Hybrid solutions – the best of all possible worlds?

Glänzel, Wolfgang[1,2] und Thijs, Bart[1]

[1] Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven, Leuven, Belgium

[2] Dept. Science Policy & Scientometrics, LHAS, Budapest, Hungary

Abstract

Recently the combination of text- and citation-based methods is increasingly applied to structural analysis since they compensate several severe shortcomings of their individual components while the combination takes full advantage of their individual strengths. In the present study we will focus on three major application fields, particularly, (1) document clustering, (2) topic representation and (3) subject delineation in order to discuss the real strength of this paradigmatic approach in a coherent system of the structural analysis of document spaces for the application to scientific information, scientometrics and research evaluation.

1. Introduction

The wide availability of electronic versions of abstract and full-text databases and the tremendously fast development of information technology has recently facilitated and broadened the application possibilities of textual data mining and analysis on a large scale. The time when simple combinations of subject headings, keywords or title terms where used to describe the relationship of underlying documents has long passed. Thus text mining has become an important tool in bibliometrics be it as independent instrument or as an additional component in combination with citation-based methods. In the present study opportunities and limitations of text- and link-based methods are discussed and recent results in the application of their combination are presented. In this context we will focus on three major fields of applications, namely

1. the peculiarities of global and local clustering of document spaces,
2. bibliometrics-aided retrieval for the delineation of subjects and topics, and
3. the use of core-documents for the representation of clusters, topics and disciplines

Despite their advantages, scientometric text- and link-based techniques have specific properties that are inherited from their origins and that results in characteristic drawbacks. Among those, the general issues are discussed in this introduction since they have effect on all three mentioned scopes. The more specific issues will be discussed in separate sections.

1.1. The link approach

In scientometrics, the link approach generally refers to all relations that are directly or indirectly created by citations. One distinguishes between direct links that are whether directed (as citations to and references from a document, respectively), or cross-citations among document sets, if the direction of information flow does not matter, and indirect links, which are established by the extent of information that is jointly used by two different documents manifested by whether references (called bibliographic coupling, cf. Fano, 1956, Kessler, 1963, Glänzel and Czerwon, 1996) or citations (called co-citations, cf. Small, 1973, Marshakova, 1973). Since citations directly point to documents that are considered relevant by an author for the own research, link-based constructs tend to be strongly discriminative and to indicate rather false negative then false positive relationship. At the same time, citations are selective as they also express affinity to a certain community or even a particular "school". This property is inherited even if indirect links are used. Identified relationship based on this approach is thus reliable but often enough not all relations are recognised. These are two sides of the same coin, one of which could be considered a great advantage, the other one a severe shortcoming.

1.2. The lexical approach

The application of textual components goes back to the use of combinations of title words (Garfield, 1969), keywords (e.g., Turner et al., 1988), or subject headings (Todorov, 1992). Initially, this resulted in binary relations, i.e., merely based on presents or absence of terms and their co-occurrence. Characteristics of text-based methods were similar to those of the link approach, the results were, however, often different. Already at that time a combination of these two types of methods was proposed to overcome this discrepancy. Braam et al. (1990) and Zitt and Bassecoulard (1994) were among the firsts advocating such combination. As a result of the IT revolution storage capacity as well as CPU and processing speed dramatically increased. Recording, storing and processing abstracts and even full texts of documents has become feasible with the effect of having to cope with a flood of textual information. Text-mining techniques, e.g., database tomography (Kostoff et al., 1997) made it possible to extract and analyse terms and phrases from large computerised corpora. Not only the mere occurrence of terms and phrases but also their frequency in the document as well as that of the documents containing these text components could be taken into account. However, scientific text is based on natural language and within the same (larger) subject the same or similar vocabulary and paradigms are in use. As a consequence, text-based similarities proved to be less discriminative (as compared with their link-based counterparts), result in less unrecognised relationships but produce more false positives than citation links. Therefore so-called hybrid solutions, that is, the combination of text- and citation-based similarities are increasingly applied to compensate their individual shortcomings and, at the same time, to synergise their advantages to make the best of the two worlds. The superiority of such hybrid methods over text-based and link-based approaches has recently been shown, among others, by Glenisson et al. (2005) and Boyack and Klavans (2010).

2. Methodology of hybrid techniques

As mentioned in the previous section, citation and textual similarities are based on different structures. This complicates the direct combination of distances or similarity measures based on those. In order to obtain a proper representation of the hybrid space, two simple but efficient methods are available, one each for the Vector-Space Model (VCM) and the graph model (GM). Document spaces can usually be represented by vector spaces or graphs. In the first case documents are represented by vectors, while the angle between these vectors determines their similarity. In particular, parallel vectors have maximum similarity while perpendicular vectors have no similarity at all. In the second case, documents are represented by vertices and those edges. The weight of the edges connecting documents expresses their similarity, if two vertices are not connected, the corresponding edge has weight zero. In the framework of the graph model, a combination of two different (e.g., textual and citation-based) similarities can readily be obtained by graph integration or graph coupling (cf. Liu et al., 2011). However, if a method is needed that also controls for the weight of the components, this can advantageously be developed in the VSM. The easiest way of combining the two worlds is, in fact, the *convex combination of the underlying angles*. One obtains a hybrid (cosine) similarity measure r as the cosine of the convex combination of the underlying angles, i.e.,

$$r = \cos\left(\lambda \cdot \operatorname{arccos}(\eta) + (1-\lambda) \cdot \operatorname{arccos}(\xi)\right), \quad \lambda \in [0,1],$$

where η is the cosine similarity defined on bibliographic coupling and ξ the corresponding textual similarity, while $\operatorname{arccos}(\eta)$ and $\operatorname{arccos}(\xi)$ denote the two underlying angles. The λ parameter is used to adjust the weight of the components (Glänzel and Thijs, 2012).

This method has two crucial advantages. One arises from the necessity of fine-tuning of the two components' weight in their combination if the method is to be applied to various fields in the sciences but also in the social sciences and humanities, where citations plays a less important role as in the science. In general, this approach makes it possible to include documents whenever citation links are weak or even missing. The second advantage results from the property that extracted terms can, beyond their role in calculating similarities, be used for labelling and describing the analysed structures.

One of the most popular applications of hybrid similarities is the clustering of document spaces. The objectives of clustering exercises vary from a representative mapping of the epistemological structure of science, subject classification issues to the sophisticated detection and monitoring of the emerging topics and disciplines. In the following section we will shed some light on different facets of document clustering and specific requirements to be met by the hybrid method.

3. Global and local clustering of document spaces – what is different?

The issue of the common vocabulary in larger fields has already been mentioned in the first section. Hence the question arises of what effects on the applicability of textual similarities are produced by the language use at different levels of aggregation. While in global clustering, that is, in the clustering of a complete database or a larger parts of it, the textual component can still be used to label the obtained clusters, in local clustering, such as the analysis of rather narrow disciplines, the common vocabulary might distort this procedure. At lower levels of aggregation, terms and phrases might become less specific since they express a common knowledge base and vocabulary. Others might gain more ‘‘information value’’. This has been explained by Glänzel and Thijs (2011) using the example of the terms ‘algebra’ and ‘group’, which have different significance in the global and local environment. This immediately results in the necessity of adjusting the weight of the two components in the hybrid similarity measures if those are applied at different levels. Hybrid techniques are long used in the structural analysis of document spaces and their applicability to large and medium-large sets has been proven. Nevertheless, clustering of very small coherent or less coherent sets is possible as well, if the combination of the two components is balanced enough to compensate for both the peculiarities of the common vocabulary and the missing citation links. Its practicability has been shown by recent studies of Hirsch-related literature (Lin et al., 2012) or ‘entrepreneurship research’ (Meyer et al., 2012).

A second consequence is that the use of terms to label the obtained clusters will no longer properly work at the topic level. Labelling using TF-IDF terms still works at the “medium” level, e.g., for the discipline “information science’ (Janssens, 2007). The term network visualised in Figure 1 according to mean TF-IDF weights still truly reflects the content of the obtained clusters. However, at the level of research topics such as *biofuels* within the subject category ‘energy and fuels’, *brain-computer interface (BCI)* within ‘biomedical engineering’, *prenatal diagnosis* within ‘obstetrics and gynaecology’ or *state & region* within ‘geography’ the borderline for the applicability of terms for the contentual description of the topics is reached, as has shown in recent papers by the authors (e.g., Glänzel and Thijs, 2011; Glänzel and Thijs, 2012). Here we mention the four most frequent terms (‘signals’, ‘blood flow’, ‘classification’, ‘EEG’) from the BCI cluster just as an example. At this level, the most important TF-IDF terms are thus not specific and discriminative enough to provide a sufficiently precise groundwork for the description of the content of the topic in question as most of those rather reflect the general context and textual environment shared with other related topics. This effect, which is typical of local clustering in general, leads us to the second important application of hybrid similarities. This will be discussed in the following section.

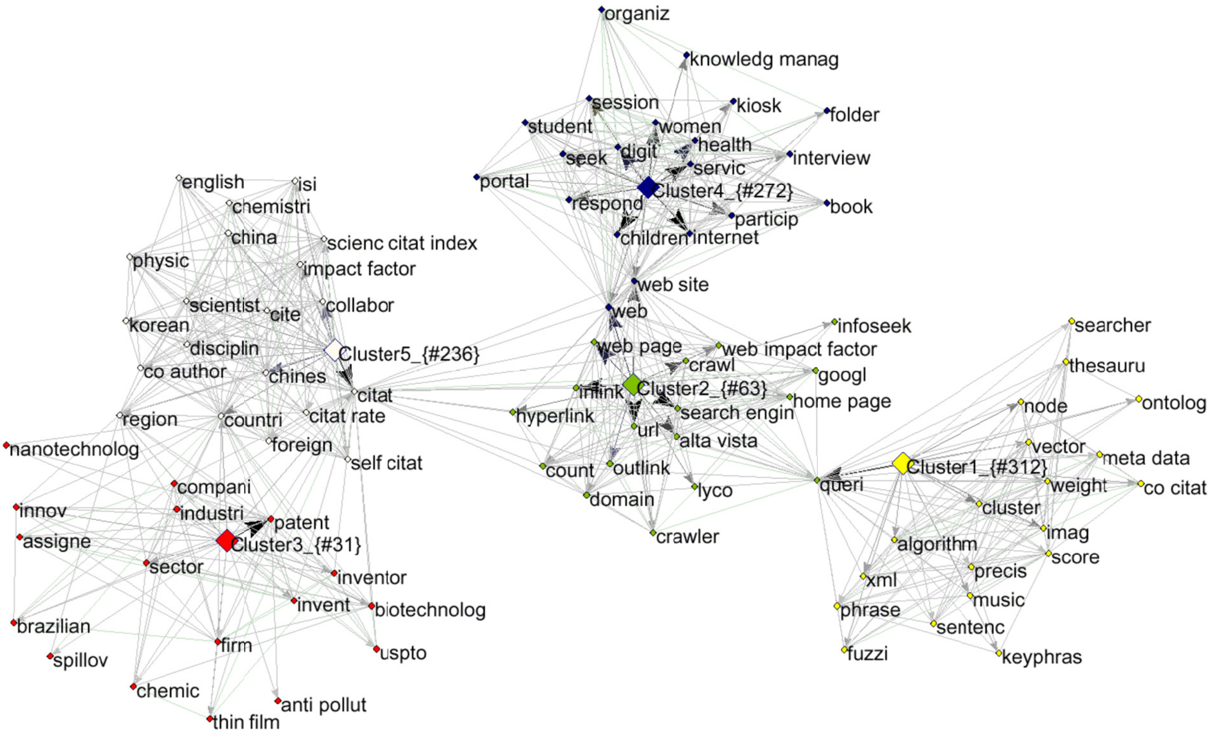

Figure 1. Term networks with for each of five clusters the best 20 stemmed terms or phrases from titles and abstracts according to mean TF-IDF scores according to Janssens et al. (2008)

4. Core-documents for the representation of clusters, topics and disciplines

The problem of an adequate cluster representation in narrower disciplines was already tackled in the context of clustering bibliometrics and bioinformatics. One possibility is the use of medoids, which can easily be determined from the clustering exercise (Janssens, 2007). Medoids provide, strictly speaking, the most typical documents in a cluster, the most central vertices in a network. However, bibliometric networks are rather decentralised and tend to embrace more or less strongly interlinked substructures with their own local central nodes. That is why we reconsidered a method that was introduced somewhat more than 15 years ago and goes back to the notion of co-citations and bibliographic coupling proposed even some decades earlier. The idea of identifying the 'core' of literature was first proposed in the context of co-citation analysis (Small, 1973), where documents belonging to such a cluster, by definition, formed a set of considerably cited papers. The term 'core documents' was anew introduced by Glänzel and Czerwon (1996) in the context of bibliographic coupling to identify those papers which form important nodes in the network of scholarly communication. The notion of core documents is less "centralistic" than the medoid approach, although they also proved to express (local) centrality (Zhang et al., 2009), and, as we will see, it is completely independent of any clustering. In fact, the above notion can readily be extended to hybrid techniques, where core documents can, of course, be used in the context of document clustering, as well, to represent networks and their internal structures. First we provide their definition: 'Core documents' are documents that have at least $n > 0$ links of at least a given strength $r \in (0, 1)$ according to the predefined similarity measure. The determination of the two parameters n and r is practically based on experience. Both parameters should be chosen so that core documents represent the order of magnitude of 1% of the total. At lower levels of aggregation, such as local clustering, the h-index of the underlying network (cf. Schubert et al., 2009) can be used instead of an arbitrarily predefined value n. Glänzel (2012a) has given empirical evidence that this choice results in an adequate document representation that is by two orders of magnitude lower that the total.

Core documents are, by definition, strongly linked with many other documents and thus represent the most interconnected part of the network. Figure 2 shows the environment of a (randomly) selected core document; the core document is distinguished from the other publications though its size.

Examples for the application of core documents for the representation of clusters and topics and for detecting and labelling new emerging topics can be found in recent studies by Glänzel and Thijs (2011 and 2012, respectively).

Apart from the representation of clusters and networks, the strong interconnectivity of core documents implies an interesting by-effect that has already been pointed to in the original study on this issue (Glänzel and Czerwon, 1996). The degree distribution in a scale-free network approximately follows a power law (cf. Newman, 2003). As a consequence, the most connected 1% of the high-degree nodes are, depending on the distribution's parameter, responsible for 10% to 25% of all degrees in the network (cf. Glänzel, 2012b). Thus following the large number of their links, including those of the documents connected with them, already covers a considerable part of related publications. This property can be useful for the retrieval of information and for the delineation of subjects, and thus directly leads to the third issue, which will be discussed in Section 5.

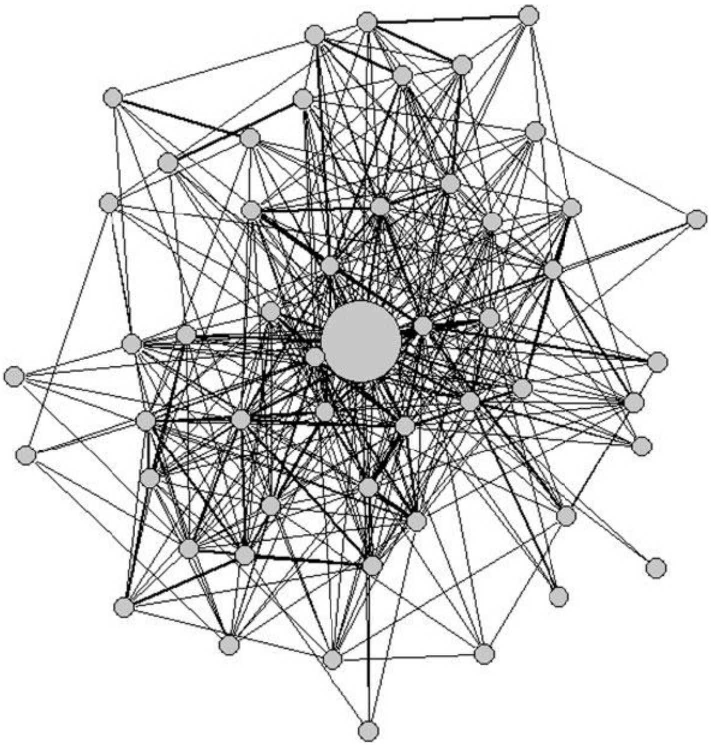

Figure 2. Visualisation of the link environment of a 'core document' according to Glänzel and Thijs (2012)

5. Bibliometrics-aided retrieval and subject delineation

The third issue refers to the extension of traditional information retrieval by means of hybrid techniques (cf., Zitt and Bassecoulard, 2006; Glänzel et al., 2006). The methods introduced by these authors are quite general. They have in common that the strategy for retrieval or subject delineation proceeds from an initial document set called 'seed' of literature (Zitt and Bassecoulard) or 'core literature' (Glänzel et al.) that covers at least a certain part the subject in question well and truly. Thus the basic idea of the strategy is the use of two parts, the first of which is assumed to result in an *incomplete* but truly *relevant* set of documents. This first part is mostly based on traditional retrieval or delineation, for instant, based on core journals (such as *JASIST* for information science, or *Scientometrics* for bibliometrics) and/or lexical queries. The second part then aims at extending this set by potentially relevant documents on the basis of so-called conditional criteria, for instance, papers published in related fields or in non-core journals. In order to define a valid strategy and to increase the probability of the relevance of the additionally retrieved documents, further conditions defined on bibliometric properties must be met, that is, only that part of the second group will be included that has, from the bibliometric viewpoint, close relations with the initial set. Thus the procedure proceeds from high-precision but low-recall set and supplements it by adding "purified" items from a low-precision and high-recall sets. The result is a considerable increase of both precision *and* recall. In verbal terms, the final document set is built around a truly relevant seed or core by adding further documents on the basis of thematic similarity. Figure 3 presents a schematic overview of this procedure.

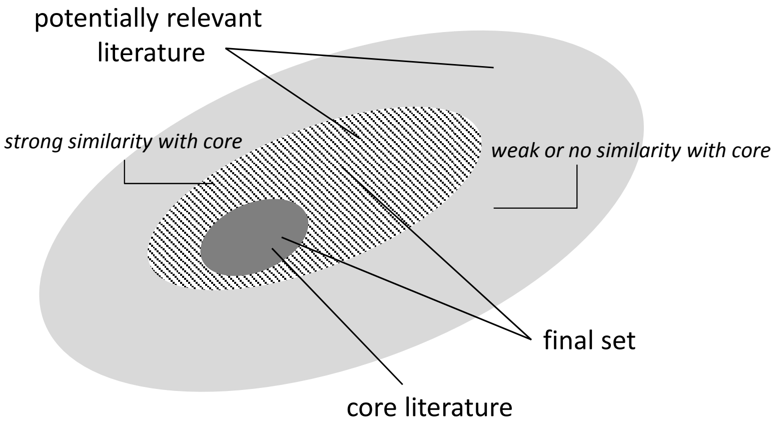
potentially relevant literature
strong similarity with core
weak or no similarity with core
final set
core literature

Figure 3. Schematic overview of bibliometrics aided retrieval and subject delineation

Possible methods of extending core sets (or seed literature) are the use of direct citation links (e.g., Zitt and Bassecoulard, 2006; Glänzel et al., 2009) or of bibliographic coupling. We just mention in passing, that a simple version of the latter method was already offered by the *Institute for Scientific Information* (ISI – now part of Thomson Reuter) in the CD Edition of the Science Citation Index (SCI). In particular, for each retrieved document the user could choose among so-called related records, which were identified on the basis of references shared with the retrieved document. However, particular similarity measures or thresholds for the number of shared references were not given at that time.

In the case of subject delineation within clusters, for instance, obtained from a hybrid clustering as described in the third section, the extension of 'seed' or 'core' sets by using the same hybrid similarities, as has been used for the clustering, presents itself as being most obvious. The complete cluster can be regarded as a "natural" potentially relevant set. It should be stressed that the search strategy defined on a hybrid retrieval algorithm is, as such, independent of any clustering. A potentially relevant set can always be retrieved based on simple search strategies using journals, corporate addresses and lexical queries. Threshold for hybrid similarities can then be used to fine-tune delineation in order to optimise filtering of truly relevant items. Further elaboration and the practical application of this part is, however, left to future research.

6. Conclusions

The fields of applications of the described hybrid methods are manifold. They help improve the delineation of complex and interdisciplinary fields and topics, allow clustering at the local level, even in fields where citations do not play an important role, and detecting and labelling new emerging topics.

The question of whether hybrid solutions stand for the best of all possible worlds can, of course, not be answered but they stand for a versatile state-of-the-art approach that provides a bundle of techniques using the same algorithms for solving complex tasks. A methodologically consequent implementation of combining textual and link components makes the best of the two worlds at least. Consequently, this reduces the arbitrariness and danger of drawing on different, possibly not always compatible sources.

Acknowledgement

Figures 1–2 are reproduced with permission of the journal Scientometrics.

References

Boyack, K.W., Klavans, R. (2010), Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *JASIST*, 61(12), 2389–2404.

Fano, R.M. (1956), *Information theory and the retrieval of recorded information*. In J. H. Shera, A. Kent, J.W. Perry (Eds.), Documentation in action. New York: Reinhold Publ. Co., pp. 238–244.

Garfield, E. (1969), Permuterm Subject Index – the primordial dictionary of science. *Current Contents*, 12(22), 4.

Glänzel, W., Czerwon, H. J. (1996), A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37(2), 195–221.

Glänzel, W., Janssens, F., Speybroeck, S., Schubert, A., Thijs, B., (2006), *Towards a Bibliometrics-Aided Data retrieval for scientometric purposes*. Poster presented at the 9th International Conference on Science and Technology Indicators, Leuven, Belgium. Book of Abstracts, 206–208.

Glänzel, W., Janssens, F., Thijs, B. (2009), A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics. *Scientometrics*, 79(1), 109–129.

Glänzel, W., Thijs, B. (2011), Using 'core documents' for the representation of clusters and topics. *Scientometrics*, 88(1), 297–309.

Glänzel, W., Thijs, B. (2012), Using 'core documents' for detecting and labeling new emerging topics. *Scientometrics*, 91(2), 399–416.

Glänzel, W. (2012a), The role of core documents in bibliometric network analysis and their relation with h-type indices. *Scientometrics*, 93 (1), 113-123.

Glänzel, W. (2012b), *High-End performance or outlier? Evaluating the tail of scientometric distributions*. H.N. Choi, H.S. Kim, K.R. Noh, S.H. Lee, H.J. Kang, H. Kretschmer (Eds), Proceedings of the 8th International Conference on Webometrics, Informetrics and Scientometrics (WIS) & 13th COLLNET Meeting, Seoul, Korea, October 23–26, 2012, KISTI.

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

Janssens, F. (2007), *Clustering of scientific fields by integrating text mining and bibliometrics*. Doctoral dissertation. Faculty of Engineering, Katholieke Universiteit Leuven, Belgium.

Kostoff, R.N., Eberhart, H.J., Toothman, D.R. (1997), Database tomography for information retrieval. *Journal of Information Science*, 23(4), 301–311.

Kessler, M.M. (1963), Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25.

Liu, X.H., Glänzel, W., De Moor, B. (2012), Optimal and hierarchical clustering of large-scale hybrid networks for scientific mapping. *Scientometrics*, 91(2), 473–493.

Marshakova, I.V. (1973), System of connections between documents based on references (as the science citation index). *Nauchno-Tekhnicheskaya Informatsiya Seriya*, 2(6), 3–8.

Newman, M.E.J. (2003), The structure and function of complex networks. *Siam Review*, 45(2), 167–256.

Schubert, A., Korn, A., Telcs, A. (2009), Hirsch-type indices for characterizing networks. *Scientometrics*, 78(2), 375–382.

Small, H. (1973), Co-citation in the scientific literature: A new measure of the relationship between two documents. *JASIS*, 24(4), 265–269.

Todorov, R. (1992), Displaying content of scientific journals: A co-heading analysis. *Scientometrics*, 23(2), 319–334.

Turner W.A., Chartron G., Laville F., Michelet M. (1988), *Packaging information for peer review: new co-word analysis techniques*. In: A. van Raan (Ed.) Handbook of Quantitative Studies of Science and Technology, Elsevier, North Holland.

Zhang, L., Glänzel, W., Liang, L. (2009). Tracing the role of individual journals in a cross-citation network based on different indicators. *Scientometrics*, 81(3), 821–838.

Zitt, M., Bassecoulard, E. (2006), Delineating complex scientific fields by hybrid lexical-citation method: an application to nanoscience. *Information Processing & Management*, 42(6), 1513–1531.