Int J Digit Libr (2018) 19:139–149
https://doi.org/10.1007/s00799-017-0217-7

Bag of works retrieval: TF*IDF weighting of works co-cited with a seed

Howard D. White[1]

Received: 11 October 2016 / Revised: 29 April 2017 / Accepted: 1 May 2017 / Published online: 19 May 2017

Abstract Although not presently possible in any system, the style of retrieval described here combines familiar components—co-citation linkages of documents and TF*IDF weighting of terms—in a way that could be implemented in future databases. Rather than entering keywords, the user enters a string identifying a work—a seed—to retrieve the strings identifying other works that are co-cited with it. Each of the latter is part of a "bag of works," and it presumably has both a co-citation count with the seed and an overall citation count in the database. These two counts can be plugged into a standard formula for TF*IDF weighting such that all the co-cited items can be ranked for relevance to the seed, given that the entire retrieval is relevant to it by evidence from multiple co-citing authors. The result is analogous to, but different from, traditional "bag of words" retrieval, which it supplements. Some properties of the ranking are illustrated by works co-cited with three seeds: an article on search behavior, an information retrieval textbook, and an article on centrality in networks. While these are case studies, their properties apply to bag of works retrievals in general and have implications for users (e.g., humanities scholars, domain analysts) that go beyond any one example.

Keywords Co-citation · Relevance ranking · Seed documents · Models of users

This paper revises and considerably expands on one that appeared in the *Proceedings of the 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR2016)*, pp 63–72 http://ceur-ws.org/Vol-1567/paper7.pdf.

✉ Howard D. White
whitehd@drexel.edu

[1] College of Computing and Informatics, Drexel University, Philadelphia, PA 19104, USA

1 Introduction

The option of starting searches with works as seeds is built into the major citation indexes, Web of Science, Scopus, and Google Scholar. In cited reference searches in WoS, for example, strings that uniquely identify cited articles, books, and other publications can be both entered and retrieved. The following is typical:

BATES MJ, 1989, V13, P407, ONLINE REV

Strings like this need to be translated into full references, of course, but it is enough for present purposes that they can function as seeds and retrievals in WoS or similar databases (as can DOI's). Assume, then, that all such strings constitute a "bag of works." By contrast, in paradigmatic information retrieval (IR), the documents the strings represent are seen as a "bag of words"—that is, content-bearing words from titles, abstracts, or full texts on which algorithms operate to rank them by topical closeness to the query (as in, e.g., [1]). Bag of works (BoW) retrievals could be implemented in databases with bibliographic records that list the items each document cites. The bag of works in such databases is the total set of strings that denote cited works [2]. Retrievals involve single seed documents as queries (as in, e.g., [3]), a co-citation relevance metric, and a standard version of TF*IDF weighting [4: 109], [5: 543], in which logged term frequencies (TF) are multiplied by logged inverse document frequencies (IDF).

Citation-based retrievals such as these supplement more conventional topical retrievals. The examples to come were produced with the Dialog Classic search system, formerly on ProQuest but defunct since 2013. Dialog hosted the same citation databases as the Web of Science. However, Dialog's RANK command enabled users to do something not possible in WoS: They could enter a seed document as a query and then more or less instantly generate a list of all

the works co-cited with the seed, ranked high to low by their co-citation counts. A co-citation count here reflects how many later works have cited the seed and another work jointly. If numerous, the ranked documents formed core-and-scatter distributions of the sort long familiar in bibliometrics. But Dialog had more: The same RANK command could also be made to list every work's total citation count in the database. Given both co-citation counts and citation counts from the same database, one has the raw data for TF*IDF weights of the sort long familiar in information retrieval. The counts used in the examples below are from the Social Sciences Citation Index on Dialog and are quite robust.

The value of TF*IDF weighting in this case is that it reranks a lengthy list of co-cited documents by the *specificity* of their relation to the seed. What this means for word-blind strings denoting works will be shown.

The present paper does not experimentally test the BoW model. It simply illustrates it with detailed commentary on works actually retrieved in three sample distributions. The goal is to introduce *recurrent features of BoW structures as verbal objects*—in particular, the exploitable structure that IDF creates. For example, high-ranking works in BoW retrievals are like what typical recommender systems display, but lower-ranked works, including those bottom-ranked, may also interest certain users—that is, users who want to examine items not merely "similar to" the seed, but items related to it for other reasons. The entire BoW retrieval is relevant to the seed in varying degrees by empirical co-citation evidence from multiple authors. Thus, a BoW retrieval does not stand or fall on the basis of topical similarity to the seed, if that is taken as a gold standard.

Seeds may be works of any kind—articles, books, research reviews, and so on—as long as they have a record (preferably substantial) of being cited. Works retrieved by the seed may also be of any kind. They may or may not resemble the seed in global topic. Since seeds merely imply topical content, their semantic relations with retrieved works will be more various and less predictable than those obtained by term-matching or query expansion based on term-matching. Yet when titles are added, BoW retrievals have broadly predictable features:

- The relevance of many high-ranked works to the seed will be relatively easy for a user to detect, especially a domain-literate one. Sometimes this relevance will be obvious from exact or partial matches between the seed's topical indicators (e.g., title, descriptors, journal name) and those of the retrieved works. Other times, the connection must be inferred by a user with appropriate background knowledge. The feasibility of such inferences is not a mere conjecture, however, because co-citers have already made similar ones.
- The relevance of retrieved works to a seed may also be indicated by author-matches, regardless of whether the retrievals resemble the seed topically.
- The relevance of works to the seed will be progressively less easy to detect over the whole distribution. A substantial segment of low-ranked works will be hard to relate to the seed because they appear to have little in common with it (terms in common with its title, for example). Their connection must be learned from the document that cites them both, and it may be only distant. One predictor is the citation window—the amount of body text that separates references to the seed and to some other item [6,7]. But works not like the seed *are* co-cited with it in the overall context set by the citing document and thus bear consideration. If nothing else, they suggest connections that might never occur to someone who retrieved only works whose relation to the seed is obvious.
- Aside from topical relations, works co-cited with the seed indicate the seed's intellectual neighbors as perceived by citers over time—its historical reception. This is not a consideration in standard IR experiments, but a user with domain knowledge may well see it as a bonus.

To unpack the retrieval formula:

TF Term frequencies in this case are counts of documents co-cited with the seed document in later works: the higher the counts, the greater the predicted relevance of the documents to the seed. This parallels the bag of words model, in which the more times a query term appears in a document, the more relevant to the query that document is predicted to be. A seed such as the BATES MJ string above can retrieve the strings co-cited with it, regardless of the natural language they contain or how indexers have described them.

IDF In standard topical retrieval, the IDF factor weights nonstopped words in the database progressively lower as the number of documents containing them increases, because frequently used words are relatively poor discriminators of subject matter. In bag of works retrieval, IDF functions in the same way but is interpreted differently. The raw DF values are the total citation counts for documents in the database. The higher the DF count, the more well-known a document is and the greater its breadth of implication, its general applicability. IDF, which inverts the DF count, tends to favor works that are *narrowly and specifically* related to the seed over works that are more broadly or less immediately related. The promotion of specific terms and the demotion of nonspecific terms is what Karen Sparck Jones intended the IDF factor to do when she proposed it [8]—she called it "statistical specificity"—except that she and almost everyone since have used IDF weighting on *words* rather than *works*. Yet on word-blind strings denoting co-cited works, IDF performs no less well.

*TF*IDF* The very simple formula uses base-10 logs, and N is estimated with a rounded count of the records in the database. For any co-cited document string:

$$\text{Relevance to the seed} = (1 + \log\text{TF}) * (\log(N/\text{DF}))$$

The seed and any number of retrieved works can be ranked by their individual TF*IDF weights. Relevance varies directly with the TF factor and inversely with the IDF factor. That is, TF*IDF weighting raises works whose co-citation counts (TF) with the seed are high relative to their overall citation counts (DF), and it lowers works whose co-citation (TF) counts with the seed are low relative to their overall citation counts (DF).

With respect to operational systems, CiteSeer^x automatically returns a small, nontransparent selection of the titles co-cited with a seed, but it is the exception. In the Web of Science, Scopus, and Google Scholar, no co-citation retrievals of any kind are possible. In citation databases, algorithms take a seed document as input and return the documents that cite it, a linkage known as direct citation. The documents in this retrieved set—call it Set A—are by default ranked high to low by their own citation counts (in Google Scholar) or by recency of publication (in the Web of Science and Scopus). However, the direct citation relationship does not allow the documents in Set A to be ranked by their relevance to the seed, because each simply lists the seed once among its references, and so its score with respect to the seed is always one. All citing documents thus appear equally relevant to the seed. By contrast, the documents *co-cited* with the seed can be ranked for relevance to it, because their variable co-citation counts can be treated as relevance scores. This requires the further step of retrieving the co-cited documents as Set B, which then become available for TF*IDF (that is, BoW) weighting.

2 Related work

Carevic and Schaer [9] used the iSearch test collection in physics to experiment with BoW retrieval as presented in White [10]. In iSearch, documents come with both cited references and assessors' relevance ratings on a four-point scale. The authors were looking for title cohesion between their seeds, which were documents prescored by assessors as relevant to a topic, and the documents those seeds retrieved by TF*IDF-weighted co-citation. The authors intended to evaluate BoW retrieval in the style of mainstream IR research. That proved unfeasible, however, because the co-citation counts they found in iSearch were small or nonexistent. But in examples from two search topics, the top-ranked co-cited documents did cohere with the seed documents in their title terms.

Although document recommender systems use a variety of ranking measures [11], BoW weighting has apparently not been proposed before. Nor, apparently, has relative co-citation frequency [12: 269], a much older measure that yields a similar but not identical ranking. The closest analogue to BoW in actual recommender systems, aside from raw co-citation counts, is weighting by CCIDF—"the common citations between any pair of documents weighted by the inverse frequency of citation" [13: 70]. CCIDF, developed for CiteSeer^x, can also be called IDF-weighted bibliographic coupling (or co-coupling) strength; it is used in CiteSeer^x to recommend short lists of seed-related titles under the heading "Active Documents." Bibliographic coupling, known as "Related Records" in WoS, measures the similarity of any two works by counting the references they share. The added IDF factor "downplays the importance of common citations to highly cited methodological papers, for example" [13: 70], because sharing a methodology does not indicate how similar papers are in their global topics.

The retrieval effectiveness of CCIDF was slightly improved, according to Huynh et al. [14], by augmenting it with co-citation data for 16 seed papers in computer science. However, in at least three other experimental tests of recommender systems, co-citation counts, co-coupling counts, and CCIDF were all outperformed by more sophisticated measures. The best retrievals in Liang et al. [15] made use of three dependency relations between citing papers and cited papers (*comparable*, *based on*, and *general*); also of indirect as well as direct citation links. In the system described in Küçüktunç et al. [16], a user enters a whole file of papers (e.g., an article's references) as seeds; from these a citation graph of references from and citations to the seeds is formed. Operating on the graph, a variant of PageRank gave the most relevant results. Pan et al. [17] sought to recommend the papers most similar to a target paper. Their best solution was obtained by combining citation graphs with TF*IDF-weighted content words ("heterogeneous graphs"), so as to take advantage of the strengths of each kind of information.

In Beel et al. [18], the authors report that CCIDF performed no better than CC alone in the recommender system built into their Docear reference management software. They extracted seeds from Docear users' document collections and used them to retrieve bibliographically coupled items. On a random basis, about half of the retrievals were weighted with IDF; the other half were not. No significant difference was found between the percentages of items that users clicked for examination in the two kinds of retrieval (just over 6% in each). This finding is carefully hedged, owing to characteristics of Docear and choices made in the study design.

The systems just described seek answers to the question: Which weights produce papers that are most related to the input and thus best to recommend? The present paper

has, to repeat, a different emphasis: Since all the works in the retrieval are related to the input, how does the relationship change as the weights change, and how might this be exploited? In this view, for example, the "highly cited methodological paper" downplayed in [13: 70] is not simply a candidate for burial; it, too, might have value for the user.

3 First example: the berrypicking paper

Copied with light editing from Dialog output, Table 1 shows four lines of raw data in which the seed was an influential 1989 paper by Marcia J. Bates, "The design of browsing and berrypicking techniques for the online search interface" [19]. Commands not shown formed Set A—the set of all documents directly citing the Bates paper in the online Social Sciences Citation Index (SSCI, File 7). Dialog's RANK command was used (with the DETAIL option) to form Set B—the cited references (CR's) co-cited with the seed by at least three documents (an arbitrary threshold) in Set A. Some 706 such references were retrieved as types. Their tokens in the CR field numbered 11,550.

Under "Term" in Table 1 are truncated strings identifying these references, with Bates at top. Under "Items Ranked" is the co-citation count of each of the strings with the seed. Under "Items in File" is the overall citation count for each of the strings in the database. Again, the co-citation counts become the TF factor, and the citation counts become the IDF factor. For seeds, the two counts are generally identical. The *N* in the IDF factor for the SSCI in 2013 was estimated at three million records.

Bates 1989 is actually cited in 279 documents in Set A, but the commonest string identifying it is cited in 264, and so that count is used here for simplicity. The other strings are minor variants cited at most a few times each. Fragmented ID strings that affect counts are a long-standing problem in citation databases.

Table 2 displays some calculations for high-end and low-end Bates documents, Here, the top TF*IDF weights do not much alter the ranking produced by the raw TF counts, but large changes in rank can occur, as will be briefly taken up later.

Ranked by their weights, the retrieved works form a log-normal distribution. A normal-distribution scale may thus be used to identify groups for comparison. Weights converted to z-scores, as in Table 2, place some works at the distribution's midpoint or in one of its tails. The titles of works from those positions will be compared in specimen retrievals below.

In her 1989 paper, Bates argued that online search interfaces could be improved by replacing the dominant model of literature searching with a more realistic one. In the dominant model, searchers express an information need with a single query, submit it to a database, and retrieve a single set of documents to be judged for relevance. In the Bates model, searchers behave more like someone picking berries in a forest: "Typical search queries are not static, but rather evolve. Searchers commonly gather information in bits and pieces instead of one grand best retrieved set. Searchers use a wide variety of search techniques, which extend beyond

Table 1 Sample raw data from a citation file on Dialog Classic

```
DIALOG RANK Results  (Detailed Display)
----------------------------------------
RANK: S4/1-279   Field: CR=  File(s): 7

RANK No  Items in File  Items Ranked    Term
-------  -------------  ------------    ----
    1          264          264          BATES MJ, 1989, V13,-
    2          203           61          ELLIS D, 1989, V45, -
    3          357           60          KUHLTHAU CC, 1991, V-
    4          274           53          BELKIN NJ, 1982, V38-
```

Table 2 TF*IDF ranking of top 3 and bottom 3 works co-cited with Bates (1989)

Strings identifying works in SSCI	TF	DF	Log TF	Log IDF	TF* IDF	*Z* score
BATES MJ, 1989, V13, P407, ONLINE REV	264	264	3.42	4.06	13.88	5.11
ELLIS D, 1989, V45, P171, J DOC	61	203	2.79	4.17	11.61	3.12
BATES MJ, 1990, V26, P575, INFORM PROCESS MANA	31	94	2.49	4.50	11.22	2.78
BELKIN NJ, 1982, V38, P61, J DOC	53	274	2.72	4.04	11.00	2.58
LINCOLN YS, 1985, NATURALISTIC INQUIRY	4	6023	1.60	2.70	4.32	−3.30
LAVE J, 1991, SITUATED LEARNING LE	3	4555	1.48	2.82	4.16	−3.44
KUHN TS, 1970, STRUCTURE SCI REVOLU	3	5680	1.48	2.72	4.02	−3.56

those commonly associated with bibliographic databases. Searchers use a wide variety of sources other than bibliographic databases" [19: 214].

The relative ease of relating high-ranked works to Bates 1989 can be seen in Table 3, where the top 12 items are spelled out as titles. (Italicized titles are books.) Word counts from the titles indicate what the Bates paper connotes: 12, *information*; 7, *retrieval*; 4, *design, search, seeking*; 3, *interface, interact-*; 2, *browsing, online, user*. A researcher familiar with this area could discern in them a coherent theme—something like "psychological and behavioral factors in designing user-oriented interfaces for online document retrieval." The titles express the theme with considerable variety, but that is a recurrent feature of co-citation retrieval, which captures citers' implicit understanding of connections in ways that keyword matching and expansion do not. Co-citation ties may also cause thematically salient authors to recur. For example, Table 3 has two more papers by Bates and three by Nicholas J. Belkin. Another indicator of consistency is that the papers in Table 3 were all published in library and information science journals.

The mid-ranked titles in Table 3 (median TF*IDF = 8.06) are also all contributions to library and information science, and two authors from the top 12—Ellis and Bates herself—reappear here. Even so, TF*IDF predicts them to be less relevant to the seed. The titles are topically more miscellaneous, and they bring out the thematic connotations of the seed less clearly. While several might furnish examples pertinent to online interface design (e.g., Stoan, Case, Larson, Covi, Chen, Järvelin), that is not their main focus.

The mid-ranked titles also mix two schools of IR: the user oriented and the system oriented. TF*IDF here captures citers' intuitions that system-oriented works are somewhat less relevant to the berrypicking paper than user-oriented works. Like the seed, the 12 top-ranked items all exemplify the user-oriented school, but only some of the 12 mid-ranked items do. The two schools have many points of contact, but differ in their central concerns and methods. User-oriented studies (e.g., Case, Covi, Chen, Järvelin) tend to discuss search behavior and techniques in a philosophical, qualitative way, even when interviews or surveys of people are involved. In contrast, system-oriented studies (e.g., Beaulieu, Ellis 1996, Maron, Tague-Sutcliffe) tend toward depersonalized topics, such as search algorithms, evaluation measures for retrieval experiments, mathematical properties of indexing systems, and the like. A comparable effect will be seen in the next section.

The 12 titles at the low end of the Bates distribution are prominent theoretical or methodological items, mostly books, that are relevant to many research specialties. Some are from sociology, and others tell how to do qualitative research; none is from information science. It is here that BoW retrieval most clearly departs from what is customary in paradigmatic IR. It is hard to imagine assessors in TREC experiments marking any of the works in Table 3 as relevant to the berrypicking paper (assuming they were presented). Yet each has been co-cited with it at least three times. On that ground, a researcher or teacher examining Bates's intellectual world might find them valuable—perhaps even more so than closely similar works.

Authors of seed papers are themselves candidates for such information. To illustrate, Marcia Bates read the present paper in an earlier draft and wrote: "I think someone studying the intellectual development of a field could use your approach to great effect. I find the end-of-the-list co-cited papers to be a really intriguing set. First, it says something about what kind of research/philosophical point of view co-exists with my writing. Also, though there is some overlap in the thinking among the writers, they represent some significant differences in philosophy that make them possibly distinct theory streams." She goes on to speculate why various end-of-the-list works appear, concluding that it is "not accidental that most of the last items are methodological" [personal communication, February 2016].

In another paper [20], two historians comment on retrievals co-cited with seeds they themselves supplied. They found the retrievals to be readily intelligible and could see a place for them in humanities scholarship.

4 Second example: an IR textbook

Table 4 has specimen titles from the distribution of works co-cited with *An introduction to information retrieval* [4], a standard textbook by Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. (Henceforth, "Manning.") They show once more how TF*IDF weighting elevates titles thematically close to the seed, and how this thematization effect lessens as the ranks descend.

Works co-cited with Manning were again retrieved in SSCI on Dialog Classic in 2013. The seed work happened to be designated by five variant strings, whose counts this time were combined. Set A, the works citing Manning, numbered 527. Set B, the items co-cited with the textbook at least three times, came to 592.

Manning may be a somewhat unusual seed, given its broad title and coverage. But as a major textbook, it could be chosen to represent IR in general, in hopes of retrieving other works that organize or are important to the field. Significantly, it does retrieve both editions of the field's other principal textbook, *Modern information retrieval.* Additional top titles (e.g., Salton, Blei, Deerwester) convey foundational or contemporary topics prominent in IR. The majority of the titles share terms with four of Manning's chapter-headings (italicized):

Table 3 Top, middle, and bottom 12 titles co-cited with Bates (1989) as seed

TF*IDF	Sole or first author, year, and title of co-cited work
13.88	BATES MJ, 1989, The design of browsing and berrypicking techniques for the online search interface
11.61	ELLIS D, 1989, A behavioral approach to information retrieval design
11.22	BATES MJ, 1990, Where should the person stop and the information search interface start?
11.00	BELKIN NJ, 1982, ASK [anomalous states of knowledge] for information retrieval Part 1
10.90	KUHLTHAU CC, 1991, Inside the search process: Information seeking from the user's perspective
10.88	BELKIN NJ, 1995, Cases, scripts and information seeking strategies: Design of interactive information retrieval systems
10.84	MARCHIONINI G, 1995, *Information Seeking in Electronic Environments*
10.75	BELKIN NJ, 1993, BRAQUE: Design of an interface to support user interaction in information retrieval
10.68	COVE JF, 1988, Online text retrieval via browsing
10.66	BATES MJ, 1979, Information search tactics
10.57	INGWERSEN P, 1992, *Information Retrieval Interaction*
10.54	BELKIN NJ, 1980, Anomalous states of knowledge as a basis for information retrieval
10.47	TAYLOR RS, 1968, Question negotiation and information seeking in libraries
8.08	HARTLEY RJ, 1990, *Online Searching: Principles and Practice*
8.08	STOAN SK, 1984, Research and library skills: An analysis and interpretation
8.06	CASE DO, 1991, Conceptual organization and retrieval of text by historians: The role of memory and metaphor
8.06	BEAULIEU M, 1996, Evaluating interactive systems in TREC
8.06	ELLIS D, 1996, The dilemma of measurement in information retrieval research
8.06	LARSON RR, 1991, Between Scylla and Charybdis: Subject searching in the online catalog
8.06	COVI LM, 1999, Material mastery: Situating digital library use in university research practices
8.04	MARON ME, 1960, On relevance, probabilistic indexing and information retrieval
8.03	BATES MJ, 1998, Indexing and access for digital libraries and the Internet: Human, database, and domain factors
8.03	CHEN HC, 1990, User misconceptions of information retrieval systems
8.03	JARVELIN K, 2004, Information seeking research needs extension toward tasks and technology
8.03	TAGUE-SUTCLIFFE J, 1995, *Measuring Information: An Information Services Perspective*
4.90	DAVIS FD, 1989, Perceived usefulness, perceived ease of use, and user acceptance of information technology
4.87	GLASER BG, 1967, *The Discovery of Grounded Theory*
4.87	SIMON HA, 1955, A behavioral model of rational choice
4.85	PUTNAM RD, 1995, *Bowling Alone: America's Declining Social Capital*
4.80	STRAUSS A, 1998, *Basics of Qualitative Research*
4.74	GRANOVETTER MS, 1973, The strength of weak ties
4.73	GIDDENS A, 1984, *The Constitution of Society: Outline of the Theory of Structuration*
4.67	GARFINKEL H, 1967, *Studies in Ethnomethodology*
4.62	PATTON MQ, 1990, *Qualitative Evaluation and Research Methods*
4.32	LINCOLN YS, 1985, *Naturalistic Inquiry*
4.16	LAVE J, 1991, *Situated Learning: Legitimate Peripheral Participation*
4.02	KUHN TS, 1970, *The Structure of Scientific Revolutions*

Table 4 Top, middle, and bottom 12 titles co-cited with Manning et al. (2008) as seed

TF*IDF	Sole or first author, year, and title of co-cited work
15.90	MANNING CD, 2008, *Introduction to Information Retrieval*
11.02	CHIRITA PA, 2007, Personalized query expansion for the Web
10.89	BAEZA-YATES R, 2011, *Modern Information Retrieval: The Concepts and Technology behind Search* [2nd ed]
10.82	BAEZA-YATES R, 1999, *Modern Information Retrieval*
10.58	SALTON G, 1975, A vector space model for automatic indexing
10.55	ZHAI CX, 2004, A study of smoothing methods for language models applied to information retrieval
10.53	SALTON G, 1988, Term-weighting approaches in automatic text retrieval
10.47	PORTER MF, 1980, An algorithm for suffix stripping
10.47	BLEI DM, 2003, Latent Dirichlet allocation
10.46	SUN R, 2006, Mining dependency relations for query expansion in passage retrieval
10.35	PONTE JM, 1998, A language-modeling approach to information retrieval
10.30	DEERWESTER S, 1990, Indexing by latent semantic analysis
10.19	LEE KS, 2008, A cluster-based resampling method for pseudo-relevance feedback
8.12	HEARST MA, 2006, Clustering versus faceted categories for information exploration
8.12	APHINYANAPHONGS Y, 2005, Text categorization models for high-quality article retrieval in internal medicine
8.12	FIDEL R, 2004, The many faces of *accessibility:* Engineers' perception of information sources
8.12	EUZENAT J, 2007, *Ontology Matching*
8.10	DAMASHEK M, 1995, Gauging similarity with n-grams: Language-independent categorization of text
8.10	GERSTBERGER PG, 1968, Criteria used by research and development engineers in the selection of an information source
8.10	CROFT WB, 2003, *Language Modeling for Information Retrieval*
8.10	JANSEN BJ, 2008, Determining the informational, navigational, and transactional intent of Web queries
8.09	JELINEK F, 1980, Interpolated estimation of Markov source parameters from sparse data
8.08	SHI JB, 2000, Normalized cuts and image segmentation
8.06	JOACHIMS T, 1999, Making large-scale support vector machine learning practical [book chapter]
8.06	FOUSS F, 2007, Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation
5.26	KASS RE, 1995, Bayes factors
5.26	CORTES C, 1995, Support vector networks
5.18	BARABASI AL, 1999, Emergence of scaling in random networks
5.14	NEWMAN MEJ, 2003, The structure and function of complex networks
5.08	LANDIS JR, 1977, The measurement of observer agreement for categorical data
5.04	PEARL J, 1988, *Probabilistic Reasoning in Intelligent Systems*
4.89	ALBERT R, 2002, Statistical mechanics of complex networks
4.72	SCHWARZ G, 1978, Estimating the dimensions of a model
4.54	ZADEH LA, 1965, Fuzzy sets
4.40	PRESS WH, 1992 *Numerical Recipes in C, the Art of Scientific Computing*
3.95	ALTSCHUL SF, 1990, Basic local alignment search tool
3.71	ALTSCHUL SF, 1997, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs

- Salton's two papers with *Scoring, term weighting and the vector space model*
- Chirita, Sun, and Lee with *Relevance feedback and query expansion*
- Zhai and Pointe with *Language models for information retrieval*
- Deerwester with *Matrix decompositions and latent semantic indexing*

Although their titles do not match chapter-headings, Porter is discussed in the subchapter on stemming, and Blei, in a note to the chapter on language modeling. In fact, Manning's bibliography includes all the top-ranked titles except Chirita, Sun, and Lee (counting Baeza-Yates as one work and a 1975 book by Salton as incorporating his 1975 paper).

The 12 mid-ranked titles (median TF*IDF = 8.10) match the Manning chapter-headings less well than those at the top, and their implied topics are again more miscellaneous. Only three are among Manning's own references (Joachims, Hearst, Croft). Some are easy to relate to the seed (Aphinyanaphongs, Croft); others are less so (Jelinek, Shi). A number of titles are mid-ranked because their IDF weights have been lowered by relatively high citation counts. They may be both highly co-cited with the seed and well-cited and well-known themselves (e.g., Joachims). In contrast, Chirita, Sun, and Lee are top-ranked with the Manning textbook because they are highly co-cited with it but not highly cited otherwise; thus they relate to it very specifically.

Manning represents system-oriented IR as it is construed in computer science, with its emphasis on algorithms, mathematics, and retrieval system measurement. All the top-ranked works in Table 4 belong to this school. However, as with the Bates example, the mid-ranked titles include items from both schools of IR, except this time it is items from the *user-oriented* school (Fidel, Gerstberger, Jensen) that are pushed down. The worlds of Bates and Manning are each explicitly characterized as "information retrieval," yet to someone with domain knowledge they are fairly distinct, and again blind TF*IDF weighting of citation data nicely differentiates their intellectual styles and concerns.

The bottom-ranked works in Table 4, like those in Table 3, extend well beyond the seed in scope (Kass, Cortes, Schwarz, Zadeh, Press). Having been co-cited with Manning from three to five times, they are manifestly relevant to IR, but they connect with it through content-neutral ideas that can be employed in a wide range of quantitative fields. As evidence of the latter, all of them have citation counts in the thousands or tens of thousands. Several by physicists (Barabási, Newman, Albert) represent the importance of mathematical network science as it applies to IR, especially to Web retrieval techniques.

5 Third example: a paper on centrality

In Table 5, BoW retrieval produces a distribution with features like those in the first two examples, although a different discipline, sociology, is involved. The seed is a classic 1987 contribution to the literature on social networks, Phillip Bonacich's "Power and centrality: A family of measures" [21]. Its abstract is brief:

"Although network centrality is generally assumed to produce power, recent research shows that this is not the case in exchange networks. This paper proposes a generalization of the concept of centrality that accounts for both the usual positive relationship between power and centrality and Cook et al.'s recent exceptional results."

Many information scientists will be familiar with measures such as degree centrality, betweenness centrality, and closeness centrality. All were originally developed to analyze how patterns of ties among persons—more generally, "actors"— in networks relate to properties such as status or power. In 1972 Bonacich defined eigenvector centrality [22], in which an actor's own centrality (and power) in a network is weighted by the centrality (and power) of others to whom that actor has ties. The measure is positive because every actor is assumed to have at least some power. PageRank in IR is a similar measure. The novelty of the 1987 seed paper is that it introduces "beta-centrality," an extension of the eigenvector measure to accommodate actors with negative weights—that is, situations in which an actor's power is increased by having ties to *powerless* actors. The paper Bonacich refers to in his abstract, Cook 1983 in Table 5, describes situations of that sort.

Works co-cited with the seed were once more drawn from SSCI on Dialog Classic in 2013. Set A consisted of the 416 articles or other pieces in journals that cited Bonacich 1987. Set B consisted of 554 works that were co-cited with it at least five times (another arbitrary threshold).

Titles of the 12 top-ranked works amplify the theme set by the seed: six mention *networks*; five, *centrality*; four, *status* or *power*; three, *markets*. Certain terms suggest quantitative research methods: *measures, factoring, weighting, scores, index, experimental results*. Other terms suggest social inequality: *asymmetric relations, key player*. Papers by Freeman and Katz that established earlier, complementary measures of centrality are drawn in, and so are Cook 1983 and Stephenson's account of information-based centrality. As in the Bates example, authors recur in the top 12: There are two other papers by Bonacich himself and two papers by Podolny that apply his 1987 measure to markets.

The 12 mid-ranked works (median TF*IDF = 7.98) continue the network theme but do not mention centrality in their titles. Many of them apply network theory to businesses, professions, and organizations. Two (Bollen, Newman 2001) apply it to bibliometrics. Such practical applications would

Table 5 Top, middle, and bottom 12 titles co-cited with Bonacich (1987) as seed

TF*IDF	Sole or first author, year, and title of co-cited work
13.96	BONACICH P, 1987, Power and centrality: A family of measures
11.77	BONACICH P, 1972, Factoring and weighting approaches to status scores and clique identification
11.77	FRIEDKIN NE, 1991, Theoretical foundations for centrality measures
11.46	STEPHENSON K, 1989, Rethinking centrality: Methods and examples
11.35	BONACICH P, 2001, Eigenvector-like measures of centrality for asymmetric relations
11.06	PODOLNY JM, 2001, Networks as the pipes and prisms of the market
10.99	JENSEN M, 2003, The role of network resources in market entry: Commercial banks' entry into investment banking, 1991-1997
10.98	FREEMAN LC, 1979, Centrality in social networks: Conceptual clarification
10.90	BALLESTER C, 2006, Who's who in networks. Wanted: The key player
10.83	PODOLNY JM, 1993, A status-based model of market competition
10.73	COOK KS, 1983, The distribution of power in exchange networks: Theory and experimental results
10.73	KATZ L, 1953, A new status index derived from sociometric analysis
10.71	BORGATTI SP, 2005, Centrality and network flow
8.04	BAUM JAC, 1992, Institutional embeddedness and the dynamics of organizational populations
8.02	MARIOLIS P, 1975, Interlocking directorates and control of corporations: The theory of bank control
8.01	BOLLEN J, 2006, Journal status
8.00	GALASKIEWICZ J, 1985, Professional networks and the institutionalization of a single mind set
7.98	MOODY J, 2005, Dynamic network visualization
7.98	NEWMAN MEJ, 2003, Mixing patterns in networks
7.98	NEWMAN MEJ, 2001, The structure of scientific collaboration networks
7.98	MEHRA A, 2001, The social networks of high and low self-monitors: Implications for workplace performance
7.98	SCHILLING MA, 2007, Interfirm collaboration networks: The impact of large-scale network structure on firm innovation
7.97	HAGEDOORN J, 1994, The effect of strategic technology alliances on company performance
7.96	GORMAN M, 1989, What do venture capitalists do?
5.23	ROGERS EM, 2003, *Diffusion of Innovations* [5th ed]
5.15	FESTINGER L, 1954, A theory of social comparison processes
5.13	HEIDER F, 1958, *The Psychology of Interpersonal Relations*
5.06	PORTER ME, 1980, *Competitive Strategy: Techniques for Analyzing Industry and Competitors*
5.01	GIDDENS A, 1984, *The Constitution of Society: Outline of the Theory of Structuration*
4.81	COHEN J, 1983, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* [2d ed]
4.65	AIKEN LS, 1991, *Multiple Regression: Testing and Interpreting Interactions*
4.63	JENSEN MC, 1976, Theory of the firm: Managerial behavior, agency costs and ownership structure
4.60	WHITE H, 1980, A heteroscedasticity consistent covariance matrix estimator and a direct test for heteroscedasticity
4.45	KAHNEMAN D, 1979, Prospect theory: An analysis of decision under risk
4.02	RADLOFF L, 1977, The CES-D scale: A self-report depression scale for research in the general population
3.89	BARON RM, 1986, The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations

probably interest certain researchers—e.g., writers of literature reviews—even though they are well down in the distribution.

The items bottom-ranked here were co-cited with Bonacich no less than five to eight times, but, as in the Bates and Manning examples, they are no longer directly relatable to the seed through their titles. Six are books; all are works of disciplinary or multidisciplinary sweep. Expositions of research methods and statistical textbooks appear with theoretical treatises. The Giddens book is so widely used for its theories that it also turns up among the bottom 12 works co-cited with Bates.

Earlier it was said that items top-ranked by TF*IDF weights may differ from items top-ranked in the original data. This is quite common (see, e.g., [10]). In Table 5, only five of the top titles duplicate those in the top 12 of Dialog Classic, which ranks items by their raw co-citation counts with the seed. What has happened is that some works with raw citation counts in the thousands—e.g., Wasserman and Faust's *Social network analysis*, Burt's *Structural holes*, and Granovetter's 1973 article "The strength of weak ties"—have been automatically demoted by IDF. At the same time, IDF has promoted to the top certain less famous works that deal specifically with network centrality measures. The famous works are not lost; they are just no longer among the first items to meet the eye in a ranked list.

It would be good in bag of works retrieval if users could quickly discover works whose co-citation counts and citation counts are both relatively high. That would combine predicted relevance to the seed with an indicator of breadth of influence and use. Toward this end, several papers by White [10,20,23–26] have introduced a graphic display called a pennant diagram. These diagrams plot TF and IDF values separately on two axes, so that each work appears as a point in two-dimensional space. The overall shape of the plot resembles a pennant with the seed at the tip at right. Increasingly relevant items are moved rightward toward the seed; increasingly influential and widely used (but less specific) items are moved downward. Unfortunately, even short labels for extensive distributions of works, such as those discussed here, greatly overpack standard display space. Pennants serve best when relatively few points are involved. But it is not hard to imagine some of their information presented as text. BoW distributions could also be partitioned so as to respond to individual interests. For example, weighted titles might be exhibited for particular authors, journals, or periods of time.

6 Discussion: possible users and uses

It seems an unwritten rule in IR that knowledge of works should not be presumed. The default assumption is that users will represent their interests through topical terms because that is what they routinely submit. Using a document as one's search term requires domain knowledge of the sort possessed only by certain text-oriented scientists and scholars. It moreover assumes familiarity with the conventions of citation databases, which even learned researchers may lack. Note, then, that topical terms can function just like works in retrieving co-cited items. For example, one or more topical terms can retrieve Set A as full records from WoS; from those, software external to WoS can extract Set B. That is how data for maps of co-cited works or authors are now generated. Yet it may still be the case that:

- The user can represent an interest through at least one seed document in addition to topical terms. Many thousands of people have enough domain expertise to do this and thus might find uses for BoW retrievals.
- The user can represent an interest *only* through one or more seed documents. Suppose, for instance, one wants to explore Bates's berrypicking idea at length; how can her metaphor be transferred to nonmetaphorical contexts? With BoW retrieval, the question answers itself, as the titles in Table 3 show.
- The user's interest is the seed document itself. Here, the user is not doing a conventional literature search but seeking information on *the seed document's use by citers over time*. This possibility differs strikingly from the model of users in paradigmatic IR and, once again, BoW retrieval is pertinent.

Paradigmatic IR systems are designed for users who know "needs" rather than documents, and whose needs are met mainly by documents hitherto unknown. This design accommodates both nonscientists and scientists who read primarily to have their questions answered and not because of their interest in documents as texts *per se.* As Bates [27] points out, the typical scientist wants to keep up with relevant research findings but frequently does so through an interpersonal network well before they are published. The actual literature is regarded as archival, and many contributions to it may go unread. In marked contrast, the typical humanities scholar's research is centered on texts as ends in themselves, to be mastered in all their unique particulars. Bates's data show that humanists already know the literature in their specialties so well that they are surprised if a literature search turns up even a few new items. However, BoW retrievals for such persons could reveal something new: how masses of citers have received and contextualized known works.

Take, for example, Virginia Woolf's *Mrs. Dalloway* as a seed in Arts and Humanities Citation Index. One might expect that the items top-ranked with it would be studies of Woolf and of that novel. Not so; down much of the distribution, the majority of items are writings by Woolf herself. (The same is true of another Woolf novel, *Orlando.*) The

items pushed to lower ranks by the IDF factor include such "co-studied" works as *Ulysses, The Sound and the Fury,* and *The Waste Land.* Obviously, the relevance of these works to the seed is not topical, but part of the history of scholarship on it. BoW retrieval thus in a small way supports intellectual history.

In this regard, BoW retrieval bears on citation-based domain analysis. Domain analysts can often name one or more documents that initiated a particular line of research. Given well-chosen "foundational" seeds, Set A and Set B are both significant portrayals of a domain. Set A may contain one or more of the domain's research fronts—clusters of relatively recent documents that define emerging research areas. Set B, which includes the seed, is the domain's intellectual base—older documents that have proved widely useful within a particular paradigm [28]. So bag of works retrieval can in some cases also be understood as *intellectual base* retrieval. Because every document in Set B is ranked for relevance to the seed, thresholds can be set for extracting the most important documents in the base, as evidenced by their citedness. Most researchers are probably unaware that retrievals involving *intellectual base* documents as against *research front* documents are possible. But that is because they have not seen the idea implemented as a well-described option in a retrieval system.

References

1. Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L.: Interpreting tf-idf term weights as making relevance decisions. ACM Trans. Inf. Syst. **26**(3), 13 (2008)
2. Huang, W., Kataria, S., Caragea, C., Mitra, P., Giles, C.L., Rokach, L.: Recommending citations: translating papers into references. In: Proceedings of the 21st International Conference on Information and Knowledge Management, pp. 1910–1914 (2012)
3. Nascimento, C., Laender, A.H.F., da Silva, A.S., Gonçalves, M.A.: A source independent framework for research paper recommendation. In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital libraries, pp. 297–306 (2011)
4. Manning, C.D., Raghavan, P., Schütze, H.: An Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
5. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
6. Eto, M.: Evaluations of context-based co-citation searching. Scientometrics **94**, 651–673 (2013)
7. Liu, S., Chen, C.: The proximity of co-citation. Scientometrics **91**, 495–511 (2012)
8. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. **28**, 11–21 (1972)
9. Carevic, Z., Schaer, P.: On the connection between citation-based and topical relevance ranking: Results of a pretest using iSearch. In: Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval, pp. 37–44 (2014)
10. White, H.D.: Some new tests of relevance theory in information science. Scientometrics **83**, 653–667 (2010)
11. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research paper recommender systems: a literature survey. Int. J. Digit. Libr. **17**(4), 305–338 (2016)
12. Small, H.: Co-citation in the scientific literature: a new measure of the relationship between two documents. J. Am. Soc. Inf. Sci. **24**, 265–269 (1973)
13. Lawrence, S., Giles, C.L., Bollacker, K.: Digital libraries and autonomous citation indexing. IEEE Comput. **32**(6), 67–71 (1999)
14. Huynh, T., Hoang, K., Do, L., Tran, H., Luong, H., Gauch, S.: Scientific publication recommendations based on collaborative citation networks. In: Proceedings of the International Conference on Collaboration Technologies and Systems (CTS), pp. 316–321 (2012)
15. Liang, Y., Li, Q., Qian, T.: Finding relevant papers based on citation relations. In: Wang, H., Li, S., Oyama, S., Hu, X., Qian, T. (eds.) Lecture Notes on Computer Science, vol. 6897, pp. 403–414 (2011)
16. Küçüktunç, O., Saule, E., Kaya, K., Çatalyürek, U.V.: Towards a personalized, scalable, and exploratory academic recommendation service. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 636–641 (2013)
17. Pan, L., Dai, X., Huang, S., Chen, J.: Academic paper recommendation based on heterogeneous graph. In: Sun, M., Liu, Z., Zhang, M., Liu, Y. (eds.) Lecture Notes on Computer Science, vol. 9427, pp. 381–392 (2015)
18. Beel, J., Breitinger, C., Langer, S.: Evaluating the CC-IDF citation-weighting scheme: how effectively can 'Inverse Document Frequency' (IDF) be applied to references? In: Proceedings of the 12th iConference (in press) (2017)
19. Bates, M.J.: The design of browsing and berrypicking techniques for the online search interface. Online Review 13: 407–424 [Quoted as reprinted in her (2016) Information users and information system design. Ketchikan Press, Berkeley, California, pp 195–216] (1989)
20. White, H.D.: Co-cited author retrieval and relevance theory: examples from the humanities. Scientometrics **102**, 2275–2299 (2014)
21. Bonacich, P.: Power and centrality: a family of measures. Am. J. Sociol. **92**, 1170–1182 (1987)
22. Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. J. Math. Sociol. **2**, 113–120 (1972)
23. White, H.D.: Combining bibliometrics, information retrieval, and relevance theory, part 1: first examples of a synthesis. J. Am. Soc. Inf. Sci. Technol. **58**, 536–559 (2007)
24. White, H.D.: Combining bibliometrics, information retrieval, and relevance theory, part 2: some implications for information science. J. Am. Soc. Inf. Sci. Technol. **58**, 583–605 (2007)
25. White, H.D.: Pennants for Strindberg and Persson. In: Celebrating Scholarly Communication Studies: A Festschrift for Olle Persson at his 60th Birthday, pp. 71–83 (2009). http://www.issi-society.org/ollepersson60/
26. White, H.D., Mayr, P.: Pennants for descriptors. Paper presented at the 12th International Conference on Theory and Practice of Digital Libraries. arXiv:1310.3808 (2013)
27. Bates, M.J.: Document familiarity, relevance, and Bradford's Law: the Getty Online Searching Project report no. 5. Information Processing & Management **32**, 697–707 [Reprinted in her (2016) Information users and information system design. Ketchikan Press, Berkeley, California, pp. 283–300], (1996)
28. Jarneving, B.: A comparison of two bibliometric methods for mapping of the research front. Scientometrics **65**, 245–263 (2005)