Intelligent Identification of Domain Development Paths Based on Thematic Evolution: A Case Study of Artificial Intelligence (Postprint)

Authors: Zhou Yuan, Zhang Chao, Tang Jie, Liu Yufei, Zhang Yutao

Date: 2023-08-27T00:00:00+00:00

Abstract

[Purpose/Significance] Identifying domain development paths is of great significance to technological innovation. However, existing methods such as expert interviews and citation analysis cannot adapt to the current situation of explosive growth in literature. To address this problem, this paper proposes a domain development path identification method based on topic evolution. [Method/Process] The method can automatically obtain data from the Aminer platform, construct a keyword-scholar matrix, and comprehensively employ KMeans++ and spectral clustering algorithms to identify research topics and related scholars. It establishes associations between different topics through similarity calculation, ultimately obtaining and visualizing the development path of the research domain. [Results/Conclusion] Through empirical analysis of the artificial intelligence domain, the results demonstrate that the proposed method can effectively reflect the evolution of domain research topics, help researchers quickly locate research hotspots and priorities in the domain, and enrich research methods related to domain development paths.

Full Text

Intelligent Identification of Domain Development Trajectories Based on Topic Evolution: A Case Study of Artificial Intelligence

Zhou Yuan[1], Zhang Chao[2], Tang Jie[3], Liu Yufei[4], Zhang Yutao[3]

[1]School of Public Policy and Management, Tsinghua University, Beijing 100084
[2]School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074

[3]Department of Computer Science and Technology, Tsinghua University, Beijing 100084
[4]The CAE Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088

Abstract

[**Purpose/Significance**] Identifying domain development trajectories is crucial for scientific and technological innovation. However, existing methods such as expert interviews and citation analysis cannot cope with the current explosion of literature. To address this problem, this paper proposes a novel method for identifying domain development trajectories based on topic evolution. [**Method/Process**] This method automatically retrieves data from the AMiner platform, identifies research topics and related scholars by constructing a keyword-scholar matrix using K-Means++ and spectral clustering algorithms, establishes associations between different topics through similarity calculation, and finally obtains and visualizes the domain development trajectory. [**Result/Conclusion**] Through empirical analysis of the artificial intelligence domain, the results demonstrate that this method can effectively reflect the evolution of research topics, help researchers quickly locate hotspots and key focuses in the field, and enrich research methods related to domain development trajectories.

Keywords: domain development trajectory; topic evolution; K-Means++; spectral clustering; artificial intelligence

2 Related Research

2.1 Topic Evolution

Topic evolution, also known as topic transition, typically employs data mining techniques to explore changes in topic content and intensity over time, as well as interactions between different topics. With the explosive growth of literature, topic evolution analysis faces challenges of large data volume and complex data types. How to quickly and accurately extract domain development 脉络 from massive amounts of data is a common concern for researchers and information professionals. Scholars have proposed numerous models to address this issue, with methods relevant to this paper falling into two main categories: clustering analysis-based approaches and probabilistic topic model-based approaches.

Clustering analysis-based methods for topic evolution are primarily applied in Topic Detection and Tracking (TDT) and bibliometrics. TDT defines topics as events occurring at specific times and places, focusing on topic evolution in news text streams, including both content and intensity evolution. Bibliometric methods include citation analysis and co-word analysis. Cui Lei et al. used citation co-citation clustering to study domain development history and summarized research topics through co-occurrence clustering of high-frequency words. Tang

Guoyuan et al. reviewed domestic research on co-word analysis-based topic evolution, dividing the analytical process into five steps: data source identification, evolution stage division, analysis object determination, co-word matrix construction and normalization, and topic evolution analysis. Liu Zhihui et al. proposed the Author Keyword Coupling Analysis (AKCA) model, which can discover implicit relationships between authors and identify domain topic changes through variations in keyword coupling strength as authors publish more papers. These three methods are relatively simple and applicable to different domains, with mature analysis tools available (such as CiteSpace and NEViewer). However, citation analysis suffers from time lag issues and faces problems of massive link counts and resource consumption; co-word analysis is highly sensitive to keyword selection, potentially yielding significantly different results based on word choice; and AKCA focuses on analyzing author collaboration relationships, indirectly obtaining research topics through authors' high-frequency keywords.

Probabilistic topic model-based methods for topic evolution have gained increasing attention in recent years. Li Xiangdong et al. used the LDA model and JS divergence to study changes in topic content and intensity over time. Ni Liping et al. employed the LDA model to identify technical topics in different time slices and formed domain development trajectories by clustering global topics using the AP clustering algorithm. D.M. Blei et al. proposed the Dynamic Topic Model (DTM), which assumes that topics continuously evolve over time, dividing documents into time slices based on publication time, extracting topics from each slice separately, and calculating similarity between topic distributions across time slices using KL divergence. Topic model-based methods can solve inherent time lag issues in citation analysis and uncover deep semantic information compared to co-word analysis. However, they only reveal statistical probability-level semantic relationships, require pre-specified topic numbers, and cannot dynamically adjust them.

Most existing topic evolution methods analyze internal document features (i.e., document content itself). To incorporate external features of academic literature (such as authors and journals) into topic analysis, scholars have proposed corresponding algorithmic models for different features. A typical method is the AT model (author-topic model) proposed by Rosen-Zvi, which adds author latent variables to the LDA model to obtain author-topic distributions. However, this model implicitly assumes each author has only one topic, which does not adapt to the reality of diverse and continuously changing author research topics. D.M. Blei et al. proposed the RTM model (Relational Topic Model), which achieves better word prediction and link prediction by jointly modeling document content and links between documents. However, RTM is somewhat complex when link prediction is not required. The above two topic models can better identify topics in different time slices, and topic associations can be achieved by calculating similarity between topics using KL distance or JS divergence, with final topic evolution results obtained after association filtering.

2.2 Path Visualization

To facilitate intuitive understanding of domain development trajectories, topic evolution requires visualization. Topic evolution and visualization are inseparable. Existing path visualization methods are numerous, among which Chen Chaomei developed CiteSpace based on Java, featuring citation analysis and temporal network visualization. S. Havre et al. introduced the ThemeRiver model, which reflects changes in topic intensity over time through the "river" width. Microsoft Research Asia proposed the TextFlow method, adding topic merging and splitting information to text analysis. This paper employs weighted Jaccard similarity for topic association and draws inspiration from the essence of the TextFlow method to design a visualization approach compatible with our path identification method.

In summary, domain research topic evolution involves multiple factors, such as research content evolution and scholar transitions. Since experts and scholars are the main body of scientific research, outstanding scholars often lead domain development. By mining scientific literature published by outstanding scholars, we can discover main research topics in scientific fields and connections between topics. Keywords reflect scholars' research content; when certain keywords are mentioned by a scholar in published papers, it may indicate certain associations between these keywords. If these keywords are mentioned by different authors, it may suggest that different authors recognize associations between these keywords. Compared with co-word analysis, constructing a keyword-author matrix with authors as the co-occurrence unit yields keyword clusters with deeper-level relationships. K-Means++ adapts to massive text clustering scenarios, and combined with spectral clustering, it can dynamically adjust the number of categories compared to topic models with fixed topic numbers, offering greater flexibility. Using similarity threshold methods to associate topics within adjacent time periods and visualizing topic evolution results using D3.js can clearly display domain development trajectories.

3 Methodology

Figure 1 [Figure 1: see original paper] shows the overall process of our method, mainly comprising four steps: data source and preprocessing, topic identification, topic association, and association network visualization.

3.1 Data Source and Preprocessing

The data source selected is the AMiner scientific database, which provides accurate domain expert recommendations through data mining and social network analysis methods. The platform is open-source, allowing convenient retrieval of required expert and paper data via API. This paper proposes a general keyword-expert-based method: AMiner is used to identify domain-related experts, and APIs are used to retrieve all papers published by these experts. The data preprocessing pipeline consists of three steps: keyword extraction, lemmatization,

and stopword removal.

First, keyword extraction is performed. Due to language characteristics, English literature does not require word segmentation. However, single words lack specific meaning; phrases are more meaningful semantic structures than individual words. This paper uses the RAKE algorithm to extract phrases of two or three words from titles and abstracts, with author-provided keywords directly added to the keyword list. In English, keywords typically consist of multiple words without punctuation and rarely contain function words such as "an," "this," or "but." The RAKE algorithm divides documents into several clauses based on punctuation, then splits sentences into phrases using stopwords as potential keywords. Each phrase's score is accumulated from its constituent words: score(w) = wordDegree(w) / wordFrequency(w), where score represents word w's score, wordDegree represents the degree of word w (degree increases by 1 whenever it co-occurs with another word), and wordFrequency represents the total occurrences of word w in the document.

Second, lemmatization is applied. Since English words have multiple forms, keywords require lemmatization to merge words with the same meaning but different forms. For lemmatization, this paper adopts the Stemming algorithm in NLTK.

Third, stopwords are removed. After lemmatization, stopwords must be eliminated: removing overly broad keywords such as "artificial intellig" and "data mine"; removing meaningless words such as "case studi" and "data sourc." After stopword removal, the final keyword list is obtained.

3.2 Topic Identification Based on Scholar Characteristics

To identify topics in different time periods, the time series must first be divided into several time slices of length L, with literature assigned to corresponding time slices based on publication time. Previous keyword clustering mostly used co-word matrices. This paper attempts to use scholars as features: keywords as row vectors and scholars as column vectors, using the K-Means++ algorithm to cluster keywords within a single time slice. The resulting keyword clusters can be considered research topics of that time slice. Keywords can be represented using the Vector Space Model (VSM), where each scholar is a dimension, mapping each keyword t to: v(tj) = (a_1, tjfi; …; ai, tjfi; …; an, tjfn), where ai (i = 1, 2, …, n) represents the i-th author, and tjfi represents the frequency of keyword tj in all articles by author i.

The keyword-author association matrix is as follows: Matrix = [t_1f_1 … $t_1$fi … $t_1$fn; tj f_1 … tj fi … tj fn; tm f_1 … tm fj … tm fn].

3.2.1 Topic Identification Using K-Means++ Algorithm K-Means++ is an improved algorithm based on traditional K-Means. K-Means uses distance as the criterion for classification, producing clusters with high intra-class similarity and low inter-class similarity. However, K-Means requires manual speci-

fication of initial cluster centers before clustering. The K-Means++ algorithm solves this drawback by selecting better cluster centers through maximizing distances between initial cluster centers (centroids). The K-Means++ algorithm features fast computation, convenient parameter tuning, and easy result interpretation, making it suitable for massive text clustering scenarios.

3.2.2 Spectral Clustering for Merging Adjacent Topics This paper uses graph structure-based spectral clustering algorithm to merge similar nodes within the same time slice, ultimately obtaining research topics of the domain within time slices. Both K-Means and K-Means++ algorithms face the problem of requiring pre-specified cluster numbers. In reality, even experts often cannot determine the appropriate number of categories, though they can specify an approximate range—the maximum number of topics within a time slice. Therefore, during initial clustering, experts determine the topic number range, after which spectral clustering performs secondary clustering on topics already identified by K-Means++, merging similar topics within the same time slice to obtain final topics.

The general spectral clustering algorithm process is: (1) Construct adjacency matrix. In spectral clustering, the adjacency matrix W (i.e., topic similarity matrix) must first be built. Weighted Jaccard similarity algorithm is used to calculate similarity between topic nodes: using keyword frequency as weights, calculating similarity between each topic and all topics one by one, ultimately obtaining the similarity matrix. (2) Let D be the degree matrix, calculate Laplacian matrix L = D - W, and normalize the Laplacian matrix using equation (4): LN = D^(-1/2) L D^(1/2). (3) Compute the eigenvectors v_1, v_2, …, v corresponding to the k smallest eigenvalues of LN, let V = [v_1, v_2, …, v], where matrix V's row count equals the number of topic nodes and column count equals k. (4) Use K-Means algorithm to cluster matrix V into s classes.

3.2.3 Identification of Topic-Related Authors To explore the most important topic changes in a domain, this paper selects core research scholars in the field and collects their published papers. If a keyword appears multiple times in a scholar's articles, it indicates that the scholar has conducted extensive research on the topic represented by the keyword and indirectly demonstrates the scholar's influence in that domain.

After two rounds of clustering, all keywords are assigned to several topics. Since in the keyword vector (a_1, tj f_1; …; ai, tj fi; …; an, tj fn), tjfi represents the number of times keyword tj is mentioned in scholar ai's articles, we can consider that the scholar corresponding to the maximum value in the vector has researched the keyword more extensively and has greater influence in the domain represented by the keyword. Therefore, by calculating each topic's centroid and examining the magnitude of each value in the centroid vector, scholars with greater influence on the topic can be identified.

3.3 Topic Association Based on Similarity Calculation

After two rounds of clustering, topics from different time windows are obtained. However, these topics are independent of each other. To further analyze topic evolution, topics identified from different time windows must be associated. The essence of topic evolution is the change in topic content, and topics in adjacent time slices can be associated through similarity calculation.

The topics obtained in this paper consist of a series of keywords, and similarity between different topics can be calculated using weighted Jaccard similarity to associate topics and obtain a series of topic pairs. Due to high similarity between these topic pairs, they can be considered to have topic evolution relationships. Similarity calculation may produce some invalid associations with unclear content continuation, so associated topics need to be filtered to highlight core topic evolution. This paper uses a threshold-based method for association filtering: let the sum of similarities between all topics in time period s and topic T^(s+1)_j be sum, and be the relative threshold. If sim(T^s_i, T^(s+1)_j) / sum < , the association can be considered invalid. The filtered topic pairs have strong associations and can represent topic evolution.

Weighted Jaccard similarity algorithm is used here to calculate similarity between topics across different time windows. Let the i-th topic in time period s be T^s_i, and the j-th topic in time period s+1 be $T^{(s+1)j}$. *The frequency of keyword t in time period s is: w(t, T^s_i) = Σ{t* ^{T}s_i} freq(t). The intersection frequency of keyword t in topics T^s_i and T^(s+1)_j is: w(t, T^s_i T^(s+1)_j) = min(w(t, T^s_i), w(t, T^(s+1)_j)). The union frequency is: w(t, T^s_i T^(s+1)_j) = max(w(t, T^s_i), w(t, T^(s+1)_j)). The similarity is: sim(T^s_i, $T^{(s+1)j}$ = Σ{t ^{T}s_i T^(s+1)_j} w(t, T^s_i $T^{(s+1)j}$ / Σ{t ^{T}s_i T^(s+1)_j} w(t, T^s_i T^(s+1)_j).

3.4 Association Network Visualization Design

This paper uses D3.js to visualize association results, enabling researchers and managers to intuitively understand domain development dynamics. Through similarity calculation, topics from different time windows establish associations, and visualization facilitates understanding and analysis of technical topic development.

Topic evolution involves four types of information: topic intensity, topic content, association relationships, and topic-related scholars. The visualization design includes two display elements: points and lines, where points represent topics on time slices and lines represent associations between topics. To display more information, the domain development trajectory generated by this method can be viewed via the Web with added interactivity: hovering the mouse over a topic point displays the Top 5 high-frequency words and 5 most relevant scholars; holding and dragging nodes is possible.

(1) Topic intensity. Topic intensity represents the popularity of a research

topic. Since topics identified in this paper are collections of keywords, topic intensity is measured by the total keyword frequency within the topic. In visualization design, topic intensity is represented by node width: node_{width}^s_i = w(T^s_i) × num_{doc}(s), where node_{width}^s_i represents the intensity of the i-th topic on time slice s, w(T^s_i) represents the total word frequency of that topic, and num_{doc}(s) is the number of papers in time slice s.

(2) Topic content. This includes topic names and keywords within topics. This paper uses the highest-frequency keyword in a topic as the topic name and displays the Top 5 high-frequency words in a pop-up box.

(3) Association relationships. Topics in adjacent time slices connect to form evolution relationships. Since the same keywords occupy different proportions in different topics, the connection lines have different thicknesses at both ends. Forward width is the line width displayed on time slice s+1: forward_{width} = node_{width}^(s+1)_j × sim(T^s_i, T^(s+1)j) / sim(T^(s+1)j). Backward width is the line width displayed on time slice s: backward*{width} = node*{width}S_i × sim(T^s_i, T^(s+1)_j) / sim(T^s_i).

(4) Topic-related scholars. The pop-up box displays the Top 5 scholars with influence in the research topic.

4 Experiments and Results Analysis

4.1 Case Selection and Data Processing

To verify the effectiveness of the domain development trajectory identification method, this paper takes "artificial intelligence" as the keyword and retrieves literature from AMiner via API. Artificial intelligence has a long development history. To identify the core changes from complex domain development, this paper selects the Top 100 experts in the field and retrieves all their paper data through API. Since the term "artificial intelligence" originated at the 1956 Dartmouth Conference, this paper limits the time span from 1956 to 2017, obtaining a total of 25,614 articles.

Based on Wang Liya's finding of a 5.5-year half-life for computer science papers and Tian Jinping's summary of five development stages of artificial intelligence, this paper divides literature into 11 time slices with 6-year intervals for convenience in mapping time slices to corresponding development stages, with 2016-2017 forming a separate time slice. The number of articles in each period is shown in Figure 2 [Figure 2: see original paper]. The number of publications increased slowly from 1956-1991, grew exponentially from 1992-2009, reflecting scholars' intensified efforts and rapidly increasing research interest in this field. The 2016-2017 time slice contains only two years and has fewer articles. Expert and paper data can be obtained via API (http://doc.aminer.org/en/latest/).

4.2 Topic Identification and Association

After retrieving literature, preprocessing is first performed. Keywords are extracted through the RAKE algorithm and lemmatized. The top 1,000 most frequent keywords from each time slice are selected, and a stopword list is constructed manually. To observe macro-level topic changes, the initial clustering K value can be set smaller, here defined as 10. Topics in each time slice are obtained through the K-Means++ algorithm. Similar topics are merged using spectral clustering, with the K value for spectral clustering also set to 10. After initial clustering, if the 10 topics in a time slice have low correlation, the number of topics in that time slice can remain 10 after spectral clustering. If some topics are highly correlated, they can be merged into one topic through spectral clustering, thereby achieving dynamic adjustment of topic numbers. Topic identification results are shown in Table 1 .

A total of 68 topics were identified across 11 time slices. Next, topic association is performed using weighted Jaccard similarity. Through experimentation, a similarity threshold of 0.2 yields the best display effect, with similarity greater than the threshold indicating an association relationship between topic pairs.

4.3 Domain Development Trajectory Analysis

After visualization processing, the artificial intelligence domain development trajectory is shown in Figure 3 [Figure 3: see original paper]. The trajectory graph contains two elements: points and lines, where points represent specific research topics on time slices, and lines represent associations between topics. From left to right, as time progresses, topics continuously emerge, die out, inherit, merge, and split. According to the theory of five development stages of artificial intelligence, this paper divides the 11 time slices accordingly: [1956-1961] as Stage 1, [1962-1967] and [1968-1973] as Stage 2, [1974-1979], [1980-1985], and [1986-1991] as Stage 3, [1992-1997] as Stage 4, and [1998-2003], [2004-2009], [2010-2015], and [2016-2017] as Stage 5.

Since the [2016-2017] time slice contains only two years and has fewer articles, and node width in the visualization design logic is positively correlated with the number of articles in the time slice, the topic evolution graph shows a "convergence" effect at the last time node. The artificial intelligence domain development trajectory graph can be accessed via Web (URL: http://118.24.155.51:8080/trend_{ai}/). Nodes can be dragged in the webpage, and hovering over a node displays its core 5 topic keywords and 5 main experts.

Given that Stage 3 plays a connecting role in AI's development history, this paper selects Stage 3 (1974-1991) for in-depth topic content analysis due to space limitations. The enlarged view of Stage 3 is shown in Figure 4 [Figure 4: see original paper]. Figure 4 intuitively shows Stage 3's research hotspots, including expert systems, neural networks, knowledge representation, natural language processing, and parallel processing.

4.3.1 Topic Intensity Change and Association Analysis Expert systems appear twice in Stage 3 with thick connecting lines, where expert systems in T2 are strongly related to knowledge representation in T3, indicating that expert systems integrated multiple previous research results and became the core research topic during this period. In reality, from the 1960s to the late 1970s, the DENDRAL mass spectrometry analysis system developed by Stanford University marked the emergence of expert systems, ushering AI research into a new domain. Neural networks appear in the T3 time slice, associated with decision support systems, natural language processing, and information extraction in T2, indicating that neural network research benefited from the development of these three disciplines. Knowledge representation appears in the T2 time slice. In the 1980s, Japan launched the fifth-generation computer development program capable of large-scale parallel processing, and research on knowledge engineering entered a prosperous period. Pattern recognition in the T2 time slice is weakly associated with research topics in adjacent time slices, appearing as an isolated node in the graph.

4.3.2 Topic Content Evolution Analysis Through two rounds of clustering, core research topics and topic-related keywords in each time window are obtained. The top 5 most frequent keywords in each topic represent the topic's detailed content for in-depth analysis. All topics in Stage 3 are shown in Table 2 . The model can identify research topics in each time slice. Table 2 shows that each topic has a clear direction, with keywords indicating the topic's research content. Taking neural networks in T3 as an example, the five most frequent words—neural network, information retrieval, intelligent system, bioinformatics, and biomedical research—indicate that early neural network research was closely related to biotechnology research, reflecting the fact that neural networks originated from human research on brain neurons. Intelligent systems illustrate neural networks' initial application scenarios.

Table 2 shows that research topic content continuously changes over time. Taking expert systems as an example, expert systems in T2 contain rule-based, spectrum, knowledge engine, and knowledge acquisition, indicating that some systems similar to those developed from mass spectrometry analysis collectively formed a new generation of expert systems. At this time, expert systems were rule-based and integrated knowledge. From T2 to T3, expert systems and knowledge representation in T2 merged to form a new generation of expert systems containing knowledge representation, knowledge-based systems, logic programs, and knowledge engines, with greater emphasis on knowledge. The subtle changes in expert systems show their evolution from rule-oriented to knowledge-oriented.

4.3.3 Topic-Related Scholar Evolution Analysis The model clusters based on scholars and keywords, with each cluster's centroid pointing to the core scholars of that topic, thereby revealing scholar changes during topic merging and splitting processes. Scholar changes can indirectly reflect changes in research directions within topics. The pop-up box shown in Figure 5 [Figure

5: see original paper] contains key topic information and core scholars.

Selecting scholars related to expert systems for further analysis, expert systems appear as a separate topic for the first time in the 1980-1985 time slice and last appear in the 1992-1998 time slice, appearing in 3 time stages total, as shown in Table 3 .

In the 1980-1985 time slice, B.G. Buchanan is a professor in the Computer Science Department at the University of Pittsburgh. He co-developed the first-generation expert system DENDRAL with scientists including E.A. Feigenbaum, a mass spectrometry analysis expert system. C. Djerassi extensively applied mass spectrometry analysis in chemical research, expanding expert systems' application scenarios. E.H. Shortliffe's research is medical-related, publishing numerous articles on medical consultation system design and expert system performance optimization during this time slice. Thus, expert systems in this time slice were highly specialized, aiming to solve specific problems in particular domains.

In the 1986-1991 time slice, H. Prade and D. Dubois's research relates to fuzzy theory. H. Prade published articles on improving original rule-based expert systems using possibility theory. P.R. Cohen, Y. Wilks, and S.C. Shapiro focused on information extraction and natural language understanding, with P.R. Cohen improving performance degradation issues in the GRANT expert system. The second-generation expert systems introduced information extraction, fuzzy theory, and uncertain reasoning techniques, offering greater generality and solving the over-specialization drawback of first-generation expert systems to some extent.

In the 1992-1998 time slice, N.R. Jennings's research relates to multi-agent systems for solving problems that single expert systems cannot address. E. Horvitz conducted research on decision theory and possibility models during this period. T.M. Mitchell, a core scholar in AI, published articles on knowledge extraction from the Internet. Core experts' research directions in this stage were relatively dispersed, indicating that traditional expert system research was declining, with content gradually shifting toward large-scale knowledge acquisition, knowledge representation, and multi-expert system collaboration.

4.4 Validation of Domain Development Trajectory Identification Effectiveness

Validating method effectiveness is a common challenge in topic evolution research. For topic content and intensity evolution, there are currently no universal standards to assess result validity, nor effective quantitative comparison methods across different topic evolution models. Beyond quantitative metrics, we are more concerned with whether identified topic evolution can truly reflect actual domain development. Therefore, an expert consultation meeting was convened to evaluate the results. The attending experts included two associate senior computer domain experts, one PhD student, and two master's students.

After evaluation, experts believe that the identified artificial intelligence domain development trajectory is generally accurate, with reasonable associations between topics across time slices and clear visualization, offering certain reference value for analyzing AI technology development. However, some issues exist: (1) Some topics appear at slightly deviated time nodes; (2) Some recent topics are missing, such as speech recognition and computer vision; (3) Topics have hierarchical relationships, but all topics are arranged in parallel in the graph.

Upon analysis, the author believes that temporal deviations may occur because knowledge diffusion requires time, leading to differences between expert judgments and data mining-based judgments. For missing recent topics, manual retrieval in the original data revealed relatively few relevant articles, possibly because these topics are less studied in academia but more in industry. To address this, other data sources (such as patents) need to be introduced to correct results. Regarding hierarchical relationships between topics, the author plans to introduce Wikipedia tree-structured knowledge in future work to consider keyword granularity information during clustering. The author also found that some results do not align with expert cognition, mainly due to deviations caused by single data sources. Therefore, future research will expand data sources and use multi-source data to optimize domain development trajectory identification.

References

[1] RADZIEMSKI L J. From LASER to LIBS, the path of technology development [J]. Spectrochimica Acta part B: atomic spectroscopy, 2002, 57(7): 1109-1113.
[2] WANG Wei. Research on technology forecasting combining Delphi method and technology roadmap: taking Taiyuan's "12th Five-Year" technology development forecast as an example [J]. China Science and Technology Forum, 2011(4): 103-107.
[3] XU Ji, ZHANG Weiguo, LUO Jun. Enterprise technology development path planning based on technology roadmap analysis and AHP [J]. Science of Science and Management of S.& T., 2010, 31(11): 103-107.
[4] PAN Ying. Research on key technology development path based on patent citation strength [J]. Information Studies: Theory & Application, 2014, 37(12): 71-75.
[5] XU Guannan, XIE Mengjiao, PAN Meijuan, et al. Evolution and prediction of 3D printing industry technology: research based on patent main path analysis [J]. Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition), 2016, 18(4): 77-85.
[6] HAN Yi, JIN Bihui. A new perspective on citation network structure analysis based on connectivity: main path analysis [J]. Studies in Science of Science, 2012, 30(11): 1634-1640.
[7] QIN Xiaohui, LE Xiaoqiu. Domain topic evolution research based on LDA topic association filtering [J]. New Technology of Library and Information Service, 2015, 31(3): 18-25.

[8] ZHAO Yingguang, HONG Na, AN Xinying. Application research of topic models in topic evolution methods [J]. Library and Information Service, 2014, 30(10): 63-69.
[9] HONG Yu, ZHANG Yu, LIU Ting, et al. A survey on evaluation and research of topic detection and tracking [J]. Journal of Chinese Information Processing, 2007, 21(6): 71-87.
[10] CUI Lei, WANG Xiaoning. Deep mining analysis of discipline topic evolution: taking general surgery as an example [J]. Journal of Medical Informatics, 2009, 30(8): 5-10.
[11] TANG Guoyuan, ZHANG Wei. Research progress and analysis of discipline topic evolution based on co-word analysis [J]. Library and Information Service, 2015, 59(5): 128-136.
[12] LIU Zhihui, ZHANG Zhiqiang. Author keyword coupling analysis method and empirical research [J]. Journal of the China Society for Scientific and Technical Information, 2010, 29(2): 268-275.
[13] WU Ruomei, KONG Yuefan. Comparative study of co-word analysis and co-citation analysis methods [J]. Information and Documentation Services, 2010(1): 26-29.
[14] LI Gang, BA Zhichao. Research on several issues in co-word analysis process [J]. Journal of Library Science in China, 2017, 43(4): 93-113.
[15] LI Xiangdong, ZHANG Jiao, YUAN Man. Research on topic evolution of scientific journals based on LDA model [J]. Journal of Intelligence, 2014(7): 115-121.
[16] NI Liping, LIU Xiaojun, MA Chiyu. Topic evolution analysis based on LDA model and AP clustering [J]. Computer Technology and Development, 2016, 26(12): 6-11.
[17] BLEI D, LAFFERTY J D. Dynamic topic models [C]//Proceedings of the 23rd international conference on machine learning. Pittsburgh: ACM, 2006: 113-120.
[18] ROSEN-ZVI M, GRIFFITHS T, STEYVERS M, et al. The author-topic model for authors and documents [C]//Proceedings of the 20th conference on uncertainty in artificial intelligence. Arlington: AUAI Press, 2004: 487-494.
[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): 912-919.
[20] CHANG J, BLEI D. Relational topic models for document networks [C]//Proceedings of the artificial intelligence and statistics. Clearwater Beach: JMLR.org, 2009: 81-88.
[21] LI Jie, CHEN Chaomei. CiteSpace: Text Mining and Visualization in Scientific Literature [M]. Beijing: Capital University of Economics and Business Press, 2016.
[22] HAVRE S, HETZLER E, WHITNEY P, et al. ThemeRiver: visualizing thematic changes in large document collections [J]. IEEE transactions on visualization & computer graphics, 2002, 8(1): 9-20.
[23] CUI W, LIU S, TAN L, et al. TextFlow: towards better understanding

of evolving topics in text [J]. IEEE transactions on visualization & computer graphics, 2011, 17(12): 2412-2421.
[24] LIU Ping, GUO Yuepei, GUO Yiting. Using author keyword networks to detect author similarity [J]. New Technology of Library and Information Service, 2013(12): 62-69.
[25] LI H, ABE N. Word clustering and disambiguation based on co-occurrence data [C]//International conference on computational linguistics. Montreal: Association for Computational Linguistics, 1998: 749-755.
[26] TANG J. AMiner: mining deep knowledge from big scholar data [C]//International conference companion on world wide Web. San Francisco: International World Wide Web Conferences Steering Committee, 2016: 373-373.
[27] ROSE S, ENGEL D, CRAMER N, et al. Automatic keyword extraction from individual documents [M]//BERRY M W, KOGAN J. Text mining: applications and theory. Chichester: John Wiley & Sons, Ltd, 2010: 1-20.
[28] ARTHUR D, VASSILVITSKII S. k-means++: The advantages of careful seeding [C]//Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms. Philadelphia: Society for Industrial and Applied Mathematics, 2007: 1027-1035.
[29] ZHOU Aiwu, YU Yafei. Research on K-Means clustering algorithm [J]. Computer Technology and Development, 2011, 21(2): 62-65.
[30] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and algorithm [C]//Advances in neural information processing systems. Vancouver: NIPS Foundation, 2002: 849-856.
[31] LIU Zhiwei. Research on similarity matrix in spectral clustering [J]. Modern Computer, 2010(15): 67-69.
[32] LYU Zeyu. History, current situation and future of artificial intelligence [J]. China Computer & Communication, 2016(13): 166-167.
[33] WANG Liya. Analysis of half-life of computer science discipline based on CNKI [J]. Library and Information, 2015(1): 100-105.
[34] TIAN Jinping. Overview of artificial intelligence development [J]. Science Mosaic, 2007(1): 230-232.
[35] LINDSAY R K, BUCHANAN B G, FEIGENBAUM E A, et al. DENDRAL: a case study of the first expert system for scientific hypothesis formation [J]. Artificial intelligence, 1993, 61(2): 209-261.
[36] SAMPLES, DJERASSI C. Mass spectrometry in structural and stereochemical problems [J]. Journal of the American Chemical Society, 1966, 88(9): 1937-1943.
[37] SUWA M, SCOTT A C, SHORTLIFFE E H. An approach to verifying completeness and consistency in rule-based expert systems [J]. AI magazine, 1982, 3(4): 16.
[38] FARRENY H, PRADE H, WYSSSE. Approximate reasoning in a rule-based expert system using possibility theory: a case study [M]. Paris: Laboratoire des Langages et systèmes informatiques, 1985: 407-414.
[39] KJELDSEN R, COHEN P R. Evolution and performance of the GRANT system [J]. IEEE Expert-intelligent systems and their applications, 1987, 2(2):

73-79.
[40] JENNINGS N R. On agent-based software engineering [J]. Artificial intelligence, 2000, 117(2): 277-296.
[41] PAEK T, HORVITZ E. Uncertainty, Utility, and misunderstanding: a decision-theoretic perspective on grounding in conversational systems [C]//Proceedings of the aaai fall symposium on psychological models of communication in collaborative systems. Cape Cod: AAAI Press, 1999: 85-92.
[42] CRAVEN M, DIPASQUO D, FREITAG D, et al. Learning to extract symbolic knowledge from the World Wide Web [J]. Coastal management, 1998, 31(2): 121-126.
[43] ZHOU H, YU H, HU R. Topic evolution based on the probabilistic topic model: a review [J]. Frontiers of computer science, 2017, 11(5): 786-802.

Author Contributions

Zhou Yuan: Responsible for framework design, paper revision, and writing guidance.
Zhang Chao: Responsible for literature review and paper writing.
Tang Jie: Responsible for framework design.
Liu Yufei: Responsible for viewpoint refinement and paper revision.
Zhang Yutao: Responsible for programming, related experiments, and paper revision.

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