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

In bag of works retrieval…

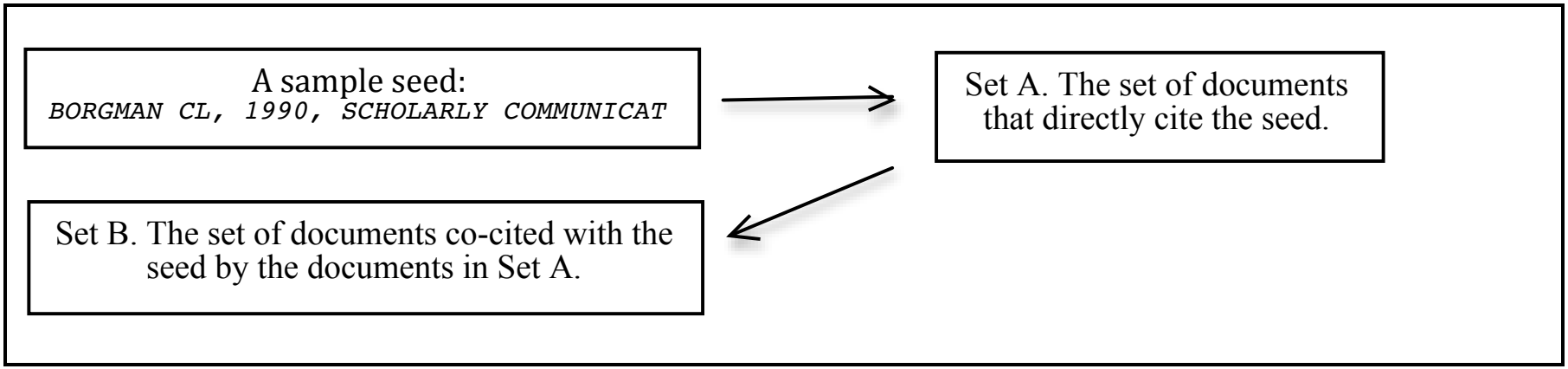
- Query is not one or more phrases implying a topical interest, but a work implying a topical interest—a seed.
- Database is "a bag" of citing and cited works, identified by brief strings.
- Such strings must be translatable into full bibliographic records of the works.
- Retrieval is a set of works (often large) that are all relevance-ranked with respect to the seed.
- *Any* retrieved work may interest the searcher, even the lowest ranked.

TF*IDF weighting of *works* instead of *words*

- TF weighting is based on the *co-citation counts* of retrieved works with the seed.
- IDF weighting is based on the *overall citation counts* of the retrieved works in the database.
- TF*IDF weighting *pushes up* works that are *more specifically and more obviously* related to the seed.
- TF*IDF weighting *pushes down* works are are *less specifically and less obviously* related to the seed.
- All TF*IDF predictions of relevance to the seed are based on empirical co-citation evidence.

Current situation

- Bag of works retrieval must be conducted in a database that allows:
  - a string identifying a work to be entered as a query;
  - retrieval of works that *directly cite* that work [Set A];
  - retrieval of works *co-cited with the seed* in Set A [Set B].
- With RANK command, It was possible to create Set A and Set B in Thomson Reuters databases on Dialog Classic (defunct as of 2013).
- It is not possible now because Web of Science, Scopus, and Google Scholar are not programmed for co-citation retrieval. (They can't create Set B.)



TI- THE INTELLECTUAL BASE AND RESEARCH FRONTS OF JASIS 1986-1990
AU- PERSSON O
JN- JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 1994, V45, N1, P31-38

One citing document retrieved in Set A.

CR- *BORGMAN CL, 1990, SCHOLARLY COMMUNICAT*
GARFIELD E, 1979, CITATION INDEXING
KESSLER MM, 1963, V14, P10, AM DOC
MCCAIN KW, 1908, V37, P111, J AM SOC INFORM SCI
PERSSON O, 1992, REPRESENTATIONS SCI
SALTON G, 1979, V22, P146, IEEE T PROFESSIONAL
SMALL H, 1985, V7, P391, SCIENTOMETRICS
SMALL HG, 1974, V4, P17, SCI STUD
VLADUTZ G, 1984, V21, P204, P AM SOC INFORM SCI
WHITE HD, 1981, V32, P163, J AM SOC INFORM SCI

A few cited references (CR's) retrieved in Set B. One is Borgman, the seed; the others are works that Persson co-cites with Borgman. TF's are counts of each Borgman-Other pair in all CR's of Set B.

An article as seed in bag of works retrieval

Seed:

Bates, M.J. (1989) The design of browsing and berrypicking techniques for the online search interface. Online Review, 13, 5, 407-424.

Seed as Social Sciences Citation Index string:

CR=BATES MJ, 1989, V13, P407, ONLINE REV

Seed as DOI:

http://dx.doi.org/10.1108/eb024320

First four Dialog RANK results (detailed display)
Social Sciences Citation Index

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-...

etc.

Top 3 and bottom 3 works co-cited with seed ranked by TF*IDF weight

Works	*TF*	*DF*	*Log TF*	*Log IDF*	*TF* IDF*
BATES MJ, 1989, V13, P407, ONLINE REV [seed]	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

TF*IDF = (1 + logTF) * (log(*N*/DF))

TF*IDF measures the relevance of the co-cited work to the seed over the entire retrieval. Any number of items may be ranked.

Top-ranked works have specific and obvious implications for the seed and its field.

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

Bottom-ranked works have implications for many fields beyond the seed's—but still are relevant to it.

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*

A book as seed in bag of works retrieval

Seed:

Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. (2008). *Introduction to Information Retrieval*. New York and Cambridge, UK: Cambridge University Press.

Seed as a Social Sciences Citation Index string:

CR=MANNING CD, 2008, INTRO INFORM RETRIEV

Top-ranked works have specific and obvious implications for the seed and its field

TF*IDF	Sole or First Author, Date, and Title of Co-cited Work
15.90	MANNING CD, 2008, *Introduction to Information Retrieval* [seed]
11.02	CHIRITA PA, 2007, Personalized query expansion for the Web
10.82	BAEZA-YATES, 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

Bottom-ranked works have implications for many fields beyond the seed's—but are still relevant to it

TF*IDF	Sole or First Author, Date, and Title of Co-cited Work
15.90	MANNING CD, 2008, *Introduction to Information Retrieval* [seed]
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: Networks of Plausible Reasoning*
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

Model of user of typical bag of *words* retrievals in paradigmatic IR

- Wants a question answered or an interest satisfied.
- Knows “a need” but not relevant works.
- Will search by spontaneously entering words that imply need.
- Can’t or won’t use ID’s of known works to search citation indexes.
- Retrieved documents are valued as information sources, not as ends in themselves.

Potential users of bag of *works* retrieval

- Someone who can imply an interest with at least one known seed work in addition to words.
- *Or:* Someone who can represent an interest *only* with a seed work.
- *Or:* Someone interested in how citers have used *the seed work itself* over time.
  - Scholars and domain analysts studying intellectual history who want to know how co-citation has contextualized a particular seed work.
  - Authors of a seed work who want to know how that work has been contextualized by others.

Bag of works retrieval is based on *implicit* content.

- It operates on ID strings of works such as

  CR= BATES MJ, 1989, V13, P407, ONLINE REV

  CR= MANNING CD, 2008, INTRO INFORM RETRIEV

- Once made explicit, the retrievals can be both *highly relevant* and *different from* retrievals made with bag of words algorithms.
- The two approaches are complementary.

Carevic, Zeljko, and Philipp Schaer. 2014. On the connection between citation-based and topical relevance ranking: results of a pretest using iSearch.

In *Bibliometric-enhanced information retrieval: BIR 2014; proceedings of the First Workshop on Bibliometric-Enhanced Information Retrieval, co-located with 36th European Conference on Information Retrieval (ECIR 2014)*, edited by Philipp Mayr, Philipp Schaer, Andrea Scharnhorst, Birger Larsen, and Peter Mutschke.

CEUR worksshop proceedings 1143, 37-44. Aachen: RWTH Aachen. http://ceur-ws.org/Vol-1143/paper5.pdf.

Thanks

1. White, H.D. (2007) 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.
2. White, H.D. (2007) Combining bibliometrics, information retrieval, and relevance theory, Part 2: Some implications for information science, *American Society for Information Science and Technology* 58, 4, 583-605.
3. White, H.D. (2009) Pennants for Strindberg and Persson. Special volume of the *E-Newsletter of the International Society for Scientometrics and Informetrics* S-5, 71-83.
4. White, H.D. (2010) Some new tests of relevance theory in information science. *Scientometrics* 83, 3, 653-667.
5. White, H.D. (2011) Relevance theory and citations. *Journal of Pragmatics* 43, 14, 3345-3361.
6. White, H.D. (2014) Co-cited author retrieval and relevance theory: Examples from the humanities. *Scientometrics* 102, 3, 2275-2299.