Towards Expert Preference on Academic Article Recommendation Using Bibliometric Networks

Conference Paper · May 2020

CITATIONS
2

READS
144

4 authors:

Yu Zhang
UNSW Canberra
23 PUBLICATIONS **110** CITATIONS
SEE PROFILE

Min Wang
UNSW Sydney
34 PUBLICATIONS **641** CITATIONS
SEE PROFILE

Morteza Saberi
University of Technology Sydney
180 PUBLICATIONS **4,479** CITATIONS
SEE PROFILE

Elizabeth Chang
Curtin University
416 PUBLICATIONS **7,932** CITATIONS
SEE PROFILE

Towards Expert Preference on Academic Article Recommendation Using Bibliometric Networks

Yu Zhang[1], Min Wang[1], Morteza Saberi[2], and Elizabeth Chang[1]

[1] University of New South Wales, Canberra, ACT 2604, Australia
yu.zhang, min.wang, e.chang@adfa.edu.au
[2] University of Technology Sydney, Sydney, NSW 2007, Australia
morteza.saberi@uts.edu.au

Abstract. Expert knowledge can be valuable for academic article recommendation, however, hiring domain experts for this purpose is rather expensive as it is extremely demanding for human to deal with a large volume of academic publications. Therefore, developing an article ranking method which can automatically provide recommendations that are close to expert decisions is needed. Many algorithms have been proposed to rank articles but pursuing quality article recommendations that approximate to expert decisions has hardly been considered. In this study, domain expert decisions on recommending quality articles are investigated. Specifically, we hire domain experts to mark articles and a comprehensive correlation analysis is then performed between the ranking results generated by the experts and state-of-the-art automatic ranking algorithms. In addition, we propose a computational model using heterogeneous bibliometric networks to approximate human expert decisions. The model takes into account paper citations, semantic and network-level similarities amongst papers, authorship, venues, publishing time, and the relationships amongst them to approximate human decision-making factors. Results demonstrate that the proposed model is able to effectively achieve human expert-alike decisions on recommending quality articles.

Keywords: Academic article recommendation · Human expert decision · Bibliometric networks · Data management.

1 Introduction

Academic article recommendation has been critical for readers who are seeking for articles with high prestige [12, 14], and it would be particularly valuable if the recommendation can be provided by domain experts. However, achieving article recommendation that is approximate to expert decisions has been hardly considered.

Hiring collective human intelligence, known as crowdsourcing, has been frequently adopted as an overarching method for large-scale database annotation and evaluation tasks. It usually requires human input in terms of natural language annotation, computer vision recognition, and answering cognitive questions [2]. The applications of crowdsourcing techniques in these domains have demonstrated its relatively high accuracy and reliability [11].

Although having expert judgement on academic articles would be an ideal way to rank them and recommend the top ones, it is a huge workload for human to manually evaluate each article and rank them accordingly, especially when human resources are rather limited while the workload is heavy. Automatic paper ranking and recommendation system has been a sound solution to handle the enormous and ever growing volume of articles. Classic algorithms include PageRank [5], CoRank [15], FutureRank [7], and P-Rank [10]. In addition, more advanced algorithms, such as HITS [8] and W-Rank [13], have also been proposed recently to integrate more bibliometric information and factors. However, there is a lack of analysis of these algorithms towards human expert decisions. Some studies adopted crowdsourcing approaches to rank only a small amount of top articles and compared the ranking results with those generated by the algorithms for evaluation purposes [9]. Asking crowd workers to investigate the top articles makes sense to certain extent since the essential aim of article ranking is to recommend valuable articles to readers, nonetheless, it remains infeasible for the workers to retrieve the top ones from large-scale article databases. Therefore, it is necessary to design a method which can automatically generate a reasonable article ranking list that approximates to domain expert decisions.

In this study, domain expert decision on ranking academic articles is investigated and a computational model is proposed to approach the expert decisions. Specifically, we analyse the decisions of a group of experts in the research domain of intrusion detection (cyber security) by comparing their ranking results with those generated by state-of-the-art algorithms. Based on the analysis, a computational model is proposed to approximate expert decision on recommending prestige articles using a heterogeneous bibliometric network. The proposed method is fully automatic and is able to generate reasonable ranking results approaching domain expert decisions. The contribution of this study can be summarised into two folds. Firstly, the analysis of domain expert decisions and the comparative study remedy the gap that the existing literature has barely considered to approximate expert decision on recommending quality academic articles. Secondly, the proposed method is able to automatically generate reasonable ranking results close to those from domain experts, breaking through the limitations of human's capability in handling large-scale bibliometric databases.

2 Materials and Methods

2.1 Domain expert recommendation analysis

In this study, three domain experts in the research field of intrusion detection (cybersecurity) are hired, and accordingly, we use bibliometric data in the same field for analysis and evaluation. The data includes 6,428 articles, 16,284 citations, 12,093 authors and 1,048 venues collected from the Microsoft Academic Graph (MAG) database from year 2000 to 2017. A historical time point is set at year 2008 to divide the database into two parts. The articles published from 2000 to 2008 are used for testing the algorithms, and the remaining articles from 2009 to 2017 are used for evaluating the future trends.

Since it is infeasible for human to deal with such large-scale databases, we use citation count and five popular ranking algorithms, including PageRank, Co-Rank, P-Rank, FutureRank, and HITS as baselines to rank papers in the first partition. Afterwards, only the 30 top articles ranked by each baseline are collected and summarised into a list of 120 articles after removing overlaps, then fed to the experts because recommending top articles is the essential target for academic article recommendation.

Statistical methods are used to evaluate the expert decisions and converge them into one result. Specifically, the Joglekar's algorithm [3] is used to generate confidence intervals for expert decision error rate estimates by assuming a Bernouli distribution for paper scores made by each expert. Suppose there are M experts and N papers, the error rate of expert i, denoted as ϵ_i, is calculated as follows:

$$\epsilon_i \leftarrow \frac{1}{2} - \sqrt{\frac{\prod_{j \in Q_1, Q_2}(a_{ij} - 1/2)}{2(a_{Q_1 Q_2} - 1/2)}}; \quad M = Q_1 \cup Q_2 \cup \{i\} \tag{1}$$

where Q_1 and Q_2 are two disjoint sets for expert i, a_{ij} denotes the normalised number of times that expert i and expert j agree on the scores. The agreement of score on paper n between experts i and j is defined as follows:

$$a_{i,j}(n) = \begin{cases} 1 & \text{if } |s_i(n) - s_j(n)| \leq \theta \\ 0 & \text{otherwise;} \quad n \in \{1, ..., N\}; i, j \in \{Q_1, Q_2\} \end{cases} \tag{2}$$

where θ is a threshold set to 0.05. Exhaustive search strategy is adopted to select the optimal disjoint sets Q_1 and Q_2 which provide the minimum value for $z_t\sqrt{(\epsilon_i(1-\epsilon_i))/N}$, where z_t represents the tth percentile of the standard normal distribution.

Algorithm 1: Converging human expert scores

Input: Score list by each expert: S_i; error rate for each expert: ϵ_i, $i \in \{1, ..., M\}$

Output: Normalised final score list: S

$i^* \leftarrow \text{Argmin}\, \epsilon_i$
$Q_1, Q_2 \leftarrow ExhaustiveSearch(i^*, M)$
for $n = 1 : N$ **do**
  $d_{ij}(n) \leftarrow |s_i(n) - s_j(n)|; i, j \in \{i^*, Q_1, Q_2\} = M$
  if $size(\text{Argmin}\, d_{ij}(n)) \leq 1$ **then**
    $\{i, j\} \leftarrow \text{Argmin}\, d_{ij}(n)$
    $S(n) \leftarrow \frac{(1-\epsilon_i)S_i(n) + (1-\epsilon_j)S_j(n)}{2}$
  else
    $S(n) \leftarrow \frac{\sum_{i \in M}(1-\epsilon_i)S_i(n)}{3}$
return S

Finally, majority voting scheme is used for converging the scores made by multiple experts since it is able to determine a more reasonable score for every article based on the opinions of the majority. The processing procedure is summarised in Algorithm 1.

2.2 Computational model

Heterogeneous bibliometric network The bibliometric information including articles, authors, venues (journals and conferences), and the relationship amongst them form a heterogeneous network $\mathcal{G}$ which, as illustrated in Fig. 1, can be described with a set of nodes $\mathcal{N}$ and their links $\mathcal{L}$ as follows:

$$\mathcal{G} = \mathcal{G}_{P-A} \cup \mathcal{G}_{P-P} \cup \mathcal{G}_{P-V} \tag{3}$$

$$= \{\mathcal{N}, \mathcal{L}\} = \{\mathcal{N}_A \cup \mathcal{N}_P \cup \mathcal{N}_V, \mathcal{L}_{P-A} \cup \mathcal{L}_{P-P} \cup \mathcal{L}_{P-V}\} \tag{4}$$

where P, A and V denote article, author, and venue, respectively. Considering the citation relevance, the citation network is further updated to $\mathcal{G}_{P-P} = \{\mathcal{N}_P, \mathcal{L}_{P-P}, \mathbf{W}\}$, where $\mathbf{W} \in \mathbb{R}^{N \times N}$ is the adjacency matrix of the citation network and $N = |\mathcal{N}_P|$ is the number of articles in it. The adjacency matrix $\mathbf{W}$ is a representative description of the citation network structure with its entries, denoted as $w_{i,j}$, referring to the relevance of a citation link from article i to article j.

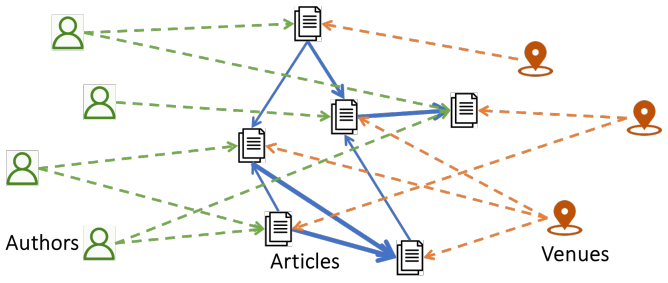

Fig. 1: The heterogeneous bibliometric network

In this study, we follow the definition and implementation of citation relevance in W-Rank [13] which interprets this concept from two perspectives, namely the semantic similarity of the articles' content and the network-level similarity evaluating the mutual links in the citation network. Specifically, we extract the titles and abstracts of articles (as they usually contain the key information of an article), and calculate the semantic similarity of these lexical items using a sense-level algorithm named 'align, disambiguate and walk' [6] which has been demonstrated to be flexible in handling lexical items in different lengths and effective in comparing the meaning of the lexical items. For network-level

similarity, we use Cosine similarity to measure the mutual links in the citation network according to the following equation:

$$Cosine(P_i, P_j) = \frac{|L_{P_i} \cap L_{P_j}|}{\sqrt{|L_{P_i}| \times |L_{P_j}|}} \tag{5}$$

where P_i and P_j denote two papers, L_P denotes the links that connect to node P in the citation network, and $L_{P_i} \cap L_{P_j}$ represents the links connecting to both P_i and P_j regardless of the link direction. Finally, the citation relevance is formulated as an integration of the semantic similarity and network-level similarity as follows:

$$w_{i,j} = \alpha \cdot Semantic(P_i, P_j) + \beta \cdot Cosine(P_i, P_j) \tag{6}$$

where α and β are coefficients defined by an exponential function $e^{\lambda(Similarity-\tau)}$, where λ is set to 6 in favour of the similarity values which are greater than the threshold, and the threshold τ is adjusted to be the median values of the two types of similarities, respectively. The α and β are normalised so that $\alpha + \beta = 1$.

Ranking algorithm The article ranking algorithm is designed to obtain a score for each article and this is accomplished by propagating between article authority scores S and hub scores H from three types of nodes (paper P, author A, and venue V) in the heterogeneous network.

Specifically, for author-paper network and paper-venue network, the hub score of an author or a venue node X_i is computed according to Equation 7 and the corresponding authority score propagated from the hub score is computed by Equation 8.

$$H(X_i) = \frac{\sum_{P_j \in Out(X_i)} S(P_j)}{|Out(X_i)|} \tag{7}$$

$$S_X(P_i) = Z^{-1}(X) \sum_{X_j \in In(P_i)} H(X_j) \tag{8}$$

where $Out(X_i)$ represents the paper nodes linked from the author or venue node X_i in the network, and $In(P_i)$ denotes all the nodes linked to paper node P_i, and $Z(\cdot)$ is a normalisation term.

For citation network, the calculation of the hub score and authority score follows a similar procedure, except that the citation network is a weighted network thus requires the following updates:

$$H(P_i) = \frac{\sum_{P_j \in Out(P_i)} w_{i,j} S(P_j)}{\sum_{P_j \in Out(P_i)} w_{i,j}} \tag{9}$$

$$S_P(P_i) = Z^{-1}(P) \sum_{P_j \in In(P_i)} H(P_j) w_{i,j} \tag{10}$$

where $In(P_i)$ and $Out(P_i)$ denote the nodes linked to and from paper node P_i, respectively.

In addition, we consider publishing time as a factor to promote new papers which are often underestimated by citation-based models due to inadequate citations. An exponential function is used to favour those which are close to the 'Current' time: $S_T(P_i) = Z^{-1}(T)e^{-\rho(T_{Current}-T_{P_i})}$, where $\rho = 0.62$, $T_{Current}$ is the current time of evaluation, and Z is a normalisation term. Finally, the paper authority score S is updated considering the above four components as follows:

$$S(P_i) = \alpha \cdot S_P(P_i) + \beta \cdot S_A(P_i) + \gamma \cdot S_V(P_i) + \delta \cdot S_T(P_i) + (1-\alpha-\beta-\gamma-\delta) \cdot \frac{1}{N_p} \quad (11)$$

where N_p is the total number of papers in the collection, and the last term represents a random jump. We set the four parameters as $\alpha + \beta + \gamma + \delta + \theta = 0.85$ to allow a 0.15 probability of random jumps. Algorithm 2 summarises the corresponding details.

Algorithm 2: Ranking algorithm

Input: heterogeneous network: $\mathcal{G}_{P-A} \cup \mathcal{G}_{P-P} \cup \mathcal{G}_{P-V}$; publishing time: T_P
Output: paper authority score: S
Parameter: $\alpha, \beta, \gamma, \delta, \tau, \rho$

initialise: $S \leftarrow \{1/N_p, 1/N_p, ..., 1/N_p\}$; $old = 1$; $new = -1$;
calculate time score: $S_T(P) \leftarrow exp(-\rho(\tau - T_P))$
while $any(abs(old - new) > 0.0001)$ **do**
    update hub score and authority score:
    $H(A) \leftarrow GetHubScore(\mathcal{G}_{P-A}, S)$
    $H(V) \leftarrow GetHubScore(\mathcal{G}_{P-V}, S)$
    $H(P) \leftarrow GetHubScore(\mathcal{G}_{P-P}, S)$
    $S_A(P) \leftarrow GetScore(\mathcal{G}_{P-A}, H(A))$
    $S_V(P) \leftarrow GetScore(\mathcal{G}_{P-V}, H(V))$
    $S_P(P) \leftarrow GetScore(\mathcal{G}_{P-P}, H(P))$
    update paper authority score: $S \leftarrow Integrate(\alpha S_P, \beta S_A, \gamma S_V, \delta S_T, \frac{1}{N_p})$
    $old = new$; $new = S$;
return S;

3 Results

Based on the database partition, we respectively calculate the citation count (CC), future citation count (FC) and weighted future citations (WFC) generated before and after the historical time point, and then compare them with the domain experts' decisions. Since the objective of this study is to approximate expert decisions on recommending top articles, the ground truth for ranking the articles is the rank list aggregated from the human experts' decisions. Spearman's rank correlation coefficient (ρ) [4] is used to assess the similarity between the ground truth list and the rank list obtained by the selected baselines. We

experimentally use the optimal parameters for all the tested methods. The corresponding 0.95 confidence interval (CI) of the estimated ρ is calculated by Fisher transformation [1]: $CI = \tanh(\text{Arctanh}\,\rho \pm z_{\alpha/2}/\sqrt{n-3})$, where N is the sample size and $z_{\alpha/2} = 1.96$ is the two-tailed critical value of the standard normal distribution with $\alpha = 0.05$. In addition, the detection error trade-off (DET, FNR/FPR) curves are generated to compare the classification performance at different thresholds which divide the articles into the highly ranked (positive group) and the lower ranked (negative group).

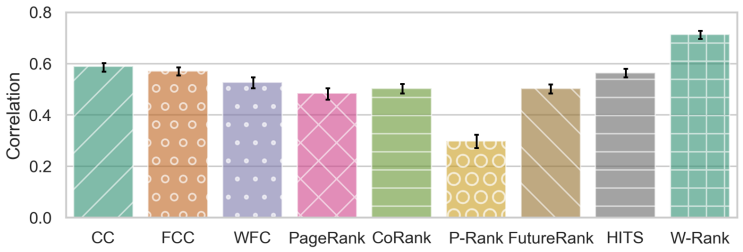

Fig. 2: Spearman's rank correlation between domain expert decision and each method (the error bar indicates a CI of 0.95)

The Spearman's rank correlation results are summarised in Fig. 2. Amongst all the methods, the proposed W-Rank achieved the highest correlation with human expert decisions at 0.714, outperforming the existing algorithms by at least 41.7%. This improvement is also demonstrated by the DET curves shown in Fig. 3. A deeper investigation on the algorithm settings shows that amongst all the bibliometric factors, citation and publication time play a more important role than other factors in approximating ranking results to human expert decisions. However, as the results suggested, it is insufficient to only use citation number to rank articles. Instead, various bibliometric factors should be considered simultaneously, and this is consistent with the decision making process of

Fig. 3: DET curves of each method

human experts. In addition, a dramatic drop can be found at the correlation of the P-Rank algorithm in Fig. 2, and the DET curve of the P-Rank in Fig. 3 also covers the largest area under the curve (AUC) indicating poor performance as well. Since P-Rank explored the bibliometric combination of citation, author and venue while all the other baselines either did not use venue factor or added time and other factors, it can be concluded that the venue factor is a redundant factor in approximating expert decisions.

4 Conclusion

Approaching domain expert decisions is a promising direction in designing automatic ranking system for academic article recommendation. In this study, we investigated expert decisions on quality article recommendation by performing a comparative analysis between the results provided by the experts and generated by existing article ranking algorithms. In addition, a computational model was proposed to automatically generate quality article recommendations based on heterogeneous bibliometric networks considering the information of citations, authorship, venues, publishing time, and their relationships. The comparative study and experimental results demonstrate the effectiveness of the proposed model in generating reasonable recommendations that are approximate to expert decisions. It brings an inspiring perspective of integrating both human and machine intelligence into bibliometric ranking and crowdsourcing. In future studies, we will improve the approaches of seeking optimal parameter setting, complementary and redundant features from the bibliometric information for W-Rank using more advanced machine learning methods rather than being experimentally measured.

References

1. Fisher, R.A.: Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika **10**(4), 507–521 (1915)
2. Jin, Y., Du, L., Zhu, Y., Carman, M.: Leveraging label category relationships in multi-class crowdsourcing. In: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 128–140. Springer (2018)
3. Joglekar, M., Garcia-Molina, H., Parameswaran, A.: Evaluating the crowd with confidence. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 686–694. ACM (2013)
4. Myers, J.L., Well, A.D., Lorch Jr, R.F.: Research design and statistical analysis. Routledge (2013)
5. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
6. Pilehvar, M.T., Jurgens, D., Navigli, R.: Align, disambiguate and walk: A unified approach for measuring semantic similarity. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1341–1351 (2013)

7. Sayyadi, H., Getoor, L.: Futurerank: Ranking scientific articles by predicting their future pagerank. In: Proceedings of the 2009 SIAM International Conference on Data Mining. pp. 533–544. SIAM (2009)
8. Wang, Y., Tong, Y., Zeng, M.: Ranking scientific articles by exploiting citations, authors, journals, and time information. In: Twenty-seventh AAAI conference on artificial intelligence (2013)
9. Wang, Z., Liu, Y., Yang, J., Zheng, Z., Wu, K.: A personalization-oriented academic literature recommendation method. Data Science Journal **14** (2015)
10. Yan, E., Ding, Y., Sugimoto, C.R.: P-rank: An indicator measuring prestige in heterogeneous scholarly networks. Journal of the American Society for Information Science and Technology **62**(3), 467–477 (2011)
11. Zhang, X., Shi, H., Li, Y., Liang, W.: SPGLAD: A self-paced learning-based crowdsourcing classification model. In: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 189–201. Springer (2017)
12. Zhang, Y., Saberi, M., Wang, M., Chang, E.: K3S: Knowledge-driven solution support system. In: Proceedings of the twenty-seventh AAAI conference on artificial intelligence. vol. 33, pp. 9873–9874 (2019)
13. Zhang, Y., Wang, M., Gottwalt, F., Saberi, M., Chang, E.: Ranking scientific articles based on bibliometric networks with a weighting scheme. Journal of Informetrics **13**(2), 616–634 (2019)
14. Zhang, Y., Wang, M., Saberi, M., Chang, E.: From big scholarly data to solution-oriented knowledge repository. Frontiers in Big Data **2**, 38 (2019)
15. Zhou, D., Orshanskiy, S.A., Zha, H., Giles, C.L.: Co-ranking authors and documents in a heterogeneous network. In: Seventh IEEE international conference on data mining (ICDM 2007). pp. 739–744. IEEE (2007)