 23 May 2005

Abstract

In the present study results of an earlier pilot study by Glenisson, Glänzel and Persson are extended on the basis of larger sets of papers. Full text analysis and traditional bibliometric methods are serially combined to improve the efficiency of the two individual methods. The text mining methodology already introduced in the pilot study is applied to the complete publication year 2003 of the journal Scientometrics. Altogether 85 documents that can be considered research articles or notes have been selected for this exercise. The outcomes confirm the main results of the pilot study, namely, that such hybrid methodology can be applied to both research evaluation and information retrieval. Nevertheless, Scientometrics documents published in 2003 cover a much broader and more heterogeneous spectrum of bibliometrics and related research than those analysed in the pilot study. A modified subject classification based on the scheme used in an earlier study by Schoepflin and Glänzel has been applied for validation purposes.

Keywords: Automatic indexing; Full text analysis; Text-based clustering; Mapping of science; Bibliometrics

1. Introduction

In an earlier paper by Glenisson, Glänzel, and Persson (2005) full text analysis and traditional bibliometric methods were combined to improve the efficiency of the individual methods. This methodology

* Corresponding author. Address: Katholieke Universiteit Leuven, Steunpunt O&O Statistieken, Dekenstraat 2, B-3000 Leuven, Belgium.
E-mail addresses: patrick.glenisson@econ.kuleuven.be (P. Glenisson), wolfgang.glanzel@econ.kuleuven.ac.be (W. Glänzel), frizo.janssens@esat.kuleuven.ac.be (F. Janssens).
[1] Tel.: +32 016 32 57 13.

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2005.03.021

was applied to a special issue of Scientometrics, particularly to the dedicated Scientometrics issue[2] made up of selected papers presented at the *9th International Conference on Scientometrics and Informetrics* held in Beijing (China) on August 25–29 2003. The study was based on 19 selected papers that could readily be assigned to five categories, namely to the category of *Mathematical models*, the category of *Advances in Scientometrics* devoted to methodological papers, the category of *Policy relevant contributions*, the category *Webometrics/Informetrics* and finally the category of contributions devoted to bibliometric studies of *Collaboration in science*. The latter was created to represent the material of the *4th COLLNET Meeting* organised as post-event of the Beijing Conference. This category proved—as expected—rather heterogeneous. The described subject classification was used by the guest-editors of the special issue to organise, structure and present the otherwise broad and heterogeneous material of the conference in an adequate manner, but it was also used for validation purposes in the above-mentioned study. The outcomes have shown that such hybrid methodology can be applied to both research evaluation and information retrieval. Because of the limited number of papers underlying the study, the paper, however, has to be considered a pilot study that is to be extended in the present study on the basis of a larger set of papers. For this extension, the authors have chosen the complete publication year 2003, i.e., vols. 56–58 of the journal Scientometrics. The category scheme used for the Chinese proceedings issue was tailor-made to represent the scope of the conference. The 2003 issues of Scientometrics, however, represent a research profile that somewhat deviates from that special issue. Among these issues of vols. 56–58, we find, for instance, part of the proceedings of the *7th International Conference on Science and Technology Indicators* held in Karlsruhe (Germany) in 2002[3] and the special issue on the *4th Triple Helix Conference* in Copenhagen (Denmark) in 2002[4]. The latter one does not represent the mainstream of bibliometric research. This was the reason why we have adopted a modified version of the classification used by Schoepflin and Glänzel (2001). This scheme will be discussed in the following methodological section.

The idea of studying the full text of scientific literature by means of mathematical statistics, and combining these tools with bibliometric methods is not new. Mullins, Snizek and Oehler (1988; Snizek, Oehler, & Mullins, 1991) began studying structural and textual characteristics of a scientific paper already fifteen years ago. Also the idea of combining bibliometric methods with the full text analysis of a scientific paper is not new. It has its roots in modern co-word analysis developed by Callon for purposes of evaluating research (e.g., Callon, Courtial, & Laville, 1991). Braam, Moed, and Van Raan (1991) suggested combining co-citation with word analysis in the context of evaluative bibliometrics to improve efficiency of co-citation clustering. The word analysis by Braam et al. used publication ''word-profiles'' that were based on indexing terms and classification codes. Not much later, Noyons and Van Raan (1994) and Zitt and Bassecoulard (1994) demonstrated the appeal of plunging into contents by using keywords from both patent—and scientific literature to characterise the science-technology linkage. Many of these early studies were based on descriptors such as indexing terms, subject headings or keywords extracted from titles and/or abstracts.

The full text analysis applied in the study by Glenisson et al. (2005) was also supplemented by the above-mentioned traditional approach based on indexing terms extracted from titles and abstracts. The authors obtained—so to speak—as a by-product a direct comparison between the power of full text and title/abstract based text analysis.

The bibliometric part of the earlier study was restricted to simple statistical functions obtained from the papers' reference lists, particularly the mean reference age and the share of references to serial literature. Taking the outcomes of the previous papers and the opportunities offered by a medium-sized set of underlying papers into account, the following research questions will be answered:

[2] Scientometrics (2004), vol. 60, issue 3, pp. 273–534.
[3] Scientometrics (2003), vol. 57, issue 2, pp. 153–320.
[4] Scientometrics (2003), vol. 58, issue 2, pp. 191–467.

- What structures in bibliometric research as represented by the journal Scientometrics can be revealed with the help of text-mining methods?
- Are titles and abstracts descriptive enough to allow acceptable analyses based on terms or is full text to be preferred?
- Can bibliometric measures be assumed to reflect formal characteristics of documented scientific communication that might supplement results obtained from content-based analyses?
- In how far can text-mining extend, improve and explain structures found on basis of bibliometric methods?

In particular, this study aims at analysing in how far the cognitive structure of contemporary bibliometrics and informetrics is reflected by coherent clusters found in a representative selection of papers, in how far these clusters might have adequate bibliometric characteristics and in how far the two methods can supplement each other.

2. Methods

2.1. Text representation

We adopted the common vector space model to encode a document in a k-dimensional term space where each component w_{ij} represents the weight of term t_j in document d_i. The grammatical structure of the text is hereby neglected and therefore it is also referred to as a 'bag-of-words' representation. The set of all terms t_j is called a vocabulary or thesaurus.

The TF–IDF term weighting scheme is defined as follows:

$$w_{ij} = f_{ij} \log\left(\frac{N}{n_j}\right)$$

where f_{ij} is the number of occurrences of t_j in d_i and is referred to as term frequency (TF). N represents the total number of documents and n_j is the number of documents containing term t_j in the collection. The logarithm is called inverse document frequency (IDF).

We express similarity between pairs of documents d_{i1} and d_{i2} as the cosine of the angle between the corresponding vector representations as introduced by Salton and McGill (1986):

$$\mathrm{sim}(d_{i_1}, d_{i_2}) = \frac{d_{i_1} \cdot d_{i_2}}{\|d_{i_1}\| \cdot \|d_{i_2}\|}$$

The underlying hypothesis states that high similarity equals strong relevance (see Baeza-Yates & Ribeiro-Neto, 1999). We can rewrite the TF-IDF scheme as a transformation of the $n \times m$ document-term matrix A containing the raw counts: $A = PA_{tf-idf}Q$, where P is a $n \times n$ diagonal matrix with normalization constants for each document and Q an $m \times m$ diagonal matrix holding the inverse IDF's for each term. A is placed on the left side of the equation to draw the analogy with the LSI formulation described below.

To find the major associative patterns of word usage in the document collection and to get rid of the 'noise variability' in it, we used a transformation of A based on Latent Semantic Indexing (LSI). Here we assume that there is some underlying or latent structure in the word usage that is partially obscured by variability in word choice. By means of a Singular Value Decomposition (SVD) we compose another matrix A_k that is an 'optimal' approximation of A, but with rank k much lower than the column or row dimension of A (Berry, Dumais, & O'Brien, 1995; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990).

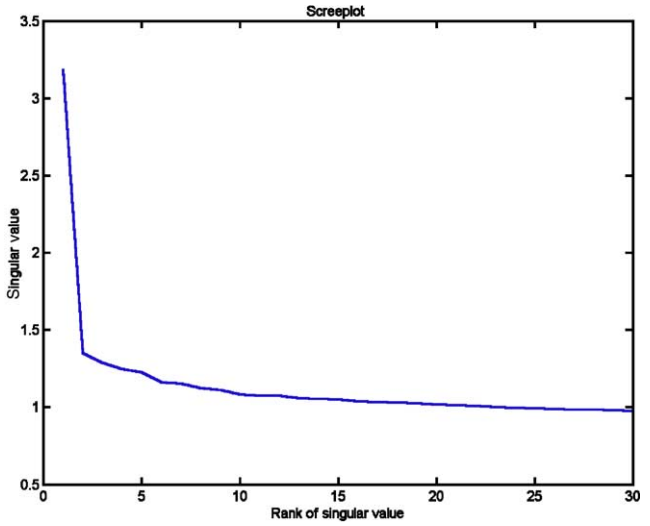

Fig. 1. Screeplot of the SVD of the document-term matrix.

More specifically, the SVD of the $n \times m$ document-term matrix is written as

$$A = U \sum V^{\mathrm{T}}$$

where $\sum = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_{\min(n,m)}})$, with $\lambda_1 \geqslant \cdots \geqslant \lambda_{\min(n,m)}$ sorted eigenvalues, and U,V orthogonal $n \times n$ and $m \times m$ matrices respectively. The best rank k approximation of A is defined by $A_k = U_k \sum_k V_k^{\mathrm{T}}$ with U_k and V_k the first k columns of U and V, and Σ_k the $k \times k$ diagonal matrix containing the k largest singular values of A.

Choosing a rank k from a screeplot (see Fig. 1) that models the semantic structure of a collection in an optimal way remains an open question and is governed mostly by empirical testing (Berry et al., 1995). The interesting effect of Latent Semantic Indexing is the fact that synonyms or different term combinations describing the same concept will be mapped onto the same factor based on the common context in which they generally appear—even for terms that do not co-occur in any document. Besides the implicit relating of synonyms, also the problem of polysemy is partly addressed by LSI. Other concept indexing methods exist, but are outside the scope of this investigation.

The indexing process itself is carried out using an in-house adaptation of the Jakarta Lucene indexing platform, which is a high-performance, open source, full-featured text search engine library written entirely in Java.[5]

2.2. Preprocessing

As in most data mining endeavours, data preprocessing, acquisition, and cleansing jointly represent up to 80% of the overall effort distribution. Preprocessing steps include the removal of stopwords and author names, stemming and the detection of phrases. Stopword removal is the process of eliminating words that have little or no semantic value, such as articles and prepositions. Author names were extracted from the references and removed from the vocabulary so to eliminate author co-citation influences in the clustering

[5] http://jakarta.apache.org/lucene/

process. Stemming involves the removal of word suffixes such as plurals, verb tenses and deflections, and the replacement by their canonized equivalent. A popular stemmer for English is the Porter stemmer (Porter, 1980), which uses a simple rule-based scheme to process the most common English words. The dimensionality of the vector space is hereby reduced, but a disadvantage is the lesser value of stems for interpretation purposes. Bigrams, or phrases composed of two words, were detected using the Dunning likelihood ratio test (Dunning, 1993; Manning & Schutze, 2000). The 500 bigrams that scored best according to this test were selected, and subsequently verified manually. After elimination of words that occur only in a single document, we finally combined the withheld words and phrases into a final thesaurus by means of which all documents were indexed. Salton's cosine on the subsequent LSI matrix was used as a basis for distance calculation.

2.3. Overall framework

An overview of the subsequent text-based and bibliometric analysis is presented in Fig. 2: we cluster the documents under consideration using a hierarchical method and compare these results with the expert

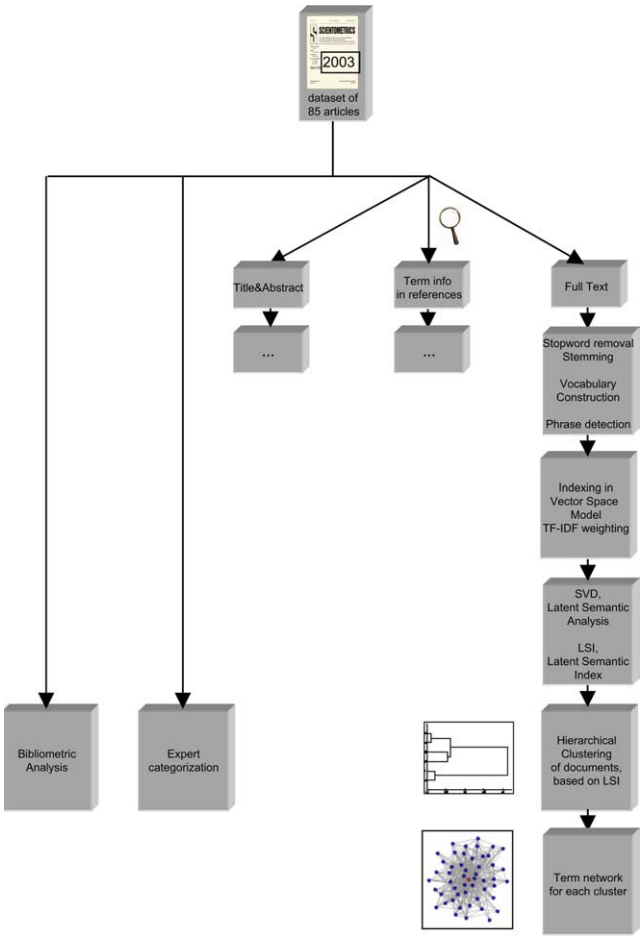

Fig. 2. Overview of the presented analysis.

category assignments as well as with a bibliometric analysis. To aid the extraction of underlying themes from the resulting clusters, we perform a co-word analysis on top-scoring terms from each cluster.

The resulting term networks are mainly intended to provide a qualitative rather than quantitative way of cluster evaluation. Biolayout Java (Enright & Ouzounis, 2001), an application originally developed for laying out and displaying interactive representations of biological graphs, is used for visualizing the networks. Two different types of edge weights exist in our term network. An edge from the large central node (red) to a term is weighted according to the importance of the term in the cluster. Intuitively, this weighting scheme forms imaginary concentric circles around the central node with more important terms appearing on a circle with smaller radius. When two terms in a network are linked by an edge, they both co-occur in the same document of the corresponding cluster, but within a given lexical distance, set to 8.

Due to the lack of ground truth and the difficulties to define a crisp categorization, we will provide an in-depth analysis how bibliometric, text-based and expert category information provide different views on the thematic structure of the document collection. Moreover, we compare the outcome of using full text information to similar approaches based on Title and Abstract, and term-based information from the References. We mention again that author names were systematically excluded from the thesauri used in our analyses.

3. Materials

As already stated in the outset, the complete publication year 2003 of the journal Scientometrics has been selected. This publication year comprises vols. 56–58 with three issues each. Only papers have been selected the organisation of which is basically in line with that of a research article. Letter to the editor, items on individuals, news items, editorial material and reviews have therefore been omitted. Our assignment is not perfectly in keeping with that used by the *Institute for Scientific Information* (ISI–Thomson Scientific, Philadelphia, PA, USA), as in three cases the assignment differs. Altogether 85 papers could thus be selected.

We have already mentioned in the introduction that we have adopted a modified version of the classification used by Schoepflin and Glänzel (2001). In particular, the classification scheme used in previous work by Glenisson et al. (2005) proved imperfect since, above all, the heterogeneous category *Collaboration in science*, used because of the COLLNET workshop, proved unnecessary for the 2003 data. The study by Schoepflin and Glänzel aimed at monitoring and characterising structural changes in the research profile in bibliometrics in the period 1980–1997. The authors created five categories, *Mathematical models/informetric laws, Case studies, Advances in Scientometrics, Indicator engineering, Sociological approaches* and *Policy relevant issues*. The term Webometrics did not yet appear in this scheme since at that time it was not yet established as a sub-discipline of scientometrics/informetrics. Allowing for structural changes we came up with the scheme in Table 1.

Table 1
Category scheme for scientometrics papers and distribution of papers over categories

Abbreviation	Description	Share (%)
A	Advances in scientometrics	31.8
E	Empirical papers/case studies	34.1
M	Mathematical models	2.4
P	Political issues	17.6
S	Sociological approaches	3.5
I	Informetrics/Webometrics	10.6

Besides the introduction of the new category I the only change in the scheme suggested by Schoepflin and Glänzel combines the former categories Empirical papers/Case studies and Indicator engineering/Data presentation to the new one denoted by E. According to the study by Schoepflin and Glänzel mainly the share of empirical and methodological papers increased between 1980 and 1997. In particular, the share of category A grew from roughly 1/5 to 1/3, whereas category E increased from about 1/4 to 1/2. This trend is contrasted by the decrease of political issues (P). The explanation for these trends is very simple. The discussion over how bibliometric indicators could best be used as tools for science policy and research management was to a large extent replaced by concrete bibliometric studies on politically relevant questions. The categories M and S remained somewhat in the periphery of bibliometric mainstream research.

In the mirror of the above results we had expected a large share of papers belonging to categories A and E. This is only in part reflected by the situation in 2003 (cfr. Table 1). Although the categories are clearly predominant as they make up about 2/3 of all papers, the share of P-class papers is somewhat unexpected. The question arises of whether there is a new structural change ongoing in our field. The fact that the Triple Helix special issue does not cover bibliometric mainstream research does only partially answer this question. The answer is much more complex. The formerly by and large clear borderlines between indicator research, sociology of science, informetric laws and science policy become more and more fuzzy, and are gradually fading away. The case of the Lamirel paper (Lamirel, Francois, Al Shehabi, & Hoffmann, 2004) discussed in our pilot study might just serve as an example for this link between Informetrics/webometrics and mathematics. Furthermore, in several papers tailor-made methods are developed and immediately applied to concrete questions in a policy-relevant context. Mathematical or statistical models are established in methodological studies. On the other hand, phenomena on the web or in information science have stimulated mathematical research. An increasing number of papers requires double or even triple assignment. In several cases a simultaneous assignment to the categories E, A and P would be most appropriate. The category assignment remains thus imperfect. We expect a positive effect of the combination of bibliometric and text-mining methods on monitoring, describing and understanding the structure of our field.

4. Clustering of full text articles

4.1. Latent semantics

We aim at laying out research themes covered by the journal Scientometrics in 2003 by constructing a *full text*-based document map, from which term networks are subsequently derived. As outlined in the Methods section, we process and index each of 85 articles' title, abstract and full text body—with exemption of the reference list. The resulting 85×3589 document-by-term matrix A is transformed to a reduced rank 6 matrix A_6 by a Singular Value Decomposition (SVD). Rank 6 was chosen through an 'educated guess' based on the decay of the singular values (Fig. 1) and is, admittedly, slightly arbitrary in the sense that other neighbouring values may be appropriate as well. We recall that this is inherent to Latent Semantic Indexing. A_6 has the same dimensions as A but is no longer sparse due to this transformation. This property induces non-zero weights on terms that are related to a document, but not occurring in it. For example, top terms for the Braun, Szabadi-Peresztegi, and Kovács-Némethl (2003) paper,

"*No-bells for ambiguous lists of ranked Nobelists as science indicators of national merit in physics, chemistry and medicine, 1901–2001*",

include 'award', 'nobel prize' and 'laureate' in the original matrix A, whereas A_6 returns 'Matthew effect', 'mean observed citation rate (mocr)' and 'scientometric'. Indeed, the first line of the article's introduction reads:

'*Rewards and recognition in scientific research can take several important aspects*',

and confirms that it is related to the Matthew effect in a scientometric context, although Matthew effect and mocr are absent in the full text of the article.

4.2. Exploratory document map

From A_6 we create a 85×85 distance matrix using the Salton cosine measure. To visualise the interrelatedness between all documents, we plot a three-dimensional MDS map and overlay the assigned category information in Fig. 3.

The Informetrics group (I) is reasonably distinguishable. The Methodological Advances (A) and Empirical case studies (E) are scattered throughout the map and give a first indication of heterogenity. The two deep-mathematical contributions (M) seem to be closely related, whereas the three sociological studies (S) are further apart. Finally, some of the policy-relevant (P) contributions are close, whereas others are, again, scattered over the cloud. Although the map constitutes a low dimensional approximation of mutual distances, we already can see that the underlying topics are likely to be intertwined or, at least closely connected, and that the subject classification is, in part, insufficient. To gain more insight in the underlying topic structure, we proceed with a detailed cluster and co-word analysis.

4.3. Document cluster analysis

Based on the computed distances we cluster the data using Ward's hierarchical clustering and cut the dendrogram at $k = 6$. To determine a statistically optimal number of clusters we used the Stability method as proposed in Ben-Hur, Elisseeff, and Guyon (2002) and used in Glenisson, Mathijs, Moreau, and De Moor (2003). In this method the optimal number of clusters k is determined by inspection of a stability diagram as in Fig. 4. The plot essentially shows a cumulative distribution of overlaps between two, recurrently computed, cluster solutions. More specifically we measure for 1000 times the overlap between

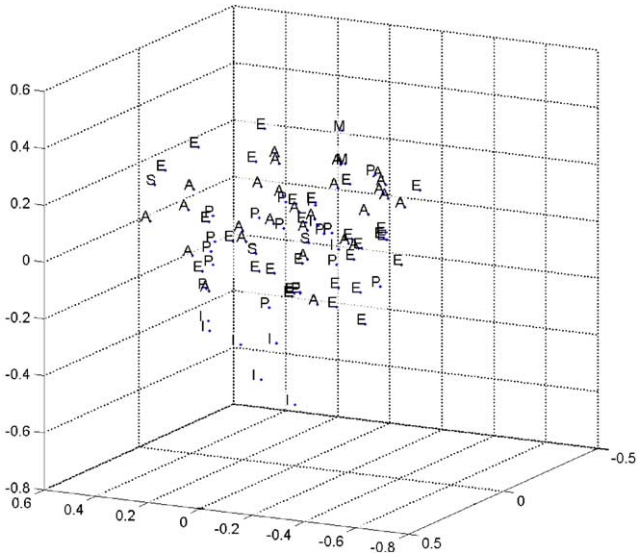

Fig. 3. MDS map of all documents in the 2003 Scientometrics issue.

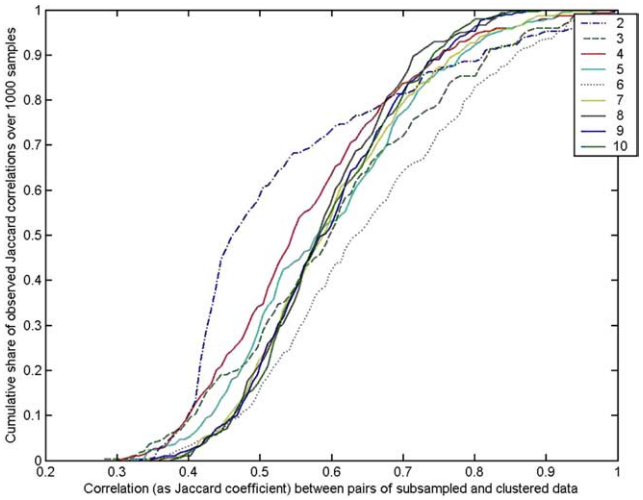

Fig. 4. Stability diagram for determination of the number of clusters k following the method in Ben-Hur et al. (2002).

clustering solutions of pairs of random subsamples from the data matrix (i.e., sampled rows from A_6). Each subsample comprises 72 documents, corresponding to 85% of 85 datapoints. We thus adopt the parameter settings $f = 0.85$ and $k = 10$ in Ben-Hur et al. (2002). The overlap, or correlation between the pair of solutions, is quantified by the Jaccard coefficient. The figure shows that for higher k the distances between the curves decrease and form a band. Typically one looks for a transition curve to this wider set of distributions.

For small k, there are several interesting observations to make. First of all the solutions are not monotonic over k, which supports the observation in the MDS map that the underlying structure is non-trivial. The cases $k = 2, 3, 4$ appear to produce solutions that are less stable than those of very fine-grained segmentations (higher k) in over 60–70% of the subsampling runs. Based on these observations we chose $k = 6$ as the most stable solution.

Silhouette values constitute the basis for checking the quality of a clustering solution when resorting solely to the data that generated it (Jain & Dubes, 1988). Silhouette values are used as a metric-independent measure that describes the ratio between cluster coherence and cluster separation for each point:

$$s_{ik} = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the average distance of member i to all other members of cluster k, and $b(i)$ the average distance of i to the points in the nearest cluster. We can summarize the silhouette values for all points in a cluster by taking an average. Likewise we can compute a score for an entire solution by subsequently averaging out over all clusters.

In Fig. 5 we illustrate the Silhouette profiles for the solution $k = 6$. For each cluster k it shows a tilted histogram of s_{ik} for all its members. We observe homogeneous structure in clusters 1, 3, 4 and 5. Cluster 2 appears to cover a broader set of topics, whereas in Cluster 6 we can expect themes that strongly incline to other clusters as several negative Silhouette values occur. These observations will be elaborated in detail in the next subsection.

To relate the clustering outcome to the expert categorization we use the Rand index (Jain & Dubes, 1988), engineered to quantify a clustering outcome with a 'gold standard' partitioning. The Rand index

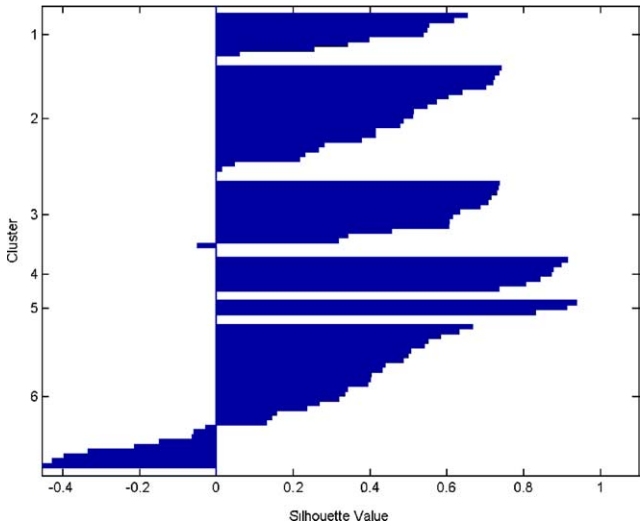

Fig. 5. Silhouette profiles resulting from full text clustering. For each member of each cluster the Silhouette value, which captures the contrast between within-cluster and between-cluster similarity, is plotted.

Table 2
Confusion table for full text clustering solution compared with expert-defined classes

Class cluster	A	E	I	M	P	S
1	7	1	0	0	1	0
2	4	13	0	0	5	0
3	8	1	3	2	0	0
4	1	0	6	0	0	0
5	0	3	0	0	0	0
6	7	11	0	0	9	3

This table is used for computation of the Rand index.

measures the correspondence between a cluster solution and an external partitioning by examining all pairs of objects: pairs that end up in the same cluster for both the computed and the expert solution are considered an agreement. The same goes for pairs that are allocated to different clusters in both outcomes. All other pairs are considered a disagreement. The statistic takes on values in the interval [0,1] where 1 indicates perfect correspondence. For the proposed clustering solution, we report a value of 0.1127 for the Rand index, which is quite low, but still significant (p-value $< 10^{-3}$).

To understand the relatively low value of the Rand index, we show the confusion table, on which the computation of the Rand index is based, in Table 2. We see classes S, M, I and P, admittedly all of smaller size, moderately to well conserved in the text-based cluster structure. Conversely, papers assigned to the larger classes A and E are heavily shifted around the text clusters. To better understand this discrepancy, we proceed with looking into the content of each cluster by:

- Examining each cluster's ranked list of documents with respect to the cluster's medoid (i.e., representative element). This is printed in Table 3.
- Analyzing the content structure of the document clusters through a co-word analysis as described in the Method section.

Table 3
Ranked papers per cluster according to distances to cluster's medoid

Distance to medoid	Title (assigned class)
Cluster 1	
0	'A macro study of self-citation(A)'
0.0374	'Better late than never? On the chance t(A)'
0.0577	'The Holy Grail of science policy: Explor(A)'
0.0866	'A new classification scheme of science f(A)'
0.0899	'No-bells for ambiguous lists of ranked N(P)'
0.1159	'The decline of Swedish neuroscience: Dec(A)'
0.1899	'Relations of relative scientometric indi(A)'
0.1974	'About Abels and similar international aw(E)'
0.2184	'Patents cited in the scientific literatu(A)'
Cluster 2	
0	'Changing trends in publishing behaviour (E)'
0.0190	'Neuroscience output of China: A MEDLINE-(E)'
0.0226	'Evaluating two Austrian university depar(P)'
0.0266	'A scientometric study of the research pe(P)'
0.0347	'The contribution of women in Brazilian s(E)'
0.0370	'Developing English-language academic jou(P)'
0.0383	'The visibility of Italian journals(E)'
0.0507	'Difficulties and challenges of Chinese s(E)'
0.0590	'One step further in the production of bi(A)'
0.0765	'More reprint requests, more citations?(A)'
0.0891	'Neo-colonial science by the most industr(E)'
0.0984	'A comparison between domestic and intern(E)'
0.1074	'Scientific cooperation between Chile and(E)'
0.1096	'Internationalization of mathematical res(E)'
0.1176	'Can scientific impact be judged prospect(A)'
0.1195	'Scientometrics of the international jour(E)'
0.1260	'Indias collaboration with Peoples Republ(E)'
0.1334	'Citation patterns in the Kuwaiti journal(E)'
0.1378	'Science from the periphery: Collaboratio(P)'
0.1420	'Exploring a pseudo-regression model of t(A)'
0.1435	'Assessing stem cell research productivit(E)'
0.2110	'The influence of cultural factors on sci(P)'
Cluster 3	
0	'OLAP and bibliographic databases(I)'
0.0331	'Abstracts, introductions and discussions(A)'
0.0358	'Mathematical model of delay in the secon(M)'
0.0434	'Informetric studies using databases: Opp(I)'
0.0611	'The diffusion of scientific publications(A)'
0.0662	'Hypothesis generation guided by co-word (A)'
0.0665	'Zipfs law and the diversity of biology n(I)'
0.0781	'The effect of statistical methods and st(E)'
0.0931	'Monitoring elasticity between science an(A)'
0.1009	'Bridging citation and reference distribu(A)'
0.1045	'Correcting glasses help fair comparisons(A)'
0.1308	'Defining a core: Theoretical observation(M)'
0.1526	'Co-citation analysis and the search for (A)'
0.2526	'Patterns in journal citation data reveal(A)'

Table 3 (*continued*)

Distance to medoid	Title (assigned class)
Cluster 4	
0	'A method for identifying clusters in set(I)'
0.0129	'Data mining in a closed Web environment(I)'
0.0228	'A vector space model as a methodological(I)'
0.0269	'Linguistic patterns of academic Web use (I)'
0.0472	'The relationship between the WIFs or inl(I)'
0.0563	'Disciplinary and linguistic consideratio(I)'
0.0602	'Mapping communication and collaboration (A)'
Cluster 5	
0	'Journal co-citation analysis of semicond(E)'
0.0069	'Author co-citation analysis of semicondu(E)'
0.0769	'The nature and relationship between the (E)'
Cluster 6	
0	'The Triple Helix as a model to analyze I(P)'
0.0065	'Entrepreneurial universities and the dyn(E)'
0.0147	'Porter vs. Porter: Modeling the technolo(P)'
0.0370	'Regional R&D activities and interactions(E)'
0.0476	'Intellectual property and public researc(S)'
0.0544	'Patterns of knowledge production: The ca(P)'
0.0781	'Towards hybrid Triple Helix indicators: (A)'
0.0964	'Publications and patents in nanotechnolo(E)'
0.0975	'Science cited in patents: A geographic f(A)'
0.1083	'Interdisciplinarity and knowledge inflow(P)'
0.1230	'Measuring the relationship between high (P)'
0.1384	'Seismology as a dynamic, distributed are(E)'
0.1460	'Quantifying the benefits of participatin(P)'
0.1471	'Do sciencetechnology interactions pay of(A)'
0.1569	'Tracing technological change over long p(E)'
0.1644	'Bibliometric analysis on additionality o(P)'
0.1722	'The Triple Helix of university industry(S)'
0.1869	'Potential science-technology spillovers (A)'
0.1979	'The mutual information of university-ind(A)'
0.2137	'Interdisciplinary information input and (E)'
0.2195	'Age profile, personnel costs and scienti(P)'
0.2348	'Constructing a multi-objective measure o(A)'
0.2420	'Large firms and the sciencetechnology in(E)'
0.2423	'A quantitative view on the coming of age(E)'
0.2483	'Critical and emerging technologies in Ma(E)'
0.2519	'Age effects in scientific productivity T(E)'
0.2678	'Studying the brain drain: Can bibliometr(S)'
0.2852	'Constructing a patent citation map using(P)'
0.2873	'Characterising intellectual spaces betwe(E)'
0.3235	'Swarming of innovations, fractal pattern(A)'

5. Scientometrics in 2003 through the eyes of text mining

The results of the co-word analysis on all documents in each cluster are shown in Figs. 6–11. The map in Fig. 6 represents the content structure of cluster 1 with altogether 9 papers. This cluster represents publications that are concerned with methodological questions related to bibliometric indicators.

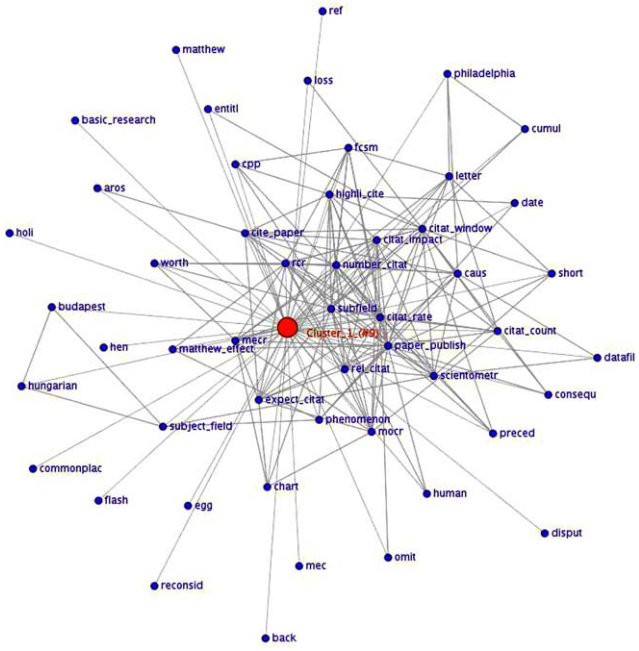

Fig. 6. Co-word map of document cluster 1.

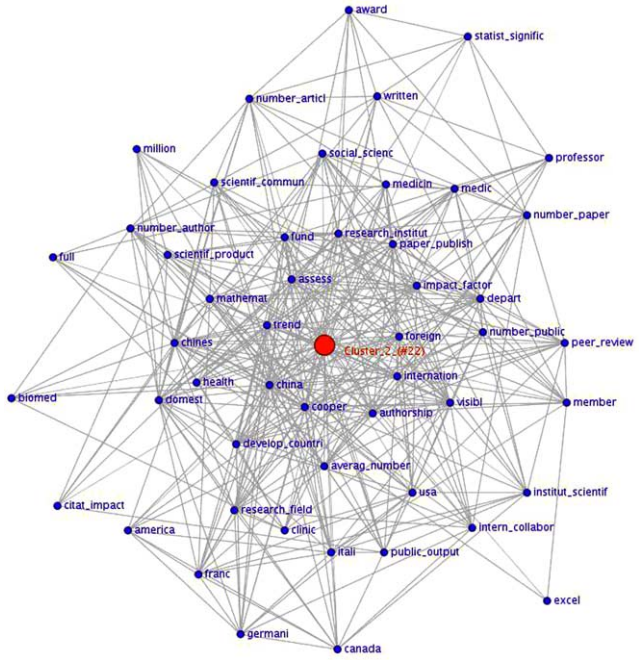

Fig. 7. Co-word map of document cluster 2.

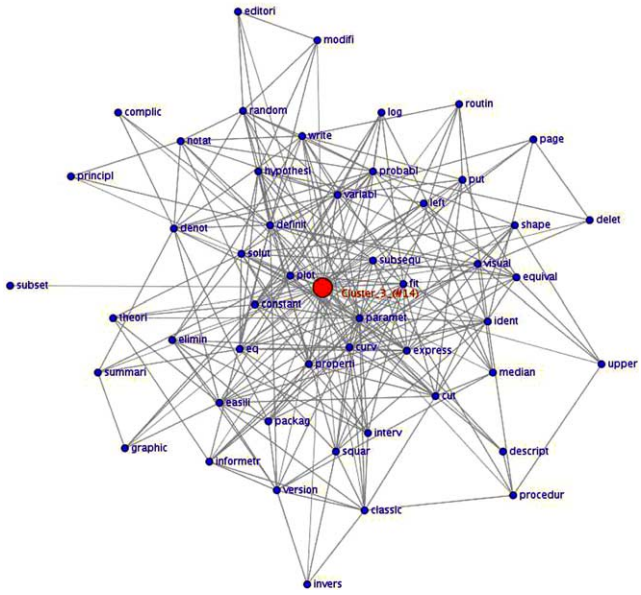

Fig. 8. Co-word map of document cluster 3.

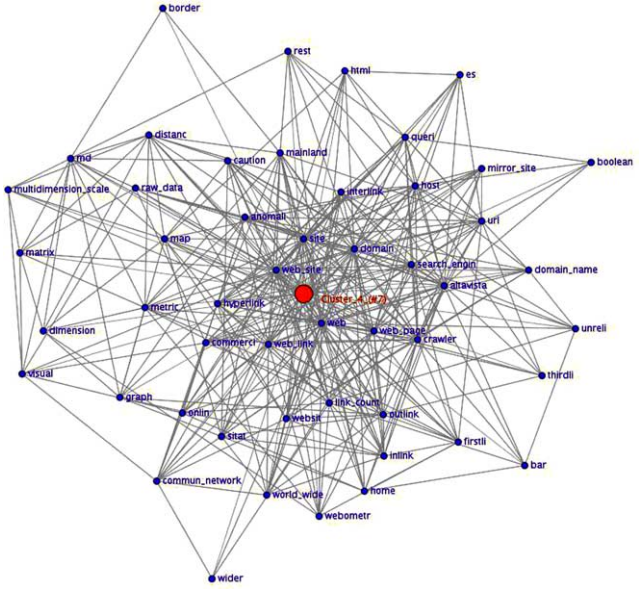

Fig. 9. Co-word map of document cluster 4.

Indicator-related terms such as indicator names and terms relevant in the context of measuring publication activity and citation impact are close to the centre, and strongly interlinked. The acronym ‘mec’ located in

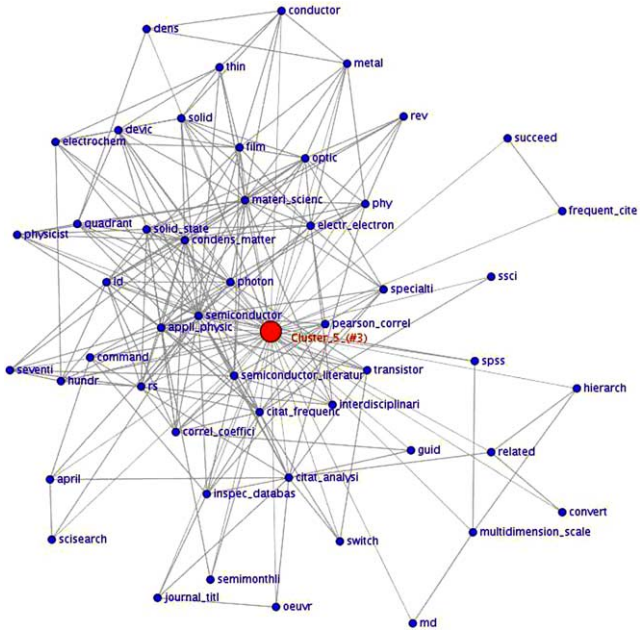

Fig. 10. Co-word map of document cluster 5.

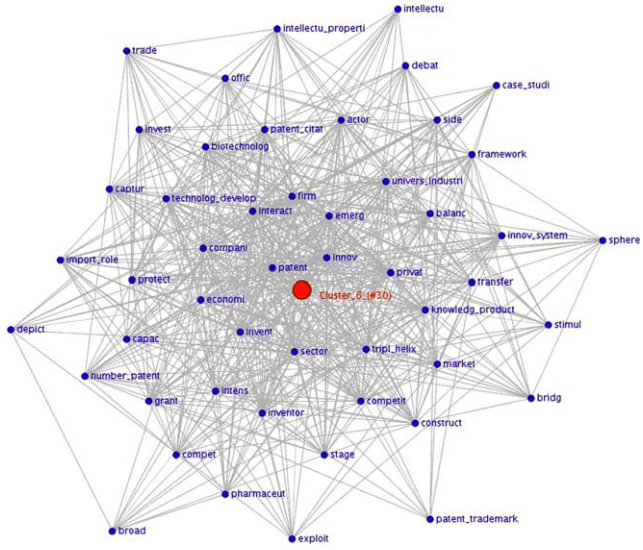

Fig. 11. Co-word map of document cluster 6.

the south of the map stands for 'Matthew Effect for Countries'. One could consider this cluster representing *methodol-ogical indicator research*.

Cluster 2 is dominated by empirical papers and case studies (cf. Table 3). The content structure in this cluster is very dense. The terms in this map are presented in Fig. 7 and relate above all to national and institutional aspects as well as to science fields. This is the cluster of *case studies and traditional bibliometric applications*.

Cluster 3 is a second theoretical/methodological cluster. Unlike the first one, this cluster relates to more advanced methodological techniques, such as informetric laws, frequency distributions and multivariate statistics. This cluster could be characterised as *theoretical and mathematical issues in bibliometrics*. The term structure is presented in Fig. 8.

Cluster 4 presented in Fig. 9 clearly represents webometrics and network-related issues. All terms are strongly interlinked. This cluster corresponds by and large to the category of *Webometrics/Informetrics*.

Cluster 5 with 3 papers is the smallest one. Co-citation analysis and the analysis of other citation statistics are the topic of these papers. The term structure (cf. Fig. 10) reflects the statistical vocabulary used in these studies. This cluster covers *specific applications of statistical methods*.

The last cluster with 30 papers (see Fig. 11) is by far the largest one. It comprises technology and innovation related studies, the science-technology interface and almost the complete Triple Helix issue can be found here (cf. Table 3). Also the sociological approaches are covered by this cluster. This cluster can be considered a borderland of classical scientometrics, namely the interdisciplinary approaches such as *sociological, policy relevant and technology related issues*.

The comparison of the topic structure based on articles given in Table 3 as well as the content structure presented in Figs. 6–11 with the category assignment in Table 1 shows only a partial accordance. The two large categories A and E covering 65% of all papers proved heterogeneous. Category A has (jointly with category M) three sub-clusters, namely, Cluster 1, 3 and 6, whereas Category E falls apart into three other sub-clusters: Cluster 2, 5 and 6. Policy relevant issues are also covered by clusters 2 and 6. Only Category I is represented by a corresponding co-word cluster, namely cluster 4. The full text analysis substantiates that both methodological and empirical research have nowadays at least two different main focuses each, one is based on scientometric standard techniques such as classical indicators, the other ones are clearly broadening the scope of traditional bibliometrics.

The phenomena discussed here is also in line with the situation visualised in Fig. 3 where only the Webometrics/Informetrics group is clearly distinguishable. Especially the groups A and E appear as heterogeneous clouds which seem—jointly with the P-papers—to fall apart. Both approaches thus substantiate that there are additional evolutions within the empirical and methodological research in bibliometrics.

6. Added value of full text with respect to Title and Abstract and reference-based term information

We extended our text analysis by comparing the outcome using full text to similar analyses using information from Title and Abstract on the one hand, and keywords present in the article's Reference section on the other hand. Table 4 shows a comparison of the Silhouette values per cluster and the Silhouette values for the entire solution in the case of $k = 6$. Note that there is no direct correspondence between the clusters across these three solutions; i.e., cluster 1 from full text does not necessarily correspond to cluster 1 from Title and Abstract, etc.

We see that the information captured in the reference titles gives rise to a less pronounced cluster structure (0.1447) than Title and Abstract, which, in turn, fares worse (0.2105) than full-text (0.5321). These trends are confirmed when assessing correspondence to the expert categories with the Rand index printed below in Table 4.

We remark, however, that it is unlikely that the structure of the data in the three cases is such that a parameterization $k = 6$ is equally probable. Hence to ensure the consistency of the observed trend over

Table 4
Silhouette values for three considered sources of information in the articles

$k = 6$ solution	Full text	Title & Abstract	(Non-author) terms from References
Silhouette value per cluster			
Cluster 1	0.3485	−0.1483	−0.006
Cluster 2	0.5665	0.3655	−0.0254
Cluster 3	0.5411	−0.1106	0.0897
Cluster 4	0.8072	0.3988	0.2343
Cluster 5	0.8034	0.4625	0.4435
Cluster 6	0.126	0.2953	0.132
Overall Silhouette for $k = 6$	0.5321	0.2105	0.1447
Rand index for $k = 6$	0.6754	0.6457	0.5896

Note that there is no formal correspondence between clusters across the three cases.

Table 5
Overall Silhouette coefficients averaged out over 9 cluster solutions for each representation

$k = 2$–10 solutions	Full text	Title & Abstract	(Non-author) terms from References
Average overall Silhouette over solutions $k = 2$–10	0.4206	0.2870	0.160
Average Rand over solutions $k = 2$–10	0.6817	0.6237	0.5968

other parameterizations in each of the cases, we averaged out the overall Silhouette and Rand values over all values $k = 2$–10 in Table 5. We see that full text still produces, on average, better cluster structures (0.4206) than Title and Abstract (0.2870) and Reference info (0.160), whereas the advantage in favour of full text is again less pronounced when considering the Rand index over various k (0.6817 versus 0.6237).

The question remains whether there are hidden mechanisms that drive this advantage. It could well be that it is the adopted clustering method, and not the information nuggets, that accounts for the observed effect over the three experiments. To test this hypothesis, a sample of clustering methods should be applied to the three methods. Three more hierarchical methods (single-link, average-link and complete-link), and one divisive method (K-medoids) were additionally applied to the three information sources. We kept the parameterization $k = 6$ fixed for each of them and show the results in Table 6. We do not discuss the difference in performance between the clustering methods as it is outside the scope of this paper. Despite the low sample size, we tested as alternative hypothesis H_1 whether the Rand index stemming from the Title and Abstract, and Reference based sources differs significantly with respect to the full text approach. The second observation in Table 6 (single-link clustering) prevents a rejection of the null hypothesis

Table 6
Comparison of Rand index over various $k = 6$ clustering solutions for the three information sources

$k = 6$ solution	Full text	Title & Abstract	(Non-author) terms from References
Ward	0.6754	0.6457	0.5896
Single link	0.3199	0.3941	0.3020
Average link	0.6697	0.6644	0.3667
Complete link	0.6754	0.6445	0.6342
K-medoids	0.6748	0.6611	0.6496

($p = 0.9907$). The inferior quality of single-link clustering, however, is known (Jain & Dubes, 1988), so we can consider it an outlier making the p-value drop to 0.0097 in favour of our research hypothesis. Nevertheless, too few samples (i.e., the four clustering methods) are now included for rigorous inference, so we cannot strictly rule out the possibility that the observed differences are due to artefacts rather than the structure of the data sources.

However, as mentioned in the discussion of Table 3, the expert classification scheme is not fully geared at classifying documents according to the scientific realms in which they reside (e.g., the category 'case studies (E)' covers quite a broad range of topics–see supra). Also, when manually comparing co-word maps across the three data structures, we found that the use of full text included more relevant phrases for interpretation (results not shown). As extra supportive evidence we provide the complete clustering solutions with $k = 6$ for Title and Abstract and Reference text in Tables A1 and A2.

Although our hypotheses might seem straightforward, it was not known a priori how informative keywords from references were for mapping purposes. Likewise, we expected that the benefits of full text information would be invisible due to the inclusion of excessive noise. The results on this data set point in the opposite direction: given appropriate preprocessing, the use of full text in the presented framework can be a valuable asset in mapping disciplines and subfields.

7. Combined text mining and bibliometrics

The statistical analysis based on the full text provided a relational chart of the structure represented by the documents under study. As already used in our pilot study (Glenisson et al., 2005), the mean reference age and the share of serials in all references can be used to characterise fields and sub-disciplines in the sciences and social sciences. In the following we will check whether these indicators can be used to characterise the six categories and/or the clusters found on the basis of the statistical full text analysis. Fig. 12 presents the plot of *Mean Reference Age* vs. *Share of Serials* on basis of the categories. This can be considered a traditional bibliometric approach although the classification of articles was peer based.

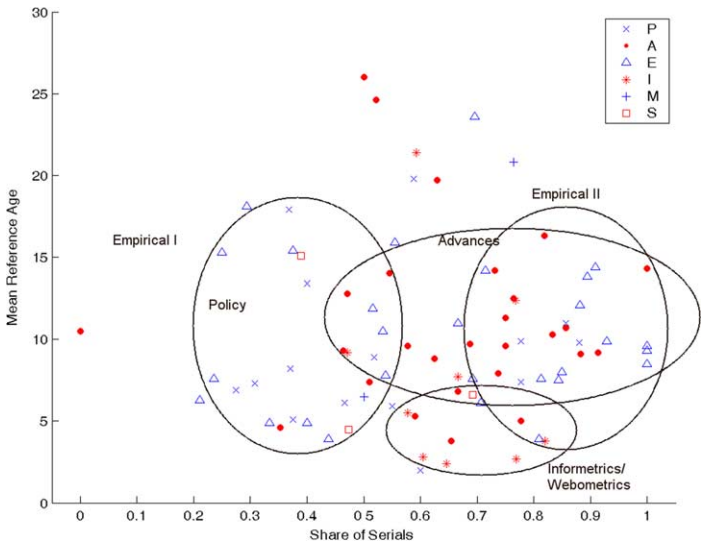

Fig. 12. Plot of mean reference age vs. share of serials on basis of the categories.

As already seen in the pilot study, Webometrics is characterised by low reference age and medium–high share of serials (cf. Glenisson et al., 2005). Most of the policy related issues are characterised by relatively low share of serials. Nevertheless, there is a group of papers with clearly higher share, too. This confirms the results of the full text analysis, namely that this category practically forms two sub-clusters. The category Advances in Scientometrics proves strikingly homogeneous with several outliers only. Most of the A-class papers have, however, a mean reference age ranging between 5 and 15 years, with medium–high share of serials ranging between 50% and 90%. The empirical groups proved heterogeneous, indeed. Regarding the share of serials this class forms two distinct sub-classes, particularly, one with low share ($\leqslant$55%) and one with relatively high share ($\geqslant$67%). The class with lower share has similar characteristics as the policy relevant class. The classes and subclasses are visualised by ellipses (see Fig. 12). The other two groups M and S are very small as compared with the other classes, and do not show any characteristic patterns.

In what follows, we combine the bibliometric approach with results of the full text analysis. For this purpose the clusters in Table 3 are used. They are represented by their medoids (i.e., representative elements). Fig. 13 presents the plot of *Mean Reference Age* vs. *Share of Serials* on basis of the co-word clusters. Similar bibliometric characteristics of clusters are much less pronounced than in the category-based approach. The number of outliers is also somewhat larger. We have indicated the two special issues (Triple Helix Conference and S&T Indicators Conference) by ellipses. These issues form surprisingly homogeneous groups. Moreover, cluster 2 is characterised by medium Mean Reference Age (MRA). Cluster 3 can be subdivided into two sub-clusters with regard to the MRA using a threshold of about 12 years. Those with older MRA have a more theoretical/informetrical focus.

On the other hand, cluster 6 consists of two sub-clusters with regard to the share of serials. Here a share of 60% forms an appropriate threshold. This threshold seems to separate more methodologically oriented scientometrics/technometrics papers and policy relevant applications. Papers with similar "content" might thus have different bibliometric characteristics depending on the target readership and the field of application.

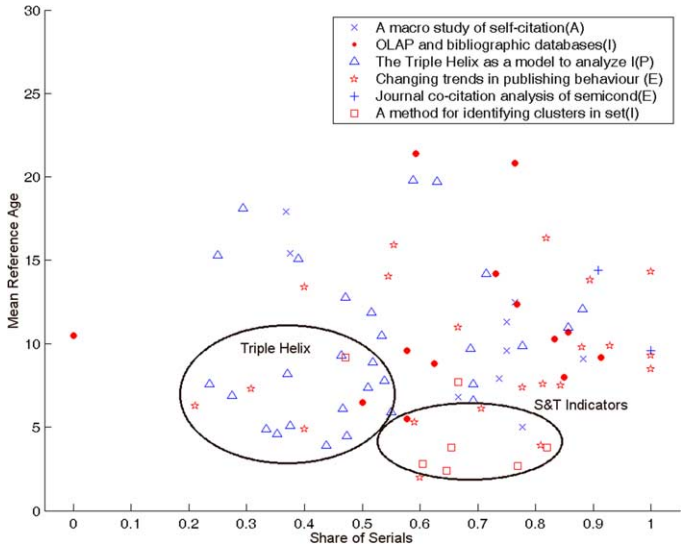

Fig. 13. Plot of mean reference age vs. share of serials on basis of the co-word clusters. Clusters are represented by their medoid documents.

8. Conclusion

The results of this study have much deepened the findings of the pilot study. The pilot study was based on a small selection of papers presented at an International Conference on Scientometrics and Informetrics. Besides quality, representative coverage was the main criterion for selection. Due to limited space in the journal, less than 20 papers could be selected, a small set of papers that showed the picture of a clearly structured discipline. Unlike this special issue, the 2003 publications in Scientometrics represent almost the complete and heterogeneous spectrum of scientometric, informetric and technometric research activity also covering topics beyond the mainstream in the field. Nevertheless, the combination of text-mining and bibliometric techniques proved an appropriate tool to answer the questions addressed in the outset.

Text-mining provided reliable results in representing structural aspects of bibliometric research if methods are based on full text. At the level of six clusters, five large clusters and a smaller but very specific one could be identified. The term structure of each cluster gave a clear picture about the research profile of the sub-disciplines represented by the cluster.

The restriction to (non-author) reference terms, on the other hand, proved to yield the weakest results. Whether to prefer full text over Title and Abstract remains, however, an issue of discussion. We found that the cohesion of clusters, their interpretability towards science mapping and correspondence to the expert classification do not deteriorate when implying full text in our mining approach. Given the limited scope of our investigation, we remain cautiously optimistic about the traceable benefits with respect to Title and Abstract information.

The clusters found through application of text mining provided additional information that can be used to extend, improve and explain structures found on basis of bibliometric methods. The full text analysis has thus shown that within the categories, such as methodological or empirical research, substantial differences in profile and orientation can occur. The question how bibliometric measures can, in turn, be assumed to reflect formal characteristics of documented scientific communication that might supplement results obtained from content-based analyses could also be answered in a positive way. Reference-based citation measures can help to fine-structure clusters determined on basis of co-word analysis. Among others, bibliometric indicators can provide information how "theoretical" or "applied", how "hard" or "soft" research within the same topic is. *Hybrid methodologies* combining data-mining techniques and bibliometric methods will therefore probably prove valuable tools to facilitate endeavours in mapping fields of science in the future.

Acknowledgements

The authors acknowledge support from the Flemish Government (Steunpunt O&O Statistieken), Research Council K.U. Leuven (GOA-Mefisto-666, GOA-Ambiorics, IDO), the Fonds voor Wetenschappelijk Onderzoek-Vlaanderen (G.0115.01, G.0240.99, G.0407.02, G.0413.03, G.0388.03, G.0229.03, G.0241.04), the Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie Vlaanderen (STWW-Genprom, GBOU-McKnow, GBOU-SQUAD, GBOU-ANA), the Belgian Federal Science Policy Office (IUAP V-22), and the European Union (FP5 CAGE, ERNSI, FP6 NoE Biopattern, NoE Etumours). The authors would like to thank the reviewers for their useful sugg- estions.

Appendix A

Tables A1 and A2.

Table A1
Ranked papers per cluster according to distances to cluster's medoid when resorting only to Title & Abstract information

Distance to medoid	Title (assigned class)
Cluster 1	
0	'Evaluating two Austrian university depar(P)'
0.0693	'Mathematical model of delay in the secon(M)'
0.1287	'Constructing a multi-objective measure o(A)'
0.1296	'Zipfs law and the diversity of biology n(I)'
0.1757	'A quantitative view on the coming of age(E)'
0.1952	'The Holy Grail of science policy: Explor(A)'
0.2084	'Interdisciplinary information input and (E)'
0.2349	'Better late than never? On the chance t(A)'
0.2667	'Assessing stem cell research productivit(E)'
0.3157	'Relations of relative scientometric indi(A)'
0.4961	'About Abels and similar international aw(E)'
0.5523	No-bells for ambiguous lists of ranked N(P)'
Cluster 2	
0	'Patents cited in the scientific literatu(A)'
0.0846	'Measuring the relationship between high (P)'
0.0956	'Tracing technological change over long p(E)'
0.0985	'Informetric studies using databases: Opp(I)'
0.1119	'Monitoring elasticity between science an(A)'
0.1209	'Characterising intellectual spaces betwe(E)'
0.1277	'Science cited in patents: A geographic f(A)'
0.1284	'OLAP and bibliographic databases(I)'
0.1399	'A new classification scheme of science f(A)'
0.1780	'Bridging citation and reference distribu(A)'
0.1790	'Large firms and the sciencetechnology in(E)'
0.1796	'Constructing a patent citation map using(P)'
0.1881	'Porter vs. Porter: Modeling the technolo(P)'
0.269	'Studying the brain drain: Can bibliometr(S)'
0.2768	'Defining a core: Theoretical observation(M)'
0.2998	'Critical and emerging technologies in Ma(E)'
0.3135	'A vector space model as a methodological(I)'
Cluster 3	
0	'The visibility of Italian journals(E)'
0.047	'More reprint requests, more citations?(A)'
0.1216	'Difficulties and challenges of Chinese s(E)'
0.1284	'The nature and relationship between the (E)'
0.1915	'The effect of statistical methods and st(E)'
0.2036	'The diffusion of scientific publications(A)'
0.3920	'Journal co-citation analysis of semicond(E)'
0.4273	'Author co-citation analysis of semicondu(E)'
Cluster 4	
0	'One step further in the production of bi(A)'
0.0553	'Age profile, personnel costs and scienti(P)'
0.0718	'Scientometrics of the international jour(E)'
0.0723	'Correcting glasses help fair comparisons(A)'
0.0755	'Indias collaboration with Peoples Republ(E)'
0.1080	'The contribution of women in Brazilian s(E)'
0.1266	'Can scientific impact be judged prospect(A)'
0.1281	'A comparison between domestic and intern(E)'
0.1480	'Neuroscience output of China: A MEDLINE-(E)'
0.1551	'Science from the periphery: Collaboratio(P)'

Table A1 (*continued*)

Distance to medoid	Title (assigned class)
0.1620	'A scientometric study of the research pe(P)'
0.1729	'A macro study of self-citation(A)'
0.1766	'Age effects in scientific productivity T(E)'
0.2358	'Developing English-language academic jou(P)'
0.2412	'Abstracts, introductions and discussions(A)'
0.2482	'Scientific cooperation between Chile and(E)'
0.2581	'Neo-colonial science by the most industr(E)'
0.2950	'Changing trends in publishing behaviour (E)'
0.3013	'Internationalization of mathematical res(E)'
0.3188	'The decline of Swedish neuroscience: Dec(A)'
0.3311	'Citation patterns in the Kuwaiti journal(E)'
0.3902	'Seismology as a dynamic, distributed are(E)'
Cluster 5	
0	'Potential science-technology spillovers (A)'
0.0487	'The Triple Helix of university industry(S)'
0.0488	'Patterns of knowledge production: The ca(P)'
0.0640	'Do sciencetechnology interactions pay of(A)'
0.0692	'Towards hybrid Triple Helix indicators: (A)'
0.0768	'Intellectual property and public researc(S)'
0.1013	'The mutual information of university-ind(A)'
0.1079	'Regional R&D activities and interactions(E)'
0.1098	'Entrepreneurial universities and the dyn(E)'
0.1184	'Swarming of innovations, fractal pattern(A)'
0.1574	'The Triple Helix as a model to analyze I(P)'
0.1669	'Publications and patents in nanotechnolo(E)'
0.2114	'Bibliometric analysis on additionality o(P)'
0.2155	'Interdisciplinarity and knowledge inflow(P)'
0.2822	'Quantifying the benefits of participatin(P)'
Cluster 6	
0	'Exploring a pseudo-regression model of t(A)'
0.0799	'Co-citation analysis and the search for (A)'
0.1150	'Mapping communication and collaboration (A)'
0.1861	'Patterns in journal citation data reveal(A)'
0.1878	'Data mining in a closed Web environment(I)'
0.1881	'Hypothesis generation guided by co-word (A)'
0.1883	'Disciplinary and linguistic consideratio(I)'
0.2118	'A method for identifying clusters in set(I)'
0.2340	'The relationship between the WIFs or inl(I)'
0.3187	'Linguistic patterns of academic Web use (I)'
0.4395	'The influence of cultural factors on sci(P)'

Table A2
Ranked papers per cluster according to distances to cluster's medoid when resorting only to Reference-based text information

Distance to medoid	Title (assigned class)
Cluster 1	
0	'Correcting glasses help fair comparisons(A)'
0.6998	'Bridging citation and reference distribu(A)'
0.7308	'Relations of relative scientometric indi(A)'

(*continued on next page*)

Table A2 (*continued*)

Distance to medoid	Title (assigned class)
0.7482	'Zipfs law and the diversity of biology n(I)'
0.8141	'A macro study of self-citation(A)'
0.8169	'Seismology as a dynamic, distributed are(E)'
0.8403	'Age effects in scientific productivity T(E)'
0.8417	'Changing trends in publishing behaviour (E)'
0.8558	'The visibility of Italian journals(E)'
0.8609	'The decline of Swedish neuroscience: Dec(A)'
0.8621	'Mapping communication and collaboration (A)'
0.8652	'Studying the brain drain: Can bibliometr(S)'
0.8790	'The Holy Grail of science policy: Explor(A)'
0.8815	'Better late than never? On the chance t(A)'
0.8840	'A scientometric study of the research pe(P)'
0.8880	'Bibliometric analysis on additionality o(P)'
0.8915	'Scientometrics of the international jour(E)'
0.8993	'Constructing a multi-objective measure o(A)'
0.9004	'Exploring a pseudo-regression model of t(A)'
0.9063	'Neo-colonial science by the most industr(E)'
0.9155	'Indias collaboration with Peoples Republ(E)'
0.9180	'Assessing stem cell research productivit(E)'
0.9327	'About Abels and similar international aw(E)'
0.9343	'Science from the periphery: Collaboratio(P)'
0.9361	'Can scientific impact be judged prospect(A)'
0.9423	'Mathematical model of delay in the secon(M)'
0.9450	'Patterns in journal citation data reveal(A)'
0.9497	'Internationalization of mathematical res(E)'
0.9534	'No-bells for ambiguous lists of ranked N(P)'
Cluster 2	
0	'The contribution of women in Brazilian s(E)'
0.3937	'One step further in the production of bi(A)'
0.8335	'A comparison between domestic and intern(E)'
0.8758	'Intellectual property and public researc(S)'
0.9352	'Age profile, personnel costs and scienti(P)'
0.9500	'Evaluating two Austrian university depar(P)'
0.9549	'The effect of statistical methods and st(E)'
0.9582	'Tracing technological change over long p(E)'
0.9607	'Quantifying the benefits of participatin(P)'
0.9639	'Interdisciplinarity and knowledge inflow(P)'
0.9645	'Patterns of knowledge production: The ca(P)'
0.9680	'Abstracts, introductions and discussions(A)'
0.9762	'The influence of cultural factors on sci(P)'
0.9785	'Porter vs. Porter: Modeling the technolo(P)'
0.9817	'A new classification scheme of science f(A)'
0.9844	'Measuring the relationship between high (P)'
0.9852	'More reprint requests, more citations?(A)'
0.9925	The diffusion of scientific publications(A)'
0.9988	'A quantitative view on the coming of age(E)'
Cluster 3	
0	'Author co-citation analysis of semicondu(E)'
0.5160	'Journal co-citation analysis of semicond(E)'
0.5763	'Co-citation analysis and the search for (A)'
0.7999	'The nature and relationship between the (E)'

Table A2 (*continued*)

Distance to medoid	Title (assigned class)
0.8022	'Citation patterns in the Kuwaiti journal(E)'
0.8414	Hypothesis generation guided by co-word (A)'
0.8778	'OLAP and bibliographic databases(I)'
0.8910	'Informetric studies using databases: Opp(I)'
0.9206	'Defining a core: Theoretical observation(M)'
Cluster 4	
0	'The relationship between the WIFs or inl(I)'
0.3711	'A method for identifying clusters in set(I)'
0.4163	'Disciplinary and linguistic consideratio(I)'
0.4479	'Linguistic patterns of academic Web use (I)'
0.6665	'A vector space model as a methodological(I)'
0.7071	'Data mining in a closed Web environment(I)'
0.9731	'Scientific cooperation between Chile and(E)'
Cluster 5	
0	'Difficulties and challenges of Chinese s(E)'
0.2700	'Neuroscience output of China: A MEDLINE-(E)'
0.3926	'Developing English-language academic jou(P)'
0.5666	'Constructing a patent citation map using(P)'
Cluster 6	
0	'Science cited in patents: A geographic f(A)'
0.4202	'Do sciencetechnology interactions pay of(A)'
0.5824	'Towards hybrid Triple Helix indicators: (A)'
0.6021	'Entrepreneurial universities and the dyn(E)'
0.6029	'The Triple Helix as a model to analyze I(P)'
0.6100	'The Triple Helix of university industry(S)'
0.6325	'Large firms and the sciencetechnology in(E)'
0.6804	'Potential science-technology spillovers (A)'
0.6934	'Characterising intellectual spaces betwe(E)'
0.7028	'Patents cited in the scientific literatu(A)'
0.7304	'The mutual information of university-ind(A)'
0.7521	'Regional R&D activities and interactions(E)'
0.7703	'Swarming of innovations, fractal pattern(A)'
0.7788	'Interdisciplinary information input and (E)'
0.7792	'Monitoring elasticity between science an(A)'
0.9002	'Publications and patents in nanotechnolo(E)'
0.9051	'Critical and emerging technologies in Ma(E)'

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. ACM Press.

Ben-Hur, A., Elisseeff, A., & Guyon, I. (2002). A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing, 7*, 6–17.

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

Braam, R., Moed, H., & Van Raan, A. (1991). Mapping of science by combined co-citation and word analysis. 2. Dynamical aspects. *Journal of the American Society for Information Science., 42*(4), 252–266.

Braun, T., Szabadi-Peresztegi, Z., & Kovács-Némethl, E. (2003). No-bells for ambiguous lists of ranked Nobelists as science indicators of national merit in physics, chemistry and medicine 1901–2001. *Scientometrics, 56*(1), 3–28.

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), 155–205.

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*, 391–407.

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

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

Glenisson, P., Mathijs, J., Moreau, Y., & De Moor, B. (2003). Meta-clustering of gene expression data and literature-extracted information, SIGKDD explorations. *Special Issue on Microarray Data Mining, 5*(2), 101–112.

Glenisson, P., Glänzel, W., & Persson, O. (2005). Combining full text analysis and bibliometric indicators. A pilot study. *Scientometrics*, forthcoming.

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

Lamirel, J. C., Francois, C., Al Shehabi, S., & Hoffmann, M. (2004). New classification quality estimators for analysis of documentary information, application to patent analysis and web mapping. *Scientometrics, 60*(3), 445–462.

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

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

Noyons, E. C. M., & Van Raan, A. F. J. (1994). Bibliometric cartography of scientific and technological developments of an R&D field. The case of Optomechatronics. *Scientometrics, 30*, 157–173.

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

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

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

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.