Scientometrics (2015) 102:2215–2222
DOI 10.1007/s11192-014-1480-7

Bibliometrics-aided retrieval: where information retrieval meets scientometrics

Wolfgang Glänzel

Received: 19 October 2014 / Published online: 25 November 2014
© Akadémiai Kiadó, Budapest, Hungary 2014

Abstract This paper attempts to sketch the interrelation between information retrieval and scientometrics pointing at possible synergy effects provided by some recently developed bibliometric methods in the context of subject delineation and clustering. Examples of specific search strategies based on both traditional retrieval techniques and bibliometric methods are used to illustrate this approach. Special attention is paid to hybrid techniques and the use of 'core documents'. The latter ones are defined merely on the basis of bibliometric similarities, but have by definition properties that make 'core documents' also interesting and attractive for information retrieval.

Keywords Bibliometrics-aided retrieval · Subject delineation · Complex search strategies · Hybrid techniques · Core documents

Introduction

At first sight, *Information Retrieval* (IR) and *scientometrics* seem, despite their common roots, to follow different paths in their evolution as sub-disciplines of information science. This might be considered a certain divergence in terms of both methodology and goals; this divergence is governed by the prevailing objectives bibliometrics has taken in the previous decades in favour of various services for science policy and research management in the context of the evaluation of research performance. This departure from the traditional goals of information science, described as "perspective shift" (e.g., Glänzel 2006), entailed an own, specific orientation in its methodology as well. The main focus was now laid on the

W. Glänzel (✉)
Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven, Belgium
e-mail: Wolfgang.Glanzel@kuleuven.be

W. Glänzel
Department of Science Policy and Scientometrics, Library of the Hungarian Academy of Sciences, Budapest, Hungary

development of indicators for evaluative purpose, and principles of data collection that are determined by the respective needs and standards of the evaluative context and benchmarking. As a consequence of this perspective shift, new fields of applications and challenges opened to scientometrics; but many tools were still designed for use in scientific information, in particular, in information retrieval and library science and services and became later used in a context for which these were not designed. The journal Impact Factor, originally designed as a measure for comparing journals independently of size and to help select journals for the *Science Citation Index* (Garfield and Sher 1963), might serve as the perhaps most popular example of this change. This might just illustrate such kind of "divergence" in the basic orientation of sub-disciplines of information science. Nevertheless the question arises of whether there is still some common platform with methodological components that could cross-pollinate each other and provide essential synergetic effects for the two sub-disciplines. In the following, this will be sought for using the example of bibliometrics-aided retrieval.

Information retrieval for bibliometrics?

Due to the dynamics in evaluative scientometrics, the focus has also shifted away from the formerly popular macro level towards meso and micro studies of both actors and topics. One consequence of this dynamics is the necessity of proper *subject delineation* that has, among others, become a central issue in so-called "domain studies", bibliometric studies of *interdisciplinary research* and the identification and analysis of *emerging topics*. In particular, science policy has addressed new emerging and complex interdisciplinary topics the delineation of which has proved particularly difficult. Appropriate subject delineation is also necessary to find correct reference standards for benchmarking the research performance of the actors in the topic under study.

Sufficiently fine-grained subject-classification schemes can help define a broader scope within which the actual subject can be delineated. However, using pre-set disciplines or topics usually results in noise that is too large for obtaining acceptable coverage with both high precision and recall. Even scientific journals are too coarse for subject delineation since the distribution of relevant documents over journals is usually very skewed (Bradford 1934).

Subject delineation, on the one hand, strongly relies on Information Retrieval methods through complex search strategies that are often composed of core journals and lexical terms such as keywords and phrases, but, as a consequence of the specific tasks in research assessment, goals and methods of advanced subject delineation essentially differ from those of traditional retrieval.

Bibliometrics for information retrieval?

Bibliometrics, in turn, provides important techniques to improve the power of Information Retrieval by incorporating bibliometric components. Similarity or distance measures defined on direct citations, bibliographic coupling, lexical relationship and 'core documents' (in the sense of the definition by Glänzel and Czerwon 1996) can facilitate and improve the retrieval of scientific information. Both Bibliometrics and Information Retrieval may thus serve as mutual input and can be combined in an iterative way. This combination will be discussed in the following.

Bibliometrics-aided retrieval

The initial situation

Bibliometrics, in general, requires specific retrieval. The borderline between relevant and not relevant documents is fuzzy and often determined by users or actors in the domain in question. Sometimes this borderline has to be adjusted according to the actual tasks. Figure 1 visualises this situation. The red circle stands for the truly relevant documents, the blue area marks not relevant documents and the purple zone represents that part of literature, which might be included in the bibliometric study depending on the actual needs. The scope of the study or report in question then decides whether documents in the red circle or which part of the purple area are to be used for the bibliometric analysis.

Bibliometrics also allows adding 'metric' components to the search strategy. In particular, thresholds of the strength of (co-)citation, bibliographic coupling or textual links can be used to fine-tune the metric component. And this can be done even during the retrieval process, quasi as "bibliometric feedback" on the results of the retrieval. This is visualised in Fig. 2. This combination of traditional search strategies with advanced bibliometric methods is called *bibliometrics-aided retrieval* (Glänzel et al. 2006).

When bibliometrics meets information retrieval …

Methodology of bibliometrics-aided retrieval is based on specific "search strategies" that have been developed and applied to domain studies, for instance, by Glänzel et al. (2004), Glänzel and Veugelers (2006) as well as by recent papers in bibliometrics (e.g., Noyons et al. 2003) and information science (e.g., Zitt and Bassecoulard 2006). Here we summarise and outline the general model combining components from lexical and bibliometric components (see Glänzel et al. 2006; Glänzel and Thijs 2012a).

The methods introduced in the above-mentioned studies 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 2006) or 'core' (Glänzel et al. 2006) that covers at least a certain part of 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 instance, 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 in the context of the respective bibliometric study 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 by adding further documents on the basis of thematic similarity.

The basic methodological idea of the search strategy is the use of two parts, which, in turn, include further components. The first part of the retrieval strategy, which is assumed to result in an incomplete set of relevant documents, comprises *unconditional criteria* (UC_1 … UC_k) with $k \geq 0$. 'Unconditional' here means in a rather formal way that no further

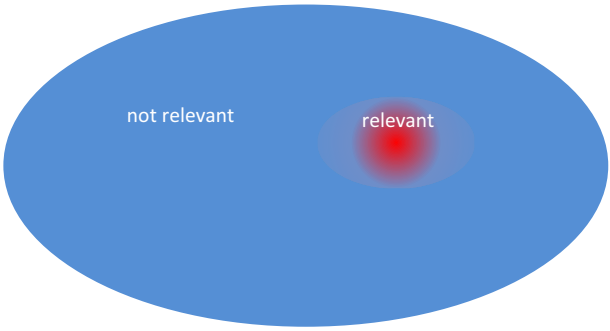

Fig. 1 Visualisation of the initial situation of "bibliometric" retrieval

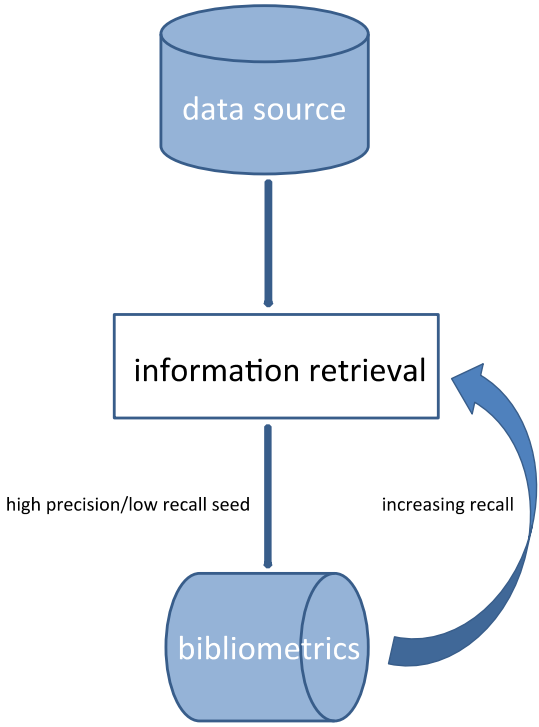

Fig. 2 Visualisation of the interrelation between IR and bibliometrics in bibliometrics-aided retrieval

conditions need to be attached to these criteria to obtain relevant hits. In particular, that means that (nearly) all results retrieved on the basis of these criteria are expected to be *relevant*. The precision of this strategy is therefore expected to be high while the recall is—because of the strict criteria—rather low.

The second part aims at extending this set by potentially relevant documents and includes so-called *conditional criteria* (CC_1 … CC_m … CC_{m+n}) with m, $n > 0$, or $m = n = 0$, respectively, in the trivial case that no conditional criteria are required. The term 'conditional' points to the fact that the application of these criteria results in high recall but low precision and further conditions are needed to obtain acceptable and useful results.

In order to define a valid strategy, the condition $k + m > 0$ must be met, that is, we have at least one unconditional criterion or two conditional criteria, if $k = 0$. Restrictions for the latter group obtained by logical combination among these criteria are to increase the probability of the relevance of the retrieved documents. The bibliometric retrieval (BR) can then be defined as the following logical combination

$$\mathrm{BR} = (\mathrm{UC}_1 \vee \ldots \vee \mathrm{UC}_k) \vee ((\mathrm{CC}_1 \vee \ldots \vee \mathrm{CC}_m) \wedge (\mathrm{CC}_{m+1} \vee \ldots \vee \mathrm{CC}_{m+n})).$$

Both the unconditional and the conditional part of the BR can include different components such as lexical terms and citation-based criteria. Note that each UC or CC itself can be defined as a logical combination of specific criteria, for instance, $\mathrm{UC}_i = \mathrm{C}_1\neg(\mathrm{C}_2 \vee \mathrm{C}_3)$ with C_1, …, C_3 being some criteria, where the negation is used to exclude noise.

Note that the above formula goes far beyond a simple logical combination of queries as it combines complex structures instead of individual search terms. Practically all types of search fields, including keywords, terms, subject headings, journal titles, citations and references, even corporate addresses and author names/identifiers can be incorporated into the retrieval strategy. This allows the inclusion of advanced bibliometric methods, such as direct citations, bibliographic coupling, co-citations, textual similarity, and their various combinations to form complex strategies. The strategy might even extend to the combination of different databases as will be shown in the second example below.

Bibliometric retrieval can then be fine-tuned by extending or reducing the sets of unconditional and conditional criteria and by adjusting the thresholds for the bibliometric components of the criteria such as number or share of references, coupling units, etc. The strategy can thus aim at defining subjects in a narrower or broader sense by including or excluding related research topics as sketched in Fig. 1.

In the following we give two examples to illustrate this logical combination and the option of fine-tuning. The first example refers to the topic *stem-cell* research (see Glänzel et al. 2004), the second one to *bioinformatics* (e.g., Glänzel et al. 2009).

1. Example: *Stem cells* (Glänzel et al. 2004)

UC1	Journal in WoS = STEM CELLS
UC2	Address word = STEM CELL*
UC3	Keywords = (STEM CELL* OR STEM (ES) CELL* OR PROGENITOR* CELL* OR HEMATOPOI* CELL*)
CC1	Journal = JOURNAL OF HEMATOTHERAPY & STEM CELL RESEARCH
CC2	Keywords = (BONE-MARROW OR UMBILICAL-CORD-BLOOD OR UCB OR HUCB OR CYTOPOI* OR MEGAKARYOPOI* OR ERYTHROPOI* OR MYELOPOI* OR THROMBOPOI* OR STROMAL CELL* OR PRECURSOR CELL*)
CC3	Cited source documents = UC1 OR UC2 OR UC3
Options	Papers citing 3–5 other papers classified as unconditionally relevant making up at least 40 % of all SCIE references, or 6–10 UC papers making up at least 30 % of all SCIE references, or citing more than 10 UC papers

The search strategy resulting from the above conditions can be formulated as

$$\mathrm{BR} := (\mathrm{UC1} \vee \mathrm{UC2} \vee \mathrm{UC3}) \vee ((\mathrm{CC1} \vee \mathrm{CC2}) \wedge \mathrm{CC3})$$

2. Example: *Bioinformatics* (Glänzel et al. 2009)

UC1	Journal in WoS = BIOINFORMATICS (formerly COMPUTER APPLICATIONS IN THE BIOSCIENCES), JOURNAL OF COMPUTATIONAL BIOLOGY, BRIEFINGS IN BIOINFORMATICS, BMC BIOINFORMATICS
UC2	Journal in Medline = IN SILICO BIOLOGY, PSB ON-LINE PROCEEDINGS, APPLIED BIOINFORMATICS, PLOS COMPUTATIONAL BIOLOGY
CC1	Keywords in title = BIOINFORMATICS, COMPUTATIONAL BIOLOG*, SYSTEMS BIOLOGY
CC2	Related records of UC1
CC3	Cited or citing source of UC1
Options	Different rules for citations—both directions—can be defined

The search strategy resulting from the above conditions can be formulated as

$$\mathrm{BR} := (\mathrm{UC1} \vee \mathrm{UC2}) \vee (\mathrm{CC1} \wedge (\mathrm{CC2} \vee \mathrm{CC3}))$$

In conclusion, it should be stressed that the objective of bibliometric retrieval is not to provide complex bibliometric solutions but an appropriate groundwork for further analysis by retrieving a tailor-made dataset according to the needs of the corresponding bibliometric study.

New research lines: hybrid methods and 'core documents' in bibliometrics-aided retrieval

The method described above can readily be applied on the large scale, i.e., for the retrieval in a complete database and preferably for large topics. In order to facilitate the retrieval, especially within rather small subject areas, bibliometrics-aided retrieval can be extended by using *hybrid similarities*, that is, the direct combination of text- and citation-based methods for building document similarities. This means, instead of the combination of lexical search terms with citation links or with "related records" based on bibliographic coupling (e.g., Glänzel et al. 2009; Laurens et al. 2010), similarities, that are directly based on so-called hybrid textual-citation methods, can be applied to some of the conditional criteria. This might help increase the efficiency and avoid too many steps in the logical BR algorithm as well as too complex textual search constructs. The bibliometric retrieval (BR) can then be re-defined as follows.

A document is considered relevant if it meets some conditional criterion (CC_j), and is strongly linked based on a hybrid similarity measure to at least a certain number of documents meeting an unconditional criterion (UC_i).

This definition immediately leads us to the notion of *core documents*. The term 'core documents', originally proposed by Small (1973) in the context of co-citation analysis, was re-introduced by Glänzel and Czerwon (1996) on the basis of bibliographic coupling to identify those papers which form important nodes in the network of scholarly communication. In fact, their definition can readily be extended to hybrid techniques as well [cf., Glänzel and Thijs (2012a)]:

'Core documents' are documents that have at least $n > 0$ links of at least a given strength r 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.

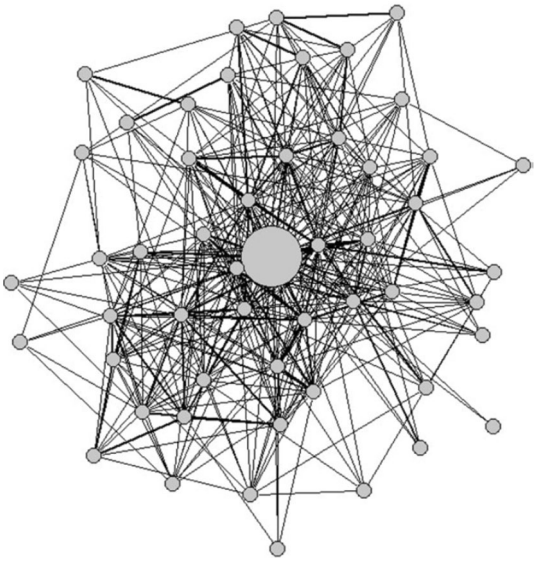

Fig. 3 Visualisation of the link environment of a 'core document' (according to Glänzel and Thijs 2012b)

'Core documents' do not only play an important role in structural bibliometrics for detecting and describing emerging topics and for network representation; they are useful instruments for the retrieval of documents, namely to identify further relevant documents by following their strong and medium-strong links. The link environment of a typical core document is shown in Fig. 3. Following the links outgoing from the core document, i.e., from the large circle in the centre of the picture, lead to related documents, which, in turn, are connected with further relevant documents. In this manner, one or several related 'core documents' stand for a particular research theme and the link structure of 'core documents' can be used to represent the network structure of important topics within the subject under study (Glänzel and Thijs 2011). Thus 'core documents' can also be considered an efficient tool for reduction of dimensionality.

Concluding discussion

Bibliometrics-aided retrieval (BR) in the context of subject delineation for evaluative purposes essentially differs from traditional information retrieval. The fact that bibliometric domain studies are mostly focused on complex interdisciplinary fields is one of the most important reasons. The second reason is that in bibliometric retrieval there is no end-user that could remove possible noise from the large sets of retrieved documents. The aim of bibliometrics-aided retrieval is thus a representative and adjustable coverage of the field under study. Metrics can be used for fine-tuning search strategies and to stop retrieval at any level.

The underlying document set will, of course, never be exhaustive nor will it be completely free of noise, but fine-tuning the retrieval through adjusting the search criteria and setting thresholds for the metric components can help meet both purists' and generalists' needs. BR is a powerful tool to develop and adjust search strategy at any level of aggregation.

It facilitates the delineation of very complex and interdisciplinary research fields and topics. Adjustable hybrid (i.e., text/citation-based) techniques allow bibliometrics-aided retrieval even in fields where citations do not play an important role, for instance, in the applied sciences, in most fields of the social sciences and in the humanities (cf., Glänzel and Thijs 2011).

'Core documents' represent the most interlinked papers in a set. Following their strong and weaker links might help retrieve relevant information without formulating further search queries. This feature provides a strong added value that might make 'core documents' interesting and attractive for information retrieval too.

Acknowledgments The present study is an extended version of a paper presented at the 14th International Conference on Scientometrics and Informetrics, Vienna (Austria), 15–19 July 2013 (Glänzel, 2013). Figure 3 is reproduced from Glänzel and Thijs (2012b) with permission of the publisher.

References

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering, 137*, 85–86.

Garfield, E., & Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation, 14*(3), 195–201.

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. (2006). The 'perspective shift' in bibliometrics and its consequences. In *I international conference on multidisciplinary information sciences & technologies (InScit2006)*, Mérida, Spain, 25–28 October 2006. http://de.slideshare.net/inscit2006/the-perspective-shift-in-bibliometrics-and-its-consequences.

Glänzel, W., Janssens, F., Speybroeck, S., Schubert, A., Thijs, B., & Rafols, I. (2006). Towards a bibliometrics-aided data retrieval for scientometric purposes. In *Poster presented at the 9th international conference on science and technology indicators*, Leuven, Belgium, 7–9 September 2006. Book of Abstracts, pp. 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. (2012a), Hybrid solutions—the best of all possible worlds? *Bibliometrie & Praxis und Forschung*, 1, 3. URN:urn:nbn:de:bvb:355-152-4.

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

Glänzel, W., Verbeek, A., Du Plessis, M., van Looy, B., Magerman, T., Thijs, B. et al. (2004). *Stem cells—analysis of an emerging domain of scientific and technological endeavour*. https://www.ecoom.be/sites/ecoom.be/files/downloads/stemcells_domain_study.pdf.

Glänzel, W., & Veugelers, R. (2006). Science for wine: A bibliometric assessment of wine and grape research for wine producing and consuming countries. *American Journal of Enology and Viticulture, 57*(1), 23–32.

Laurens, P., Zitt, M., & Bassecoulard, E. (2010). Delineation of the genomics field by hybrid citation-lexical methods: Interaction with experts and validation process. *Scientometrics, 82*(3), 647–662.

Noyons, E. C. M., Buter, R. K., van Raan, A. F. J., Schmoch, U., Heinze, T., Hinze, S. et al. (2003). *Mapping excellence in science and technology across Europe. Nanoscience and Nanotechnology*, Draft report of project EC-PPN CT-2002-0001 to the European Commission, University Leiden. Accessible via: http://studies.cwts.nl/projects/ec-coe/downloads/Final_report_13112003_nano.pdf.

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

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