Partitioning Highly, Medium, and Lowly Cited Publications

Yong Huang

Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, Wuhan, Hubei, 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, Bloomington, IN., U.S.A.

Ying Ding

School of Information, University of Texas, Austin, TX, U.S.A.

Dell Medical School, University of Texas, Austin, TX, U.S.A.

Wei Lu (**corresponding author; weilu@whu.edu.cn**)

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

Yong Huang and Yi Bu equally contribute to this article.

Partitioning Highly, Medium, and Lowly Cited Publications

Abstract: Dividing papers based on 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 paper proposes a novel approach to partition highly, medium, and lowly cited publications based on their citation distribution. We utilize 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.

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 (i.e., highly, medium, or lowly cited) (e.g., [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; and
(3) 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, Yu, and Yu [11] used 40 and 275 as the thresholds for determining lowly, medium, and highly cited publications, while Wadhwa *et al.* [12] utilized 5 and 20. Bu, Waltman, and Huang [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 paper, we propose an approach to assist researchers to divide publications into groups based on proper statistical steps instead of arbitrary decisions.

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.

METHODOLOGY

Problem Statement

The initial aim of the present study is to divide publications into three groups, i.e., 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, publications whose number of citations is x_{max} or more are classified as highly cited publications, and the remainder

constitute medium cited publications. Suppose that we have N publications in a given dataset, p_1, p_2, …, p_N, and publications p_i $(1 \leq i \leq N)$ has 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.

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, the left sub-figure of Figure 1 shows 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 colored, 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 the current paper, 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]

[1] Clauset, Shalizi, and Newman [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

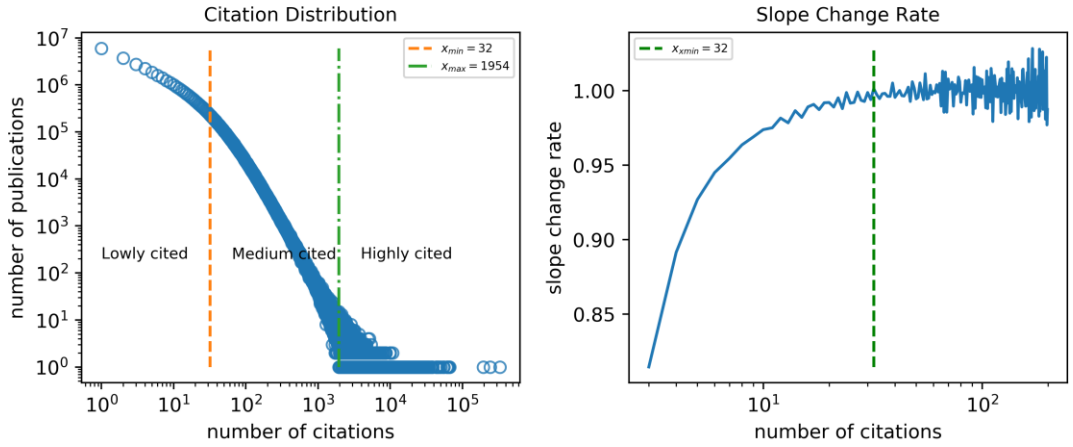

Figure 1. Citation distribution and publication grouping result of the whole WoS dataset (left) and the slope change rate (right). In the left sub-figure, the red and the green dotted lines (vertical) represent x_{min} and x_{max}, respectively. In the right sub-figure, the slope change rate indicates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

Determining x_{min} *and* x_{max}

Let $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ 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 the total number of points in the distribution. To determine $\mathrm{x_{min}}$, the threshold between lowly and medium cited publications, we investigate the change of slope of lines generated by adjacency points in the right sub-figure of Figure 1 and select the point that have the greatest change of slope. To this end, we annotate the slope of the line connecting the ith and the jth points as $k_{i,j}$ $(1 \leq i, j \leq max\ (C_{p_i}))$:

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.

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

x_{min} is defined as:

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

where $\frac{k_{i-1,i}}{k_{1,i}}$ measures the slope change. The fact that $\frac{k_{i-1,i}}{k_{1,i}}$ equals one indicates that the slope of the line connecting two adjacent points does not change. The first point (from left to right) where $\frac{k_{i-1,i}}{k_{1,i}} = 1$ 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 horizon axis value) if one examines points one by one from the left to the right. Mathematically:

$$x_{max} = \arg\min_x (y = 1) \quad (4)$$

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 1,000.

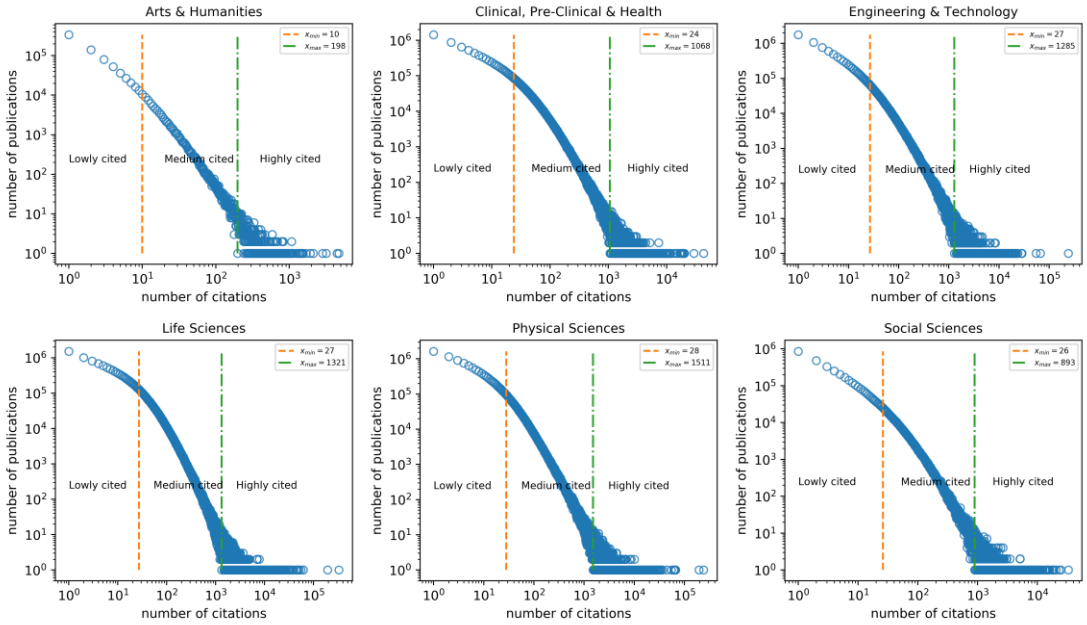

Figure 2. Citation distribution and publication grouping result of each discipline.

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 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 $x_{\max}$, the strategies II and III have 120 and 220, but ours is much greater than theirs (1954). Correspondingly, there are only 0.02% publications

that are categorized 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 to 12.52% and 9% in the two strategies, respectively.

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

Grouping strategy	x_{min}	x_{max}	% of lowly cited publication	% of medium cited publication	% of 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

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.

The current paper 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 utilizing 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, the current paper 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.

SUPPLEMENTARY INFORMATION

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

ACKNOWLEDGEMENTS

This article is financially supported by the major program of the Social Science Foundation of China (No. 17ZDA292) and China Postdoctoral Science Foundation Funded Project (No. 2019M662729). The authors acknowledge the Indiana University Pervasive Technology Institute for providing KARST, a high-performance computing system in Indiana University, that have contributed to the research results reported within this paper. This research was supported in part by the Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute, and in part by the Indiana METACyt Initiative. The Indiana METACyt Initiative at Indiana University was also supported in part by the Lilly Endowment, Inc. The authors thank Matthew Alexander Hutchinson and Xiaoran Yan for setting up empirical environments. The authors are also grateful to two anonymous reviewers for their insightful suggestions.

REFERENCES

[1] Garfield E and Merton RK. (1979). *Citation indexing: Its theory and application in science, technology, and humanities (Vol. 8)*. New York: Wiley.

[2] Bornmann L and Daniel H-D. What do citation counts measure? A review of studies

on citing behavior. *Journal of Documentation 2008*; 64(1): 45-80.

[3] Lu C, Bu Y, Dong, X, Wang J, Ding Y, Larivière, V., …, Zhang C. Analyzing linguistic complexity and scientific impact. Journal of the Informetrics. 2019 August;13(3).

[4] Huang Y, Bu Y, Ding Y, Lu W. From zero to one: A perspective on citing. *Journal of the Association for Information Science and Technology*. 2019.

[5] Huang Y, Bu Y, Ding Y, Lu W. Direct citations between citing publications. arXiv preprint arXiv:1811.01120. 2018 Nov 2.

[6] Guo C, Milojević S and Liu X. How are academic articles cited over time? In *Proceedings of the iConference*, 2015.

[7] Bornmann 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. *Journal of Informetrics 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. (2003). Characteristics of highly cited papers. *Research Evaluation 2003; 12*(3): 159-170.

[11] Wang M, Yu G, Yu D. Mining typical features for highly cited papers. *Scientometrics*. 2011 Jun 1;87(3):695-706.

[12] Wadhwa NK, Tewari DK, Walke R, Yadav AK, Dhawan SM. Bibliometric Analysis of NPL Papers Published during 1981–1985 and 2001–2005: Case Study.

[13] Bu Y, Waltman L, Huang Y. A multidimensional perspective on the citation impact of scientific publications. arXiv preprint arXiv:1901.09663. 2019 Jan 28.

[14] Redner S. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B 1998; 4*(2): 131-134.

[15] Clauset A, Shalizi CR, Newman ME. Power-law distributions in empirical data. SIAM review. 2009 Nov 6;51(4):661-703.

[16] Hirsch JE. An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences. 2005 Nov 15;102(46):16569-72.