A theoretical model of scientific impact based on citations

Hui Fang
School of Electronic Science and Engineering,
State Key Laboratory of Analytical Chemistry for Life Science,
Nanjing University, Nanjing 210023, CHINA
e-mail: fanghui@nju.edu.cn

ABSTRACT

Number of citations is a basic bibliometric index for evaluating academic articles and assessing researchers, research groups, institutions, and journals. It has several derivative indices which typically assume all citations to have the same importance. Recent research has shown that this is not always true. In this paper, a theoretical relationship between the scientific impact of a research paper and the citations it receives is established. The mathematical model considers the following factors: the number of citations it has received, the role it plays in the papers citing it, and the scientific impact of the papers citing it. This theoretical model depends on the citation matrix and a content-based citation analysis. Limited by current technology, a simplified model referring to normalized citation is provided. The paper further discusses how to apply the method to estimate the scientific impact of researchers, research groups, institutions, and journals. The intended future work for this model is also presented.

Keywords: Citations; Scientific impact; Publication quality indicators; References; Content-based citation analysis

INTRODUCTION

Academic papers reflect researchers' work. Researchers write papers to share their scientific results and opinions with the scientific community (Gilbert 1977). Thus, the number of papers written by a researcher is an index of the quantity of his or her scientific work. Gross and Gross (1927) first introduced the use of citation counts to evaluate scientific journals. Citation counts gradually became an index to represent the scientific impact of a publication and can be regarded as credits given to the author(s). Different citation measures reflect a researcher's impact on his or her peers. Examples include the total number of citations, citations per paper (Lehmann, Jackson and Lautrup 2006), highly cited publications indicator (Waltman and van Eck 2012), and the *h*-index (Hirsch 2005) and its variants (Bornmann et al. 2011).

Citation counts are affected by factors such as time (Garfield 1981), field and journal (Moed et al. 1985), article (Lawani 1986; Laband 1990), author or reader (Mählck and Persson 2000; Cronin 2005), availability of publications (Silverman 1985), and incorrect citing (Broadus 1983; Liang, Zhong nad Rousseau 2014). These factors may cause errors when estimating the impact of scientific activity by citation counts (Garfield 1972).

Recent research has attempted to explain the relationship between citation and scientific impact. Assuming that citations from experts are more valuable, Ding and Cronin (2011) investigated the popularity and prestige of researchers based on whether the citations they received were from highly cited papers. Ding (2011) applied the PageRank algorithm to author citation networks. Highly cited publications do not always have a large scientific impact, owing to some random citation factors (Waltman, van Eck and Wouters 2013). Some researchers have proposed citation modifications such as normalized citation (Waltman and van Eck 2013) and indirect citation (Fragkiadaki and Evangelidis 2014).

To accurately assess the scientific impact of a paper, a theoretical model of scientific impact using citations is established. Indirection citation models such as PageRank can be explained using this theoretical model. Limited by the technology available today, a simplified version of the model referring to the underlying idea of normalized citation is proposed. The author provides guidance on the selection of parameter values for the model. At the end of this paper, the author discusses the model and proposes future work.

SCIENTIFIC IMPACT OF RESEARCH

Direct and Indirect Scientific Impact

Different references play different roles in a research article, and they have different levels of contribution to the citing paper (Ding et al. 2013). References may be the basis of the work, enlighten the author(s), or provide supporting evidence. Suppose the scientific impact of paper P_c is $Scim(P_c)$, which can be determined according to its citation record (as described in the following sections). Both the work of P_c's author(s) and P_c's references contribute to the achievement of P_c. Therefore, we can allocate the credit of P_c, $Scim(P_c)$, to P_c's author(s) and its references, according to their contribution to P_c.The scientific impact allocated to all its references is

$$Scim_R(P_c) = \beta(P_c)Scim(P_c), \quad (1)$$

Where $\beta(P_c)$ is the proportion of contribution by P_c's references to P_c.

$Scim_R(P_c)$ can be further allocated to every reference of P_c according to their contribution. Suppose P_{Ri} is the i-th reference ($i = 1, 2, ..., n_R$) of P_c, n_R is the number of references of P_c. Defined $Scim_R(P_{Ri}, P_c)$ as the scientific impact allocated to P_{Ri} due to P_c citing it:

$$Scim_R(P_{Ri}, P_c) = w(P_{Ri}, P_c)Scim_R(P_c), \quad (2)$$

where $w(P_{Ri}, P_c)$ is the weight of the scientific impact carried by P_{Ri} among all references of P_c. Clearly, $\sum_{i=1}^{n_R} w(P_{Ri}, P_c) = 1$. Then we have

$$Scim_R(P_{Ri}, P_c) = \beta(P_c)w(P_{Ri}, P_c)Scim(P_c), \quad (3)$$

Eq. (3) implies that a reference can influence other papers through the citing paper. We call this an indirect scientific impact. If P_c is cited by some papers, it attains scientific impact $Scim(P_c)$. Eq. (3) accounts for the indirect scientific impact of P_{Ri} to the papers citing P_c. In general, the indirect scientific impact that paper P obtained due to P_c citing it is

$$Scim_{cid}(P,P_c) = \beta(P_c)w(P,P_c)Scim(P_c)\,. \tag{4}$$

On the other hand, P_c is directly impacted by all its references. Traditional methods give each of P_c's references one citation. Recently, normalized citation (Moed 2010; Leydesdorff and Bornmann 2011; Glänzel et al. 2011) has been proposed to diminish differences caused by varying citation behavior among different fields (Bornmann and Daniel 2008). Leydesdorff and Opthof (2010) gave all the references of a paper one citation in total, with each reference obtaining an equal fractional citation.

Adopting this idea of normalized citation, considering the contributions of all references and the different importance among the references to a paper, the direct scientific impact P obtained due to P_c citing it is

$$Scim_{cd}(P,P_c) = \beta(P_c)\times w(P,P_c)\times 1\,. \tag{5}$$

The '1' in Eq. (5) reflects normalized citation.

Scientific Impact of a Paper

The scientific impact P obtained due to P_c citing it is the summation of the direct and indirect scientific impacts.

$$Scim_c(P,P_c) = Scim_{cd}(P,P_c) + \theta Scim_{cid}(P,P_c)\,, \tag{6}$$

where θ represents the relative importance of the indirect scientific impact compared with the direct scientific impact.

The entire indirect scientific impact of paper P is the summation of the indirect scientific impact caused by all the citations it received:

$$Scim_{id}(P) = \sum_{i=1}^{n_c} Scim_{cid}(P,P_{ci})\,, \tag{4'}$$

where n_c is the number of citations paper P received, and P_{ci} is the i-th paper citing P.

Similarly, the entire direct scientific impact of paper P is:

$$Scim_d(P) = \sum_{i=1}^{n_c} Scim_{cd}(P,P_{ci})\,. \tag{5'}$$

The scientific impact of paper P is the summation of its scientific impact caused by all the citations it received:

$$Scim(P) = \sum_{i=1}^{n_c} Scim_c(P,P_{ci}) = Scim_d(P) + \theta Scim_{id}(P)\,. \tag{7}$$

Incorporating Eq. (6) into Eq. (7), we obtain

$$Scim(P)=\sum_{i=1}^{n_c}\beta(P_{ci})w(P,P_{ci})[1+\theta Scim(P_{ci})]\,. \tag{8}$$

That is to say, the scientific impact of a paper is a linear combination of the scientific impact of all the papers citing it. The coefficient of the scientific impact of P_{ci} ($i=1,2,.....,n_c$) is the contribution proportion of P to P_{ci}. If none of the citing papers have themselves been cited, then their scientific impact is zero. In this case, P only has direct scientific impact.

Scheme to Determine Scientific Impact of a Paper

The scientific impact of paper P, expressed as in Eq. (8), consists of direct and indirect scientific impacts. According to Eq. (5'), the direct scientific impact can be estimated using information from each paper citing P, P_{ci} ($i=1,2,.....,n_c$). According to Eq. (4'), the indirect scientific impact includes information from each paper citing each P_{ci}, which requires further recursion.

The author uses a hypothetical citation relationship, as shown in Figure 1, to describe the recursion process that determines a paper's indirect scientific impact. As defined in Rousseau (1987), Hu, Rousseau and Chen (2011) and Fragkiadaki and Evangelidis (2014), the papers directly citing P are called the first citation generation papers. The first citation generation papers are directly cited by the second citation generation papers. In general, the k+1-th citation generation papers directly cite the k-th citation generation papers.

In Figure 1, P is cited by P_i ($i=1,\ 2,\ .....,\ n_c$). P_2 has not been cited. So its scientific impact is 0. Thus, the indirect scientific impact of P caused by P_2 is 0, or $Scim_{cid}(P,P_2)=0$. As for the citation of P by P_1, $Scim_{cid}(P,P_1)=\beta(P_1)w(P,P_1)Scim(P_1)$. P_1 is cited by $P_{1,1}$ and $P_{1,2}$, so $Scim(P_1)=Scim_{cd}(P_1,P_{1,1})+Scim_{cd}(P_1,P_{1,2})+\theta Scim_{cid}(P_1,P_{1,1})+\theta Scim_{cid}(P_1,P_{1,2})$. $Scim(P_{1,1})=0$ as $P_{1,1}$ has not been cited. Thus, the indirect scientific impact of P_1 only depends on $Scim(P_{1,2})$ and the role of P_1 in $P_{1,2}$. The direct scientific impact of $P_{1,2}$ is determined by the roles of $P_{1,2}$ in $P_{1,2,1}$, $P_{1,2,2}$, and $P_{1,2,3}$. The indirect scientific impact of $P_{1,2}$ only depends on $Scim(P_{1,2,2})$ and the role of $P_{1,2}$ in $P_{1,2,2}$, because $P_{1,2,1}$ and $P_{1,2,3}$ have not been cited. $P_{1,2,2}$ only impacts $P_{1,2,2,1}$, and $P_{1,2,2,1}$ has not been cited. So $P_{1,2,2}$ has no indirect scientific impact, and its scientific impact is its direct scientific impact.

The determination of the indirect scientific impact of P caused by papers P_3 to P_{n_c} is similar to that caused by P_1, therefore it is not discussed here.

The author defines a paper that has not been cited as an end in the citation network, for example P_2 and $P_{1,2,1}$ in Figure 1. As the scientific impact of an end is 0, its references have only direct scientific impact due to the citation by this paper. Exactly determining the indirect scientific impact of a paper requires us to track every path along the citation network to all of the ends indirectly linking to it. In short, determining the scientific impact of a paper requires information from its first citation generation papers, including their scientific impact. Determining the scientific impact of the first citation generation papers requires information

from the second citation generation papers. The procedures repeat until they meet ends. The only exception is a mutual citation, which is discussed in the next section.

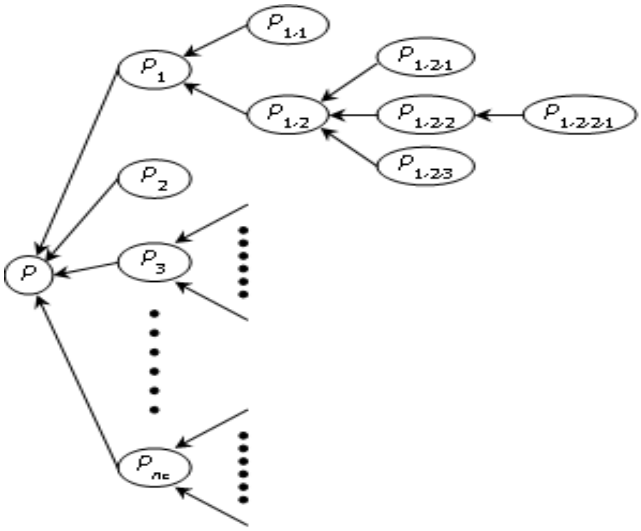

Figure 1: A hypothetical Citation Relationship (Arrow lines represent citations. For example paper *P* is cited by P_i ($i = 1, 2, ..., n_c$). The papers connecting P_3 to P_{n_c} either directly or indirectly are omitted).

Mutual Citation

Mutual citation occurs when two papers contain related research work and had authors who were in close communication, or when they were written by the same researcher(s). Supposed papers P_1 and P_2 mutually cite. According to Eq. (8), P_1 influences P_2 through P_2 citing P_1, and vice versa. But citation is used to reflect one paper's impact on others. The author does not consider the influence of one paper on itself, directly or indirectly. Thus, the indirect scientific impact through mutual citation is omitted.

$$\begin{aligned} Scim(P_1) &= Scim_{\mathrm{O}}(P_1) + \beta(P_2)\, w(P_1, P_2)\left[1 + \theta Scim_{\mathrm{O}}(P_2)\right] \\ Scim(P_2) &= Scim_{\mathrm{O}}(P_2) + \beta(P_1)\, w(P_2, P_1)\left[1 + \theta Scim_{\mathrm{O}}(P_1)\right] \end{aligned} \tag{9}$$

where $Scim_{\mathrm{O}}(P_1)$ is the scientific impact of P_1 due to citations other than that from P_2, defined $Scim_{\mathrm{O}}(P_1) = \sum_{i=1, P_{ci} \neq P_2}^{n_{c1}} \beta(P_{ci}) w(P_1, P_{ci})[1 + \theta Scim(P_{ci})]$, $Scim_{\mathrm{O}}(P_2)$ is the scientific impact of P_2 owing to citations other than that from P_1, defined $Scim_{\mathrm{O}}(P_2) = \sum_{i=1, P_{ci} \neq P_1}^{n_{c2}} \beta(P_{ci}) w(P_2, P_{ci})[1 + \theta Scim(P_{ci})]$, and n_{c1} and n_{c2} are the number of papers citing P_1 and P_2, respectively.

In the aforementioned recursion for determining the scientific impact of one paper, if it is found that P_2 cites P_1 and P_1 cites P_2, as shown in Figure 2, then it can be concluded that they mutually

cite. Then, there is no need to calculate $Scim_O(P_1)$ for the second P_1, because it has already been determined for the first P_1. The recursion path of P_2 citing P_1, and P_1 citing P_2 then terminates.

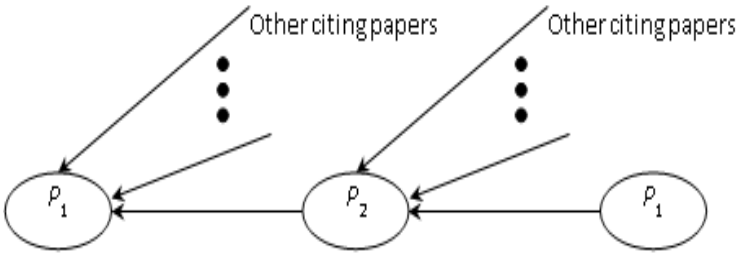

Figure 2: Mutual Citation

RELATED WORK

In this section, the indicators to reflect the influence of papers are discussed. Some similar indicators which do not reflect that of papers are not discussed here, such as eigenfactor which reflects the influence of journals.

Normalized Citation

Different research fields have different citation cultures, which leads to differences in the average length of references per paper. In fields with more references per paper, papers have a higher chance of being cited than those in fields with fewer references. Normalized citations have been proposed to impartially compare research work among different fields. When considering a certain paper, normalized citation gives every reference in a paper a fractional citation equal to the reciprocal of the number of references in the citing paper (Leydesdorff and Opthof 2010). The normalized citation for paper *P* is

$$c_N(P) = \sum_{i=1}^{n_c} \frac{1}{N_R(P_{ci})}, \tag{10}$$

where $N_R(P_{ci})$ is the number of references in the *i*-th paper citing *P*, P_{ci}. This fractional citation equally allocates the credit (the entire influence on a citing paper) to all the references of the citing paper. This is a practical solution for allocating credit among references, because current technology cannot effectively differentiate between the contributions of different references to the citing paper.

PageRank

PageRank estimates the importance score of a web page according to the number of high PageRank pages linking to it. The recursive formula of PageRank is

$$PR_A = (1-d) + d \times \sum_i \frac{PR_i}{N_i}, \tag{11}$$

where PR_A is the score of page *A*, *d* is the damping factor, PR_i is the score of the *i*-th page citing page *A*, and N_i is the number of pages cited by PR_i.

In bibliometrics, papers for pages in Eq. (11) to evaluate a paper's impact can be substituted (Sidiropoulos and Manolopoulos 2005; Ma, Guan and Zhao 2008), as an empirical model,

$$PR(P) = (1-d) + d \times \sum_{i=1}^{n_c} \frac{PR(P_{ci})}{N_R(P_{ci})}, \tag{11'}$$

Eq. (11') appears to only represent indirect citations. However, it implicitly reflects direct citations. Expanding $PR(P_{ci})$ in Eq. (11') results in,

$$PR(P) = (1-d) + d(1-d)\sum_{i=1}^{n_c} \frac{1}{N_R(P_{ci})}\left[1 + \frac{d}{1-d}\sum_{j=1}^{n_{ci}} \frac{PR(P_{ci,j})}{N_R(P_{ci,j})}\right] \tag{12}$$

where n_{ci} is the number of papers citing P_{ci}, and $P_{ci,j}$ is the *j*-th paper citing P_{ci}. The term '1' in the summation function corresponds to the direct scientific impact in Eq. (8). It is generated by the constant item $(1-d)$ in Eq. (11'), which gives any paper a basic score $(1-d)$ even if it has not been cited.

SCEAS Rank

When calculating citation or scientific impact, a non-cited paper should have a score of zero. The Scientific Collection Evaluator with Advanced Scoring (SCEAS) Rank (Sidiropoulos and Manolopoulos 2005) is a modified PageRank that explicitly expresses direct and indirect citations:

$$S(P) = \sum_{i=1}^{n_c} \frac{S(P_{ci}) + b}{N_R(P_{ci})} a^{-1} \quad (a \geq 1,\ b > 0), \tag{13}$$

where *b* is the direct citation enforcement factor, and *a* denotes the speed that an indirect citation enforcement converges to zero. Rewriting Eq. (13) results in,

$$S(P) = a^{-1} b \sum_{i=1}^{n_c} \frac{1}{N_R(P_{ci})}\left[1 + \frac{1}{b} S(P_{ci})\right] \quad (a \geq 1,\ b > 0). \tag{13'}$$

The direct citation enforcement factor *b* corresponds to the reciprocal of θ in Eq. (8). The damping coefficient $a^{-1}b$ represents that references only contribute to part of a paper.

A SIMPLIFIED MODEL OF SCIENTIFIC IMPACT

In Eq. (8), there are three parameters of each citing paper P_{ci} $(i = 1, 2, \ldots\ldots, n_c)$ that determine *P*'s scientific impact. They are $\beta(P_{ci})$, $w(P, P_{ci})$, and θ.

The first two parameters vary among different papers. In short, they depend on the importance of P to P_{ci}. To exactly determine them, a content-based citation analysis at the semantic level is required (Small 2011). This could potentially calculate the different importance of each reference. However, this is not currently possible. Further study is required to improve semantic analysis techniques and quantification methods for the importance level of different kinds of citations. Under present conditions, this can be determined using some approximate assumptions.

One simple way to estimate $w(P,P_c)$ is to ignore the differences in contributions among the references to the citing paper, as done in normalized. Then,

$$w(P,P_c)=1/N_R(P_c)\,. \tag{14}$$

Now, Eq. (8) can be rewritten as

$$Scim(P)=\sum_{i=1}^{n_c}\beta(P_{ci})\frac{1}{N_R(P_{ci})}[1+\theta Scim(P_{ci})]\,. \tag{8'}$$

Expanding $PR(P_{ci,j})$ in Eq. (12), results in,

$$PR(P)=(1-d)+d(1-d)\sum_{i=1}^{n_c}\frac{1}{N_R(P_{ci})}\left\{1+d\sum_{j=1}^{n_{ci}}\frac{1}{N_R(P_{ci,j})}\left[1+\frac{d}{1-d}\sum_{k=1}^{n_{ci,j}}\frac{PR(P_{ci,j,k})}{N_R(P_{ci,j,k})}\right]\right\} \tag{12'}$$

Where $n_{ci,j}$ is the number of papers citing $P_{ci,j}$, and $P_{ci,j,k}$ is the k-th paper citing $P_{ci,j}$. Ignoring the first term $(1-d)$ of Eq. (12'), the coefficient before the first summation sign ($d(1-d)$) corresponds to the damping coefficient in Eq. (13') ($a^{-1}b$), which corresponds to the fact that all references only contribute to part of a paper. To simplify the problem, it is assumed that the proportion of contribution to a paper by all its references is the same for different papers for now. Then, $\beta(P_{ci})=\beta$, and Eq. (8') can be rewritten as

$$Scim(P)=\beta\sum_{i=1}^{n_c}\frac{1}{N_R(P_{ci})}\left\{1+\theta\beta\sum_{j=1}^{n_{ci}}\frac{1}{N_R(P_{ci,j})}\left[1+\theta Scim(P_{ci,j})\right]\right\}\quad . \tag{8''}$$

By comparing the coefficients before the first and second summation sign in Eq. (8'') with those in Eq. (12'), the following equation is obtained.

$$\begin{aligned}\beta&=d(1-d)\\ \theta\beta&=d\end{aligned}\quad . \tag{15}$$

Therefore,

$$\theta=(1-d)^{-1}\,. \tag{16}$$

In web evaluation, d is usually set to be 0.85. Ma et al. (2008) used an empirical study to find that there is no significant difference in the results of estimating papers when using $d=0.5$ or $d=0.85$. Thus θ can take the values between 2 and 6.67.

In PageRank, only one empirical parameter (*d*) is fixed for all papers. In the simplified equation of *Scim* in Eq. (8'), θ is fixed for all papers. The earlier estimation of θ is found by fixing $\beta(P_{ci})$ to β in Eq. (8'').

However, $\beta(P_{ci})$ varies according to each paper. Eq. (8') uses the normalized citation assumption, which supposes that the contribution of each reference to the citing paper is the same. Therefore, $w(P,P_{ci})$ is simply a function of $N_R(P_{ci})$. Similar treatment can be applied to $\beta(P_{ci})$. In general, the impact that a paper's references have on it increases with the number of references. First, consider a paper with only one reference, i.e., $N_R(P_{ci})=1$. Suppose the average contribution of a reference to the citing paper is β_1 $(0<\beta_1<1)$. Then the contribution of the author(s)' average effort to the citing paper is $1-\beta_1$. Suppose that the author(s)' average effort in a paper is not related to the number of references. Thus, for a citing paper with $N_R(P_{ci})$ references,

$$\beta(P_{ci})=\frac{N_R(P_{ci})\beta_1}{1-\beta_1+N_R(P_{ci})\beta_1}=\frac{N_R(P_{ci})\beta_1}{1+\left[N_R(P_{ci})-1\right]\beta_1}. \tag{17}$$

Obviously, $\beta(P_{ci})\le 1$, which tends to be 1 when $N_R(P_{ci})=\infty$, although the number of references is limited by the space of a journal.

In practice, when the number of references in a paper increases, the contributions of some references are similar or overlap; for example, when references are cited in one sentence. So Eq. (17) is modified to:

$$\beta(P_{ci})=\frac{\left[N_R(P_{ci})\right]^{\lambda}\beta_1}{1-\beta_1+\left[N_R(P_{ci})\right]^{\lambda}\beta_1}, \tag{17'}$$

where $\lambda \le 1$, which limits the increase of the proportion of contribution of the references by their amount. Obviously, $\beta(P_{ci})/N_R(P_{ci})\le\beta_1$.

Figure 3 shows $\beta(P_{ci})$ vs. $N_R(P_{ci})$ when $\beta_1=0.05$. The curve has a larger slope when $N_R(P_{ci})$ is small, and flattens when $N_R(P_{ci})$ is large. When λ is larger (such as 1), $\beta(P_{ci})$ approaches 1 when $N_R(P_{ci})$ is approximately 300. For a smaller λ (such as 2/3), $\beta(P_{ci})$ reaches approximately 0.8 when $N_R(P_{ci})$ is 1000. Obviously, a larger λ is suitable for fields with a smaller average number of references, and a smaller value is suitable for fields with more references per paper. $\beta(P_{ci})$ with λ = 2/3 is about 80% of its value when λ = 0.8, for $N_R(P_{ci})<1000$. So the impact of λ on *Scim* is much less than that of $N_R(P_{ci})$, when $N_R(P_{ci})$ is large.

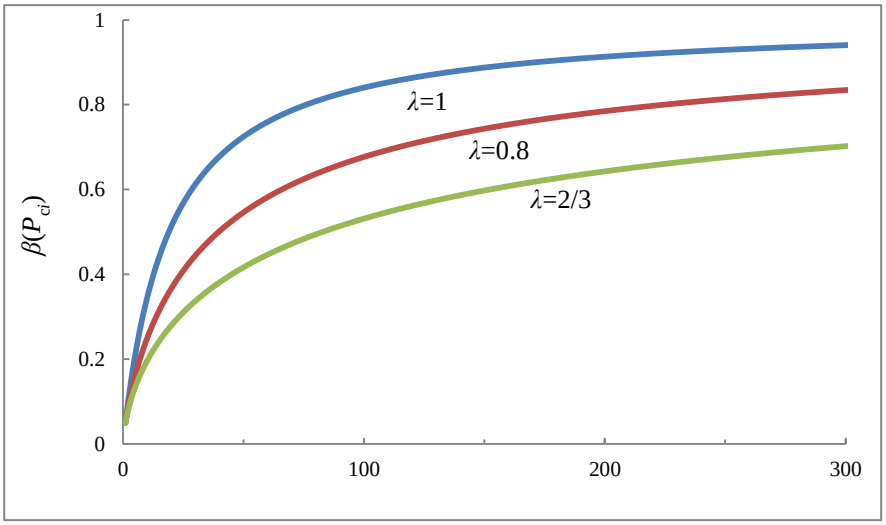

Figure 3: $\beta(P_{ci})$ vs. $N_R(P_{ci})$ with β_1 = 0.05, as defined in Eq. (17')

Any change in the *Scim* of a paper will change the *Scim* of the papers it cites directly or indirectly, along the path of citation from higher citation generation to lower. If a new publication cites $P_{1,2,1}$ in Figure 1, then *Scim*($P_{1,2,1}$) will be larger than 0, and it further increases *Scim*($P_{1,2}$), *Scim*(P_1), and *Scim*(P). Sidiropoulos and Manolopoulos (2005) proposed that a change in the score of a node should decay along the citation path, and converge to zero if the citation generations between the two papers is large enough. For example, they suggest that the score of a paper will be affected little by a new citation to one of its 7th citation generation papers, as illustrated in Figure 4. Because $\beta(P_{ci}) / N_R(P_{ci}) \le \beta_1$, the change in the *Scim* of a paper due to its citing paper is maximized if the citing paper only has one reference. Consider the case in Figure 4(a) where every paper directly or indirectly citing P has only one reference. A change in the *Scim*(P_i) ($i \ge 1$) affects *Scim*(P) more than if any node from P_1 to P_i had more than 1 reference.

In Figure 4(a), $Scim(P) = \beta_1 + \beta_1^2\theta + \beta_1^3\theta^2 + \ldots + \beta_1^7\theta^6$. If a new paper P_{7_1} cites P_6, as shown in Figure 4(b), then $Scim(P) = \beta_1 + \beta_1^2\theta + \beta_1^3\theta^2 + \ldots + 2\beta_1^7\theta^6$. The difference in the *Scim*(P) of these two cases, $\Delta Scim(P)$, is $\beta_1^7\theta^6$. If the $\Delta Scim(P)$ caused by P_6 having one more citation must be less than ε, then $\beta_1 < (\varepsilon / \theta^6)^{1/7}$. For example, if $\theta = 6.67$ and $\beta_1 = 0.05$, then $\Delta Scim(P) = 6.86 \times 10^{-5} < 0.01\%$.

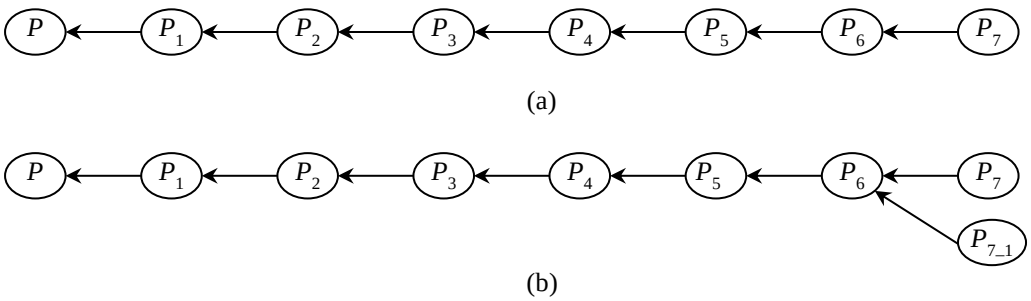

Figure 4: A Hypothetical Citation Relationship

DISCUSSION AND FUTURE WORK

According to Eq. (8), the scientific impact of a paper depends on the number of papers citing it, the scientific impact of the citing papers, and the role it plays in the citing papers. In general, the more citations a paper receives, the higher *Scim* the papers citing it have, and the more important it is in its citing papers, the higher *Scim* it has. Thus, scientific impact takes into consideration the quality of the citing papers. PageRank and SCEAS Rank are two simplified empirical models of scientific impact that make two assumptions: all references have equal importance to the citing paper, and every paper has a fixed proportion of contribution from all its references, except papers with no reference. The relative importance of indirect scientific impact to direct scientific impact in *Scim* is estimated according to PageRank. Similar to the idea of normalized citation, which is also used in PageRank and SCEAS Rank, the author has modified the damping coefficient to incorporate the number of references in the citing paper, as shown in Eq. (17').

In this paper, the scientific impact of individual papers is studied. The method presented can also be applied to evaluate the scientific impact of a researcher, a research group, an institute, an area, or a journal. The scientific impact of a researcher can be the summation of the scientific impact of his or her research articles. This method considers both the quantity and quality of academic output. The same method can be used for a group, an institute, or an area. When considering the scientific impact of a journal, the average *Scim* of its papers can be used. Different journals have different amounts of papers, so the average quality of papers ensures the impartiality of evaluation.

There are different kinds of citations. The simplified model of *Scim* expressed by Eqs. (8') and (17') is an approximate method that ignores the differences in the importance of references to the citing papers, and is practical in terms of current technology. An accurate calculation of *Scim* using Eq. (8) requires a content-based citation analysis at the semantic level, which would calculate the differing importance of references. This requires improvements to semantic analysis techniques, and to the quantification of the importance level of different kinds of citations. The quantization of the importance level of different kinds of citations may cause controversy. For example, criticism or negation citations often emerge in scientific disputes. It is not possible to ascertain which side of the dispute is correct before direct evidence is found. Simply reducing the importance of criticism or negation citations will underestimate the *Scim* of valid papers when papers with opposite opinions negate them. In addition, even discredited papers may inspire future valid papers. Until a comprehensive and detailed investigation on the importance of different kinds of citations has been completed, it is suggested that every reference is assigned the same weight.

To calculate the *Scim* of a paper, one needs to search all the papers citing it directly and indirectly and calculate the *Scim* of all the citing papers. The key operation is the search of all papers along the trail of citations in a database. However, the aforementioned recursion process is tedious and time consuming if it is done manually. All the records of related subject categories can be downloaded and the score of all the papers in the citation network can be calculated. However, the indirectly citing papers may span many research fields. Then, the citation matrix may be very large for some popular disciplines, and disciplines that span many fields. Previous work investigating indirect citations has been confined to a certain field or small database (Sidiropoulos and Manolopoulos 2005; Ma et al. 2008; Ding 2011). Further comprehensive investigations on the scientific impact of a paper would benefit from a system that can automatically calculate the *Scim* of a given paper. The functions required by this system include an automatic search of papers in the database that directly or indirectly cite the given paper. Future study will focus on these problems.

ACKNOWLEDGMENT

This research project is funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

REFERENCES

Bornmann, L., and Daniel, H.D. 2008. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation,* Vol. 64, no. 1: 45–80.

Bornmann, L., Mutz, R., Hug, S.E., and Daniel, H.D. 2011. A multilevel meta-analysis of studies reporting correlations between the *h* index and 37 different *h* index variants. *Journal of Informetrics,* Vol. 5, no.3: 346–359.

Broadus, R.N. 1983. An investigation of the validity of bibliographic citations. *Journal of the American Society for Information Science,* Vol. 34, no. 2: 132−135.

Cronin, B. 2005. *The hand of science. Academic writing and its rewards*. Lanham, MD., USA: Scarecrow Press.

Ding, Y. 2011. Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology,* Vol. 62, no. 2: 236–245.

Ding, Y., and Cronin, B. 2011. Popular and/or prestigious? Measures of scholarly esteem. *Information Processing and Management,* Vol. 47, no. 1: 80–96.

Ding, Y., Liu, L., Guo, C., and Cronin B. 2013. The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics,* Vol. 7, no. 3: 583–592.

Fragkiadaki, E., and Evangelidis G. 2014. Review of the indirect citations paradigm: theory and practice of the assessment of papers, authors and journals. *Scientometrics,* Vol. 99, no. 2: 261–288.

Garfield, E. 1972. Citation analysis as a tool in journal evaluation - journals can be ranked by frequency and impact of citations for science policy studies. *Science,* Vol. 178, no. 4060: 471−479.

Garfield, E. 1981. Citation classics - four years of the human side of science. *Essays of an Information Scientist,* Vol. 5, no.22: 123−134.

Gilbert, G.N. 1977. Referencing as persuasion. *Social Studies of Science,* Vol. 7, no. 1: 113−122.

Glänzel, W., Schubert, A., Thijs, B., and Debackere, K. 2011. A priori vs. a posteriori normalisation of citation indicators. The case of journal ranking. *Scientometrics,* Vol. 87, no. 2: 415–424.

Gross, P.L.K., and Gross E.M. 1927. College libraries and chemical education. *Science,* Vol. 66, no. 1713: 385−389.

Hirsch, J.E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences,* Vol. 102, no. 46: 16569–16572.

Hu, X., Rousseau, R., and Chen, J. 2011. On the definition of forward and backward citation generations. *Journal of Informetrics*, Vol. 5, no.1: 27–36.

Laband, D.N. 1990. Is there value-added from the review process in economics? Preliminary evidence from authors. *Quarterly Journal of Economics,* Vol. 105, no. 2: 341−352.

Lawani, S.M. 1986. Some bibliometric correlates of quality in scientific research. *Scientometrics,* Vol. 9, no. 1−2: 13−25.

Lehmann, S., Jackson, A.D., and Lautrup, B.E. 2006. Measures for measures. *Nature,* Vol. 444, no. 7122: 1003–1004.

Leydesdorff, L., and Bornmann, L. 2011. How fractional counting of citations affects the impact factor: normalization in terms of differences in citation potentials among fields of science. *Journal of the American Society for Information Science and Technology,* Vol. 62, no.2: 217–229.

Leydesdorff, L., and Opthof, T. 2010. Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *Journal of the American Society for Information Science and Technology,* Vol. 61, no: 11: 2365–2369.

Liang, L., Zhong, Z., and Rousseau R. 2014. Scientists' referencing (mis)behavior revealed by the dissemination network of referencing errors. *Scientometrics,* Vol. 101, no. 3: 1986-1993.

Ma, N., Guan, J., and Zhao, Y. 2008. Bringing PageRank to the citation analysis. *Information Processing & Management,* Vol. 44, no. 2: 800–810.

Mählck, P., and Persson, O. 2000. Socio-bibliometric mapping of intra-departmental networks. *Scientometrics,* Vol. 49, no. 1: 81−91.

Moed, H.F. 2010. Measuring contextual citation impact of scientific journals. *Journal of Informetrics,* Vol. 4, no. 3: 265–277.

Moed, H.F., Burger, W.J.M., Frankfort, J.G., and van Raan, A.F.J. 1985. The use of bibliometric data for the measurement of university research performance. *Research Policy,* Vol. 14, no. 3: 131−149.

Rousseau, R. 1987. The Gozinto theorem: Using citations to determine influences on a scientific publication. *Scientometrics*, Vol.11, no. 3-4: 217–229.

Sidiropoulos, A., and Manolopoulos, Y. 2005. A citation-based system to assist prize awarding. *SIGMOD Record,* Vol. 34, no. 4: 54–60.

Silverman, R.J. 1985. Higher education as a maturing field? Evidence from referencing practices. *Research in Higher Education,* Vol. 23, no. 2: 150−183.

Small, H. 2011. Interpreting maps of science using citation context sentiments: A preliminary investigation. *Scientometrics,* Vol. 87, no. 2: 373–388.

Waltman, L., and van Eck, N.J. 2012. The inconsistency of the *h*-index. *Journal of the American Society for Information Science and Technology,* Vol. 63, no. 2: 406–415.

Waltman, L., van Eck, N.J., and Wouters P. 2013. Counting publications and citations: Is more always better? *Journal of Informetrics*, Vol. 7, no.3: 635–641.

Waltman, L., and van Eck, N.J. 2013. Source normalized indicators of citation impact an overview of different approaches and an empirical comparison. *Scientometrics,* Vol. 96, no.3: 699–671.