Scientometrics (2015) 102:2323–2345
DOI 10.1007/s11192-014-1485-2

Science models for search: a study on combining scholarly information retrieval and scientometrics

Peter Mutschke · Philipp Mayr

Received: 31 October 2014 / Published online: 27 November 2014

Abstract Models of science address statistical properties and mechanisms of science. From the perspective of scholarly information retrieval (IR) science models may provide some potential to improve retrieval quality when operationalized as specific search strategies or used for rankings. From the science modeling perspective, on the other hand, scholarly IR can play the role of a validation model of science models. The paper studies the applicability and usefulness of two particular science models for re-ranking search results (Bradfordizing and author centrality). The paper provides a preliminary evaluation study that demonstrates the benefits of using science model driven ranking techniques, but also how different the quality of search results can be if different conceptualizations of science are used for ranking.

Keywords Information retrieval · Scientometrics · Bradfordizing · Network analysis · Betweenness · Science models

Introduction: Science models and scholarly information retrieval

Models of science address statistical properties of the science system under study in order "to gain insights into the inner workings of science" (Scharnhorst et al. 2012). The major focus of these models, referred to as *science models* in the following, is to capture basic patterns and mechanisms of scholarly activities and structures in science as well as their evolution over time. Thus, science modeling asks for the "when (temporal), where (spatial), what (topical), or with whom (network analysis)" (Scharnhorst et al. 2012) science appears and works. Science models have a long tradition in the field of scientometrics

P. Mutschke (✉) · P. Mayr
GESIS–Leibniz-Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany
e-mail: peter.mutschke@gesis.org

P. Mayr
e-mail: philipp.mayr@gesis.org

which produces a number of valid models such as growth laws of scientific productivity (Lotka 1926), publications and citations (de Price 1976), the distribution of papers over journals (Bradford 1934), the emergence and evolution of scientific paradigms (Goffman and Warren 1969), and statistical regularities of co-authorship networks (Barabasi et al. 2002; Newman 2001, 2004). Nowadays, the availability of large amounts of data as well as novel methods to extract and visualize structure and dynamics in science lead to improved possibilities to gain new insights into the nature of scholarly activities and structural phenomena in science (cf. Scharnhorst et al. 2012).

Scholarly digital libraries (DLs) can be considered as particular representations of science systems. Hence, searching in DLs can be seen as a particular way of interacting with the specific science system. The purpose of information retrieval (IR) systems is to help users accomplish a search task. From this perspective science models predicting structures in science may provide some potential, in particular in a growing scientific information spaces, to improve scholarly IR when operationalized as specific scholarly search strategies (cf. Bates 1990) or used for rankings. From the science modeling perspective, on the other hand, IR can play the role of a validation model of science models, i.e. a kind of "litmus test" for the adequacy of science models for understanding, forecasting and communicating the science system. The more a science model incorporated in a scholarly IR system contributes to its overall usefulness in solving a scholarly information seeking problem the more the science model in question gains in reliability as regards the patterns and mechanisms of science described by the model. Thus, we propose to consider the contribution of a science model to the usefulness (Cole et al. 2009) of a scholarly IR system as an evaluation criterion of the adequacy of the model under study. We expect that the investigation of science model driven search and ranking services might contribute to both.

In this paper we study the applicability and usefulness of particular science models for a specific scholarly activity: *the search for scientific information*. The paper is inspired by recent discussions at two workshops, namely the ISSI 2013 workshop "Combining Bibliometrics and Information Retrieval"[1] and the ECIR 2014 workshop "Bibliometric-enhanced Information Retrieval"[2] (Mayr et al. 2014). While we introduced the general idea of combining science modeling and IR together with three examples of science-model driven IR in brief in Mutschke et al. (2011) in this paper we focus more detailed on two non-textual ranking approaches that reason about global structures in scientific communities:[3] Bradfordizing and author centrality. Bradfordizing focusses on a concentration pattern of papers in journals (called Bradford law of scattering) yielding in a coreness measure for journals. When used for re-ranking, papers of core journals appear on the top of the ranking. Author centrality addresses the strategic position of authors in co-authorship networks such as betweenness. When used for re-ranking, papers authored by high betweenness researchers are ranked on top. Our major hypothesis is that these models represent different conceptualizations of science and thus provide quite different views to a scientific information space when used as IR services (such as re-ranking, as proposed here).

Our approach is motivated by the fact that traditional IR approaches fail at a crucial point where the application of science models may help: The perceived expectation of users searching the web is that retrieval systems should list the most relevant or valuable

[1] http://www.gesis.org/en/events/events-archive/conferences/issiworkshop2013/.

[2] http://www.gesis.org/en/events/events-archive/conferences/ecirworkshop2014/.

[3] Instead of addressing just local features such as citation counts of papers or author productivity.

documents in the result list first (so-called relevance ranking). However, users are often overburdened by the amount of results returned by classical IR systems that typically do not structure result sets according to usefulness criteria of the user. Moreover, using just pure text-statistical rankings such as TF-IDF (term frequency–inverse document frequency) seem to stagnate in terms of relevance improvement (Armstrong et al. 2009). This strongly suggests to enhance classical IR approaches (such as TF-IDF) by the use of ranking models that take structural attributes of the science system into account. Google pagerank and its derivations (see e.g. Lin 2008) or Google Scholar's citation rank are just two popular examples for informetric-based rankings applied in internet search engines. In the following the paper discusses the two models from both perspectives, science modeling as well as IR.

Case study

Models

Bradford distribution of journals as a bibliometric model of science

In most disciplines scientific journals still play a major role in the scientific communication process. They appear periodically, they are topically focused, they have established domain-specific standards of quality control and often they are involved in the academic gratification system. Metrics like the famous impact factor are aggregated on the journal level. In disciplines like medical sciences journals are the main platform for a scientific community to communicate and discuss new research results. These examples shall illustrate the impact journals bear in the context of science models. Journals and journal publication characteristics are important as regards modeling science, understanding the functioning of science respectively. Here, Bradford's law of scattering can be seen as a particular approach to science modeling insofar as it contributes to a better understanding of publication patterns in science.

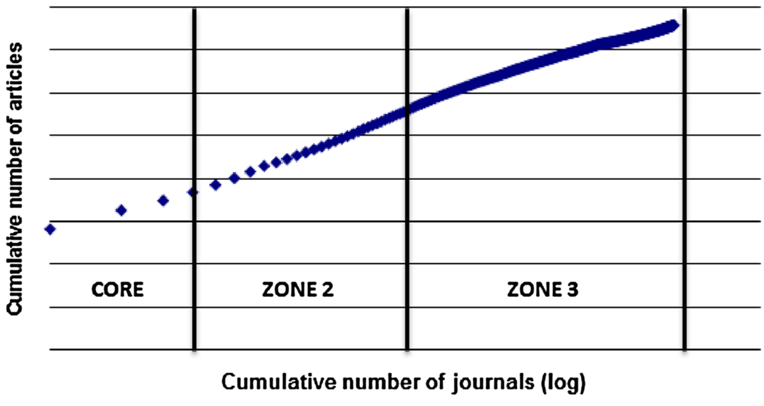

Fig. 1 A typical Bradford distribution: Core, zone 2 and zone 3 (so-called periphery). The cumulative number of journals (*x-axis*) is displayed on a logarithmic scale

Bradford's law of scattering bases on literature observations the librarian Bradford carried out in 1934. His findings and the formulation of the bibliometric model stand for the beginning of the modern documentation (Bradford 1948)—a documentation which founds decisions on quantifiable measures and empirical analyses. The early empirical laws described by Lotka, Zipf, and of course Bradford, are landmark publications which still influence research in scientometrics (Bookstein 1990), but also in other research communities like computer science or linguistics. In brief, bibliometric and informetric research investigates the mathematical descriptions and models of regularities of all observable objects in the library and information science area. These objects include authors, publications, references, citations, all kinds of texts etc. Bradford's work bases on analyses of journal publications on different subjects in the sciences. Fundamentally, Bradford's law states that literature on any scientific field or subject-specific topic scatters in a typical way. A core or nucleus with the highest concentration of papers—normally situated in a set of few so-called core journals—is followed by zones with loose concentrations of paper frequencies (see Fig. 1 for a typical Bradford distribution). The last zone covers the so-called periphery journals which are located in the model far distant from the core subject and normally contribute just one or two topically relevant papers in a defined period. Bradford's law as a general law in informetrics can successfully be applied to most scientific disciplines, and especially in multidisciplinary scenarios (Mayr 2009). Bradford describes his model in the following:

"The whole range of periodicals thus acts as a family of successive generations of diminishing kinship, each generation being greater in number than the preceding, and each constituent of a generation inversely according to its degree of remoteness" (Bradford 1934).

Bradford provides in his publications (1934, 1948) just a graphical and verbal explanation of his law. A mathematical formulation has been added later by early informetric researchers. Bradford's original verbal formulation of his observation has been refined by Brookes (1977) to

$$G(r) = k \ln\left\{\frac{a + r}{a}\right\},$$

where $G(r)$ is the cumulative distribution function, k and a are constants, and r is the rank 1,2, …, n. The result of the application of this formula is often called a rank-order distribution of the items in the samples. In the literature we can find different names for this type of distribution, e.g. "long tail distribution", "extremely skewed", "law of the vital few" or "power law" which all show the same properties of a self-similar distribution. Thus, Bradford's law is relevant for scholarly information systems due to its structuring ability and the possibility to reduce a large document set into a core and succeeding zones.

Power-law distribution of author betweenness as a network model of science

Network models describe and analyse relational data such as, in the case of scholarly activities and structures, relationships between co-occurring topics as well as their life-cycle within a network of topics (Callon et al. 1983; Grivel et al. 1995; Mutschke and Quan Haase 2001), the structure and growth of scholarly collaboration networks (Barabasi et al. 2002; Newman 2001, 2004), or the nature of citation relationships between scientific publications (cf. Scharnhorst et al. 2012). The background of author betweenness as a network pattern of scholarly activity is the perception of science as "a social network of researchers that generate and validate a network of knowledge"

(Scharnhorst et al. 2012; see also He 2009; Mali et al. 2012). Those networks are seen as "one representation of the collective, self-organized emerging structures in science" (Börner and Scharnhorst 2009). De Haan (1997) distinguishes six different forms of scientific collaboration that provide different "window(s) on patterns of collaboration within science" (Mali et al. 2012), such as shared editorship, shared supervision of projects, shared organization of conferences, common participation in research programs, common authorship of research proposals and co-authorship of scientific publications. Due to the increasing complexity of nowadays research issues collaborative and inter-disciplinary research becomes more and more important in order to solve research problems that can hardly be solved by a single scientist or within a single discipline (Jiang 2008; Mali et al. 2012).

The increasing significance of collaboration in science does not only correlate with an increasing amount of collaborative papers (Lu and Feng 2009; Leydesdorff and Wagner 2008) but also, and more importantly, with an increasing impact of multi-authored papers. Beaver (2004) pointed out that co-authored papers are much more likely to be cited than single authored papers (see also Lang and Neyer 2004) and concludes that "collaborative research produces significantly more authoritative research". He (2009) showed that co-authoring is significantly associated with the impact of a paper, given by the relative number of citations received. Based on a study of the impact of collaboration, Börner et al. (2005) discovered that "a global brain comprised of larger highly successful co-authorship teams is developing (that) … will be able to dynamically respond to the increasing demands on information processing and knowledge management".

Hence, collaboration in science is mainly represented by co-authorships. Transferred to a whole scientific community, co-authorships form a network as a particular "prototype of a social network" (Yin et al. 2006) reflecting the overall collaboration structure in science on the base of the final outcome of research (manifested in scientific publications). Thus, it can be argued that "social interaction structures (in science) are best described by co-authorship networks" (Mali et al. 2012). According to the notion of a *social network* (Wassermann and Faust 1994), co-authorship networks are described as a graph $G = (V, E)$, where the set V of vertices represents authors, and the set E of edges represents symmetric ties between authors who co-authored a paper.

Co-authorship networks have been intensively studied. Most of the studies focus mainly either on general network properties (see Newman 2001; Barabasi et al. 2002) or on empirical investigation of particular networks (Yin et al. 2006; Liu et al. 2005). However, yet not enough attention has been paid to co-authorship networks as general models of science. From the perspective of science modeling it is important to note that, as co-authorship also indicate the share of knowledge among authors, "a co-authorship network is as much a network depicting academic society as it is a network depicting the structure of our knowledge" (Newman 2004). A crucial effect of being embedded within such a networked collaboration structure is that its structure affects the flow of information through the community as well as the opportunities to collaborate (Freeman 1977; Yin et al. 2006). On the level of the individual scientists it is also evident that different authors also occupy different positions in a co-authorship network, according to their different roles within the scientific community, which are, as a consequence, mutually affected by the structure of the network: "Some individuals are more influential and visible than others as a result of their position in the network" (Yin et al. 2006). Thus, as inequality of positions is a structural property in social networks in general, locating strategic positions in scientific collaboration structures becomes an important issue also in science modeling (see also Jiang 2008; Lu and Feng 2009; Liu et al. 2005).

This perception of collaboration in science corresponds directly with the idea of *structural centrality* (Bavelas 1948; Freeman 1977, 1978/79) which characterizes centrality as a property of the strategic position of nodes within the relational structure of a network. As the concept of centrality addresses the level of involvement of a node within the relational structure of a network it provides an index of the degree of "importance" or "prominence" (Wassermann and Faust 1994; cf. Mutschke 2010) of the node in a network. Many measures of centrality have been proposed in the literature. Generally accepted is the perception that centrality measures reflect the opportunities of actors to influence interaction and information flow in the network. Given this, central actors are actors who occupy strategic positions in the network that possess a high potential of influence with respect to the network flow. The differences among the measures result primarily from different assumptions about the structure of that network flow (cf. Borgatti and Everett 2006; Friedkin 1991; Mutschke 2010). Interestingly, collaboration in science is mainly characterized in terms that match a concept of centrality widely used in social network analysis, namely the *betweenness* of actors in a social network. Yin et al. 2006 see co-authorship as a "process in which knowledge flows among scientists", and Chen et al. (2009) characterize "scientific discoveries as a brokerage process (which) unifies knowledge diffusion as an integral part of a collective information foraging process". This betweenness-related role of collaboration in science was also confirmed by a number of empirical studies. Yan and Ding (2009) discovered a high correlation between citation counts and the betweenness of authors in co-authorship networks. Liu et al (2005) discovered a strong correlation between program committee membership and betweenness in co-authorship networks. Mutschke and Quan-Haase (2001) observed a high correlation of betweenness in co-authorship networks and betweenness of the authors' topics in keyword networks.

Betweenness (Freeman 1977) evaluates the degree to which a node is positioned *between* others on shortest paths in a graph. Thus, the measure focuses on the structural dependency of any pair of nodes (i, j) from a third node k which is located between i and j. However, that dependency is reduced to the extent to which there are shortest paths between i and j which do not contain k. Therefore, the pair dependency $\delta_{ij}(k)$ of two nodes i and j from a third node k is defined as the ratio of the number $\sigma_{ij}(k)$ of shortest paths connecting i and j via k to the total number σ_{ij} of shortest paths between i and j: $\delta_{ij}(k) = \sigma_{ij}(k)/\sigma_{ij}$. The betweenness $C_B(k)$ of k is then given by the sum of all pair dependencies with k for all unordered node pairs (i, j): $C_B(k) = \sum_{i,j}\delta_{ij}(k)$, $i, j, k \in V$. The normalized measure, due to the maximum is: $C'_B(k) = C_B(k)/(n^2 - 3n + 2)$. The more a node plays such an intermediary role for other pairs of nodes, the more the node is central by betweenness, i.e. the more a node is needed as an "interface" between other nodes in the graph or even needed as a "cutpoint" between fractions of the graph, which would be unconnected otherwise, the higher its betweenness. Thus, betweenness emphasizes the bridge or brokerage role of a node in a network (Freeman 1977, 1978/79, 1980; cf. Mutschke 2010).

Applied to co-authorship networks, a high betweenness is supposed to be achieved by authors who connect other authors, primarily authors from different communities. That outstanding position holds a great deal of "power" in the community in terms of control of information and knowledge flow through the network and, more importantly, the sharing of knowledge among different fractions of the scientific community (cf. Lu and Feng 2009; Leydesdorff and Wagner 2008; Liu et al. 2005; Yin et al. 2006). High betweenness authors are therefore characterized as "pivot points of knowledge flow in the network" (Yin et al. 2006). They can be seen as the main driving forces not only for just bridging gaps between different communities but also, by bringing different authors together, for community making processes and finally for progress in science. Thus, it is assumed that an author's impact on a scientific field can be quantified by his/her betweenness in co-authorship networks (see also Yan and Ding 2009).

According to the phenomenon of unequal distribution of scientific publications saying that a relatively small number of scientists author a larger fraction of publications (de Price 1976) co-authorship networks usually exhibit a scale-free structure whose degree distribution follows a power-law (Albert and Barabasi 2002), i.e. the social structure of a co-authorship network is usually governed by a small number of authors having a high number of co-authors and a large number of authors having a small number of co-authors. The same phenomenon can also be seen for betweenness distributions in co-authorship networks (Yan and Ding 2009; Newman 2001, 2004). The strong inverse relationship indicates that there is usually only a fairly small number of influential scientists playing broker roles in a scientific community (Newman 2001, 2004) such that a ranking model based on author centrality in co-authorship networks is given a high selectivity on publications of the central actors of a field.

Accordingly, an index of betweenness of authors in a co-authorship network can be seen as an index of the relevance of authors for the domain under study. This strongly suggests the use of author betweenness for re-ranking large result sets in scholarly IR systems (Mutschke 1994, 2001; see also Zhou et al. 2007). In the following chapter we describe an IR prototype which has been introduced in Mutschke et al. (2011). The prototype implements the two science models Bradfordizing and Author Centrality.

Proof-of-concept prototype

Both science model based re-ranking approaches were implemented in a proof-of-concept retrieval prototype (developed in the IRM projects,[4] see Fig. 2) that uses the models as particular search stratagems (Bates 1990) to enhance retrieval quality. As a test collection the SOLIS database from GESIS was taken that contains about 370,000 metadata records on social science literature (including title, abstract, controlled keyword etc.). As a basic retrieval engine the open source search engine Solr was used which provides a standard term frequency based ranking mechanism (TF-IDF). The two re-ranking models proposed work as retrieval add-ons implemented as Solr plugins. Both address the problem of oversized results since they re-order search results on-the-fly according to the ranking criterion of the science model used, i.e. either by journal coreness or by author betweenness. In the following the science model driven re-ranking methods are explained and compared with the standard ranking model.

Standard ranking

For comparison, let us first explain the standard TF-IDF ranking model used in classical IR engines. TF-IDF is a pure term frequency oriented method which evaluates the relationship between the frequency of a term in a document and its frequency in the entire corpus (Salton et al. 1983). TF-IDF yields high values when a term t occurs many times in only a few documents d of the collection. The more often a term appears in all documents, the lower the TF-IDF value. Thus, the basic approach of TF-IDF is to evaluate the discriminating power of terms in an information space. By summing up all the TF-IDF weights of

[4] http://www.gesis.org/en/research/external-funding-projects/archive/irm/.

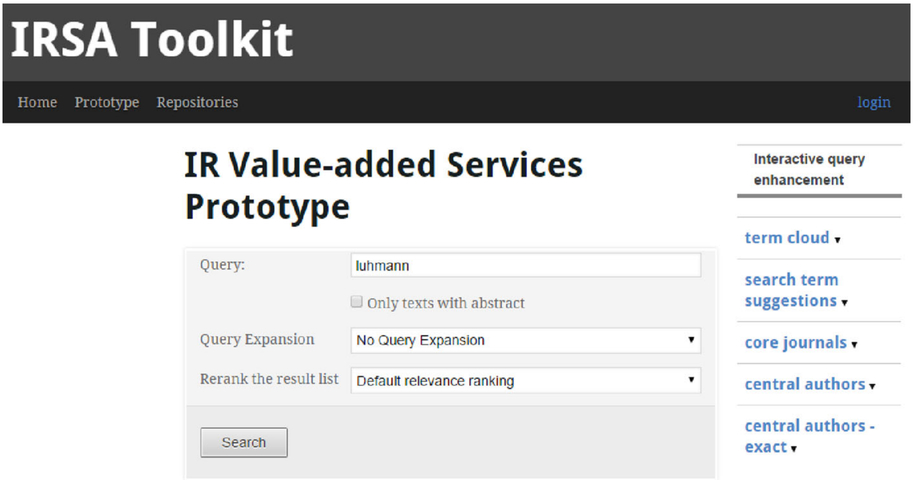

Fig. 2 IRM research prototype, see demo under http://multiweb.gesis.org/irsa/IRMPrototype

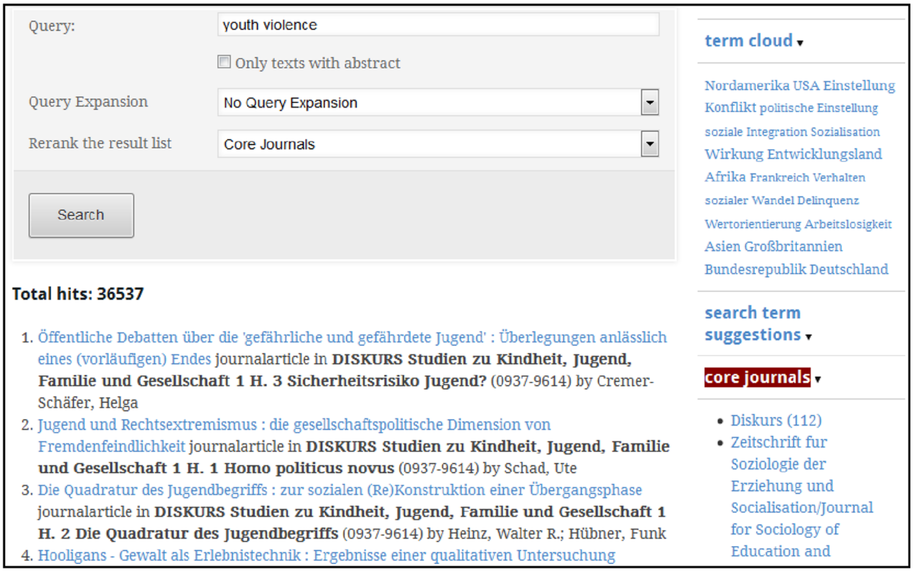

Fig. 3 A Bradfordized search by the IRM research prototype. ISSN numbers of core journals for the search term "youth violence" are displayed on the *right side* of the screen, together with their productivity (article counts). The result set is ranked by coreness of journals

terms in a query a relevancy score for each document in the collection with respect to the query can be calculated. Many variations of this basic approach have been proposed which mainly address the way the weight is calculated, but the key approach of assigning a weight for each term in each document is not altered in these variations. In our prototype TF-IDF is used as the initial ranking model, delivered by the search engine Solr. This initial ranking is then rearranged by metrics derived from the science models described above, i.e. journal coreness and author betweenness. The following sections describe the two re-ranking models proposed.

Re-ranking by journal coreness

Bradfordizing re-ranks a result set of journal articles according to the frequency of journals in the result set such that articles of core journals are ranked ahead (see example in Fig. 3). This is a robust and quick way of sorting the central publication sources for any query to the top positions of a result set.

The Bradfordizing procedure (see White 1981) is implemented in the IRM prototype as a Solr plugin. In a first step the search results are filtered by their ISSN numbers.[5] The next step aggregates all results with an ISSN number. For this step we use a build-in functionality of our prototype engine Solr, the Solr faceting mechanism. Facets in Solr can be defined on any metadata field, in our case the "source" field of our databases. The journal with the highest ISSN count gets the top position in the result. The second journal gets the next position, and so on (see example in Fig. 3). This procedure is an exact implementation of the original Bradfordizing approach. In the last step, the document ranking step, our current implementation works with a simple boosting mechanism. The frequency counts of the journals are used as boosting factors for documents in these journals. The numerical ranking value from the original TF-IDF ranking of each document is multiplied with the frequency count of the journal. The result of this multiplication will be taken as the ranking value for the final document ranking.

In principle, this ranking technique can be applied to any search result providing qualitative metadata (e.g. journal articles in literature databases). Generally, Bradfordizing needs 100 or more documents because smaller document sets often show too little scattering to divide the result into meaningful zones. Bradfordizing can be applied to document types other than journal articles, e.g. monographs (cf. Worthen 1975; Mayr 2009). Monographs e.g. provide ISBN numbers which are also good identifiers for the Bradfordizing analysis. To conclude, our implementation of re-ranking by Bradfordizing is a simple approach which is generic, adaptable to various document types and quickly implementable as a build-in functionality. The only precondition for the application is the existence of qualitative metadata to assure precise identification and access to the documents.

Bradfordizing a result set shows the following advantages: (a) a structured view on a result set which is ordered by journals; (b) an alternative view on publication sources in an information space which is intuitively closer at the research process than statistical methods (e.g. best match ranking) or traditional methods (e.g. exact match sorting); (c) an approach to switch between the search modus, e.g. starting with directed term searching and changing to a browsing mode (Bates 2002). A recent analysis of the improvements of relevance distribution between the journal zones after Bradfordizing has been recently investigated (Mayr 2013, 2014).

Re-ranking by author betweenness

Re-rankings based on author betweenness re-order a set of publications retrieved such that publications of high betweenness authors are ranked on top (Mutschke 1994, 2001; Zhou et al. 2007). This is done by the following procedure: (0) The Solr engine returns a result

[5] ISSNs are stable identifiers for journals.

set due to the user's query by matching query terms to indexing terms. (1) For each document in the result set co-authorships are detected, according to the list of authors assigned to a document.[6] (2) Authors and co-authorships provided by step (1) are then sequentially added to the graph $G = (V, E)$, where the set V of vertices represents authors involved in a co-authorship ("co-authors"),[7] and the set E of edges represents co-authorships, i.e. a co-author becomes a node $v \in V$ and a co-authorship becomes an edge $e = (i, j) \in E$, $i, j \in V$, in G. (3) For each node in the graph its betweenness centrality is measured, according to the formula explained above. (4) Each document in the result set is assigned a weight given by the maximum betweenness value of its authors. Single authored publications are captured by this method if their authors appear in the graph due to other publications they have published in co-authorship. (5) The result set is re-sorted in descending order by the betweenness weight of the documents (delivered by step 4) such that documents of high betweenness authors appear on the top of the list (see Fig. 4). Publications from pure single-authors are finally added to the end of the list in order to preserve the original result list.

Several studies have demonstrated that re-rankings of retrieval results based on network analysis methods can improve the retrieval performance significantly (Bogers and Bosch 2006; Yaltaghian and Chignell 2002; Mutschke 2001, 2004). However, apart from Google that use link structures among Web pages for ranking, enhanced network models of science have not been widely used in scholarly information systems so far. Our model of author centrality based ranking originates from the model initially proposed by Mutschke (1994) which has been re-implemented for a real-life IR environment, to our knowledge before anyone else, within the Daffodil system (Mutschke 2001; Fuhr et al. 2002).

Currently, both re-ranking models are implemented by the sowiport[8] portal from GESIS. In the following sections we will present and discuss preliminary results of an evaluation of the retrieval quality of the models proposed.

Evaluation results

The benefit of the two science model driven ranking approaches was evaluated by a retrieval test. For this, a relevance assessment of the top ranked papers was conducted with 73 participants (students in library and information science) where each assessor had to choose one topic out of 10 different predefined topics from the CLEF corpus (see Table 1).

In our experiment the two science model based re-ranking models proposed, author betweenness (AUTH) and journal coreness (BRAD), were compared to the Solr standard relevance ranking algorithm (based on TF-IDF) which was taken as the baseline (SOLR). For each of the ten CLEF topics and each of the three retrieval models SOLR, AUTH, BRAD the n = 10 top ranked SOLIS documents were taken as the pool of documents to be assessed (cf. Voorhees and Harman 2005). The assessors then judged the documents in the pool as relevant or not relevant without knowing the originating ranking model. The size of the subsets returned by the three models can be seen in Table 2.

The assessors did 43.78 single assessments in average which sums up to 3,196 single relevance judgments in total, with an overall agreement of 62 % (Schaer et al. 2010). To

[6] Actually, co-authorships are computed during indexing time in advance and are retrieved by the system via particular facets added to the user's query.

[7] To reduce computation effort pure single-authors are not added to the graph.

[8] http://sowiport.gesis.org/. Furthermore, sowiport provides simple re-ranking models such as citation counts of papers and author productivity as well.

Fig. 4 Re-ranking done by author centrality in the IRM research prototype. Central authors are displayed on the *right side* of the screen, in descending order of their betweenness in the co-authorship graph. The result set is ranked by the betweenness of authors in the co-authorship graph

rate the reliability and consistency of agreement between the different assessments Fleiss's Kappa measure of inter-grader reliability (Fleiss 1971) was used. Fleiss's Kappa expresses the extent to which the observed amount of agreement exceeds the expected random degree of agreement which is given if all raters made their ratings completely randomly. Kappa scores can range from 0 (no agreement) to 1.0 (full agreement). All Kappa scores κ in our experiment range between 0.20 and 0.52 (see Table 2) which are fair up to moderate levels of agreement according to Landis and Koch (1977). In seven cases the Kappa scores yield acceptable values according to the more conservative interpretations of Greve and Wentura (1997). Thus, the results are reliable enough to make further considerations as regards the precision quality of the models proposed.

The precision of each service was then calculated by dividing the total number of documents assessed as relevant by the total number of all assessed documents (relevant and not relevant). The resulting precision values (in %) for each model as well as their improvement in precision, compared to the other two models under study, can be seen in Table 3.

The average precision of SOLR is 57 %, which forms the baseline in the following considerations. The precision values of SOLR range between 17 and 76 %. Ignoring the 17 % value all other values are stable around the baseline. The two alternative re-ranking mechanisms BRAD and AUTH achieved an average precision that is near the baseline, namely 60 % for AUTH and 57 % for BRAD. AUTH yielded an even higher, but not significantly higher average precision than SOLR. Both re-rank mechanisms show a relatively stable behavior (except from also just one outlier for each of the models). Thus, the precision of the two alternative non-textual re-ranking models is quite similar to the standard ranking model.

The more important observation is that the three models perform very differently (see Table 3). For five topics (83, 84, 93, 105, 153) the baseline SOLR achieved the best result.

Table 1 Ten selected CLEF topics and their description (in English) and the derived queries (in German)

ID	Topic	Description	Query
83	Media and war	Find documents on the commentatorship of the press and other media from war regions	(medien AND krieg*)
84	New media in education	Find documents reporting on benefits and risks of using new technology such as computers or the Internet in schools	("neue medien" AND unterricht*) OR ("neue medien" AND schule*)
88	Sports in Nazi Germany	Find documents about the role of sports in the German Third Reich	(sport* AND nationalsozialismus*)
93	Burnout syndrome	Find documents reporting on the burnout syndrome	burnout*
96	Costs of vocational education	Find documents reporting on the costs and benefits of vocational education	(kosten* AND berufsausbildung*)
105	Graduates and labour market	Find documents reporting on the job market for university graduates	(akademik* AND arbeitsmarkt*)
110	Suicide of young people	Find documents investigating suicides in teenagers and young adults	((selbstmord* OR suizid*) AND jugend*)
153	Childlessness in Germany	Information on the factors for childlessness in Germany	(kinderlos* AND deutsch*)
166	Poverty	Research papers and publications on poverty and homelessness in Germany	(armut* AND deutschland*)
173	Propensity towards violence among youths	Find reports, cases, empirical studies and analyses on the capacity of adolescents for violence	(gewaltbereit* AND jugend*)

Table 2 Number of documents retrieved by each service[a] and Inter-grader reliability expressed by Fleiss's Kappa (κ)

Topic ID	SOLR	AUTH	BRAD	κ
83	1,136	1,136	488	0.522
84	389	389	103	0.305
88	110	110	33	0.528
93	185	185	81	0.416
96	159	159	60	0.488
105	923	923	318	0.466
110	176	176	48	0.223
153	343	343	136	0.203
166	2,883	2,883	838	0.439
173	755	755	227	0.411

[a] The number of documents retrieved by BRAD is lower than the number of documents retrieved by SOLR and AUTH since BRAD considers only the fraction of journal articles in a result set

However, only in one case (topic 84) a clear improvement of more than 10 % compared to the second best ranking can be seen. AUTH performed best for four topics (96, 110, 166, 173), with a clear distance to the second best ranking in three cases (topics 96, 166, 173). BRAD performed best for one topic (88), with a clear distance to the second best ranking. Only in two cases (topics 93 and 153) all three models perform in a nearly similar manner, whereas in all other cases larger differences in precision can be observed: In three cases AUTH performs significantly better than SOLR (topics 96, 166, 173). In two cases even more than 100 % better (topics 96, 166). BRAD performs significantly better than SOLR in four cases (topics 88, 96, 166, 173), in one case even more than 100 % better (topic 166). SOLR performs significantly better than AUTH in two cases (topics 84, 88), in both cases even more than 100 %, and significantly better than BRAD in four cases (topics 83, 84, 105, 110), in one case (topic 83) even more than 100 %. In all other cases the precision values of AUTH and BRAD are similar to the precision values of SOLR. Thus, when we compare the two alternative rankings (AUTH, BRAD) to the standard model (SOLR) we find a 3:2 success relationship between AUTH and SOLR, and a success relationship of 4:4 between BRAD and SOLR. A strong diversity in success can be also observed when we compare AUTH and BRAD, however with a major predominance for the benefit of AUTH. AUTH performs clearly better than BRAD in five cases (topics 83, 96, 110, 166, 173), whereas BRAD is much better than AUTH only in two cases (topics 84, 88).[9]

This strongly demonstrates how different the quality of search results can be if different conceptualizations of science are used as retrieval and ranking models. Interestingly, the different models also point to very different relevant documents. A comparison of the relevant top 10 documents returned by the three models showed that the result sets are nearly disjoint. Out of the total amount of 300 top ranked documents for all models and

[9] There are three extreme cases of very low precision values which need some explanation: A reasonable explanation for the low precision of SOLR in the case of topic 166 is most likely the low selectivity of the search term 'Deutschland' (Germany) in SOLIS which might negatively affect the precision of TF-IDF. A possible explanation for the low precision of BRAD in this case of topic 83 is the lower coverage of media related journals in SOLIS. Likewise, the low precision of AUTH in the case of topic 88 can be explained by the rather fragmentary representation of historical science topics in SOLIS which lead to less representative networks. However, much more detailed research needs to be done here.

Table 3 Precision values and improvements (in %) of AUTH, BRAD and SOLR for the ten topics (best precision values and improvements of more than 10 % displayed in bold)

Topic ID	Precision AUTH (%)	Precision BRAD (%)	Precision SOLR (%)	AUTH > SOLR (%)	BRAD > SOLR (%)	AUTH > BRAD (%)	BRAD > AUTH (%)	SOLR > AUTH (%)	SOLR > BRAD (%)
83	70.00	26.76	**76.00**	–	–	**162**	–	9	**184**
84	34.26	64.49	**74.04**	–	–	–	**88**	**116**	**15**
88	15.00	**63.33**	53.33	–	**19**	–	**322**	**256**	–
93	72.00	74.00	**74.49**	–	–	–	3	3	1
96	**86.36**	63.64	40.00	**116**	**59**	**36**	–	–	–
105	65.12	59.09	**66.67**	–	–	10	–	2	**13**
110	**74.00**	56.00	70.00	6	–	**32**	–	–	**25**
153	57.58	57.45	**63.27**	–	–	0	–	10	10
166	**65.12**	55.17	17.24	**278**	**220**	**18**	–	–	–
173	**56.38**	48.94	38.71	**46**	**26**	**15**	–	–	–
avg.	**59.58**	56.89	57.37	4	–	5	–	–	1

Table 4 Coverage of author betweenness based rankings and network structures

Topic ID	Precision[a] (%)	Number docs retrieved[b]	Number docs of co-authors[c]	Ratio docs co-aut. (%)	Number vertices	Number components	Giant size	Ratio giant (%)	Avg. size comp.
83	70	1,136	1,104	97	590	216	14	2,4	2,7
84	34	389	371	95	346	112	14	4	3,1
88	15	110	107	97	64	22	15	23,4	16
93	72	185	181	98	201	69	7	3,5	2,9
96	86	159	133	84	114	33	27	23,7	3,5
105	65	923	818	89	690	189	96	13,9	3,7
110	74	176	170	97	78	32	8	10,3	2,4
153	57	343	333	97	243	79	19	7,8	3,1
166	65	2,883	2,699	94	1,679	403	396	23,6	4,2
173	56	755	726	96	512	142	35	6,8	3,6
Avg.	59	706	664	94	452	130	63	11,9	4,5

[a] Cf. Table 3

[b] Cf. Table 2

[c] Number of documents from authors contained in the co-authorship graph, i.e. all documents except from documents authored by pure single-authors (due to the procedure described above)

Table 5 Rank orders of precision, result set size and network properties

Rank	Rank by precision	Rank by number docs	Rank by number docs of co-aut.	Rank by ratio docs co-aut.	Rank by number vertices	Rank by number components	Rank by giant size	Rank by ratio giant	Rank by avg. size comp.
1	96	166	166	93	166	166	166	96	88
2	110	83	83	83, 88, 110, 153	105	83	105	166	166
3	93	105	105	173	83	105	173	88	105
4	83	173	173	84	173	173	96	105	173
5	105, 166	84	84	166	84	84	153	110	96
6	153	153	153	105	153	153	88	153	84, 153
7	173	93	93	96	93	93	83, 84	173	93
8	84	110	110		96	110	110	84	83
9	88	96	96		110	96	93	93	110
10		88	88		88	88		83	

topics (3 models × 10 per model × 10 topics) only 22 intersections in total could be found (cf. Mutschke et al. 2011). Thus, there is no or very little overlap between the sets of relevant top documents obtained from the different rankings. AUTH and SOLR as well as AUTH and BRAD have just three and BRAD and SOLR have just six relevant documents in common (for all 10 topics). That confirms the hypothesis that retrieval and ranking services modeling different conceptualizations of science indeed provide quite different views to a scientific information space.

However, is there also a relationship between the statistical pattern described by the science model and its retrieval quality when used for re-ranking? Here, we took a closer look at AUTH, i.e. the question whether a relationship between network structure and retrieval performance of author betweenness based rankings can be identified or not. First of all, we observed a high coverage of re-rankings done by AUTH. More than 90 % of the retrieved documents were captured by AUTH (see Table 4, column 5). The rest were publications from pure single-authors.However, there is obviously no clear relationship between the precision of AUTH and basic network properties that might explain the differences in precision per topic. If one considers the rank orders of the nine features from Table 4 it is evident that the low precision for topic 88 (having the lowest rank order by precision, see Table 5) might be induced by the low size of the result set and, accordingly, the low size of the network (also lowest rank order for both features, see Table 5, columns 3 and 6). On the other hand, topic 88 is relatively high on proportion of giant size to the entire network (third highest rank order by ratio of giant size). Topic 96 having the highest rank order by precision, in contrast, is very low as regards result set and network size (similar to topic 88), but captures the highest rank as regards proportion of giant size. The topic having the second highest rank order by precision (topic 110) is also low on result set and network as well as giant size. Topic 93, having the third highest rank as regards precision, has the lowest rank as regards giant size. Overall, there are small numbers of components and small giant sizes for topics with low precision of AUTH as well as for topics with a very high AUTH precision and vice versa (compare, for instance, topics 83, 88 and 110).

Thus, the positive finding is that betweenness based rankings are obviously not (too) sensitive to the level of fragmentation of the co-authorship network under study. Even for fragmented networks publications of high betweenness authors yield a high precision. The overall ranking quality of AUTH is well on average. However, its quality varies a lot from topic to topic, and it is obviously difficult to predict the usefulness of AUTH by taking into account network properties.

Discussion

The precision tests turned out that science-model driven search services provide beneficial effects to information retrieval quality which are on average as high as the precision of standard rankings. They provide a particular view to the information space that is quite different from traditional retrieval methods such that the models open up new access paths to knowledge spaces. Insofar as an evaluation of the retrieval quality of those search services is seen as a litmus test of the adequacy of the science models investigated in predicting major scholarly patterns or activities, the two models studied are verified as expressive models of science. Moreover, it turned out that the results provided by the investigated models differ to a great extent indicating that the models obviously highlight very different dimensions of scientific activity: co-authoring (captured by network oriented

IR models like author betweenness), and publishing (captured by bibliometric models of IR like Bradfordizing).

Thus, it could be shown how structural models of science can be used to improve retrieval quality. By applying those different conceptualizations of science as IR models the user is provided by different result cutouts containing other relevant documents in the first section of the result set than other models, which additionally offers the opportunity to switch between different models, such as term-based search services and structure-oriented ranking models (as proposed in a general form by Bates (2002). This is an important perspective due to a further result of the study, namely that the retrieval quality of the models differs immensely from topic to topic. This might be an effect of the general problem of modeling science into a core and a periphery—the general approach of all ranking models—that always runs the risk of disregarding important scientific developments outside the core (cp. Hjorland and Nicolaisen 2005). Therefore, combining diverse models of science in a retrieval process might enhance its quality immensely.

The other way around, the IR experiment turned out that to the same extent to which science models contribute to IR (in a positive as well as negative sense), science-model driven IR might contribute to a better understanding of different conceptualizations of scholarly activities and patterns. Recall and precision values of retrieval results obtained by science model oriented search and ranking techniques seem to provide important indicators for the adequacy of science models in representing and predicting structural phenomena in science. The high precision of author betweenness based rankings, for instance, strongly indicates that a network model approach to information retrieval issues, the concept of centrality in particular, adequately addresses an author's scientific impact (cf. Yan and Ding 2009). An explanation of this phenomenon might be that authors of high betweenness are supposed to be of high community driving relevance for the science system. Actually, Mutschke and Quan-Haase (2001) found that central authors are strongly associated with the mainstream topics of a research field. This suggests that publications of high betweenness authors, which finally base their central positions, are of particular relevance for the scientific community. This perception of the strategic role of highly central actors in science might also explain the high precision of rankings done by author betweenness: Authors of high betweenness apparently address the key topics of a field. This is reflected by their publications which therefore have obviously a strong tendency to represent the main issues of the field. As pivot points of research high betweenness actors pass on and integrate different research interests emerging from different local groups which make them attractive for collaboration (respectively, for occupying positions *between* different groups by collaboration) because of their greater ability or willingness to merge different research issues and perspectives. Both their strategic position as pivot points of network flow and their key issue oriented focus make central authors to main players of the field. As a consequence, those actors in turn gain in value for scientific collaborations which mutually reinforce the network effect.

This perspective might also explain a further relevant result of the evaluation, namely the fact that author centrality based rankings favor other documents than pure term-frequency based rankings. Thus, by focusing on publications of actors who occupy strategic positions governing the knowledge flow in a scientific community a network model approach of ranking provides a particular view to the document space that is quite different from the term-oriented view of traditional ranking methods. On the other hand, no clear relationship between network structures and the precision of rankings done by author betweenness could be observed.

As regards Bradfordizing we likewise discovered positive as well as negative effects of this method. Some added-values appear very clearly. On an abstract level, re-ranking by Bradfordizing can be used as a compensation mechanism for enlarged search spaces with interdisciplinary document sets. Bradfordizing can be used in favor of its structuring and filtering facility. Our analyses show that the hierarchy of the result set after Bradfordizing is a completely different one compared to the original ranking. The user gets a new result cutout with other relevant documents which are not listed in the first section (in our experiment the top 10 documents) of the original list. Furthermore, Bradfordizing can be a helpful information service to positively influence the search process, especially for searchers who are new on a research topic and don't know the main publication sources in a research field (see Mayr 2014). The opening up of new access paths and possibilities to explore document spaces can be a very valuable facility. Additionally, re-ranking via bradfordized documents sets offer an opportunity to switch between term-based search and the search mode browsing. Thus, the approach provides an alternative ranking option, as one additional way or stratagem to access topical documents (cf. Bates 2002). Interesting in this context is a statement by Bradford himself where he explains the utility of the typical three zones. The core and zone 2 journals are in his words "obviously and a priori relevant to the subjects", whereas the last zone (zone 3) is a very "mixed" zone, with some relevant journals, but also journals of "very general scope" (Bradford 1934). Pontigo and Lancaster (1986) come to a slightly different conclusion of their qualitative study. They investigated that experts on a topic always find a certain significant amount of relevant items in the last zone. This is in agreement with quantitative analyses of relevance assessments in the Bradford zones (Mayr 2009, 2013). The study shows that the last zone covers significantly less often relevant documents than the core or zone 2. The highest precision can constantly be found in the core.

To conclude, modeling science into a core and a periphery like the Bradford approach always runs the risk and critic of disregarding important developments outside the core. Hjorland and Nicolaisen (2005) started a first exploration of possible side effects and biases of the Bradford methods. They criticized that Bradfordizing favors majority views and mainstream journals and ignores minority standpoints. This is a serious argument, because by definition, journals which publish few papers on specific topics have very little chance to get into the core of a more general topic. A counter-argument could be that the Bradfordizing approach is just an application which is working on existing document sets. The real problem is to be found beforehand, in the development of a data set, especially in the policy of a database producer.

On the other hand, an evaluation of the method and its effects carried out in two laboratory-based information retrieval experiments (CLEF and KoMoHe) using a controlled document corpus and human relevance assessments (see Mayr 2013; Ingwersen and Järvelin 2005 for pros and cons of this methodology) turned out that Bradfordizing is a very robust and promising method for re-ranking the main document types (journal articles and monographs) in today's digital libraries. The IR tests show that relevance distributions after re-ranking improve at a significant level if articles in the core are compared with articles in the succeeding zones. The items in the core are significantly more often assessed as relevant than items in zone 2 or zone 3. The largest increase in precision can typically be observed between core and zone 3. This has been called the Bradfordizing effect (see Mayr 2013). The results of our study can also be seen as a coalescence of Bradford Law insofar as Bradford did not postulate or observe a relevance advantage in the core. In Bradford's eyes all documents in his bibliographies were "relevant to a subject". His focus was the scattering of documents across journals, not the relevance distribution between document zones. According to Saracevic (1975), Bradford was one of the first persons who used the term relevant in our context ("relevant to a subject"). The results in Mayr (2013) show that articles in core journals are valued more often as relevant than articles in succeeding zones (cf. Garfield 1996). This is an extension to the original conception of relevance distribution in the zones by Bradford. As we can empirically see, bibliometric distributions like Bradford distributions can also be described as "relevance related distributions" (Saracevic 1975). The relevance advantage in the core can probably be explained by the fact that (a) core journals publish more state-of-the-art articles, (b) core journals are more often reviewed by peers in a certain field and (c) core journals cover more aspects of the searched topic than journals in the peripheral zones. Further research is needed to clarify these questions.

Open issues

Our study turned out that science model driven ranking models can improve retrieval quality immensely, but also how different the retrieval quality can be depending on the topic. Thus, a lot of research effort needs to be done to make more progress in coupling science modeling with IR. Apart from further evaluation studies with larger amounts of topics and users for different scientific domains the major challenge that we see is to find more reliable patterns of structures in science that allow to better predict the usefulness of a science model for IR purposes: What are the parameters that provide a valid and reliable assessment of the usefulness of a particular model? Is there, for instance, a particular relationship between network properties (such as amount of co-authorship in a field, core-periphery structure of a co-authorship network, amount of clustering, characteristic path length, average degree etc.) and the accuracy of an author centrality based ranking? Do we need parameters to punish certain entities, e.g. multidisciplinary and very frequently publishing journals in a Bradford distribution? What are the particular constraints of the model under study when used for IR purposes? Re-rankings done by author betweenness, for instance, are mostly constrained by the problem of author name ambiguity (homonymy, synonymy) in bibliographic databases. In particular the potential homonymy of names may misrepresent the true social structure of a scientific community. A further problem is the computation effort of complex calculations like betweenness for which fast solutions need to be found that can be used for IR purposes. A further important issue is to provide scholarly IR users with visual representations of structures in science (like knowledge maps[10]) that may help the user with a structural overview of the information space in question that is derived from specific patterns in science. The probably most challenging issue, however, results from the fact that structures in science evolve over time: the dynamic mechanisms that underlie the formation of structures in science need to be related to dynamic features in scholarly IR as well.

Acknowledgments We thank Philipp Schaer and Thomas Lüke who were the main developers and our co-investigators in the IRM projects. This work was funded by DFG, Grant No. INST 658/6-1 and SU 647/5-2.

[10] This issue is addressed by the COST action KNOWeSCAPE (http://knowescape.org/).

References

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*, 47–97.

Armstrong, T. G., Moffat, A., Webber, W., & Zobel, J. (2009). Improvements that don't add up: Ad-hoc retrieval results since 1998, . In: *Conference on information and knowledge management* (pp. 601–610). doi:10.1145/1645953.1646031.

Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A, 311*, 590–614.

Bates, M. J. (1990). Where should the person stop and the information search interface start? *Information Processing and Management, 26*(5), 575–591.

Bates, M. J. (2002). Speculations on browsing, directed searching, and linking in relation to the Bradford distribution. *Paper presented at the fourth international conference on conceptions of library and information science* (CoLIS 4).

Bavelas, A. (1948). A mathematical model for group structure. *Applied Anthropology, 7*, 16–30.

Beaver, D. (2004). Does collaborative research have greater epistemic authority? *Scientometrics, 60*(3), 309–408.

Bogers, T., & van den Bosch, A. (2006). Authoritative re-ranking in fusing authorship-based subcollection search results. In: *Proceedings of the sixth Belgian–Dutch information retrieval workshop* (pp. 49–55). DIR-2006 (2006).

Bookstein, A. (1990). lnformetric distributions. Part I: Unified overview. *JASIS, 41*(5), 368–375.

Borgatti, S. P., & Everett, M. (2006). A graph-theoretic perspective on centrality. *Social Networks, 28*, 466–484.

Börner, K., Dall'Asta, L., Ke, W., & Vespignani, A. (2005). Studying the emergine global brain: Analysing and visualizing the impact of co-authorship teams. Complexity. Special issue on understanding. *Complex Systems, 10*(4), 57–67.

Börner, K., & Scharnhorst, A. (2009). Visual conceptualizations and models of science. *Journal of Informetrics, 3*, 161–172.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering, 137*(3550), 85–86.

Bradford, S. C. (1948). *Documentation*. London: Lockwood.

Brookes, B. C. (1977). Theory of the Bradford law. *Journal of Documentation, 33*(3), 180–209.

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

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics, 3*, 191–209.

Cole, M., Liu, J., Belkin, N. J., Bierig, R., Gwizdka, J., Liu, C., Zhang, J., & Zhang, X. (2009). Usefulness as the criterion for evaluation of interactive information retrieval. In *Proceedings of the third human computer information retrieval workshop, Washington, DC.*

De Haan, J. (1997). Authorship patterns in Dutch sociology. *Scientometrics, 39*(2), 197–208.

de Price, D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*, 292–306.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry, 40*, 35–41.

Freeman, L. C. (1978/79). Centrality in social networks: Conceptual clarification. *Social Networks 1*,(pp. 215–239).

Freeman, L. C. (1980). The gatekeeper, pair-dependency and structural centrality. *Quality and Quantity, 14*, 585–592.

Friedkin, N. E. (1991). Theoretical foundations for centrality measures. *American Journal of Sociology, 96*, 1478–1504.

Fuhr, N., Schaefer, A., Klas, C.-P., & Mutschke, P. (2002). Daffodil: An integrated desktop for supporting high-level search activities in federated digital libraries. In Agosti, M., Thanos, C. (Eds.), *Research and advanced technology for digital libraries. 6th European Conference, EDCL 2002. Proceedings. Lecture Notes in Computer Science* (Vol. 2458, pp. 597–612). Berlin, Heidelberg, New York: Springer.

Garfield, E. (1996). The significant scientific literature appears in a small core of journals. *The Scientist, 10*(17), 13. Retrieved from http://www.garfield.library.upenn.edu/commentaries/tsv10(17)p13y090296.html.

Goffman, W., & Warren, K. S. (1969). Dispersion of papers among journals based on a mathematical analysis of two diverse medical literatures. *Nature, 221*(5178), 1205–1207.

Greve, W., & Wentura, D. (1997). *Wissenschaftliche Beobachtung: Eine Einführung*. Weinheim: PVU/Beltz.
Grivel, L., Mutschke, P., & Polanco, X. (1995). Thematic mapping on bibliographic databases by cluster analysis: A description of the SDOC environment with SOLIS. *Knowledge Organisation, 22*, 70–77.
He, Z.-L. (2009). International collaboration does not have greater epistemic authority. *JASIST, 60*(10), 2151–2164.
Hjørland, B., & Nicolaisen, J. (2005). *Bradford's law of scattering: Ambiguities in the concept of "subject"*. Paper presented at the 5th international conference on conceptions of library and information science.
Ingwersen, P., & Järvelin, K. (2005). The turn-integration of information seeking and retrieval in context. In W. B. Croft (Ed.), *The Kluwer international series on information retrieval*. Dordrecht: Springer.
Jiang, Y. (2008). Locating active actors in the scientific collaboration communities based on interaction topology analysis. *Scientometrics, 74*(3), 471–482.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.
Lang, F. R., & Neyer, F. J. (2004). Kooperationsnetzwerke und karrieren an deutschen hochschulen. *KfZSS, 56*(3), 520–538.
Leydesdorff, L., & Wagner, C. S. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics, 2*(4), 317–325.
Lin, J. (2008). Pagerank without hyperlinks: Reranking with PubMed related article networks for biomedical text retrieval. *BMC Bioinformatics*, *9*. Retrieved from http://www.biomedcentral.com/1471-2105/9/270.
Liu, X., Bollen, J., Nelson, M. L., & van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing and Management, 41*(2005), 1462–1480.
Lotka, A. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*(12), 317–323.
Lu, H., & Feng, Y. (2009). A measure of authors' centrality in co-authorship networks based on the distribution of collaborative relationships. *Scientometrics, 81*(2), 499–511.
Mali, F., Kronegger, L., Doreian, P., & Ferligoj, A. (2012). Dynamic scientific co-authorship networks. In A. Scharnhorst, et al. (Eds.), *Models of science dynamics. Encounters between complexity theory and information sciences* (pp. 195–232). Berlin, Heidelberg: Springer.
Mayr, P. (2009). *Re-ranking auf basis von Bradfordizing für die verteilte Suche in digitalen bibliotheken*. Berlin: Humboldt-Universität zu Berlin.
Mayr, P. (2013). Relevance distributions across Bradford zones: Can Bradfordizing improve search? In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, & H. Moed (Eds.), *14th International society of scientometrics and informetrics conference* (pp. 1493–1505). Vienna, Austria. Retrieved from http://arxiv.org/abs/1305.0357.
Mayr, P. (2014). Are topic-specific search term, journal name and author name recommendations relevant for researchers? In *4th European symposium on human–computer interaction and information retrieval*. London. Retrieved from http://arxiv.org/abs/1408.4440.
Mayr, P., Scharnhorst, A., Larsen, B., Schaer, P., & Mutschke, P. (2014). Bibliometric-enhanced information retrieval. In M. de Rijke et al. (Ed.), *Proceedings of the 36th European conference on information retrieval* (ECIR 2014) (pp. 798–801). Amsterdam: Springer. http://arxiv.org/abs/1310.8226.
Mutschke, P. (1994). Processing scientific networks in bibliographic databases. In H. H. Bock, et al. (Eds.), *Information systems and data analysis. Prospects-foundations-applications. Proceedings 17th annual conference of the GfKl 1993* (pp. 127–133). Heidelberg, Berlin: Springer.
Mutschke, P. (2001). Enhancing information retrieval in federated bibliographic data sources using author network based stratagems. In Constantopoulos, P., Sölvberg, I. T. (Eds.), *Research and advanced technology for digital libraries: 5th European conference*, ECDL 2001, Proceedings. Lecture notes in computer science (Vol. 2163, pp. 287–299). Berlin, Heidelberg, New York: Springer.
Mutschke, P. (2004). Autorennetzwerke: Verfahren der Netzwerkanalyse als Mehrwertdienste für Informationssysteme. Bonn: Informationszentrum Sozialwissenschaften. (IZ-Arbeitsbericht Nr. 32). http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/iz_arbeitsberichte/ab_32.pdf.
Mutschke, P. (2010). Zentralitäts- und Prestigemaße. In R. Häußling & C. Stegbauer (Eds.), *Handbuch Netzwerkforschung*. Wiesbaden: VS-Verlag für Sozialwissenschaften.
Mutschke, P., Mayr, P., Schaer, P., & Sure, Y. (2011). Science models as value-added services for scholarly information systems. *Scientometrics, 89*(1), 349–364. doi:10.1007/s11192-011-0430-x.
Mutschke, P., & Quan-Haase, A. (2001). Collaboration and cognitive structures in social science research fields: Towards socio-cognitive analysis in information systems. *Scientometrics, 52*(3), 487–502.
Newman, M. E. J. (2001). The structure of scientific collaboration networks. *PNAS, 98*, 404–409.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *PNAS, 101*, 5200–5205.

Pontigo, J., & Lancaster, F. W. (1986). Qualitative aspects of the Bradford distribution. *Scientometrics, 9*(1–2), 59–70.

Salton, G., Fox, E. A., & Wu, H. (1983). Extended Boolean information retrieval. *Communications of the ACM, 26*(11), 1022–1036.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science, 26*(6), 321–343.

Schaer, P., Mayr, P., & Mutschke, P. (2010). Implications of inter-rater agreement on a student information retrieval evaluation. In Atzmüller, M., Benz, D., Hotho, A., Stumme, G. (Eds.), *LWA2010: Lernen, Wissen and Adaptivität; Workshop Proceedings; Kassel*, 4–6 Oct 2010. http://arxiv.org/ftp/arxiv/papers/1010/1010.1824.pdf.

Scharnhorst, A., Börner, K., & van den Besselaar, P. (Eds.). (2012). *Models of science dynamics. Encounters between complexity theory and information science*. Berlin, Heidelberg: Springer.

Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval*. Cambridge: The MIT Press.

Wassermann, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.

White, H. D. (1981). 'Bradfordizing' search output: how it would help online users. *Online Review, 5*(1), 47–54.

Worthen, D. B. (1975). The application of Bradford's law to monographs. *Journal of Documentation, 31*(1), 19–25.

Yaltaghian, B., & Chignell, M. (2002). Re-ranking search results using network analysis. A case study with Google. In *Proceedings of the 2002 conference of the Centre for Advanced Studies on Collaborative Research*.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *JASIST, 60*(10), 2107–2118.

Yin, L., Kretschmer, H., Hannemann, R. A., & Liu, Z. (2006). Connection and stratification in research collaboration: An analysis of the COLLNET network. *Information Processing and Management, 42*, 1599–1613.

Zhou, D., Orshansky, S. A., Zha, H., & Giles, C. L. (2007). Co-ranking authors and documents in a heterogeneous network. In *Proceedings of the 2007 seventh IEEE international conference on data mining* (pp 739–744).