Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 44 (2008) 800–810
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Bringing PageRank to the citation analysis

Nan Ma [a], Jiancheng Guan [b,*], Yi Zhao [c]

[a] School of Management, Beijing University of Aeronautics and Astronautics, Beijing 100083, PR China
[b] School of Management, Fudan University, 670 Guoshun Road, Shanghai 200433, PR China
[c] School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, PR China

Received 3 April 2007; received in revised form 16 June 2007; accepted 19 June 2007
Available online 20 August 2007

Abstract

The paper attempts to provide an alternative method for measuring the importance of scientific papers based on the Google's PageRank. The method is a meaningful extension of the common integer counting of citations and is then experimented for bringing PageRank to the citation analysis in a large citation network. It offers a more integrated picture of the publications' influence in a specific field. We firstly calculate the PageRanks of scientific papers. The distributional characteristics and comparison with the traditionally used number of citations are then analyzed in detail. Furthermore, the PageRank is implemented in the evaluation of research influence for several countries in the field of Biochemistry and Molecular Biology during the time period of 2000–2005. Finally, some advantages of bringing PageRank to the citation analysis are concluded.

Keywords: Citation analysis; Citation network; PageRank; Internal Citations

1. Introduction

Basically, all research performance evaluation systems deal with the concept of quality. Among the various methods that have been proposed and used to complement the traditional research performance evaluation measures, citation analysis was and still is one of the most widely used (Chubin & Hackett, 1990; Martin, 1996; Thomas & Watkins, 1998).

In most previous researches of citation analysis (Garfield, 1979, 1983; Glänzel, 1996; Koenig, 1982, 1983; Kostoff, 1996; Lawani & Bayer, 1983; Narin, 1976; Narin & Hamilton, 1996), the metric used to quantify the importance of scientific publications has been largely based on integer counting of their citations. Although, the number of citations gives a direct approximation of a paper's importance, we may still experience some situations when citations do not seem to provide a full picture of the academic prestige of a paper. In this regard, several studies (Moed, 2002; van Raan, 1996, 1997) have yielded to construct alternative metrics

* Corresponding author.
E-mail addresses: guanjianch@buaa.edu.cn, guanjianch@126.com, guanjianch@sina.com (J. Guan).

0306-4573/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2007.06.006

for measuring research influence in an advanced way, such as the indicators of *CPP*, *JCSm*, *FCSm*, *CPP/JCSm* and *CPP/FCSm*,[1] etc.

While most of the citation analyses have been based on how many times the papers were cited, we are thus motivated to study from another perspective which concerns more about who actually cited the papers and the prestige they have transferred to the cited papers. And the Google's PageRank algorithm (Page, Brin, Motwani, & Winograd, 1998) is such an existing metric used for the evaluation of Web pages. The PageRank algorithm is an important component of the Google's search engine, and boosted several studies on the link analysis of the Web (Kleinberg, 1999; Snyder & Rosenbaum, 1999). Actually, the principle of PageRank algorithm can be mostly approximated to the indicator of "*journal influence*" suggested by Pinski and Narin (1976), which evaluates the influence of journals by taking into account not simply the number of citations from one journal to the other, but also the influence of the citing journal. As suggested by Bollen, Rodriguez, and van de Sompel (2006), this procedure is on the other hand related to the "inherited status" defined in the social network science (Bonacich, 1987). In the most recent study, Bollen et al. (2006) bring this idea into an evaluative study on scientific journals by using the weighted PageRank.

By way of context, we present here a brief reason for choosing the field of "Biochemistry and Molecular Biology" in this bibliometric study. Molecular biology has been emerged as a fast growing discipline since the elucidation of DNA molecule by Watson and Crick in 1953 (Dalpé, 2002). Considering the potentially important influence on disease detection, treatment for higher plants and animals, as well as future economic development, many countries have listed the field of "Molecular Biology" as high priority in their national S&T plans. As molecular biology becomes more complex and sophisticated, greater attention has been paid to the publication practices within the discipline (Anderson, 1992; Anon., 1992; He, Zhang, & Teng, 2005; Herbertz & Müller-Hill, 1995; Pendlebury, 1990).

This study demonstrates the application of PageRank algorithm in the citation network and empirically analyzes the scientific productivity and academic influence of individual countries based on publications and PageRank values in the context of "Molecular Biology" subject field. Typically, the evaluations of scientific impact are built on the citation counts of papers (Kostoff, 1996; Lawani & Bayer, 1983; Moed, 2002; Narin & Hamilton, 1996; van Raan, 1996). Yet, the citation counting is kind of an approximation of a paper's popularity (Bollen et al., 2006), which could only reflect the academic influence in a limited area. In this case, we would like to bring PageRank, which is based on a different ranking mechanism, to conventional evaluative study on national scientific impact. Since the PageRank algorithm bears more information about the linking relationships between citing and cited papers, we may hopefully shed light on the national scientific impact from another prestige-oriented perspective.

2. The PageRank algorithm

In such complex networks as World Wide Web, an important attribute of a node is the in-degree (out-degree); namely the number of inbound (outbound) links on the node (Cohen, Havlin, & ben-Avraham, 2002). The in-degree of a given page could be considered as an approximation of a page's importance or quality (Page et al., 1998). Actually, the underlying idea is borrowed from the efforts in bibliometric study to define the research influence in terms of citations, which is simply read: citation counts are a measure of importance (Garfield, 1979). The PageRank algorithm (Brin & Page, 1998) has extended this idea by not counting the inbound links from all pages equally, but by normalizing via both the importance and the number of outbound links of the neighboring pages. In this respect, the PageRank value could serve as a better measure of importance, as it incorporates the paper's visibility and authority at the same time by taking both the number of citations and prestige of the citing papers into account.

[1] *CPP* denotes the indicator of Citation per paper; *JCSm* and *FCSm* denotes the corresponding Journal and Filed mean Citation Scores; Self-citations are excluded in the calculation of the ratios *CPP/JCSm* and *CPP/FCSm*, to prevent the ratios are affected by divergent self-citation behavior.

Page et al. (1998) defined the PageRank of a Web page A, denoted by $PR(A)$, using the following equation:

$$PR(A) = (1-d) + d \cdot \sum_i \frac{PR(T_i)}{C(T_i)} \tag{1}$$

where $PR(T_i)$ denotes the PageRank of page T_i which has connection with page A; $C(T_i)$ denotes the number of outbound links on page T_i; and d is a damping factor which can be set between 0 and 1.

In Eq. (1), we see that the PageRank of A is recursively defined by the PageRank of those pages that link to page A. Within the algorithm, the PageRank of pages T_i is always weighted by the number of outbound links $C(T_i)$, leading thereby to a smaller PageRank value transferred from pages T_i to the recipient page A. It is also assumed that any additional inbound link to a recipient page A will always increase A's PageRank.

There is a second version of PageRank algorithm as follows:

$$PR(A) = \frac{(1-d)}{N} + d \cdot \sum_i \frac{PR(T_i)}{C(T_i)} \tag{2}$$

where N is the total number of pages on the web.

Actually, the second version of the algorithm does not differ largely from the first one. However, it can better explain the metaphor of the original Random Surfing model suggested by Brin and Page (1998), in which the PageRank of a page is conceived as being the probability for a surfer visiting the page after clicking on many links. Thus, the probability for a surfer keeping clicking on links is given by the damping factor d, which is, depending on the degree of probability, set between 0 and 1. Since the surfer jumps to another page at random after he stops licking links, the probability therefore is implemented as the complementary part $(1-d)$ into the algorithm.

Due to the huge size of actual web, an approximate iterative computation is usually applied to calculate the PageRank. This means that each page is assigned an initial starting value and the PageRanks of all pages are then calculated in several computation circles based on Eq. (1) or (2). Take Eq. (1) for an example, the minimum PageRank of a page is given by $(1-d)$; while the maximum PageRank is determined as $dN + (1-d)$. This maximum can theoretically achieve, only if all Web pages solely link to one page, and this page also solely links to itself.

3. Data and methodology

In this work, we apply Eq. (1) of PageRank algorithm to the citation network with the goal of measuring the importance of scientific publications from another perspective. Using the seed journals in the category of "Biochemistry and Molecular Biology" as an entrance to the SCI papers during the time period 2000–2005, a relevant subject environment is established by taking all these journal papers into consideration. Although the JCR's coverage of journals in each subject category has been changing year by year, we find that the included journals have remained at 261 during 2003–2005 in the field of "Biochemistry and Molecular Biology". In this context, we determine the seed journals to be the 261 journals simply based on the 2005 JCR Science Edition.

Given the selected 261 journal names and predetermined time period 2000–2005, the 236,517 papers were searched in ISI databases, limiting to the document type of "article". Then, all the searching results were downloaded to the local computer with full record in a "tab delimited" form provided by ISI databases. To implement the computation, we compiled a computer program under the environment of Visual Studio.NET, using Microsoft SQL Server 2000 as the supporting database. The linking relationships between scientific papers were established through analyzing one of the bibliographic items called "cited reference". The extracted references were then matched up to the source papers item by item; those references, which had not find counterpart in the current dataset, would not be considered in the following calculation. Once the cited–citing relationships had been built up, a citation network came into existence.

A citation network is represented as heterogeneous, multi-relational papers, in which nodes are individual scientific papers and edges are citation links between them. In this study, the citation network is consisted of 236,517 nodes representing all articles published in 261 journals in the field of the "Biochemistry and Molecular Biology" during the time period of 2000–2005, and 511,212 links representing the citations (references) produced by all these papers.

During the calculation, we started with a uniform PageRank value equal to 1 for all papers and then iterated in the database. The convergence condition of iterations was set down using the residual vector. Let PR_i be a vector of PageRanks of all papers at the ith iteration. For an intermediate vector PR_i, let $\text{Residual}_i = PR_{i+1} - PR_i$. We treated $\|\text{Residual}_i\|$ as an indicator for how well PR_i approximates PR^*. $\|\text{Residual}_i\|$ is expected to be less than $\exp \times 10^{-5}$ after an adequate number of iterations. Eventually, a steady state set of PageRanks, namely PR^*, for all nodes of the network would be reached.

The damping factor d in the original study by Brin and Page (1998) was set as 0.85. This value was determined by their observation that an individual surfer will typically follow an order of 6 hyperlinks, bringing to a stopping probability $(1 - d) = 1/6 \approx 0.15$. However, in Chen, Xie, Maslov, and Redner's (2006) empirical study, it was found that scientific papers usually follow a shorter path of about average two links. Thus, we chose $d = 0.5$ to be our damping factor in the current citation network. Whatsoever, the value of damping factor d and its effect on the performance of PageRank algorithm still need further discussions in future studies.

4. Results and discussion

4.1. The introduction of "Internal Citations"

First, one should keep in mind the citations in the ISI databases are counted over all publications in a total of 105 subject categories; whereas we only made calculations for the subset of 261 journals included in "Biochemistry and Molecular Biology" during the time period of 2000–2005.

Due to the restricted range of seed journals and limited time period, we could not find the whole list of "cited reference" in the current dataset. According to the authors' calculation, the citations collected in the current network represent 1/5–1/6 of all citations recorded in ISI databases for all these papers included. In this respect, we brought forward a concept of Internal Citations, which represent the number of citations only by papers included in the limited seed journals and time period. As illustrated in Fig. 1, where nodes are papers and arrows represent the citing–cited relationship, the statistic of citations increases with growing amount of publications indexed by the databases. We may clearly observe that, the "ISI Database Citations" are actually kind of "Internal Citations" if compared with "All Citations".

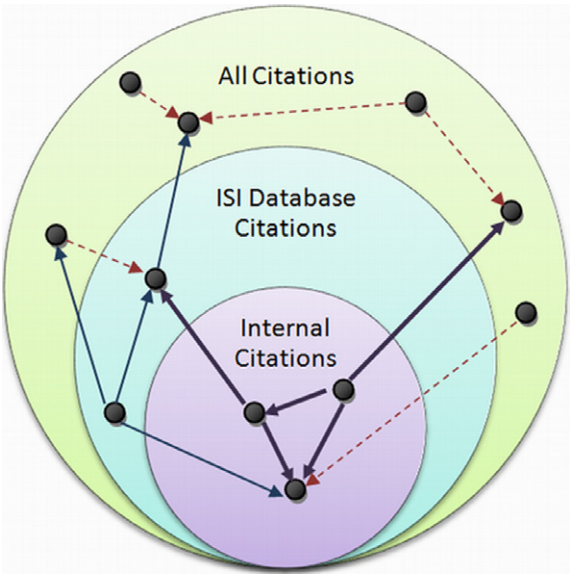

Fig. 1. The sketch map of various definitions of "citations".

4.2. The consistency between PageRanks and Internal Citations

To see how the PageRank algorithm performs, we may first take a look at the distributional characteristic of the PageRanks compared with the Internal Citations (see Fig. 2). Redner (1998), studying the citation distribution of 783,339 papers cataloged by ISI and 24,296 papers published in "*Physical Review D*" between 1975 and 1994, found that the probability a paper is cited k times follows a power law $P(k) \propto k^{-\gamma}$ with exponent $\gamma_{\text{cite}} = 3$. It indicates that the in-degree distribution of the citation network follows a power law. Another study by Vázquez (2001) extended the study to the out-degree distribution as well, finding that it also has an exponential tail.

In the following part, we focus on the distribution of PageRanks, viz. the number of papers $N(k)$ that are calculated to have a PageRank value of k. The figure is then produced on a semi-logarithmic scale. Instead of displaying all data points individually, we binned individual data points into separate groups, which can be a better guide to the eye, and the marker size was decided by the number of data points in the bins. It is indicated by Fig. 2 that the distribution of PageRanks has indeed the typical form of Power-Law as the data points basically fit the estimated exponential curve. The overall PageRanks follows a distribution of $N(k) \propto k^{-\gamma}$ with 0.95 for γ_{PR} and 0.903 for *R square*. In this case, we infer that the PageRank is also a substitutive indicator, which can also reflect the scale-free characteristic in such complex network.

We then calculate Kendall's τ and Spearman's ρ between the ranks of Internal Citations and PageRank value for each year during 2000–2005 (see Table 1) in order to test if there exists correlation relationship between them. The nonparametric correlations are adopted instead of the parametric correlation

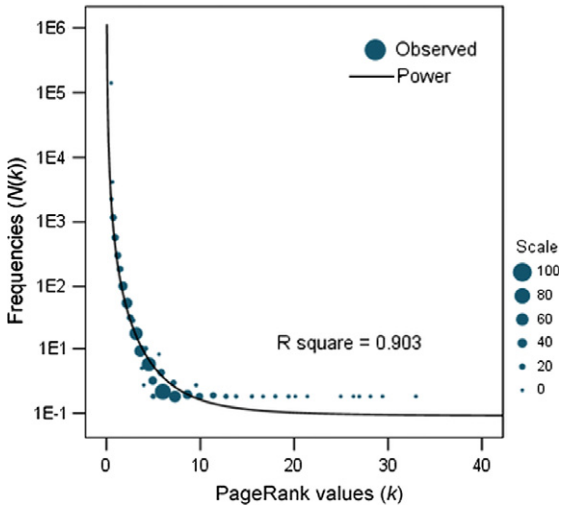

Fig. 2. The distribution of PageRank values.

Table 1
Correlations between the ranks of Internal Citations and PageRanks

	2000	2001	2002	2003	2004	2005
Number of publication	37,884	38,474	38,666	39,800	41,151	40,542
Kendall's τ						
Correlation coefficient	0.900[a]	0.890[a]	0.884[a]	0.890[a]	0.914[a]	0.975[a]
Significant (two-tailed)	0.000	0.000	0.000	0.000	0.000	0.000
Spearman's ρ						
Correlation coefficient	0.980[a]	0.976[a]	0.972[a]	0.973[a]	0.982[a]	0.998[a]
Significant (two-tailed)	0.000	0.000	0.000	0.000	0.000	0.000

[a] Correlation is significant at the 0.01 level (two-tailed).

(Pearson's γ) because the current data do not follow the Gaussian distributional assumption required by the Pearson's correlation coefficient. It is indicated in Table 1 that the two correlation coefficients have both reached about 0.9 at the 0.01 significant level (two-tailed), which suggests a highly correlated relationship between the PageRank and Internal Citations.

Given the above two points, the PageRank can be considered as a reliable indicator representing the importance of scientific publication in such citation network, and is strongly correlated with the traditional indicator of Internal Citations. This observation ensures that the evaluation studies based on PageRank could never lead to perverse results.

4.3. The disparity between PageRanks and Internal Citations

As two different measures of importance, the distinctions between them should be studied. For the sake of simplicity and clarity, the average PageRank for each group of papers with k Internal Citations, namely $PR(k)$ is calculated instead of individual PageRanks for each paper (see Fig. 3). It is indicated in Fig. 3 that the plot of $PR(k)$ versus k nicely fits the dashed line for smaller k, while the dispersion in $PR(k)$ becomes evident when k gets larger.

Focused on the dispersed part in Fig. 3, the top 10 highest ranking papers according to their PageRank values are listed in Table 2, together with their ranks of Internal Citations and brief bibliographic information.

While most of the papers listed in Table 2 are highly cited papers, we can also find one modestly cited paper that is highly ranked in the PageRank. Smardon's paper "EGO-1 is related to RNA-directed RNA polymerase and functions in germ-line development and RNA interference in C-elegans", which deals with the RNA interference technique, is listed in the 7th highest ranking papers according to PageRank values, while it ranks 204 based on Internal Citations. We happened to notice that the two laureates of 2006 Nobel Prize in Physiology or Medicine have been awarded because of their contribution in RNA interference techniques, which is "coincidentally" similar to the topic of Smardon's paper. We may believe that it is the potential important topic which lets Smardon's paper receive much concern from academic authorities and thus become a valuable research effort; otherwise it would probably be ignored due to limited number of citations. In this sense, the PageRank algorithm could identify several influential papers suffered from low citations, and furthermore make them visible to the research community.

In summary, we suggest that PageRank is a better indicator serving as a substitution of the number of citations for measuring papers' influence. On the one hand, it has high relevancy with the traditional citation measures and will hardly lead to perverse results. On the other hand, it could incorporate the importance of citing papers to a specific cited paper, therefore excavating several important papers that may suffer from low citations.

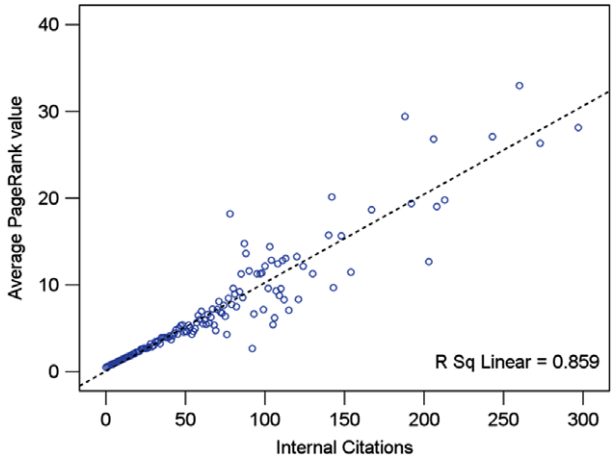

Fig. 3. The average PageRank value as a function of the Internal Citations.

Table 2
The top 10 highest ranking papers according to PageRank values

Rank of $PR(k)$	$PR(k)$	Rank of k	k	First author	Title	Publication year	Source
1	32.965	3	260	Du CY	Smac, a mitochondrial protein that promotes ...	2000	Cell
2	29.409	11	188	Schluenzen F	Structure of functionally activated small ...	2000	Cell
3	28.131	1	297	Kelley LA	Enhanced genome annotation using ...	2000	J. Mol. Biol.
4	27.075	4	243	Emanuelsson O	Predicting subcellular localization of proteins ...	2000	J. Mol. Biol.
5	26.795	8	206	Zamore PD	RNAi: Double-stranded RNA directs the ...	2000	Cell
6	26.313	2	273	Krogh A	Predicting transmembrane protein topology with ...	2001	J. Mol. Biol.
7	24.972	204	58	Smardon A	EGO-1 is related to RNA-directed RNA ...	2000	Curr. Biol.
8	21.404	6	213	Verhagen AM	Identification of DIABLO, a mammalian protein that ...	2000	Cell
9	20.123	17	142	Eskes R	Bid induces the oligomerization and ...	2000	Mol. Cell Biol.
10	19.753	15	148	Schwartz S	PipMaker – A Web server for aligning two genomic ...	2000	Genome Res.

4.4. Exploratory analyses of scientific performance based on PageRank

It is previously mentioned in Section 1 that the PageRank algorithm is supposed to provide us an integrated picture of academic influence for the evaluation of national scientific performance. Thus, we are motivated to conduct the following analyses based on the calculated PageRank values and make further conclusions on the research status of individual countries. In addition, local reorder of countries with respect to the PageRank is discussed and compared with the rankings of citation counts.

According to authors' calculation, there are altogether 144 countries contributed to the publication activity in the field "Biochemistry and Molecular Biology" as reported in the current dataset. For the publications affiliated to more than one country, we assign 1 to each country uniformly and take them as the output of all the affiliated countries. In the present study, we concentrate on the top eight countries[2] and 25 members of the European Union[3] which together account for 90% of the world's total publications in the field of "Biochemistry and Molecular Biology".

Fig. 4 provides an overview of the publication contribution from the top eight productive countries as well as the European Union. In Fig. 4, the sub-pie chart on the right hand is drawn from the "EU 25" part of the left-hand pie chart. The percentage in brackets denotes the ratio of each country's publications to the world's total. It indicates that USA is the second largest shareholder in terms of publications covered by "Biochemistry and Molecular Biology", while EU takes the leading position as a research union. Among all 25 member states of EU, the former 15 members take 93% of the EU's total publications and 35% of the world's total publications. It is indicated in the sub-pie chart that the United Kingdom, Germany and France do play a leading role in the European Union and even enter the top five on the world stage. The Asian countries, like Japan, PR China, South Korea and India, are still active contributors as they always perform in many other scientific fields, such as computer science (Guan & Ma, 2004) and nanoscience and nanotechnology (Zhou & Leydesdoff, 2006), etc.

[2] The names of the eight countries have been presented in abbreviation for the sake of clarity. And they are USA (US), Japan (JP), Canada (CA), PR China (CN), South Korea (KR), Australia (AU), India (IN) and Russia (RU).

[3] Among the present 25 members of the European Union, 15 of them were discriminated to be the former members of the EU before May 2004, including France (FR), Germany (DE), Italy (IT), Netherlands (NL), Belgium (BE), Luxemburg (LU), Denmark (DK), Ireland (IE), United Kingdom (GB), Greece (GR), Spain (ES), Portugal (PT), Austria (AT), Finland (FI) and Sweden (SE); 10 more countries have later on joined in the EU, like: Cyprus, Malta, Poland, Hungary, Czech Republic, Slovakia, Slovenia, Estonia, Latvia and Lithuania. The reason for discriminating the former 15 countries lies in the fact that they were and still are playing a leading role in this organization, which we will also see in the following analyses.

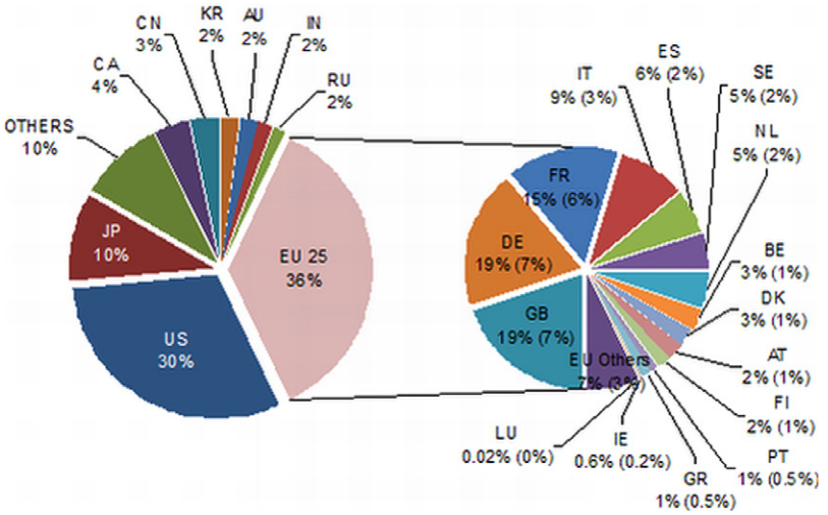

Fig. 4. The publication contribution of top eight countries and the European Union in the period of 2000–2005.

As a surrogate indication of scientific impact, the average PageRank values of scientific publications have been then calculated at the national level for all the above-mentioned countries, including: USA, Japan, Canada, PR China, South Korea, Australia, India, Russia and former 15 member states of EU, as well as the left 10 member states of EU represented by ''EU 10'' as one community. In the following part, ''research status'' is defined to depict the two-dimensional scientific performance, consisting of scientific productivity and academic influence. And the two dimensions are separately measured by the number of publications and average PageRank values as illustrated in Fig. 5.

The countries in Fig. 5 are represented by different shapes, e.g. hollow squares for the former 15 member states of EU, solid dots for the four Asian countries, and solid triangles for the other countries together with the left 10 member states of EU as a whole.

In Fig. 5, USA can again be characterized by a preponderance of both scientific productivity and academic influence in the field of ''Biochemistry and Molecular Biology''. Focused on the upper-left part where countries do high-quality researches under a limited productivity level, there are altogether 11 out of former 15 member states of EU. This suggests that the European Union has provided its member states with a barrier-free platform on which they could easily communicate and thus improve their research level.

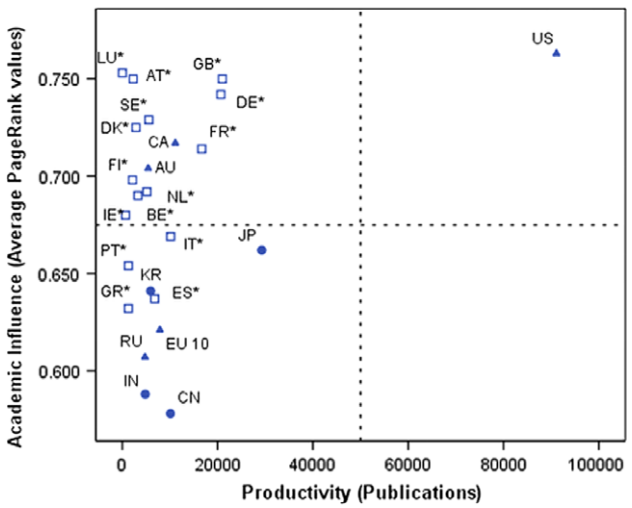

Fig. 5. The research ''status'' of countries as reflected by the scientific productivity and its corresponding academic influence.

Table 3
The comparison of two importance measures aggregated on the national level

Country	Number of publications	Internal citations per paper (ICPP)	Rank of ICPP	PageRank per paper (PRPP)	Rank of PRPP
USA	91,079	2.991	1	0.763	1
Luxembourg*	24	1.750	14	0.753	2
UK*	20,984	2.790	3	0.752	3
Austria*	2287	2.798	2	0.750	4
Germany*	20,680	2.640	4	0.742	5
Sweden*	5584	2.384	8	0.729	6
Denmark*	2889	2.550	5	0.725	7
Canada	11,124	2.448	6	0.717	8
France*	16,674	2.404	7	0.714	9
Australia	5442	2.096	10	0.704	10
Finland*	2150	2.062	11	0.698	11
Netherlands*	5144	2.132	9	0.692	12
Belgium*	3283	2.048	12	0.690	13
Ireland*	692	2.036	13	0.680	14
Italy*	10,158	1.741	15	0.669	15
Japan	29,256	1.725	16	0.662	16
Portugal*	1303	1.498	17	0.654	17
South Korea	5944	1.457	19	0.641	18
Spain*	6801	1.438	20	0.637	19
Greece*	1285	1.487	18	0.632	20
EU 10	7891	1.187	21	0.621	21
Russia	4757	1.042	22	0.607	22
India	4821	0.790	23	0.588	23
PR China	10,111	0.729	24	0.578	24

Note: The former 15 member states of EU are marked with ∗.

On the other hand, four Asian countries have all located in the lower-left part in Fig. 5, which suggests a pessimistic research status with lower prestige and fewer publications. Actually, the research efforts in Asian countries have been suffering from the low international visibility, due to some rooted barriers of language, culture, communication channels, etc. In a sense, they are still peripheric countries toward the mainstream research community.

Finally, a comparison between PageRank and Internal Citations is made on the national level. The items in Table 3 are sorted in an ascending order of the Rank of average PageRank values.

According to Table 3, Luxembourg has remarkably high influence in ''Biochemistry and Molecular Biology'' as reflected by the average PageRank value. Regarding its few publications, the research quality of Luxembourg is extraordinarily high compared with most other countries. Focusing on the distinction between two indicators, one may find the ranking of countries based on PageRank has reordered compared with that of average Internal Citations. For countries that rank higher on the PageRank than on Internal Citations, like Luxemburg, Sweden, etc., we may suggest that these countries are potentially influential in this specific field. For the countries that rank lower on the PageRank than on Internal Citations, like Denmark, Canada, France, etc., their importance might be slightly overestimated by the citation counts for their research publications are not adequately cited by influential papers.

However, we should always keep in mind the PageRank indicator and its corresponding citation network in the current study was established on the basis of a specific field during a limited time period. The result must be local and field-specific. Yet we may expect that the PageRank indicator will be more universalized with expanded data source and long stretching time period.

5. Concluding remarks

The present study attempts to measure the importance of scientific publications from another perspective. The PageRank algorithm provides a meaningful extension to the traditionally used number of citations for individual papers or aggregation of papers at different levels.

The citation network of scientific publications is an important resource that would provide additional information compared to the traditional citation analysis. The PageRank is an objective measure of its citation importance that corresponds well with people's subjective idea of importance. Because of this correspondence, PageRank is an excellent way to prioritize the influence of research papers suffering from relatively lower citations, therefore excavating the potential influence of scientific publications.

The evidence from national comparison of scientific performance in the field of ''Biochemistry and Molecular Biology'' shows that US always plays the leading role in both publication productivity and academic influence during 2000–2005. Eleven countries out of the former 15 member states of European Union have been conducting high quality scientific researches at a relatively lower level of production. By comparison, the research influences of Asian countries have been much lower when their scientific productivity remains low either.

A meaningful advantage of PageRank is that it could largely eliminate the flattery of academic influence caused by author self-citations. Nowadays, while the ISI databases could automatically filter out self-citations from total citation counts, bibliometricians are able to easily operate the citation analysis in a ''self-citation-free'' environment. However, it is somewhat reckless to completely ignore the effect of author self-citations, as there probably are some thoughtful considerations and instructive nexus when authors cite their previous works. In this case, the PageRank algorithm treats the author self-citations with more consciousness. The PageRank contribution of a author's paper T_i to his own paper A could be strongly diluted if the papers (T_i) neither are important (large $PR(T_i)$) nor have a short list of references (small $C(T_i)$). Otherwise, the self-citation links bearing valuable academic relationships would still contribute a significant part to the ranking of an author's own cited paper.

However, the robustness of PageRank algorithm with respect to the free parameter d is still under discussion. As previously mentioned, we choose $d = 0.5$ based on Chen's empirical study that most scientists normally follow an order of 2 citation links in a citation network. Besides, we also have calculated the PageRank values on different situations of damping factor d. For $d = 0.85$, as it was in the original Google algorithm, some local changes have been observed for those highly ranked papers when $d = 0.5$. According to our statistics, all the top 10 PageRank papers calculated under $d = 0.5$ have remained in the top 35 PageRank papers when $d = 0.85$. In other words, there is little global reordering of papers when the damping factor d changes from 0.5 to 0.85. Thus, we may conclude that PageRank indicator is a robust measurement representing the academic influence of scientific papers.

Acknowledgments

This research is funded by National Natural Science Foundation of China (Project No. 70573009). Authors are grateful for the valuable comments and suggestions of both the anonymous reviewers and the editor, which significantly improved the paper. The authors, naturally, bear responsibility for any remaining shortcomings.

References

Anderson, A. (1992). Molecular biology: US juggernaut overwhelms divided European elite. *Science, 256*, 460–464.

Anon. (1992). Top 50 research institutions in molecular biology ranked by citation impact 1981–1991. *Science Watch, 3*(7).

Bollen, J., Rodriguez, M. A., & van de Sompel, H. (2006). Journal status. *Scientometrics, 69*(3), 669–687.

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology, 92*(5), 1170–1182.

Brin, S., & Page, L. (1998). *The anatomy of a large-scale hypertextual web search engine*. Retrieved May 22, 2006. Available from http://www.db.stanford.edu/pub/papers/google.pdf.

Chen, P., Xie, H., Maslov, S., & Redner, S. (2006). *Finding scientific gems with Google*. Retrieved May 22, 2006. Available from http://arxiv.org/abs/physics/0604130.

Chubin, D. E., & Hackett, E. J. (1990). *Peerless science: Peer review and US science policy*. Albany, NY: State University of New York Press.

Cohen, R., Havlin, S., & ben-Avranham, D. (2002). *Structural properties of scale-free networks*. Berlin: Wiley–VCH Verlag.

Dalpé, R. (2002). Biliometric analysis of biotechnology. *Scientometrics, 55*(2), 189–213.

Garfield, E. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*. New York, NY: Wiley.

Garfield, E. (1983). How to use citation analysis for faculty evaluations, and when is it relevant? (Part 1). *Current Contents, 44*, 5–13.

Glänzel, W. (1996). The needs for standards in bibliometric research and technology. *Scientometrics, 35*, 167–176.
Guan, J. C., & Ma, N. (2004). A comparative study of research performance in computer science. *Scientometrics, 61*, 339–359.
Herbertz, H., & Müller-Hill, B. (1995). Quality and efficiency of basic research in molecular biology: A bibliometric analysis of thirteen excellent research institutes. *Research Policy, 24*, 959–979.
He, T., Zhang, J., & Teng, L. (2005). Basic research in biochemistry and molecular biology in China: A bibliometric analysis. *Scientometrics, 62*(2), 249–259.
Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46*(5), 604–632.
Koenig, M. E. D. (1982). Determinants of expert judgment of research performance. *Scientometrics, 4*, 361–378.
Koenig, M. E. D. (1983). Bibliometric indicators versus expert opinion in assessing research performance. *Journal of the American Society for Information Science, 34*, 136–145.
Kostoff, R. N. (1996). Performance measures for government-sponsored research: Overview and background. *Scientometrics, 36*, 281–292.
Lawani, S. M., & Bayer, A. E. (1983). Validity of citation criteria for assessing the influence of scientific publications: New evidence with peer assessment. *Journal of the American Society for Information Science, 34*, 59–66.
Martin, B. R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics, 36*, 343–362.
Moed, H. F. (2002). Measuring China's research performance using the Science Citation Index. *Scientometrics, 53*(3), 81–96.
Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Cherry Hill, NJ: Computer Horizons.
Narin, F., & Hamilton, K. S. (1996). Bibliometric performance measures. *Scientometrics, 36*, 293–310.
Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The PageRank citation ranking: Bringing order to the web*. Retrieved May 22, 2006. Available from http://dbpubs.stanford.edu/pub/1999-66.
Pendlebury, D. (1990). Cold Spring Harbor tops among independent labs. *Scientist*, 20.
Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with applications to the literature of physics. *Information Processing and Management, 12*(5), 297–312.
Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *European Physics Journal B, 4*, 131–134.
Snyder, H., & Rosenbaum, H. (1999). Can search engines be used for web-link analysis? A critical review. *Journal of Documentation, 55*(4), 375–384.
Thomas, P. R., & Watkins, D. S. (1998). Institutional research rankings via bibliometric analysis and direct peer review: A comparative case study with policy implications. *Scientometrics, 41*, 335–355.
van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer-review based evaluation and foresight exercises. *Scientometrics, 36*, 397–420.
van Raan, A. F. J. (1997). Scientometrics: State of the art. *Scientometrics, 38*, 205–218.
Vázquez, A. (2001). *Statistics of citation network*. Retrieved June 13, 2006. Available from http://arxiv.org/abs/cond-mat/0105031.
Zhou, P., & Leydesdoff, L. (2006). The emergence of China as a leading nation in science. *Research Policy, 35*(1), 83–104.