EasyChair Preprint

№ 4529

An Improved Framework for Identifying Emerging Author Keywords

Yang Jinqing, Huang Shengzhi, Wei Yuhan, Liu Zhifeng and Lu Wei

November 7, 2020

An improved framework for identifying emerging author keywords

1.Introduction

Keywords are consensus expressions of a concept formed by researchers in a specific field (Allan, 1998). They are the words that express the main topics of your research article. The emergence of new keywords performs better in representing the innovative state of scientific research in a specific discipline. Therefore, identifying emerging keywords can monitor the evolution of research topics and find valuable researches early.

In previous studies, the bibliometric method has been widely adopted into the pioneering works of identifying emerging topic or technology. Novelty and growth are the earliest adopted and most basic characteristics of an emerging topic. Time feature is the essential element of measuring novelty (Price, 1963), and it is an objective physical variable that is not affected by external factors. However, it is inadequate for different keywords to characterize novelty only considering the time factor, because different keywords may have personalized novelty at the same time point. Therefore, we proposed an approach that can measure the dynamic novelty of the individual keyword.

2. Proposed method

In this study, the new metric with the acceleration attenuation factor is proposed to measure the dynamic novelty of the individual keyword. We first regard the intersection of the novelty curve and growth curve of a keyword as the niche position in the entire experimental dataset, and then compute slope value in the niche position. Finally, we computed the medians of slope values of all keywords in a niche position to form the niche baseline. The slope values in a niche position are defined as a niche position value. If the niche position value of a keyword is higher than the niche baseline for the first time, the keyword is emerging at the time point, otherwise, it is still in the bud. The purpose of this study is to earlier find the research topics that have sustained growth in the years that followed. High growth rate before niche position (Front growth ratio) can verify the validity of results and strong sustained growing trend after niche position (Later growth ratio) can prove foresight of results. The framework is showed in Fig.1.

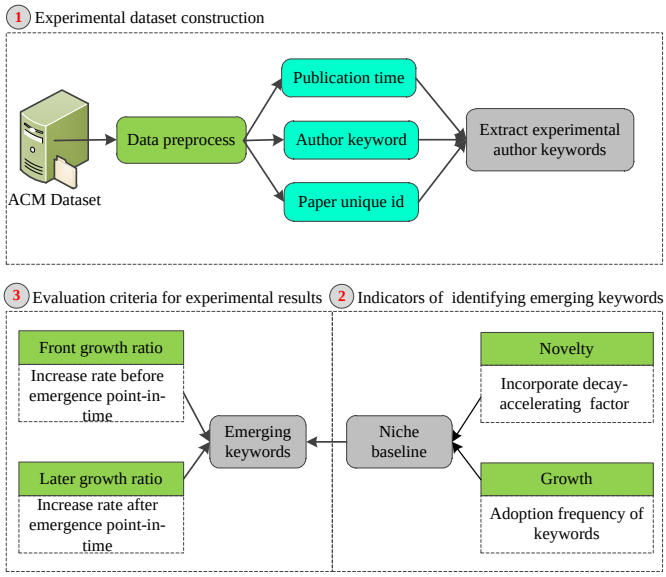

Fig.1 The framework for identifying emerging keywords

3. Results

To verify the feasibility and reliability of this framework, we took the pioneering work (Tu,

2012) that did not integrate the acceleration attenuation factor as the baseline for experimental comparison. The first task of our work was to evaluate the ability of this framework on distinguishing emerging keywords from general keywords. According to the order of cumulative keyword frequency, we selected the fore top-n as positive samples and the rear top-n as negative samples which were used to construct the experimental dataset. If the adoption frequency of keywords from the rear part has too low adoption frequency, the evaluation effect will be undifferentiated. Therefore, we selected the keywords with a cumulative frequency greater than 10 in the negative samples. To properly evaluate the experimental results, the precision and recall of the two frameworks were computed based on our constructed experiment data. By adopting the dataset construction strategy mentioned above, we built five datasets with different top-n to evaluate our framework. The experimental results are shown in Table 1.

Table 1 Evaluation of the two frameworks' performance

Top n%	N	Baseline			Our proposed framework		
		Precision	Recall	F1	Precision	Recall	F1
0.05%	119*2	0.5022	0.9664	0.6609	0.7532	0.9747	0.8498
0.1%	238*2	0.4898	0.9580	0.6561	0.7541	0.9664	0.8471
0.5%	1190*2	0.4857	0.9244	0.6368	0.7006	0.9142	0.7933
1%	2381*2	0.4800	0.9034	0.6269	0.6784	0.8727	0.7634
1.5%	3571*2	0.4741	0.8804	0.6163	0.6517	0.8421	0.7348

As shown in Table 1, our proposed framework has a better performance in identifying emerging topics. The F1 value of our framework for all top-n data was much better than that of baseline. Specifically, the precision of our framework was always higher. Although the recall rate sometimes was slightly lower, it was still acceptable. On the contrary, the baseline was unable to distinguish positive and negative samples. We intended to help decision-makers and scholars to promising research topics, and we believe the emerging topic should be a high-value research topic in the future. However, this binary classification which just evaluated the capability of distinguishing emerging keywords and general ones is unable to verify the timeliness of identifying emerging keywords.

We fulfill the second evaluation task from three perspectives, (1) emergence time point (2) front growth ratio before emergence time point, and (3) later growth ratio after emergence time point. We selected the 1850 keywords identified by both frameworks as this experimental dataset its results are shown in the following Fig.2.

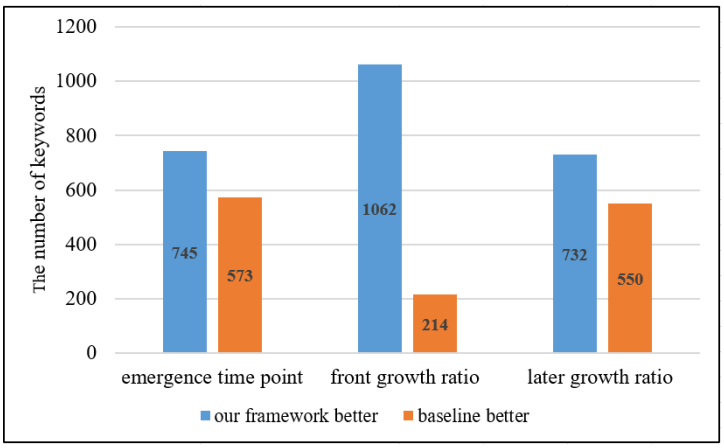

Fig.2 Results of the second evaluation between frameworks

As shown in Fig.2, our proposed method can earlier identify more emerging keywords than the baseline method. It also has a higher front growth ratio before niche position, and more emerging keywords identified have a stronger sustained growing trend than baseline. In detail, 40.27% of emerging keywords were identified earlier by our framework than that of baseline, while 30.89% of emerging keywords from our framework have a later emergence time point. Also, 57.41% of emerging keywords had faster growth before niche position than the baseline, which is consistent with the original intention of our framework. Finally, 39.57% of emerging keywords had faster growth after niche position than the baseline, and only 29.72% of emerging keywords fell behind. It means that emerging keywords identified by our framework have a stronger sustained growing trend.

4. References

Allan, J., Carbonell, J. G., Doddington, G., et al. (1998, February). Topic Detection and Tracking Pilot Study: Final report[C]// In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, USA. San Francisco, CA, Morgan Kaufmann Publishers, Inc, 194-218.

Carley, S. F., Newman, N. C., Porter, A. L., et al. (2018). An indicator of technical emergence. *Scientometrics*, 115(1), 35-49.

Garner, J., Carley, S., Porter, A. L., et al. (2017, July). Technological emergence indicators using emergence scoring. In *2017 Portland international conference on management of engineering and technology (PICMET)* (pp. 1-12). IEEE.

Price D J S. (1963). Little Science, Big Science-- And Beyond. *Von Der Studierstube Zur*, 7(3-6), 443-458.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research policy*, 44(10), 1827-1843.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research policy*, 43(8), 1450-1467.

Tu, Y. N., & Seng, J. L. (2012). Indices of novelty for emerging topic detection. *Information processing & management*, 48(2), 303-325.