Classification of Paper Values Based on Citation Rank and PageRank

Wataru Souma^{1†}, Irena Vodenska^{2,3}, Lou Chitkushev²

¹College of Science and Technology, Nihon University, Funabashi, 274-8501, Japan

²Metropolitan College, Boston University, Boston, MA 02215, USA

³Center for Polymer Studies, Boston University, Boston, MA 02215, USA

Citation: Souma, Wataru, Irena Vodenska, Lou Chitkushev. "Classification of paper values based on citation rank and pagerank." *Journal of Data and Information Science*, vol. 5, no. 3, 2020, pp. 57–70. <https://doi.org/10.2478/jdis-2020-0031>

Received: Jan. 31, 2020

Revised: May 14, 2020

Accepted: Jun. 11, 2020

Abstract

Purpose: The number of citations has been widely used to measure the significance of a paper. However, there is a need in introducing another index to determine superiority or inferiority of papers with the same number of citations. We determine superiority or inferiority of papers by using the ranking based on the number of citations and PageRank.

Design/methodology/approach: We show the positive linear correlation between Citation Rank (the ranking of the number of citation) and PageRank. On this basis, we identify high-quality, prestige, emerging, and popular papers.

Findings: We found that the high-quality papers belong to the subjects of biochemistry and molecular biology, chemistry, and multidisciplinary sciences. The prestige papers correspond to the subjects of computer science, engineering, and information science. The emerging papers are related to biochemistry and molecular biology, as well as those published in the journal "Cell." The popular papers belong to the subject of multidisciplinary sciences.

Research limitations: We analyze the Science Citation Index Expanded (SCIE) from 1981 to 2015 to calculate Citation Rank and PageRank within a citation network consisting of 34,666,719 papers and 591,321,826 citations.

Practical implications: Our method is applicable to forecast emerging fields of research subjects in science and helps policymakers to consider science policy.

Originality/value: We calculated PageRank for a giant citation network which is extremely larger than the citation networks investigated by previous researchers.

Keywords Number of citation; PageRank; High-quality papers; Prestige papers; Emerging papers; Popular papers

[†] Corresponding author: Wataru Souma (Email: wataru.soma@gmail.com).

1 Introduction

The number of citations is considered as the most frequently used measure to evaluate the significance of papers. However, the following question has been arisen: which paper is the most important among those with the equal number of citations? Several additional measures have been introduced to address this question, one of them is PageRank proposed by Brin and Page (1999).

Then, Bollen, Rodriguez, and Van de Sompel (2006) described the Institute for Scientific Information impact factor (IF), which was defined as the mean number of citations that a journal received over two years and intended as a metric of popularity, while Google PageRank was developed as a metric of prestige. Chen et al. (2007) calculated the number of citations and the Google PageRank number for all papers in the Physical Review family of journals published in the period from 1893 to 2003. They observed a linear relationship between the number of citations and the Google PageRank number. Additionally, they discovered that several outliers in this linear relationship corresponded to the papers ranked as outstanding according to Google PageRank but with the modest number of citations and were universally familiar to physicists due to their considerable scientific impact. Therefore, they denoted these papers as scientific “gems” and concluded that this index could be used successfully as a measure of scientific quality. These scientific “gems” were also investigated by Maslov and Redner (2008). Ma, Guan, and Zhao (2008) confirmed the applicability of this structure to the citation networks of biochemistry and molecular biology.

These previous studies have investigated the citation networks corresponding to the selected scientific fields; however, no study has been conducted with regard to applying the concept of PageRank to all papers in all scientific fields. Therefore, the aim of the present study is to identify the prestige papers (Souma & Jibu, 2018) in all fields of science. Additionally, by employing the number of citations and the Google PageRank number of each paper published in each journal, we calculated the mean values of the number of citations and the Google PageRank number for each journal and proposed a new measure of journal influence (Souma, Vodenska, & Chitkushev, 2019a; 2019b).

The remainder of this paper is organized as follows. In Section 2, we describe the data used in the present study and calculate the Citation Rank and PageRank indices for each paper. We also confirm the presence of the linear correlation between Citation Rank and PageRank. Subsequently, by considering the observed linear correlation, we identify the high-quality, prestige, emerging, and popular papers. The last section is devoted to the summary and discussion of results.

2 Data, Citation Rank, and PageRank

In the present study, we employ the Science Citation Index Expanded (SCIE) provided by Clarivate Analytics Co., Ltd, US. We utilize the SCIE data for the period from 1981 to 2015. This dataset contains 34,666,719 papers and 591,321,826 citations.

By considering papers as nodes and citations from a citing paper to a cited paper as directed links, we can represent the dataset of citations as a directed network. We denote this network as the citation network, which consists of numerous connected components. The giant weakly connected component (GWCC) comprises 34,428,322 nodes, which contribute to 99.3% of the total number of papers mentioned in the dataset, and 591,177,607 directed links, which constitute 99.98% of the total number of citations represented in the dataset. We focus on GWCC as described below.

Brin and Page (1999) proposed the so-called PageRank to obtain the appropriate ranking of a web page in the World Wide Web (WWW). PageRank of paper i is derived from the Google PageRank number, g_i , defined according to the following recursion formula (Chen et al., 2007):

$$g_i = (1 - d) \sum_{j \in \tilde{k}_i} \frac{g_j}{\tilde{k}_j} + \frac{d}{N}. \quad (1)$$

Here, $N = 34,428,322$ denotes the total number of papers contained in GWCC, and $\tilde{k}_j$ is the total number of citing papers of node j . The sum is taken over the neighboring nodes j , which are the link points to node i . In Equation (1), d denotes a free parameter that controls the convergence and effectiveness of the recursive calculation. In the case of citation networks, the direction of links is usually oriented toward the past. Therefore, if we consider only the first term of Equation (1), the Google PageRank numbers are accumulated in old papers. The second term of Equation (1) is included to prevent this accumulation effect.

In the original calculation of PageRank, $d = 0.15$ was adopted for the case of WWW (Brin & Page, 1999). Then, $d = 0.5$ was adopted in the case of the citation network (Chen et al., 2007). Following Chen et al. (2007), we set $d = 0.5$ in this study. As shown by Souma and Jibu (2018), although the distribution of PageRank depends on d , the PageRank values of all considered papers are close to each other in the case of $d = 0.15$ and $d = 0.5$.

In the left panel of Figure 1, k_i represents the number of citations of paper i , and g_i represents the Google PageRank number of paper i . This figure represents a double-logarithmic scale scatter plot of k_i and g_i . Here, a black dot represents one paper. The gray solid line represents the average value $\langle g \rangle$, which is calculated for bins of the logarithmically equal width against k . This figure shows that the graph

Research Paper

of $\langle g \rangle$ versus k indicates a smooth and positive linear correlation in the high k range. Therefore, we conclude that there is a linear correlation between k_i and g_i in the high k range.

We define the CitationRank of paper i as the ranking of the number of citations and denote it as $r_{k,i}$. The PageRank of paper i is the ranking of the Google PageRank number and is denoted by $r_{g,i}$. By using $r_{k,i}$ and $r_{g,i}$, we can obtain the right panel of Figure 1. In this figure, the gray solid line represents $r_g = r_k$. Similarly, as in the case of the left panel, the right panel of Figure 1 also shows the presence of the linear correlation between $r_{k,i}$ and $r_{g,i}$. Furthermore, by analyzing this figure, we can determine superiority or inferiority of papers with the same number of citations in terms of quality. Namely, a paper with the high PageRank value is considered as superior with respect to that with low PageRank, even if the papers have the same ranking value of r_k .

Figure 1. Left: the scatter plot of the number of citations, k , and the Google PageRank number, g , for each paper (black dots). The gray solid line represents the average value, $\langle g \rangle$, which is calculated for bins of the logarithmically equal width against k . Right: the scatter plot of CitationRank (the ranking of the number of citations), r_k , and PageRank, r_g , for each paper (black dots). The gray solid line represents the standard line $r_g = r_k$.

3 Classification of the values of papers

The relation $r_g = r_k$ is the standard equation used to determine superiority or inferiority of papers. On its basis, we define the papers corresponding to the following categories: high-quality, prestige, emerging, and popular papers.

3.1 High-quality papers

We consider that high-quality papers are characterized by high CitationRank and high PageRank, and therefore, we define the ranking of a high-quality paper according to the average value of $r_{k,i}$ and $r_{g,i}$ as follows:

$$r_i = \frac{1}{2}(r_{k,i} + r_{g,i}). \quad (2)$$

The list of the identified top 10 high-quality papers is presented below:

1. Piotr Chomczynski and Nicoletta Sacchi. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical biochemistry*, 162(1): 156–159, 1987.
2. George M Sheldrick. A short history of SHELX. *Acta Crystallographica Section A: Foundations of Crystallography*, 64(1):112–122, 2008.
3. Axel D. Becke. Density functional thermochemistry. iii. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648– 5652, 1993.
4. Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37:785–789, Jan 1988.
5. John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.
6. Julie D Thompson, Desmond G Higgins, and Toby J Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. *Nucleic acids research*, 22(22): 4673–4680, 1994.
7. J Martin Bland and Douglas G Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.
8. Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.
9. Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.
10. Zbyszek Otwinowski and Wladek Minor. Processing of X-ray diffraction data collected in oscillation mode. In *Methods in enzymology*, 276, 307–326. Elsevier, 1997.

From this list, it can be seen that the selected papers belong to the subjects of biochemistry and molecular biology, chemistry, and multidisciplinary sciences.

The high-quality papers are also extracted by using the constraint defined as follows:

$$r_g \leq -r_k + a, \quad (3)$$

where a is a free parameter. Equation (2) is orthogonal to the standard $r_g = r_k$. Although we can apply Equation (2) to the whole range of $r_{k,i}$, we consider the range $r_k \leq 10^5$. This is because the papers with low CitationRank do not correspond to the high-quality papers. Figure 2 shows the top 10 subjects related to the high-quality papers extracted by varying the parameter $a = (1, 2, \dots, 10) \times 10^4$. In this figure, it can be seen that the ratio of these 10 subjects is close to be stable among different values of a .

Figure 2. Top 10 subjects of the high-quality papers.

Figure 3 represents the correlation between CitationRank $r_{k,i}$ and PageRank $r_{g,i}$ for the top four subjects in the case of $a = 10^4$. These figures show that the papers are certainly distributed in the high CitationRank and the high PageRank ranges. However, in these ranges, many papers are distributed over the standard $r_g = r_k$. This fact indicates that many papers belonging to the subjects of multidisciplinary sciences, medicine, chemistry, and biochemistry and molecular biology have low PageRank, even if CitationRank is high. Therefore, in these cases, proportionality of citation, and value is less exhibited.

3.2 Prestige papers

We consider that papers distributed under the standard $r_g = r_k$ can be classified as the prestige papers. The farther away we move from the standard below, the higher is the prestige of a dissertation. To identify high-prestige papers, we introduce the ratio of CitationRank $r_{k,i}$ and PageRank $r_{g,i}$:

$$y_i = \frac{r_{k,i}}{r_{g,i}}, \quad (4)$$

and then, we define the conditional PageRank given as follows:

$$r_{g,i}(x) = r_{g,i}(y_i \geq x). \quad (5)$$

Figure 3. Top four subjects of the popular papers.

Here, x represents the distance from the standard $r_g = r_k$. Similarly, as in the case of high-quality papers, we consider the range $r_k \leq 10^5$.

Figure 4 represents the top 10 subjects of the prestige papers against x . In this figure, it can be seen that the ratio of each subject depends on x ; however, the ranking of these subjects is stable. Figure 4 shows that with an increase in x , the ratio of the subjects of computer science and engineering increases as well.

Figure 5 represents the distribution of the CitationRank and PageRank values corresponding to the subjects of computer science and engineering in the case of $x = 10$. Compared to the case of the high-quality papers, the prestige ones are distributed in the range below the standard. This means that the prestige papers have high ranking in terms of PageRank, even if CitationRank is low.

Research Paper

Figure 4. Components of the prestige papers.

Figure 5. Top two subjects of the prestige papers.

The list of the top 10 prestige papers selected when $x = 10$ is presented below:

1. J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95 – International Conference on Neural Networks*, 4, 1942–1948,1995.
2. S. M. Alamouti. A simple transmit diversity technique for wireless communications. *IEEE Journal on Selected Areas in Communications*, 16(8):1451–1458, 1998.

3. I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. Wireless sensor networks: a survey. *Computer Networks*, 38(4):393 – 422, 2002.
4. Zdzislaw Pawlak. Rough sets. *International Journal of Computer & Information Sciences*, 11(5):341–356, 1982.
5. I. F. Akyildiz, Weilian Su, Y. Sankarasubramaniam, and E. Cayirci. A survey on sensor networks. *IEEE Communications Magazine*, 40(8):102–114, 2002.
6. Thomas R Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.
7. Piyush Gupta and Panganmala R Kumar. The capacity of wireless networks. *IEEE Transactions on information theory*, 46(2):388– 404, 2000.
8. Sally Floyd and Van Jacobson. Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on networking*, (4):397–413, 1993.
9. Giuseppe Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on selected areas in communications*, 18(3):535–547, 2000.
10. Simon Haykin. Cognitive radio: brain-empowered wireless communications. *IEEE journal on selected areas in communications*, 23(2):201–220, 2005.

From this list, it can be seen that papers belong to the subjects of computer science, engineering, and information science.

3.3 Emerging and popular papers

Preparatory to defining the concepts of emerging and popular papers, we investigate the dependence between the CitationRank and PageRank and the year of publication. Figure 6 represents the changes in CitationRank and PageRank from 2015 to 1981. The papers published in 2015 are distributed in the range of low CitationRank and low PageRank. However, the distribution moves to the direction of high CitationRank in the range above the standard line, i.e., in the range $r_g \leq r_k$. After 2000, the distribution moves toward the direction of high PageRank, and almost all papers are distributed below the standard in 1981. Therefore, we can conclude that the CitationRank increase first and Pagerank increase after that.

To confirm the conclusion derived from the results presented in Figure 6 we calculate the average values of CitationRank, $\langle r_k \rangle_t$ and that of PageRank, $\langle r_g \rangle_t$ for each published year t . Figure 7 represents the changes in the indices through the considered time period. As expected from the results presented in Figure 6, these average values move in the range above the standard up to 2000. Then, after 2000, they move to the range under the standard towered over the high PageRank. Although the average values move as described here, many papers remain in the range of high CitationRank and low PageRank. We define the emerging paper as the paper with a high growth rate of the number of citations, high Citationrank, and low PageRank. On the other hand, we define popular paper as the paper with a low growth rate of the number of citations, high Citationrank, and low PageRank. Therefore, we can consider that the papers distributed above the standard line are the mix of emerging and popular papers.

Figure 6. Dependence between CitationRank and PageRank in the published year t . Although we omit the label and scale of abscissa and ordinate in these figures, they are the same as in the right panel of Figure 1.

Figure 7. Changes in the average value of CitationRank $\langle r_k \rangle_t$ and that of PageRank $\langle r_g \rangle_t$. The dotted line is the standard $r_g = r_k$.

We consider that the papers distributed over the standard $r_g = r_k$ are the mix of emerging and popular papers. The farther away we move from the standard to over, the higher is the emergence of the dissertation. To identify the high-prestige papers, we introduce the ratio of PageRank $r_{g,i}$ and CitationRank, $r_{k,i}$ defined as follows:

$$\tilde{y}_i = \frac{r_{g,i}}{r_{k,i}}, \quad (6)$$

and define the conditional CitationRank given by:

$$r_{k,i}(x) = r_{g,i}(\tilde{y}_i \geq x). \quad (7)$$

Here, x represents the distance from the standard $r_g = r_k$. Similarly, as in the case of the high-quality and prestige papers, we consider the range $r_k \leq 10^5$.

Figure 7 represents the top 10 subjects corresponding to the prestige papers against x . We selected them at $x = 1$. From this figure, it can be seen that the ratio of each subject strongly depends on x . Figure 7 shows that the ratio of biochemistry and molecular biology, multidisciplinary science, and chemistry increase, as x increase.

Research Paper

Figure 8. Components of the emerging and popular papers.

The list of the emerging and popular papers selected when $x = 5.5$ is presented below:

1. Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.
2. Brad T Sherman, Richard A Lempicki, et al. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44–57, 2009.
3. Yan Zhao and Donald G Truhlar. The m06 suite of density functional for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four m06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts*, 120(1-3):215–241, 2008.
4. David P Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2): 215–233, 2009.
5. Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, 2005.
6. Thomas Jenuwein and C David Allis. Translating the histone code. *Science*, 293(5532): 1074–1080, 2001.
7. Peng Li, Deepak Nijhawan, Imawati Budihardjo, Srinivasa M Srinivasula, Manzoor Ahmad, Emad S Alnemri, and Xiaodong Wang. Cytochrome c and dATP-dependent formation of apaf-1/caspase-9 complex initiates an apoptotic protease cascade. *Cell*, 91(4): 479–489, 1997.
8. Zhengui Xia, Martin Dickens, Jel Raingeaud, Roger J Davis, and Michael E Greenberg. Opposing effects of ERK and JNK-p38 map kinases on apoptosis. *Science*, 270(5240): 1326–1331, 1995.

9. Rosalind C Lee, Rhonda L Feinbaum, and Victor Ambros. The *c. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5): 843–854, 1993.
10. Alan Hall. Rho GTPases and the actin cytoskeleton. *Science*, 279(5350):509–514, 1998.

These papers belong to the subject of biochemistry and molecular biology, chemistry, and multidisciplinary science. Moreover, the five papers belonging to biochemistry and molecular biology were published in the journal “Cell” and top three papers among them were published after 2005. However, the four papers belonging to multidisciplinary science were published in the journal “Science” before 2001. Therefore, we can consider that the former three papers are emerging papers, and the latter four papers correspond to the popular ones.

4 Conclusion

In the present study, we calculated CitationRank and PageRank based on the SCIE data for the period of 35 years (from 1981 to 2015) and identified the high-quality, prestige, emerging, and popular papers. We found that the high-quality papers belong to the subjects of biochemistry and molecular biology, chemistry, and multidisciplinary sciences. The prestige papers correspond to the subjects of computer science, engineering, and information science. The emerging papers are related to biochemistry and molecular biology, as well as those published in the journal “Cell.” The popular papers belong to the subject of multidisciplinary sciences.

However, we may have simply identified the dependencies between the subjects and the citation patterns. Therefore, we also calculated CitationRank and PageRank for each subject and have classified the value of papers. In addition, as suggested by Mariani, Medo, and Zhang (2015) and Mariani, Matúš, and Zhang (2016), we focused our attention on applying PageRank to the growing network. Therefore, we applied the new PageRank-based algorithm proposed by them to obtain a more concrete classification of the value of papers.

Although we considered extremely prestige papers, if we had chosen interdisciplinarity as the most important factor, we would have been able to calculate the betweenness centrality and investigate the correlation with CitationRank and PageRank. For the future research, it may be also useful to define indices by integrating the CitationRank, the PageRank, and the BetCentRank (the ranking of betweenness centrality).

Acknowledgments

This work is supported by Nihon University College of Science and Technology Grants-in Aid 2012 and 2016.

Research Paper**Author contributions**

Proposing the research problems: Wataru Souma (wataru.soma@gmail.com); funding acquisition: Wataru Souma, designing the research framework: Irena Vodenska (vodenska@bu.edu) and Lou Chitkushev (ltc@bu.edu); writing and revising the manuscript: Wataru Souma, Irena Vodenska, and Lou Chitkushev.

References

- Bollen, J., Rodriquer, M.A., & Van de Sompel, H. (2006). Journal status. *Scientometrics*, 69, 669–687.
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30, 107–117.
- Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, 1(1), 8–15.
- Ma, N., Guan, J., & Zhao, Y. (2008). Bringing PageRank to the citation analysis. *Information Processing & Management*, 44(2), 800–810.
- Maslov, S., & Redner, S. (2008). Promise and pitfalls of extending Google's pagerank algorithm to citation networks. *Society for Neuroscience*, 28(44), 11103–11105.
- Souma, W., & Jibu, M. (2018). Progress of studies of citations and PageRank. In *Scientometrics* (pp. 213–231). IntechOpen.
- Souma, W., Vodenska, I., & Chitkushev, L. (2019a). New measures of journal impact based on citation network. In *Proceedings of the 17th International Conference on Scientometrics and Informetrics, (ISSI) 2019, Rome, Italy, September 2–5, 2019*, 2524–2525.
- Souma, W., Vodenska, I., & Chitkushev, L. (2019b). New measures of journal impact based on the number of citations and PageRank. In *Science & Technology Metrics Proceedings*, 112–120.
- Mariani, M., Medo, M., & Zhang, Y. (2015). Ranking nodes in growing networks: When PageRank fails. *Scientific Reports*, 5, 16181.
- Mariani, M., Matúš, M., & Zhang, Y. (2016). Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics*, 10(4), 1207–1233.

This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).