Identifying Your Representative Work Based on Credit Allocation

Peng Bao, Jiahui Wang
School of Software Engineering, Beijing Jiaotong University
CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology
{baopeng,17126211}@bjtu.edu.cn

ABSTRACT

With the rapid development of scientific impact quantification in the field of science of success, the ability to identify the representative work of a researcher has important implications in a wide range of areas, including hiring, funding, and promotion systems. In this paper, we propose a two-step credit allocation algorithm (TSCA) for identifying the representative work of a researcher. This algorithm explicitly captures the importance of a paper, its relevance to other papers, and the unequally distributed contribution of each citation. We validate TSCA by applying it on the citation data from American Physical Society (APS) in the scenario of identifying the Nobel prize winning papers of the Nobel laureates. Experiments demonstrate that the proposed algorithm can significantly outperform the existing methods.

CCS CONCEPTS

• **Information systems** → **Content ranking**; • **Human-centered computing** → *Reputation systems*;

KEYWORDS

representative work, credit allocation, citation network

ACM Reference Format:
Peng Bao, Jiahui Wang. 2018. Identifying Your Representative Work Based on Credit Allocation. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3184558.3186901

1 INTRODUCTION

In the last decades, a large number of scientific papers are published online, accelerating the publication of research findings and development of modern science. How to distinguish the scientific impact of a scientist's publications has been an important issue for revealing the quality of academic works [1]. More importantly, identifying the representative work of individual researchers becomes a practical and challenging problem. It is crucial for quantitative assessment of researchers, which affects decisions in a wide range of areas, including hiring, funding, and promotion [2].

The most straightforward and widely adopted metric is the citation count. However, it ignores the fact that the importance of each citation is different, which can not objectively reflect the quality of papers. Accordingly, Google's PageRank algorithm was introduced to address this issue, along with its variations [3, 4]. Despite their initial success in revealing the importance of papers, they are not appropriate for identifying the representative work of individual researcher since the contribution of coauthors in a paper is different [5, 6]. Recently, in order to incorporate both the importance of a paper and its relevance to other papers, Niu et al. [7] have proposed a self-avoiding preferential diffusion process to rank a scientist's papers. However, they still consider the contribution of each citation equally.

In this paper, we propose a two-step credit allocation algorithm (TSCA) for identifying the representative work of a researcher. Based on the scientific credit allocation mechanism, this algorithm explicitly captures the importance of a paper, its relevance to other papers, and the unequally distributed contribution of each citation, distinguishing itself from the method presented in [7]. We validate TSCA by applying it on the citation data from American Physical Society (APS) in the scenario of identifying the Nobel prize winning papers of the Nobel laureates. Experimental results demonstrate that the proposed algorithm can significantly outperform the state-of-the-art methods.

2 THE ALGORITHM

In this paper, we consider the *representative work* of a researcher as an important paper in his/her field of expertise, which is similar with the definition in [7].

Firstly, we formally express the problem of representative work identification. Suppose we have M authors with the i-th author denoted as a_i and N papers with the j-th paper denoted as p_j. We denote the author-paper matrix as $\mathbf{A} \in \mathbb{R}^{M\times N}$, with its (i,j)-th entry $A_{ij} = 1$ if a_i is an author of p_j. Here we denote $\mathbf{a_i}$ as the i-th row of matrix $\mathbf{A}$. The paper-paper matrix is denoted as $\mathbf{P} \in \mathbb{R}^{N\times N}$, with its (i,j)-th entry $P_{ij} = 1$ if p_i cites p_j, and the diagonal entry $P_{ii} = 0$. Consequently, the cocitation matrix can be computed as $\mathbf{C} = \mathbf{P}\mathbf{P}^\top$, where $\mathbf{C} \in \mathbb{R}^{N\times N}$ and its (i,j)-th entry C_{ij} denotes the number of papers co-cited by p_i and p_j. Note that we use d_i^{in} and d_i^{out} to denote the number of citations and references of p_i, respectively.

In the proposed algorithm TSCA, the scores of target author a_t's papers can be expressed as

$$\mathbf{r_t} = \mathbf{a_t}\mathbf{W}, \tag{1}$$

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186901

ALGORITHM 1: Identify representative work

```
Input: matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, matrix $\mathbf{P} \in \mathbb{R}^{N \times N}$
Output: scores of $a_t$'s papers
Initialize matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$;
Compute cocitation matrix $\mathbf{C} = \mathbf{P}\mathbf{P}^{\top}$;
for each paper $p_i$ do
    Step 1: Compute $U_{ik} = \frac{P_{ki}}{d_i^{in}}$;
    Compute cocitation strength $s_{ij} = C_{ij} / \sqrt{d_i^{out} d_j^{out}}$;
    Step 2: Compute $W_{il}^1$, $W_{il}^2$ according to Eq. (2), (3);
    Set $W_{ii}^2 = 0$ and Compute $W_{il}$ according to Eq. (4);
end
$\mathbf{r_t} = \mathbf{a_t}\mathbf{W}$, where $\mathbf{a_t}$ is the $t$-th row of $\mathbf{A}$;
return $\mathbf{r_t}$;
```

where the matrix $\mathbf{W}$ is the credit matrix with its (i, j)-th entry W_{ij} representing the credit value that p_j receives from a two-step credit allocation starting from p_i. The pseudocode is listed in Algorithm 1. The allocation starts with initial credit 1 on paper p_i. In the first step, the initial credit on p_i is evenly distributed to the papers which cite it. Therefore, the credit that p_k receives from p_i can be computed as $U_{ik} = P_{ki}/d_i^{in}$. In the second step, the credit on paper p_k will be unequally allocated back to the papers cited by it based on the cocitation strength, distinguishing from the evenly redistribution method in [7]. For simplicity, we define the cocitation strength s_{ij} between p_i and p_j as $s_{ij} = C_{ij}/\sqrt{d_i^{out} d_j^{out}}$, which explicitly captures the relevance between two paper and guide us to locate the author's field of expertise.

Given a paper p_l which is cited by p_k, the credit that p_l receives in the second step is

$$W_{il}^1 = \sum_{k=1}^{N} s_{il} \frac{P_{kl} U_{ik}}{d_k^{out}}. \tag{2}$$

Due to the problem of data sparsity, we further consider the credit assigned from a paper $p_{l'}$ which is cited by p_k but not written by a_t. This credit is denoted by W_{il}^2 and computed as

$$W_{il}^2 = \sum_{k=1}^{N} \sum_{l'=1}^{N} s_{il} \frac{P_{l'l} P_{kl'} U_{ik}}{d_k^{out} d_{l'}^{out}}. \tag{3}$$

Note that we set $W_{ii}^2 = 0$ to ensure that the credit starting from one paper cannot return to itself. In addition, by incorporating the well-known preferential attachment mechanism [1], W_{il} can be finally expressed by

$$W_{il} = d_l^{in} (W_{il}^1 + W_{il}^2). \tag{4}$$

3 EXPERIMENTS

The dataset used in this paper comprises the papers published in all the journals in APS from 1893 to 2009, consisting of 245,365 authors, 463,344 papers, and 4,692,026 citations. Firstly, we identify the Nobel laureates whose articles published in APS are awarded the Nobel prize. For each laureate, the Nobel winning paper is widely accepted as his/her representative work. In total, we collect 37 laureates along with their Nobel prize winning papers, which will be used as our ground truth in the following experiments. As the Nobel prize will significantly attract the attention to the prize winning paper, a more meaningful and much harder problem is to identify the Nobel prize winning paper among the Nobel laureate's papers before he/she receives the prize. Therefore, we focus on ranking each laureate's papers one year before the Nobel prize is received.

Table 1: Prediction performance of four methods.

Methods	Citation	PageRank	SPD	TSCA
mean rank	3.97	3.77	3.67	**3.10**

In order to quantitatively examine the effectiveness of TSCA, we compute the *mean rank* of these Nobel prize winning papers in each Nobel laureate's personal ranking list. The smaller the mean rank is, the better the ranking method performs. Three existing methods are selected as baselines: citation count, PageRank [3], and SPD [7].

As shown in Table 1, we find that TSCA exhibits the smallest mean rank among all the methods, with an improvement of 21.9%, 17.8%, and 15.5% compared with citation count, PageRank, and SPD, respectively.

4 CONCLUSIONS

In this paper, we propose a two-step credit allocation algorithm for representative work identification. We apply it in the scenario of identifying the Nobel prize winning papers of the Nobel laureates on a APS dataset. Experimental results demonstrate that this algorithm can significantly outperform the existing methods such as citation count, PageRank, and SPD. In the future, time effect will be further investigated.

ACKNOWLEDGMENTS

The work was supported by the National Natural Science Foundation of China under grant number 61702031, Beijing Excellent Talents Supporting Program under grant number 2017000020124G054, and the Fundamental Research Funds for the Central Universities under grant number 2015RC031. Peng Bao is also funded by CAS Key Lab of Network Data Science and Technology. The authors would also like to thank the anonymous referees for their valuable comments.

REFERENCES

[1] D. Wang, C. Song, and A.-L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154), 127C132, 2013.

[2] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354(6312), 596, 2016.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(98), 107-117, 1998.

[4] J. Zhou, A. Zeng, Y. Fan, and Z. Di. Ranking scientific publications with similarity preferential mechanism. *Scientometrics*, 106, 805-816, 2016.

[5] H. W. Shen and A.-L. Barabási. Collective credit allocation in science. *Proc. Natl. Acad. Sci.*, 111(34), 12325C12330, 2014.

[6] P. Bao and C. Zhai. Dynamic credit allocation in scientific literature. *Scientometrics*, 112, 595-606, 2017.

[7] Q. Niu, J. Zhou, A. Zeng, Y. Fan, and Z. Di. Which publication is your representative work? *Journal of Informetrics*, 10(8): 842-853, 2016.