Prophet-Based Prediction-Correction Model for Topic Intensity Evolution: A Postprint Empirical Study in the Stem Cell Field

Authors: Zhang Xin, Wen Yi, Xu Haiyun, Liu Zhongyu

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] Topic evolution plays a crucial role in detecting scientific and technological frontiers and formulating innovation strategies. [Method/Process] The topic evolution analysis process is decomposed into three key steps: topic representation, similarity association, and intensity evolution calculation. We propose a topic intensity evolution and prediction model that employs the LDA model for topic representation, introduces dimensions such as content similarity, co-occurrence similarity, and trend similarity for topic association calculation, and incorporates a Prophet-based prediction-correction model for forecasting topic evolution trends. An empirical analysis of evolution is conducted using the stem cell field as a case study. [Results/Conclusion] Experimental results demonstrate that when using the Logistic growth model for prediction, the R2Score for each research topic exceeds 0.90, indicating that the Logistic growth model in Prophet aligns well with the growth patterns of topics in this field and can effectively fit the evolution trends of topic intensity. The proposed topic evolution model offers valuable insights for topic distribution and evolution analysis within specialized domains.

Full Text

Prophet Prediction-Correction Topic Intensity Evolution Model: A Case Study in the Stem Cell Field

Zhang Xin, Wen Yi, Xu Haiyun, Liu Zhongyu

Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041

Abstract: [**Purpose/Significance**] Topic evolution analysis plays a crucial role in detecting technological frontiers and deploying innovation strategies.

[**Method/Process**] This paper decomposes the topic evolution analysis process into three key steps: topic representation, similarity correlation, and intensity evolution calculation. We propose a topic intensity evolution and prediction model that employs the LDA model for topic representation, introduces content similarity, co-occurrence similarity, and trend similarity for topic correlation calculation, and incorporates a Prophet-based prediction-correction model for forecasting topic evolution trends. The stem cell field is used as a case study for empirical evolution analysis. [**Result/Conclusion**] Experiments demonstrate that using the Logistic growth model for prediction yields R^2 scores above 0.90 for each research topic, indicating that Prophet' s Logistic growth model aligns well with the growth patterns of topics in this domain and can effectively fit the evolution trends of topic intensity. The proposed topic evolution model offers valuable insights for topic distribution and evolution analysis within specialized fields.

Keywords: topic evolution; topic similarity; time series; Prophet

Classification Number: G251

---

1. Research Background

Research topic evolution involves the mining, analysis, and visualization of the emergence, diffusion, and development processes of research topics. It enables intelligence analysts and science and technology managers to comprehensively and objectively grasp the patterns of innovation and development within a field, making it a foundational and core task in technological frontier detection, technology foresight, and roadmap development. A profound understanding and accurate grasp of the laws of scientific and technological innovation and evolution trends, along with systematic planning of new innovation pathways, are essential for technological frontier prediction and innovation strategy deployment. In the era of big data, the explosive growth of scientific literature has made deep, automated processing and mining of massive scientific documents through large-scale knowledge computing theories and methods the mainstream approach for research topic evolution.

Research topic evolution methods can be categorized into three types: qualitative research, quantitative research, and combined qualitative-quantitative approaches. Expert knowledge-based methods primarily fall under qualitative research, while citation and text mining methods belong to quantitative research.

Expert Knowledge-Based Methods. Traditional research topic identification in disciplines mainly relies on expert interpretation, employing methods such as expert interviews, Delphi method, TRIZ method, and morphological analysis [1]. These methods are highly subjective and costly, but due to the credibility of domain experts, they remain widely adopted and achieve the best

accuracy rates.

Citation-Based Methods. Since citation information effectively represents knowledge inheritance, it plays a crucial role in research topic discovery and evolution analysis. Representative methods include the citation main path approach used by N. P. Hummon [2], A. Martinelli [3], and L. Y. Y. Liu et al. [4], and the citation clustering approach employed by A. Pilkington et al. [5] and R. J. Lai et al. [6].

Text Mining Methods. With advancements in deep learning and natural language processing technologies, along with improved computing capabilities, text mining methods have become increasingly important in research topic evolution analysis. These methods can be further divided into keyword co-occurrence-based methods [7], syntactic structure analysis-based methods [7], and probabilistic topic model-based methods [8-10].

A comparison of the three topic evolution analysis methods is shown in Table 1 .

Table 1. Main Methods for Research Topic Evolution Analysis

Method Category	Principle	Advantages	Disadvantages
Expert Knowledge-Based	Constructing domain research topic evolution trends through expert interviews	Authoritative expert interpretation	Labor-intensive, high cost
Citation Analysis	Calculating document and topic similarity through citation relationships	Quantitative description of multiple citation relationships, facilitating discovery of topic inheritance and development	Narrow citation coverage, citation motivation interference, sometimes lagging

Method Category	Principle	Advantages	Disadvantages
Text Mining	Obtaining topic similarity through topic word distribution and distribution distance	Greatly promotes automation and efficiency of topic evolution analysis, aids topic relevance measurement	Dependent on computer processing, computational steps affect results

Current topic evolution analysis methods emphasize status analysis but neglect future prediction. Particularly, trend analysis in recent time slices is often inaccurate due to publication delays or incomplete data collection by publishers. We argue that recent time slice data has two key characteristics: (1) **High data value**—recent publications are most relevant to current research and best reflect recent topic distributions, making them too valuable to discard entirely; (2) **Data incompleteness**—recent publication data is incomplete, and using this incomplete data directly for display and prediction leads to incorrect analysis and forecasting results.

2. Methodology Framework

The overall research process for topic evolution in this paper is divided into two stages: the first stage is data processing and research topic identification, with topic representation and extraction as its core; the second stage is research topic trend analysis and visualization, with topic correlation and trend prediction as its core. The relationship between the two stages is that topic representation and extraction form the foundation and prerequisite for subsequent trend analysis—only with good topic representation can the subsequent trend analysis results be robust and interpretable—while topic trend analysis represents the goal and outcome, serving intelligence analysis tasks such as research situation analysis and technology decision-making deployment. The specific process is illustrated in Figure 1 [Figure 1: see original paper].

(1) Data Processing and Research Topic Identification. Data is retrieved from database providers using specific search strategies, then processed through deduplication, missing value handling, stop word removal, and stemming to form a cleaned domain corpus. Fields for analysis (keywords, titles, abstracts, or full text) are extracted and divided into time slices according to certain rules. Topic modeling methods described in Section 2.1 are used for topic extraction and semantic enhancement, yielding two matrices: document-topic relationships and topic-word relationships.

(2) Research Topic Trend Analysis and Visualization. For the extracted research topics, different similarity calculation methods are employed for topic

correlation. Topic intensity is calculated across different time slices to obtain corresponding time series. Time series analysis methods described in Section 2.3 are applied to data from complete and accurate stages for trend prediction, while incomplete recent data is used for prediction correction to obtain topic trends. Guided by research topic lifecycle theory and combined with domain experts, topic trends are analyzed and interpreted, then visualized as line charts, topic river maps, and other forms to display research topic intensity evolution.

2.1 Topic Representation Modeling

The LDA topic model, proposed by D. M. Blei et al. in 2003 [9], has become the most widely used model in topic extraction and representation due to its ability to effectively extract latent topics from documents. While many subsequent variants such as DTM [11] and TOT [12] have been proposed, these algorithms have higher computational complexity and are not easily integrated into intelligence analysis tools. The selection of optimal topic numbers can be guided by perplexity metrics [9], and non-parametric topic models like HDP [15] introduce hierarchical features for automatic selection, but they suffer from high computational complexity and suboptimal performance.

The LDA model is a Bayesian probabilistic model that assumes documents are composed of several latent topics, and topics are composed of words. Specifically, given a document collection D with M documents d_1, d_2, ···, d , where document m has length N , the document generation process of the LDA model is: (1) Sample the topic distribution for document d from a Dirichlet distribution with parameter α; (2) Sample the topic z , for the j-th word w , in document d from a multinomial distribution parameterized by ; (3) Sample the word distribution Φ_z , for topic z , from a Dirichlet distribution; (4) Sample the final word w , from the word multinomial distribution Φ_z , . Let the dictionary size be N_d and the number of topics be N_T (symbols are the same in the following sections).

The joint probability distribution of words and topics can be expressed as:

$$P(w, z|\alpha, \beta) = \prod_{i=1}^{M} \int P(\theta_i|\alpha) \prod_{j=1}^{N_i} \sum_{z_{i,j}} P(z_{i,j}|\theta_i) P(w_{i,j}|z_{i,j}, \beta) d\theta_i$$

Model parameters can be estimated using Formula (1). The original paper by D. M. Blei used the E-M method for parameter estimation, which was slow. I. Porteous et al. [13] proposed the Collapsed Gibbs Sampling method, significantly accelerating topic model training and facilitating practical applications. M. Hoffman et al. [14] proposed the Online Learning method for LDA models, using batch updates and merging for training, making topic training feasible for big data. The widely used Python Gensim toolbox employs this method.

Two issues require discussion for better practical application of topic models:

(1) Topic Model Parameter Selection. The main parameters to specify during training are α, β, and topic number K. Recent research typically uses indices like Topic Coherence for topic evaluation, such as D. Mimno et al. [16]. Some studies use fused features for topic number selection, such as Wang Tingting et al. [17]. We combine perplexity and coherence metrics for topic number selection. This approach is relatively fast, yields satisfactory extraction results, and requires fewer additional features. It provides only a recommendation for topic number, with the final decision requiring expert interpretation of extraction results.

(2) Topic Semantic Enhancement. Traditional topic models use word sets for topic representation, which are often difficult to interpret. To address this, scholars have proposed semantic enhancement methods such as TNG [18], CITPM [19], PhraseLDA [20], and Chunk-LDAvis [21], which use phrases for topic representation with stronger readability. We employ Bi-Gram for topic semantic enhancement, enabling extracted topics to contain both words and common Bi-Gram phrases. This method is fast, requires minimal manual intervention, and only needs one-time processing during data preprocessing without secondary replacement operations after extraction, offering higher performance efficiency for large-scale corpus extraction than other methods.

2.2 Topic Association

2.2.1 Topic Intensity After topic extraction, topic intensity calculation is required. Topic intensity is a statistical attribute of topics representing their degree of attention. Current calculation methods mainly include three approaches: document count-based, corpus probability-based, and text saliency-based. Sun Mengmeng et al. [22] compared these three methods and concluded that they yield consistent results for long texts, with the first method producing more significant results. Therefore, we adopt the first method for topic intensity characterization.

Topic Intensity Definition: In time slice u with document count D_u, the intensity of topic j is defined as the number of articles belonging to topic j:

$$ST_j = \sum_{d \in D_u} I(\text{topic}(d) = j)$$

2.2.2 Topic Similarity Measurement Methods After topic extraction, we aim to explore relationships between topics. In this paper, these relationships are described using topic similarity. Traditional topic similarity is calculated from the topic content distribution dimension. Additionally, we propose two new perspectives—topic co-occurrence and temporal trends—to measure similarity, introducing co-occurrence similarity and trend similarity metrics. We conduct consistency analysis among these three similarity measurement approaches.

(1) Content Similarity. Content similarity characterizes the structural similarity of research topics in terms of content. Specifically, it measures topic similarity using the similarity of topic word distribution representations. Various methods exist to represent distribution similarity, such as Kullback-Leibler (KL) divergence, Hellinger distance, Jaccard distance, and Jensen-Shannon (JS) divergence. JS divergence has the advantage of symmetry, making it more suitable for topic similarity calculation scenarios. We select JS divergence as the content similarity measure.

Let topic T_i (i [1, N_T]) have probability ϕ_{ik} for word w_k (k [1, N_d]) in the dictionary. The content similarity calculation formula is:

$$simContent(T_i, T_j) = 1 - JS(\phi_i || \phi_j) = 1 - \left[\frac{1}{2} KL(\phi_i || \frac{\phi_i + \phi_j}{2}) + \frac{1}{2} KL(\phi_j || \frac{\phi_i + \phi_j}{2}) \right]$$

where the KL function is KL divergence, calculated as:

$$KL(\phi_i || \phi_j) = \sum_{k=1}^{N_d} \phi_{ik} \log \frac{\phi_{ik}}{\phi_{jk}}$$

(2) Co-occurrence Similarity. Research topics have attribute features beyond content structure that can characterize topic similarity. We propose using the co-occurrence frequency of research topics in documents to represent co-occurrence similarity. For document d_m (m [1, M]), let the probability of topic T_i be _{mi} and topic T_j be _{mj}. The co-occurrence degree of these two topics in this document is min(_{mi}, _{mj}). The co-occurrence similarity between two research topics across the entire document collection is:

$$simCoocur(T_i, T_j) = \sum_{m=1}^{M} \min(\theta_{mi}, \theta_{mj})$$

(3) Trend Similarity. Trend similarity measures the similarity of different technical topics in temporal evolution trends. Each research topic' s intensity across different time slices constitutes a time series. We propose defining trend similarity through the similarity between these time series:

$$simTrends(T_i, T_j) = (1 + dist(T_i, T_j))^{-1}$$

where dist(T_i, T_j) is the distance measure between time series {ST_i}_U and {ST_j}_U formed by topics T_i and T_j. Time series similarity measurement methods include lock-step metrics (e.g., Euclidean distance, Mahalanobis distance) and elastic metrics (e.g., Dynamic Time Warping). DTW can overcome Euclidean distance limitations, support sequence shifting, and flexibly

handle multi-phase sequences, making it the most commonly used time series metric. The DTW distance calculation uses dynamic programming:

$$dist(T_i, T_j) = comDist(T_i^1, T_j^1) + \min(dist(R(T_i), R(T_j)), dist(T_i, R(T_j)), dist(R(T_i), T_j))$$

where R(T_i) represents the remaining sequence of T_i, and comDist(T^1_i, T^1_j) denotes the distance between the first time points of two sequences, which can use Euclidean distance in practice.

2.2.3 Consistency Among Three Topic Similarity Measurement Methods The three similarity methods measure topic similarity from different perspectives, requiring different data and computational complexities. Are the results from these three methods consistent? We use edit distance to measure the consistency of results from different topic similarity measurement methods. Using the similarity measurement methods described in Section 2.2.2, we obtain sequences ordered by similarity to a given topic from largest to smallest. By comparing the similarity of sequences generated by different topic similarity measurement methods, we can measure method consistency. We employ the edit distance proposed by Vladimir Levenshtein to measure sequence similarity, representing the minimum number of operations (insertion, deletion, or substitution) required to transform one sequence into another. Smaller edit distances between sequences generated by two similarity measurement methods indicate stronger consistency, and vice versa.

2.3 Topic Trend Prediction

Each topic' s intensity forms a time series. With one year as a time slice, incomplete data from recent time slices is first removed, and the preceding data is used for modeling. However, recent time slice data better reflects current topic trends, so this data is used for prediction correction and future trend forecasting.

Problem Description: In time slice u, the intensity of topic j is ST_{uj}. For each topic j, topic intensities across different time slices constitute a time series {ST_{1j}, ST_{2j}, ST_{3j}, ···, ST_(U-1)j}. We aim to predict the topic intensity $\hat{\text{ST}}$_{Tj} in time slice T (T > U).

(1) Basic Model. Current mainstream time series models include ARIMA and LSTM neural networks [23]. ARIMA is effective for short-term prediction, while LSTM performs better for long-term prediction. In this specific problem, with yearly time units and limited training data, LSTM neural network models struggle to converge.

In 2018, after Facebook open-sourced the Prophet forecasting tool [24], it quickly became popular for time series analysis. As of July 22, 2019, the project had

376 watchers, 8,888 stars, and 2,172 forks on GitHub. Prophet is an additive model assuming observed variables follow:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

where g(t) is the non-periodic growth trend term, s(t) is the periodic factor term, h(t) is the holiday factor term, and _t is an error term following a normal distribution. Compared to previous models, Prophet offers advantages including automation, interpretability, scalability, and fast training. For this study, research topics show clear growth trends with relatively few data points. Compared to the classical ARIMA model, Prophet better predicts growth trends and converges more easily than LSTM models requiring large training samples.

Prophet uses both saturated growth (logistic) and piecewise linear (linear) trend models. It also explicitly introduces change point detection. Letting C(t) represent the carrying capacity that changes over time, δ be the rate change vector at S change points, γ be the adjustment vector, and a(t) be an indicator vector {0,1}^S, the saturated growth model formula is:

$$g(t) = \frac{C(t)}{1 + \exp(-(k + a(t)^T\delta)(t - (m + a(t)^T\gamma)))}$$

The linear model growth formula is:

$$g(t) = (k + a(t)^T\delta)t + (m + a(t)^T\gamma)$$

In this study, we introduce the Prophet trend growth model for topic trend prediction. Specifically, we use yearly time slices, temporarily ignoring periodic and holiday factors by setting weekly_{seasonality}=False and daily_{seasonality}=False. With relatively small trend data volumes, we set the number of change points to a smaller value (3 in this study), using default values for other parameters.

(2) Prediction Correction. Since all research topic articles have the same collection time, we can approximate that each research topic has the same missing proportion in the final time slice. Based on this proportion, we propose correcting the values predicted by the Prophet model (see Formula 11), where ST_{Tj} is the actually observed topic intensity at time T, $\hat{\text{S}}$T_{Tj} is the topic intensity at time T predicted by the Prophet model, and T_j is the corrected topic intensity at time T:

$$ST^*_{Tj} = ST_{Tj} \times \frac{\sum_{j=1}^{N_T} ST_{Tj}}{\sum_{j=1}^{N_T} \hat{ST}_{Tj}}$$

After incorporating the correction model, the overall prediction process becomes a three-stage model: (1) Remove incomplete data and use the Prophet model for preliminary prediction; (2) Use Formula (8) to correct preliminary prediction results based on recent time slice data; (3) Conduct subsequent time slice topic intensity evolution trend prediction based on the prediction-corrected data.

3. Empirical Research

Stem cell and regenerative medicine research has brought revolutionary changes to disease treatment, including cancer, and has been selected nine times as one of the top ten scientific breakthroughs by *Science* magazine. It is also a current research hotspot in biomedical fields both domestically and internationally, with relevant projects repeatedly deployed in major national science and technology programs such as the National Key R&D Plan. Therefore, we selected the stem cell field for empirical research. Using the search query (TI=Stem Cells) in ISI Web of Knowledge in May 2019, we retrieved 43,469 articles. The number of articles by year is shown in Figure 3 [Figure 3: see original paper], revealing a trend of slow growth followed by rapid growth and then saturation. The apparent decline in the last two years is due to incomplete data collection.

3.1 Data Preprocessing (1) Keyword Extraction. In the data preprocessing stage, we first extract title and abstract fields from papers. Regular expression matching is then used to remove punctuation marks, numbers, email addresses, and other special characters from the original literature. The simple_{preprocess} tool in the gensim package is used for initial tokenization. Subsequently, the industrial-grade natural language processing tool SpaCy is employed for part-of-speech analysis, extracting nouns, verbs, adjectives, and adverbs as objects for topic extraction.

(2) Semantic Enhancement. Semantic enhancement can be performed either during preprocessing or after topic extraction. Following a similar approach to reference [17] and considering computational time complexity, we use Bi-Gram for semantic enhancement. The Phrases tool in gensim is used to extract Bi-gram phrases and add them to the original text, enabling extracted topics to contain more interpretable phrase information. This method is fast, requires minimal manual intervention, and needs only one-time processing during data preprocessing without secondary post-extraction replacement operations, offering higher performance efficiency for large-scale corpus extraction than other methods.

3.2 Topic Extraction in the Stem Cell Field Using the LDA model in Gensim with alpha set to 'auto' , we employ both perplexity and coherence metrics to guide topic number selection. Perplexity is calculated using the log_{perplexity} function, and coherence is calculated using the models.coherencemodel function in Gensim.

As shown in Figure 4 [Figure 4: see original paper], as the number of topics increases, the perplexity metric gradually decreases, stabilizing when there are about a dozen technical topics. The coherence metric reaches its best performance with 14-16 technical topics, then stabilizes or even declines. According to reference [16], coherence generally provides better evaluation than perplexity. Combined with expert interpretation of topics under different topic numbers, we determined the number of research topics to be 15. The content and structure of extracted topics are shown in Table 2 , where the first column' s topic content labels were provided after consulting domain experts, and the topic structure column shows each keyword and its weight in the topic model' s word distribution.

3.3 Topic Intensity Calculation and Correlation Analysis (1) Results of Three Similarity Measurement Methods. Using formulas (3), (4), and (5), we calculate similarity matrices for different measurement methods, representing similarity between research topics as heatmaps (see Figure 5 [Figure 5: see original paper]), where darker colors indicate higher similarity values.

In Figure 5, darker grayscale cells indicate stronger content similarity between the row and column topics. Diagonal elements represent self-similarity (value of 1). Content similarity is symmetric, shown by identical grayscale shades in symmetric cells about the diagonal. Cells for topic8-topic12 and topic10-topic13 are darker, indicating strong content similarity.

Figure 6 [Figure 6: see original paper] shows co-occurrence strength results, with diagonal elements representing self-co-occurrence strength (topic frequency). Co-occurrence strength is also symmetric. Figure 6 shows darker cells for topic8-topic13 and topic8-topic10, indicating high co-occurrence frequencies. Topics with high frequencies (e.g., topic8, topic10) also show relatively high co-occurrence strength with other topics.

Figure 7 [Figure 7: see original paper] shows trend similarity results, with diagonal elements representing self-trend similarity (value of 1). Trend similarity is symmetric, with identical grayscale shades in symmetric cells. Darker cells for topic5-topic6 and topic3-topic4 indicate similar trends among these topics.

(2) Consistency Analysis Results of Three Similarity Measurement Methods. Table 3 presents the consistency analysis results. The last column shows average edit distances across 15 research topics. Table 3 clearly shows: (1) Large edit distances between different similarity judgment formulas indicate weak consistency among content similarity, co-occurrence similarity, and trend similarity—topics with strong content/co-occurrence similarity do not necessarily share consistent development trends; (2) The consistency ranking among the three similarity methods is: (content, co-occurrence) > (co-occurrence, trend) > (content, trend).

3.4 Topic Intensity Evolution Analysis and Prediction We use three metrics to measure deviation between observed and predicted values, evaluating prediction quality:

RMSE (Root Mean Square Error):

$$RMSE(ST_i, \hat{ST}_i) = \sqrt{\frac{1}{U}\sum_{u=1}^{U}(ST_{ui} - \hat{ST}_{ui})^2}$$

MAE (Mean Absolute Error):

$$MAE(ST_i, \hat{ST}_i) = \frac{1}{U}\sum_{u=1}^{U}|ST_{ui} - \hat{ST}_{ui}|$$

R^2 Score (Coefficient of Determination): R^2 ranges from 0 to 1, with values closer to 1 indicating better fit.

$$R^2(ST_i, \hat{ST}_i) = 1 - \frac{\sum_{i=1}^{U}(ST_{ui} - \hat{ST}_{ui})^2}{\sum_{i=1}^{U}(ST_{ui} - \bar{ST}_{ui})^2}$$

Table 4 compares results from three trend prediction methods. The first column shows results from the ARIMA time series analysis model (using Auto-ARIMA tool), the second column shows LSTM time series analysis results, and the last three columns show Prophet model predictions. Each cell contains (RMSE, MAE, R^2 Score) values. The table reveals:

(1) For all research topics, Prophet models achieve R^2 scores above 0.90, outperforming ARIMA and LSTM models, demonstrating Prophet' s ability to effectively fit research topic evolution trends. The log-transformation approach does not improve prediction accuracy. Since topic distributions show clear growth trends (non-stationary series), direct ARIMA application performs poorly. With yearly time slices and limited data, LSTM models are inadequately trained, perform poorly, and are prone to overfitting. Prophet' s Logistic growth pattern aligns well with the domain' s topic growth patterns, and its fewer parameters make it easier to achieve good results.

(2) Different research topics exhibit inconsistent temporal evolution patterns. Most topics follow Logistic trends, but Topics 5, 7, 11, and 12 are better fitted with linear trends, possibly because these topics are in rapid growth phases and have not yet reached saturation.

Figure 8 [Figure 8: see original paper] compares original prediction and prediction-correction model results for the same technical topic. The left figure shows fitting and prediction results for 2018 raw data, while the right figure

shows results after prediction-correction. The prediction-correction model clearly produces more stable subsequent trend predictions with less volatility, better conforming to topic evolution growth patterns.

3.5 Visualization of Topic Intensity Evolution Analysis and Prediction Results Using the Prophet-based prediction-correction model, we fit models for all 15 research topics, with results shown in Figure 9 [Figure 9: see original paper]. The figure shows that most research topics follow model growth patterns, but outlier data—especially from recent time slices—significantly impacts subsequent predictions, increasing volatility (e.g., topic2, topic10). Outliers can also widen confidence intervals (e.g., topic6, topic15).

We visualize topic intensity evolution using topic river maps, overlaying trend analysis charts for all research topics in Figure 10 [Figure 10: see original paper]. Each grayscale band represents a research topic, with band width indicating topic intensity. Figure 10(1) shows raw data without prediction modeling. Figures 10(2) and 10(3) show results after introducing the Prophet-based prediction-correction model. In Figure 10(2), band width represents topic intensity, while in Figure 10(3), it represents relative topic intensity (the ratio of a specific topic' s intensity to the sum of all topic intensities in that time slice). Relative intensity better reflects structural changes among topics within the domain. The right-side boxes show post-2018 results. Figures 10(2) and 10(3) better conform to evolutionary development patterns, largely mitigating the impact of incomplete recent data on trend analysis.

Figure 10 clearly shows that stem cell research topics overall exhibit growth trends. However, Figure 10(3) reveals distinct differences among topics: Topic 2 (stem cells and disease diagnosis/treatment) shows the most rapid development, with its proportion in overall research gradually increasing as stem cells find applications in more disease treatments. Conversely, Topic 15 (hematopoietic stem cells) represents relatively mature theoretical research, with its proportion in overall research showing a declining trend as some mature technologies transition from theory to application.

Around 1990 and 2000, Figure 10(2) shows significant changes in growth trends for various research topics (similar to inflection points), while Figure 10(3) shows oscillations in topic distributions. These time points align with major discoveries such as James Thomson' s isolation of human embryonic stem cells in 1988, the U.S. listing stem cells as a top ten scientific breakthrough in 1999, and Japan' s stem cell program proposal in 2000. This suggests that major scientific discoveries may not only change topic growth trends but also alter topic structures.

4. Discussion and Conclusions

Research topic analysis forms the foundation for scientific and technological decision-making. Current topic trend analysis primarily focuses on status de-

scription, with insufficient research on regular pattern analysis and prediction. This paper proposes a research topic trend analysis and prediction method with a three-stage model: topic extraction and representation, topic correlation and similarity analysis, and topic trend analysis and prediction.

In the topic extraction and representation stage, we use the LDA model, which offers advantages like automatic extraction of document latent topic representation and remains a classic method. However, LDA has limitations in representation capability and interpretability, requiring domain expert interpretation. LDA also assumes consistent topic numbers across time slices—a common approach in practical topic evolution systems that offers computational efficiency but inadequately represents emerging topics. Future work could consider combining pre-trained models for improved research topic representation.

In the topic correlation and similarity analysis stage, we use conventional content similarity and propose co-occurrence similarity and trend similarity metrics. We explore relationships among these similarity measures, finding the consistency ranking: (content, co-occurrence) > (co-occurrence, trend) > (content, trend).

In the topic trend analysis and prediction stage, we introduce the Prophet model for intensity evolution trend analysis and prediction, comparing it with classical models like ARIMA and LSTM. To address incomplete recent data in intensity evolution, we propose a two-stage prediction-correction model. Experiments demonstrate that this model effectively fits current status and future evolution trends.

Model advantages, disadvantages, and future work:

(1) The stem cell field empirical analysis uses data reaching hundreds of thousands of records, with topics showing clear trend characteristics after time slicing, enabling time series model fitting. However, data coverage remains insufficient. Future work will consider additional typical domains for empirical validation and comparative analysis across different domains to further verify the proposed method' s applicability.

(2) In topic extraction and representation, while LDA is a classic method, its representation capability and interpretability need improvement. Future work will explore pre-trained models for enhanced research topic representation.

(3) In topic intensity trend prediction, the Prophet prediction-correction model offers high automation, easy implementation, and alignment with topic growth patterns. This study uses yearly time slices, finer than the typical 3-5 year slices in informatics research, better capturing trend features. However, it remains coarse-grained and doesn' t fully utilize Prophet' s rich periodic and holiday modeling capabilities. Future work will use larger datasets at finer granularity for trend prediction and analysis.

(4) Classifying research topics by their different growth patterns for emerging and hot topic mining is a worthwhile future direction. Inflection or oscillation points in topic evolution curves can be used to infer the emergence of major discoveries or disruptive technologies.

References

[1] Luo Wenxin, Wang Yuanyuan. Review of research methods for technology topic evolution [J]. Knowledge Management Forum, 2018, 3(5): 255-265.

[2] HUMMON N P, DEREIAN P. Connectivity in a citation network: the development of DNA theory [J]. Social networks, 1989, 11(1): 39-63.

[3] MARTINELLI A. An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry [J]. Research policy, 2012, 41(2): 414-429.

[4] LU L Y Y, LIU J S. A survey of intellectual property rights literature from 1971 to 2012: the main path analysis [C]//Proceedings of PICMET' 14 conference: portland international center for management of engineering and technology; infrastructure and service integration. Piscataway: IEEE, 2014: 1274-1280.

[5] PILKINGTON A, MEREDITH J. The evolution of the intellectual structure of operations management-1980-2006: a citation/co-citation analysis [J]. Journal of operations management, 2009, 27(3): 185-202.

[6] LAI R J, LI M F. Technology evolution of lower extremity exoskeleton from the patent perspective [J]. Key engineering materials. 2014, 625: 536-541.

[7] WANG Z Y, LI G, LI C Y, et al. Research on the semantic-based co-word analysis [J]. Scientometrics, 2012, 90(3): 855-875.

[8] Hu Zhengyin, Liu Chunjiang, Wei Ling, et al. Design and practice of domain patent technology mining system for TRIZ [J]. Library and Information Service, 2017, 61(1): 117-124.

[9] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of machine learning research, 2003, 3(1): 993-1022.

[10] Fan Shaoping, An Xinying, Shan Lianhui, et al. Research on topic evolution type and evolution path identification methods based on medical literature [J]. Information Studies: Theory & Application, 2019, 42(3): 114-119.

[11] BLEI D M, LAFFERTY J D. Dynamic topic models [C]//Proceedings of the 23rd international conference on Machine learning. New York: ACM, 2006: 113-120.

[12] WANG X, MCCALLUM A. Topics over time: a non-Markov continuous-time model of topical trends [C]//Proceedings of the 12th ACM SIGKDD inter-

national conference on Knowledge discovery and data mining. New York: ACM, 2006: 424-433.

[13] PORTEOUS I, NEWMAN D, IHLER A, et al. Fast collapsed gibbs sampling for latent dirichlet allocation [C]//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM, 2008: 569-577.

[14] HOFFMAN M, BACH F, BLEI D M. Online learning for latent dirichlet allocation [C]//Advances in neural information processing systems 23. Vancouver: Curran Associates Inc., 2010: 856-864.

[15] GRIFFITHS T L, JORDAN M I, TENENBAUM J B, et al. Hierarchical topic models and the nested Chinese restaurant process [C]//Advances in neural information processing systems. Vancouver: Curran Associates Inc., 2004: 17-24.

[16] MIMNO D, WALLACH H M, TALLEY E, et al. Optimizing semantic coherence in topic models [C]//Proceedings of the conference on empirical methods in natural language processing. Edinburgh: Association for Computational Linguistics, 2011: 262-272.

[17] Wang Tingting, Han Man, Wang Yu. Optimization of LDA model and topic number selection: a case study of scientific literature [J]. Data Analysis and Knowledge Discovery, 2018, 2(1): 29-40.

[18] WANG X, MCCALLUM A, WEI X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval [C]//IEEE International Conference on Data Mining. Piscataway: IEEE, 2007: 697-702.

[19] LI B, WANG B, ZHOU R, et al. CITPM: A cluster-based iterative topical phrase mining framework [C]//International conference on database systems for advanced applications. Dallas: Springer International Publishing, 2016: 197-213.

[20] Zhang Qin, Zhang Zhixiong. Research on topic phrase mining method based on PhraseLDA model [J]. Library and Information Service, 2017, 61(8): 120-125.

[21] Liu Ziqiang, Xu Haiyun, Yue Lixin, et al. Research on core technology topic identification method based on Chunk-LDAvis [J]. Library and Information Service, 2019, 63(9): 73-84.

[22] Sun Mengmeng. Research on topic extraction algorithm based on noun phrase extraction and term weight analysis [D]. Hangzhou: Zhejiang University, 2014.

[23] GRAVES A. Supervised sequence labelling [M]//Supervised Sequence Labelling with Recurrent Neural Networks. Berlin: Springer, 2012: 5-13.

[24] TAYLOR S J, LETHAM B. Forecasting at scale [J]. The American statistician, 2018, 72(1): 37-45.

Author Contributions

Zhang Xin: Proposed research ideas, wrote the paper.
Wen Yi: Proposed research questions, revised the paper.
Xu Haiyun: Revised the paper.
Liu Zhongyu: Evaluated model results.

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