Concise Papers

Comprehensive Citation Index for Research Networks

Henry H. Bi, Jianrui Wang, and Dennis K.J. Lin

Abstract—The existing *Science Citation Index* only counts direct citations, whereas *PageRank* disregards the number of direct citations. We propose a new *Comprehensive Citation Index* (*CCI*) that evaluates both direct and indirect intellectual influence of research papers, and show that *CCI* is more reliable in discovering research papers with far-reaching influence.

Index Terms—Citation analysis, citation networks, comprehensive citation index, PageRank, science citation index.

1 Introduction

As an essential part of research papers, citation serves two broad functions: 1) it directs readers to the sources of knowledge that has been drawn upon in one's work, and enables readers to assess the knowledge claims in the cited sources for themselves; and 2) it maintains intellectual traditions (such as giving credit to the cited works) and provides peer recognition in the research community [1], [2]. Consequently, citation has been used as a tool for searching research papers [3], [4], [5], and assessing research productivity [6]. The most popular citation analysis method is probably *Science Citation Index* (*SCI*) [4]. *SCI* ranks research papers according to the number of direct citations that papers receive: the more citations a paper has, the more significant the paper is. To demonstrate *SCI* [4], Garfield originally gives an example of a citation network [7] consisting of 15 papers, as reproduced in Fig. 1a. According to *SCI*, Paper 2 is the most influential paper in this citation network because it has more citations than any other papers.

Because *SCI* is restricted to direct citations, there are two serious concerns. First, not all citations are equally important. For example, Paper 1 in Fig. 1a is cited by Papers 2, 3, 4, 6, and 15; Paper 1's citations from Papers 2 and 4, which have more citations themselves, should carry more weights than its citations from Papers 3, 6, and 15, which have fewer citations themselves. The subgraphs in Figs. 1b, 1c, and 1d clearly show the citations of Papers 2, 3, 4, 6, and 15. Second, direct citations only reflect the immediate impact of papers, but the overall influence of papers should not be limited to direct citations. This is because many papers' far-reaching intellectual influence over years and decades cannot be explained solely by their direct citations.

2 Comprehensive Citation Index

2.1 Mathematical Formulation

In general, each paper's intellectual influence is passed on to its citing papers, to the papers that cite its citing papers, to the papers that cite the citing papers of its citing papers, and so on. Hence, a paper's overall intellectual influence should consist of both 1) direct influence on its citing papers and 2) indirect influence through citation links on those papers that do not directly cite it, and such indirect influence decreases through each citation link.

To model a paper's overall influence in terms of citations, let the weight $\beta (0 \leq \beta < 1)$ be the portion of influence that each paper distributes evenly to all the papers that it cites. $\beta < 1$ is consistent with the fact that, in general, although each paper is influenced by the papers that it cites, its unique intellectual merit (which is represented by the portion $1 - \beta$) should be greater than zero and should not be attributed to the papers that it cites.

Then, a paper's overall influence in a citation network can be modeled as

$$x_i = |J_i| + \beta \sum_{j \in J_i} \frac{x_j}{r_j} = \sum_{j \in J_i} \left(1 + \beta \frac{x_j}{r_j}\right), \tag{1}$$

where x_i is Paper i's *Comprehensive Citation Index (CCI)* value, which represents Paper i's overall influence in terms of citations; J_i is the set of papers that directly cite Paper i; $|J_i|$ is the cardinality of set J_i, and is the number of direct citations (i.e., direct influence) that Paper i has; r_j is the number of papers (including Paper i) directly cited by Paper j; $\beta \frac{x_j}{r_j}$ is the portion of Paper j's influence attributed to Paper i; $\beta \sum_{j \in J_i} \frac{x_j}{r_j}$ is the total amount of Paper i's indirect influence on the papers in this citation network.

Equation (1) can be represented in a matrix form for all papers in a citation network as follows:

$$\begin{aligned}\mathbf{x} = \mathbf{He} + \beta\mathbf{Gx} = &\begin{pmatrix} h_{11} & \dots & \dots & h_{1n} \\ h_{21} & \dots & & h_{2n} \\ \dots & & \dots & \dots \\ h_{n1} & \dots & \dots & h_{nn} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix} \\ &+ \beta \begin{pmatrix} g_{11} & \dots & \dots & g_{1n} \\ g_{21} & \dots & & g_{2n} \\ \dots & & \dots & \dots \\ g_{n1} & \dots & \dots & g_{nn} \end{pmatrix} \mathbf{x},\end{aligned} \tag{2}$$

where $\mathbf{x}$ is the *CCI* vector (i.e., overall influence); $\mathbf{H}$ is the citation network matrix such that $h_{ij} = 1$ if Paper j cites Paper i and $h_{ij} = 0$ otherwise; $g_{ij} = \frac{h_{ij}}{r_j}$ for $\mathrm{r}_j \neq 0$ and $g_{ij} = 0$ otherwise; and $\mathbf{e}$ is a vector of ones.

Equation (2) can be rewritten as $(\mathbf{I} - \beta\mathbf{G})\mathbf{x} = \mathbf{He}$, where $\mathbf{I}$ is an identity matrix. $\mathbf{I} - \beta\mathbf{G}$ is called an M-matrix [8], which is nonsingular when $0 \leq \beta < 1$. Therefore, (2) has a unique solution $\mathbf{x} = (\mathbf{I} - \beta\mathbf{G})^{-1}\mathbf{He}$. When $\beta = 0$, *CCI* is the same as *SCI*.

2.2 An Illustrative Example

We use the simple citation network in Fig. 1a to intuitively illustrate the rationale of *CCI*. Table 1 shows the computation results of both *SCI* and *CCI* for this citation network. The main insights are summarized as follows:

1. As shown in Fig. 1b, Paper 2 cites Paper 1 and almost half of Paper 2's citing papers also cite Paper 1. Hence, Paper 1 has both direct and indirect influence on those citing papers of Paper 2.
2. Fig. 1c shows that Paper 1 has direct influence on Paper 4 as well as indirect influence on Paper 4's citing papers.
3. *SCI* has the same ranking for Papers 1 and 4 (each of which has five direct citations), and ranks Paper 2 (which has seven direct citations) higher than Paper 1. But based on 1) and 2) above, it is likely that Paper 1 is more influential than Papers 2 and 4. This observation is confirmed by the *CCI* rankings in Table 1 with $\beta = 0.3$. Note that the sensitivity analysis of β will be conducted in Section 4.

- H.H. Bi is with the Atkinson Graduate School of Management, Willamette University, 900 State Street, Salem, OR 97301. E-mail: hbi@willamette.edu.
- J. Wang is with the Syncsort Inc., 50 Tice Boulevard, Woodcliff Lake, NJ 07677. E-mail: jianrui@gmail.com.
- D.K.J. Lin is with the Department of Statistics, The Pennsylvania State University, 317 Thomas Building, University Park, PA 16802-2111. E-mail: DennisLin@psu.edu.

Manuscript received 12 Jan. 2008; revised 3 Nov. 2009; accepted 15 Mar. 2010; published online 7 Sept. 2010.
Recommended for acceptance by M. Garofalakis.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2008-01-0018.
Digital Object Identifier no. 10.1109/TKDE.2010.167.

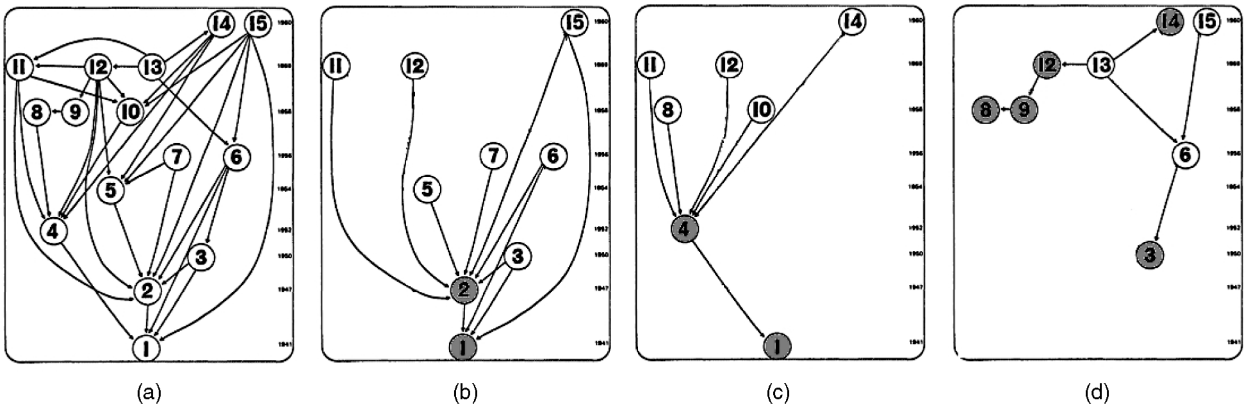

Fig. 1. A citation network consisting of 15 papers: (a) is directly adopted from [4]; (b), (c), and (d) are subgraphs of (a).

4. As shown in Fig. 1d, *SCI* has the same ranking for Papers 3, 8, 9, 12, and 14 (each of which has one direct citation), but *CCI* ranks some of those papers differently in Table 1. The differences can be explained by the fact that those papers are cited by papers that have different influences. For example, the *CCI* ranking of Paper 3 is higher than that of Paper 12, because Paper 3's citing paper (i.e., Paper 6 with $CCI = 2.00$) is more influential than Paper 12's citing paper (i.e., Paper 13 with $CCI = 0$).

This example shows that *CCI* has better resolution than *SCI* and is capable of differentiating the importance of different citations. This distinctive feature of *CCI* is useful for precisely evaluating the different influences of papers, which may have the same or similar number of direct citations.

3 Related Works

So far we have used *SCI* to explain our motivation why we develop a new citation analysis method. Now we will discuss related works to justify the novelty of *CCI*.

3.1 PageRank

PageRank [8], [9], [10] in link analysis [8], [11], [12] considers that in a network, each incoming link is different such that an incoming link has more value if it comes from a more important node. The *PageRank* algorithm [9], [10] has been used to rank webpages.

TABLE 1
Comparison between *SCI* and *CCI* for the Citation Network in Fig. 1a

Paper	*SCI*	*SCI* Ranking	*CCI* ($\beta = 0.3$)	*CCI* Ranking	Ranking Change (*SCI* Ranking – *CCI* Ranking)
1	5	2	10.16	1	1
2	7	1	8.88	2	-1
3	1	8	1.20	9	-1
4	5	2	7.06	3	-1
5	4	4	4.15	5	-1
6	2	6	2.00	7	-1
7	0	13	0.00	13	0
8	1	8	1.32	8	0
9	1	8	1.05	10	-2
10	4	4	4.36	4	0
11	2	6	2.05	6	0
12	1	8	1.00	11	-3
13	0	13	0.00	13	0
14	1	8	1.00	11	-3
15	0	13	0.00	13	0

PageRank is defined as [10], [13]:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in I(p_i)} \frac{PR(p_j)}{O(p_j)}, \quad (3)$$

where $p_1, p_2, \ldots, p_N$ are the pages; N is the total number of pages under consideration; $I(p_i)$ is the set of pages that link to p_i; $O(p_j)$ is the number of outbound links from p_j; d is a damping factor that is the probability that, at any step, a person will continue clicking on links. Note that $\beta \sum_{j \in J_i} \frac{x_j}{r_j}$ in the *CCI (1)* has a similar form as $d \sum_{p_j \in I(p_i)} \frac{PR(p_j)}{O(p_j)}$ in the *PageRank (3)*. This is because in *CCI*, each paper distributes a portion of its overall influence evenly to all the papers that it cites, while in *PageRank*, "the rank of a page is divided among its forward links evenly to contribute to the ranks of the pages they point to" [10, p. 4].

Although the application of *PageRank* has proven that it is an effective algorithm in ranking webpages, it is improper to apply *PageRank* to citation analysis, because *PageRank* disregards the number of direct citations. As explicitly pointed out by the developers of *PageRank*, "there are a number of significant differences between webpages and academic publications" [10, p. 1]. In particular, "simple backlink (i.e., incoming link or direct citation) counts have a number of problems on the web. Some of these problems have to do with characteristics of the web which are not present in normal academic citation databases" [10, p. 2]. In addition, links among webpages do not necessarily represent any intellectual influence between pages. As a result, the incoming link counts (i.e., direct citations) of a page p_i are not included in p_i's *PageRank* $PR(p_i)$ in *(3)*. Moreover, because $\frac{1-d}{N}$ in *(3)* is less than one, it does not represent incoming link counts. $\frac{1-d}{N}$ represents the probability that when a random surfer arrives a webpage with no outbound link, the surfer picks another webpage at random and continues surfing again. But such randomness does not exist in citation.

Different from links among webpages that do not represent intellectual influence between webpages, citations reflect direct and indirect intellectual influence from a paper to its citing papers, to its citing papers' citing papers, and so on. Direct intellectual influence is the fundamental part in citations. Hence, even when indirect influence is considered, the importance of direct citations still must be sufficiently evaluated. *CCI* properly captures direct citations as $|J_i|$ in (1).

3.2 Status or Rank Prestige

In social network analysis, a method has been proposed "to measure the prestige of the actors in a set of actors" by considering "the prominence of the individual actors who are doing the 'choosing'" [14, p. 205]. Specifically, an "actor's rank depends on the ranks of those who do the choosing; but note that the ranks of those who are choosing depend on the ranks of the actors who

TABLE 2
The *CCI*, *SCI*, and *PageRank* Rankings of the Top-10 Most Influential Papers Published in *Management Science* between 1954 and 2003

No.	Title (in alphabetical order)	*CCI* rankings with different weight β										*PageRank* rankings with different damping factor d								
		0 (*SCI* ranking)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	A New Product Growth for Model Consumer Durables	7	8	7	7	7	7	7	7	7	6	7	7	8	9	10	12	19	26	55
2	A Suggested Computation for Maximal Multi-Commodity Network Flows	320	210	155	119	95	87	82	76	70	67	335	261	224	206	202	220	266	327	535
3	Dynamic Version of the Economic Lot Size Model	24	25	25	24	26	25	27	27	29	29	53	59	64	77	97	121	169	279	472
4	Games with Incomplete Information Played by 'Bayesian' Players, I: The Basic Model	22	15	15	13	12	12	11	11	11	12	14	13	11	10	7	6	7	8	9
5	Information Distortion in a Supply Chain: The Bullwhip Effect	11	11	11	11	11	11	12	12	13	14	18	18	21	24	27	35	48	75	157
6	Jobshop-Like Queueing Systems	26	24	17	17	16	16	16	16	16	16	34	24	18	14	13	11	9	11	11
7	Linear Programming under Uncertainty	56	46	37	32	29	30	28	25	27	26	67	65	62	64	79	94	120	171	330
8	Models and Managers - Concept of a Decision Calculus	71	59	52	48	43	40	41	40	39	37	73	72	69	71	81	97	121	172	328
9	Optimal Policies for a Multi-echelon Inventory Problem	23	20	16	16	15	15	14	13	12	11	39	32	23	18	16	15	15	13	17
10	The LaGrangian Relaxation Method for Solving Integer Programming-Problems	15	13	12	12	13	13	13	14	14	17	17	15	14	17	19	27	39	57	126
	Average:	**57.5**	**43.1**	**34.7**	**29.9**	**26.7**	**25.6**	**25.1**	**24.1**	**23.8**	**23.5**	**65.7**	**56.6**	**51.4**	**51.0**	**55.1**	**63.8**	**81.3**	**113.9**	**204.0**

choose them, and so on" [14, p. 206]. The *rank prestige* $P_R(n_i)$ for actor n_i within a set of g actors is defined as [14, p. 206]:

$$P_R(n_i) = x_{1i}P_R(n_1) + x_{2i}P_R(n_2) + \cdots + x_{gi}P_R(n_g),$$
$$\text{where } x_{ji} = \begin{cases} 1, & \text{if actor } n_j \text{ chooses actor } n_i \\ 0, & \text{otherwise.} \end{cases} (1 \le i, j \le g). \quad (4)$$

However, (4) is improper for evaluating the impact of papers in citation networks. This is because (4) inappropriately implies that each paper has no unique intellectual merit since (4) attributes each paper's overall influence completely to the papers that it cites. In comparison, the *CCI* (1) does not have this problem.

3.3 Y-Factor

Y-factor is proposed to rank journals [13]. *Y-factor* is defined as a product of a journal's impact factor and that journal's Weighted *PageRank*. Although impact factor and Weighted *PageRank* may make sense separately, the meaning of their product is not clear, just as the developers of *Y-factor* point out explicitly that the "definition of the *Y-factor* rankings may not be scientifically convincing" [13, p. 686].

3.4 h-Index and g-Index

h-index [15] is proposed for quantifying the scientific productivity of individuals. If an individual has published N papers, then she has index h if h of her N papers have at least h citations each and the other $(N - h)$ papers have $\le h$ citations each. *g-index* [16] is similar to *h-index*. For an individual, if her papers are listed in the decreasing order of the number of citations that they received, then this individual's *g-index* is the largest number such that the top g papers together received at least g^2 citations. Clearly, *h-index* and *g-index* have a focus on the impact of individual researchers, which is different from *CCI* that evaluates the impact of individual papers.

4 Evaluation and Analysis

In this section, we use a benchmark to evaluate *CCI* in comparison with *SCI* and *PageRank*. Here, we use peer review as the benchmark. This is because peer review is broadly used in practice [17], and peer review provides an alternative assessment based on human inputs (in contrast with *CCI*, *SCI*, and *PageRank* based on computation).

We evaluate and compare *CCI*, *SCI*, and *PageRank* by applying them to a large citation network. From 1/31/2007 to 2/23/2007, we collected from http://scholar.google.com a citation data set that contains 288,404 entries between 1950 and 2004. This data set includes 5,003 papers published in the journal of *Management Science*, their cited papers and citing papers, the cited papers of their cited papers, the citing papers of their citing papers, and so on, which may or may not be published in *Management Science*. Although all entries in this data set have been used in calculating *CCI*, *SCI*, and *PageRank*, only the papers published in *Management Science* are included in the *CCI*, *SCI*, and *PageRank* rankings.

The reasons that we use this citation network include: first, in 2004, the INFORMS members chose the top-10 most influential papers published in *Management Science* between 1954 and 2003 [18]. Those top-10 papers are the results of peer review by a large number of INFORMS members. Ideally, the peer-review rankings of those top-10 papers are $1, 2, \ldots, 10$ with the average ranking $= 5.5$. Second, this citation network is large enough to provide reliable information. Finally, the same paper may appear in Google multiple times for various reasons. To improve the accuracy of paper rankings, manual cleaning work has to be performed to combine duplicate entries that represent the same paper into one. This

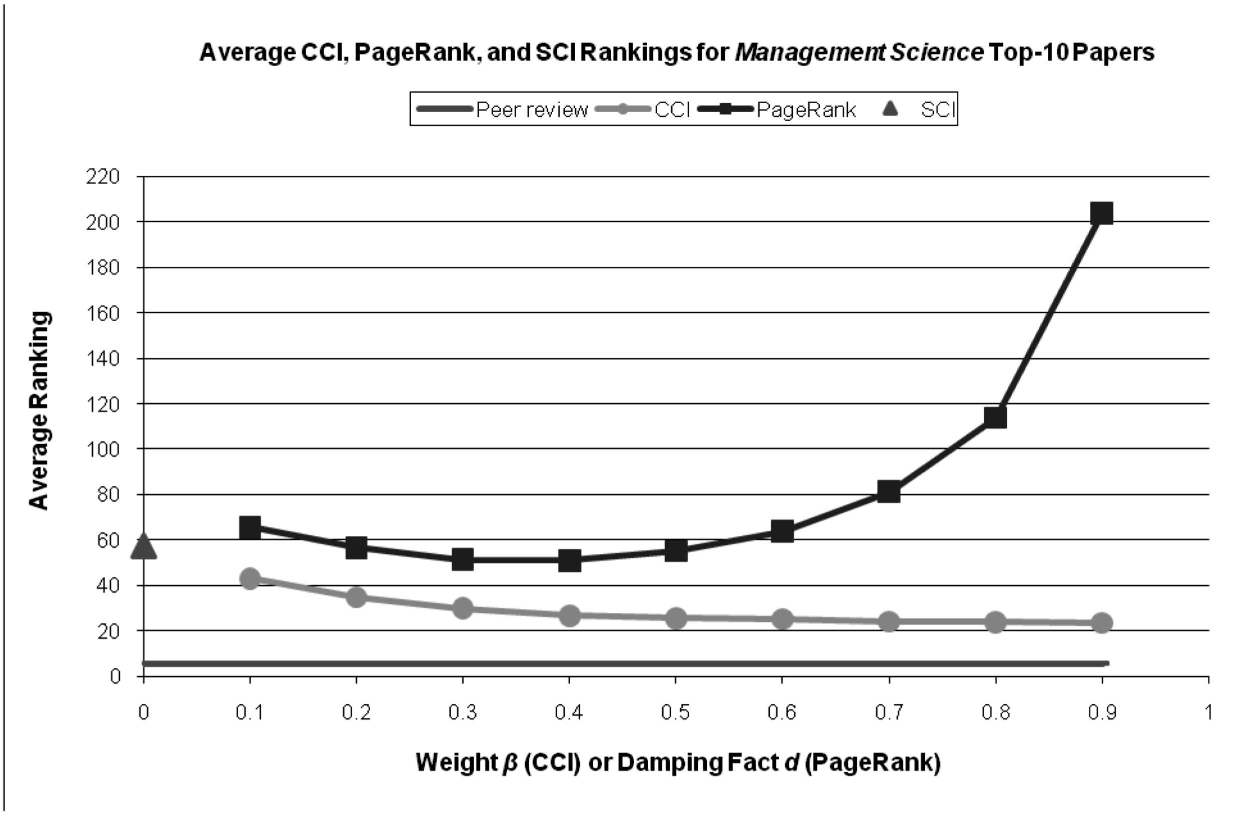

Fig. 2. Sensitivity analysis of the average *CCI*, *PageRank*, and *SCI* rankings in Table 2.

citation network is also "small" enough for us to possibly go through all entries to do cleaning work.

Table 2 shows the *CCI*, *SCI*, and *PageRank* rankings of those top-10 papers among 5,003 papers published in *Management Science*. Those rankings are based on the calculation of *CCI* and *PageRank* values (using (1) and (3), respectively), which are not shown in Table 2 for brevity. Table 2 and Fig. 2 also provide sensitivity analysis for the different values of weight β (*CCI*) and damping factor d (*PageRank*). When $\beta = 0$, *CCI* rankings are the same as *SCI* rankings.

Table 2 and Fig. 2 provide some useful insights. First, the *CCI* rankings of those top-10 papers are consistently closer to the peer review results (i.e., the average peer-review ranking $= 5.5$) and better than both *SCI* and *PageRank* rankings. Second, the average *CCI* ranking of those top-10 papers is improved gradually from $\beta = 0.1$ to $\beta = 0.9$ and is very stable when $0.3 \leq \beta \leq 0.9$. Finally, the *PageRank* algorithm requires that $d < 1$ for possible convergence [8, p. 47]; when $d = 0$, all papers have the same *PageRank*, which is trivial and not shown in Fig. 2.

Note that in the *CCI* (1), the weight β represents the portion of intellectual influence that Paper j distributes evenly to all the papers that it cites; that is, this portion of intellectual influence (i.e., existing knowledge) is originally created by all the papers that Paper j cites, not created by Paper j itself. The portion of intellectual influence (i.e., new knowledge) created by Paper j is represented by $1 - \beta$. Therefore, the characteristics of specific citation networks should be considered when choosing β for different citation networks. In general, if papers in a citation network are largely based on previous research works, then β may be given a large value; if papers in a citation network typically involve innovative research, then giving β a small value is more appropriate.

It is worth noting that to examine whether *CCI* is robust to noises, we have cleaned the *Management Science* data set (which contains 288,404 entries) by deleting noisy entries that have no citation and no publication year. Those noises include lecture slides, course notes, speeches, white papers, etc., which are not typical research publications like journal or conference papers and, thus, are not useful for research citation analysis. The cleaned data set contains 219,634 entries (about 76 percent of the original total). The detailed calculation displayed in a chart similar to Fig. 2 shows that the two *CCI* curves (before and after cleaning) are very close to each other with a similar shape and trend. This demonstrates the robustness of the *CCI* method against noises. If noises are mainly due to lecture slides, course notes, and so on that do not have direct citations, we believe that the *CCI* method is rather robust because it takes both direct and indirect influence of research papers into account.

5 Conclusion

Evaluating the influence of research publications is a challenging issue. In this paper, we have proposed a new citation analysis method—*Comprehensive Citation Index*—by incorporating both direct and indirect intellectual influence of research papers into a simple linear model. Importantly, *CCI* overcomes the limitations of *SCI* and *PageRank* in citation analysis that *SCI* neglects the indirect influence of papers and that *PageRank* does not count the number of direct citations.

When peer review is not feasible for assessing a large number of papers, data-driven citation analysis methods seem to be the best alternative. Among such methods, *CCI* rankings are closer to peer review results than *SCI* and *PageRank* rankings. Because research is a long process and research papers' direct and indirect intellectual influence on other papers is gradually released during knowledge accumulation, *CCI* is more reliable than *SCI* and *PageRank* in discovering papers that have far-reaching influence over years and decades. In the future, we will apply the *CCI* method to find significant research papers in different research areas.

Acknowledgments

The authors sincerely thank the editors and reviewers for their valuable comments that have greatly contributed to improving this paper.

References

[1] R.K. Merton, "Matthew Effect in Science," *Science,* vol. 159, no. 3810, pp. 56-63, 1968.
[2] R.K. Merton, "The Matthew Effect in Science II: Cumulative Advantage and the Symbolism of Intellectual Property," *Isis,* vol. 79, no. 299, pp. 606-623, 1988.
[3] E. Garfield, "The History and Meaning of the Journal Impact Factor," *J. Am. Medical Assoc.,* vol. 295, no. 1, pp. 90-93, 2006.
[4] E. Garfield, "Citation Indexing for Studying Science," *Nature,* vol. 227, no. 5259, pp. 669-671, 1970.
[5] E. Garfield, "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, pp. 471-479, 1972.
[6] S.M. Lawani, "Citation Analysis and Quality of Scientific Productivity," *Bioscience,* vol. 27, no. 1, pp. 26-31, 1977.
[7] D.J.D. Price, "Networks of Scientific Papers," *Science,* vol. 149, no. 3683, pp. 510-515, 1965.
[8] A.N. Langville and C.D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings.* Princeton Univ. Press, 2006.
[9] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks,* vol. 30, nos. 1-7, pp. 107-117, 1998.
[10] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," technical report, Stanford Digital Library Technologies Project, 1998.
[11] C.L. Borgman and J. Furner, "Scholarly Communication and Bibliometrics," *Ann. Rev. Information Science and Technology,* vol. 36, pp. 3-72, 2002.
[12] M. Thelwall, "Interpreting Social Science Link Analysis Research: A Theoretical Framework," *J. Am. Soc. for Information Science and Technology,* vol. 57, no. 1, pp. 60-68, 2006.
[13] J. Bollen, M.A. Rodriguez, and H.V. de Sompel, "Journal Status," *Scientometrics,* vol. 69, no. 3, pp. 669-687, 2006.
[14] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications.* Cambridge Univ. Press, 1994.
[15] J.E. Hirsch, "An Index to Quantify an Individual's Scientific Research Output," *Proc. Nat'l Academy of Sciences USA,* vol. 102, no. 46, pp. 16569-16572, 2005.
[16] L. Egghe, "Theory and Practice of the G-Index," *Scientometrics,* vol. 69, no. 1, pp. 131-152, 2006.
[17] P. Ball, "Achievement Index Climbs the Ranks," *Nature,* vol. 448, no. 7155, p. 737, 2007.
[18] W.J. Hopp, "Ten Most Influential Papers of Management Science's First Fifty Years," *Management Science,* vol. 50, no. 12, pp. 1763-1764, 2004.