Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 63, No. 1 (2005) 163–180

Combining full-text analysis and bibliometric indicators. A pilot study

PATRICK GLENISSON,[a] WOLFGANG GLÄNZEL,[a,b] OLLE PERSSON[c]

[a] *Katholieke Universiteit Leuven, Steunpunt O&O Statistieken, Leuven (Belgium)*
[b] *Hungarian Academy of Sciences, Institute for Research Policy Studies, Budapest (Hungary)*
[c] *Inforsk, Department of Sociology, Umeå University, Umeå (Sweden)*

In the present study full-text analysis and traditional bibliometric methods are combined to improve the efficiency of the individual methods in the mapping of science. The methodology is applied to map research papers from a special issue of *Scientometrics*. The outcomes substantiate that such hybrid methodology can be applied to both research evaluation and information retrieval. The subject classification given by the guest-editors of the special issue is used for validation purposes. Because of the limited number of papers underlying the study the paper is considered a pilot study that will be extended in a later study on the basis of a larger corpus.

Introduction

Bibliometric methods proved valuable tools to monitor and chart scientific processes; the validity of bibliometric indicators is, however, somewhat limited as secondary representations of full text documents are used. Indeed, when considering publications as atomic entities in scientometric studies, one can readily describe and analyse the bibliometric relationship between elements of a given set of scientific publications, but lexical connections remain covered when taking this viewpoint. Mullins, Snizek and Oehler (MULLINS et al., 1988; SNIZEK et al., 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 et al., 1991). BRAAM et al. (1991a,b) 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

Received November 5, 2004

Address for correspondence:
PATRICK GLENISSON
Katholieke Universiteit Leuven, Steunpunt O&O Statistieken
Dekenstraat 2, B-3000 Leuven, Belgium
E-mail: Patrick.Glenisson@econ.kuleuven.ac.be

0138–9130/US $ 20.00

and classification codes. Not much later, NOYONS & VAN RAAN (1994) and ZITT & BASSECOULARD (1994) demonstrated the appeal of plunging into contents by using keywords from both patent- and scientific literature to characterize 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 analysis of full-text analyses took recently a sharp rise soon as large textual databases became available in electronic form.

While in the above papers word analysis was used to complement bibliometric techniques to improve the performance of bibliometric methods, also the converse direction is possible. Thus Kostoff uses bibliometric methods to supplement results from *database tomography* (e.g., KOSTOFF, 2001). Especially text-mining techniques developed for the information retrieval may gain new fields of application in research evaluation through combination with traditional bibliometric methods.

Terms are the building blocks to organize, store and access information and they hold a key position in the field of information retrieval. In what follows, we show how a approach of indexing full-text scientific articles combined with an exploratory statistical analysis can complement bibliometric approaches in the mapping of science. In this study we present a systematized methodology to index textual documents and further characterize them using standard data mining techniques. The statistical analysis based on the full text of a set of documents provides a relational analysis of the topicality represented by these documents; the bibliometric component of the analysis adds characteristics describing their position in the set. Finally, we have added a topic classification based on peer evaluation of the documents as a third component serving at the same time as a validation of the two applied methods.

Although most of the individual steps used in following analysis are common ground to most scientists involved in information retrieval or statistics, properly combining and leveraging them to presentable results often remains rather art than science – especially when considering full-text articles.

In the past we have successfully operationalised a similar, document-centered, text mining approach in the biomedical field (GLENISSON, 2003a; 2003b). Here, we conduct a systematic lexical analysis on all full-text articles in the special *Scientometrics* issue* made up of selected papers presented at the *9th International Conference on Scientometrics and Informetrics* held in Beijing (China) on 25-29 August 2003. This conference was organised under the auspices of the International Society for Scientometrics and Informetrics (ISSI) and locally of the Chinese Association for Science of Science and S&T Policy (CASTP). The previous eight events in this series of biannual International Conferences on Scientometrics and Informetrics have been held in Belgium (1987), Canada (1989), India (1991), Germany (1993), USA (1995),

* *Scientometrics*, 60 (3) (2004) pp. 273–534.

Israel (1997), Mexico (1999) and Australia (2001). These ISSI Conferences provide an umbrella for the presentation of research results in all topics in bibliometrics, webometrics and related specialities. The special issue could thus be seen as representative of the research trends in scientometrics, and consequently of interest to the audience of the corresponding journal. Moreover, in the future, we plan to extend the results to other full-text papers of *Scientometrics*. These factors constitute the rationale of selecting this dataset for our pilot study.

To conclude, the main objective of this pilot study is to analyse in how far the cognitive structure of contemporary bibilometrics and informetrics, as published in the journal *Scientometrics*, is reflected by coherent, text-based clusters found in a representative selection of papers, and in how far these clusters have adequate bibliometric characteristics. We check whether the found attributes are in keeping with intellectual assignments made by the guest-editors of the dedicated issue based on both content evaluation and quantitative metrics.

Data sources

All scientific papers published in *Scientometrics,* 60 (3) (2004) have been used as source for this study. In all, nineteen research papers representing all topics in our fields were selected for the Beijing issue. According to their topics, these papers have been assigned to five sections by the guest editors (cf., ***Glänzel et al.****). The topic structure of the dedicated issue is presented by Table 1. This documents classification serves as the third, peer-based component in the following analysis.

The issue sections in Table 1will be called *classes* in the following analysis.

For the full-text analysis all figures and tables as well as mathematical equations have been removed. In addition, the abstracts have undergone a supplementary text analysis. The reference lists of the papers served as the source for the bibliometric indicators. We summarize the adopted analytic approach as follows:

1. Use a document-centered text clustering framework to map full-text research papers.
2. Assess correspondence of the outcome with the guest editor's class assignments to the nineteen documents under study.
3. Compare the results when restricting the analysis to the articles' title and abstract.
4. Compare the text-based mapping to classical bibliometric analysis and investigate complementarity.

* Author names in ***bold-italics*** refer to their contribution in *Scientometrics*, 60 (3) (2004) as given in the Appendix.

Table 1. Guest editor's thematic structure of the 'Beijing issue' in *Scientometrics* 60 (3) (2004)

Section code	Section name	Paper
I	Advances in Scientometrics	***Havemann et al.***
		Moed & Garfield
		Small
		Yue & Wilson
II	Policy relevant issues	***Negishi et al.***
		Shelton & Holdridge
		Markusova et al.
		Wu et al.
III	Bibliometric approaches to collaboration in science	***Beaver***
		Kretschmer
		Persson et al.
		Yoshikane & Kageura
IV	Advances in Informetrics and Webometrics	***Lamirel et al.***
		Qiu & Chen
		Tang & Thelwall
		Vaughan & Wu
V	Mathematical models in Informetrics and Scientometrics	***Egghe***
		Glänzel
		Shan et al.

The full-text analysis

Text representation

Vector space model. A systematic way to encode textual information in a computer-amenable format is to represent a document as an object in a k-dimensional term vector space. As a result each document d_i has k components w_{ij}, which correspond to the weights of term t_j in d_i. As the grammatical structure of the text is neglected in this process, such a vector space model is often referred to as a 'bag-of-words' representation of text. The TF-IDF weighing scheme is defined as follows:

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

where f_{ij} is the number of occurrences of t_i in d_i and is often referred to as term frequency (TF). N represents the total number of documents and n_j is the number of documents containing term j in the collection. The logarithm is often called inverse document frequency (IDF). Indexing is the calculation of these term weights for each document and the result of this process is a document-by-term matrix.

We express similarity between pairs of documents d_{i_1} and d_{i_2} or between a text document d_{i_1} and a query document d_{i_2} as the cosine of the angle between the

corresponding vector representations as introduced by Salton:

$$sim(d_{i_1}, d_{i_2}) = \frac{d_{i_1} \cdot d_{i_2}}{\left\| d_{i_1} \right\| \cdot \left\| d_{i_2} \right\|}.$$

The underlying hypothesis states that high similarity equals strong relevance (see BAEZA-YATES, 1999). Salton's measure has an advantage over the Pearson correlation in that the similarity is insensitive to the number of zeros (LARSEN, 2002). In the mapping of publications the IDF weighting scheme is used.

Thesaurus construction. The construction of a literature index starts with the collection of a set of documents in ASCII format. These documents might be abstracts, full-text reports, database entries, emails,… The document corpus is processed by removing punctuation, case and document structure. Standard English stemming using the Porter stemmer canonises the words according to morphological and inflexional endings (e.g., plurals, tenses …) and helps to reduce to a certain extent the dimensionality as well as the dependency between terms. Although it has been recently reported (KOSTOFF, 2003) that stemming is context-dependent, we ignore this in the currently presented approach. A removal of words including articles, prepositions and conjunctions is desirable to reduce noise and is done using analysis of term distribution, and/or using a handcrafted stopword list. This process yields a first thesaurus or 'bag-of-words'.

Phrases are terms consisting of several word (e.g., 'information retrieval') and although little is known on how, or if, they affect the performance of learning text (e.g., document classification, document retrieval), they are important when extracting comprehensible information (e.g., keyword-based summarization of a document). We adopt the log-likelihood ratio (see MANNING, 2000) to computationally detect statistically overrepresented bigram phrases. We manually pruned the 900 top-scoring phrases to a smaller list of 434 and appended them to the thesaurus, the five highest scoring being: dimensional_informetrics, citation_impact, diachronous_prospective, web_site and co_authorship. All in all we obtain 4568 single terms and 434 bigrams.

During our analysis we discard the following lexical concepts, but we define them here as they bear some relevance to the discussion of our results: Synonyms and acronyms are different terms conveying the same meaning or referring to the same object (e.g., 'tumour'/'tumor', 'SCI'/'Science Citation Index'), whereas polysemy refers to terms conveying different meanings according to the context they appear in (e.g., 'CD' as compact disk, Crohn's disease, cytosine deaminase …).

Exploratory analysis

Data projection and visualization. Using our corpus-derived thesaurus, we indexed the Beijing issue's 19 articles using the IDF representation, hence obtaining a

(19x3610) document-term matrix. Subsequently we calculate a cosine-based distance matrix *D* (19x19) to assess mutual similarity between documents. To visualize these interrelations, which constitute a high-dimensional document-map, we apply classical metric multidimensional scaling on *D*, and plot the projection of the articles onto the two principal dimensions in Figure 1. The points in the chart represent all 19 articles which are labelled with the corresponding authors. To assess the correspondence with the guest-editors' classification, we use different markers for each topic class (see Figure legend).

We see that human judgment on the topic structure is surprisingly well reflected in this low dimensional visualization. Especially classes II and V form coherent clusters that are located in diametrically opposite quadrants. Also the papers of class IV form a distinct cluster although the paper by ***Lamirel et al.*** seems in a way to link this cluster to other classes, above all to class V. In the ***Lamirel*** paper, information retrieval methods are applied to the Web, whereas the three other papers (***Qiu & Chen, Tang & Thelwall, Vaughan & Wu***) are concerned with the analysis of patterns of university and company weblinks. The ***Lamirel*** paper uses Galois lattices as underlying model in information retrieval what might explain the observed tendency and the relative outlier position of the paper in question. Nevertheless, these three clusters reflect a clear polarisation where the classes form a triad. The other two classes cannot be separated by corresponding clusters. This result does not really surprise: although Collaboration in Science has become an important research topic in our field, this topic is quite heterogeneous. It has, for instance, research components from sociology of science, from policy relevant evaluative studies and from mathematical models of co-author networks as well. Nevertheless, the ***Kretschmer*** paper dealing with structural and ***Yoshikane & Kageura*** paper concerned with dynamic aspects of individual collaboration networks and the paper by ***Persson et al.*** analysing the interaction of collaboration with other characteristics of scientific communication form a coherent subgroup of class III. On the other hand, the papers by ***Moed & Garfield*** analysing patterns of documented scientific communication as reflected by references to frequently cited papers, by ***Small*** on highly cited papers and by ***Yue & Wilson*** concerned with the bibliometric analysis of scientific journals form a cluster within class I. Only the papers by ***Beaver*** and ***Havemann et al.*** proved outliers in their classes.

Summarising one can conclude that the selection made from the Beijing contributions reflects a strong polarisation between the mathematical-theoretical approaches and the policy-relevant issues. The two poles are linked by the methodological issues in scientometrics which, in turn, include the topic collaboration in science. Webometrics forms a distinctly separate cluster which might drift apart from the other classes and evolve to an own specialty in future.

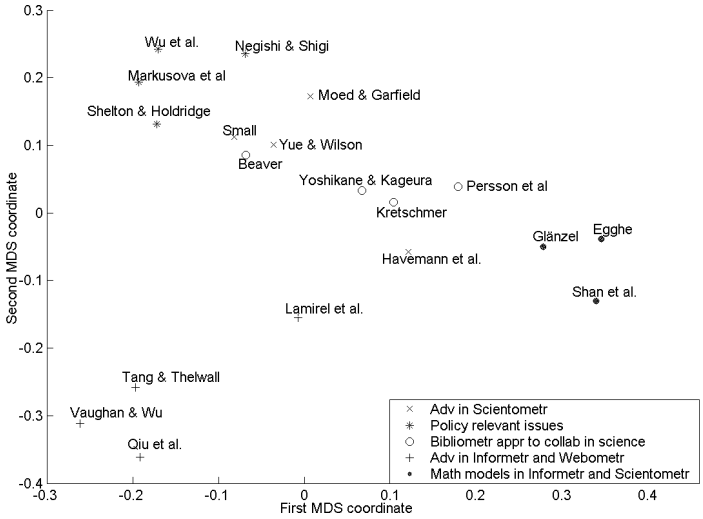

Figure 1. Projection of 19 articles onto 2D space

Hierarchical clustering. Apart from a visual exploration of the interrelations – which can at times be misleading when class labels are known - we test how well the topics reflected in the special issue's papers can be discovered by unsupervised clustering. To this end we apply Ward's hierarchical clustering (see JAIN, 1988) to *D*.

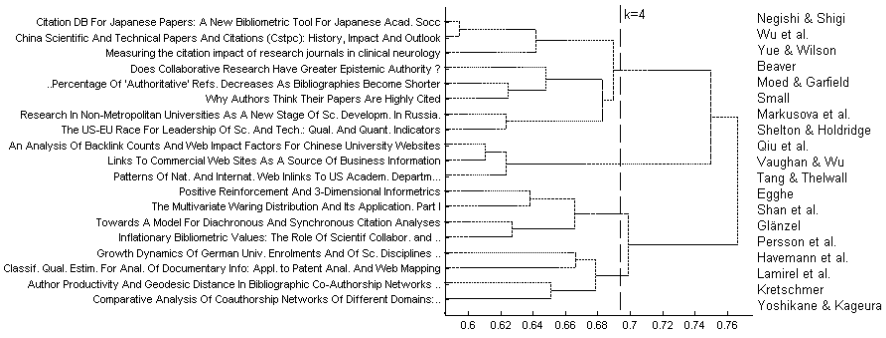

Figure 2. Dendrogram of this issue's contributions clustered using full-text analysis
(The cutoff of the tree, which directly determines the number of clusters is shown as a dashed line.
The cluster number is indicated near each cutoff point.)

The resulting dendrogram is plotted in Figure 2. Automatically determining the number of clusters, and hence delineating underlying themes, is a complex issue in unsupervised learing. We chose $k = 3$ or $k = 4$, which is indeed a plausible choice given the shape of the dendrogram. It can be argued that the number of points to be clustered is rather small, and thus compromising stability of the clustering solution. However, given the fact that the documents under study can be considered as representative samples of the field, we adopt the reasonable assumption that overall results will not be influenced drastically by increasing the number of points.

To measure how well the solution $k = 4$ corresponds to the thematic structure in Table 1 we show the confusion in Table 2. Each entry in this table contains the number of papers shared by the inferred cluster and the corresponding class. For example, cluster 1 contains three papers from class V and a single paper from class III. Cluster 2 almost corresponds to entire class IV: Advances in Informetrics and Webometrics (up to 1 paper, see dendrogram). Cluster 3 encompasses the class V: Mathematical Models in Informetrics and Scientometrics plus one extra paper (see dendrogram). The polarisation found in Figure 1 becomes even clearer if the dendogram is read. The two remaining clusters 1 and 4 are formed by science policy and methodological papers.

Table 2. Confusion table of unsupervised clustering (k=4) with the guest editors' classification of papers into topics

Cluster \ Class	I	II	III	IV	V
1	3	4	1	0	0
2	0	0	0	3	0
3	0	0	1	0	3
4	1	0	2	1	0

Full-text versus abstract-only clustering

To assess to which extent the abstracts in the special issue's contributions constitute a condensation of the actual themes, we redo the above cluster exercise using only title and abstract information. In Figure 3 we plot the 2D projection of the selected papers using title and abstract information only.

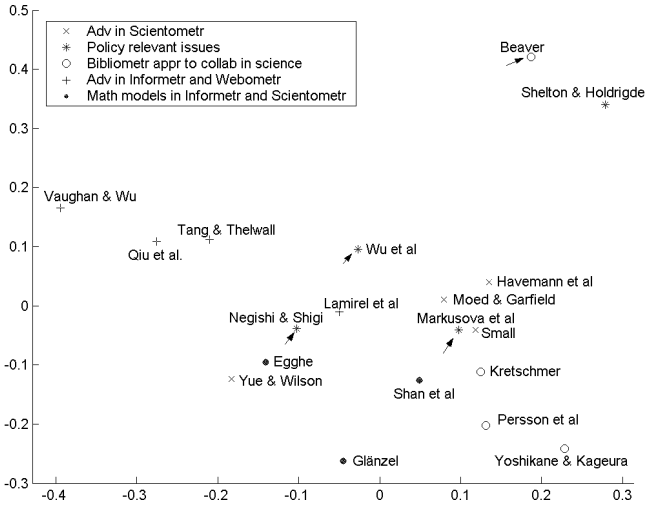

Figure 3. Projection of 19 articles onto 2D space using information only from title and abstract
(Articles marked with an arrow display sharp deviations in the keyword-based information content captured in the abstract and the full-text.)

With respect to Figure 1 we observe that, apart from four instances, the relative positioning of the articles has not changed dramatically. However, the articles marked with an arrow display significant differences in information content of abstract and full-text respectively. An interesting observation reveals that all papers significantly different from full-text approach are concerned with policy-relevant topics or are close to that class (cf. Figure 1).

Besides the qualitative analysis above, we measure the ability of *abstract-only* versus *full-text* clustering to produce the thematic grouping from Table 1 by comparing the two solutions using the Rand index (JAIN, 1988), which is engineered to quantify a clustering outcome with a 'gold standard' partitioning. 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 as a disagreement. The statistic takes on values between [0, 1] where 1 indicates perfect correspondence. The full-text clustering result gives a Rand index of 0.7778, which corroborates our earlier observations. Conversely, adopting the same pre-processing steps, term weighting schemes and clustering parameters, the abstract-only clustering returns a value of 0.6257. Nevertheless, these two values give poor insight in the significance of

the cluster solutions with respect to random patterns. Therefore, Rand indices for 200 clustering permuted cluster solutions (both for *abstract-only* and *full-text* respectively) were computed. The one-sided *p*-values of the observed Rand indices with respect to these sampled empirical distributions are $<10^{-3}$ and 0.464 for the full-text and abstract-only setup respectively. We thus see that the correspondence of the abstract-only clustering with the subject classification fails to be statistically significant, which provides us a quantitative argument of implying full-text in the methodology.

The above results show the advantage of using full-text over title and abstract in our exercise to measure correspondence of text-based document clusters with external expert categories. However, more detailed, quantitative studies on the information content in structured documents exist (e.g., LOSEE, 1996; KOSTOFF, 2004). We see these type of analyses as complementary to our approach as they enable the identification of those parts of full-text documents that are maximally informative by some relevance metric. In future work, less informative parts of a full-text document could be excluded upfront from the document index.

Information extraction

For the extraction of relevant keywords, we used the TF-IDF representation (instead of IDF). The underlying reason is that frequency of keywords within a document are not strongly relevant to model *interrelatedness*, but, on the contrary, give important clues on the topicality of a single document. In our experiments, probably due to the limited number of documents to properly model the term distributions, we found the IDF-representation superior for stable and good clustering results, whereas TF-IDF was more appropriate for extracting meaningful keywords. In Table 3 we provide text *profiles* for (arbitrarily chosen) representative documents from each topic class.

For each of the articles we see the occurrence of theme-relevant keywords and their weights. However, acronyms are not taken into account in the analysis, giving rise to tautologies such as EU (European Union), gTLD (generic Top Level Domains). Some of the acronyms contain dots (e.g., U.S.); as a result they remain undetected as they cross the sentence boundary. Neither did we accommodate for spelling variations such as 'per cent' and 'percent' or stemming peculiarities such as 'psychologi' (psychology) or 'pl' (PLS, acronym for 'partial least squares'). Also, the small size of corpus gives rise to erroneous phrases that pass our filter such as 'percent_most', or 'frequent_cite'.

Nevertheless, the text profiles provide clear clues on the aboutness of each of the articles and it is fair to state that (a combination of) keywords from the profile are good starting points to retrieve related documents or to extract salient sentences. In a similar manner we could aggregate several documents of each category and collect keywords to profile the corresponding subdomains, but the restricted number of instances in this dataset (3-4 articles per category) would yield overly specific terms.

Table 3. Text profiles and term weights of selected representative documents

Author(s): ***Persson et al.*** Class: III		Author(s): ***Glänzel*** Class: V	
co_author	0.417794	diachronous_prospect	0.492265
collabor*	0.287652	synchronous	0.377403
domest*	0.208460	synchronous_retrospect	0.360994
self_citat*	0.185298	age	0.250921
explan*	0.170916	diachronous_prospect	0.238375
Growth	0.154099	technic*_reliabl*	0.180497
Reference_list	0.151925	citat*_process	0.150553
intern*_collabor*	0.151925	life_time	0.147679
reference_behaviour	0.151468	impact_measur*	0.125460
inflationari	0.151468	random_select*	0.114862

Author(s): ***Moed & Garfield*** Class: I		Author(s): ***Shelton & Holdridge*** Class: II	
research_field	0.358836	EU	0.638957
authorit*_docum*	0.281942	WTEC	0.346503
authorit*	0.241017	panel	0.224208
docum*	0.197558	output_indic*	0.142678
referenc*	0.179418	NAS	0.142678
percent_most	0.179418	leadership	0.142678
refer*_list	0.176746	world	0.119689
refer*	0.165171	input	0.114998
frequent*_cite	0.156779	row	0.102220
persuasion	0.153787	panelist	0.101913

Author(s): ***Tang & Thelwall*** Class: IV	
department	0.420497
intern*_inlink	0.315920
gTLD	0.273798
public_impact	0.189552
disciplin*	0.148494
psychologi	0.145234
command	0.145234
region	0.135706
histori	0.123676
disciplinari_differ*	0.105307

Bibliometric analysis

The statistical analysis based on the full text provided a relational chart of the topicality represented by the documents under study. The information extraction in the previous section provided information about the content of selected representative of each class. The extension of this analysis would in most cases not reveal any common properties (keywords) of papers within one and the same class. The three papers in class V are, for instance, concerned with completely different topics: ***Egghe*** showed how compound informetric phenomena can be expressed through the composition of

different processes, ***Glänzel*** introduced a field of special stochastic citation processes to model prospective and retrospective aspects of the ageing of scientific literature and ***Shi et al.*** analysed properties of the multivariate Waring distribution as the basis for the application to author productivity. In the light of these considerations it is useful to point out why these papers still form a cluster in our, essentially keyword-based, topic map. We therefore look at the terms display highest pointwise correlation $\sigma_j = w_{i_{1,j}} w_{i_{2,j}}$ – and hence contribute mostly to Salton's measure – between ***Egghe*** and ***Glänzel***. Among these terms we find entries obviously referring to mathematical concepts such as 'finite' ($\sigma_j = 0.95594$), 'discrete', 'cumulative', 'linear', 'modelling' ($\sigma_j = 0.64261$), 'formula' ($\sigma_j = 0.45791$), 'curve' and 'inverse' ($\sigma_j = 0.33615$). But less obvious contributions occur such as 'brief overview' ($\sigma_j = 0.95594$), 'stress' ($\sigma_j = 0.64261$) or 'special case' ($\sigma_j = 0.45791$) which are due to the limited sample of our study and the nature of the keyword-based approach. With these restrictions in mind, it is left to a bibliometric component to add characteristics that might describe the position of papers belonging to the same class in the set of all selected documents.

As introduced by GLÄNZEL & SCHOEPFLIN (1999) and applied by SCHOEPFLIN & GLÄNZEL (2001) to selected volumes of the journal *Scientometrics* published in the period 1980–1997, 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. These indicators have been constructed for individual papers and the authors have shown that while the contribution of sub-disciplines in scientometrics was still well-balanced in 1980, papers dealing with case studies and methodology became predominant by 1997. In the following we will check whether these indicators can be used to characterise the classes defined by the guest-editors and/or the clusters found on the basis of the statistical full-text analysis. Figure 4 presents the distribution of Reference age for all publications grouped by the classes defined in the special issue. Three papers in class II have extremely low mean reference age with low standard deviation. The paper by ***Wu et al.*** is the only exception with a somewhat higher age of references. This is followed by class IV. The explanation for that is readily found: the sub-specialty is extremely young and there is practically no much relevant literature older than a couple of years to be referred to. Class V is characterised by high reference age; alone the ***Glänzel*** paper might be considered a slight outlier. As expected, the indicator values of the other two classes (I and III) range between those of class II and V.

If we aggregate the data over classes we obtain interesting results concerning the age of references in sub-domains defined by the editors. This is depicted in Figure 5. Policy relevant issues were not older than about three years on the average, followed by webometrics with roughly four years. The sub-domain with slowest ageing was mathematical models in bibliometrics with a mean reference age of about ten years.

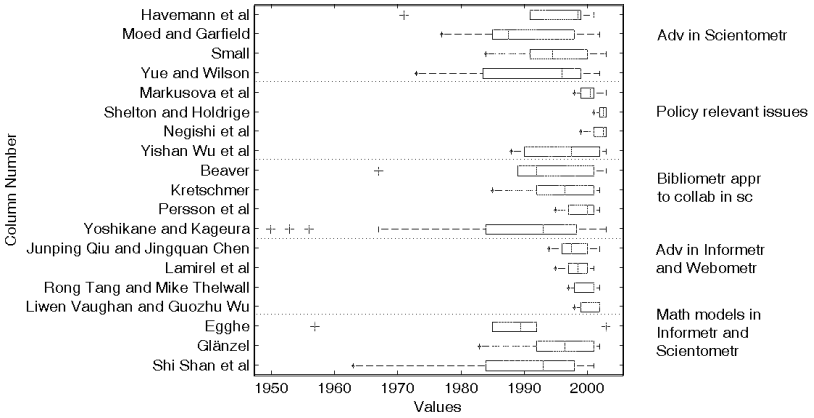

Figure 4. Distribution of reference age for all publications grouped by thematic groups

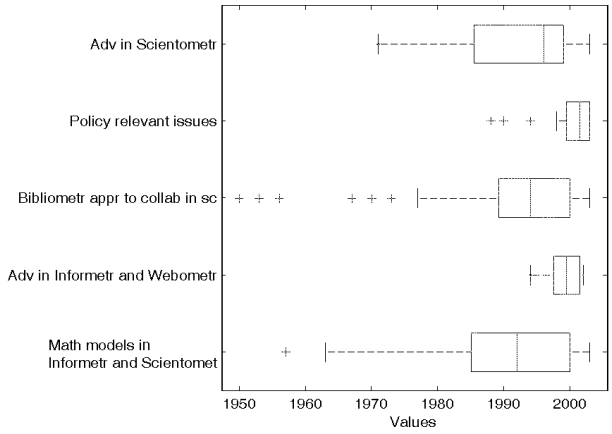

Figure 5. Histograms of reference age (pooled over the publications in each group)

The scatter plot of Share of Serials versus Mean Reference Age presented in Figure 6, finally, separates all defined classes (see also GLÄNZEL & SCHOEPFLIN, 1999). Theoretical, methodological papers and indicator engineering have a significantly higher share of serials than science policy issues. Probably the sociological component is responsible for the share of serials of class III that is significantly lower that that of class I (cf. SCHOEPFLIN & GLÄNZEL, 2001).

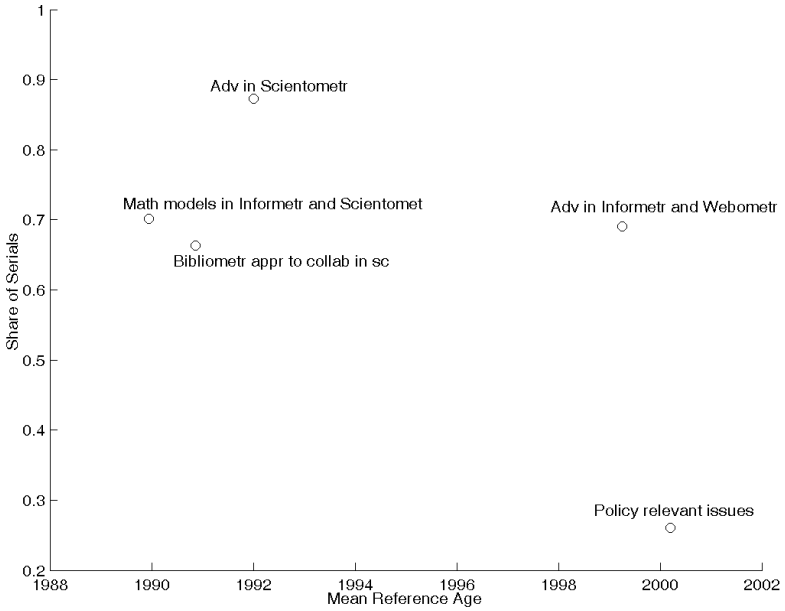

Figure 6. Scatter plot of Share of Serials versus Mean Reference Age for the identified sub-domains

Discussion

Although the idea that mere keywords can characterize the complexity of human discourse seems to be somewhat naïve, we observe that documents can be reasonably well categorised by algorithms that rely on this very assumption. Evidently, keywords are of also great use to retrieve information but, although the field of IR has a long history standing, often experts rely on extensive search strings to browse bibliographic databases. We strengthen in this study earlier text mining results that weighted text profiles provide good descriptions of the particular topic of each individual paper. The challenge to scale these desirable properties up to more massive data sets, however, remains.

The fact that 21% of all papers would not be correctly assigned if the analysis were restricted to titles and abstracts only suggests that whenever possible a full-text analysis should be preferred. However, if the informativeness of full-text information is to hold over those of abstracts in our approach, we foresee that the overall document structure (sections, subsections, paragraphs) should be increasingly exploited. Indeed, we already

obtained several weak indications of relevant keywords 'drowning' in methodological terminology. As mentioned above we can draw upon previous work to cope with such issues.

Both the projection onto 2D space based on multidimensional scaling and the unsupervised clustering procedure nicely show the ongoing polarisation in the field as well as the literature interlinking the poles. The topic Collaboration in Science proved an interdisciplinary sub-domain within the field of bibliometrics. The bibliometric component added by the application of an analysis of the reference lists of the papers gave some more precision to the structure obtained through the full-text analysis. It is difficult to say whether the bibliometric analysis should supplement the text analysis or vice versa. In any case, the combination of the two methods might essentially improve the interpretability of outcomes. A further interesting question arose, namely if the sub-domain webometrics will drift away from bibliometrics and informetrics, and – what has not yet been analysed in this pilot study – where quantitative technology studies appear on the charts and how the landscape will change in time.

Of course, the extent to which the selection of papers published in the special *Scientometrics* issue can be considered representative is unknown.

Although the size of the corpus was rather limited, trustable statistical results could still be obtained. And, more importantly, it allowed for an in-depth analysis – at the document-level – of both text mining and bibliometric results. In the future, we therefore plan extending this exercise to more challenging corpora spanning multiple issues over various years.

As a conclusion, we believe that *hybrid methodologies* combining data-mining techniques and bibliometric methods such as proposed above are valuable tools to facilitate endeavours in mapping fields of science. Moreover, these techniques might also be used to improve the otherwise difficult subject delineation in interdisciplinary research fields thus opening new perspectives in the evaluation of research, too.

References

BAEZA-YATES, R., RIBEIRO-NETO, B. (1999), *Modern Information Retrieval*. ACM Press.

BEAVER, D. D. (2004), Does collaborative research have greater epistemic authority?, *Scientometrics,* 60 (3) : 399–408.

BRAAM, R., MOED, H., VAN RAAN, A. (1991a), Mapping of science by combined co-citation and word analysis. 1. Structural aspects. *Journal of the American Society for Information Science*, 42 (4) : 233–251.

BRAAM, R., MOED, H., VAN RAAN, A. (1991b), 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.

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.

EGGHE, L. (2004), Positive reinforcement and 3-dimensional informetrics, *Scientometrics,* 60 (3) : 497–509.

GLÄNZEL, W., SCHOEPFLIN, U. (1999), A bibliometric study of reference literature in the sciences and social sciences, *Information Processing & Management*, 35 (1) : 31–44.

GLÄNZEL, W. (2004), Towards a model for diachronous and synchronous citation analyses, *Scientometrics,* 60 (3) : 511–522.

GLÄNZEL, W., JIANG, G. H., ROUSSEAU, R., WU, Y. S. (2004), Preface, *Scientometrics*, 60 (3) : 281–282.

GLENISSON, P., ANTAL, P., MATHYS, J., MOREAU, Y., DE MOOR, B. (2003a), Evaluation of the vector space representation for text-based gene clustering, In: *Proceedings of the Eighth Annual Pacific Symposium on Biocomputing 8 (PSB 2003)*, pp. 240–251.

GLENISSON, P., MATHIJS, J., MOREAU, Y., DE MOOR, B. (2003b), Meta-clustering of gene expression data and literature-extracted information, *SIGKDD Explorations,* Special Issue on Microarray Data Mining, 5 (2) : 101–112.

HAVEMANN, F., HEINZ, M., WAGNER-DÖBLER, R. (2004), Growth dynamics of German university enrolments and of scientific disciplines in the 19th century, Scaling behaviour under weak competitive pressure, *Scientometrics,* 60 (3) : 283–294.

JAIN, A., DUBES, R. (1988), *Algorithms for Clustering Data*. Prentice Hall.

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.

KOSTOFF, R. N. (2003), The practice and malpractice of stemming, *Journal of the American Society for Information Science and Technology,* 54 (10) : 984–985.

KOSTOFF, R. N., BLOCK, J. A., STUMP, J. A., PFEIL, K. M. (2004), Information content in Medline record fields, *International Journal of Medical Information,* 73 (6) : 515–527.

LARSEN, B., INGWERSEN, P. (2002), *The Boomerang Effect: Retrieving Scientific Documents Via the Network of References and Citations*. Paper presented at *SIGIR'0*2, August 11-15, Tampere, Finland.

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.

LOSEE, R. M, (1996), Text windows and phrases differing by discipline, location in document, and syntactic structure, *Information Processing & Management*, 32 (6) : 747–767.

MANNING, C. D., SCHUTZE, H. (2000), *Foundations of Statistical Natural Language Processing*. MIT Press.

MARKUSOVA, V. A., MININ, V. A., LIBKIND, A. N., JANSZ, C. N. M., ZITT, M., BASSECOULARD-ZITT, E. (2004), Research in non-metropolitan universities as a new stage of science development in Russia, *Scientometrics,* 60 (3) : 365–383.

MOED, H. F., GARFIELD, E. (2004), In basic science the percentage of 'authoritative' references decreases as bibliographies become shorter*, Scientometrics,* 60 (3) : 295–303.

MULLINS, N., SNIZEK, W., OEHLER, K. (1988), The structural analysis of a scientific paper. In: VAN RAAN, A. F. J. (Ed.), *Handbook of Quantitative Studies of Science and Technology*. New York: Elsevier Science, pp. 81–105.

NEGISHI, M., SUN, Y., SHIGI, K. (2004), Citation database for Japanese papers, A new bibliometric tool for Japanese academic society, *Scientometrics,* 60 (3) : 333–351.

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.

PERSSON, O., GLÄNZEL, W., DANELL, R. (2004), Inflationary bibliometric values, The role of scientific collaboration and the need for relative indicators in evaluative studies*, Scientometrics,* 60 (3) : 421–432.

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

QIU, J. P., CHEN, J. Q., WANG, Z. (2004), An analysis of backlink counts and Web Impact Factors for Chinese university websites, *Scientometrics,* 60 (3) : 463–473.

SCHOEPFLIN, U., GLÄNZEL, W. (2001), Two decades of Scientometrics – An interdisciplinary field represented by its leading journal, *Scientometrics*, 50 (2) : 301–312.

SHAN, S., JIANG, G. H., JIANG, L. (2004), The multivariate Waring distribution and its application, *Scientometrics,* 60 (3) : 523–535.

SHELTON, R. D., HOLDRIDGE, G. M. (2004), The US-EU race for leadership of science and technology, Qualitative and quantitative indicators, *Scientometrics,* 60 (3) : 353–363.

SMALL, H. (2004), Why authors think their papers are highly cited, *Scientometrics,* 60 (3) : 305–316.
SNIZEK, W., OEHLER, K., MULLINS, N. (1991), Textual and nontextual characteristics of scientific papers: Neglected science indicators. *Scientometrics*, 20 (1) : 25–35.
TANG, R., THELWALL, M. (2004), Patterns of national and international Web inlinks to US academic departments, An analysis of disciplinary variations, *Scientometrics,* 60 (3) : 475–485.
VAUGHAN, L. W., WU, G. Z. (2004), Links to commercial websites as a source of business information, *Scientometrics,* 60 (3) : 487–496.
WU, Y. S., PAN, Y. T., ZHANG, Y. H., MA, Z., PANG, J. G., GUO, H., XU, B., YANG, Z. Q. (2004), China Scientific and Technical Papers and Citations (CSTPC), History, impact and outlook, *Scientometrics,* 60 (3) : 385–397.
YOSHIKANE, F., KAGEURA, K. (2004), Comparative analysis of coauthorship networks of different domains, The growth and change of networks, *Scientometrics,* 60 (3) : 433–444.
YUE, W. P., WILSON, C. S. (2004), Measuring the citation impact of research journals in clinical neurology, A structural equation modelling analysis, *Scientometrics,* 60 (3) : 317–332.
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.

Appendix
Papers published in *Scientometrics*, 60 (3) (2004)
(in alphabetical order of the first authors)

BEAVER, D. D. (2004), Does collaborative research have greater epistemic authority?, *Scientometrics,* 60 (3) : 399–408.
EGGHE, L. (2004), Positive reinforcement and 3-dimensional informetrics, *Scientometrics,* 60 (3) : 497–509.
GLÄNZEL, W. (2004), Towards a model for diachronous and synchronous citation analyses, *Scientometrics,* 60 (3) : 511–522.
GLÄNZEL, W., JIANG, G. H., ROUSSEAU, R., WU, Y. S. (2004), Preface, *Scientometrics*, 60 (3) : 281–282
HAVEMANN, F., HEINZ, M., WAGNER-DÖBLER, R. (2004), Growth dynamics of German university enrolments and of scientific disciplines in the 19th century, Scaling behaviour under weak competitive pressure, *Scientometrics,* 60 (3) : 283–294.
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.
MARKUSOVA, V. A., MININ, V. A., LIBKIND, A. N., JANSZ, C. N. M., ZITT, M., BASSECOULARD-ZITT, E. (2004), Research in non-metropolitan universities as a new stage of science development in Russia, *Scientometrics,* 60 (3) : 365–383.
MOED, H. F., GARFIELD, E. (2004), In basic science the percentage of 'authoritative' references decreases as bibliographies become shorter, *Scientometrics,* 60 (3) : 295–303.
NEGISHI, M., SUN, Y., SHIGI, K. (2004), Citation database for Japanese papers, A new bibliometric tool for Japanese academic society, *Scientometrics,* 60 (3) : 333–351.
PERSSON, O., GLÄNZEL, W., DANELL, R. (2004), Inflationary bibliometric values, The role of scientific collaboration and the need for relative indicators in evaluative studies, *Scientometrics,* 60 (3) : 421–432.
QIU, J. P., CHEN, J. Q., WANG, Z. (2004), An analysis of backlink counts and Web Impact Factors for Chinese university websites, *Scientometrics,* 60 (3) : 463–473.

SHAN, S., JIANG, G. H., JIANG, L. (2004), The multivariate Waring distribution and its application, *Scientometrics,* 60 (3) : 523–535.

SHELTON, R. D., HOLDRIDGE, G. M. (2004), The US-EU race for leadership of science and technology, Qualitative and quantitative indicators, *Scientometrics,* 60 (3) : 353–363.

SMALL, H. (2004), Why authors think their papers are highly cited, *Scientometrics,* 60 (3) : 305–316.

TANG, R., THELWALL, M. (2004), Patterns of national and international Web inlinks to US academic departments, An analysis of disciplinary variations, *Scientometrics,* 60 (3) : 475–485.

VAUGHAN, L. W., WU, G. Z. (2004), Links to commercial websites as a source of business information, *Scientometrics,* 60 (3) : 487–496.

WU, Y. S., PAN, Y. T., ZHANG, Y. H., MA, Z., PANG, J. G., GUO, H., XU, B., YANG, Z. Q. (2004), China Scientific and Technical Papers and Citations (CSTPC), History, impact and outlook, *Scientometrics,* 60 (3) : 385–397.

YOSHIKANE, F., KAGEURA, K. (2004), Comparative analysis of coauthorship networks of different domains, The growth and change of networks, *Scientometrics,* 60 (3) : 433–444.

YUE, W. P., WILSON, C. S. (2004), Measuring the citation impact of research journals in clinical neurology, A structural equation modelling analysis, *Scientometrics,* 60 (3) : 317–332.