Identification of Research Topic Transition Characteristics Around Researchers' Career Peaks (Postprint)

Authors: Lixue Chen, Teng Guangqing, Lü Jing, Tuo Rui

Date: 2023-04-01T16:02:54+00:00

Abstract

[Purpose/Significance] Investigating researchers' career development and the evolution patterns of their research topics can not only reveal the intrinsic mechanisms underlying the development of scientific productivity, but also provide improved policy guidance and support for the advancement of science. [Method/Process] Based on data from representative disciplines in natural sciences, social sciences, arts, and humanities, we identify the career peaks of researchers. Building upon this, we employ the Top2Vec topic modeling method from natural language processing to identify research topics, using career peaks as a demarcation criterion for academic careers, and analyze the topic similarity and topic transition probabilities of research topics across different career stages. [Results/Conclusions] The results indicate that researchers across disciplines generally exhibit increased frequency of topic transitions after reaching their career peaks, while elite scholars, conversely, demonstrate greater specialization in their research topics after career peaks.

Full Text

Identification of Characteristics of Topic Change before and after Career Peak of Scientists

Chen Lixue, Teng Guangqing, Lü Jing, Tuo Rui
School of Information Science and Technology, Northeast Normal University, Changchun 130117

Abstract:
[Purpose/Significance] Exploring the career development of scientists and the changing patterns of their research topics can not only reveal the internal mechanisms of scientific productivity development, but also help provide better policy guidance and support for the advancement of scientific undertakings.

[Method/Process] Based on representative disciplinary data from natural sciences, social sciences, and arts and humanities, this study identifies the career peaks of scientists. Using the career peak as the basis for dividing scientists' academic careers, the Top2Vec topic modeling method from natural language processing is employed to identify research topics, and topic similarity and topic transition probability of research topics at different stages of scientists' academic careers are analyzed. [Result/Conclusion] The results show that scientists across disciplines generally change research topics more frequently after experiencing their career peaks, while elite scholars have more specific research topics after experiencing their career peaks.

Keywords: scientists; career peak; Top2Vec; topic change; topic similarity

Classification Number: G250.2
DOI: 10.13266/j.issn.0252-3116.2021.16.009

The study of career change patterns and topic transitions of scientists has long been a research hotspot in the field of library and information science, with particular attention paid to top scientific talents from both society and academia [1]. According to the Matthew Effect [2], outstanding achievements in a scientist's career can bring reputation and recognition, which often translate into tangible assets that in turn contribute to future career success. A recent study published in *Nature* also found that scientists' careers typically involve a "hot streak" period [3], during which individual scientists' performance far exceeds their normal level, with the most significant characteristic being that their research 成果 receive high attention (frequently cited). Although existing research has found similar peak periods in scientists' careers, few studies have deeply explored what specific changes occur in individual scientists' research work before and after these career peaks, particularly regarding how research topics change for scientists and elite scholars around these periods. In June 2019, the General Office of the Central Committee of the Communist Party of China and the State Council issued "Opinions on Further Promoting the Spirit of Scientists and Strengthening Style and Academic Construction" [4], which states: "Increase stable support for outstanding scientific and technological workers and innovation teams to accelerate cultivation and promote the healthy development of scientific and technological undertakings." From this perspective, studying the activity mechanisms of scientists, especially outstanding ones, also aims to provide better policy guidance and support for the further development of scientific undertakings. Therefore, it is necessary to conduct detailed exploration and analysis of the characteristics of research activities of scientists, especially outstanding researchers, in the context of implementing national science and technology development strategies.

Since knowledge development is continuous, fluid, and multi-disciplinary, changes in scientists' research topics reflect the evolving nature of information collection and knowledge transfer [5]. Additionally, with the rapid development of scientific knowledge in recent years, new problems and knowledge emerge constantly. Given this, the authors attempt to combine two dimensions—

scientists' career peaks and research topics—to analyze changes in research topics from the perspective of career peaks, using natural language processing (NLP) methods and selecting different disciplinary fields from natural sciences, social sciences, and arts and humanities, in order to gain clearer understanding and deeper insights into the changing characteristics of research topics for scientists and elite scholars before and after career peaks.

2 Related Research Status

Understanding the mechanisms of individual scientists' research activities and important milestones in their academic careers helps explore the dynamic patterns of scientific productivity. From sociological theory, young scientists as "marginal people" in academia have invested less in specific ideas or schools of thought, accumulated less reputation, and therefore need not worry excessively about losses from research failure, making them more likely to achieve results. Young scientists are also better at approaching old problems from new perspectives, with broader interests, more energy, and higher academic enthusiasm. Although they lack experience, their research is highly original. Older scientists, while superior in accumulated research experience, independent judgment, and handling contradictions, lack passion, produce many uninspired works, and are less likely to make major breakthroughs [6,7]. B.F. Jones et al. [8] found through studying Nobel laureates' careers that young people with ideas are more likely to make major breakthroughs in hard sciences.

Additionally, much research has examined scientists' career peaks and corresponding achievements [9-12]. While these studies attach great importance to scientists' career development, they lack unified definitions of career peaks and relatively single research perspectives, without focusing on what changes occur in research work accompanying career peaks. In the latest 2020 study, researchers found that Nobel laureates have more publications and higher citation rates early in their careers than other scientists, while also discovering a brief "Nobel Dip" phenomenon of declining research impact after winning the prize [13]. This suggests that interesting changes occur in scientists' specific scientific work after experiencing career peaks, with changes in research topics before and after career peaks becoming a forward-looking topic of academic concern.

A potentially forward-looking research topic may lead to high-impact research outcomes, which can not only improve scientists' reputation but also create research opportunities for the entire field. Given the impact of research topics on individual scientists' academic careers and on disciplines and innovation policies, there is an urgent need to adopt quantitative methods to understand how scientists' research topics change throughout their academic careers [14-16]. In recent years, scholars at home and abroad have focused on quantifying and simulating the evolution of research topics in scientists' academic careers [17-20]. Although frequent topic changes may bring risks of failure and decreased productivity, research also shows that a stable and focused research team helps

scientists maintain productivity but is not conducive to innovation [21,22].

Generally, scientists cannot maintain unchanged research topic content throughout their academic careers. Scientists' topic changes may result from trade-offs between conservatism and risk-taking [23]. A. Hoonlor et al. [24] analyzed journals and conference papers in computer science and found that scientists' research focus changes approximately every 10 years, with only a small proportion of researchers publishing year after year on the same topic long-term. A. Rzhetsky et al. [25] modeled disciplinary knowledge as networks and found that scientists increasingly pursue conservative research strategies in biomedicine, tending to explore local neighborhoods of central topics rather than making large-span topic transitions. T. Jia et al. [26] used classification codes in physics to find that physicists' research interests change dramatically from the beginning to the end of their academic careers. A. Zeng et al. [27] recently found that today's scientists switch between topics more frequently than earlier researchers, and high transition rates early in academic careers are associated with lower overall productivity.

In summary, academia has accumulated certain results on scientists' career peaks and research topic transitions respectively. However, few studies have connected individual scientists' career peaks with their topic transitions for analysis. Therefore, this study selects representative disciplines from natural sciences, social sciences, and arts and humanities to conduct in-depth research on the changing characteristics of research topics at different stages of academic careers for scientists and elite scholars, hoping to provide reference for revealing the mechanisms of scientific productivity development.

3 Theoretical Foundations

3.1 Research Topic Identification for Scientists

Identifying scientists' research topics mainly involves natural language processing (NLP) of their published literature to discover potential semantic structures in large document collections, also known as topic classification. Currently, the most widely used topic modeling methods include Probabilistic Latent Semantic Analysis (PLSA) [28] and Latent Dirichlet Allocation (LDA) [29]. Although these methods are popular in academic research, they have some drawbacks. For example, to achieve optimal model performance, preprocessing is often required, such as customizing stop word lists, stemming, lemmatization, and spending considerable effort presetting appropriate topic numbers. Additionally, most topic modeling methods rely on bag-of-words representation, ignoring word order and semantics. To overcome these drawbacks, this study adopts the Top2Vec [30] topic modeling method proposed in 2020 to conduct topic modeling on scientists' published literature for research topic identification.

Top2Vec, as a distributed topic vector model, utilizes semantic embeddings of documents and words to find topics. The number of dense regions discovered in semantic space is considered the number of prominent topics. Topic vectors

are calculated from dense regions of documents, where dense regions consist of very similar documents, and topic vectors are obtained by calculating the "centroid"—the arithmetic mean of all document vectors in the same dense cluster. The "centroid" can well represent the topic vector of a document dense region, and the words closest to this topic vector are those that can best describe it semantically. The resulting topic vectors are co-embedded with document and word vectors, with distances between word vectors representing semantic similarity. Top2Vec-generated topics have also been proven to be more informative than probabilistic generative models and contain more representative corpora. This model does not require stop word removal, stemming, or lemmatization preprocessing, and can automatically find the number of topics. Its main operation process is shown in Figure 1 [Figure 1: see original paper].

In Figure 1, the specific operation steps of Top2Vec topic modeling are as follows:

Create semantic embeddings. Use SentenceTransformer to create embedded document and word vectors. Figure 1(a) shows an example of a semantic space, where gray points are documents and hollow points are words. Words are closest to the documents they best represent, and similar documents are also close to each other. Use UMAP (uniform manifold approximation and projection) to create low-dimensional embeddings of document vectors. Document vectors in high-dimensional space are very sparse, and dimensionality reduction helps discover dense regions, where each point is a document vector. Use HDBSCAN to find dense regions of documents and calculate topic vectors. Hierarchical density-based spatial clustering (HDBSCAN) is applied to document vectors, using the number of clusters to replace the number of topics. HDBSCAN, as a density-based clustering method, does not use identified outliers to calculate the "centroid" and does not force every document to be assigned to a category, instead setting these documents not assigned to topic sets as outliers. In Figure 1(b), light gray points are outliers not belonging to specific clusters. "Centroid" calculation is performed for each group of document vectors belonging to dense clusters, generating a topic vector for each topic (black points in Figure 1(b)).

Keyword extraction based on C-TF-IDF. After completing topic clustering, it is necessary to explore how one topic differs from another based on content. A variant of TF-IDF [31] called C-TF-IDF is used for topic word exploration. C-TF-IDF is a class-based TF-IDF process, where C represents CLASS. It can extract generative characteristics of text documents based on their topic categories. Unlike traditional TF-IDF, C-TF-IDF is not used to compare word importance between different documents but treats all documents in a single topic category as a single document processing, with each category considered a very long document. The resulting C-TF-IDF scores can reflect the weight of important words in a topic. It can extract elements that make each topic unique relative to other topics. Formula (1) is as follows:

$$C\text{-}TF\text{-}IDF_i = \frac{tf_{ti}}{w_i} \times \log\left(\frac{m}{n_t}\right) \quad \text{Formula (1)}$$

In Formula (1), for each topic category i, the frequency of each word t is extracted as tf_{ti}, divided by the total number of words w_i in that topic, which is a normalized form of high-frequency vocabulary in the topic. Then the total number of documents m is divided by the total occurrence frequency of word t in all classes n, transformed into logarithmic form and multiplied by the previous term, thereby completing the identification of scientists' research topics.

3.2 Topic Similarity and Topic Transition Probability

This study selects two indicators—topic similarity score and topic transition probability—to measure changes in scientists' research topics. Similarity score can measure how extensively scientists have migrated topics during topic transitions; topic transition probability is used to determine the frequency of scientists' research topic changes. The study uses cosine similarity to calculate topic similarity, a method proven to be the most widely used semantic distance measurement method in current natural language processing.

Cosine Similarity algorithm judges the similarity between word vectors based on the cosine angle between two word vectors. The closer the cosine value is to 1, the closer the angle is to 0 degrees, indicating the two vectors are more similar. When the angle equals 0, the two vectors are equal. Formula (2) is as follows:

$$\text{Similarity} = \cos(\theta) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad \text{Formula (2)}$$

Where A_i and B_i represent the features of document word vectors contained in two types of topics, $\|A\|$ and $\|B\|$ are the L2 norms of the two word vectors, θ is the angle between word vectors A and B, and Similarity represents the final cosine similarity score. Based on completed topic identification, the above formula is used to calculate similarity between different topics. According to topic similarity scores, a similarity score matrix for individual scientists' research topics is constructed, as shown in Table 1 .

In Table 1, n represents the total number of documents, w_n represents the nth document of a scientist, and s_{ij} represents the similarity score between document i and document j based on cosine similarity. On this basis, Formula (3) is used to calculate the topic similarity score for all documents of an individual scientist during a certain period.

$$SIM_{au} = \frac{2\sum_{i,j=1}^{n} s_{ij}}{n(n+1)} \quad (1 \le i \le j \le n) \quad \text{Formula (3)}$$

In Formula (3), SIM_{au} represents the similarity score of all documents of an individual scientist during a certain period. The lower the similarity score of

a scientist's published paper topics during a period, the greater the span of research topics during that period. Additionally, the topic transition probability formula for individual scientists is as follows:

$$\text{SwitchProbability} = \frac{t_n - 1}{n - 1}, \quad n \neq 1 \quad \text{Formula (4)}$$

Where n represents the total number of papers published by a scientist, and t_n represents the number of topics contained in the scientist's documents. SwitchProbability represents topic transition probability. Higher topic transition probability indicates more frequent transitions between different research topics; lower topic transition probability indicates more specific research topics.

Additionally, considering the limitations of the original Top2Vec algorithm's TensorFlow-Text installation package on Windows systems, to make the research method more generalizable and reproducible, this study uses the Top2Vec-based topic modeling method under the PyTorch deep learning framework. Compared with the original Top2Vec modeling method, it not only retains the core of the original model but also has better system compatibility.

4 Research Methods and Process

4.1 Data Sources and Process Framework

To detect characteristics of research topic changes before/after the peak period of scientists from a multidisciplinary perspective (natural sciences, social sciences, arts and humanities), more factors need to be considered than in previous single-discipline studies. Single-discipline studies do not need to consider document quantity factors, but vastly different document quantities between disciplines in a multidisciplinary perspective may bias topic modeling and statistical results, hindering horizontal comparison between disciplines. For this reason, this study selects mycology, library and information science, and philosophy—three disciplines with roughly equivalent document quantities—as representatives of natural sciences, social sciences, and arts and humanities respectively. Using the Web of Science Core Collection as the basic data source, advanced searches were conducted with search formulas "SU = MYCOLOGY", "SU = INFORMATION SCIENCE & LIBRARY SCIENCE", and "SU = PHILOSOPHY". The retrieval date was November 1, 2020, with the time period from 1985 to present, document type limited to "Article", and language limited to "English", ultimately obtaining 158,446 documents. Among them, mycology has 43,000 documents, library and information science has 65,961 documents, and philosophy has 49,485 documents. On this basis, authors contained in the documents were extracted and grouped by discipline. According to ORCID identifiers, duplicate authors were manually verified without double counting, resulting in 266,388 authors total: 113,241 in mycology, 106,730 in library and information science, and 46,417 in philosophy.

Traditional wisdom holds that important awards and high-level achievements can serve as markers of scientists' career peaks. However, important awards are rare and insufficient for evaluating the broader scientific workforce. Moreover, awards emphasize academic recognition of previous achievements rather than scientific research itself peaking at the time of award. Therefore, academia primarily uses highly cited papers as the basis for identifying scientists' career peaks, especially when studying a single discipline over long time periods, mainly by setting uniform time periods (e.g., 10 years) for citation frequency [3,32]. Considering this study spans three disciplinary categories and paper half-life is influenced by multiple factors such as document type and discipline nature, using uniform time periods for citation frequency is inappropriate. Additionally, factors such as "sleeping beauty" papers are considered. Therefore, this study uses the paper with the highest absolute citation frequency as the marker of a scientist's career peak, and treats the publication year of the highest-cited paper as when the scientist reached their career peak.

Specific research work follows this process: Use the Top2Vec model for topic modeling of obtained literature and merge niche topics; Calculate topic similarity and topic transition probability within respective periods before/after the peak period, and comparatively analyze topic transition characteristics within each period before/after the peak for the overall scientist population; Screen elite scholars, calculate the difference in topic similarity and topic transition probability before and after the peak period for elite scholars, and analyze topic transition characteristics of elite scholars before and after experiencing career peaks.

4.2 Overall Topic Distribution Overview

The research work takes abstracts of literature in mycology (43,000 papers), library and information science (65,961 papers), and philosophy (49,485 papers) as the basic corpus data, using the Top2Vec-based topic identification method for topic modeling. During modeling, a threshold is preset to limit each topic cluster to contain no fewer than 20 documents. The distribution results of research topics from 1985 to present for the three disciplines are shown in Figure 2 [Figure 2: see original paper].

In Figure 2, (a), (b), and (c) show the visualization results of topic distribution in mycology, library and information science, and philosophy respectively. Nodes in the figure represent published documents, areas of different shades represent different topics, light gray represents outliers not belonging to any topic, and high-weight topic words for prominent topics are marked. These topic words represent key information of their respective topics.

Mycology obtained 56 topics through topic modeling. From Figure 2(a), it is clear that this discipline has very high topic dispersion (filled with numerous light gray outliers), lacks topics that aggregate massive numbers of documents, and has many outlier documents. Topics with relatively more documents in-

clude topic26, topic29, and topic51. Among them, high-weight topic words such as “patient,” “infection,” “treatment,” and “antifungal” indicate that topic26 mainly involves infectious disease research; high-weight topic words such as “protein,” “cerevisiae,” “gene,” and “mutant” reflect that topic29 concerns genetic research; high-weight topic words in topic51 including “describe,” “new,” “genus,” and “phylogenetic” indicate this topic relates to biological species research. Besides these, other topics contain fewer documents, but the total number of research topics in this discipline is large, demonstrating to some extent the complexity and diversity of mycology.

Library and information science topic modeling resulted in 46 topics. In Figure 2(b), the most prominent topics are topic14, topic44, topic26, topic39, topic40, and topic9. Based on high-weight topic words, each topic’s research focuses on medical informatics (topic14), knowledge organization (topic44), network information dissemination (topic26), bibliometrics (topic39), university libraries (topic40), and knowledge services (topic9). These six topics constitute the core research topics of library and information science. In addition to core topics, the discipline also has other niche topics containing fewer documents, most of which drift in peripheral areas outside central topics, indicating large differences between niche topics and core topics.

Philosophy topic modeling resulted in 41 topics, as shown in Figure 2(c). The visualization shows this discipline’s topics have high aggregation, with the vast majority of documents assigned to topic40 in the central region. High-weight topic words for this topic include “science,” “theory,” “knowledge,” and “history,” belonging to basic theoretical research in philosophy. Other prominent topics include topic9, topic17, and topic31. Comparing high-weight topic words shows research content mainly focuses on philosophy of medicine, philosophy of agriculture, philosophy of science, etc. This is also why such topic distributions are in edge areas outside large document clusters—these topics do not have high topic correlation with any other single topic.

Comprehensively, philosophy topics show the highest document aggregation and the fewest internal topics; mycology topics are the most dispersed with the most total topics; library and information science falls between the two, but its clustering results include the most topics containing large numbers of documents.

Considering that in the topic distribution results obtained by the Top2Vec topic modeling method in this study, many topics across disciplines contain numerous niche topics based on few documents. To reduce the impact of niche topics on experimental results, C-TF-IDF is used to compress topic numbers. By iteratively calculating cosine similarity between topics, comparing C-TF-IDF vectors between topics, merging the most similar vectors, and finally recalculating C-TF-IDF vectors to update original topic representations, the goal of merging edge topics containing few documents with the most similar topics is achieved. Ultimately, mycology research topics were reduced to 24, library and information science topics to 24, and philosophy topics to 20. The top 5 topics with the most documents in each discipline are shown in Table 2 .

Table 2 is arranged in descending order by the number of documents contained in each discipline's topics. However, discovering the characteristics and differences in topic distribution across the three disciplines is not the purpose of this study. The topic distribution characteristics and topic consolidation results obtained here serve only as the basis for subsequent identification of scientists' research topic transitions and migrations.

5 Research Results

5.1 Macro Analysis of Topic Similarity and Transition Probability

To ensure experimental validity, the previously obtained 266,388 authors were further screened. First, data with missing values were deleted, then scientists with no fewer than 5 published papers were selected. This resulted in 5,427 authors in mycology, 3,912 in library and information science, and 1,371 in philosophy. Using the publication year of the highest-cited paper as the criterion for determining career peak, with the publication year of the highest-cited paper as the Career Peak (CP), Formula (3) is used to calculate topic similarity scores before, during, and after the career peak. To further analyze scientists' topic transition characteristics, Formula (4) is used to calculate topic transition probabilities for scientists in different periods across disciplines. The results are shown in Table 3 .

Note: pre-CP represents before career peak; CP represents career peak; post-CP represents after career peak; similarity represents topic similarity; probability represents topic transition probability.

Table 3 records the average topic similarity scores of research outcomes published by scientists in different disciplines before reaching the career peak, during the career peak year, and after the career peak, as well as the probability of individual scientists' research topics changing. From the overall similarity of research topics across different disciplines, philosophy has the highest topic similarity scores ($ 0.840).$Simultaneously, this disciplinealsohasthelowesttopictransitionprobabilitycomparedtotheothertwod$ indicating the largest research topic span within stages for mycology researchers. Library and information science researchers' topic similarity scores across stages show their topic span is slightly smaller than mycology. However, from topic transition probability calculation results, library and information science researchers have greater probability of topic transition than mycology both before and after the peak period, with topic transition probability during the peak year also smaller than mycology.

The above analysis shows certain differences exist among different disciplines' scientists in topic similarity and transition probability. From the calculation results of average topic transition probability for scientists in different periods within disciplines, scientists' topic transition probability after the peak period is higher than before the peak period across all three disciplines. Mycology scientists' topic transition probability after the peak increased by 12.2% compared

to before the peak, library and information science by 21.9%, and philosophy by 25%. This result indicates that the overall scientist population changes research directions between different topics more frequently after experiencing career peaks. Of course, this also reflects from another side that before reaching career peaks, scientists have relatively high specificity in research topics. Before reaching career peaks, scientists tend to focus on what they are good at or specialized research during that period; after career peaks, scientists begin to have higher career freedom, no longer limited to relatively concentrated research topics, thus the frequency of topic transitions increases.

5.2 Topic Transition Characteristics of Elite Scholars Before and After Career Peak

Elite scholars in each discipline are typically the leading force for scientific and technological progress in their disciplines. While academia has already noted differences between elite scholars and ordinary scholars in academic careers and creativity [33], the government has also issued policies to increase encouragement and support for top scientific talents and outstanding scientific workers [4]. This part of the study further examines the topic transition characteristics of elite scholars before and after career peaks, hoping to provide scientific basis for national science and technology policy formulation and implementation. Currently, academia identifies elite scholars based on indicators such as quantity of scientific contributions (high publication) and academic recognition (high citation). In this specific research, scholars ranking in the top 1% in both publication quantity and average citation frequency per paper in each discipline are selected. To ensure universality of results, "flash-in-the-pan" researchers are excluded, ensuring that high-publication and high-citation researchers who have engaged in scientific research for no less than 10 years are analyzed as elite scholars in the field. According to these criteria, 170 elite scholars in mycology, 246 in library and information science, and 97 in philosophy were obtained.

If previous analysis focused on the magnitude and frequency of research topic transitions within each time period, this part focuses more on differences in research topic transitions before and after the career peak as a dividing line. Literature published by each elite scholar before and after their career peak is integrated into two long documents respectively, with topic modeling conducted separately, and Formula (2) used to calculate topic similarity scores before and after the peak. The distribution of topic similarity scores before and after career peaks for elite scholars is shown in Figure 3 [Figure 3: see original paper], with the horizontal axis representing the proportion of scholars and the vertical axis representing similarity score intervals.

Figure 3(a)(b)(c) show the distribution of research topic similarity scores for elite scholars in mycology, library and information science, and philosophy before and after reaching career peaks. Overall, elite scholars across disciplines still maintain high similarity with their pre-peak research topics after experiencing career peaks. The proportion of elite scholars in each discipline whose post-peak

research topics have similarity of 0.5 or above ($ $0.5) with pre-peak topics all reach over 97% (97.1%, 99.4%, 97.3%). Under the condition of topic similarity $ $0.7 before and after the peak, the proportions of elite scholars in the three disciplines are 83.9%, 80.1%, and 64.9% respectively. The topic similarity score intervals with the highest proportions of elite scholars in each discipline are [0.8,0.9), [0.8,0.9), and [0.7,0.8). This result indicates that most elite scholars in each discipline can maintain continuity in research topics after career peaks, and even when some degree of topic migration occurs, they still choose topics very similar to early research (high similarity scores).

Obviously, the above results do not completely match the previous analysis results for the overall scientist population. Therefore, the study further calculates the difference in topic transition probability before and after career peaks for elite scholars in each discipline. With the horizontal axis representing the proportion of scholars and the vertical axis representing difference intervals, the changes in topic transition probability before and after career peaks for elite scholars are shown in Figure 4 [Figure 4: see original paper].

In Figure 4, being closer to the dashed line centered means smaller changes in topic transition probability after career peaks compared to before. The upper region farther from the line indicates more increased topic transition probability after the peak (difference close to 1); the lower region farther from the line indicates more decreased topic transition probability after the peak (difference close to -1). The pie chart embedded in Figure 4 shows the proportions of elite scholars with increased (difference > 0), decreased (difference < 0), and unchanged (difference $= 0$) topic transition probability. From the embedded chart results, most elite scholars across disciplines show decreased topic transition probability after career peaks compared to before ($57.7\% > 39.4\%$, $63.6\% > 32.4\%$, $60.8\% > 28.4\%$). Among them, philosophy elite scholars show the most obvious pattern, with those having decreased topic transition probability (60.8%) more than double those with increased probability (28.4%). Moreover, Figure 4(c) also reflects that philosophy elite scholars have the largest magnitude of topic transition probability decrease (more scholars with differences near -1). This result indicates that elite scholars tend to engage in more specialized scientific research after experiencing academic career peaks than before.

6 Conclusion and Discussion

This study adopts a combination of bibliometrics and document topic modeling to explore career peaks and related research topic transition characteristics of scientists in mycology, library and information science, and philosophy. Based on the above analysis results, preliminary conclusions are drawn as follows:

(1) Scientists generally change topics more frequently after experiencing career peaks. In the analysis of the overall scientist population, although differences in topic similarity before and after career peaks are not obvious, the topic transition probability indicator shows clear differences be-

fore and after career peaks. Scientists across disciplines have higher topic transition rates after career peaks than before (see Table 3). This result indicates that for the overall scientist population, scientists who have not yet reached career peaks do not change research topics frequently, while after experiencing career peaks, scientists change research topics more frequently than before. Of course, this also reflects from another perspective that before reaching career peaks, scientists have relatively high specificity in research topics.

(2) Elite scholars have more specific research topics after experiencing career peaks. Topic similarity before and after peaks for elite scholars shows that most elite scholars have high similarity between pre-peak and post-peak research topics (see Figure 3), and post-peak topic transition probability is lower than pre-peak (see Figure 4). This result indicates that elite scholars show almost opposite characteristics to the overall scientist population: the more outstanding the elite scholars are in scientific research, the more they tend toward more specialized research directions after experiencing career peaks, and their research topics increasingly favor "ten years of sharpening one sword."

In today's era of rapid scientific and technological development, discovering and revealing patterns and characteristics in the development process of scientists' academic careers helps reveal mechanisms of scientific productivity development and promotes better scientific research policy formulation to guide scientists toward scientific and technological innovation. The study also has some limitations: selecting one discipline from natural sciences, social sciences, and arts and humanities respectively is insufficient to cover broader scientific research fields. Future research will include broader scientific fields and adopt more detailed analysis methods for deeper investigation.

References

[1] Zhou Jianzhong, Yan Hao, Sun Li. Research on the growth trajectory and influencing factors of Chinese scientific researchers' careers [J]. Science Research Management, 2019, 40(10): 126-141.
[2] MERTON R K. The matthew effect in science [J]. International journal of dermatology, 1968, 27(3810): 56-63.
[3] LIU L, WANG Y, SINATRA R, et al. Hot streaks in artistic, cultural, and scientific careers [J]. Nature, 2018, 559(7714): 396-399.
[4] Central Committee of the Communist Party of China, State Council. Opinions on Further Promoting the Spirit of Scientists and Strengthening Style and Academic Construction [EB/OL]. [2021-07-18]. http://www.gov.cn/zhengce/2019-06/11/content_{5399239}.htm.
[5] RUAN W, HOU H, HU Z. Detecting dynamics of hot topics with alluvial diagrams: a timeline visualization [J]. Journal of data and information science, 2017, 2(3): 37-48.
[6] Qiu Junping, Yu Houqiang. Comprehensive analysis of influencing factors

of scientists' golden age [J]. Journal of Intelligence, 2014, 33(3): 11-15, 5.
[7] COLE S. Age and scientific performance [J]. American journal of sociology, 1979, 84(4): 958-977.
[8] JONES B F, WEINBERG B A. Age dynamics in scientific creativity [J]. Proceedings of the national academy of sciences, 2011, 108(47): 18910-18914.
[9] SIMONTON D K. Career landmarks in science: individual differences and interdisciplinary contrasts [J]. Developmental psychology, 1991, 27(1): 119.
[10] SIMONTON D K. Age and outstanding achievement: what do we know after a century of research? [J]. Psychological bulletin, 1988, 104(2): 251.
[11] BRODETSKY S. Newton: scientist and man [J]. Nature, 1942, 150(3816): 698-699.
[12] STEPHAN P, LEVIN S. Age and the Nobel Prize revisited [J]. Scientometrics, 1993, 28(3): 387-399.
[13] LI J, YIN Y, FORTUNATO S, et al. Scientific elite revisited: patterns of productivity, collaboration, authorship and impact [J]. Journal of the royal society interface, 2020, 17(165): 20200135.
[14] JONES B F. The burden of knowledge and the "death of the renaissance man": is innovation getting harder? [J]. The review of economic studies, 2009, 76(1): 283-317.
[15] COKOL M, IOSSIFOV I, WEINREB C, et al. Emergent behavior of growing knowledge about molecular interactions [J]. Nature biotechnology, 2005, 23(10): 1243-1247.
[16] SINATRA R, DEVILLE P, SZELL M, et al. A century of physics [J]. Nature physics, 2015, 11(10): 791-796.
[17] PETERSEN A M, FORTUNATO S, PAN R K, et al. Reputation and impact in academic careers [J]. Proceedings of the national academy of sciences, 2014, 111(43): 15316-15321.
[18] PETERSEN A M. Quantifying the impact of weak, strong, and super ties in scientific careers [J]. Proceedings of the national academy of sciences, 2015, 112(34): e4671-e4680.
[19] Shi Qingwei, Qiao Xiaodong, Xu Shuo, et al. Author-topic evolution model and its application in research interest evolution analysis [J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(9): 912-920.
[20] Chen Lixue, Guo Siyue, Teng Guangqing, et al. Research on the focus and migration of scientists' research topics [J]. Digital Library Forum, 2019(12): 9-17.
[21] UZZI B, MUKHERJEE S, STRINGER M, et al. Atypical combinations and scientific impact [J]. Science, 2013, 342(6157): 468-472.
[22] GUIMERA R, UZZI B, SPIRO J, et al. Team assembly mechanisms determine collaboration network structure and team performance [J]. Science, 2005, 308(5722): 697-702.
[23] BOURDIEU P. The specificity of the scientific field and the social conditions of the progress of reason [J]. Social science information, 1975, 14(6): 19-47.
[24] HOONLOR A, SZYMANZKI B K, ZAKI M J. Trends in computer science research [J]. Communications of the ACM, 2013, 56(10): 74-83.

[25] RZHEETSKY A, FOSTER J G, FOSTER I T, et al. Choosing experiments to accelerate collective discovery [J]. Proceedings of the national academy of sciences, 2015, 112(47): 14569-14574.
[26] JIA T, WANG D, SZYMANZKI B K. Quantifying patterns of research-interest evolution [J]. Nature human behaviour, 2017, 1(4): 78.
[27] ZENG A, SHEN Z, ZHOU J, et al. Increasing trend of scientists to switch between topics [J]. Nature communications, 2019, 10(1): 1-11.
[28] HOFMANN T. Probabilistic latent semantic indexing [C]//Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 1999: 50-57.
[29] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of machine learning research, 2003, 3(1): 993-1022.
[30] ANGELOV D. Top2Vec: distributed representations of topics [EB/OL]. [2021-02-18]. https://arxiv.org/pdf/2008.09470.
[31] SALTON G, YU C T. On the construction of effective vocabularies for information retrieval [J]. Acm sigplan notices, 1973, 10(1): 48-60.
[32] SINATRA R, WANG D, DEVILLE P, et al. Quantifying the evolution of individual scientific impact [J]. Science, 2016, 354(6312): 596.
[33] FROSCH K H. Workforce age and innovation: a literature survey [J]. International journal of management reviews, 2011, 13(4): 414-430.

Author Contribution Statement:
Chen Lixue: Data collection and analysis, paper writing;
Teng Guangqing: Proposed research ideas, designed research plan, paper writing and revision;
Lü Jing: Data analysis;
Tuo Rui: Data analysis.

Identification of Characteristics of Topic Change before and after Career Peak of Scientists

Chen Lixue, Teng Guangqing, Lü Jing, Tuo Rui
School of Information Science and Technology, Northeast Normal University, Changchun 130117

Abstract:
[Purpose/significance] Exploring the individual career development of scientists and the transforming laws of research topics can not only reveal the internal mechanism of scientific productivity development, but also help provide better policy guidance and support for scientific undertakings. [Method/process] Based on representative discipline data from natural science, social science, art and humanities, this article identifies the career peaks of scientists. The career peak is used as the basis for dividing scientists' academic careers. The Top2Vec topic modeling method in natural language processing is used to identify research topics, and topic similarity and topic transfer probability of research topics at different stages of scientists' academic careers are measured. [Result/conclusion] The research results show that scientists in various disciplines generally change research topics more frequently after experiencing their career peaks, while elite

scholars have more specific research topics after experiencing their career peaks.

Keywords: scientists; career peak; top2vec; topic change; topic similarity

Note: Figure translations are in progress. See original paper for figures.