Can Scientific Publication's Network Structural Features Predict its Citation?

Zhuoran Luo[†], Jiangen He[‡*], Jiajia Qian[†], Yuqi Wang[†], Junying Chen[†] ,Wei Lu[†*]
†School of Information Management, Wuhan University
‡Department of Information Science, Drexel University
{zoraluo,weilu,jiajiaqian,yuqi.wang2014, junyingchenvip}@whu.edu.cn
{jiangen.he}@drexel.edu

ABSTRACT

The citation relationship between scientific publications constitutes a huge and complex citation network, which is of great significance for hotspot analysis and cutting-edge prediction in different fields. Nevertheless, how to evaluate the novelty and impact of a scientific publication in its early stages remains an open question. To address this issue, we apply a network representation learning approach (struc2vec) to represent the full complexity of citation network structure, explore the extent to which an emerging science publication has changed the network structure of existing knowledge, and explain the relationship between this change and the paper's cited numbers from both clustering and network visualization perspectives. We found that the structural features captured by struc2vec can predict future citations of scientific publications to some extent. The predictive effects can be interpreted by how a new publication connects to and alters the existing structure of scientific knowledge in our visual analytics.

CCS CONCEPTS

• Computing methodologies →Artificial intelligence→Knowledge representation and reasoning;

KEYWORDS

Representation Learning, Structural Features, Citation Prediction

ACM Reference format:

Zhuoran Luo, Jiangen He, Jiajia Qian, Yuqi Wang, Junying Chen and Wei Lu. 2020. Can scientific publication's network structural features predict its citation? *JCDL '20, August 1–5, 2020, Virtual Event, China, 2 pages. https://doi.org/10.1145/3383583.3398575*

1 INTRODUCTION

Citation network is one of the important media of scientific knowledge diffusion [1] and is important for researchers to understand the mechanisms by which certain knowledge is generated, developed, and formed. Uzzi B. et al. [2] found that the most influential scientific studies were mainly combined with previous conventional work through citation network analysis. Chen [3] measured the novelty of newly introduced literature by analyzing the boundary crossover effects of citation networks. However, in the face of a large citation network, traditional scientometric methods cannot effectively exploit the local structural features of the citation network. In recent years, deep learning and non-linear descending dimensional techniques have been applied to graphical representation learning methods but they often do not work well in the classification task of determining whether the local topologies of distant nodes in a citation network are similar. Struct2vec [4] is an algorithm that defines node similarity in terms of network structural similarity and is shown to be effective in capturing nodes with similar local structures in a network. How to understand the relationship between these changes induced by new publications and the novelty and future impact they represent is still an open question. To address these questions, by integrating approaches of machine learning and network visualization, we explore the extent to which the emergence of a new scientific paper changes original the structure of knowledge networks, and analyze the relationship between the change and the future impact of the article. The schematic of our research process is shown in Figure 1.

2 REPRESENTATION OF THE CITATION NETWORK

In the core database of the Web of Science, we searched 1343 papers on autophagy from 1973 to 2005, including 216 articles in 2005. First, we generate the citation network G based on the paper's citation relationships. Then we capture the features of G using the struct2vec model and represent each node in G with a 128-dimensional embedding. However, it is difficult to interpret the characteristics of the nodes in the network using 128-dimensional data, and for this reason, we choose to downscale and visualize the embedding results using the t-SNE algorithm. Finally, we use the K-means algorithm to cluster the nodes represented by the embedding to obtain groups of nodes with similar citation structures. We use a boxplot to count and display the distribution of paper citations in each cluster, as shown in Figure 2. Based on the length of the boxplot, we find that the data frame lines for clusters 1, 7, and 19 are shorter, and the mean value of each cluster and median is higher, which indicate that the Struct2vec algorithm can capture the nodes with a similar network structure, and that article nodes with similar citation network structure features have relevant patterns for future citation cases.

* Corresponding authors.
† Author from School of Information Management, Wuhan University
‡Author from Department of Information Science, Drexel University

JCDL '20, August 1–5, 2020, Virtual Event, China

ACM ISBN 978-1-4503-7585-6/20/08. https://doi.org/10.1145/3383583.3398575

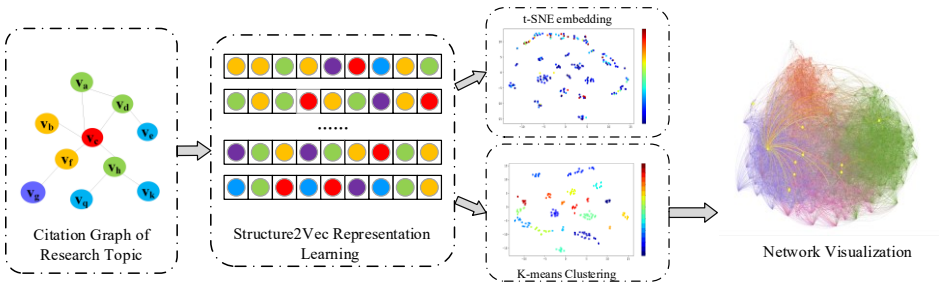

Figure 1. The schematic of our research process

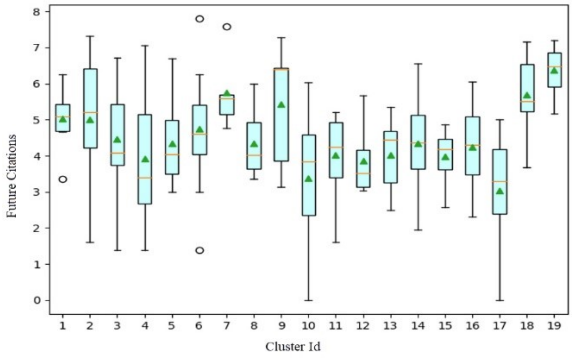

Figure 2. Boxplot of cited count for each group

3. CITATION NETWORK VISUALIZATION

To further explain our findings, we explored from a visual perspective how articles published in the field of autophagy in 2005 were added as new content to the pre-2005 citation network. We imported information from 1343 articles from 1973 to 2005 in the topic of cellular autophagy into Citespace, and performed layout and association clustering of the co-citation network to obtain 16 modules, as shown in Figure 3(a). Here we select three groups from the 19 groups obtained from the embedding clusters for structural variance analysis.

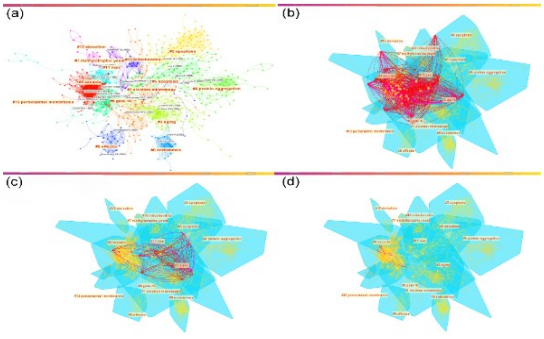

Figure 3. Structural variance of the co-citation network

Among them, Class19 is the highly cited group, and it can be seen from the new connections in Figure 3(b) that the papers in this group bring a large number of new connections to the network, with a large structural change to the original network. Class1 is the medium cited group, and it can be seen from Figure 3(c) that the papers in its class bring some new connections to the co-referenced network, but it does not perform as significantly as Class19 in terms of number and class span contribution. Class15 is the low cited group, and it can be seen from Figure 3(d) that the papers in this group bring the least structural change to the network. We can see that the structure of the co-citation network varies greatly between groups, also, it should be noted that this structural variation is strongly consistent within the same group.

4. CONCLUSION

Inspired by representation learning and graph embedding studies, we apply the network representation learning approach to citation network studies, and then use representation learning techniques to mine the literature with similar local network topologies in citation networks to explore whether the similarity in network structure can be used to predict the citation of papers. Through our research, we found that the structural features of the citation network captured by the representational learning model predicted the citation of the article and the novelty of the knowledge and methods represented by the paper to a certain extent.

This is an ongoing study, and in the future, our research data will cover papers in more areas and further validate the strengths of our approach through statistical models.

ACKNOWLEDGMENTS

This work was supported by the Major Project of National Social Science Foundation of China[17ZDA292].

REFERENCES

[1] Rogers, E. M. 1995. Diffusion of innovations(4th. ed). New York: Simon and Schuster.

[2] Uzzi, B., Mukherjee, S., Stringer, M. J., & Jones, B. 2013. Atypical Combinations and Scientific Impact. Science, 342(6157), 468-472. DOI: https://doi.org/10.1126/science.1240474.

[3] Chen, C. 2012. Predictive effects of structural variation on citation counts. Journal of the Association for Information Science and Technology, 63(3), 431-449. DOI: https://doi.org/10.1002/asi.21694.

[4] Ribeiro, L. F., Saverese, P. H., & Figueiredo, D. R. 2017. struc2vec: Learning node representations from structural identity. In SIGKDD. 385-394. DOI: https://doi.org/10.1145/3097983.3098061.