


Partitioning highly, medium and lowly cited publications

Journal of Information Science
1–6
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0165551520917655
journals.sagepub.com/home/jis


Yong Huang

Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, China

Yi Bu

Department of Information Management, Peking University, Beijing, China; Center for Complex Networks and Systems Research, Luddy School of Informatics, Computing, and Engineering, Indiana University, USA

Ying Ding

School of Information, University of Texas, USA; Dell Medical School, University of Texas, USA

Wei Lu

Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, China

Abstract

Dividing papers based on their numbers of citations into several groups constitutes one of the most common research practices in bibliometrics and beyond. However, existing dividing methods are both arbitrary and subject to bias. This article proposes a novel approach to partition highly, medium and lowly cited publications based on their citation distribution. We utilise the whole Web of Science (WoS) dataset to demonstrate how to apply this approach to scholarly datasets and examine the robustness of our algorithm in each of the six disciplines under the WoS dataset. The codes that underlie the algorithm are available online.

Keywords

Bibliometrics; citation distribution; informetrics; scientometrics

1. Introduction

Citations have long been viewed as an important indicator of publications' impact [1]. Studying highly cited publications has become a tradition in bibliometrics, and most articles in bibliometrics have to deal with the choice of partitioning publications into different categories (e.g. highly, medium or lowly cited) [2 – 5]. However, the majority of these studies have chosen these different categories by establishing artificial thresholds. The employed approaches mainly include

1. Dividing all papers into three groups so that the total number of citations received by publications in each group remains the same [6];
2. Manually setting percentages of citation count ranking to divide papers into groups [7,8]. For example, this may constitute placing all publications in descending order according to their numbers of citations, and arbitrarily setting the first 1% as highly cited publications, 1%–10% as medium cited publications and the remainder as lowly cited publications;

Corresponding author:

Wei Lu, Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, Wuhan 430072, Hubei, China.
Email: weilu@whu.edu.cn

- Setting thresholds to divide publications into groups based on the authors' empirical experience. For instance, Aversa [9] arbitrarily set 10 and 30 as the minimum numbers of citations for highly cited articles. Aksnes [10] also defined highly cited publications as those whose citation count is 17 times that of the average citation count of all publications in a given field. Wang et al. [11] used 40 and 275 as the thresholds for determining lowly, medium and highly cited publications, while Wadhwa et al. [12] utilised 5 and 20. Bu et al. [13] set 100 as the thresholds between highly cited and non-highly cited publications.

However, method 1 lacks theoretical supports on why the same total number of citations in each group makes sense. Methods 2 and 3 are subjective because the percentages or thresholds are determined arbitrarily based on the researchers' empirical experience without considering the distribution of citation counts or fitting details statistically. In this article, we propose an approach to assist researchers to divide publications into groups based on statistical distributions instead of arbitrary decisions.

2. Dataset

We used the whole Web of Science (WoS) dataset housed by Indiana University Network Science Institute (IUNI) as our empirical dataset. This dataset contains 69,326,157 scientific articles ranging from 1900 to 2018 and 1,397,532,215 citing relationships among these publications.

3. Methodology

3.1. Problem statement

The initial aim of this study is to divide publications into three groups, that is, highly, medium and lowly cited publications. To achieve this, we need to identify two thresholds, x_{\min} and x_{\max} , so that publications whose number of citations is lower than x_{\min} are classified as lowly cited, publications whose number of citations is x_{\max} or more are classified as highly cited and the remainder constitute medium cited publications. Suppose that we have N publications in a given dataset, p_1, p_2, \dots, p_N , and publications p_i ($1 \leq i \leq N$) have C_{p_i} ($C_{p_i} \geq 0$) citations. The assigned category (group) of publications p_i is identified as

$$G(p_i) = \begin{cases} l, & C_{p_i} < x_{\min} \\ m, & x_{\min} \leq C_{p_i} < x_{\max} \\ h, & C_{p_i} \geq x_{\max} \end{cases} \quad (1)$$

where l , m and h represent lowly, medium and highly cited groups, respectively.

3.2. Citation distribution

We plot the citation distribution of the whole WoS dataset in the left sub-figure of Figure 1, in which we can find that the distribution is almost a straight line in a double-logarithmic coordinate system. However, it can be seen from the left sub-figure of Figure 1 that the points indicating publications with a few citations deviate from the line *downwards* (most lowly cited publications), and those indicating publications with a large number of citations deviate from the skewed line *upwards* (most highly cited publications). Only the middle part of the curve (the area between two coloured, dotted and vertical lines) looks straight. Network scientists and physicists (such as Redner [14]) believe that different types of distributions reflect distinct mechanisms of the formation of curves. Inspired by this, in this article, the publications positioned in the middle section (the straight part) are partitioned as medium cited publications (mechanism I); those positioned in the downward section are partitioned as lowly cited publications (mechanism II) and those positioned in the upward section are partitioned as highly cited publications (mechanism III).¹

3.3. Determining x_{\min} and x_{\max}

Let $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ be the points in the citation distribution plot of a certain dataset where (x_1, y_1) is the most top left point, (x_n, y_n) is the most bottom right point and n is the total number of points in the distribution. To determine x_{\min} , the threshold between lowly and medium cited publications, we investigate the change of slope of lines

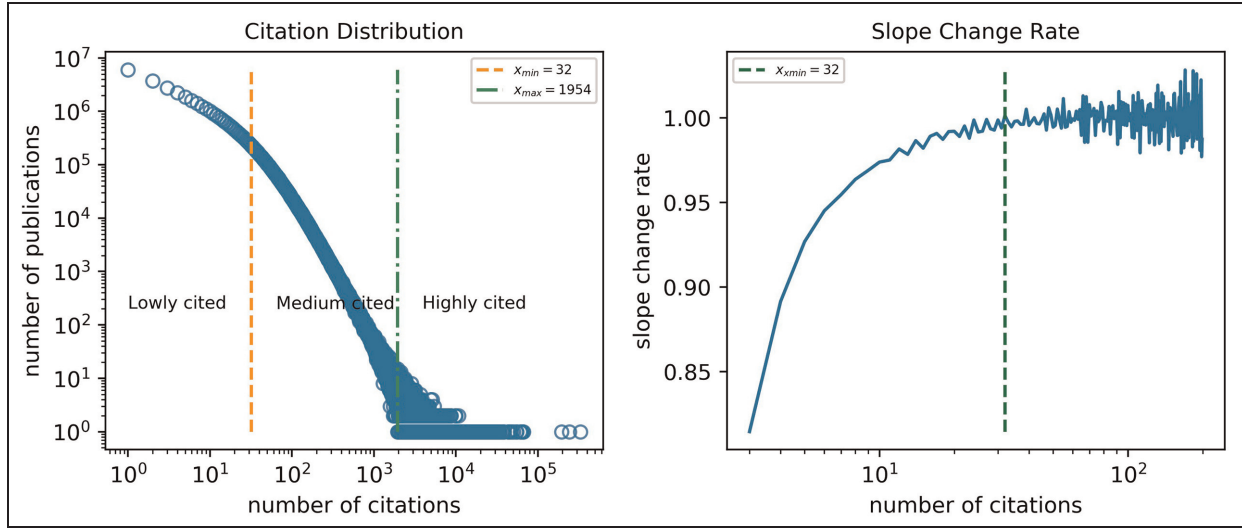


Figure 1. Citation distribution and publication grouping result of the whole WoS dataset (left) and the slope change rate (right). In the left sub-figure, the red and green dotted lines (vertical) represent x_{\min} and x_{\max} , respectively. In the right sub-figure, the slope change rate indicates the change of slope of lines generated by adjacency points in the left sub-figure; see details in formulas (2) and (3).

generated by adjacency points in the right sub-figure of Figure 1 and select the point that has the greatest change of slope. To this end, we annotate the slope of the line connecting the i th and the j th points as $k_{i,j}$ ($1 \leq i, j \leq \max(C_{p_i})$)

$$k_{i,j} = \frac{y_i - y_j}{x_i - x_j} \quad (2)$$

x_{\min} is defined as

$$x_{\min} = \arg_x \min \left(\frac{k_{i-1,i}}{k_{1,i}} = 1 \right) \quad (3)$$

where $k_{i-1,i}/k_{1,i}$ measures the slope change. The fact that $k_{i-1,i}/k_{1,i}$ equals one indicates that the slope of the line connecting two adjacent points does not change. The first point (from left to right) where $(k_{i-1,i}/k_{1,i}) = 1$ corresponds to the point with the greatest value of curvature.

We expect x_{\max} as the turning point where there are many visible points in the distribution that are ‘piled up’ horizontally. To this end, we stipulate x_{\max} as the first point whose vertical axis value equals one (i.e. only one paper in the dataset that has the number of citations corresponding to the horizontal axis value) if one examines points one by one from the left to the right. Mathematically

$$x_{\max} = \arg_x \min(y = 1) \quad (4)$$

4. Results

The left sub-figure of Figure 1 shows the citation distribution and the publication grouping results based on the whole WoS dataset, where one can see that $x_{\max} = 1954$ and $x_{\min} = 12$. To test the robustness of our proposed algorithm, we duplicate it on the six disciplines of WoS according to the labels of WoS publications. Figure 2 shows the distribution of each discipline, namely, ‘Arts & Humanities’, ‘Clinical, Pre-Clinical, & Health’, ‘Engineering & Technology’, ‘Life Sciences’, ‘Physical Sciences’ and ‘Social Sciences’. We observe that, in each sub-figure, citation distributions are similar and the ‘downward’ and ‘upward’ phenomena occur in all sub-figures. The publication grouping results of x_{\min} are 11 or 12 in most disciplines except arts and humanities. In arts and humanities, publications with fewer than five citations are regarded as lowly cited articles; this is quite reasonable, because the number of citations of arts and humanities scientific publications is averagely fewer than that in other disciplines based on Figure 2. In terms of x_{\max} , we find that arts and humanities and social sciences have a lower value of x_{\max} , while other disciplines have values over 1000.

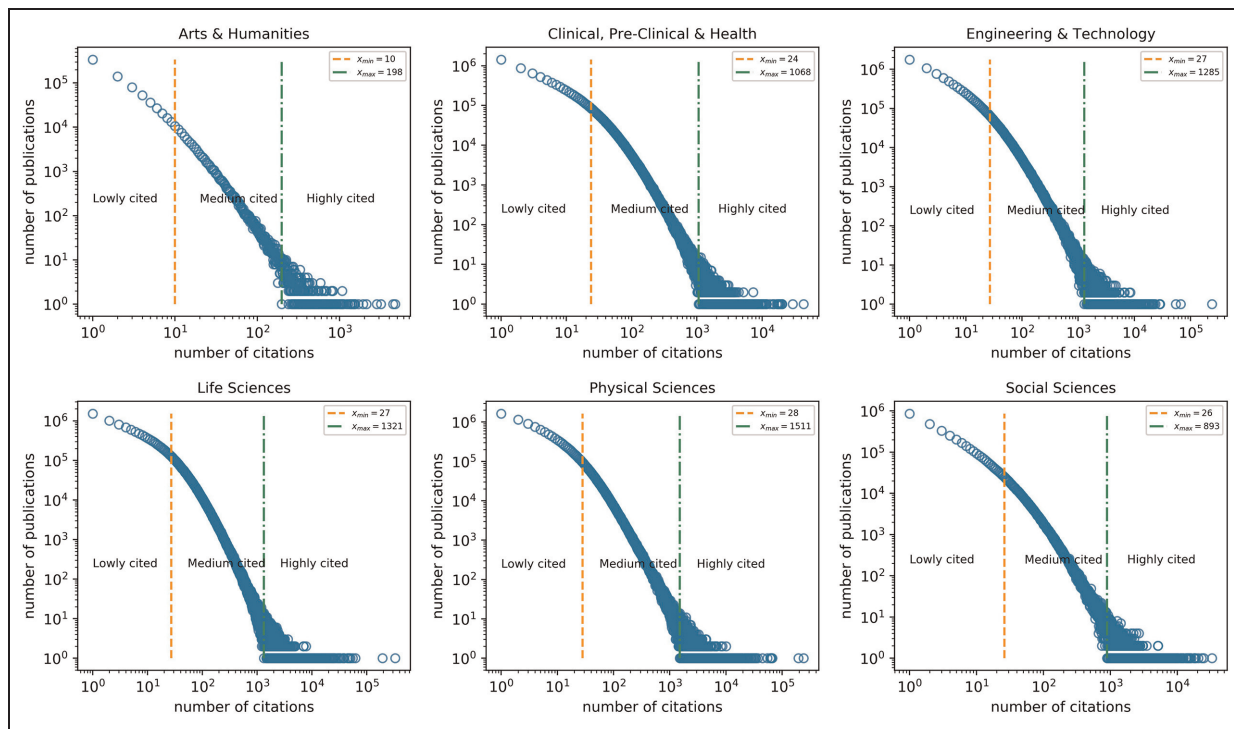


Figure 2. Citation distribution and publication grouping result of each discipline.

Table 1. Comparison among different publication grouping strategies (I: our proposed method, II: the grouping strategy where each group of publications has an equal number of total citations [6], and III: highly cited publications as the top 1%, medium cited publications as 1%–10% and lowly cited as 10%–100% [3]).

Grouping strategy	x_{min}	x_{max}	%Lowly cited publication	%Medium cited publication	%Highly cited publication
I	32	1954	82.51	17.47	0.02
II	36	120	84.64	12.52	2.84
III	51	220	90.00	9.00	1.00

We also compare our proposed method (annotated as strategy I) with two existing methods (strategies II and III). In strategy II, we follow Guo et al. [6] and stipulate that each group of publications has an equal number of total citations. In strategy III, we define highly cited publications as the top 1%, medium cited as 1%–10% and lowly cited as those after 10% [3]. The empirical results of the three strategies are shown in Table 1. We find that, in strategy II, x_{min} equals 36, which is similar to ours (32); in strategy III, x_{min} is 51. As for x_{max} , strategies II and III have 120 and 220, but ours is much greater than theirs (1954). Correspondingly, there are only 0.02% of publications that are categorised as highly cited, which is much smaller than the other two strategies. In our strategy, however, medium cited publications occupy more than 17%; this equals 12.52% and 9% in the two strategies, respectively.

5. Summary

This work proposes a novel approach to partitioning publications with different citation counts based on their citation distributions. The biggest advantage of the proposed method is that we determine the thresholds (x_{min} and x_{max}) purely based on the citation distributions instead of manually.

This article demonstrates how to adopt this approach to publication partitioning. Besides directly duplicating this method in bibliometric research, future studies can also consider expanding this approach. For instance, similar to publications, authors in a given dataset could be divided into three groups based on their total number of citations or h index [16] by utilising the method provided here. Meanwhile, more advanced statistical indicators might be considered for fitting the power-law or log-normal distributions in this process more accurately.

Furthermore, this article uses quite a large dataset to implement the empirical study, but our proposed method may not be applied in a relatively small dataset. This is because sometimes x_{\min} may not exist and x_{\max} can occur in any place based upon relatively small datasets.

Acknowledgements

The authors acknowledge the Indiana University Pervasive Technology Institute for providing KARST, a high-performance computing system in Indiana University, that has contributed to the research results reported within this paper. The authors are also grateful to two anonymous reviewers for their insightful suggestions. The authors thank Matthew Alexander Hutchinson and Xiaoran Yan for setting up empirical environments.

Author contribution

Y.H. and Y.B. equally contribute to this article.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This article is financially supported by the major program of the Social Science Foundation of China (No. 17ZDA292).

ORCID iDs

Yi Bu <https://orcid.org/0000-0003-2549-4580>

Wei Lu <https://orcid.org/0000-0002-0929-7416>

Supplemental material

More details about the dataset, experiments and codes can be found online at <https://github.com/hyyc116/paper-grouping>.

Note

1. Clauset et al. [15] understood power-law distribution empirically. However, they transited the raw distribution of data and, therefore, the new fitted line is straight instead of in a three-phase style like ours. In order to achieve the goal of dividing all publications into three groups based on their citation counts, we do not implement the strategies in Clauset et al. [15], as we have different purposes.

References

- [1] Garfield E and Merton RK. *Citation indexing: its theory and application in science, technology, and humanities* (vol. 8). New York: Wiley, 1979.
- [2] Bormmann L and Daniel H-D. What do citation counts measure? A review of studies on citing behavior. *J Doc* 2008; 64(1): 45–80.
- [3] Lu C, Bu Y, Dong X et al. Analyzing linguistic complexity and scientific impact. *J Informetr* 2019; 13(3): 817–829.
- [4] Huang Y, Bu Y, Ding Y et al. From zero to one: a perspective on citing. *J Assoc Inf Sci Tech* 2019; 70: 1098–1107.
- [5] Huang Y, Bu Y, Ding Y et al. Direct citations between citing publications. arXiv preprint arXiv:1811.01120, 2 November 2018.
- [6] Guo C, Milojević S and Liu X. How are academic articles cited over time? In: *Proceedings of the iConference*, 2015, https://www.ideals.illinois.edu/bitstream/handle/2142/73747/439_ready.pdf?sequence=2
- [7] Bormmann L, de Moya Anegón F and Leydesdorff L. Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis. *PLoS ONE* 2010; 5(10): e13327.
- [8] Glänzel W. Characteristic scores and scales: a bibliometric analysis of subject characteristics based on long-term citation observation. *J Informetr* 2007; 1(1): 92–102.
- [9] Aversa E. Citation patterns of highly cited papers and their relationship to literature aging: a study of the working literature. *Scientometrics* 1985; 7(3–6): 383–389.
- [10] Aksnes DW. Characteristics of highly cited papers. *Res Evaluat* 2003; 12(3): 159–170.
- [11] Wang M, Yu G and Yu D. Mining typical features for highly cited papers. *Scientometrics* 2011; 87(3): 695–706.

-
- [12] Wadhwa NK, Tewari DK, Walke R et al. Bibliometric analysis of NPL papers published during 1981–1985 and 2001–2005: case study. <https://pdfs.semanticscholar.org/c57d/c82f18851aecfbfa51a9dba0f76f7fca1fb6.pdf>
- [13] Bu Y, Waltman L and Huang Y. A multidimensional perspective on the citation impact of scientific publications. arXiv preprint arXiv:1901.09663, 28 January 2019.
- [14] Redner S. How popular is your paper? An empirical study of the citation distribution. *Eur Phys J B* 1998; 4(2): 131–134.
- [15] Clauset A, Shalizi CR and Newman ME. Power-law distributions in empirical data. *SIAM Rev* 2009; 51(4): 661–703.
- [16] Hirsch JE. An index to quantify an individual's scientific research output. *P Natl Acad Sci USA* 2005; 102(46): 16569–16572.