When Research Topic Trend Prediction Meets Fact-Based Annotations

Jiachen Wang · Jiajie Xu · Wei Chen · Lei Zhao

Received: 16 June 2022 / Revised: 2 September 2022 / Accepted: 26 September 2022 / Published online: 12 October 2022
© The Author(s) 2022

Abstract

The unprecedented growth of publications in many research domains brings the great convenience for tracing and analyzing the evolution and development of research topics. Despite the significant contributions made by existing studies, they usually extract topics from the titles of papers, instead of obtaining topics from the authoritative sessions provided by venues (e.g., AAAI, NeurIPS, and SIGMOD). To make up for the shortcoming of existing work, we develop a novel framework namely RTTP (Research Topic Trend Prediction). Specifically, the framework contains the following two components: (1) a topic alignment strategy called TAS is designed to obtain the detailed contents of research topics in each year, (2) an enhanced prediction network called EPN is designed to capture the research trend of known years for prediction. In addition, we construct two real-world datasets of specific research domains in computer science, i.e., database and data mining, computer architecture and parallel programming. The experimental results demonstrate that the problem is well solved and our solution outperforms the state-of-the-art methods.

Keywords Research topic trend prediction · Citation analysis · Transformer

1 Introduction

We have witnessed the unprecedented growth of publications in recent years, such as the papers published in AAAI, IJCAI, and NeurIPS, which is of great convenience for researchers to keep up with the status of research topics they are interested in. However, the overwhelming amount of academic papers also leads to the waste of time following the development of research topics. Therefore, tracing the evolution of research topics is an interesting and significant task. The studies for this problem can be divided into the following two categories. The first focuses on analyzing how research topics have changed in the past [1–5]. The second concentrates on the prediction of the future trend of research topics [6, 7]. Especially, inspired by the senior performance of deep learning techniques in modeling sequential information, many corresponding methods are proposed in recent years [8–10] to predict the research topic trend.

The analysis of research topic trend is essentially the study on the popularity of corresponding publications. To quantify the influence of a particular paper, one of the most intuitive and commonly adopted measures is the number of citations, a.k.a. citation count, which has been widely used as the basis of many scientific influence indicators such as h-index [11] and impact factor [12]. With this consideration, we adopt the integrated features of citation counts of papers to quantify the popularity of a certain research topic. Another difference from existing studies is that we obtain the research topics from the authoritative sessions provided by venues, rather than extract research topics from the titles of papers. Despite the importance of titles, there are lots of cases where the titles are weakly related to the research topics of the papers. For instance, “What makes a chair a chair?” [13], published in the top computer vision conference CVPR in 2011. We can hardly recognize the research topic from the title, which reflects that extracting research topics from titles of papers is not always effective. On the other hand, the authoritative sessions provided by various venues are the summarized experience of many experts and are highly related to the topics of accepted papers, such as Data Integration and Cleaning given by SIGMOD 2018[1]. Although the sessions may change slightly every year, extracting trend information from these fact-based annotations is of great significance for understanding the evolution...
of research topics. Unfortunately, none of the existing works to analyze and predict research topic trend take full advantage of this feature.

To predict the research topic trend more accurately, we propose a unified framework entitled RTTP, which consists of an innovative topic alignment strategy TAS and a deep learning-based enhanced prediction network EPN. Specifically, we formulate the task as a regression problem and construct the trend sequences from earlier years (e.g., 5 years) to predict the future trend of various research topics. TAS takes both global and local semantic information into account to obtain the detailed contents of research topics in each year. EPN is developed to capture research trend development and mutual correlation among research topics, and further predict the future trend of research topics.

To sum up, the contributions of this paper are as follows:

- To the best of our knowledge, we are the first to study the research topic trend prediction problem based on fact-based annotations, which provides a new perspective for researchers to keep up the development of research topics.
- We formally define the problem as a sequence prediction problem and propose a unified framework called RTTP. The framework consists of a topic alignment strategy TAS that is designed to obtain the detailed contents of research topics in each year, and a deep learning-based prediction network EPN, which is developed to capture potential trend information of known sequence.
- We conduct experiments on two real-world datasets to investigate the effectiveness of the proposed framework RTTP, and provide some in-depth analysis. The experimental results demonstrate the higher performance of RTTP compared with those of baselines.

The rest of this paper is organized as follows. Section 2 presents the related work including research topic trend analysis and citation count prediction. The preliminary and problem definition are introduced in Sect. 3. The proposed framework is introduced in Sect. 4. Section 5 reports the experiment details, followed by the conclusion and future work in Sect. 6.

2 Related Work

2.1 Analysis of Research Topic Trend

Analysis of research topic trend aims to reflect the popularity of various research topics. Some studies on research topic trend focus on exploring and understanding how the research topics changed in the past [1–5]. Despite the superior performance of the above methods, they lack the ability to predict future trends. To satisfy the prediction and analysis for future topics, extensive studies have emerged on traditional machine learning (ML) methods and deep learning (DL) models.

Some effective prediction methods based on machine learning algorithms have been proposed in the field of research topic trend. Charnine et al. employ the CatBoost method to realize long-term prediction of research trending topics [14]. To determine persistent and emerging research topics in scientific literature, Balili et al. carefully design a framework that leverages SVM, XGBoost, and Logistic Regression [7]. In addition, Abuhay et al. first utilize the non-negative factorization topic modeling method to discover topics, and then utilize the ARIMA prediction method to predict time series [6]. By constructing a temporal scientific knowledge network according to INSPEC controlled indexing, Behrouzi et al. transform the process of trend prediction into a link prediction problem in keywords networks [15]. To solve the problem, they propose a machine learning-based link prediction algorithm and fuse various features such as topology information and nodes clustering coefficient.

There is another stream of analysis for research topic trend focusing on deep learning model, which has shown excellent performance for fitting complex functions. Especially, considering that recurrent neural network (RNN) and its variants have been well applied in various sequence modeling tasks [16–18], Chen et al. [19] propose a gated recurrent unit (GRU) based model to predict the trending topics of mutually influenced conferences, which can capture the sequential properties of research evolution in each conference and discover the dependencies among different conferences simultaneously. Xu et al. [9, 10] follow Chen’s work and propose models based on long short-term memory (LSTM) networks to address research topic trend prediction influenced by peer publications. Taheri et al. [8] utilize fields of study from the Microsoft Academic to predict the upcoming years’ computer science trends based on LSTM. However, most existing studies use title words to represent research topics [9, 10, 14, 15, 19], which is extremely limited and unconvincing. The reason is that the titles of many papers are very unconventional and even have little correlation with the research topic of the paper. In contrast, sessions provided by various venues are summarizing the expression of papers’ research topics, which makes the study on research topic trend analysis convincing and meaningful.

2.2 Citation Count Prediction

A variety of criteria exist in the literature for evaluating the influence of a scientific paper, but one of the most important evaluation metrics is the citation count which records the number of citations to the considered paper [20–22].

Springer
Moreover, citation count has been used as the basis of many other metrics such as h-index [11], g-index [23], impact factor [12] and other evaluation metrics for journals, conferences, researchers or other research institutes [24, 25]. In recent years, citation count prediction has been widely studied. For example, Ru et al. utilize several fundamental characteristics of papers to predict the citation count as the popularity degree of each paper in the future [26, 27]. Ma et al. applied a vanilla RNN-based Sequence-to-Sequence (Seq2seq) model to predict citation counts of paper [28]. A novel citation prediction model that combined paper metadata text with early citations [29] is proposed, which extracts high-level semantic features by using Bi-LSTM to improve prediction performance. To predict the in-text citation count from each structural function of a paper separately, the latest work proposes a novel fine-grained prediction model for citation counts [30].

### 3 Preliminary and Problem Definition

In this section, we introduce several preliminary concepts and then formulate the problem of research topic trend prediction.

**Definition 1 Raw Topic.** A raw topic \( r \) refers to one specific session provided by a venue, and the set of which is denoted as \( R = \{ r_1, r_2, \ldots, r_n \} \), which represents the raw topics of all considered years and venues in a specific research domain.

\( W = \{ w_1, \ldots, w_m \} \) represents the considered consecutive years, in which \( w_j \in W \) is a unique year (e.g., 2010).

**Example 1** The set of raw topics provided by ICDE 2021 is \( R = \{ \text{Data Integration and Cleaning}, \text{Graph Data Management}, \ldots, \text{Analysis and ML over Graphs} \} \), where “Data Integration and Cleaning” is a specific raw topic.

**Definition 2 Research Topic.** Each research topic \( t_i \) is essentially the set of clusters of semantically similar raw topics across years, obtained by Topic Alignment Strategy. In detail, the cluster of corresponding raw topics in year \( w \) of \( t_i \) is denoted as \( t_i^w \). \( P_i^w \) w.r.t. \( t_i^w \) is composed of a set of papers, and each paper \( p_k \in P_i^w \) is embedded by the citation counts of this paper.

**Example 2** As present Fig. 1, \( t_i^w \) is composed of a set of raw topics in \( R \), such as \( t_i^{2010} = \{ \text{Probabilistic Databases}, \text{Probabilistic and Uncertain Data} \} \). It represents the specific expression of research topic \( t_i \) in 2010.

**Definition 3 Topic Alignment Strategy.** Given the set of raw topics \( R \) of a specific research domain, Topic alignment strategy is to find a function \( \Theta \) to partition \( R \) into the following set across years,

\[ \Theta(R) \rightarrow T = \{ T_1, \ldots, T_m \} \]

\[ \text{Fig. 1 An example of research topic} \]

2 https://dblp.org/db/conf/icde/icde2021.html.
where \( T^w \) represents the set of all research topics in year \( w_j \), as presented in Fig. 1.

**Definition 4.** Research Topic Popularity Score Chain. Research topic popularity score \( S^w_{t_j} \) is defined as the sum of citation counts of papers appearing in \( P^w_{t_j} \), that is, the corresponding score of \( t_j \) is \( S^w_{t_j} \). Research topic popularity score chain \( S^w \) is defined as an ordered sequence of \( S^w_{t_j} \), i.e., \( S^w = \{ S^w_{t_j}, ..., S^w_{t_j}, ..., S^w_{t_j} \} \). \( S^w \) is the corresponding score chain of \( T^w \).

**Problem Formulation** Given a set of raw topics \( R \) and corresponding papers, \( T = \bigcup_{w_j=1}^{m} T^w \) is obtained by **Topic Alignment Strategy**. Based on the research topic popularity score chains of \( m \) years \( S = \bigcup_{j=1}^{m} S^w \), we extract continuous time steps \( S' = \bigcup_{p=1}^{q} S^w \) to represent the known research topic trend, where \( 1 \leq p < q < m \). The object is to predict the future research topic popularity score chain \( S^{w\eta+1} \) in year \( \eta+1 \), i.e.,

\[ S' \xrightarrow{\sigma} S^{w\eta+1} \]

where \( \sigma \) is a mapping function.

**4 Proposed Framework**

The authoritative sessions provided by venues are the summarized experience of many experts and are highly related to the topics of accepted papers. This motivates us to propose a unified framework RTTP that can generate annual research topic trends from fact-based annotations and predict the future research topic trend with an enhanced deep neural network. The overview of our proposed framework is presented in Fig. 2. The framework consists of two parts: topic alignment strategy TAS and enhanced prediction network EPN.
4.1 Topic Alignment Strategy

To track research evolution, we design an innovative topic alignment strategy taking global and local information into account. We first construct a set of research topics, in which each research topic contains all semantically similar raw topics for the years in W. Then research topic trend for each year is presented as the specific expression of all research topics for that year.

4.1.1 BERT-as-service for Global Information

BERT is an advanced pre-trained model for language embedding and outperforms in understanding the semantics of user searches due to the contextual information. Raw topics need to be embedded into vector space to measure the semantic similarity while BERT can obtain higher quality semantic representations. BERT-as-service [31] hosts BERT as a service with the advantages of low latency and high scalability. Thus, the semantic similarity can be measured as a service with the advantages of low latency and high scalability. Therefore, the semantic similarity can be measured as a service with the advantages of low latency and high scalability.

\[ D = \| e_r - e_o \| = \sqrt{\sum_{i=1}^{d} (e_{r_i} - e_{o_i})^2} \]  

where \( e_r, e_o \in \mathbb{R}^d \) are the embedding vectors of \( r \) and \( o \) through BERT-as-service, \( e_r, e_o \) are the values of \( j \)-th dimension of \( e_r \) and \( e_o \), respectively.

4.1.2 Word Feature for Local Information

Local feature considers the effect of common words from two raw topics on alignment. Inverse Document Frequency (IDF) [32] measures the universal importance of a word. The more raw topics containing one word, the lower score for that word, which explains the word is ubiquitous. On the contrary, the fewer raw topics, the higher score of the word, which indicates the unique contribution of the word to a particular term.

4.1.3 Strategy Implementation

We design a variant of K-means [33] algorithm covering global and local factors above. \( Z = \{ z_1, \ldots, z_r, \ldots \} \) represents the set of unique words appearing in \( R \) after removing stop words. \( Y = \{ y_1, \ldots, y_s, \ldots \} \) corresponding to \( Z \) represents the set of IDF scores for each word, and \( \tilde{Y} = \{ \tilde{y}_1, \ldots, \tilde{y}_s, \ldots \} \) is the normalized representation of \( Y \), where \( y_s \) and \( \tilde{y}_s \) are defined as:

\[ y_s = \log_{10} \left( \frac{n}{\sum_{r=1}^{n} \phi(r)} \right), \quad \tilde{y}_s = \frac{y_s - y_{\min}}{(y_{\max} + \Delta y) - y_{\min}} \]  

\( \phi \) is a function, if \( z_s \in r, \phi(r) = 1, \) else, \( \phi(r) = 0 \). \( y_{\min} \) and \( y_{\max} \) are the minimum and maximum values in \( Y \), respectively. \( \Delta y \) is used to prevent the denominator being zero. The distance between raw topic \( r \) and \( o \) is updated as:

\[ Dis = \left( 1 - \sum_{z_s \in (r \cap o)} \tilde{y}_s \right) \cdot D \]  

where \( D \) and coefficient \( \left( 1 - \sum_{z_s \in (r \cap o)} \tilde{y}_s \right) \) denote global and local information of TAS, respectively.

If pairwise raw topics have no common words (i.e., \( \sum_{z_s \in (r \cap o)} \tilde{y}_s \) equals to zero), the distance \( Dis \) depends entirely on global information, otherwise, we take IDF scores of common words as the coefficient. Considering the quality of common words, if the score of one word is high, the word only belongs to certain raw topics. When two raw topics contain the word at the same time, we hold that the similarity of the two raw topics is high. Considering the quantity of common words, the more words in common, the more similar the two raw topics are. The number of clusters (also the number of research topics) \( l \) is the average of the number of raw topics available among all considered years and venues. We randomly select \( l \) raw topics as the initial cluster centers and regard the nearest raw topic object as new cluster center at each iteration.

To get the specific expression of each research topic for different years, we partition each one of \( l \) clusters into \( m \) sub-clusters, that is, the \( i \)-th cluster is divided into \( \bigcup_{m=1}^{m} T_i^m \). We can further obtain the detailed research topics \( T'_w = \bigcup_{m=1}^{m} T_i^w \). To obtain the research topic trend for each year, we organize \( T'_w \rightarrow T = \{ T_{w_1}, \ldots, T_{w_{w'}} \} \), where \( T_{w_i} \) represents the research topic trend in year \( w_i \) containing the specific expression of all \( l \) research topics, i.e., \( \bigcup_{m=1}^{m} T_i^w \).

4.2 Enhanced Prediction Network

Apart from the topic alignment strategy, another core part of our framework is how to capture potential trend information from the known sequence to predict the future research topic trend. To achieve the goal, we develop an enhanced prediction network EPN inspired by transformer [34], which has been proven powerful in mining inherent correlation of known knowledge. As presented in Fig. 2, the structure of EPN can be refined into an embedding module, an encoder, a decoder, and a prediction module, the details of which are introduced as follows.
4.2.1 Research Evolution Embedding

Based on TAS, we can obtain the research topic popularity score chains of all considered years, i.e., \( S = \bigcup_{m=1}^{\mu} S^m \). We extract the research topic popularity chains of continuous time steps as known research topic trends and simplify the notation of input sequence \( S' = \{S^m, \ldots, S^n\} \) as \( s = (s_1, \ldots, s_n) \).

Research topic trend at time step \( i \) is represented as an embedding vector \( s_i \in \mathbb{R}^l \), where \( l \) is the number of research topics. The mapping matrix \( \Phi \) is used to project the input into hidden space. The sum of \( \Phi s_i \) and corresponding positional encoding \( p_i \) is used to represent research evolution embedding \( e_i \), which is formally defined as:

\[
e_i = \Phi s_i + p_i \tag{4}
\]

where \( \Phi \in \mathbb{R}^{dh \times l} \), \( e_i \in \mathbb{R}^{dh} \), \( p_i \in \mathbb{R}^{dh} \). The positional encoding utilizes the periodicity of the sine and cosine functions to control the relative position of the input sequence and further keeps the sequential property of research topic trends. Formally, it is defined as:

\[
p_i(2\alpha) = \sin\left(\frac{i}{10000^{2\alpha/dh}}\right) \quad p_i(2\alpha+1) = \cos\left(\frac{i}{10000^{2\alpha/dh}}\right) \tag{5}
\]

where \( i \) implies the position of vector in the input sequence and \( \alpha = 0, 1, 2, \ldots, \frac{d}{2} - 1 \) controls the current dimension of the positional encoding. Thus, we can obtain the research evolution representation \( E = [e_1, \ldots, e_n] \in \mathbb{R}^{n \times dh} \).

4.2.2 Multi-Head Self-Attention

Research topic trend for each year is presented as the specific expression of all research topics for that year. There are two perspectives of potential information from the known sequence, research trend development along the timeline, and mutual correlation among different research topics of the current time step. Self-attention is highly capable of capturing internal correlations of data which has been widely used in various research fields [35, 36]. With this method, Self-attention can “see” the whole sequence representing research topic trends and assign different attention weights to different content to make model focus on interested information. Formally, we have:

\[
\text{SelfAttention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} V\right) \tag{6}
\]

where \( Q, K, V \) are the query, key and value matrices, respectively, and essentially the copies of \( E \in \mathbb{R}^{n \times dh} \).

Multi-head self-attention can capture richer information since it allows the model to jointly attend to information from different representation subspaces at different positions. Multi-head self-attention projects \( Q, K, V \) \( h \) times with different linear projections and perform the attention mechanism in parallel. Each projection represents a \( \text{head} \), then \( h \)-heads are concatenated and projected once again. In this way, the model can fully exploit the potential information of the known sequence. The concepts mentioned above are formally defined as:

\[
\text{head}_i = \text{SelfAttention}\left( EW_i^Q, EW_i^K, EW_i^V\right) \tag{7}
\]

\[
\text{MultiHead}(E) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O
\]

where the projections are parameter matrices \( W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{dh \times dh}, W^O \in \mathbb{R}^{hd \times dh}, \) and \( h \) is the number of \( \text{heads} \). To avoid degradation of weight matrix and gradient vanish, we employ residual connections around each of the attention sublayers, followed by layer normalization,

\[
H = LN(E + \text{MultiHead}(E)) \tag{8}
\]

where \( H \in \mathbb{R}^{n \times dh} \) represents the output through current layer, \( LN(\cdot) \) is the layer normalization operation.

4.2.3 Position-wise Feed-forward Network

The output of attention layer \( H \) is then fed into a fully connected feed-forward network, which consists of two linear transformations with a ReLU activation in between. The hidden state is updated as:

\[
H = \max\left(0, HW_1 + b_1\right)W_2 + b_2 \tag{9}
\]

where \( W_1, W_2 \) are the weight matrices of Position-wise Feed-forward Network. In addition, we also employ the residual connection and layer normalization in Eq. (8) here.

4.2.4 Enhanced Perception Module

We expect the hidden state \( H \) after Multi-head attention to be able to represent the historical research topic trend. In order to strengthen the association between the known research trend sequence and the future target, we design an enhanced perception module EPM where the decoder receives the sequence \( s' = (s_2, \ldots, s_{n+1}) \) containing future information. Specifically, EPM combines historical and enhanced information as the new input of decoder, which is formally defined as:

\[
\hat{s}' = \alpha s' + (1 - \alpha)HW_3 \tag{10}
\]

where \( W_3 \in \mathbb{R}^{dh \times l} \) and \( \alpha \) is a balancing coefficient, usually \( \alpha = 0.5 \). Note that, we completely utilize the historical
information from the encoder as decoder input when performing the test.

4.2.5 Prediction Module

The final hidden state $H$ through encoder-decoder interaction of multi-head attention contains two parts of information. One is prior information before time step $\eta + 1$ from the encoder and decoder, which contains the potential research topic trend information of known sequence. The other is the enhanced prediction information of time step $\eta + 1$ from the decoder, which is closely related to the research trends of prior time steps. Lastly, the final hidden state is fed into a multilayer perceptron (MLP) to predict the future research topic trend $\hat{s}_{\eta+1}$,

\[
\hat{s}_{\eta+1} = MLP(H) \tag{11}
\]

4.3 Optimization Objective

The model is trained by minimizing the Mean Squared Error loss function, which is defined as:

\[
L = \frac{1}{N} \sum_{i=1}^{N} (\hat{s}_{\eta+1} - s_{\eta+1})^2 \tag{12}
\]

where $\hat{s}_{\eta+1}$ is the predicted vector, $s_{\eta+1}$ is the actual vector representing research topic popularity score chain $S_{w_{\eta+1}}$ in year $w_{\eta+1}$. Finally, the objective is optimized with the Adam optimization algorithm [37].

5 Experiment

5.1 Data Preparation

We select conferences from two active research domains in computer science to evaluate the proposed framework. Specifically, in database and data mining domain (DBDM), four conferences including CIKM, ICDE, SIGIR, SIGMOD are selected, and in computer architecture and parallel programming domain (CAPP), five conferences including ASPLOS, ATC, ISCA, MICRO, PPoPP are considered. The selected conferences are all the international top ones in their specific area and provide the raw topics. We obtain the full research papers of conferences from DBLP and add the raw topic label and the citation label from the official conference website or Google Scholar. Notably, the collected research papers are published from 1997 to 2020. The detailed statistics of datasets are presented in Table 1.

| Dataset | Number of papers | Number of raw topics |
|---------|------------------|-----------------------|
| DBDM    | 7932             | 2317                  |
| CAPP    | 3630             | 1024                  |

Table 1 Statistics of datasets DBDM and CAPP

5.2 Experimental Settings

5.2.1 Compared Methods

To investigate the effectiveness of the proposed framework, we conduct comparative experiments against the following two aspects of methods. Firstly, the methods MEY, LR, and NNCP focusing on citation count prediction have presented relatively good performance in existing studies. Secondly, we select four representative time series models, including one classical time series model ARIMA, two widely used deep learning models RNN and GRU which have senior performance in modeling sequential information, and one latest research trend prediction model DNTP.

- **MEY**: The mean of early years method [28, 30] is a simple but effective prediction method which always uses the average of citation counts of known early years as the predicted citation counts in the future.
- **LR**: Linear regression model has been widely used as a baseline for citation count prediction [27, 28, 38, 39].
- **NNCP**: NNCP [28] is proposed to solve citation count prediction based on Seq2seq model. We feed the predicted sequence into a MLP to obtain the research popularity score for the next year while reproducing the model NNCP.
- **ARIMA**: ARIMA [6] has been widely used in time series prediction tasks. For each research topic, its popularity score in each year is regarded as the time series.
- **RNN**: Deep learning technology presents extraordinary talent in a variety of prediction tasks. Recurrent Neural Network has been proven to perform well in research topic trend prediction [8–10, 19], which has an outstanding ability to capture the sequential property.
- **GRU**: GRU is an enhanced variant of RNN, which leads to effective performance and affordable computation cost [19].
- **DNTP**: DNTP [8] is a state-of-the-art study about research trend prediction based on long short-term memory neural networks. Five different time-series values of each research topic are fed into the model and the one-dimensional output layer presents the prediction outcome.
for the next year. The number of internal units is also set to 100.

5.2.2 Parameter Settings

We use the pre-trained BERT model with 12 layers. The dimension of the embedding vector is 768. The number of research topics on datasets DBDM and CAPP is 24 and 10, respectively. For all baselines, we adjust parameters to their respective optimal results and set the length of the known sequence to 5. In our method, the number of multi-head self-attention layers is set to 6, and the number of heads is set to 8. The dimension of the hidden state is set to 512. All experiments are conducted on a single NVIDIA Geforce 1080 Ti GPU.

5.2.3 Evaluation Methodology

To evaluate the prediction performance, we organize the raw topics and papers of each conference in the temporal order. The first 70% of data are used for training, the following 10% of data are used for validation, and the remaining 20% of data are used for testing. We adopt three popular and widely used metrics in regression problems to evaluate the performance of the proposed method: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Normalized Discounted Cumulative Gain (NDCG). RMSE and MAE measure the precision of the predicted research topic popularity scores. RMSE is more sensitive to the exception value while MAE measures the deviation between the predicted and the actual values. NDCG@K concentrates on the ranking of the predicted trending topics. The smaller RMSE as well as MAE, and the larger NDCG@K are preferred.

5.3 Result Analysis

The prediction performances of all methods in terms of RMSE, MAE, and NDCG@K are presented in Table 2. The comparisons of prediction time on the testing set are presented in Table 3. Overall, RTTP achieves higher performances than all compared methods in both datasets. Based on the results, some important findings are summarized as follows.

Transformer can effectively capture mutual correlation of research trend evolution. The transformer-based framework RTTP and the other three time series models significantly exceed the classical time series method ARIMA, since ARIMA individually models the time series of each research topic instead of modeling the whole prior research trend sequence. Additionally, RTTP also outperforms the RNN-based models including RNN, GRU, and NNCP. We can observe that the proposed framework obtains a performance improvement of 3.87%, 3.07%, 3.54% over RNN, GRU, NNCP, respectively, on RMSE on DBDM, and 1.22%, 0.9%, 2.98% on CAPP. The performance of the research trend prediction model DNTP using long short-term memory on different metrics is not stable enough on two datasets.
Fig. 3 Performance of RTTP with varying $\eta$. a, b, and c represent NDCG@1, NDCG@5, and NDCG@10 on two datasets, respectively. d represents NDCG@15 and NDCG@20 on DBDM.

| $\alpha$ | RMSE | MAE | NDCG@1 | NDCG@5 | NDCG@10 | NDCG@15 | NDCG@20 |
|---------|------|-----|--------|--------|---------|---------|---------|
| 0.0     | 354.82 | 286.87 | 0.8565 | 0.8620 | 0.8936 | 0.9114 | 0.9374 |
| 0.2     | 357.60 | 289.12 | 0.8625 | 0.8658 | 0.8932 | 0.9114 | 0.9376 |
| 0.4     | 349.38 | 280.86 | 0.8651 | 0.8768 | 0.9005 | 0.9161 | 0.9417 |
| 0.5     | 343.01 | 279.87 | 0.8733 | 0.8877 | 0.9085 | 0.9207 | 0.9443 |
| 0.6     | 351.77 | 283.20 | 0.8760 | 0.8789 | 0.9041 | 0.9159 | 0.9418 |
| 0.8     | 348.58 | 280.45 | 0.8725 | 0.8803 | 0.9087 | 0.9217 | 0.9437 |
| 1.0     | 350.80 | 283.17 | 0.8505 | 0.8715 | 0.8972 | 0.9125 | 0.9383 |
The results demonstrate that transformer has a better ability to capture the potential properties of the research trend evolution since transformer utilizes the multi-head self-attention to extract interested information from global sequence rather than receive sequential information in order. We also observe that the RNN-based models exhibit high capabilities, where RNN and GRU both achieve better performance than MEY and LR, indicating that the RNN-based models are strong baselines on sequence modeling and prediction tasks.

**RTTP works effectively on top research topic trend prediction.** Top-\(K\) research topics can guide the promising branches of the current domain. For a more detailed analysis, we pay more attention to the ranking of various research topics in the future using the ranking metric NDCG@\(K\). Illustrated from Table 2, RTTP achieves the best results for all values of \(K\). It reveals that our method has a stronger ability in predicting top research topics. For the same \(K\), the performance on CAPP tends to be higher than that on DBDM, since the former only contains 10 research topics, while 24 research topics are considered on DBDM, and NDCG@10 means the ranking of all research topics for CAPP. Furthermore, we find that the improvements of RTTP on DBDM are more obvious than those on CAPP, which infers that the granularity of dividing research topics may affect top-\(K\) research topic prediction accuracy.

**RTTP is competitive in prediction time.** Observed from Table 3, ARIMA costs much more prediction time than other methods, since ARIMA needs to perform predictions for each research topic rather than generate the research topic popularity scores of all research topics once. In addition, RNN achieves the best performance on prediction time among all compared methods. RTTP gets extremely close prediction time to the optimal on DBDM while worse on CAPP. The reason is that the scale of DBDM is larger than CAPP, the advantages of parallel computing from transformer-based methods manifest as the size of datasets increases. Self-attention mechanism can compute all positions in the sequence in parallel while recurrent neural networks compute each position in the sequence one by one. RTTP achieves affordable prediction time and optimal prediction performance on smaller datasets. We argue that RTTP is more competitive on large-scale datasets in prediction time.

### 5.4 Parameter Study

\(\eta\) is the length of known years, i.e., the length of a sequence, and the corresponding NDCG@\(K\) on two datasets by varying \(\eta\) are presented in Fig. 3. Intuitively, RTTP always achieves the best performance when \(\eta\) is set to 5. The prediction performance is generally positively correlated with \(\eta\) before \(\eta=5\) while negatively on the contrary. With the decrease of known years, the model perceives less information about research topic trend evolution, thus it weakens the ability to extract valid information from early topics development pattern. When \(\eta=1\), the model degenerates to predict the research topic trend only with the information of previous one year. The performance shows a relatively weak downward tendency after \(\eta=5\) since some noise information may be introduced. Additionally, we observe a rebound tendency with the increase of \(\eta\) on dataset DBDM, possibly caused by the model capturing the long-term citation pattern that works for trend prediction.

To obtain the trade-off between external enhanced information and historical trend information, we also perform experiments over different \(\alpha\). Table 4 illustrates the results w.r.t. all evaluation metrics on dataset DBDM. It can be observed that RMSE and MAE fluctuate within a certain range. NDCG of different \(\alpha\) increases with \(\alpha\) varying from 0 to 0.8 since external enhanced information becomes more. Comprehensive consideration based on several metrics, we set \(\alpha\) to 0.5 intending to obtain the best performance for RTTP.

### 6 Conclusion and Future Work

In this paper, we propose a novel framework called RTTP to predict the research topic trend with fact-based annotations. Specifically, a novel topic alignment strategy considering both global and local information is first designed to overcome the semantic differences between cross-year and cross-venue, and further obtain the detailed contents of research topics in each year. Next, an enhanced prediction network EPN, which consists of an embedding module, an encoder, a decoder and a prediction module, is used to capture the research trend of known years for prediction. The experimental results on two real-world datasets demonstrate the effectiveness of proposed framework. What’s more, we propose a new perspective to obtain research topics from scientific papers. In the future work, we can further improve the performance of our framework by involving more information. Considering the differentiated impact of cited papers is an interesting idea to extend. The influence of authors should also be considered in research topic trend prediction, as authors or teams with high influence are more likely to lead the development of a branch topic of research domain. Furthermore, inspired by the internet of everything and the development of interdisciplinary, it may be a potential research direction to aware influence of the interaction between different research
domains (such as artificial intelligence and data mining) on research topic trend.

**Funding** This work was supported by the National Natural Science Foundation of China under Grant No. 61902270.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

**References**

1. Chen B, Tsutsui S, Ding Y, Ma F (2017) Understanding the topic evolution in a scientific domain: an exploratory study for the field of information retrieval. J Inf 11(4):1175–1189
2. Coello Coello CA (2009) Evolutionary multi-objective optimization: some current research trends and topics that remain to be explored. Front Comput Sci Chin 3(1):18–30
3. Wang X, Cheng Q, Lu W (2014) Analyzing evolution of research topics with neviewer: a new method based on dynamic co-word networks. Scientometrics 101(2):1253–1271
4. Wang X, Zhai C, Roth D (2013) Understanding evolution of research themes: a probabilistic generative model for citations. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1115–1123
5. Zhou H-K, Yu H-M, Hu R (2017) Topic discovery and evolution in scientific literature based on content and citations. Front Inf Technol Electron Eng 18(10):1511–1524
6. Abuhay TM, Nigatie YG, Kovalchuk SV (2018) Towards predicting trend of scientific research topics using topic modeling. Procedia Comput Sci 136:304–310
7. Ballili C, Segev A, Lee U (2017) Tracking and predicting the evolution of research topics in scientific literature. In: 2017 IEEE international conference on big data (big Data), pp. 1694–1697. IEEE
8. Taheri S, Aliakbar S (2022) Research trend prediction in computer science publications: a deep neural network approach. Scientometrics 127(2):849–869
9. Xu M, Du J, Guan Z, Xue Z, Kou F, Shi L, Xu X, Li A (2021) A multi-rnn research topic prediction model based on spatial attention and semantic consistency-based scientific influence modeling. Computational Intelligence and Neuroscience
10. Xu M, Du J, Xue Z, Guan Z, Kou F, Shi L (2022) A scientific research topic trend prediction model based on multi-lstm and graph convolutional network. Int J Intell Syst 37(9):6331–6353
11. Hirsch JE (2005) An index to quantify an individual’s scientific research output. Proc Natl Acad Sci 102(46):16569–16572
12. Garfield E (2006) The history and meaning of the journal impact factor. jama 295(1):90–93
13. Grabner H, Gall J, Van Gool L (2011) What makes a chair a chair? In: CVPR 2011, pp. 1529–1536. IEEE
14. Charnine M, Klokov A, Kochiev L, Tishchenko A (2021) Research trending topic prediction as cognitive enhancement. In: 2021 international conference on cyberworlds (CW), pp. 217–220. IEEE
15. Behrouzi S, Sarmoor ZS, Hajisadeghi K, Kavousi K (2020) Predicting scientific research trends based on link prediction in keyword networks. J Informet 14(4):101079
16. Selvin S, Vinayakumar R, Gopalakrishnan E, Menon VK, Soman K (2017) Stock price prediction using lstm, rnn and cnn-sliding window model. In: 2017 International Conference on Advances in Computing, Communications and Informatics (icacci), pp. 1643–1647. IEEE
17. Lv Z, Xu J, Zheng K, Yin H, Zhao P, Zhou X (2018) Lc-rnn: A deep learning model for traffic speed prediction. In: IJCAI, vol. 2018, p. 27
18. Tomihira T, Otsuka A, Yamashita A, Satoh T (2020) Multilingual emoji prediction using bert for sentiment analysis. Int J Web Inf Syst
19. Chen C, Wang Z, Li W, Sun X (2018) Modeling scientific influence for research trending topic prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32
20. Oppenheim C (1995) The correlation between citation counts and the 1992 research assessment exercise ratings for british library and information science university departments. J Doc
21. Garfield E (1998) The use of journal impact factors and citation analysis for evaluation of science. In: 41st Annual meeting of the council of biology editors, Salt Lake City, UT
22. Moed HF (2006) New developments in citation analysis and research evaluation. Int Serv Use 26(2):135–137. https://doi.org/10.3233/ISU-2006-26217
23. Egghe L (2006) Theory and practise of the g-index. Scientometrics 69(1):131–152
24. Moed HF, Collledge L, Reedijk J, Moya-Anegon F, Guerrero-Bote V, Plume A, Amin M (2012) Citation-based metrics are appropriate tools in journal assessment provided that they are accurate and used in an informed way. Scientometrics 92(2):367–376
25. Wildgaard L, Schneider JW, Larsen B (2014) A review of the characteristics of 108 author-level bibliometric indicators. Scientometrics 101(1):125–158
26. Yan R, Tang J, Liu X, Shan D, Li X (2011) Citation count prediction: learning to estimate future citations for literature. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 1247–1252
27. Yan R, Huang C, Tang J, Zhang Y, Li X (2012) To better stand on the shoulder of giants. In: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 51–60
28. Abrishami A, Aliakbar S (2019) Predicting citation counts based on deep neural network learning techniques. J Informet 13(2):485–499
29. Ma A, Liu Y, Xu X, Dong T (2021) A deep-learning based citation count prediction model with paper metadata semantic features. Scientometrics 126(8):6803–6823
30. Huang S, Huang Y, Bu Y, Lu W, Qian J, Wang D (2022) Fine-grained citation count prediction via a transformer-based model with among-attention mechanism. Inf Process Manage 59(2):102799
31. Xiao H (2018) bert-as-service. https://github.com/hanxiao/bert-as-service
32. Jones KS (1972) A statistical interpretation of term specificity and its application in retrieval. J Doc
33. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. Oakland, CA, USA
34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
35. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057. PMLR
36. Tang G, Müller M, Rios A, Sennrich R (2018) Why self-attention? a targeted evaluation of neural machine translation architectures. arXiv preprint arXiv:1808.08946
37. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
38. Lovaglia MJ (1991) Predicting citations to journal articles: the ideal number of references. Am Sociol 22(1):49–64
39. Yu T, Yu G, Li P-Y, Wang L (2014) Citation impact prediction for scientific papers using stepwise regression analysis. Scientometrics 101(2):1233–1252