Journal of Informetrics 12 (2018) 679–702

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

Author ranking evaluation at scale

Marcel Dunaiski[a,b,*], Jaco Geldenhuys[b], Willem Visser[b]

[a] Media Lab, Stellenbosch University, 7602 Matieland, South Africa
[b] Department of Computer Science, Stellenbosch University, 7602 Matieland, South Africa

ARTICLE INFO

Article history:
Received 9 October 2017
Received in revised form 1 June 2018
Accepted 10 June 2018

Keywords:
Author indicators
Author ranking
Citation analysis
PageRank

ABSTRACT

We evaluate author impact indicators and ranking algorithms on two publication databases using large test data sets of well-established researchers. The test data consists of (1) ACM fellowship and (2) various life-time achievement awards. We also evaluate different approaches of dividing credit of papers among co-authors and analyse the impact of self-citations. Furthermore, we evaluate different graph normalisation approaches for when PageRank is computed on author citation graphs.

We find that PageRank outperforms citation counts in identifying well-established researchers. This holds true when PageRank is computed on author citation graphs but also when PageRank is computed on paper graphs and paper scores are divided among co-authors. In general, the best results are obtained when co-authors receive an equal share of a paper's score, independent of which impact indicator is used to compute paper scores. The results also show that removing author self-citations improves the results of most ranking metrics. Lastly, we find that it is more important to personalise the PageRank algorithm appropriately on the paper level than deciding whether to include or exclude self-citations. However, on the author level, we find that author graph normalisation is more important than personalisation.

1. Introduction

To empirically answer questions about bibliometric indicators, a representative publication database and appropriate evaluation data (or test data) are required. One problem of evaluating indicators and algorithms that measure academic quality or impact is the difficulty of obtaining appropriate test data. This problem is compounded by the subjectivity of what is considered quality or impact, and generally requires human judgment. Due to this drawback, correlation analyses are often performed which are problematic on their own (Thelwall, 2016) and only provide comparatives to some baselines, usually citation counts used as a proxy for quality.

Another option that is often employed is the use of relatively small test data sets that are based on some external knowledge. The assumption is that the author entities in the test data exhibit some property (e.g., are highly influential or well-established) that is not exclusively based on citations. Therefore, these test data sets are often used to evaluate the functionality of the ranking algorithms to identify the comprising entities and consequently their shared property. Examples of such applications are: evaluating author ranking algorithms in identifying well-established researchers using test data that comprises researchers that have received fellowship status at learned societies, have won life-time contribution awards,

* Corresponding author at: Media Lab, Stellenbosch University, 7602 Matieland, South Africa.
E-mail address: marcel@ml.sun.ac.za (M. Dunaiski).

https://doi.org/10.1016/j.joi.2018.06.004
1751-1577/

or are frequently board members of prestigious journals (Dunaiski, Visser, & Geldenhuys, 2016; Fiala, Šubelj, Žitnik and Bajec, 2015; Fiala, Rousselot, & Ježek, 2008; Fiala & Tutoky, 2017; Gao, Wang, Li, Zhang, & Zeng, 2016; Nykl, Ježek, Fiala, & Dostal, 2014); evaluating the performance of paper-level ranking algorithms in finding impactful papers using test data that comprises best paper awards or high-impact paper awards (Dunaiski et al., 2016; Dunaiski & Visser, 2012; Mariani, Medo, & Zhang, 2016; Sidiropoulos & Manolopoulos, 2005); and to showcase the applicability of newly proposed indicators (Gao et al., 2016). Very rarely, direct peer-reviewed opinions are used to evaluate metrics (i.e., Abramo & D'Angelo, 2015) since this type of information is often not readily available.

Nykl et al. (2014) analyse various PageRank approaches and the effects that author graph normalisations and self-citations have on the ranking of authors. As evaluation data they use 54 authors that have won one of two prestigious computer science awards, a set of 576 researchers that have received fellowships of the Association for Computing Machinery (ACM) (ACM, Inc., 2017b), and a list of 280 highly cited researchers. The data they use for the experiments is a subset of the Web Of Science (Clarivate Analytics, 2017) publication data consisting of 149 347 papers published in 386 computer science journals between 1996 and 2005.

Later, Nykl, Campr, and Ježek (2015) extend this research to include different schemes to answer the question of how the credit of a paper should be shared among its co-authors. They again use the ACM fellows as evaluation data and two different lists of author names in the computer science fields of "artificial intelligence" and "hardware" with 354 and 158 authors, respectively. These lists comprise authors that have won contribution awards, but also authors that have written papers that have won best paper awards or influential paper awards, which are usually handed out about 10 years after initial publication for their outstanding impact in their fields.

In this paper, we reproduce and extend the above mentioned work by Nykl et al. (2014) and Nykl et al. (2015) with a more in-depth analysis of the results. The aim of the paper is to identify the results that generalise by using two larger test data sets and two publication databases, one of which is multi-disciplinary. Furthermore, we include other author impact indicators in the evaluation such as a percentile-based indicator R6 (Leydesdorff, Bornmann, Mutz, & Opthof, 2011) and the *PR*-index (Gao et al., 2016), which combines PageRank and a variant of the h-index (Hirsch, 2005).

In addition, we analyse the impact that self-citations have on author impact indicators and evaluate different approaches of normalising the author citation graph for PageRank. Lastly, we analyse different approaches of computing impact scores for papers and how these scores should be distributed among co-authors to achieve the best ranking results in ranking well-established researchers.

With this paper, we also present a large test data set consisting of openly available information that can be used to evaluate author impact indicators. The test data comprises author lists of 27 awards handed out to 596 renowned researchers and 1000 authors that received fellowship accreditation by the ACM. We manually matched all researchers in the test data to two publication databases, the ACM's Digital Library (ACM, Inc., 2015) and Microsoft Academic Graph (MAG) (Microsoft, 2017b).

For the evaluation, we focus on variations of the PageRank algorithm (Brin & Page, 1998; Pinski & Narin, 1976) because it is frequently applied to academic citation networks to find important papers (Chen, Xie, Maslov, & Redner, 2007; Dunaiski & Visser, 2012; Hwang, Chae, Kim, & Woo, 2010) and on author citation graphs to rank authors (Dunaiski et al., 2016; Fiala & Tutoky, 2017; Nykl et al., 2015; West, Jensen, Dandrea, Gordon, & Bergstrom, 2013), and has continuously yielded good results as an impact indicator.

We use the average rank as an evaluation measure and use a new methodological approach to estimate the minimum difference required to conclude that rankings are significantly different (Dunaiski, Geldenhuys, & Visser, 2018). Applying this approach, we can compute the significance levels of the differences between two or more rankings. For example, how significant is the difference in the average rank of the authors in the test data when including or excluding self-citations for a certain metric?

With this paper we make the following contributions:

- We make a large test data set available consisting of researchers that won renowned prizes and researchers that are ACM fellows. The author names in these test data sets are matched to author entity identifiers of the ACM and MAG publication databases.
- Based on this test data, we show that using ranking algorithms based on PageRank outperform citation counts as impact indicator of well-established researchers.
- We show that almost all impact indicators are significantly improved by removing self-citations.
- We analyse the effects of different author graph normalisation approaches on the results of PageRank and find that it is more important to normalise the author citation graph than to personalise the PageRank algorithm.
- We find that evenly dividing paper scores among co-authors yields the best results by consistently ranking the authors in our test data higher, independent of which impact indicator is used to compute paper scores.

In this paper, we first review previously published work that uses either awards or fellowship information as test data to evaluate author impact indicators (Section 2). We then provide mathematical definitions of the author ranking algorithms used in this paper, as well as the definitions of the paper credit distribution functions and author citation graph normalisation