Bibliometrics, Information Retrieval and Natural Language Processing: Natural Synergies to Support Digital Library Research

Dietmar Wolfram[1]

[1]School of Information Studies, University of Wisconsin-Milwaukee
P.O. Box 413, Milwaukee, WI U.S.A. 53201
dwolfram@uwm.edu

Abstract Historically, researchers have not fully capitalized on the potential synergies that exist between bibliometrics and information retrieval (IR). Knowledge of regularities in information production and use, as well as citation relationships in bibliographic databases that are studied in bibliometrics, can benefit IR system design and evaluation. Similarly, techniques developed for IR and database technology have made the investigation of large-scale bibliometric phenomena feasible. Both fields of study have also benefitted directly from developments in natural language processing (NLP), which have provided new tools and techniques to explore research problems in bibliometrics and IR. Digital libraries, with their full text, multimedia content, along with searching and browsing capabilities, represent ideal environments in which to investigate the mutually beneficial relationships that can be forged among bibliometrics, IR and NLP. This brief presentation highlights the symbiotic relationship that exists among bibliometrics, IR and NLP.

Keywords: bibliometrics, information retrieval, digital libraries, natural language processing

1 Introduction

Both information retrieval (IR) and bibliometrics have long histories as distinct areas of investigation in information science. IR has focused on the storage, representation and retrieval of documents (text or other media) from the system and user perspectives. Bibliometrics and its allied areas (informetrics, scientometrics, webmetrics, or simply “metrics”) have focused on discovering and understanding regularities that exist in the way information is produced and used--but this simple definition belies the breadth of research undertaken. Units of analysis may be words, metadata fields, publications, authors, journals, research groups, institutions, sub-fields, disciplines or geographic regions. Applications of the study of these regularities and underlying processes extend to equally diverse areas such as science policy, subject indexing and IR system design, including digital libraries (DLs).

The documentary contents of bibliographic IR systems and their associated indexes provide much of the data that metrics researchers rely on today to conduct their research. Conversely, many of the processes within IR are directly informed by metrics research. Surprisingly, there has been little overlap in the research agendas of these two areas of study over the history of information science. This is changing with the growing recognition of the common interests and the tools and techniques used by IR and metrics researchers that can help to advance the research agendas of each field [1]. The recent international BIR workshops [2,3,4] have brought together researchers with combined interests in bibliometrics and IR. The BIR research presentations demonstrate the potential synergies that exist between these two areas. With large, full text databases now commonplace, both IR and bibliometrics have also benefitted from developments in natural language processing (NLP) and computational linguistics, where text-based techniques have improved document retrieval and the ability to explore relationships among entities of interest in metrics research.

In a sense, digital libraries represent an ideal environment in which to study the intersection of IR, bibliometrics and NLP, whether the DLs consist of repositories of formal publications or heterogeneous collections of multimedia documents. With content representation and search functionality found in more traditional IR environments, hyperlinks that mimic relationships similar to those found in citation analysis, and full text contents that lend themselves to NLP analysis. This brief overview outlines how developments in IR and bibliometrics, along with language-based methods, have helped to advance research in both areas. These intersections are most evident in bibliographic databases, but also extend to more heterogeneous digital libraries.

2 Information Retrieval Research

IR research has focused on the design of more efficient and effective systems to match documents to queries to meet the information needs of users. Early IR research was more system-centered, but user-centered approaches are now employed alongside system-centered approaches in the design and evaluation of IR systems. Bibliometrics provides useful methods for the analysis of both system content and system usage, to better understand system processes and user dynamics. In essence, IR processes represent forms of information production and use [5], where observed patterns follow classic power law distributions or possibly unimodal distributions (e.g., lognormal or Poisson-like). Furthermore, metric studies of scientific communication can provide frameworks to assist in the design of IR systems.

Aspects of IR system content that lend themselves to bibliometric modeling include index term frequency distributions, indexing exhaustivity or term assignment, term co-occurrence frequency distributions, database and index growth, and more recently, aspects of web-based IR such as frequency distributions of inlinks/outlinks and document persistence [6]. Applications of bibliometric modeling to IR systems have included the development of simulations to model IR system processes to better understand system interactions. These models also can be used to observe retrieval efficiency under different situations to identify preferred file structures or for space planning [7,8].

User interactions with IR systems as recorded by transaction logs may be similarly modeled. Characteristics that can be modeled include search terms used per query, frequency distributions of query terms, query term co-occurrence, query frequency distributions, search session length based on queries or other actions, user search and browsing regularities, and site/document visitation frequency. Search and browsing patterns may be modeled using network analysis methods, similar to those applied in citation analysis, to identify issues with interface use [9], or clustering techniques may be applied to identify larger scale search session patterns [10].

Beyond mathematical modeling of specific bibliometric characteristics of IR systems, higher level aspects of science models also can inform the search and retrieval process of scientific literature. Mutschke and Mayr [11], for instance, recognize this important observation and demonstrate that retrieval performance can benefit by ranking results based on our understanding of models of science.

Citation-based connections between bibliographic records serve as a readily exploitable data source for expanding searching and browsing options for users. These ideas have been implemented in experimental systems over the past several decades such as I^{3}R [12] and BIRS [13], with the latter also supporting query expansion through visualizations of relationships among searchable entities. The importance of citation linkages is also recognized by commercial database vendors such as

EBSCO and ProQuest, which now provide active hyperlinks to references included in full text articles and to citing articles, where available.

For much of the history of IR, the research focus was on the representation and retrieval of document surrogates with limited natural language. Representation and retrieval were based on metadata fields, keywords or controlled vocabularies. Models to support query and document matching were based on bag-of-words approaches that treated terms independently of one another. Context and semantics played no role in determining the aboutness of documents. At the same time, computational challenges increased as the size of databases grew. IR models such as the vector space model became more computationally expensive due to the high dimensionality associated with representing documents and processing document-query matching.

IR research has been more responsive to taking advantage of advanced NLP techniques to provide a more natural search environment for users and to extend retrieval evaluation beyond simple term independence models. Dimensionality reduction techniques that lower the computational overhead associated with IR processing have become commonplace since early approaches based on latent semantic analysis were introduced over 25 years ago. Successors based on language modeling [14], including topic modeling [15], have provided novel ways to tackle information search and retrieval. The applications have more recently included the recommendation of scientific articles [16] and the identification of search terms [17]. Language-based techniques have also been used to identify additional index terms for documents based on language surrounding citations appearing in citing documents [18].

3 Bibliometrics Research

The relationship between IR and bibliometrics research can be considered in some ways symbiotic [19]. As noted above, methods used in metrics research have informed IR research. Conversely, tools and techniques developed to support IR research have been adapted to support metrics research. An exemplar is the development and application of PageRank [20] used by Google to rank webpages. It was directly influenced by citation analysis methods used in bibliometrics research. The developers recognized the parallels between citations and webpage inlinks. The effectiveness of PageRank is evident in Google's longstanding success as a search engine. More recently, PageRank-based approaches have been reapplied to citation analysis research problems [21]. Other similar network-based approaches, such as Eigenfactor (URL: `http://www.eigenfactor.org`), provide additional ways to rank journals or support methods to rank or recommend articles [22]. Furthermore, the h-index, initially developed to measure author impact, has been demonstrated to have application for the ranking of webpages, similar to PageRank, but with less computational overhead [23].

Citation analysis has been a core area of bibliometrics research for over 50 years. Citation relationships in the form of direct citations, co-citations or bibliographic coupling create a network of linkages among documents, authors, and publication venues that allows one to visualize the intellectual structure of a field [24], or all of science [25]. Similarly, co-authorship-based studies provide another type of linkage that allows researchers to gain deeper insights into the dynamics of research communities [26]. One limitation inherent in citation-based studies is that relationships among entities of interest only exist if the linkages exist. The lack of direct citation, co-citation, bibliographic coupling or co-authorship should not preclude the possibility of relationships.

This situation can be addressed using approaches that integrate the vocabulary used by the entities of interest, usually employing keyword or controlled vocabulary co-occurrence, which takes bibliometric studies more into the realm of IR research. Still, there are concerns with keyword-based approaches in bibliometrics research that mirror the same bag-of-words issues found in IR research. The assumption of independence of vocabulary in bibliographic entities, whether as keywords or subject terms, creates the same issues observed earlier in IR studies. The ability to work with full text and natural language can improve these analyses. More recent research has moved beyond simple co-occurrence to identify relationships among entities of interest. Of particular note has been the application of topic modeling techniques, such as Latent Dirichlet Analysis (LDA) [27], to bibliographic records to identify relationships among bibliometric entities of interest, for instance, authors [28,29], that may not be reflected through citation or collaboration. Tang et al. [30] have developed a searchable system initially called ArnetMiner (now AMiner, URL: `http://aminer.org`) that profiles researchers based on publication content and supports expertise search.

4 Conclusion

The application of bibliometric methods for IR research and vice versa has evolved over the past forty years, from bibliometric modeling of IR system processes to the exploitation of citation relationships to provide extended browsing capabilities to identify potentially relevant documents based on citation linkages. More recently, the adoption of language-based methods from NLP and computational linguistics has benefitted both IR and bibliometrics research. Applications of language-based approaches are still relatively new in bibliometric contexts. Link-based analysis (citations, co-authorship, hyperlinks) in combination with textual analysis can capitalize on the strengths of both approaches [31]. The analysis of full text collections--whether bibliographic databases or heterogeneous, multimedia digital libraries--offers many opportunities for further study. The applications of citation-based methods and emerging language-based methods are evident in the range of presentations given at the BIRNDL workshop, which include the use of citation methods, text mining and topic modeling to enhance retrieval in full text or digital library environments, scholarly communication and our understanding of disciplinary boundaries.

5 References

1. Mayr, P., Scharnhorst, A.: Scientometrics and Information Retrieval: Weak-links Revitalized. Scientometrics 102, 2193–2199 (2015)

2. Mayr, P., Scharnhorst, A., Larsen, B., Schaer, P., Mutschke, P.: Bibliometric-enhanced Information retrieval. In Advances in Information Retrieval (pp. 798-801). Springer International Publishing (2014)

3. Mayr, P., Frommholz, I., Scharnhorst, A., Mutschke, P.: Bibliometric-enhanced Information Retrieval: 2nd International BIR Workshop. In Advances in Information Retrieval (pp. 845-848). Springer International Publishing (2015)

4. Mayr, P., Frommholz, I., Cabanac, G.: Editorial for the 3rd Bibliometric-Enhanced Information Retrieval Workshop at ECIR 2016 (2016)

5. Egghe, L.: Power Laws in the Information Production Process: Lotkaian Informetrics. Elsevier (2005)

6. Wolfram, D.: Applied Informetrics for Information Retrieval Research. Libraries Unlimited (2003)

7. Wolfram, D.: Applying Informetric Characteristics of Databases to IR System File Design, Part I: Informetric Models. Information Processing and Management, 28(1), 121-133 (1992)

8. Wolfram, D.: Applying Informetric Characteristics of Databases to IR System Design, Part II: Simulation Comparisons. Information Processing and Management, 28(1), 135-151 (1992)

9. Han, H.J., Joo, S., Wolfram, D.: Using Transaction Logs to Better Understand User Search Session Patterns in an Image-Based Digital Library. Journal of the Korean Biblia Society for Library and Information Science. 25(1), 19-37 (2014)

10. Wolfram, D., Wang, P., Zhang, J.: Identifying Web Search Session Patterns Using Cluster Analysis: A Comparison of Three Search Environments. Journal of the American Society for Information Science and Technology, 60(5), 896-910 (2009)

11. Mutschke, P., Mayr, P.: Science Models for Search: A Study on Combining Scholarly Information Retrieval and Scientometrics. Scientometrics, 102, 2323-2345 (2015)

12. Croft, W.B., Thompson, R.H.: I R: A New Approach to the Design of Document Retrieval Systems. Journal of the American Society for Information Science, 38(6), 389-404 (1987)

13. Ding, Y., Chowdhury, G. G., Foo, S., Qian, W. Bibliometric information retrieval system (BIRS): A web search interface utilizing bibliometric research results. Journal of the American Society for Information Science, 51(13), 1190-1204 (2000)

14. Ponte, J.M., Croft, W.B.: A Language Modeling Approach to Information Retrieval. In Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 275-281). ACM (1998)

15. Wei, X., Croft, W. B.: LDA-based Document Models for Ad-Hoc Retrieval. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 178-185). ACM (2006)

16. Wang, C., Blei, D.M.: Collaborative Topic Modeling For Recommending Scientific Articles. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 448-456). ACM (2011)

17. Koopman, R., Wang, S., Scharnhorst, A., Englebienne, G.: Ariadne's Thread: Interactive Navigation in a World of Networked Information. In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (pp. 1833-1838). ACM (2015)

18. Ritchie, A., Robertson, S., Teufel, S.: Comparing Citation Contexts for Information Retrieval. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 213-222). ACM (2008)

19. Wolfram, D.: The Symbiotic Relationship between Information Retrieval and Informetrics. Scientometrics, 102, 2201-2214 (2015)

20. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. `http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf` (1999)

21. Waltman, L., Yan, E.: PageRank-related Methods for Analyzing Citation Networks. In Y. Ding, R. Rousseau and D. Wolfram (Eds.). Measuring Scholarly Impact. (pp. 83-100). Springer (2014)

22. West, J. D., Wesley-Smith, I., Bergstrom, C. T.: A Recommendation System Based on Hierarchical Clustering of an Article-level Citation Network. IEEE Transactions on Big Data (Forthcoming)

23. Bar-Ilan, J., Levene, M.: The hw-rank: An h-index Variant for Ranking Web Pages, Scientometrics, 102, 2247–2253 (2015)

24. White, H.D., McCain, K.W.: Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972-1995. Journal of the American Society for Information Science, 49, 327-355 (1998)

25. Boyack, K.W., Klavans, R., Börner, K.: Mapping the Backbone of Science. Scientometrics, 64, 351-374 (2005)

26. Glänzel, W., Schubert, A.: Analysing Scientific Networks through Co-Authorship. In Handbook of Quantitative Science and Technology Research (pp. 257-276). Springer Netherlands (2004)

27. Blei, D.M., Ng, A.Y., & Jordan, M.J.: Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022 (2003)

28. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The Author-topic Model for Authors and Documents. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (pp. 487-494). AUAI Press (2004)

29. Lu, K., Wolfram, D.: Measuring Author Research Relatedness: A Comparison of Word-Based, Topic-Based and Author Co-Citation Approaches. Journal of the American Society for Information Science and Technology, 63, 1973-1986 (2012)

30. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 990-998). ACM (2008)

31. Zitt, M.: Meso-level Retrieval: IR-bibliometrics Interplay And Hybrid Citation-Words Methods In Scientific Fields Delineation, Scientometrics, 102, 2223–2245 (2015)