Journal of Informetrics 9 (2015) 334–348

Contents lists available at ScienceDirect

Journal of Informetrics

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

Do PageRank-based author rankings outperform simple citation counts?

Dalibor Fiala[a,*], Lovro Šubelj[b], Slavko Žitnik[b], Marko Bajec[b]

[a] University of West Bohemia, Department of Computer Science and Engineering, Univerzitní 8, 30614 Plzeň, Czech Republic
[b] University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 24 October 2014
Received in revised form 23 February 2015
Accepted 23 February 2015

Keywords:
PageRank
Scholars
Citations
Rankings
Importance

ABSTRACT

The basic indicators of a researcher's productivity and impact are still the number of publications and their citation counts. These metrics are clear, straightforward, and easy to obtain. When a ranking of scholars is needed, for instance in grant, award, or promotion procedures, their use is the fastest and cheapest way of prioritizing some scientists over others. However, due to their nature, there is a danger of oversimplifying scientific achievements. Therefore, many other indicators have been proposed including the usage of the PageRank algorithm known for the ranking of webpages and its modifications suited to citation networks. Nevertheless, this recursive method is computationally expensive and even if it has the advantage of favouring prestige over popularity, its application should be well justified, particularly when compared to the standard citation counts. In this study, we analyze three large datasets of computer science papers in the categories of artificial intelligence, software engineering, and theory and methods and apply 12 different ranking methods to the citation networks of authors. We compare the resulting rankings with self-compiled lists of outstanding researchers selected as frequent editorial board members of prestigious journals in the field and conclude that there is no evidence of PageRank-based methods outperforming simple citation counts.

1. Introduction and related work

Ranking researchers has become very popular due to the possible applications in various hiring, promotion, grant, or award procedures, in which manual assessment can be efficiently supplemented with automated techniques. Apart from counting the research money granted, the easiest way to evaluate a researcher's performance is to estimate the quantity and quality of scholarly publications he/she has produced. The former concentrates on production (or productivity) and the latter on impact (or influence). In its basic form, production is the number of research papers a scientist has published and impact is the number of citations from other research publications these papers have attracted. These two simple indicators may already form a basis for an easy ranking of researchers (or authors as all of these evaluations are based on the authorship of research publications). One of the drawbacks of this simplistic approach is that it does not differentiate between popularity and prestige, i.e. it considers all citations as equivalent. In the practice, however, a citation by a Nobel Prize laureate is certainly more valuable than that by a doctoral student, a citation by a scientist with a high number of citations has probably more weight than that by a scholar with only a few citations, and many citations from the same researcher are apparently less

[*] Corresponding author. Tel.: +420 377 63 24 29.
E-mail addresses: dalfia@kiv.zcu.cz (D. Fiala), lovro.subelj@fri.uni-lj.si (L. Šubelj), slavko.zitnik@fri.uni-lj.si (S. Žitnik), marko.bajec@fri.uni-lj.si (M. Bajec).

http://dx.doi.org/10.1016/j.joi.2015.02.008

worth than the same number of citations from many different scientists. All this motivated the application of "higher-order" evaluation methods (citations being a "first-order" method) such as PageRank to citation networks of authors.

The recursive PageRank algorithm by Brin and Page (1998), the founders of Google, was originally meant to evaluate the importance of webpages on the basis of the link structure of the web. The principal idea is that an important webpage is itself linked to from other important webpages. Thus, a webpage can have a high rank if it has inlinks from many webpages with low ranks but also if it has inlinks from few webpages with high ranks. The rank of a webpage depends on the ranks of the webpages linking to it. In practice, the costly calculation of PageRank in a directed graph is done in an iterative fashion and more on this will be said in the following section. Even though a similar bibliometric concept was introduced by Pinski and Narin (1976) long before Google, the PageRank's property of being applicable to any directed graph was soon utilized in the analysis of citation networks to rank journals (Bergstrom, 2007; Bollen, Rodriguez, & Van De Sompel, 2006; González-Pereira, Guerrero-Bote, & Moya-Anegón, 2010), papers (Chen, Xie, Maslov, & Redner, 2007; Ma, Guan, & Zhao, 2008; Walker, Xie, Yan, & Maslov, 2007; Yan & Ding, 2010), authors (Ding, Yan, Frazho, & Caverlee, 2009; Ding, 2011; Fiala, Rousselot, & Ježek, 2008; Fiala, 2011, 2012b, 2013a; Nykl, Ježek, Fiala, & Dostal, 2014; Radicchi, Fortunato, Markines, & Vespignani, 2009; Yan & Ding, 2011), a combination of the three (Yan, Ding, & Sugimoto, 2011), institutions (Yan, 2014), departments (Fiala, 2013b, 2014), countries (Fiala, 2012a; Ma et al., 2008), or a mixture of the above entities (West, Jensen, Dandrea, Gordon, & Bergstrom, 2013). In the many previous studies of ours we investigated various PageRank modifications with respect to the standard (baseline) PageRank and concluded that some of the variants performed better than the baseline in that they generated rankings closer to the human perception of a good ranking. In the present study, however, we consider simple citations as the baseline and the main research question is whether author rankings based on PageRank (and its variants) outperform citations in terms of better ranks assigned to outstanding researchers. If the answer was yes, the high computational cost of PageRank needed to overcome some deficiencies of citations would be well justified.

Let us remark in this place that PageRank-based (or, in general, recursive) ranking methods are only one branch of research performance evaluation techniques (in addition to standard publication and citation counts) with the other notable one being the family of h- and g-indices (Egghe, 2006; Hirsch, 2005) that combine both production and impact in a single number. These indices may obviously be used to rank authors as well, but they are not the concern of the present paper which is further organized as follows: In Section 2 we briefly recall the substance of PageRank, its modifications used in our analysis, and other related methods and refer to the relevant literature for more details. In Section 3 we describe the dataset we examined, which consists of papers from three large computer science categories (artificial intelligence, software engineering, and theory and methods). In Section 4 we present and discuss the main results of our analysis and give a negative answer to the main research question asked in the title of this article. And finally, in the last section, we summarize the most important contributions and results of this study and propose some research lines for our future work.

2. Methods

Let us define the directed author citation graph as $G=(V, E)$, where V is the set of vertices (authors) and E is the set of edges (unique citations between authors). If author v cites author u (once or more times), there is an edge $(v, u) \in E$. Then, by the recursive definition, the PageRank score PR(u) of author u depends on the scores of all citing authors in the following way:

$$PR(u) = \frac{1-d}{|V|} + d \sum_{(v,u) \in E} PR(v)\Omega \tag{1}$$

where d is the damping factor, which was set to 0.85 in the original web experiments by Brin and Page (1998), and Ω is either the multiplicative inverse of the out-degree of v like in the standard PageRank or $\sigma_{v,u}/\sum_{(v,k) \in E}\sigma_{v,k}$ like in the bibliographic PageRank by Fiala et al. (2008), where

$$\sigma_{v,k} = \frac{w_{v,k}}{[(c_{v,k}+1)/(b_{v,k}+1)]\sum_{(v,j) \in E} w_{v,j}} \tag{2}$$

with w, b, and c being various coefficients determined from both the citation and the collaboration networks of authors which will be explained below. Note that as follows from (1), an author with no citations (incoming edges) will still have a non-zero PageRank, which will be close to the multiplicative inverse of the total number of authors in the dataset. Of course, this will be influenced by the damping factor d, which was intitially determined empirically after the observation that a typical web user usually followed five links to other webpages and then chose a random webpage, e.g. by starting a new keyword search, thus resulting in about one sixth (≈ 0.15) of all transactions between webpages to be random. Indeed, the total PageRank in the system (or network) should be 1 and the individual PageRanks of vertices are then the fractions of time a random surfer spends there. We refer to the paper by Diligenti, Gori, and Maggini (2004) for an explanation of PageRank within a random walk framework. Other approaches to the PageRank problem include solving a linear system (Bianchini, Gori, & Scarselli, 2005; Langville & Meyer, 2004), but for practical reasons it is mostly computed dynamically in an iterative manner until convergence of subsequently generated rankings, which may be measured with Spearman's rank correlation coefficients. This is also the way we applied in our analysis with the maximum number of iterations set to 50, which was enough even with stricter convergence criteria and millions of nodes in the experiment by Brin and Page, and