Journal of Informetrics 10 (2016) 471–484

Contents lists available at ScienceDirect

Journal of Informetrics

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

A data analytic approach to quantifying scientific impact

Xuanyu Cao*, Yan Chen, K.J. Ray Liu

Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received 9 September 2015
Received in revised form 22 February 2016
Accepted 22 February 2016
Available online 8 April 2016

Keywords:
Citation prediction
Data analytic approach

ABSTRACT

Citation is perhaps the mostly used metric to evaluate the scientific impact of papers. Various measures of the scientific impact of researchers and journals rely heavily on the citations of papers. Furthermore, in many practical applications, people may need to know not only the current citations of a paper, but also a prediction of its future citations. However, the complex heterogeneous temporal patterns of the citation dynamics make the predictions of future citations rather difficult. The existing state-of-the-art approaches used parametric methods that require long period of data and have poor performance on some scientific disciplines. In this paper, we present a simple yet effective and robust data analytic method to predict future citations of papers from a variety of disciplines. With rather short-term (e.g., 3 years after the paper is published) citation data, the proposed approach can give accurate estimate of future citations, outperforming state-of-the-art prediction methods significantly. Extensive experiments confirm the robustness of the proposed approach across various journals of different disciplines.

1. Introduction

Citation is frequently used as a performance metric for quantifying the scientific impact of papers. In many applications, we need to know not only the current citations of papers, but also predictions of their future citations. In this section, we first motivate the study of citation prediction problem and briefly review the existing literature on citation prediction and citation distributions. Drawbacks and limitations of existing citation prediction approaches are discussed. Given these drawbacks, the principle of a novel data analytic citation prediction approach is introduced. Based on this principle, the proposed two data analytic prediction methods are epitomized.

1.1. Motivation and related works

Assessing the impact of a paper is a very important issue in academia. Besides the traditional paper awards and other subjective recognitions, an objective measure of the impact of a publication is highly desirable. It has been a trend that citation is perhaps the most often used metric to assess the scientific impact of a paper. A common argument is that the citations of a highly cited paper reflects its influence and contributions to the development of scientific advances. In fact, in addition to individual publications, citations have also been popularly used to assess the scientific impact of researchers and journals. Many popular impact measures, e.g., h-index and impact factor, rely directly on the citations of publications.

* Corresponding author.
E-mail address: apogne@umd.edu (X. Cao).

http://dx.doi.org/10.1016/j.joi.2016.02.006
1751-1577/

Despite its wide usage, the citation can only measure the current and past scientific impact of the papers, while in many scenarios people want to go beyond that to foresee the future scientific impact. Consequently, we need not only the current citations of a paper, but also a prediction of its future citations, which can reflect its future scientific impact. Unfortunately, the complex heterogeneous temporal patterns of citation dynamics make the prediction of future citations rather difficult. To make things even worse, in most of the meaningful practical applications, the task is often to predict the citations of those recently published papers, meaning that we need to make predictions based on a very short-term observation of the citation dynamics. In fact, people have already used such kind of citation prediction in practice. But their predictions are based on human heuristics (Waltman & Costas, 2014), which could be quite subjective and unreliable. All of these motivate us to study a critical yet challenging problem: can one predict the future citations of a paper based on a short-term observation of its citation dynamics?

Hitherto, many works have been devoted to the characterization of the citation distributions and fair comparison between papers from different disciplines (Eom & Fortunato, 2011; Haunschild & Bornmann, 2016; Peterson, Pressé, & Dill, 2010; Radicchi, Fortunato, & Castellano, 2008; Redner, 1998, 2005; Rodrí guez-Navarro, 2011; Schubert & Braun, 1996; Smolinsky, 2016; Stringer, Sales-Pardo, & Amaral, 2008; van Leeuwen & Moed, 2005; Zhang, 2013), yet few have considered the prediction of the future citations of individual papers. Bornmann, Leydesdorff, and Wang (2013, 2014) and Wang (2013) studied the correlation between the citation percentile of early years and that in the future and found a pessimistic result: the correlation is low. Hence, future citation prediction seems to be challenging.

In the literature, researchers have done works related to the citation prediction problem. Acuna, Allesina, and Kording (2012) predicted the future h-index of neuroscientists based on a variety of factors including number of articles written, current h-index, years since publishing the first article, number of distinct journals published in, and number of articles in several top journals. The method combined these factors with a linear regression model to predict future h-index. Furthermore, Ajiferuke and Famoye (2015) systematically studied the relations between the count response variables (e.g., numbers of citations, authors, references, views, downloads) by using several statistical models such as linear regression, lognormal regression, negative binomial regression. Hirsch (2007) compared several indicators of individual scientific achievement (e.g., h-index, total citations, citations per paper) on the task of predicting future scientific achievements. He found that h-index was the best indicator in predicting future achievements of individuals. However, Schreiber (2013) discovered that h-index was an inert indicator since it often severely depended on the growth of the citations of very old publications. This suggested that the real predictive power of h-index was limited. Petersen et al. (2014) measured the impact of authors' reputations on the future success of papers. Based on empirical observations, they argued that when a paper's current citation is low (e.g., at the early stage), the reputations of the authors are important in determining the future citations of the paper. However, if a paper's current citation is higher than a certain threshold, then the reputations of the authors are not important any more in determining the future success of the paper. Moreover, Breitzman and Thomas (2015) proposed to use the size of the inventor team to predict future citation of patents while Havemann and Larsen (2015) compared different bibliometric indicators' predictive powers on future success of young astrophysicists.

All the aforementioned existing literatures are related to the paper citation prediction problem, but none of them tackle it directly. Recently, Stegehuis, Litvak, and Waltman (2015) used the impact factor of the publishing journal and the first year citation count to predict the probability distribution of future citation of papers. However, the usage of only one single year's citation count would inevitably limit the accuracy of the prediction. Bornmann et al. (2014) made use of several other relevant factors (e.g., numbers of authors, pages, references) to predict the long-term citation percentile of papers. Yu, Yu, Li, and Wang (2014) exploited various features of papers (e.g., journal features, author features etc.) to predict the future citations with parametric regression models. But the experiments are confined to papers in the field of information science and library science. Wang, Song, and Barabási (2013) proposed a universal parametric model (hereafter the WSB model) for the temporal citation dynamics and used it to predict the future citations. The WSB model uses three parameters to characterize the citation dynamics as a function of time and explains the underlying mechanism dominating the citation process. The authors claimed that for any paper, by tuning these three parameters, the WSB model can always fit the citation dynamics well. When making predictions, given a period of citation dynamics data of a paper, the authors used it to estimate the three parameters and afterwards employed the trained WSB model to predict future citations.

However, this method has several limitations. First, since the model is parametric, the parameters need to be accurately estimated in order to make accurate predictions. To do so, they usually need a relatively long-term (usually at least 5 years, and the longer the better) observation of the citation dynamics to make meaningful predictions. If only a short-term observation (e.g., 3 years) is provided, their method does not work well, as will be shown in the later experimental results. However, as we previously mentioned, in many scenarios, the observation can be rather short-term. Hence, the usage of WSB model is limited in practice, as pointed out by Van Noorden (2013). Second, only limited experiments based on observations from high impact factor journals (e.g., *Science*, *Nature*) of fundamental sciences (e.g., chemistry, physics and biology) are conducted by Wang et al. (2013). Yet little is known about the performance on other journals such as engineering journals. Actually, according to our experiments, the WSB model performs much worse on papers in *IEEE*, which constitutes a popular journal database for electrical engineering and computer science research. Hence, the WSB model, though claimed to be universal, is not reliable for papers from different disciplines. Third, as pointed out by Wang, Mei, and Hicks (2014) and admitted by Wang, Song, Shen, and Barabási (2014), the WSB model may perform poorly on a few outliers due to severe overfitting, even with some regularization methods (Shen, Wang, Song, & Barabási, 2014). Though the outliers are minority and do not hurt the effectiveness of the WSB model too much, they somehow reduce the reliability of the prediction.