Quantifying Individual Research's Distance from the Trends based on Dynamic Topic Modeling

Meng, Jie University of Chinese Academy of Sciences, People's Republic of China | moonjaymengjie@gmail.com
Lou, Wen East China Normal University, People's Republic of China | wlou@infor.ecnu.edu.cn
He, Jiangen University of Tennessee, Knoxville, USA | jiangen@utk.edu

ABSTRACT
Research trends are the keys for researchers to decide their research agenda. However, only few works have tried to quantify how scholars follow the trends. This paper addresses this problem by proposing a novel measurement for quantifying how a scientific entity (paper or researcher) follows the hot topics in a research field. Specifically, the topic evolution and papers are vectorizing by dynamic topic modeling. Then the degree of hotness tracing is explored from three different perspectives: mainstream, short-term direction, long-term direction. Papers and researchers in the field of Computer Vision from 2006 to 2017 were selected to evaluate our method. Further study will show the results of topic evolution patterns and researchers' clusters.

KEYWORDS
Research trends; Dynamic topic modeling; Hot topics; Research behavior; NLP

INTRODUCTION
The shocking research alerted that the progress of large scientific fields may be slowed canonical (Chu and Evans, 2021), which revealed a phenomenon that many researchers would follow mainstream research over time. Quantifying the mainstream in a research field and identifying researchers with different behaviors can be challenging (Small et al., 2014). We aim to model such research trends and the distances of each scientific entity (e.g. a paper, a researcher) from the trends to explore researchers' behaviors in science communities.

METHOD

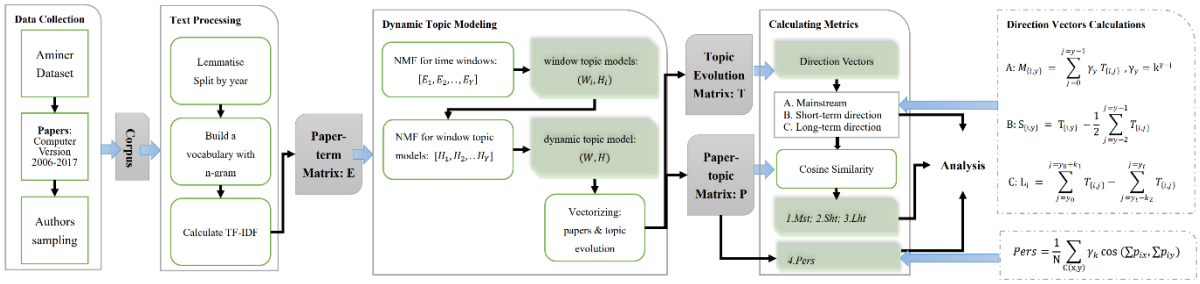

Figure 1. Research design and dynamic topic modeling process

We designed the method as four steps with three direction vectors and four metrics (Figure 1). Firstly, basic text processing techniques are utilized to create a paper-term matrix, *E*. Secondly, we quantify the trends and papers with a set of dynamic topics $DT = [dt_1, dt_2, \dots, dt_n]$, which is generated by dynamic topic modeling (DTM). DTM is implemented by three layers of Non-negative Matrix Factorization (NMF) (Greene and Cross, 2017), producing the topic model *(W, H)*. The *i-th* row of *H* represents the term distribution of dt_i. The papers P_y (for year y) can be vectorized by solving the linear system $P_y H = E_y$. Thus, a paper is represented by the distribution of *DT*. Further, we model the topic trends as a topic evolution matrix T. Two steps are included: (1) Topics with *n* largest weights are assigned to a paper, and other values in P_y are set to zero. (2) Calculating the *y-th* column of T by aggregating the topic vectors of all papers in one year $T_y = \sum_i P_{iy}$. T_y represents the distribution of *DT* in the *y-th* year of this field.

The trends of hot topics are described from three perspectives. Accordingly, three direction vectors are defined to quantify these trends. The *Mainstream (ma)* of a field is a direction vector representing established research agenda. It's a measure of the accumulated topics over the years, calculated by a weighted summing of T, as in Figure 1. γ represents decay coefficient and $k < 1$. (We set $k = 0.8$). Since papers from far-off years have less of an impact on the mainstream, the decay coefficient decreases with increasing proximity to year y. The *Short-term Direction (sd)* is to measure topics' popularity over a brief period of time (2 or 3 years). The calculation takes into account the difference between the topic vectors for year y and the average of the topic vectors for the two years prior. The *Long-term Direction (ld)* depicts a growth tendency over a long period of time, measuring the variation in average topic vectors between the starting and ending years.

85th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 1, 2022 | Pittsburgh, PA.

Once the pattern of topic trend evolution is established, we can quantify individual research's similarity with the trends. Given a direction vector *D (ma, sd, ld)*, three metrics for papers are defined: (1) *Mainstream tracing degree (Mst), (2) Short-term hotness tracing degree (Sht), (3) Long-term hotness tracing degree (Lht)*, which are derived by cosine similarity: $\cos < P_i, \mathrm{D_y} >$. For authors, these metrics are defined as the arithmetic mean of the indicator of their papers: $(1/N)\sum_{i=1}^{N} cos < P_i, D >$. For a thorough understanding of the scholar's research interests, we propose another metric, (4) *Persistence (Pers),* to quantify the continuation of the researcher's work. It is derived by weighted summing the similarity of the topic vectors of the scholars' papers from various years, as illustrated in Figure 1. $\sum p_{ix}$ represents the sum of all paper vectors in year x; $C(x, y)$ denotes all combinations of *(x,y)*; γ_k denotes the decay coefficient, $\gamma_k = k_1 + |x - y| * k_2$, $k_2 < 0$ (Here we set $k_1 = 2, k_2 = -0.25$). The decay coefficient indicates that two closer years are given more weight.

RESULT AND DISCUSSION

The dataset used in this article comes from AMiner (Wan et al., 2019), where the authors are disambiguated. AMiner assigns one or more discipline(s) to each publication. The final dataset, including of 279,875 articles, is composed of computer vision papers published between 2006 and 2017. Authors who are active in the field of computer vision are sampled by the two criteria: 1. They have more than three papers. 2. They have published in the field for more than three years. Finally, 45,203 unique authors are included.

Spearman correlation analysis (Table 1) shows that all metrics for papers are positively correlated with each other, among which *Lht* has more significant correlation with *Mst*, while correlation between *Sht* and *Mst* is relatively weaker. The results reveal that papers following long-term hotspots are more consistent with the mainstream than short-term hot topics. For authors, *Sht* has the most significant positive correlation with *Lht*, revealing that the researcher who prefers long-term hotspots also tends to follow short-term direction. Besides, *Pers* has very little correlation with other indicators, indicating that adherence to a research direction is not related to following the hot topics.

Metrics	***Papers***			***Authors***			
	Sht	***Lht***	***Mst***	***Sht***	***Lht***	***Mst***	***Pers***
Sht	1	0.592	0.379	1	0.719	0.413	0.058
Lht	0.592	1	0.711	0.719	1	0.666	0.058
Mst	0.379	0.711	1	0.413	0.666	1	0.049
Pers	\	\	\	0.058	0.058	0.049	1

Table 1. Correlations between indicators for papers and authors

CONCLUSION

The proposed approach explored how individual research in a field follows trends. This study provided a new quantitative analysis method for studying scholars' research interests and trend analysis. Future work includes optimizing the method to be applicable in a larger-scale dataset in a longer period, identifying and visualizing the exact groups in various behaviors.

REFERENCES

Chu, J. & Evans, J. (2021). Slowed canonical progress in large fields of science. *PNAS*, 118(41), e2021636118.

Greene, D., & Cross, J. (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis*, 25(1), 77-94. doi:10.1017/pan.2016.7

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

Wan, H., Zhang, Y., Zhang, J., & Tang, J. (2019). Aminer: Search and mining of academic social networks. *Data Intelligence*, 1(1), 58-76.