Measuring Scholarly Impact

Ying Ding • Ronald Rousseau • Dietmar Wolfram
Editors

Measuring Scholarly Impact

Methods and Practice

Springer

Editors
Ying Ding
School of Informatics and Computing
Indiana University
Bloomington, IN, USA

Ronald Rousseau
University of Antwerp
Antwerp, Belgium

Dietmar Wolfram
University of Wisconsin-Milwaukee
Milwaukee, WI, USA

ISBN 978-3-319-10376-1 ISBN 978-3-319-10377-8 (eBook)
DOI 10.1007/978-3-319-10377-8
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014950682

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The measurement and assessment of scholarly impact have been experiencing rapid changes over the past two decades thanks to developments in how scholarship is communicated and advances in the tools and techniques that may be used to study scholarly communication. Measures of research impact play an increasingly important role in how individuals, research groups, journals, academic departments, institutions, and countries are ranked in their respective areas of contribution to scholarship. From the beginnings of metrics-related research in the early twentieth century, when small-scale quantitative studies of scholarly communication first revealed distinct patterns in the way publications are produced, research approaches have evolved to today's methods that employ a range of tools and techniques on large-scale datasets. Research contributions from statistical sciences, scientific visualization, network analysis, text mining, and information retrieval have provided tools and techniques to investigate metric phenomena and to assess scholarly impact in new ways. The core and complementary interests of metric studies are reflected in the names that are used to describe the field, which are also used in this edited book. Authors may have preferred terms to describe what they research. Regardless of the preferred term, there is an underlying theme of exploring how the process and products of scholarly communication may be better understood. The term bibliometrics, which is still widely used, can be traced back to the 1930s when Otlet introduced the French term bibliométrie (Otlet, 1934). Bibliometrics was later defined by Pritchard (1969) as "the application of mathematics and statistical methods to books and other media of communication." Around the same time, the term scientometrics (Naukometriya) was proposed by Nalimov and Mul'chenko (1969) as "the application of those quantitative methods which are dealing with the analysis of science viewed as an information process." Later, Nacke (1979) proposed the term informetrics (Informetrie) to encompass all quantitative aspects of the study of information, its production, dissemination, and use. More recently, the term webmetrics has been used to describe the application of metric approaches to information phenomena on the Internet, and more specifically the World Wide Web (Almind & Ingwersen, 1997). Space limitations prevent us from providing a

detailed overview of these topics. Readers who are interested in finding out more about the history and scope of informetrics are encouraged to consult De Bellis (2009), Björneborn and Ingwersen (2004), as well as Egghe and Rousseau (1990).

Informetrics research has expanded beyond the evaluation of traditional units of measure such as authors and journals and now includes a broader array of units of measure and assessment. At the same time, the availability of larger, more detailed datasets has made it possible to study more granular levels of data in the production, dissemination, and use of the products of scientific communication. In the current era of data-driven research, informetrics plays a vital role in the evaluation of research entities. The focus of this research has expanded to include entities such as the datasets used in papers, genes, or drugs mentioned in papers as a focus for analysis (Ding et al., 2013). More broadly, the scientific community has been calling for scrutiny of the practice and reproducibility of research, particularly in the biomedical arena (Researching the Researchers, 2014). Techniques used in informetrics research can play a major role in this endeavor. Recently developed methods, as outlined in this book—such as data and text mining methods, network analysis, and advanced statistical techniques to reveal hidden relationships or patterns within large datasets—are quickly becoming valuable tools for the assessment of scholarly impact.

To date, there have been only a small number of monographs that have addressed informetrics-related topics. None provide a comprehensive treatment of recent developments or hands-on perspectives on how to apply these new techniques. This book fills that gap. The objective of this edited work is to provide an authoritative handbook of current topics, technologies, and methodological approaches that may be used for the study of scholarly impact. The chapters have been contributed by leading international researchers. Readers of this work should bring a basic familiarity with the field of scholarly communication and informetrics, as well as some understanding of statistical methods. However, the tools and techniques presented should also be accessible and usable by readers who are relatively new to the study of informetrics.

Each contributed chapter provides an introduction to the selected topic and outlines how the topic, technology, or methodological approach may be applied to informetrics-related research. The contributed chapters are grouped into four themes: Network Tools and Analysis, the Science System, Statistical and Text-based Methods, and Visualization. The book concludes with a chapter by Börner and Polley that brings together a number of the ideas presented in the earlier chapters.

A summary of each chapter's focus, methods outlined, software tools applied (where applicable), and data sources used (where applicable) appears below.

Network Tools and Analysis

Chapter 1
Title: Community detection and visualization of networks with the map equation framework.
Author(s): Ludvig Bohlin (Sweden), Daniel Edler (Sweden), Andrea Lancichinetti (Sweden), and Martin Rosvall (Sweden).
Topic(s): Networks.
Aspect(s): Community detection, visualization.
Method(s): Map equation.
Software tool(s) used: Infomap, MapEquation software package.
Data source: None.

Chapter 2
Title: Link prediction.
Author(s): Raf Guns (Belgium).
Topic(s): Networks.
Aspect(s): Link prediction.
Method(s): Data gathering–preprocessing–prediction–evaluation; recall–precision charts; using predictors such as common neighbors, cosine, degree product, SimRank, and the Katz predictor.
Software tool(s) used: linkpred; Pajek; VOSViewer; Anaconda Python.
Data source: Web of Science (Thomson Reuters)—co-authorship data of informetrics researchers.

Chapter 3
Title: Network analysis and indicators.
Author(s): Staša Milojević (USA).
Topic(s): Network analysis—network indicators.
Aspect(s): Bibliometric applications.
Method(s): Study of collaboration and citation links.
Software tool(s) used: Pajek; Sci2.
Data source(s): Web of Science (Thomson Reuters)—articles published in the journal *Scientometrics* over the period 2003–2012.

Chapter 4
Title: PageRank-related methods for analyzing citation networks.
Author(s): Ludo Waltman (The Netherlands) and Erjia Yan (USA).
Topic(s): Citation networks.
Aspect(s): Roles played by nodes in a citation network and their importance.
Method(s): Page-rank-related methods.
Software tool(s) used: Sci2; MATLAB; Pajek.
Data source: Web of Science (Thomson Reuters)—all publications in the journal subject category Information Science and Library Science that are of document type article, proceedings paper, or review and that appeared between 2004 and 2013.

The Science System

Chapter 5
Title: Systems Life Cycle and its relation with the Triple Helix.
Author(s): Robert K. Abercrombie (USA) and Andrew S. Loebl (USA).
Topic(s): Life cycle.
Aspect(s): Seen from a triple helix aspect.
Method(s): Technology Readiness Levels (TRLs).
Software tool(s) used: None.
Data source: From Lee et al. "Continuing Innovation in Information Technology." Washington, DC: The National Academies Press; plus diverse other sources.

Chapter 6
Title: Spatial scientometrics and scholarly impact: A review of recent studies, tools and methods.
Author(s): Koen Frenken (The Netherlands) and Jarno Hoekman (The Netherlands).
Topic(s): Spatial scientometrics.
Aspect(s): Scholarly impact, particularly, the spatial distribution of publication and citation output, and geographical effects of mobility and collaboration on citation impact.
Method(s): Review.
Software tool(s) used: None.
Data source: Web of Science (Thomson Reuters): post 2008.

Chapter 7
Title: Researchers' publication patterns and their use for author disambiguation.
Author(s): Vincent Larivière and Benoit Macaluso (Canada).
Topic(s): Authors.
Aspect(s): Name disambiguation.
Method(s): Publication patterns.
Software tool(s) used: None.
Data source: List of distinct university-based researchers in Quebec; classification scheme used by the US National Science Foundation (NSF); Web of Science (Thomson Reuters); Google.

Chapter 8
Title: Knowledge integration and diffusion: Measures and mapping of diversity and coherence.
Author(s): Ismael Rafols (Spain and UK).
Topic(s): Knowledge integration and diffusion.
Aspect(s): Diversity and coherence.

Method(s): Presents a conceptual framework including cognitive distance (or proximity) between the categories that characterize the body of knowledge under study.
Software tool(s) used: Leydesdorff's overlay toolkit; Excel; Pajek; additional software available at http://www.sussex.ac.uk/Users/ir28/book/excelmaps.
Data source: Web of Science (Thomson Reuters)—citations of the research center ISSTI (University of Edinburgh) across different Web of Science categories.

Statistical and Text-Based Methods

Chapter 9
Title: Limited dependent variable models and probabilistic prediction in informetrics.
Author(s): Nick Deschacht (Belgium) and Tim C.E. Engels (Belgium).
Topic(s): Regression models.
Aspect(s): Studying the probability of being cited.
Method(s): logit model for binary choice; ordinal regression; models for multiple responses and for count data.
Software tool(s) used: Stata.
Data source: Web of Science—Social Sciences Citation Index (Thomson Reuters)—2,271 journal articles published between 2008 and 2011 in five library and information science journals.

Chapter 10
Title: Text mining with the Stanford CoreNLP.
Author(s): Min Song (South Korea) and Tamy Chambers (USA).
Topic(s): Text mining.
Aspect(s): For bibliometric analysis.
Method(s): Provides an overview of the architecture of text mining systems and their capabilities.
Software tool(s) used: Stanford CoreNLP.
Data source(s): Titles and abstracts of all articles published in the *Journal of the American Society for Information Science and Technology* (JASIST) in 2012.

Chapter 11
Title: Topic Modeling: Measuring scholarly impact using a topical lens.
Author(s): Min Song (South Korea) and Ying Ding (USA).
Topic(s): Topic modeling.
Aspect(s): Bibliometric applications.
Method(s): Latent Dirichlet Allocation (LDA).
Software tool(s) used: Stanford Topic Modeling Toolbox (TMT).

Data source(s): Web of Science (Thomson Reuters)—papers published in the *Journal of the American Society for Information Science* (*and Technology*) (JASIS(T)) between 1990 and 2013.

Chapter 12
Title: The substantive and practical significance of citation impact differences between institutions: Guidelines for the analysis of percentiles using effect sizes and confidence intervals.
Author(s): Richard Williams (USA) and Lutz Bornmann (Germany).
Topic(s): Analysis of percentiles.
Aspect(s): Difference in citation impact.
Method(s): Statistical analysis using effect sizes and confidence intervals.
Software tool(s) used: Stata.
Data source: InCites (Thomson Reuters)—citation data for publications produced by three research institutions in German-speaking countries from 2001 to 2002.

Visualization

Chapter 13
Title: Visualizing bibliometric networks.
Author(s): Nees Jan van Eck (The Netherlands) and Ludo Waltman (The Netherlands).
Topic(s): Bibliometric networks.
Aspect(s): Visualization.
Method(s): As included in the software tools; tutorials.
Software tool(s) used: VOSviewer; CitNetExplorer.
Data source: Web of Science (Thomson Reuters)—journals *Scientometrics* and *Journal of Informetrics* and journals in their citation neighborhood.

Chapter 14
Title: Replicable science of science studies.
Author(s): Katy Börner (USA) and David E. Polley (USA).
Topic(s): Science of Science.
Aspect(s): Data preprocessing, burst detection, visualization, geospatial, topical and network analysis, career trajectories.
Method(s): Use of freely available tools for the actions described under “aspects.”
Software tool(s) used: Sci2 toolset.
Data source: Data downloaded from the Scholarly Database.

References

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to 'webometrics'. *Journal of Documentation*, *53*(4), 404–426.

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, *55*(14), 1216–1227.

De Bellis, N. (2009). *Bibliometrics and citation analysis*. Lanham, MD: Scarecrow Press.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities. *PloS One*, *8*(8), e71416.

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics*: *Quantitative methods in library*, *documentation and information science*. Elsevier Science Publishers. Retrieved from https://uhdspace.uhasselt.be/dspace/handle/1942/587.

Nacke, O. (1979). Informetrie: Ein neuer name für eine neue disziplin. *Nachrichten für Dokumentation*, *30*(6), 219–226.

Nalimov, V. V., & Mul'chenko, Z. M. (1969). На*укометрия*, *Изучение развития науки как информационного процесса* [*Naukometriya, the study of the development of science as an information process*]. Moscow: Nauka.

Otlet, P. (1934). *Traité de documentation: Le livre sur le livre*. Bruxelles, Éditions Mundaneum.

Pritchard, A. (1969). Statistical Bibliography or Bibliometrics? *Journal of Documentation*, *25*(4), 348–349.

Researching the Researchers [Editorial]. (2014). *Nature Genetics*, *46*(5), 417.

Bloomington, IN, USA — Ying Ding
Antwerp, Belgium — Ronald Rousseau
Milwaukee, WI, USA — Dietmar Wolfram

Contents