Journal of Informetrics 10 (2016) 842–853

Contents lists available at ScienceDirect

Journal of Informetrics

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

Which publication is your representative work?

Qikai Niu, Jianlin Zhou, An Zeng*, Ying Fan, Zengru Di

School of Systems Science, Beijing Normal University, Beijing, P.R. China

ARTICLE INFO

Article history:
Received 5 February 2016
Received in revised form 16 June 2016
Accepted 16 June 2016

Keywords:
Representative work
Self-avoiding preferential diffusion
Citation networks

ABSTRACT

As much effort has been made to accelerate the publication of research results, nowadays the number of papers per scientist is much larger than before. In this context, how to identify the representative work for individual researcher is an important yet uneasy problem. Addressing it will help policy makers better evaluate the achievement and potential of researchers. So far, the representative work of a researcher is usually selected as his/her most highly cited paper or the paper published in top journals. Here, we consider the representative work of a scientist as an important paper in his/her area of expertise. Accordingly, we propose a self-avoiding preferential diffusion process to generate personalized ranking of papers for each scientist and identify their representative works. The citation data from American Physical Society (APS) are used to validate our method. We find that the self-avoiding preferential diffusion method can rank the Nobel prize winning paper in each Nobel laureate's personal ranking list higher than the citation count and PageRank methods, indicating the effectiveness of our method. Moreover, the robustness analysis shows that our method can highly rank the representative papers of scientists even if partial citation data are available or spurious behaviors exist. The method is finally applied to revealing the research patterns (i.e. consistency-oriented or diversity-oriented) of different scientists, institutes and countries.

1. Introduction

With the rapid development of the Internet, nowadays most of the research papers are published online, which has significantly accelerated the publication of research findings and resulted in a large number of scientific papers. Ranking scientific publications is thus becoming increasingly important for uncovering the quality of academic works and evaluating the performance of researchers (Dorogovtsev & Mendes, 2015; Petersen, Wang, & Stanley, 2010; Radicchi, Fortunato, & Castellano, 2008; Wang, Song, & Barabási, 2013;; Zhou, Lü, & Li, 2012). Though many ranking algorithms have been proposed and intensively studied, these methods only focus on the global ranking, i.e. ranking the general significance of papers. A long-neglected practical issue for both policy makers and scientists themselves is to identify the representative work of individual scientist (Hirsch, 2005). This is actually a very practical problem as it helps universities and foundation managers to evaluate the achievement and potential of researchers, and help them finally decide who should be hired and to whom the grant should be given (Li & Agha, 2015).

In the literature, there are already many methods proposed for ranking scientific publications. The most straightforward and widely used method is the citation count (Garfield, 1955). However, this method neglects by which the paper is being

* Corresponding author.
E-mail addresses: anzeng@bnu.edu.cn (A. Zeng), yfan@bnu.edu.cn (Y. Fan).

http://dx.doi.org/10.1016/j.joi.2016.06.001

cited, so the ranking obtained by the citation count cannot objectively reflect the true quality of papers. If two papers have similar citation counts, the one cited by important papers is more likely to have higher quality than the one cited by less important papers. Accordingly, Google's PageRank algorithm (Brin & Page, 1998) was introduced to rank scientific papers (Chen, Xie, & Maslov, 2007; Fiala, Rousselot, & Ježek, 2008; Ma, Guan, & Zhao, 2008). Instead of the hyperlink network, PageRank in this case is applied to the citation network. In this algorithm, each paper is assigned with an initial score and the score of papers are updated iteratively. The paper cited by high score papers obtains a high score. The final stable scores are used as the indicator of the significance of papers. PageRank, however, still has several obvious shortcomings such as sensitive to malicious manipulations (Yao, Wei, & Zeng, 2014) and ignorant of time information (Fiala, 2012; Mariani, Medo, & Zhang, 2015). Many variations such as the PageRank with Prior (White & Smyth, 2003), LeaderRank (Lü, Zhang, & Yeung, 2011), NonlinearRank (Yao et al., 2014), CiteRank (Walker, Xie, & Yan, 2007) and SPRank (Zhou, Zeng, Fan, & Di, 2016) have been developed to address these issues. Meanwhile, the idea of PageRank has also been applied to ranking the influence of journals (Bollen, Rodriquez, & Van de Sompel, 2006; González-Pereira, Guerrero-Bote, & Moya-Anegón, 2010) and scientific achievement of researchers (Yan & Ding, 2011).

Despite the wide usage of the citation count and the success of PageRank in revealing the paper quality, these two methods are not appropriate for identifying the representative work of individual researcher (see a very recent comment in reference Ioannidis, Boyack, & Small, 2014). This is due to the fact that the contribution of authors in a paper is different (Stallings, Vance, & Yang, 2013). A high quality paper could not be the representative work of all its authors. For instance, it can happen that a physicist who mainly work in condensed matter collaborates once with an expert in biology by providing some analysis tools and finally publishes a very highly cited paper in a top biology journal. As this paper is only a side interest of the physicist, it should not be considered as his/her representative work. Therefore, the representative work of a researcher might not be simply his/her most highly cited paper nor the paper published in top journals. This requires us to consider a personalized ranking combining the quality and relevance of the papers, instead of the traditional ranking of the papers' general importance. In fact, the problem of the personalized ranking is widely studied in other fields such as the online recommendation (i.e. personalized ranking of online items) (Lü, Medo, & Yeung, 2012) and the traffic navigation (i.e. personalized ranking of routes) (Yeung, Saad, & Wong, 2013). Yet, the problem of personalized ranking in Scientometrics is still a new problem. The most relevant works are the ones in which the contribution of different coauthors in a paper is ranked (Shen & Barabási, 2014).

In this paper, we argue that the representative work of a researcher should be an important paper in his/her area of expertise. To rank papers for each scientist and identify his/her representative work, we propose a self-avoiding preferential diffusion (SPD) process on the citation network which is constructed with the citation data from American Physical Society (APS). Similar approach such as the self-avoiding random walks has been designed for searching complex network (López Millán, Cholvi, Lopez, & Fernandez Anta, 2012). To validate our method, we make use of the Nobel prize winning papers of the Nobel laureates. We find that these papers are ranked rather low in each Nobel laureate's personal ranking list if the citation count and PageRank are used. With SPD, the ranking of the Nobel prize winning papers can be significantly improved, which suggests the high effectiveness of our method. Moreover, we test the robustness of the SPD method by considering the cases where only partial publication data are available and malicious citation behaviors exist. The results show that SPD can still highly rank the true representative papers. In addition, we find that the correlation between SPD and citation ranking can be used to indicate whether the research of a scientist is consistency-oriented or diversity-oriented. The research patterns of different scientists, universities and countries are finally discussed with this correlation.

2. Related works

Uncovering the intrinsic quality of papers is one of the most fundamental issues in Scientometrics. In the literature, a large number of works have been devoted to solve this problem. The ranking algorithms such as PageRank (Brin & Page, 1998) and CiteRank (Walker et al., 2007) are based on iterative processes on citation networks. Zhou et al. (2012) designed an iterative ranking algorithm on the author-paper multiplex network with both citation links and authorship links. Besides these iterative algorithms, Medo, Cimini, and Gualdi (2011) made use of a paper's citation time series to compute the so-called fitness to estimate its intrinsic quality. Wang et al. (2013) derived a mechanistic model for the citation dynamics of individual papers, allowing the citation histories of papers from different journals and disciplines to collapse into a single curve. To achieve a more objective measure of a paper's impact, Radicchi et al. (2008) studied the universality of citation distribution, which makes the papers from different fields comparable.

Research combining authors and papers also attracted much attention. From the paper citation network, Radicchi, Fortunato, Markines, and Vespignani (2009) constructed a citation network between authors and accordingly rank the influence of scientists based on a credit diffusion process. To distinguish an author's contribution in a paper, Shen and Barabási (2014) developed a collective credit allocation model which is applied on a local citation network. De Domenico, Sole-Ribalta, Omodei, Gomez, and Arenas (2015) investigated authors' contributions to a research article with the tensor analysis in interconnected multilayer networks. Ioannidis et al. (2014) investigated with questionnaires whether an author's highest cited paper is considered as his/her best work. Other researchers also seek to uncover statistical patterns in the scientific career trajectory. Petersen, Fortunato, and Pan (2014) quantified the reputation and its impact in academic careers, and revealed that an author's reputation dominates the annual citation rate. Deville et al. (2014) studied the changing of institutions in scientists' career and found that career movements are characterized by a high degree of stratification in institutional ranking.