IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 25, 2019, accepted August 17, 2019, date of publication August 21, 2019, date of current version September 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936745

Attention-Based Deep Learning Model for Predicting Collaborations Between Different Research Affiliations

HUI ZHOU[1], JINQING SUN[1], ZHONGYING ZHAO[1], YONGHAO YANG[1], AILEI XIE[2], AND FRANCISCO CHICLANA[3,4]

[1]Shandong Province Key Laboratory of Wisdom Mine Information Technology, College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
[2]School of Education, Guangzhou University, Guangzhou 510006, China
[3]School of Computer Science and Informatics, Institute of Artificial Intelligence, De Montfort University, Leicester LE1 9BH, U.K.
[4]Andalusian Research Institute on Data Science and Computational Intelligence (DaSCI), University of Granada, 18071 Granada, Spain

Corresponding author: Zhongying Zhao (zzysuin@163.com)

This work was supported in part by the Humanities and Social Science Research Project of the Ministry of Education in China under Grant 17YJCZH262 and Grant 18YJAZH136, in part by the National Natural Science Foundation of China under Grant 61303167, Grant 61702306, Grant 61433012, Grant U1435215, and Grant 71772107, in part by the Natural Science Foundation of Shandong Province under Grant ZR2018BF013 and Grant ZR2017BF015, in part by the Innovative Research Foundation of Qingdao under Grant 18-2-2-41-jch, in part by the Key Project of Industrial Transformation and Upgrading in China under Grant TC170A5SW, and in part by the Scientific Research Foundation of SDUST for Innovative Team under Grant 2015TDJH102.

ABSTRACT It is challenging but important to predict the collaborations between different entities which in academia, for example, would enable finding evaluating trends of scientific research collaboration and the provision of decision support for policy formulation and incentive measures. In this paper, we propose an attention-based Long Short-Term Memory Convolutional Neural Network (LSTM-CNN) model to predict the collaborations between different research affiliations, which takes both the influence of research articles and time (year) relationships into consideration. The experimental results show that the proposed model outperforms the competitive Support Vector Machine (SVM), CNN and LSTM methods. It significantly improves the prediction precision by a minimum of 3.23 percent points and up to 10.80 percent points when compared with the mentioned competitive methods, while in terms of the F1-score, the performance is improved by 13.48, 4.85 and 4.24 percent points, respectively.

INDEX TERMS Relationship prediction, collaboration analysis, coauthor networks, deep learning.

I. INTRODUCTION

Scientific collaboration analysis has become a very interesting topic, which has attracted the interest of numerous researchers. This research can be applied in the assessing of the collaborative research trend between research affiliations in a country and the provision of decision support for policy formulation and incentive measures [1], [2]. The realisation of such application requires, as its first step, the prediction of potential scientific collaborations between different research affiliations. Therefore, it is necessary to design a novel methodology for predicting accurately and efficiently the collaborative relationship between different research affiliations.

The associate editor coordinating the review of this article and approving it for publication was Xiping Hu.

Collaborative relationship prediction is playing a very important role in both enterprise development and scientific area. It enables us to analyze the formation and the evolution of communities [3], [4]. Thus, it has attracted the attention of numerous researchers [5], [6]. In the evolution of homogenous co-authorship network context, for the link prediction problem Huang *et al.* proposed a hybrid approach that utilizes time-varied weight information of links [7], while Menon *et al.* designed a supervised method based on matrix factorization to learn the network topology [8]. In [9], Jin *et al.* presented a local index of path to estimate the likelihood of links between two network nodes and proposed a network model with controllable density and noise strength to predict links in homogenous social networks; and Murata *et al.* proposed in [10] a link prediction method based on weighted proximity measures of social networks.

In most of the above referred studies, researchers are concerned with homogeneous networks, i.e. network with only one type of objects and one type of link (co-authorship) [11], [12]. Nevertheless, in a real bibliographic network there are multiple types of objects (journals, topics, papers) and multiple types of links among these objects. To address this issue, Sun *et al.* [5] designed a meta-path based model to predict the co-author relationship in the heterogeneous bibliographic network. In addition, Amin and Murase [13] also studied the co-author network and added affiliation information into the network to enhance the performance of link prediction.

The above methods, though, do not consider relationship dynamics over time. Furthermore, most of the existing methods focus only on predicting the co-authoring relationship between authors. However, governments are often more interested in the current international collaborative states and the future trends of their research affiliations. Therefore, the prediction of collaborative relationships between different research affiliations within a dynamic heterogeneous network context is a timely, interesting and, at the same time, challenging research problem. To address this research problem, this paper proposes and develops an attention-based deep learning model. First, the co-author network is constructed as a heterogeneous network with time dynamics. An attention-based Long Short-Term Memory Convolutional Neural Network (LSTM-CNN) model is then presented. The proposed model takes both the influence of the research papers and their publication time (year) relationships into consideration to effectively predict the collaborative relationships between different research affiliations.

The main contributions of this paper are summarized as follows:

- A dynamic heterogeneous collaborative network is constructed to model the real world co-authored data, which fully utilizes the information of both research papers and research affiliations. The research papers' features are adopted to describe the relationships between research affiliations.
- An attention-based LSTM-CNN model, which exploits both dynamic collaboration and network structures, to predict the collaborative relationships between different research affiliations is presented.
- A comparative study of the proposed model with shallow and deep models, respectively, including SVM [14], CNN [15], and LSTM [16] is carried out. Experimental results show that the proposed attention-based hboxLSTM-CNN model significantly improves the predicting performance.

The remainder of this paper is organized as follows. Section II briefly reviews the key research works in the literature related to this area. Data modeling and problem definition are presented in Section III. In Section IV, the attention-based LSTM-CNN model is proposed to predict the collaborations between different research affiliations and its algorithm are described in detail. Extensive experimental results and discussions are reported in Section V. Finally, conclusions related to this research study are drawn in Section VI.

II. RELATED WORK

A. RELATIONSHIP ANALYSIS AND PREDICTION

Relationship identification and prediction are of great interest. It helps service providers to recommend friends and items to customers more precisely [17]. Moreover, the prediction of relationships also helps enterprises to recognize potential partners and competitors [18]. Hence, this topic has received a great deal of attention from researchers. Relationship prediction can also be considered as a link prediction problem and, within this point of view, various methods have been proposed including similarity based methods and learning based methods [19]–[21]. Li *et al.* proposed a link prediction framework in [22] by considering both nodes similarity and community information. Wang *et al.* [23] proposed an effective approach by fusing the adjacent matrix and some key topological metrics in a unified probability matrix factorization framework, which considers not only symmetric metrics but also asymmetric metrics. Xu *et al.* proposed Edge-Nodes Representation Neural Machine (ENRNM) in [24] for link prediction by capturing the abundant topological features in the network. Muniz *et al.* [25] combined contextual, temporal and topological information together to improve the performance of link prediction. Zhang *et al.* [26] focused on time evolving network and proposed an incremental dynamic link prediction algorithm. Meanwhile, the attributes in a network can contribute to the analysis and prediction of the relationship between its nodes. He *et al.* [27] proposed a Two-stage Iterative Framework (TIFIM) to obtain the maximum influential node. Based on node influence and neighbor coordination, a weighted coordination model was proposed in [28] to compute the opinions of the nodes with the change of iterations. Moreover, in [29] a 3-hop heuristic algorithm was proposed to effectively determine the top-m influential nodes. HIUD was proposed in [30] to detect the influential nodes in real social network by considering nodes feature and interactive relationships.

B. DEEP LEARNING BASED METHOD

In recent years, deep learning has been proved to be an efficient methodology to predict relationships in complex networks. Cai *et al.* [31] proposed a link prediction approach based on the recurrent neural network link prediction (RNN-LP) framework. Zhao *et al.* [32] designed a deep model equipped with improved Refresh Gate Recurrent Units (RGRU) to detect advisor-advisee relationships. Sharma and Sharma [33] adopted neural networks to predict links in academic social networks. Li *et al.* [34] proposed a novel meta-path feature-based back propagation (BP) neural network model to predict multiple types of links for heterogeneous networks. LSTM was designed in [35] as a variant of a recurrent neural network (RNN) to process and predict data with long intervals and time series delays. Results reported

in [36] showed that CNN can effectively extract lexical information from morphological information (such as the prefix or suffix of a word) and encode it as a neural representation.

III. DATA MODELING AND PROBLEM DEFINITION

A. DATA MODELING

The purpose of the proposed model is to systematically define the relationships between research affiliations in a continuous period (of time) to predict their future research collaboration. In this paper, the collaborative network is composed of two types of nodes, which represent the research papers and the research affiliations, respectively. Therefore, the features of both research affiliations and research papers will be considered when extracting the collaborative relationship between research affiliations.

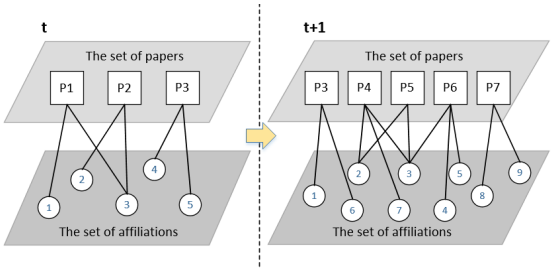


FIGURE 1. Samples of dynamic heterogeneous network at time t and $t+1$, respectively. Round nodes represent affiliations, and square nodes represent papers. The edge between affiliation (A_i) and paper (P_j) represents that at least one author of paper P_j belongs to the affiliation A_i. When two affiliations are connected to the same paper, it means that there is a collaborative relationship between them.

Based on the collaborative information, a dynamic heterogeneous network $G(A, P, E)$ is proposed, where $A = \{a_1, a_2, \ldots, a_n\}$ is the set of research institutions (affiliations), $P = \{p_1, p_2, \ldots, p_l\}$ is the set of research papers, and E is the set of undirected edges between affiliations and papers. The edge between a research institution and a research paper indicates that at least one author of the research paper is affiliated with the research institution. Thus, two research affiliations connected with the same research paper represents a collaborative relation. However, some authors may belong to two or more research affiliations at the same time. We assume that collaborations between research affiliations can only be considered when the co-authors of a paper belong to different research affiliations. A sample of such type of dynamic heterogeneous network is shown in Figure 1. As it can be seen from Figure 1, the research affiliation 3 has collaborated with research affiliation 1 and 2 (with research papers p_1, p_2) at time t; while the same research affiliation 3 collaborated with research affiliation 2, 4, 5 and 7 (with research papers p_4, p_5 and p_6) at time $t+1$.

B. PROBLEM DEFINITION

The cooperative behavior is analysed from the perspectives of the papers and collaborators. From the collaborators point of view, the following observation is noticed: if some authors have published papers jointly in the past, they are more likely to cooperate again in the future. This is also suitable for research affiliations. From the papers point of view, it has been proved that the influence of co-authored papers also affects future cooperations between research affiliations [37]. In other words, the strength of the collaborative relationship of a research institution can be described by the influence of the paper (measured by the number of citations received). Likewise, the higher the number of collaborations between affiliations in the past, the more likely they are to collaborate again in the future.

The influence of research paper p_i is associated to its citations count, $N^{p_i}_{citation}$, and downloads count, $N^{p_i}_{download}$ [38]. Because of the certain correlation between citations count and time since publication, we define the following normalized number of citation $N^{p_i}_{citation} = (n^{p_i}_c)/(t)$, where $n^{p_i}_c$ is the number of citations of research paper p_i at time (number of years) since publication, t. According to the number of citations and downloads, the influence of a research paper is defined as $I_{p_i} = softmax(N^{p_i}_{citation}) + \alpha N^{p_i}_{download} + b$, where $softmax(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}$, α is the parameter of $N^{p_i}_{download}$, and b is the bias.

Let U_x be a research affiliation having collaborative relations with affiliations $\{d_1, d_2, \ldots, d_m\}$. The cooperation matrix of U_x based on the influence of the papers is defined as:

$$I_{U_x} = \begin{bmatrix} I^0_{U_x d_1} & \cdots & I^{T-1}_{U_x d_1} \\ \vdots & \ddots & \vdots \\ I^0_{U_x d_m} & \cdots & I^{T-1}_{U_x d_m} \end{bmatrix}, \tag{1}$$

where $I^t_{U_x d_j} = \sum I^t_{p_{i U_x d_j}}, 0 \le t \le T-1$, represents the sum of the influences of all collaborative papers between U_x and d_j in year t. Meanwhile, the cooperation matrix of U_x considering time-varying collaborations is defined as:

$$N_{U_x} = \begin{bmatrix} N^0_{U_x d_1} & \cdots & N^{T-1}_{U_x d_1} \\ \vdots & \ddots & \vdots \\ N^0_{U_x d_m} & \cdots & N^{T-1}_{U_x d_m} \end{bmatrix}, \tag{2}$$

where $N^t_{U_x d_i}$ represents the total number of collaborative research papers between the affiliation U_x and d_i in year t.

Given the dynamic heterogenous network $G(A, P, E)$, we aim to predict the collaboration relation between research affiliations. The research problem in this paper is the computation of the collaborative probability matrix P_r based on the cooperative matrices I_{U_x} and N_{U_x}, which can be formalized as follows:

$$\varphi : (I_{U_x}, N_{U_x}) \rightarrow P_r. \tag{3}$$

IV. THE PROPOSED MODEL AND ALGORITHM

Figure 2 shows the overall framework of the proposed collaborative prediction model, which consists of two components: the LSTM model and the attention-based CNN model.

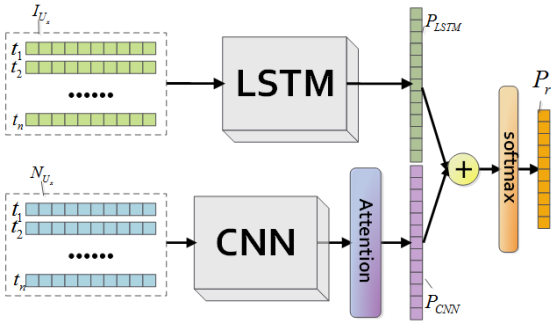


FIGURE 2. The proposed model to predict the collaborations between different affiliations.

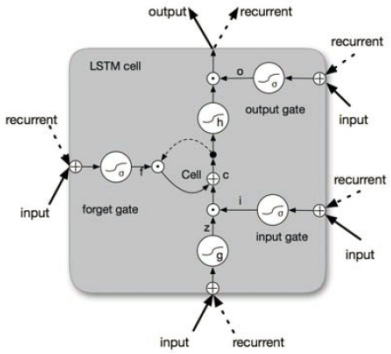


FIGURE 3. Schematic of LSTM cell.

The data of influence values of papers (I_{U_x}) is input into LSTM model as a time series, and the cooperation matrix based on the number of collaborations (N_{U_x}) is used as the input of the CNN model. Meanwhile, the attention mechanism is adopted in the CNN model to process the output features. In the following sections, we will describe the proposed model in detail.

A. COLLABORATIVE RELATIONSHIP PREDICTION MODEL BASED ON LSTM

RNN is a powerful model for capturing time dynamics but it cannot capture long-term information. LSTM is a variant of RNN designed to deal with the vanishing gradient problem. Basically, a LSTM cell consists of three multiplication gates to control the proportion of information to forget and to pass on to the next step. The basic structure of an LSTM cell is shown in Figure 3. The formulas for updating LSTM cells at time t are:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{4}$$
$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \tag{5}$$
$$\widetilde{c}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c), \tag{6}$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{7}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t, \tag{8}$$
$$h_t = o_t \odot tanh(c_t), \tag{9}$$

where,

- i_t controls the information of x_t stored in the memory cell;
- f_t is the function that controls what information will be discarded from the memory cell;
- $\widetilde{c}_t$ is the function that updates the memory cell;
- o_t controls the output based on the memory cell;
- c_t is the memory cell;
- h_t is the final output at time t;
- σ is the element-wise sigmoid function;
- $\odot$ is the element-wise product;
- x_t is the input vector at time t;
- h_t is the hidden state vector storing all the useful information at time t;
- W_i, W_f, W_c, W_o are the weight matrices for hidden state h_t;
- b_i, b_f, b_c, b_o are the bias vectors.

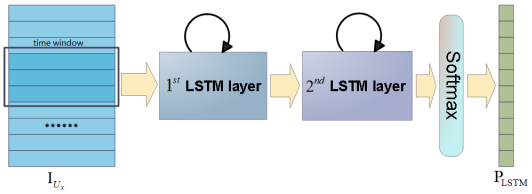


FIGURE 4. The collaborative relationship prediction model based on LSTM.

Figure 4 shows the structural details of the LSTM-based collaborative relationship prediction model. Each LSTM hidden layer contains a self loop weight, which enables cell elements in the memory module to preserve the previous information. In the proposed LSTM model, the input is the cooperation matrix I_{U_x} with T-dimensional time series, and the probability matrix is obtained through double-layer LSTM neurons and the softmax layer. Double-layer LSTM neurons can capture the information of the past and the future in two separate hidden states. We regard the hidden vector h_t output from the last step of the LSTM as the representation of the cooperation and add a softmax layer on it, which is symbolized as follows:

$$P_{LSTM} = softmax(h_t). \tag{10}$$

The model aims to predict the possibility of research cooperation between different research affiliations in the future based on the influence of the collaborative papers. We train the entire model by minimizing the cross-entropy error. It is also the submodule of the proposed attention-based LSTM-CNN model.

B. THE ATTENTION-BASED LSTM-CNN MODEL

Previous studies have shown that CNN is an effective approach for extracting morphological information from original features [36]. Thus, in this paper, a CNN model with attention mechanism to extract the collaboration information is designed as per Figure 5. As can be seen from Figure 5, the model is composed of four layers: the input layer; the convolutional layer; the pooling layer; and the attention layer.

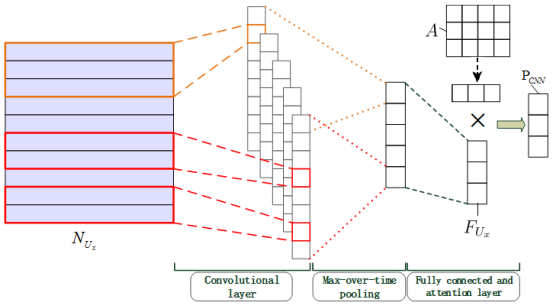


FIGURE 5. The collaborative relationship prediction model based on CNN.

The input of the designed CNN model is the cooperation matrix $N_{U_x} \in \mathbb{R}^{T\times m}$, where T denotes the size of the time window, and m denotes the number of the collaborative affiliations. The t-th row of N_{U_x} is denoted as $X_t \in \mathbb{R}^{1\times m}(0 < t < T)$, which represents the collaborative relationships of the affiliation U_x with other m affiliations in the year t. The collaborative relationships from the first year to the t-th year can be recorded as $X_{1:t} = X_1 \bigoplus X_2 \ldots \bigoplus X_t$, where $\bigoplus$ is the concatenation operator. As can be seen from Figure 5, the shapes of the filters in convolutional layer are rectangles of different sizes. Especially, we adjust shape of the filter to fit the input of the module. The width of the rectangle is fixed, matching the number of research affiliations, while the height is varied to capture the information about different fields of view. The convolution process under one filter u and one window $[i : i+h-1]$ is denoted as $c_i = f(W_u * X_{[i:i+h-1]} + b_u)$, where $W_u \in \mathbb{R}^{h\times m}$ and $b_u \in \mathbb{R}^m$ are the parameter matrix and bias vector of the filter u, respectively; h is the window size, and f is the non-linear activation function. Thus, the convolutional processes under different windows can be generated in the same process. We gather the convolutional result in each window in the vector $c = [c_1, c_2, \ldots, c_{T-h+1}]$, where $c \in \mathbb{R}^{T-h+1}$. Next, the max-over-time pooling operation [39] is adopted to select the maximum value in the convolutional results of each filter, which symbolized as $\hat{c} = max\{c_1, c_2, \ldots, c_{T-h+1}\}$. The pooling process can find out the most important feature for the next layer to predict the collaborative probability.

In the following, the attention mechanism is introduced to capture the key of collaboration information in different years. The attention matrix in the proposed model is denoted as $A \in \mathbb{R}^{m\times T}$, in which each element $a_{ij} \in A$ is initialized as follows:

$$a_{ij} = \begin{cases} 1, & if\ N^j_{U_x d_i} > 0 \\ 0, & otherwise \end{cases} \tag{11}$$

The output vector of the fully connected layer is denoted as $F_{U_x} \in \mathbb{R}^m$. Each column vector of the attention matrix is multiplied with the corresponding unit in F_{U_x} to generate the probability in the output layer. The process is formally defined as $P^l_{CNN} = A^T[l, :] \times F_{U_x}[l]$, where $l \in [1, m]$. Combined with the fully connected layer, the attention layer can effectively reduce the computing overhead of the high-dimensional data and reduce the dimension of the feature.

Finally, we employ a *softmax* function on the results derived from the double-layer LSTM model described in Section IV-A and the attention-based CNN model described above to produce the final probability distribution, which is denoted as $P_r = softmax(P_{LSTM} + P_{CNN})$.

Summarising, the proposed model considers both the influence of research papers and collaborative relationships as time series. We further formalize the above key steps of the proposed model in Algorithm 1.

Algorithm 1 The Attention-Based LSTM-CNN Algorithm for Predicting Cooperations Between Different Research Affiliations

Require: Dynamic heterogeneous graph $G(A, P, E)$.
Ensure: Collaborative probability matrix P_r.

```
Initialize the collaborative probability matrix P_r;
for each affiliation x ∈ A do
    construct matrix I_{U_x}, N_{U_x} according to Equations (1)
    and (2), respectively;
    h_t ← o_t ⊙ tanh(c_t);
    P_LSTM ← softmax(h_t);
    construct matrix A according to Equation (11);
    call CNN model to get F_{U_x};
    P_CNN ← (F_{U_x}, A);
    P_r [x, :] ← softmax(P_LSTM ⊕ P_CNN);
end for
return P_r
```

V. EXPERIMENTAL SETUP AND DISCUSSION

A. EXPERIMENTAL DATASET

In this paper, future collaborations between scientific research affiliations is predicted based on their previous collaborations. The data sets used in existing research works generally do not contain affiliations, or lack of the temporal variable of such collaborations. Therefore, a program was created to crawl data sets with affiliations and the publishing year of papers from IEEEXplore for the experimental evaluation. Meanwhile, considering that the Greater Bay Area is currently playing a very important role in both scientific and technological innovation in China and one of the authors is supported by a related research funding project, in this paper we focus on the affiliations in the Guangdong-Hong Kong-Macao Greater Bay Area. Of course, the proposed methods can also be applied in other areas once the data is available.

The data set used in this experiment contains information of 18,921 papers published in IEEEXplore from 2010 to 2017, as shown in Table 1. These papers belong to 498 research affiliations in the Guangdong-Hong Kong-Macao Greater Bay Area.

The research papers data contain information regarding the following features: their topics, published year, information of authors, number of citations and number of downloads. About 78% of the papers were cited between 0 and 9 times.

TABLE 1. Dataset statistics.

Papers	All Affiliations	Chinese Affiliations	Foreign Affiliations
18,921	943	498	445

TABLE 2. Results with different choices of size of time window.

Size of Time Window	1	2	3	4	5	6
ACC	0.857	0.864	0.879	**0.907**	0.822	0.741
Precision	0.864	0.878	0.883	**0.911**	0.872	0.825
Recall	0.848	0.845	0.867	**0.895**	0.826	0.743
F1-score	0.856	0.861	0.875	**0.903**	0.848	0.786

B. PARAMETER SETTINGS AND TRAINING

The prediction of the collaborative relationship is regarded within this study as a binary classification problem. The input of the proposed model is a dynamic heterogeneous graph $G(A, P, E)$, and the output is a collaborative probability matrix P_r. Four metrics are adopted to evaluate the performance of the experimental results: accuracy; precision; recall and F1-score.

Since the time window is an important parameter that affects the prediction results of the model, we test this parameter using the values from 1 to 6 to find the most appropriate one. The testing results are displayed in Table 2. The four metrics increase for window sizes 1, 2, 3 and 4, and decrease for window sizes 5 and 6. Thus, we set the window size to be 4 in the experiment, which means that the model uses the cooperation information of the previous four years to predict the future cooperation.

In the LSTM part of the model, the embedding size of the cooperation matrix is set to be 130 according to the crawled data, which means that one research affiliation has at most 130 collaborative affiliations. The optimal values of the parameters in the proposed model are obtained based on the empirical values adopted in the classic model and the adjustment on the crawled experimental data in this paper. As a result, the proposed model achieves the best performance when the first and second LSTM layers of the module contain 1000 and 1500 units, respectively. In addition, the dropout method is adopted to prevent overfitting, and the value of the dropout rate is set to be 0.5. In the attention-based CNN module, we adopt the multiple filters to extract the data features more comprehensively. There are three filters with different widths in this module, and each of them contains 100 feature maps. The empirical values of the width of the three filters are 3, 4 and 5, respectively. Considering that the length of the time window is small and the convolution kernel width is fixed, there is only one convolutional layer and one pool layer in this CNN model. Additionally, the dropout rate is set to be 0.5, and l_2 constraint is set to be 3 in the experiment.

C. EXPERIMENTAL RESULTS AND ANALYSES

Figure 6 shows the collaborative subnetwork of the affiliations in 2017 derived from the experimental data, in which each vertex represents an affiliation, and the size of a vertex is determined by the number of its collaborators. The undirected edge between two nodes means that the two affiliations have collaborative relationship, and the weight of the edge corresponds to the number of collaborations. According to Figure 6, we can see that the papers published form South China University of Technology have the largest number of collaborators in 2017.

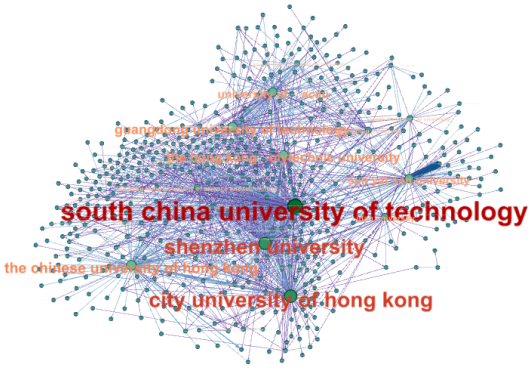


FIGURE 6. Experimental data collaborative subnetwork in 2017.

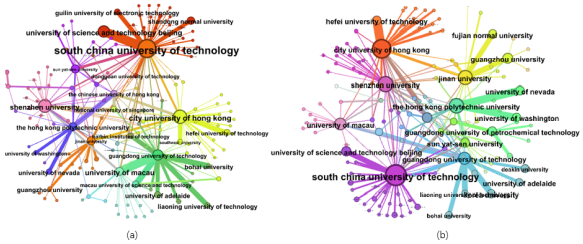


FIGURE 7. The collaborative networks based on two kinds of weights: (a) collaborative network with edge weight based on the number of citations; (b) collaborative network with edge weight based on the average number of citations.

In terms of the number of citations, we also plot collaborative networks, as shown in Figure 7. The weights of edges in Figures 7(a) and 7(b) correspond to the total number of citations and the average number of citations, respectively. The lack of edge between nodes in these two networks means the lack of collaborative papers or that the collaborative papers have zero citation and therefore the edge has zero weight and it is not shown. Additionally, the size of each node corresponds to its degree centrality measure. As can be seen from Figure 7, the larger nodes have more collaborative affiliations and have formed tight subnetworks. Thus, the collaborative papers among affiliations conform to the characteristics of ''the small world''. Moreover, institutions tend to collaborate with well-known institutions, and the influence of the collaborative papers between the well-known institutions is also higher.

We compare the proposed model with the following three competitive methods: SVM [14], CNN [15], and LSTM [16]. As can be seen from the results in Table 3, the proposed model achieved the best performance under the four evaluation metrics. In the case of the Guangdong-Hong Kong-Macao Greater Bay co-authoring network, when the input is I_{U_x},

TABLE 3. Comparative results of the proposed attention-based LSTM-CNN model with competitive baselines.

Input	Model	ACC	Precision	Recall	F1-score
I_{U_x}	SVM	72.61%	80.32%	73.6%	76.81%
	CNN	86.20%	87.61%	83.37%	85.44%
	$LSTM$	86.63%	87.89%	84.28%	86.05%
N_{U_x}	SVM	75.93%	80.91%	74.27%	77.45%
	CNN	86.02%	86.04%	84.82%	85.43%
	$LSTM$	87.23%	87.21%	85.09%	86.14%
I_{U_x} N_{U_x}	Proposed model	**90.71%**	**91.12%**	**89.47%**	**90.29%**

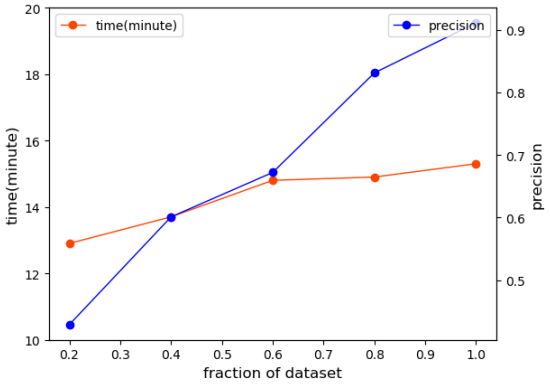


FIGURE 8. Training time and precision of the our model applied to an increasing fraction of the dataset.

the proposed model outperformed SVM by 13.48 percent points, CNN by 4.85 percent points, and LSTM by 4.24 percent points in terms of F1-score. When the input is N_{U_x}, the F1-score achieved with the proposed model outperforms the three comparative algorithms by 12.84, 4.86 and 4.15 percent points, respectively. In fact, under the two different inputs and four evaluation metrics it is noticed that LSTM outperforms CNN and CNN outperforms SVM, which indicates that the neural network performs better in processing data with complex relationships. In the proposed model, LSTM and CNN both serve as the sub-modules to promote the generation of the optimal results. It is also pointed out that the proposed model also considers important factors that influence the predicted results: influence of paper; collaborative relationships; and time series. All these properties make the proposed model the best of the four compared algorithms.

D. DISCUSSION

1) THE DISCUSSION ON THE SCALABILITY AND APPLICABILITY OF THE PROPOSED MODEL

To further analyze the scalability of the proposed model, we collect the training times for the data in different volumes, as shown in Figure 8, where the horizontal axis indicates the fraction of the dataset (i.e. the size of the data volume), and the two vertical axes indicate the training time (left) and the precision (right), respectively. As can be seen, the proposed model scales nearly linearly with the increase in the number of nodes and edges, and completes the training for all the data in about 15 minutes. In fact, it is the parallelism of the computational process in each sub-module (LSTM and CNN) which guarantees the scalability of the proposed model. In addition, the experimental precision grows rapidly with the increase of training data, which reflects the fact that training on small-scale data will lead to over-fitting of the model.

The aim of the proposed model is to systematically define the relationships between research affiliations in a continuous period to predict their future research collaboration. Although this paper focuses on analyzing the cooperative relationships in the Chinese Greater Bay Area, the rules and training methods adopted in the proposed model are universally applicable in different regions/areas at home and abroad where data is available.

2) THE DISCUSSION ON THE THREATS TO THE VALIDITY OF THE PROPOSED MODEL

For the internal validity, it is certain that the collaborative probability output from the proposed model is only related to the inputs and the training method of the model, which are completely controlled by the authors. Thus, the model possesses strong internal validity.

For the external validity, the proposed model faces a minor threat when extended to other data sets. That is, the data sets need to have the same information of the experimental data set in this paper: information of time series, collaboration papers, and research affiliations. In most cases, suitable data sets need to be crawled because there is no publicly available data that meets the requirements. In fact, it is not difficult to acquire suitable data; for this paper the relevant data from IEEEXplore had to be crawled to allow the analysis of the collaborative information for the Chinese Greater Bay Area, which can be applied to get information for other areas in the same way. In the process of crawling IEEEXplore data, we found that the page loading mode of the website is dynamic, which increases the difficulty of obtaining page information directly. To solve the problem, we adopted the request method in Urllib library to obtain the dynamic .json file of the page. In addition, the regular expression is used to extract the required information. These two approaches have improved the efficiency of data acquisition.

VI. CONCLUSION

In this paper, a dynamic heterogeneous network methodology for the collaborative relations between research affiliations was designed and an attention-based LSTM-CNN model to predict the affiliations' future collaborative relations was proposed. The dataset crawled from IEEEXplore is used to test the efficiency of the proposed model. The experimental results have shown that the proposed model achieves significant improvement when compared with the baseline algorithms of SVM, CNN and LSTM. In precision, the best result of the three baseline algorithms compared was 87.89%, while the proposed model achieved 91.12%, i.e. an increase 3.23 percent points. Meanwhile, in terms of the F1-score the proposed model improves the performance of the three compared baseline algorithms SVM, CNN and LSTM by 13.48, 4.85 and 4.24 percent points, respectively.

REFERENCES

[1] S. Aref, D. Friggens, and S. Hendy, "Analysing scientific collaborations of new zealand institutions using scopus bibliometric data," in *Proc. Australas. Comput. Sci. Week*, 2018, Art. no. 49.

[2] Y. Li, H. Li, N. Liu, and X. Liu, "Important institutions of interinstitutional scientific collaboration networks in materials science," *Scientometrics*, vol. 117, no. 1, pp. 85–103, 2018.

[3] Z. Zhao, C. Li, X. Zhang, F. Chiclana, and E. H. Viedma, "An incremental method to detect communities in dynamic evolving social networks," *Knowl.-Based Syst.*, vol. 163, pp. 404–415, Jan. 2019.

[4] Z. Zhao, S. Zheng, C. Li, J. Sun, L. Chang, and F. Chiclana, "A comparative study on community detection methods in complex networks," *J. Intell. Fuzzy Syst.*, vol. 35, no. 1, pp. 1077–1086, 2018.

[5] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks," in *Proc. Int. Conf. Adv. Social Netw. Anal. Mining*, Jul. 2011, pp. 121–128.

[6] S. Han, D. He, P. Brusilovsky, and Z. Yue, "Coauthor prediction for junior researchers," in *Proc. 6th Int. Conf. Social Comput., Behav.-Cultural Modeling Predict.*, 2013, pp. 274–283.

[7] S. Huang, Y. Tang, F. Tang, and J. Li, "Link prediction based on time-varied weight in co-authorship network," in *Proc. 18th IEEE Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, May 2014, pp. 706–709.

[8] A. K. Menon and C. Elkan, "Link prediction via matrix factorization," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, vol. 2, 2011, pp. 437–452.

[9] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 80, Oct. 2009, Art. no. 046122.

[10] T. Murata and S. Moriyasu, "Link prediction of social networks based on weighted proximity measures," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, Nov. 2007, pp. 85–88.

[11] J. Guo and H. Guo, "Multi-features link prediction based on matrix," in *Proc. Int. Conf. Comput. Design Appl.*, Jun. 2010, pp. V1-357–V1-361.

[12] M. Makrehchi, "Social link recommendation by learning hidden topics," in *Proc. ACM Conf. Recommender Syst.*, 2011, pp. 189–196.

[13] M. I. Amin and K. Murase, "Link prediction in scientists collaboration with author name and affiliation," in *Proc. 8th Int. Joint Conf. Soft Comput. Intell. Syst. (SCIS) 17th Int. Symp. Adv. Intell. Syst. (ISIS)*, Aug. 2016, pp. 233–238.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[15] L. O. Chua and T. Roska, "The CNN paradigm," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 40, no. 3, pp. 147–156, Mar. 1993.

[16] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, 2000.

[17] X. Zhu, X. Yang, C. Ying, and G. Wang, "A new classification algorithm recommendation method based on link prediction," *Knowl.-Based Syst.*, vol. 159, pp. 171–185, Nov. 2018.

[18] J. Tang, T. Lou, J. Kleinberg, and S. Wu, "Transfer learning to infer social ties across heterogeneous networks," *ACM Trans. Inf. Syst.*, vol. 34, no. 2, 2016, Art. no. 7.

[19] Z. Wang, J. Liang, R. Li, and Y. Qian, "An approach to cold-start link prediction: Establishing connections between non-topological and topological information," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 11, pp. 2857–2870, Aug. 2016.

[20] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A, Statist. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011.

[21] K. Zhou, T. P. Michalak, M. Waniek, T. Rahwan, and Y. Vorobeychik, "Attacking similarity-based link prediction in social networks," in *Proc. 18th Int. Conf. Auto. Agents MultiAgent Syst.*, 2019, pp. 305–313.

[22] L. Li, S. Fang, S. Bai, S. Xu, J. Cheng, and X. Chen, "Effective link prediction based on community relationship strength," *IEEE Access*, vol. 7, pp. 43233–43248, 2019.

[23] Z. Wang, J. Liang, and R. Li, "A fusion probability matrix factorization framework for link prediction," *Knowl.-Based Syst.*, vol. 159, pp. 72–85, Nov. 2018.

[24] G. Xu, X. Wang, Y. Wang, D. Lin, X. Sun, and K. Fu, "Edge-nodes representation neural machine for link prediction," *Algorithms*, vol. 12, no. 1, p. 12, 2019.

[25] C. P. Muniz, R. Goldschmidt, and R. Choren, "Combining contextual, temporal and topological information for unsupervised link prediction in social networks," *Knowl.-Based Syst.*, vol. 156, pp. 129–137, Sep. 2018.

[26] Z. Zhang, J. Wen, L. Sun, Q. Deng, S. Su, and P. Yao, "Efficient incremental dynamic link prediction algorithms in social network," *Knowl.-Based Syst.*, vol. 132, pp. 226–235, Sep. 2017.

[27] Q. He, X. Wang, Z. Lei, M. Huang, Y. Cai, and L. Ma, "TIFIM: A two-stage iterative framework for influence maximization in social networks," *Appl. Math. Comput.*, vol. 354, pp. 338–352, Aug. 2019.

[28] Q. He, X. Wang, C. Zhang, M. Huang, and Y. Zhao, "IIMOF: An iterative framework to settle influence maximization for opinion formation in social networks," *IEEE Access*, vol. 6, pp. 49654–49663, 2018.

[29] Q. He, X. Wang, M. Huang, J. Lv, and L. Ma, "Heuristics-based influence maximization for opinion formation in social networks," *Appl. Soft Comput.*, vol. 66, pp. 360–369, May 2018.

[30] Z. Zhao, H. Zhou, B. Zhang, F. Ji, and C. Li, "Identifying high influential users in social media by analyzing users' behaviors," *J. Intell. Fuzzy Syst.*, vol. 36, no. 6, pp. 6207–6218, 2019.

[31] X. Cai, J. Shu, and M. Al-Kali, "Link prediction approach for opportunistic networks based on recurrent neural network," *IEEE Access*, vol. 7, pp. 2017–2025, 2019.

[32] Z. Zhao, W. Liu, Y. Qian, L. Nie, Y. Yin, and Y. Zhang, "Identifying advisor-advisee relationships from co-author networks via a novel deep model," *Inf. Sci.*, vol. 466, pp. 258–269, Oct. 2018.

[33] D. Sharma and U. Sharma, "Link prediction algorithm for co-authorship networks using neural network," in *Proc. 3rd Int. Conf. Rel., Infocom Technol. Optim.*, Oct. 2014, pp. 1–4.

[34] J.-C. Li, D.-L. Zhao, B.-F. Ge, K.-W. Yang, and Y.-W. Chen, "A link prediction method for heterogeneous networks based on BP neural network," *Phys. A, Stat. Mech. Appl.*, vol. 495, pp. 1–17, Apr. 2018.

[35] A. Graves, *Long Short-Term Memory*. Berlin, Germany: Springer, 2012, pp. 37–45.

[36] J. Y. Lee and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 515–520.

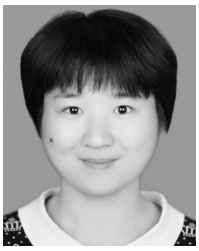
[37] F. Su, "Influence of scientific cooperation on the citation counts of journal articles," (in Chinese), *Library Inf. Service*, vol. 55, no. 10, pp. 144–148, 2011.

[38] X. Wang, M. He, Q. He, J. Guo, and D. Han, "Quantitative evaluation of scientific papers based on bibliometric research methods," (in Chinese), *Sci. Sci. Manage. S&T*, no. 4, pp. 15–18, 2004.

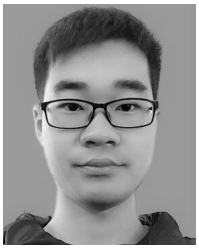
[39] N. Murray and F. Perronnin, "Generalized max pooling," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2473–2480.

HUI ZHOU received the bachelor's degree in computer science from the Shandong University of Science and Technology (SDUST), Taian, China, in 2017, where she is currently pursuing the master's degree with the Department of Computer Science and Engineering. Her research interest includes social network mining.

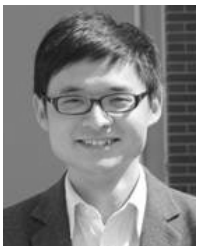
JINQING SUN received the bachelor's degree in computer science from the Shandong University of Science and Technology (SDUST), Qingdao, China, in 2016, where she is currently pursuing the master's degree with the Department of Computer Science and Engineering. Her research interest includes social network mining.



ZHONGYING ZHAO received the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2012. From 2012 to 2014, she was an Assistant Professor with the Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS). She is currently an Associate Professor with the College of Computer Science and Engineering, Shandong University of Science and Technology (SDUST). She has published 30 papers in international journals and conference proceedings. Her research interests include social network analysis and data mining. She was recognized in 2015 by SDUST with the Young Scientist for Excellence Award, and in 2016 with the High-Level Talent Award of Huangdao District.



YONGHAO YANG received the bachelor's degree in computer science from the Shandong University of Science and Technology (SDUST), Qingdao, China, in 2017, where he is currently pursuing the master's degree with the Department of Computer Science and Engineering. His research interest includes social network mining.



AILEI XIE received the Ph.D. degree from The University of Hong Kong. He is currently an Associate Professor and the Executive Director of the Bay Area Education Policy Institute for Social Development (BAEPI), Guangzhou University. He has published over 50 papers in international journals and conference proceedings. His research interests include social network analysis and social inequality. He received the University Talents Fellowship granted by Guangzhou University.

FRANCISCO CHICLANA received the B.Sc. and Ph.D. degrees in mathematics from the University of Granada, Spain, in 1989 and 2000, respectively. He is currently a Professor of computational intelligence and decision making with the School of Computer Science and Informatics, De Montfort University, Leicester, U.K. He is also a Visiting Scholar with the Department of Computer Science and Artificial Intelligence, University of Granada. He has organized and chaired special sessions/workshops in many major international conferences in research areas as fuzzy preference modeling, decision support systems, consensus, recommender systems, social networks, rationality/consistency, and aggregation. Clarivate Analytics has currently classed Prof. Chiclana as a Highly Cited Researcher in computer sciences.