Journal of Informetrics 8 (2014) 683–692

Contents lists available at ScienceDirect

Journal of Informetrics

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

Short Communication

PageRank variants in the evaluation of citation networks

Michal Nykl, Karel Ježek, Dalibor Fiala*, Martin Dostal

University of West Bohemia, Department of Computer Science and Engineering, Univerzitní 8, 30614 Plzeň, Czech Republic

ARTICLE INFO

Article history:
Received 26 February 2014
Received in revised form 6 May 2014
Accepted 10 June 2014

Keywords:
PageRank
Citation analysis
Research evaluation
Author ranking
ISI Web of Science

ABSTRACT

This paper explores a possible approach to a research evaluation, by calculating the renown of authors of scientific papers. The evaluation is based on the citation analysis and its results should be close to a human viewpoint. The PageRank algorithm and its modifications were used for the evaluation of various types of citation networks. Our main research question was whether better evaluation results were based directly on an author network or on a publication network. Other issues concerned, for example, the determination of weights in the author network and the distribution of publication scores among their authors. The citation networks were extracted from the computer science domain in the ISI Web of Science database. The influence of self-citations was also explored. To find the best network for a research evaluation, the outputs of PageRank were compared with lists of prestigious awards in computer science such as the Turing and Codd award, ISI Highly Cited and ACM Fellows. Our experiments proved that the best ranking of authors was obtained by using a publication citation network from which self-citations were eliminated, and by distributing the same proportional parts of the publications' values to their authors. The ranking can be used as a criterion for the financial support of research teams, for identifying leaders of such teams, etc.

1. Introduction

The evaluation of universities' prestige usually covers several areas such as research results, education, student satisfaction and others. When evaluating research, publications play an important role. Publications and their citations can best show the top researcher in the selected field of science. This evaluation is usually based on the number of publications indexed in e.g. the ISI Web of Science[1] (hereafter WoS) with regard to the number of citations and Journal Impact Factor (Garfield, 1972). The Impact Factor[2] of journal J in a given year (e.g. 2011) is the number of citations in this year (2011) to all items published in journal J two years before (2010 and 2009) divided by the number of journal J's citable items (i.e. excluding notes, editorials, etc.) published in those two years (2010 and 2009). Note that in the evaluation, the impact factor of citing journals is not taken into account.

Our main method for the evaluation of citation networks is the PageRank algorithm, which uses the impact of citing nodes (articles, authors and so on) for determining the importance of cited nodes. PageRank was introduced by Brin and Page (1998) to rank websites and became part of the Google search engine. From its introduction, PageRank has been

* Corresponding author. Tel.: +420 377 63 24 29.
E-mail addresses: nyklm@kiv.zcu.cz (M. Nykl), jezek_ka@kiv.zcu.cz (K. Ježek), dalfia@kiv.zcu.cz (D. Fiala), madostal@kiv.zcu.cz (M. Dostal).

[1] *ISI Web of Science* – http://www.webofknowledge.com.

[2] Computation of Journal Impact Factor in the WoS database – http://admin-apps.webofknowledge.com/JCR/help/h_impfact.htm.

http://dx.doi.org/10.1016/j.joi.2014.06.005
1751-1577/

examined for convergence, acceleration, rating prediction, etc. For example, Langville and Meyer (2006) is a good starting point for its deeper study.

PageRank has been frequently used for citation analysis. Fiala (2012) worked with the publication citation network and the authorship network to create an author citation network. The determination of edge weights with regard to the publication date and co-authorship is also solved. Other variants of bibliographic network evaluations (comprising e.g. co-citation or co-authorship network) are compared by Yan and Ding (2012). Sidiropoulos and Manolopoulos (2006) used the list of *ACM SIGMOD E. F. Codd Innovation Award* holders to compare the results of human and machine rankings of authors. We used the same approach to determine the quality of author rankings but also used some other human evaluation methods. Yu, Chen, and Chen (2012) explored a network which combines information from citations, reviews, comments, and information on the reputation of social network users who read articles and comment on them. A comparison and combination of PageRank and the journal impact factor are presented by Bollen, Rodriquez, and Van De Sompel (2006).

Our main research question was whether better evaluation results were based directly on an author network or on a publication network. We investigate several variants of author or publication citation networks. The influence of self-citations is explored and a further two variants of author ratings are proposed and studied. The author's rating can be obtained either from the weighted author citation networks or as a distribution of publication values among their authors. Other questions, therefore, concern, for example, how to determine the weights in the author network and how to distribute the publication scores among their authors. The evaluation results are compared with lists of the holders of four prestigious computer science awards. Our main contribution demonstrates that the best ranking of authors is obtained by using a publication citation network from which self-citations are eliminated and by distributing the same proportional parts of the publications' values to their authors.

The following section describes the data from the WoS collection, the used lists of prestigious awards and the construction of citation networks of papers or authors. The *Types of citation networks* section provides information on how to add weights to edges in the author citation network and how to distribute the publication scores among authors. The next section is devoted to our modifications of the PageRank algorithm. The experiments and their results are summarized in the *Experiments* section and discussed in the *Discussion* section. The conclusion and recommendation are presented in the last section.

2. Data used

All of our experiments can be run on an arbitrary bibliographic data collection, but we used the already purchased Thomson Reuters collection employed in our previous studies (Fiala, 2012). This collection consists of all publications classified as "*article*" published in Journal Citation Reports 2009 in the computer science category between 1996 and 2005. This category covers all seven WoS subcategories: Artificial Intelligence, Cybernetics, Hardware & Architecture, Information Systems, Interdisciplinary Applications, Software Engineering and Theory & Methods.

Using this data, we create two citation networks – the publication network and the author network. The networks can consider various types of self-citations (see Fig. 1). The first variant, marked *ALL*, takes into account all citations and is, therefore, the most benevolent. The second variant, marked *NOT*, removes citations between publications having at least one common author. For this reason, it is the strictest variant. The last variant is marked *PART*. It is applicable only to author networks and is created from the *ALL* variant by removing all self-loops. Other variants of self-citations are mentioned by Yan, Ding, and Sugimoto (2010), who eventually used self-citations with lower weights.

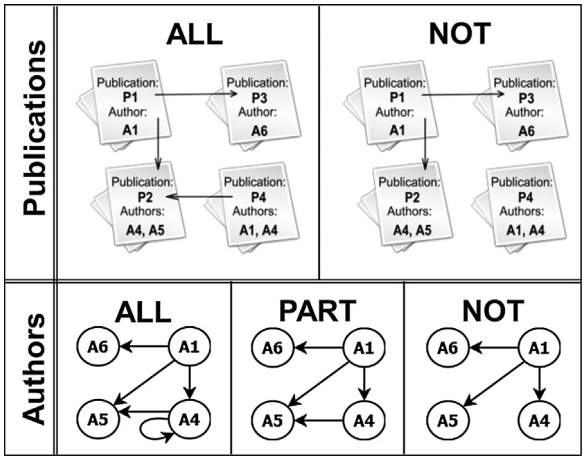

Fig. 1. Types of self-citations variants used.