Scientometrics (2020) 123:1193–1200
https://doi.org/10.1007/s11192-020-03438-0

Historical roots of Judit Bar-Ilan's research: a cited-references analysis using CRExplorer

Lutz Bornmann[1] · **Loet Leydesdorff**[2]

Received: 13 November 2019 / Published online: 6 May 2020

Abstract
Judit Bar-Ilan (JB) was an influential researcher in information science and scientometrics. She published more than 100 papers about different topics. We used the CRExplorer (see www.crexplorer.net) to investigate the historical roots of JB's research. In this program, the N_TOP10 indicator is available. We applied this indicator to identify those publications which have been very frequently cited by JB during several citing years. These might be the publications by which JB was mostly influenced in her research. Our results show that the identified publications are seminal works in information science and scientometrics as well as methodologically oriented publications dealing with text or content analyses as well as influence or distance measures.

Keywords Cited references · CRExplorer · Historical roots

Introduction

Bornmann and Marx (2013) proposed to complement the times cited perspective (the forward view in impact measurement) with the cited references perspective (the backward view; Leydesdorff and Amsterdamska 1990; Merton 1965; Zitt and Small 2008). Whereas the times cited perspective focusses on the later impact of a paper, the backward view is oriented towards the roots of a paper: which are the giants on which the research published in the paper stand (Merton 1965)? Based on the proposal of using the backwards view in impact measurement, Thor et al. (2016a) introduced the CRExplorer (see www.crexp

This paper is dedicated to the memory of Judit Bar-Ilan (1958–2019), an outstanding scholar and an inimitable friend and colleague.

✉ Lutz Bornmann
bornmann@gv.mpg.de

Loet Leydesdorff
loet@leydesdorff.net

[1] Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany

[2] Amsterdam School of Communication Research (ASCoR), University of Amsterdam, PO Box 15793, 1001 NG Amsterdam, The Netherlands

lorer.net)—a program that can be used to investigate the historical roots of various entities in science: single researchers, topics, fields, institutions, etc. (see also Thor et al. 2016b). Since its introduction, the program has been used, for instance, to investigate the roots of the field of citation analysis (Hou 2017) and the research landscape associated with Monoamine oxidases (Yeung et al. 2019).

Some month ago, the scientometrics community has lost an outstanding researcher. Judit Bar-Ilan (JB) was professor at the Department of Information Science (Bar-Ilan University, Israel) and received the Derek de Solla Price Memorial Medal in 2017 for her contributions to the fields of quantitative studies of science. As a search in Web of Science (WoS, Clarivate Analytics) using her ResearcherID (B-3452-2009) shows, she has published 117 papers between 1989 and 2018.[1] Most of the papers (87%) are in the core WoS category of scientometric research "Information Science Library Science"; nearly one quarter of the papers have been published in *Scientometrics* (Leydesdorff & Bornmann, in press). In this study, the results of a cited references analysis are presented investigating the historical roots of JB's research in information science and scientometrics.

Methods

The 117 papers, which resulted from a search in WoS using JB's ResearcherID (B-3452-2009), were downloaded as comma-separated values (CSV) and imported in CRExplorer. The dataset contained 4182 non-distinct cited references, which was reduced to 3301 distinct references. Sixty-three cited references were discarded from the set, because they did not have reference publication year information (which is necessary for conducting a cited references analysis). The minimum reference publication year is 1934 and the maximum 2018. Since cited references data are often misspelled, we used the disambiguation tools provided by CRExplorer to identify and unify the variants. This procedure reduced the set of cited references to $n = 3295$ which have been used for the statistical analysis.

Results

In this study, JB's historical roots are defined as those publications cited by JB very frequently over many citing years. For identifying these publications, Thor et al. (2018) introduced the indicator N_TOP10; it is the number of citing years in which a cited publication (reference) belongs to the 10% most frequently referenced publications. The indicator assumes that the higher this number is, the more important or influential the cited publication (reference) had been for JB's research. Note that the indicator is calculated based on only JB's publications set. N_TOP10 is not connected to the well-known $PP_{top\text{-}10\%}$ indicator or excellence rate (Bornmann et al. 2012; Waltman et al. 2012). For these indicators, reference sets are generated which are not part of the publication set in question. For calculating the indicators for a single paper in a set, the 10% most frequently cited papers in the

[1] In the WoS, slightly more papers can be found for JB. However, we focused in this study on her "curated" list of papers in Publons (Clarivate Analytics). Historical analyses identifying frequently referenced publications are relatively robust against small variations in the underlying dataset.

Table 1 Historical roots of Judit Bar-Ilan's work (cited publications with the highest number of citing years in which the publications belong to the 10% most frequently referenced publications)

Cited publication (reference)	N_TOP10
"*Accessibility of information on the web*" (Lawrence and Giles 1999): "Search engines do not index sites equally, may not index new pages for months, and no engine indexes more than about 16% of the web. As the web becomes a major communications medium, the data on it must be made more accessible"	7
"*Content analysis: an introduction to its methodology*" (Krippendorff 1980): "What matters in people's social lives? What motivates and inspires our society? How do we enact what we know? Since the first edition published in 1980, Content Analysis has helped shape and define the field. In the highly anticipated Fourth Edition, award-winning scholar and author Klaus Krippendorff introduces you to the most current method of analyzing the textual fabric of contemporary society. Students and scholars will learn to treat data not as physical events but as communications that are created and disseminated to be seen, read, interpreted, enacted, and reflected upon according to the meanings they have for their recipients. Interpreting communications as texts in the contexts of their social uses distinguishes content analysis from other empirical methods of inquiry"	7
"*The calculation of web impact factors*" (Ingwersen 1998): "This case study reports the investigations into the feasibility and reliability of calculating impact factors for web sites, called Web Impact Factors (Web-IF). The study analyses a selection of seven small and medium scale national and four large web domains as well as six institutional web sites over a series of snapshots taken of the web during a month. The data isolation and calculation methods are described and the tests discussed. The results thus far demonstrate that Web-IFs are calculable with high confidence for national and sector domains whilst institutional Web-IFs should be approached with caution. The data isolation method makes use of sets of inverted but logically identical Boolean set operations and their mean values in order to generate the impact factors associated with internal- (self-) link web pages and external-link web pages. Their logical sum is assumed to constitute the workable frequency of web pages linking up to the web location in question. The logical operations are necessary to overcome the variations in retrieval outcome produced by the AltaVista search engine"	6
"*Citation influence for journal aggregates of scientific publications: theory, with application to literature of physics*" (Pinski and Narin 1976): "A self-consistent methodology is developed for determining citation based influence measures for scientific journals, subfields and fields. Starting with the cross citing matrix between journals or between aggregates of journals, an eigenvalue problem is formulated leading to a size independent influence weight for each journal or aggregate. Two other measures, the influence per publication and the total influence are then defined. Hierarchical influence diagrams and numerical data are presented to display journal interrelationships for journals within the subfields of physics. A wide range in influence is found between the most influential and least influential or peripheral journals"	6
"*An index to quantify an individual's scientific research output*" (Hirsch 2005): "I propose the index *h,* defined as the number of papers with citation number > *h,* as a useful index to characterize the scientific output of a researcher"	5
"*Relevance: a review of and a framework for the thinking on the notion in information science*" (Saracevic, 1975): "Information science emerged as the third subject, along with logic and philosophy, to deal with relevance-an elusive, human notion. The concern with relevance, as a key notion in information science, is traced to the problems of scientific communication. Relevance is considered as a measure of the effectiveness of a contact between a source and a destination in a communication process. The different views of relevance that emerged are interpreted and related within a framework of communication of knowledge. Different views arose because relevance was considered at a number of different points in the process of knowledge communication. It is suggested that there exists an interlocking, interplaying cycle of various systems of relevances"	5
"*Automatic text processing: the transformation, analysis, and retrieval of information by computer*" (Salton 1989): a description of the content of the book is not available (see also Salton 1970)	5

Table 1 (continued)

Cited publication (reference)	N_TOP10
"*A technique for measuring the relative size and overlap of public web search engines*" (Bharat and Broder 1998): "Search engines are among the most useful and popular services on the Web. Users are eager to know how they compare. Which one has the largest coverage? Have they indexed the same portion of the Web? How many pages are out there? Although these questions have been debated in the popular and technical press, no objective evaluation methodology has been proposed and few clear answers have emerged. In this paper we describe a standardized, statistical way of measuring search engine coverage and overlap through random queries. Our technique does not require privileged access to any database. It can be implemented by third-party evaluators using only public query interfaces. We present results from our experiments showing size and overlap estimates for HotBot, AltaVista, Excite, and Infoseek as percentages of their total joint coverage in mid 1997 and in November 1997. Our method does not provide absolute values. However using data from other sources we estimate that as of November 1997 the number of pages indexed by HotBot, AltaVista, Excite, and Infoseek were respectively roughly 77 M, 100 M, 32 M, and 17 M and the joint total coverage was 160 million pages. We further conjecture that the size of the static, public Web as of November was over 200 million pages. The most startling finding is that the overlap is very small: less than 1.4% of the total coverage, or about 2.2 million pages were indexed by all four engines"	5
"*Theory and practise of the g-index*" (Egghe 2006): "The *g*-index is introduced as an improvement of the *h*-index of Hirsch to measure the global citation performance of a set of articles. If this set is ranked in decreasing order of the number of citations that they received, the *g*-index is the (unique) largest number such that the top *g* articles received (together) at least $g2$ citations. We prove the unique existence of *g* for any set of articles and we have that g^3 h. The general Lotkaian theory of the *g*-index is presented and we show that … where $a > 2$ is the Lotkaian exponent and where *T* denotes the total number of sources. We then present the *g*-index of the (still active) Price medallists for their complete careers up to 1972 and compare it with the *h*-index. It is shown that the *g*-index inherits all the good properties of the *h*-index and, in addition, better takes into account the citation scores of the top articles. This yields a better distinction between and order of the scientists from the point of view of visibility"	5
"*The anatomy of a large-scale hypertextual Web search engine*" (Brin and Page 1998): "In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at https://google.stanford.edu/ To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of Web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the Web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and Web proliferation, creating a Web search engine today is very different from three years ago. This paper provides an in-depth description of our large-scale Web search engine—the first such detailed public description we know of to date. Apart from the problems of scaling traditional search techniques to data of this magnitude, there are new technical challenges involved with using the additional information present in hypertext to produce better search results. This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext. Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want"	5

Table 1 (continued)

Cited publication (reference)	N_TOP10
"*Comparing top k lists*" (Fagin et al. 2003): "Motivated by several applications, we introduce various distance measures between 'top *k* lists.' Some of these distance measures are metrics, while others are not. For each of these latter distance measures, we show that they are 'almost' a metric in the following two seemingly unrelated aspects: (1) they satisfy a relaxed version of the polygonal (hence, triangle) inequality, and (2) there is a metric with positive constant multiples that bound our measure above and below. This is not a coincidence—we show that these two notions of almost being a metric are the same. Based on the second notion, we define two distance measures to be *equivalent* if they are bounded above and below by constant multiples of each other. We thereby identify a large and robust equivalence class of distance measures. Besides the applications to the task of identifying good notions of (dis)similarity between two top *k* lists, our results imply polynomial-time constant-factor approximation algorithms for the *rank aggregation problem* with respect to a large class of distance measures. (A correction for this article has been appended to the pdf file.)	5

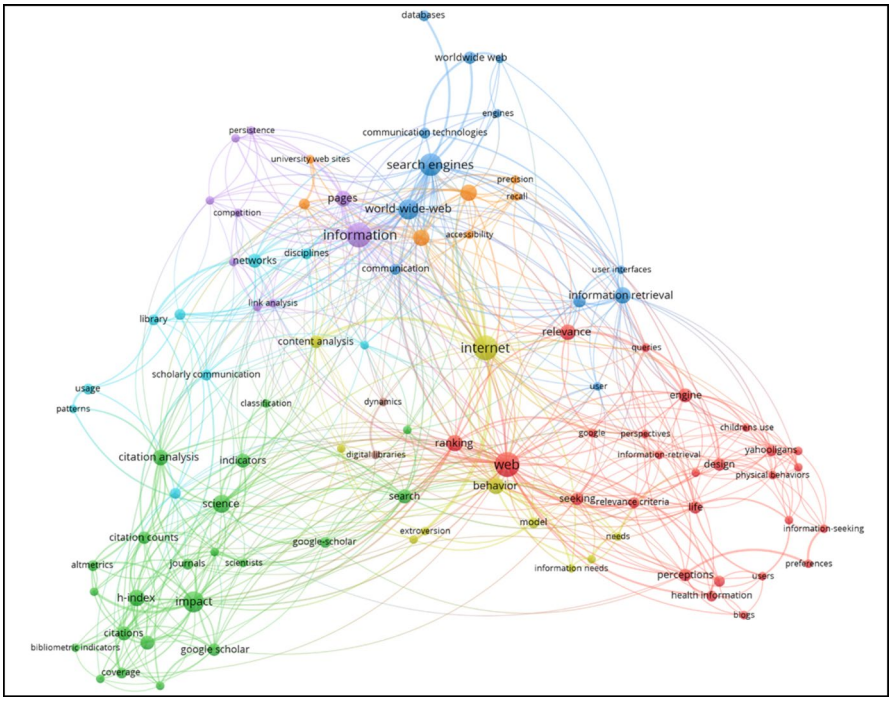

Fig. 1 Co-occurrence network of keywords attributed to 117 papers published by Judit Bar-Ilan (based on minimum number of occurrences of keywords = 2)

corresponding subject category (e.g., used in Scopus, Elsevier, or WoS) and publication year are determined (see Bornmann 2013).

Table 1 shows the title of the publications, which belong in at least five citing years to the 10% most frequently referenced publications by JB. The table includes also the abstracts of papers or short descriptions in case of books (when available). To support the interpretation of the historical root publications in Table 1, a co-occurrence network has

been generated based on the keywords (author keywords and KeyWords Plus) from JB's 117 papers. The network, which we produced with the program VOSviewer (see www.vosviewer.com), visualizes the topics of JB's research (see Fig. 1). As the network results reveal, JB was active in various topics of information science and scientometrics: information retrieval (red, dark-blue nodes), internet—world-wide-web—research (blue, yellow nodes), information behaviour (dark-blue nodes), library metrics (bright-blue nodes), altmetrics (green nodes), and *h* index (green nodes).

JB's historical roots publications in Table 1 fit very well with JB's research topics as visualized in Fig. 1: A seminal publication in information science is Saracevic (1975). Krippendorff (1980) and Salton (1989) deal with methods for analyzing the content of text documents (see also Salton 1970). These methods are relevant in research on information retrieval and information behaviour. Krippendorff (1980) is the central textbook for content analysis. Basic publications about the Internet—world-wide-web—research and search engines are Brin and Page (1998)—the paper grounding Google—and Bharat and Broder (1998), as well as Lawrence and Giles (1999). Lawrence and Giles (1999) is the locus classicus for research about search engines. The connection between the world-wide-web and the impact factor was made by Ingwersen (1998). This paper introduced the impact factor into webometrics. The *h* index has been introduced by Hirsch (2005) and Egghe (2006) proposed one of the most important *h* index variants, namely the *g* index (Bornmann and Daniel 2007; Bornmann et al. 2011). Pinski and Narin (1976) as well as Fagin et al. (2003) are methodologically oriented papers dealing with citation based influence measures and distance measures. Pinski and Narin (1976) is the classical paper about influence weights.

Discussion

JB was one of the most influential researchers in information science and scientometrics. She published more than 100 papers about different topics in both these fields. In this study, the historical roots of JB's research have been investigated using the N_TOP10 indicator: publications were identified which have been very frequently cited by JB in several citing years. These publications are mostly seminal works in information science and scientometrics as well as methodologically oriented publications dealing with text or content analyses as well as influence or distance measures.

In recent years, historical roots of various units have been investigated in many studies based on cited references data (e.g., Ballandonne 2018; Barth et al. 2014). Advanced indicators such as N_TOP10 introduced recently by Thor et al. (2018) have been seldomly used in these studies, although the indicators have the advantage of supporting the identification of landmark publications referenced in publication sets. Since the analysis of JB's publication set is a good example for the usefulness of the indicators, this study might encourage scientometricians to use them in future studies.

Acknowledgements Open Access funding provided by Projekt DEAL.

References

Ballandonne, M. (2018). The historical roots (1880–1950) of recent contributions (2000–2017) to ecological economics: insights from reference publication year spectroscopy. *Journal of Economic Methodology*. https://doi.org/10.1080/1350178X.2018.1554227.

Barth, A., Marx, W., Bornmann, L., & Mutz, R. (2014). On the origins and the historical roots of the Higgs boson research from a bibliometric perspective. *The European Physical Journal Plus, 129*(6), 1–13. https://doi.org/10.1140/epjp/i2014-14111-6.

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public web search engines. *Computer Networks and ISDN Systems, 30*(1–7), 379–388.

Bornmann, L. (2013). How to analyze percentile citation impact data meaningfully in bibliometrics: The statistical analysis of distributions, percentile rank classes, and top-cited papers. *Journal of the American Society for Information Science and Technology, 64*(3), 587–595. https://doi.org/10.1002/asi.22792.

Bornmann, L., & Daniel, H.-D. (2007). What do we know about the *h* index? *Journal of the American Society for Information Science and Technology, 58*(9), 1381–1385. https://doi.org/10.1002/asi.20609.

Bornmann, L., de Moya Anegón, F., & Leydesdorff, L. (2012). The new excellence Indicator in the World Report of the SCImago Institutions Rankings 2011. *Journal of Informetrics, 6*(2), 333–335. https://doi.org/10.1016/j.joi.2011.11.006.

Bornmann, L., & Marx, W. (2013). The proposal of a broadening of perspective in evaluative bibliometrics by complementing the times cited with a cited reference analysis. *Journal of Informetrics, 7*(1), 84–88. https://doi.org/10.1016/j.joi.2012.09.003.

Bornmann, L., Mutz, R., Hug, S., & Daniel, H. (2011). A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *Journal of Informetrics, 5*(3), 346–359. https://doi.org/10.1016/j.joi.2011.01.006.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X.

Egghe, L. (2006). Theory and practise of the *g*-index. *Scientometrics, 69*(1), 131–152. https://doi.org/10.1007/s11192-006-0144-7.

Fagin, R., Kumar, R., & Sivakumar, D. (2003). Comparing top k lists. *SIAM Journal on discrete mathematics, 17*(1), 134–160.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102.

Hou, J. (2017). Exploration into the evolution and historical roots of citation analysis by referenced publication year spectroscopy. *Scientometrics, 110*(3), 1437–1452. https://doi.org/10.1007/s11192-016-2206-9.

Ingwersen, P. (1998). The calculation of web impact factors. *Journal of Documentation, 54*(2), 236–243.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology* (1st ed.). Newcastle upon Tyne, UK: SAGE Publications.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. *Nature, 400*(6740), 107.

Leydesdorff, L., & Amsterdamska, O. (1990). Dimensions of citation analysis. *Science Technology and Human Values, 15*(3), 305–335. https://doi.org/10.1177/016224399001500303.

Leydesdorff, L., & Bornmann, L. (in press). "Interdisciplinarity" and "Synergy" in the Œuvre of Judit Bar-Ilan. *Scientometrics*.

Merton, R. K. (1965). *On the shoulders of giants*. New York, NY: Free Press.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to literature of physics. *Information Processing and Management, 12*(5), 297–312.

Salton, G. (1970). Automatic text analysis. *Science, 168*(3929), 335–343. https://doi.org/10.1126/science.168.3929.335.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Boston, MA: Addison-Wesley Longman Publishing Co., Inc.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science, 26*(6), 321–343.

Thor, A., Bornmann, L., Marx, W., & Mutz, R. (2018). Identifying single influential publications in a research field: New analysis opportunities of the CRExplorer. *Scientometrics, 116*(1), 591–608.

Thor, A., Marx, W., Leydesdorff, L., & Bornmann, L. (2016a). Introducing CitedReferencesExplorer (CRExplorer): A program for reference publication year spectroscopy with cited references standardization. *Journal of Informetrics, 10*(2), 503–515. https://doi.org/10.1016/j.joi.2016.02.005.

Thor, A., Marx, W., Leydesdorff, L., & Bornmann, L. (2016b). New features of CitedReferencesExplorer (CRExplorer). *Scientometrics, 109*(3), 2049–2051. https://doi.org/10.1007/s11192-016-2082-3.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., et al. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology, 63*(12), 2419–2432.

Yeung, A. W. K., Georgieva, M. G., Atanasov, A. G., & Tzvetkov, N. T. (2019). Monoamine oxidases (MAOs) as privileged molecular targets in neuroscience: Research literature analysis. *Frontiers in Molecular Neuroscience*. https://doi.org/10.3389/fnmol.2019.00143.

Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology, 59*(11), 1856–1860. https://doi.org/10.1002/asi.20880.