Bag of Works Retrieval: TF*IDF Weighting of Co-cited Works

Howard D. White

College of Computing and Informatics
Drexel University, Philadelphia PA, USA
whitehd@drexel.edu

Abstract

Although it is 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 novel way that could be implemented in citation-enhanced digital libraries of the future. Rather than entering keywords, the user enters a string identifying a work, called a seed, to retrieve the strings identifying other works that are co-cited with the seed. 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. The result is analogous to, but different from, traditional "bag of words" retrieval. Certain properties of the ranking are illustrated with the top and bottom items co-cited with a classic paper by Marcia J. Bates, "The design of browsing and berrypicking techniques for the online search interface." However, the 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

Bag of Works Retrieval: TF*IDF Weighting of Co-cited Works

Howard D. White

College of Computing and Informatics
Drexel University, Philadelphia PA, USA

1 Introduction

One way of expressing an interest or a question to an information retrieval system is to name a document that implies it. The implicit request is "Find-similar" or "Get me more like this" (Smucker 2008). The idea is not new: for decades, selective dissemination of information services have accepted customer profiles that contain not only subject terms (e.g., descriptors or natural-language keywords), but also names of documents already known to be of interest. The idea is usually to enter these names into citation indexes to retrieve the items that cite them, which users may also find relevant. Thus, instead of words indicating a desired subject matter, the queries are strings denoting works. The option of starting searches with works as seeds is featured in the major citation indexes, Web of Science (WoS), Scopus, and Google Scholar (GS).

In Cited Reference searches in the Web of Science, for example, strings that denote cited articles, books, and other publications can be both entered and retrieved. The following is typical:

WHITE HD 2004 APPL LINGUIST V25 P89

If strings like this are retrieved (now often followed by an analogous DOI string), they must of course be spelled out as full references. But for present purposes it is enough that such strings could function as seeds in WoS or similar databases (as could 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.

The present paper sketches a way of retrieving documents with a bag of works model as an alternative to the bag of words model. It involves seed documents, a co-citation relevance metric, and a standard version of TF*IDF weighting (Manning and Schütze 1999: 544), in which logged term frequencies (TF) are multiplied by logged inverse document frequencies (IDF). In principle, bag of works retrieval could be implemented in any digital library that has the appropriate software and data. Some bibliographic databases list the items cited by a publication as part of that publication's record. The bag of works in such databases would be the total set of strings that identify cited works; like keywords, these strings are terms that index the citing documents. Thus, the proposal here could become a regular option.

TF: Term frequencies in this case are counts of documents co-cited with the seed document in later papers: 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 like the string above can retrieve the other 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 non-stopped words in the database progressively lower as the number of documents containing them increases, because words used very frequently 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 scores are the total citation counts for documents in the database. The higher the DF count, the more well-known and widely used a document is, and the greater its breadth of implication and general applicability. IDF, which inverts the DF count, favors works that are *narrowly and specifically* related to the seed over widely used works that are more broadly and generally related.

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

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

Weighted in this way, a bag of works retrieval differs in important respects from typical retrievals in IR. Its properties include:

- *All* retrieved items are relevant to the seed in varying degrees by empirical co-citation evidence. Such evidence from multiple co-citing authors is stronger than the usual gold standard for relevance judgments, the verdict of a single assessor. Thus, *any* item is of potential interest to a domain-literate user.
- Retrieved items may be topically similar to the seed, but need not be.
- Since seeds merely imply topical content, their semantic relations with retrieved items will be more various and less predictable than those obtained by algorithmic word-matching or query-expansion based on it. Yet when spelled out as full references, all retrievals have the following broadly predictable structure (White 2010, 2011):
- A substantial segment of top-ranked items will be easy to relate to the seed in global topic (or sometimes in authorship).
- The relevance of items to the seed in global topic will be progressively less easy to see over the whole retrieval, as evidenced by the decreasing coherence of content indicators such as terms from titles and abstracts.
- A substantial segment of bottom-ranked items (those with the lowest TF*IDF weights) will be relatively difficult to relate to the seed's global topic at first glance because of their generality.

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 GS) or by recency of publication (in WoS 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 it.

By contrast, the documents *co-cited* with the seed *can* be ranked for relevance to it, because their co-citation counts vary and can be treated as relevance scores. This requires the further step of retrieving the co-cited documents as Set B. Suppose the seed is the 1990 book edited by Christine Borgman, *Scholarly Communication and*

Bibliometrics, and that it is cited in an article by Olle Persson in Set A. When the book is paired with each of the nine other items in Persson's references, each *pair* has a co-citation count of one. But over all the papers in Set A, many of these pairs would be co-cited more than once. For example, at this writing the M. M. Kessler paper that introduced bibliographic coupling in 1963 has a count of seven with Borgman's book, because seven documents in Set A have cited both it and the book in their references. It is these varying co-citation counts that are plugged into the TF factor of the TF*IDF formula in bag of works retrieval. In this case, the formula would be used to rank the relevance of documents in Set B to *Scholarly Communication and Bibliometrics*.

Paradigmatic IR researchers have delved into co-citation retrieval rather seldom. Birger Larsen, who reviewed the matter in his dissertation (2004: 49-50), concluded: "Although relatively straightforward to carry out online as demonstrated, e.g., by Chapman and Subramanyam (1981) co-citation search...does not seem to have received much attention for retrieval. Instead co-citation has been used extensively for mapping the structure of research fields..." Since he wrote, there has not been a great deal of change. Insofar as cited references are used in IR, the tendency is to use the direct citation relationship in query expansion to augment topical retrievals. The co-citation relationship does make an appearance in proposed systems for recommending papers to cite (McNee et al. 2002, Strohman et al. 2006, Huang et al. 2012, Beel et al. 2015), since acts of co-citation leave traces like those exploited in better-known recommender systems, such as co-purchasing in Amazon or co-renting in Netflix.

With respect to operational systems, CiteSeer^{x} automatically returns a small (and opaque) selection of the titles co-cited with a seed document, but it is the exception. In the Web of Science, Scopus, and Google Scholar, no co-citation retrievals of any kind are possible. For 20 years they could be carried out in the Thomson Reuters databases on DialogClassic, but that service has been defunct since 2013. Ironically, Thomson Reuters created what is now the Web of Science in the home of co-citation analysis (ISI, the Institute for Scientific Information), yet the Basic Search panel in WoS is designed mainly for retrievals by topic, author, journal, or characteristics of a single work. The secondary Cited Reference Search panel in WoS is designed to take authors or single works as input and find the items that have cited them. These capabilities are indispensable, of course, but valuable possibilities remain.

2 Example

Carevic and Schaer (2014) used the iSearch test collection in physics to experiment with bag of works retrieval as presented in White (2010). In iSearch, documents come with both cited references and assessors' relevance ratings on a four-point scale. The authors were looking for overlaps between the documents pre-scored by assessors as relevant to a topic and the documents retrieved by TF*IDF-weighted co-citation. This proved not feasible because the co-citation counts they found in iSearch were small. But in examples from two search topics, the top-ranked co-cited documents did cohere with seed documents in their title terms. The present paper further illustrates bag of works retrieval with more robust co-citation data gathered in 2013 from Thomson Reuters citation databases on DialogClassic. The intent is not to evaluate the method, but merely to present some aspects of TF*IDF-weighted co-citation not covered in Carevic and Schaer (2014) or elsewhere in paradigmatic IR sources.

Table 1 introduces the example. Copied with light editing from Dialog output, it shows four lines of raw data in which the seed document was "The design of browsing and berrypicking techniques for the on-line search interface," a well-known paper by Marcia J. Bates (1989). Commands not shown formed Set A—that is, the set of all documents directly citing the Bates paper in the online Social Sciences Citation Index (File 7). Then Dialog's RANK command was used (with the DETAIL option) to form Set B—that is, 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. 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, in bag of works weighting, 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 Social Sciences Citation Index in 2013 was estimated at three million records.

Table 1. Sample raw data from a citation file on DialogClassic

```
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-...
```

Bates (1989) is actually cited in 279 documents in Set A, but the most common version of the identifying string is cited in 264, and so that count is used here for simplicity. The others 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 specimen calculations for high-end and low-end Bates data. (Over the full dataset, these scores form a lognormal distribution, and the items shown take the extreme values in the positive and negative tails.) The documents are ranked

Table 2. Top 3 and bottom 3 works co-cited with Bates (1989)

	TF	*DF*	*Log TF*	*Log IDF*	*TF* IDF*
BATES MJ, 1989, V13, P407, ONLINE REV	264	264	3.42	4.06	13.9
ELLIS D, 1989, V45, P171, J DOC	61	203	2.79	4.17	11.6
BATES MJ, 1990, V26, P575, INFORM PROCESS MANA	31	94	2.49	4.5	11.2
BELKIN NJ, 1982, V38, P61, J DOC	53	274	2.72	4.04	11
LINCOLN YS, 1985, NATURALISTIC INQUIRY	4	6023	1.6	2.7	4.3
LAVE J, 1991, SITUATED LEARNING LE	3	4555	1.48	2.82	4.2
KUHN TS, 1970, STRUCTURE SCI REVOLU	3	5680	1.48	2.72	4.0

by TF*IDF score. 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 (see White 2010).

In bag of works retrieval, relevance varies directly with the TF factor and inversely with the IDF factor. TF*IDF weighting thus elevates works whose co-citation counts (TF) with the seed are high relative to their overall citation counts (DF). The cognitive effect is that high-ranked works in the distribution tend to be easy to relate to the seed because their verbal associations are highly specific to it. This can be seen at even the most superficial level, as in Table 3, where strings representing the top 12 items are spelled out as titles. (Books are in italicized title case.) The 12 works are all rather old, but they deal with principles of design that are relatively timeless, and, taken together, they cohere nicely for someone interested in what Bates's paper connotes. A researcher familiar with this area could readily discern a common 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 also cause thematically salient authors to recur. For example, Table 3 has two more papers by Bates and three by Nicholas J. Belkin.

Table 3. Top 12 titles co-cited with Bates (1989)

TF*IDF	Sole or First Author, Date, and Title of Co-cited Work
13.88	BATES MJ, 1989, The design of browsing and berrypicking techniques for the on-line search interface [seed]
11.61	ELLIS D, 1989, A behavioural approach to information retrieval design
11.22	BATES MJ, 1990, Where should the person stop and the information search interface start?
11	BELKIN NJ, 1982, ASK for information retrieval. Part 1.
10.9	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

At the same time, the TF*IDF weighting lowers the ranks of works whose overall citation counts (DF) are high relative to their co-citation (TF) counts with the seed. These latter works tend to be harder to relate to the seed because their associations with it are much less specific. The promotion of specific terms and demotion of non-specific terms is exactly what Karen Sparck Jones (1972) intended the IDF factor to do when she invented it—she called it "statistical specificity"—except that she and virtually everyone since have used IDF weighting on *words* rather than *works*. Yet on word-blind strings denoting works IDF performs no less well.

Table 4. Bottom 12 titles co-cited with Bates (1989)

TF*IDF	Sole or First Author, Date, and Title of Co-cited Work
4.9	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*
4.8	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 has the tail end of the 706 items in the Bates distribution. They tend to be famous theoretical or methodological items, mostly books, that are relevant to many research specialties. It is here that bag of works retrieval most clearly departs from what is customary in IR. It is hard to imagine typical assessors of relevance in TREC-style IR experiments marking any of the works in Table 4 as relevant to the Bates "berrypicking" paper (assuming they were presented). Yet each has been co-cited with it at least three times.

Granted, they may be related to the seed only very distantly in their local contexts of citation. One predictor is how widely they are separated from it in body text. (The effects of such "citation windows" have been examined by several researchers; see, e.g., Eto 2013). But they do co-occur with it in the global context set by the citing paper and thus bear consideration. If nothing else, they show connections that might never occur to someone who retrieved only works that are closely and obviously related to the seed. On that ground, a researcher or teacher examining the intellectual world of Bates's paper 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 an earlier draft of the present paper. Extracts from her comments (personal communication, February 2016) include: "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."

TF and IDF weights have been applied to ranked co-citation data before in White (2007a,b, 2009, 2010, 2014). These papers provide a number of detailed examples and extensive theoretical background. In White (2014) two historians comment like Bates on items retrieved by seeds they themselves supplied. They found the retrievals to be readily intelligible and could see a place for them in humanities scholarship.

3 Discussion

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 requires familiarity with the conventions of citation databases, which even learned researchers may lack. When Larsen (2004) built an experimental retrieval system that included direct citation linkages, he explicitly designed it so that users would not *need* a document to initiate retrieval; instead, seed documents were generated automatically from an initial subject search.

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 possess enough domain expertise to do this and thus might find uses for bag of works 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 non-metaphorical contexts? With bag of works 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 conducting 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, bag of works 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 non-specialists and scientists who read primarily to have their questions answered and not because of an interest in documents as texts *per se*. As Bates (1996) 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 empirical 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, bag of works retrievals for such persons could reveal something new: how 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. Bag of works retrieval thus in a small way supports intellectual history.

In this regard, bag of works 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. 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.

References

1. Bates, M.J. The design of browsing and berrypicking techniques for the online search interface. Online Review, 13, 5, 407-424 (1989).
2. Bates, M.J. Document familiarity, relevance, and Bradford's Law: The Getty Online Searching Project Report No. 5. Information Processing & Management 32, 6, 697-707 (1996).
3. Beel, J., et al. Research paper recommender systems: A literature survey. International Journal on Digital Libraries, pp. 1-34 (published online 2015).
4. Carevic, Z., Schaer, P. On the connection between citation-based and topical relevance ranking: Results of a pretest using iSearch. Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval, pp. 37-44 (2014).
5. Chapman, J., Subramanyam, K. Cocitation search strategy. National Online Meeting: Proceedings, pp. 97-102. Medford, NJ: Learned Information (1981).
6. Eto, M. Evaluations of context-based co-citation searching. Scientometrics 94, 2, 651-673 (2013).
7. Huang, W., et al. Recommending citations: Translating papers into references. Proceedings of the 21st International Conference on Information and Knowledge Management, pp. 1910-1914 (2012).

8. Larsen, B. References and citations in automatic indexing and retrieval systems: Experiments with the boomerang effect. PhD dissertation, Royal School of Library and Information Science, Copenhagen, Denmark (2004).
9. Manning, C., Schütze, H. Foundations of statistical natural language processing. MIT Press, Cambridge, Massachusetts (1999).
10. McNee, S., et al. On the recommending of citations for research papers. Proceedings of the ACM Conference on Computer Supported Cooperative Work, pp. 116-125 (2002).
11. Smucker, M.D. Evaluation of find-similar with simulation and network analysis. PhD dissertation, University of Massachusetts Amherst (2008).
12. Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28, 1, 11-21 (1972).
13. Strohman, T., Croft, B.W., Jensen, D. Recommending citations for academic papers. Technical Report IR466, Center for Intelligent Information Retrieval, University of Massachusetts Amherst (2006).
14. White, H.D. Combining bibliometrics, information retrieval, and relevance theory, Part 1: First examples of a synthesis. Journal of the American Society for Information Science and Technology 58, 4, 536-559 (2007a).
15. White, H.D. Combining bibliometrics, information retrieval, and relevance theory, Part 2: Some implications for information science, Journal of the American Society for Information Science and Technology 58, 4, 583-605 (2007b).
16. White, H.D. Pennants for Strindberg and Persson. In: Celebrating Scholarly Communication Studies: A Festschrift for Olle Persson at His 60th Birthday. Special volume of the E-Newsletter of the International Society for Scientometrics and Informetrics, S-5, 71-83 (2009).
17. White, H.D. Some new tests of relevance theory in information science. Scientometrics 83, 3, 653-667 (2010).
18. White, H.D. Relevance theory and citations. Journal of Pragmatics, 43, 14, 3345-3361 (2011).
19. White, H.D. Co-cited author retrieval and relevance theory: Examples from the humanities. Scientometrics, 102, 3, 2275-2299 (2014).