In Praise of Interdisciplinary Research through Scientometrics*

Guillaume Cabanac

University of Toulouse 3 – Paul Sabatier
Department of Computer Science
IRIT UMR 5505 CNRS
118 route de Narbonne
F-31062 Toulouse cedex 9

guillaume.cabanac@univ-tlse3.fr

Abstract. The BIR workshop series foster the revitalisation of dormant links between two fields in information science: information retrieval and bibliometrics/scientometrics. Hopefully, tightening up these links will cross-fertilise both fields. I believe compelling research questions lie at the crossroads of scientometrics and other fields: not only information retrieval but also, for instance, psychology and sociology. This overview paper traces my endeavours to explore these field boundaries. I wish to communicate my enthusiasm about interdisciplinary research mediated by scientometrics and stress the opportunities offered to researchers in information science.

Keywords: Scientometrics, Information Retrieval, Digital Libraries, Psychology of Science, Sociology of Science

1 Introduction

Long-established ties unite information retrieval and scientometrics/bibliometrics under the umbrella domain of information science [33,43,46]. Both rely on the quantitative study of documents a) to fulfil a user's information need or b) to reveal how knowledge is created, used, and incorporated. The BIR workshop series brings together researchers from both fields to foster the cross-fertilisation of ideas [34]. This overview paper introduces 12 cases of such interdisciplinary research [2–10,18–20]. As a companion to the keynote talk, this paper discusses these research issues with an emphasis on the data and methods used to tackle them.

* This is a companion overview paper to the keynote talk given at the Bibliometric-enhanced Information Retrieval (BIR) workshop collocated with the ECIR 2015 conference. The slides are available at http://bit.ly/birCabanac2015.

2 Research at the Crossroads of Scientometrics and . . .

My talk intends to gives a taste of the richness of research questions at the boundaries of scientometrics and other disciplines and fields. I have a bent for *descriptive* scientometrics, whose main purpose is to further our understanding of knowledge creation, sharing, and incorporation. My research is not directly concerned with *evaluative* scientometrics that regularly attracts critical comments, see, e.g., [15,36].

This section outlines my contributions that appeared in the *Journal of the Association for Information Science and Technology* and *Scientometrics*. Some studies were done in collaboration with colleagues from various backgrounds, which is a source of mutual enrichment.

2.1 . . . Information Retrieval

Before drifting apart during the past few decades, scientometrics and information retrieval (IR) were more closely related than they are nowadays. The introductory paper to the BIR@ECIR 2014 workshop recalls this tight relationship:

> “Many pioneers in bibliometrics (e.g., Goffman, Brookes, Vickery), actually came from the field of IR, which is one of the traditional branches of information science. IR as a technique stays at the beginning of any scientometric exploration, and so, IR belongs to the portfolio of skills for any bibliometrician / scientometrician.” [34, p. 799]

Some of my work in scientometrics uses IR concepts. For instance, the question tackled in [2] called on the evaluation of search effectiveness, while [5] relied on information extraction through regular expressions:

- How to tailor researcher recommendations with social clues? [2]
- How to extract and quantify eponyms from academic papers? [5]

2.2 . . . Digital Libraries

Educational materials are available from multiple sources: publishers’ websites, open access journals, preprint repositories, and so on. The mining of their usage logs reveal insights about scientists and the general public. For instance, Wang et al. graphed the working patterns of worldwide scientists based on real-time usage data from SpringerLink [42].

The development of online text-sharing platforms caught my attention. Here I studied the *Library Genesis*[1] platform hosting 25 million documents and totalling 42 terabytes in size in [6]. These documents distributed for free are mostly research papers, textbooks, and books in English. The research question I tackled was: How do ‘biblioleaks’ [12] and user crowdsourcing feed such a platform with educational and recreational materials?

[1] http://www.libgen.org

2.3 ... The Psychology of Science

The scientific thought and behaviour of individual scientists can be studied through the prism of theories and established results in psychology [13]. This is a form of reflexive research I particularly enjoy working on. My research endeavour focused on the relation between a scientist's writings and his/her gender [18] or age [19]. Psychological research also triggered questions about the temporal organisation of individual authors and gatekeepers [7], as well as on a perceptual bias affecting the bidding behaviour of conference referees [10]:

- Do men and women differ in their way of writing papers? [18]
- How does writing evolves through time: the case of James Hartley [19]
- How do order effects affect the bids on conference papers? [10]
- What is the work-life balance of academics involved in *JASIST*? [7]

2.4 ... The Sociology of Science

Another compelling research area is the study of the *social* system of science [35]. How do individual scientists organise and collaborate to produce and share knowledge? My research sought to address the following questions about collaborative academic writing [8], collaboration dynamics [4,9], and the social structure of a research field from the viewpoint of editorial boards [3]:

- Do researchers write in different ways when working alone or in groups? [8]
- What are the dynamics of lifelong careers in computer science? [9]
- Is the partnership φ-index model accurate on 1 million biographies? [4]
- What are the features of gatekeepers in the field of *Information Systems*? [3]

3 Data

The data collected to study these research questions came from a variety of sources. Here is a selection frequently used in my research:

- The *Journal Citation Reports* (JCR) is part of the *Web of Knowledge* platform run by Thomson Reuters. The JCR is released in two yearly editions: *science* and *social sciences*. Journals are listed in one or two editions under one or more categories (e.g., *Computer Science – Information Systems*). Indicators such as the Impact Factor [14] are provided for each enlisted journal. This dataset was used in [3,18].
- The *Digital Bibliography & Library Project* (DBLP) is an open dataset collecting the biographies of 1.5 million computer scientists from publisher's websites and other inputs [27]. The DBLP maintainers strive to disambiguate homonyms with social network analysis and other techniques. This dataset[2] available in XML format was used in [2,3,4,9].

[2] http://dblp.uni-trier.de/xml

- *Google Scholar* (GS) lists the publications and citations of individual researchers. The accuracy of this dataset still raised concerns [11,22,26], as GS is of less quality than commercial products, such as the *Web of Science*, and *Scopus*. This dataset was used with manual curation in [19].
- Publisher websites publish the full-text versions of papers in PDF and, sometimes, in formats easier to parse, such as HTML and XML. For instance, eponyms were extracted from *Scientometrics* papers in [5] and the occurrence of tables and figures were counted and studied in [8,18].

Compelling research questions and innovative hypotheses sometimes come to mind unexpectedly. This I experienced when realising that valuable and disregarded information exists somewhere. Here is a selection of such lesser-known data sources that I have used as input to my research.

- *Confmaster* is a conference management system. It supported the peer review process of hundreds of conferences in Computer Science (e.g., CIKM and SIGIR) and other fields. The anonymised bids placed on papers (and referee marks) of 42 such conferences were studied and the data was made publicly available [10].
- Publishers websites provide metadata about the papers included in the journals they own. For instance, the dates of submission, revision, and publication of *JASIST* papers were studied in [7] and the gatekeepers sitting on the editorial boards of 77 journals in *Information Systems* were studied in [3].
- Online text-sharing platforms host millions of educational and recreational materials. For instance, the catalogue of the *Library Genesis* with 25 million entries linking to 42 terabytes of documents was used in [6].
- The *Depositor* service[3] records all CrossRef DOIs registered with papers published in conference proceedings or journals, book chapters, books, data, and so on. These data were also used in [6].
- The *Essential Science Indicators* published by Thomson Reuters lists over 10,000 journals classified into one of 22 fields of science. This was used to uncover the topics of documents crowdsourced in a text-sharing platform [6].

4 Methods

The quantitative study of science requires one to build data processing workflows. Some components are rather stable, such as the computation of topic-based similarity measures. Other components need to be tailored for each study, such as metadata extractors from publisher websites. This section discusses some of the methods and tools I used to extract, filter, store, process, and analyse a variety of datasets.

[3] http://www.crossref.org/06members

4.1 Data Extraction

There is a growing number of open datasets providing researchers with off-the-shelf, curated, and properly formatted data (e.g., DBLP). But sometimes the data needed for a given study do not come nicely packaged and ready to use! Some studies like [7] were only made possible by programming a web scrapper with HtmlUnit[4] to extract article dates from publisher websites. In other cases, I manually collected data as in [3] about the boards of 77 journals with the name, affiliation, and gender of their 2,846 gatekeepers. Manual data curation and validation is often a necessary step in the data science process [20]. We should strive to release the valuable datasets we produce to ensure the reproducibility of the results and to foster their uptake [17].

4.2 Data Storage

For some studies, storing data in files is the most simple and efficient option. But when the analysis to perform gains in complexity, resorting to a proper database proves helpful as stressed in [28,31,44]. For example, I mapped the XML data from SQL to the Oracle relational database that was featured in [2] and other studies.

4.3 Data Processing

Depending on the underlying data model, a variety of tools and techniques are available. Command line scripts [24] are an efficient way to deal with files, as in [5]. Declarative programming languages such as SQL are concise and powerful to process complex queries, as in [9]. Imperative programming languages such as Java are also an option, albeit less concise and perhaps more difficult to master. There are also advanced spreadsheet functions (e.g., pivot tables) and off-the-shelf software like SOFA statistics[5] that proved very handy for basic data science tasks, such as generating report tables and computing statistical tests of significance as in [8]. Symbolic regression [25, Chap. 10] as implemented in the Eureqa software [39] is an example of a more advanced technique used in [5] to learn the equation of a model fitting data by maximising its goodness of fit.

4.4 Information Visualisation

Exploratory data analysis [41] relies on the visualisation of data and information resulting from data processing. Spreadsheets are simple tools to plot data, albeit cumbersome to automate. Scripting languages like Gnuplot [23] allow one to generate all sorts of graphs while minimising manual intervention. Examples of Gnuplot charts, box plots, and population pyramids appear in [3]. In addition, word clouds are an adequate visualisation to convey the topics of a text by displaying size-varying keywords, as in [3,5].

[4] http://htmlunit.sourceforge.net

[5] http://www.sofastatistics.com

5 Concluding Remarks

Are the links between information retrieval and scientometrics getting tighter? From my young observer's standpoint, this seems to be the case. Traditionally IR-oriented journals seem to publish a growing number of papers linked to scientometric issues. For example, see the recent table of contents of:

- *Foundations and Trends in Information Retrieval* [29],
- *Information Processing & Management* [45],
- *Information Retrieval* [30],
- *Information Sciences* [21],
- the *Journal of the Association for Information Science and Technology* [37],
- the *Journal of Documentation* [1],
- the *Journal of Information Science* [40],
- the *Online Information Review* [38],
- and *World Wide Web* [16].

On the other hand, *Scientometrics* published a special issue with nine papers addressing the question of "combining bibliometrics and information retrieval" [32]. The promising process of link revitalisation [33] seems to be on track.

Maybe the time is now ripe for joining forces with colleagues from other disciplines to broaden our scope and tackle further compelling research questions demanding interdisciplinary approaches.

Acknowledgements

I am grateful to the organisers of the BIR@ECIR 2015 workshop for inviting me to give this keynote talk.

References

1. Bornmann, L.: Assigning publications to multiple subject categories for bibliometric analysis. Journal of Documentation 70(1), 52–61 (2014), doi:10.1108/jd-10-2012-0136
2. Cabanac, G.: Accuracy of inter-researcher similarity measures based on topical and social clues. Scientometrics 87(3), 597–620 (2011), doi:10.1007/s11192-011-0358-1
3. Cabanac, G.: Shaping the landscape of research in Information Systems from the perspective of editorial boards: A scientometric study of 77 leading journals. Journal of the American Society for Information Science and Technology 63(5), 977–996 (2012), doi:10.1002/asi.22609
4. Cabanac, G.: Experimenting with the partnership ability φ-index on a million computer scientists. Scientometrics 96(1), 1–9 (2013), doi:10.1007/s11192-012-0862-y
5. Cabanac, G.: Extracting and quantifying eponyms in full-text articles. Scientometrics 98(3), 1631–1645 (2014), doi:10.1007/s11192-013-1091-8

6. Cabanac, G.: Bibliogifts at LibGen? Study of a text-sharing platform driven by biblioleaks and crowdsourcing. Journal of the Association for Information Science and Technology (forthcoming), doi:10.1002/asi.23445
7. Cabanac, G., Hartley, J.: Issues of work-life balance among *JASIST* authors and editors [Brief communication]. Journal of the American Society for Information Science and Technology 64(10), 2182–2186 (2013), doi:10.1002/asi.22888
8. Cabanac, G., Hartley, J., Hubert, G.: Solo *versus* collaborative writing: Discrepancies in the use of tables and graphs in academic articles. Journal of the Association for Information Science and Technology 65(4), 812–820 (2014), doi:10.1002/asi.23014
9. Cabanac, G., Hubert, G., Milard, B.: Academic careers in computer science: Continuance and transience of lifetime co-authorships. Scientometrics 102(1), 135–150 (2015), doi:10.1007/s11192-014-1426-0
10. Cabanac, G., Preuss, T.: Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees. Journal of the American Society for Information Science and Technology 64(2), 405–415 (Feb 2013), doi:10.1002/asi.22747
11. Delgado López-Cózar, E., Robinson-García, N., Torres-Salinas, D.: The Google scholar experiment: How to index false papers and manipulate bibliometric indicators. Journal of the Association for Information Science and Technology 65(3), 446–454 (2014), doi:10.1002/asi.23056
12. Dunn, A.G., Coiera, E., Mandl, K.D.: Is Biblioleaks inevitable? Journal of Medical Internet Research 16(4), e112 (2014), doi:10.2196/jmir.3331
13. Feist, G.J.: Psychology of science as a new subdiscipline in psychology. Current Directions in Psychological Science 20(5), 330–334 (2011), doi:10.1177/0963721411418471
14. Garfield, E.: The history and meaning of the Journal Impact Factor. Journal of the American Medical Association 295(1), 90–93 (2006), doi:10.1001/jama.295.1.90
15. Génova, G., Astudillo, H., Fraga, A.: The scientometric bubble considered harmful. Science and Engineering Ethics (forthcoming), doi:10.1007/s11948-015-9632-6
16. Han, D., Liu, S., Hu, Y., Wang, B., Sun, Y.: ELM-based name disambiguation in bibliography. World Wide Web 18(2), 253–263 (2015), doi:10.1007/s11280-013-0226-4
17. Hanson, B., Sugden, A., Alberts, B.: Making data maximally available. Science 331(6018), 649 (2011), doi:10.1126/science.1203354
18. Hartley, J., Cabanac, G.: Do men and women differ in their use of tables and graphs in academic publications? Scientometrics 98(2), 1161–1172 (2014), doi:10.1007/s11192-013-1096-3
19. Hartley, J., Cabanac, G.: An academic odyssey: Writing over time. Scientometrics (forthcoming), doi:10.1007/s11192-015-1562-1
20. Hartley, J., Cabanac, G., Kozak, M., Hubert, G.: Research on tables and graphs in academic articles: Pitfalls and promises [Brief communication]. Journal of the Association for Information Science and Technology 66(2), 428–431 (2015), doi:10.1002/asi.23208
21. Hurtado Martín, G., Schockaert, S., Cornelis, C., Naessens, H.: Using semi-structured data for assessing research paper similarity. Information Sciences 221, 245–261 (2013), doi:10.1016/j.ins.2012.09.044
22. Jacsó, P.: Metadata mega mess in Google Scholar. Online Information Review 34(1), 175–191 (2010), doi:10.1108/14684521011024191
23. Janert, P.K.: Gnuplot in Action: Understanding data with graphs. Manning Publications, Greenwich, CT (2010)
24. Janssens, J.: Data Science at the Command Line: Facing the Future with Time-tested Tools. O'Reilly, Sebastopol, CA (2015)

25. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA (1992)
26. Labbé, C.: Ike Antkare, one of the great stars in the scientific firmament. ISSI Newsletter 6(2), 48–52 (2010)
27. Ley, M.: The DBLP computer science bibliography: Evolution, research issues, perspectives. In: Laender, A.H.F., Oliveira, A.L. (eds.) SPIRE'02 : Proceedings of the 9th international conference on String Processing and Information Retrieval. LNCS, vol. 2476, pp. 1–10. Springer (2002), doi:10.1007/3-540-45735-6_1
28. Loudcher, S., Jakawat, W., Morales, E.P.S., Favre, C.: Combining OLAP and information networks for bibliographic data analysis: A survey. Scientometrics (forthcoming), doi:10.1007/s11192-015-1539-0
29. Lupu, M., Hanbury, A.: Patent retrieval. Foundations and Trends in Information Retrieval 7(1), 1–97 (2013), doi:10.1561/1500000027
30. Mahdabi, P., Crestani, F.: The effect of citation analysis on query expansion for patent retrieval. Information Retrieval 17(5–6), 412–429 (2014), doi:10.1007/s10791-013-9232-5
31. Mallig, N.: A relational database for bibliometric analysis. Journal of Informetrics 4(4), 564–580 (2010), doi:10.1016/j.joi.2010.06.007
32. Mayr, P., Scharnhorst, A.: Combining bibliometrics and information retrieval: Preface. Scientometrics 102(3), 2191–2192 (2015), doi:10.1007/s11192-015-1529-2
33. Mayr, P., Scharnhorst, A.: Scientometrics and information retrieval: Weak-links revitalized. Scientometrics 102(3), 2193–2199 (2015), doi:10.1007/s11192-014-1484-3
34. Mayr, P., Scharnhorst, A., Larsen, B., Schaer, P., Mutschke, P.: Bibliometric-enhanced information retrieval. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C., de Jong, F., Radinsky, K., Hofmann, K. (eds.) ECIR'14: Proceedings of the 36th European Conference on IR Research. LNCS, vol. 8416, pp. 798–801 (2014), doi:10.1007/978-3-319-06028-6_99
35. Merton, R.K.: The Sociology of Science: Theoretical and Empirical Investigations. The University of Chicago Press, Chicago, IL (1973)
36. Moustafa, K.: The disaster of the Impact Factor. Science and Engineering Ethics 21(1), 139–142 (2015), doi:10.1007/s11948-014-9517-0
37. Nicolaisen, J., Frandsen, T.F.: Bibliometric evolution: Is the *Journal of the Association for Information Science and Technology* transforming into a specialty journal? Journal of the Association for Information Science and Technology (forthcoming), doi:10.1002/asi.23224
38. Pedersen, L.A., Arendt, J.: Decrease in free computer science papers found through Google Scholar. Online Information Review 38(3), 348–361 (2014), doi:10.1108/oir-07-2013-0159
39. Schmidt, M., Lipson, H.: Distilling free-form natural laws from experimental data. Science 324(5923), 81–85 (2009), doi:10.1126/science.1165893
40. Subotic, S., Mukherjee, B.: Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles. Journal of Information Science 40(1), 115–124 (2014), doi:10.1177/0165551513511393
41. Tukey, J.W.: Exploratory Data Analysis. Behavioral Science: Quantitative Methods, Addison Wesley, Philippines (1977)
42. Wang, X., Xu, S., Peng, L., Wang, Z., Wang, C., Zhang, C., Wang, X.: Exploring scientists' working timetable: Do scientists often work overtime? Journal of Informetrics 6(4), 655–660 (2012), doi:10.1016/j.joi.2012.07.003
43. 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(4), 327–355 (1998), doi:b57vc7

44. Wolfram, D.: Applications of SQL for informetric frequency distribution processing. Scientometrics 67(2), 301–313 (2006), doi:10.1007/s11192-006-0101-5
45. Yang, S., Han, R.: Breadth and depth of citation distribution. Information Processing & Management 51(2), 130–140 (2015), doi:10.1016/j.ipm.2014.12.003
46. Zhao, D., Strotmann, A.: The knowledge base and research front of Information Science 2006–2010: An author cocitation and bibliographic coupling analysis. Journal of the Association for Information Science and Technology 65(5), 995–1006 (2014), doi:10.1002/asi.23027