The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Modeling Scientific Influence for Research Trending Topic Prediction

Chengyao Chen,[1]* Zhitao Wang,[1]* Wenjie Li,[1] Xu Sun[2,3]
[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[2]MOE Key Laboratory of Computational Linguistics, Peking University, China
[3]School of Electronics Engineering and Computer Science, Peking University, China
{cscchen, csztwang, cswjli}@comp.polyu.edu.hk, xusun@pku.edu.cn

Abstract

With the growing volume of publications in the Computer Science (CS) discipline, tracking the research evolution and predicting the future research trending topics are of great importance for researchers to keep up with the rapid progress of research. Within a research area, there are many top conferences that publish the latest research results. These conferences mutually influence each other and jointly promote the development of the research area. To predict the trending topics of mutually influenced conferences, we propose a correlated neural influence model, which has the ability to capture the sequential properties of research evolution in each individual conference and discover the dependencies among different conferences simultaneously. The experiments conducted on a scientific dataset including conferences in artificial intelligence and data mining show that our model consistently outperforms the other state-of-the-art methods. We also demonstrate the interpretability and predictability of the proposed model by providing its answers to two questions of concern, i.e., what the next rising trending topics are and for each conference who the most influential peer is.

Introduction

The research in the Computer Science (CS) discipline has made surprising progress over recent years. Accordingly, the number of research papers published on CS venues has been extremely increasing. Among all the venues, conferences are the most representative platforms to display the latest research for the fast dissemination (Vrettas and Sanderson 2015). The paper collections of conferences often hold important clues about the dynamics of research topics in according research areas. Exploring the research evolution from these conferences and being able to predict future trending topics are of great significance for a variety of communities. For instance, funding agencies can optimize the funding allocation according to the promising topics and technology companies can adjust the development strategies in advance.

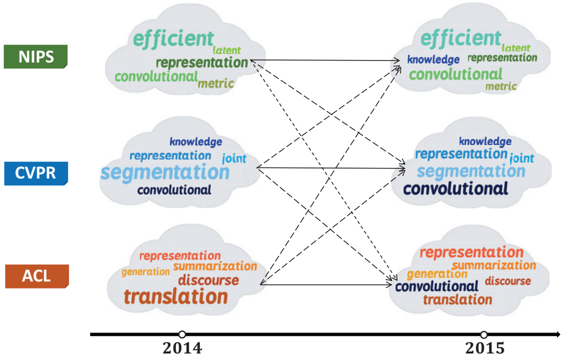

Figure 1: The research evolution of conferences in AI research area.

Considering the fast changing research trends and the growing volume of publications, keeping up with the research trends is hard even for experts. Previously, prominent efforts have been made to detect the existing research topics through topic modeling (Griffiths and Steyvers 2004; Steyvers et al. 2004) or perform a historical analysis on how research topics change in the past (Wang and McCallum 2006; Wang, Zhai, and Roth 2013). However, few studies have explored the problem of research trending topic prediction. Recently, (Prabhakaran et al. 2016) predict the rise and fall of a specific research topic using the temporal analysis on the historical curve. However, they cannot precisely produce the research topic distribution by ignoring the correlation among different research topics. Besides, they do not provide a systematic study on the intrinsic mechanisms of research preferences formation either. These limitations hinder researchers from gaining deep insights into the future research directions.

In each research area, there are often multiple top conferences promoting the research development jointly. Each conference has its own research focus and interests. In the meanwhile, inevitably, the conferences belonging to the same area (also called peer conferences) are mutually influenced. We take three top conferences, 'NIPS', 'CVPR' and 'ACL' in the Artificial Intelligence (AI) area, as an example. Figure 1 illustrates the topic evolution of the three conferences. In 2014, a number of papers published in CVPR prove the effectiveness of Convolutional Neural Network (CNN) in the image processing. The success of CNN inspires researchers who work on the Natural Language Processing (NLP) problems. The CNN model, which barely gains the

*Authors contributed equally.

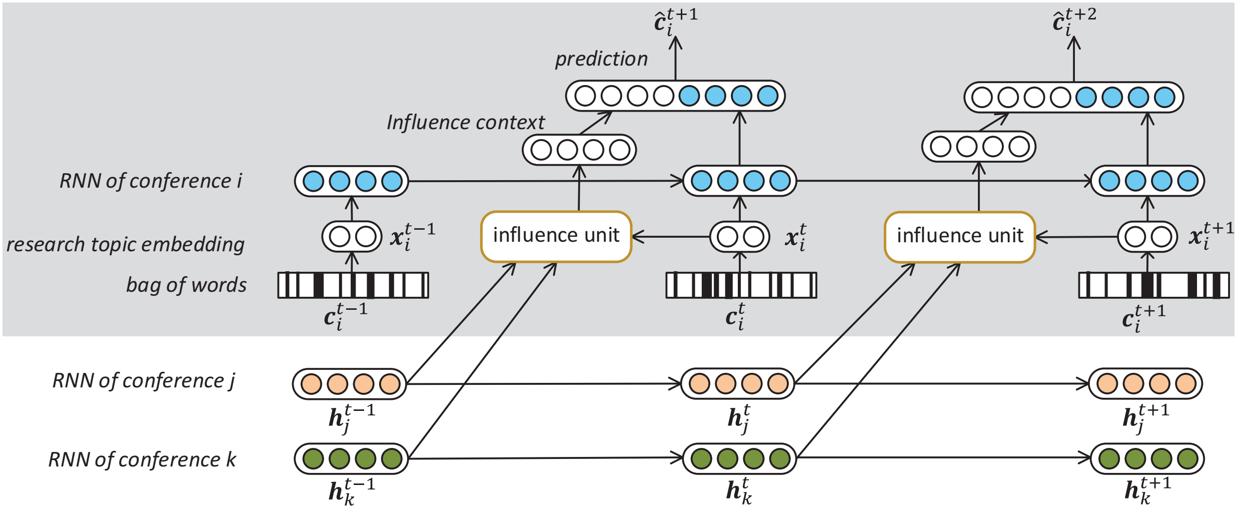

Figure 2: The framework of Correlated Neural Influence Model.

attention from NLPers before, rapidly becomes a hot topic discussed at ACL in 2015. Though several citation-based indexes, such as h-index (Hirsch 2005) or g-index (Egghe 2006), have been proposed to measure the overall impact of each conference, they cannot capture the cross-conference scientific influence on the dynamics of research topics, as illustrated above.

In this paper, we focus on the study of research trending topic prediction for mutually influenced conferences. We formulate this task as a multiple correlated sequences prediction problem and propose a novel framework named Correlated Neural Influence Model as the solution. In the framework, we use a Recurrent Neural Network (RNN) (Elman 1990) to sequentially track the research topics for each individual conference by embedding all research topics into the hidden space. The hidden states of RNN record the accumulation of sequential historical research interests at different time steps. Though RNN has been proved to have a strong ability to capture the sequential properties (Mikolov et al. 2010), it cannot uncover the cross-sequence dependencies caused by the mutual influence among conferences. To address this limitation, we propose a scientific influence unit, which correlates the multiple RNNs. It is able to construct the influence context for each conference by integrating the research hidden states of its peer conferences. From the combined vector of the hidden state and the influence context of a conference, our model produce the distribution over all topical words in the next time step. In such a way, the proposed framework jointly models the sequential research interests and external cross-conference influence in trending topic prediction.

The contributions of this paper are as follows:

- To the best of our knowledge, we are the first to systematically study the problem of research trending topic prediction through exploring the scientific influence behind the research evolution.
- Taking the scientific influence into account, we formulate the problem as a multiple mutual-influenced sequences prediction. We propose a novel framework, where multiple correlated RNN chains are developed for the research evolution modeling.
- We conduct experiments on a scientific dataset including publications of top conferences in artificial intelligence and data mining areas to demonstrate the effectiveness of the proposed model. We also demonstrate the interpretability and predictability of the proposed model by looking into the two questions of most concern, including what the next rising trending topics are and for each conference who its most influential peer is.

Correlated Neural Influence Model

The research topics of a conference would change with the research development of the other conferences in the same research area. The future research topics of a conference cannot be predicted merely based on the past publications of its own. According to the statistics on the AI conferences, about 40% new emerging words come from the other conferences. This motivates us to propose a Correlated Neural Influence Model (CONI) that can integrate the scientific influence of the peer conferences and jointly model the research evolution of all conferences in a unified framework. An overview of CONI is presented in Figure 2.

Problem Formulation

$C = \{c_1, c_2, \cdots, c_n\}$ represents the set of conferences in a research area. Let $W = \{w_1, w_2, ..., w_v\}$ be the set of v research topical words appearing in C. We use the bag-of-word vector $\boldsymbol{c}_i^t \in \mathbb{R}^v$ to represent the research topics of conference c_i at the t^{th} year. The j^{th} location of the bag-of-word representation is the normalized frequency of the word w_j occurring in c_i's publications. Given the research topics of each conference at each year, we construct a sequence of research topics for each conference c_i, i.e., $\{\boldsymbol{c}_i^1, \boldsymbol{c}_i^2, \cdots, \boldsymbol{c}_i^T\}$.

The influence relationships are represented by an adjacency matrix $\mathbf{G} \in \mathbb{R}^{n \times n}$. The i-th row $\mathbf{G}_{i*}$ represents the relationship between c_i and its peer conferences. We assume that a conference c_i can be influenced by all peer conferences in the same research area. Thus, for each $c_i \in C$, $\mathbf{G}_{ij} = 1$ if $i \neq j$, otherwise $\mathbf{G}_{ij} = 0$.

Based on the notations above, the problem studied in this work is formulated as follows. Given the research topic sequence of each conference $\boldsymbol{c}_i$ before time step T and the influence relationship matrix $\boldsymbol{G}$, the objective is to predict the future research trending topics $\boldsymbol{c}_i^{T+1}$ at the next time step $T+1$ for each conference.

Sequential Modeling of Research Evolution

In order to track the research evolution in each conference and to explore its sequential properties, the proposed CONI extends the well-known RNN framework to model the research topic sequences. It takes the research topics at the current time step as the input and iteratively encodes the research history into the hidden state to capture the research interests of the conference. Sequences of all conferences are modeled by multiple RNN chains, which share the same parameters. For ease of presentation, we take conference c_i as an example to introduce how CONI sequentially updates its hidden state of research interests according to the historical research topics.

The research topics of a conference at a specific time step are represented as a distributed vector over all the topical words. However, this bag-of-word representation may cause the 'curse-of-dimensionality' problem when the vocabulary size increases. To avoid this problem, we resort to the word embedding techniques (Mikolov et al. 2013) and convert the bag-of-word representation $\boldsymbol{c}_i^t$ into a dense and low-dimensional vector $\boldsymbol{x}_i^t$ through an embedding projection matrix $\boldsymbol{\Phi} \in \mathbb{R}^{d_w \times v}$. The embedding matrix $\boldsymbol{\Phi}$ maps each word w_j into a dense vector $\boldsymbol{\Phi}_j$. The research topics of the conference c_i at time step t are then represented by the vector $\boldsymbol{x}_i^t$.

$$\boldsymbol{x}_i^t = \boldsymbol{\Phi} \boldsymbol{c}_i^t \tag{1}$$

where $\boldsymbol{x}_i^t \in \mathbb{R}^{d_w}$.

Taking the research topics embedding $\boldsymbol{x}_i^t$ as the input, the hidden state $\boldsymbol{h}_i^t$ is iteratively updated with its previous hidden state $\boldsymbol{h}_i^{t-1}$, which is calculated as follows:

$$\boldsymbol{h}_i^t = f(\boldsymbol{h}_i^{t-1}, \boldsymbol{x}_i^t) \tag{2}$$

The recurrent function f has different variants, including the hyperbolic tangent function used in the standard RNN or other complex transformation functions like GRU (Cho et al. 2014) and LSTM (Hochreiter and Schmidhuber 1997). In CONI, we utilize GRU for its effective performance and affordable computation cost. The corresponding update formulas are:

$$\boldsymbol{h}_i^t = (1 - \boldsymbol{z}_i^t) \odot \boldsymbol{h}_i^{t-1} + \boldsymbol{z}_i^t \odot \tilde{\boldsymbol{h}}_i^t \tag{3}$$

$$\tilde{\boldsymbol{h}}_i^t = tanh(\boldsymbol{W}\boldsymbol{x}_i^t + \boldsymbol{U}(\boldsymbol{r}_i^t \odot \boldsymbol{h}_i^{t-1}) \tag{4}$$

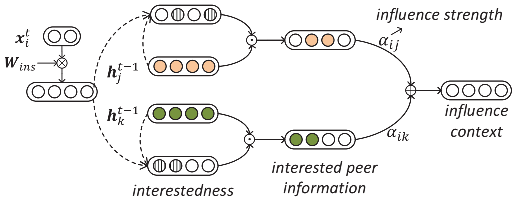

Figure 3: The scientific influence unit. The dotted lines are aimed to capture i's interestedness on the research outcomes of its peers $\boldsymbol{h}_j^{t-1}$ and $\boldsymbol{h}_k^{t-1}$, which is denoted by Eq. 7. The solid lines illustrate that after selecting the peer research outcomes with the interestedness, the interested peer information is combined under the influence strength α_{ij} and α_{ik}, forming the influence context, which is denoted by Eq. 8 and 9.

$$\boldsymbol{r}_i^t = \sigma(\boldsymbol{W}_r\boldsymbol{x}_i^t + \boldsymbol{U}_r\boldsymbol{h}_i^{t-1}) \tag{5}$$

$$\boldsymbol{z}_i^t = \sigma(\boldsymbol{W}_z\boldsymbol{x}_i^t + \boldsymbol{U}_z\boldsymbol{h}_i^{t-1}) \tag{6}$$

where $\boldsymbol{r}_i^t$ is the reset gate that controls to drop the part of historical information irrelevant to the future trends. $\boldsymbol{z}_i^t$ is the update gate that controls whether to keep the new coming information in the hidden state $\boldsymbol{h}_i^t$. σ represents the sigmoid function. $\boldsymbol{W}, \boldsymbol{W}_r, \boldsymbol{W}_z \in \mathbb{R}^{d_h \times d_w}$ and $\boldsymbol{U}, \boldsymbol{U}_r, \boldsymbol{U}_z \in \mathbb{R}^{d_h \times d_h}$.

Scientific Influence

In addition to following the within-conference research interests, a conference also follows the research development of its peer conferences. Under the cross-conference influence, the research interests of the conference will change accordingly. The RNN can well capture the historical within-conference research interests through the hidden states, but it lacks the ability to discover the dependencies on the peer conferences. In this paper, we correlate the sequences of multiple conferences by constructing a compact influence context, which combines the research information from one's peer conferences.

Two challenges arise when constructing the influence context vector. First, how to select the information a conference is interested from its peer conferences? Second, how to distinguish the influence from different peer conferences? We address these challenges by introducing a scientific influence unit, whose graphical illustration is presented in Figure 2 and elaborated in Figure 3. It selects the interested information from peer research outcomes utilizing an attention mechanism. It then linearly composites the interested information from all peer conferences based on the learned influence strength into a compact vector.

Interested Information Selection For a conference, a variety of research topics are presented in its peer conferences. However, not all topics attract its attention. We design an attention mechanism to allow the research information that a conference may be of interest to pass to it and meanwhile filter out those uninterested. Often, a conference would like to select the interested ones according to its current research topics. We expect the hidden state to be able to represent the historical research outcomes and thus measure the interestedness based on the following calculation that involves both one's current topics embedding $\boldsymbol{x}_i^t$ and the state $\boldsymbol{h}_j^{t-1}$ of its peer c_j.

$$\boldsymbol{d}_{ij}^t = softmax(\boldsymbol{W}_{ins}\boldsymbol{x}_i^t \cdot \boldsymbol{h}_j^{t-1}) \tag{7}$$

where $\boldsymbol{W}_{ins} \in \mathbb{R}^{d_h \times d_w}$ and the $softmax$ function is used to normalize the attention on each dimension. The interestedness is represented as a normalized vector, where the dimension with a higher value indicates the hidden feature receives more attention.

The interested information that a conference receives from each of its peer c_j is obtained by

$$\boldsymbol{s}_{ij}^t = \boldsymbol{d}_{ij}^t \cdot \boldsymbol{h}_j^{t-1} \tag{8}$$

Cross-Influence Composition A conference is inevitably more influenced by some peer conferences than others. To interpret the cross-influence between one conference c_i and each of its peers in $\boldsymbol{G}_i$, we aim at learning a influence parameter vector $\boldsymbol{A}_i \in \mathbb{R}^n$, where $\alpha_{ij} \in \boldsymbol{A}_i$ represents the influence strength c_i receives from its peer G_{ij}.

The influence context vector can be constructed with the linear composition of the interested research information under the cross-conference influence strength, which is denoted in Eq. 9:

$$\boldsymbol{m}_i^t = \sum_i^N (G_{ij}\alpha_{ij}) * \boldsymbol{s}_{ij}^t \tag{9}$$

In the output layer, given the hidden state $\boldsymbol{h}_i^t$, and the influence context vector $\boldsymbol{m}_i^t$, we use the $softmax$ function to output the predicted distribution of research topical words $\hat{\boldsymbol{c}}_i^{t+1} \in \mathbb{R}^v$ for each conference at the next time step $t+1$. The hidden state and the influence context vector are concatenated and fed to the $softmax$ predictor as follows:

$$\hat{\boldsymbol{c}}_i^{t+1} = softmax(\boldsymbol{W}_o[\boldsymbol{h}_i^t; \boldsymbol{m}_i^t] + \boldsymbol{b}_o) \tag{10}$$

where $\boldsymbol{W}_o \in \mathbb{R}^{v \times (d_h + d_h)}$ and $\boldsymbol{b}_o \in \mathbb{R}^v$.

Learning

We employ the generalization of multinomial logistic loss as the objective function, which minimizes the Kullback-Leibler divergence (Cover and Thomas 2012) between the predicted topic word distribution $\hat{\boldsymbol{c}}_i^{t+1}$ and the target word distribution $\boldsymbol{c}_i^{t+1}$:

$$\mathcal{L} = \sum_{i=1}^n \sum_{t=1}^T \mathbf{KL}(\hat{\boldsymbol{c}}_i^{t+1} || \boldsymbol{c}_i^{t+1}) \tag{11}$$

where

$$\mathbf{KL}(\hat{\boldsymbol{c}}_i^{t+1} || \boldsymbol{c}_i^{t+1}) = \sum_j \hat{c}_{i,j}^{t+1} log \frac{\hat{c}_{i,j}^{t+1}}{c_{i,j}^{t+1}}$$

The model is trained by minimizing the loss for the research topic sequences of all conferences. Because the hidden states are correlated through the scientific influence unit, we jointly conduct back-propagation along multiple chains (Alahi et al. 2016), and update the parameters with the Adam (Kingma and Ba 2014) algorithm.

Experiments

Dataset

We select conferences from two active research areas in CS, i.e., Artificial Intelligence (AI) and Data Mining (DM), as experimental objects to evaluate the proposed model on research trending topics prediction. In the AI area, 7 top conferences including *AAAI*, *ACL*, *CVPR*, *ICCV*, *ICML*, *IJCAI* and *NIPS* are selected, and in the DM area, 4 top conferences, i.e., *KDD*, *ICDM*, *CIKM* and *SDM* are considered.

We obtain the paper information of the above-mentioned conferences from a DBLP dataset published by (Tang et al. 2008) and updated in 2016. This dataset collects around 3.2 million research papers in CS since 1936. Each paper is associated with its author, title, venue, publication year. Based on the venue information, we recognize papers published on the target conferences and extract the titles, venues and publication years to constitute the experimental dataset.

Preprocessing The title is the most important element of a scientific article and the main indication of article's subject and topic (Jamali and Nikzad 2011). Therefore, we use the title as the topic summarization of each paper. For conference c_i at year t, we treat the title words, which occur more than once in conference publications of this year, as the research trending topics. The statistics of the dataset is presented in Table 1.

Table 1: Statistics of the dataset.

area	total papers	topical word	time period
AI	73071	3200	1995-2015
DM	9063	1562	2001-2015

Compared Methods

To demonstrate the effectiveness of the proposed model, we conduct comparative experiments against the following methods, including one basic baseline, two popular influence models, one time series model and two variants of CONI.

- **Baseline**: It takes the research topics of all conferences at current year as the predicted topics for each conference in the next year.

$$\hat{\mathbf{c}}_i^{t+1} = \frac{\sum_j^n \mathbf{c}_j^t}{n}$$

- **Degroot model** (DeGroot 1974): It is a classical influence model, which learns the interpersonal influence by exploring the correlation between a user's opinion and its friends' opinions. Following the same idea, we model the frequency of a topical word in each conference individually as follows:

$$\hat{c}_{i,k}^{t+1} = \alpha_{ii} c_{i,k}^t + \sum_{j=1}^n G_{ij}\alpha_{ij} c_{i,k}^t \tag{12}$$

where α_{ii} and α_{ij} represent the self-influence and the influence that conference i receives from conference j.

- **Flocking model** (Hegselmann and Krause 2002): It is a variant of the Degroot model assuming that a user only trusts the friends who have similar opinions with her/himself. We calculate the research topic similarities between conferences with the Euclidean distance and construct the influence relationship matrix $\boldsymbol{G}$, where $G_{ij} = 1$ if the distance between two conferences is less than 0.01. The research topics are predicted based on the Eq. 12 with the created $\boldsymbol{G}$.
- **ARMA** (Brockwell and Davis 2013): The autoregressive moving average (ARMA) model is a widely-used time series model, which has been successfully applied on citation prediction (Yogatama et al. 2011). For each conference, the frequency dynamics of each topical word at each year is regarded as the time series and the ARMA individually predicts the frequency of each word at the next year.
- **CONI_I**: It is an influence-free version of CONI. It uses the GRU-based RNN to model the research evolution of each conference and neglects the scientific influence among them.
- **CONI_V**: It is another variant of CONI, which utilizes the information from peer conferences directly instead of making selection with the attention mechanism.

Model Settings

To pre-train the initialized representations of topical words in the above-mentioned research areas. we construct a large-scale corpus by collecting papers in the areas of data mining, artificial intelligence, computational linguistics and computer vision from *arxiv.org*[1], which is a repository of electronic preprints of scientific papers. The corpus covers more than *95%* of the words occurring in the experimental dataset. Given the titles and abstracts of the collected papers (*24833 tokens*), the word representations with the dimensionality of 50 are trained using the continuous bag-of-words architecture (Mikolov et al. 2013). Words absent in the set of pre-trained embeddings are initialized randomly. We also set the dimension of hidden state as 50 for CONI, CONI_I and CONI_V. The other parameters of each compared model are set for their best performances experimentally.

Evaluation Methodology

To evaluate the prediction performance, we organize the research topics for each conference in temporal order. The first 70% data is used for training, the following 15% data for the validation and the remaining 15% for testing. We adopt three metrics. The metric **Root Mean Squared Error (RMSE)** measures the precision of the predicted word frequencies. Another two metrics, i.e., **MAP** and **NDCG@K**, are more concerned with the ranking of the predicted words.

Performance Evaluation

The prediction performances of all methods in terms of MAP and RMSE are shown in Table 2. Overall, CONI achieves higher performances than all compared methods in both research areas. Based one the results, some important findings are concluded as follows.

Table 2: Performances of different methods.

Area	Methods	MAP	RMSE
AI	Baseline	0.5073	2.327e-3
	Degroot	0.1967	1.588e-3
	Flocking	0.1941	1.615e-3
	ARMA	0.2958	3.316e-2
	CONI_I	0.5719	9.253e-4
	CONI_V	0.5717	9.150e-4
	CONI	**0.5923**	**8.897e-4**
DM	Baseline	0.4920	3.785e-3
	Degroot	0.3263	2.28e-2
	Flocking	0.3155	2.46e-3
	ARMA	0.2958	3.32 e-3
	CONI_I	0.5024	1.700e-3
	CONI_V	0.5000	1.731e-3
	CONI	**0.5128**	**1.610e-3**

RNN better captures sequential properties. The RNN-based methods including CONI_I, CONI_V and CONI significantly outperform the time series model ARMA. It demonstrate that RNN has a better ability to capture the sequential properties of the research evolution than ARMA, since RNN models the dynamics of all research topics words globally instead of individually modeling the time series of each topical word.

Cross-conference influence is important. The better performances of CONI than CONI_I support our assumption that the research development of a conference is truly influenced by its peer conferences. Besides, we also observe that both influence models Degroot and Flocking have poor performances, which denotes that the influence mechanisms for opinion formation cannot be applied to model the research evolution of mutually influenced conferences.

Attention mechanism benefits. The performance of CONI_V, which does not use the attention mechanism to select the interested information from peer conferences is worse than CONI, and even does not show a significant improvement compared with CONI_I. It reveals that the proposed attention mechanism is able to effectively capture the most relevant information that influences a conferences future research.

CONI works effectively on top topic prediction. For a more detailed analysis, we also evaluate the performances of models on top *K* topical word prediction using the ranking metric NDCG@K. The results are illustrated in Figure 4. CONI again achieves the best results, and the improvements are more significant when *K* is smaller than 30 in both research areas. It reveals that CONI has a stronger ability in predicting top trending words.

Discussion

What are the next rising trending topics? Providing insights into the topics that have the potential to become trending topics in the future is also important for researchers to catch up with the rapid progress of research. We use the pub-

[1] https://arxiv.org/

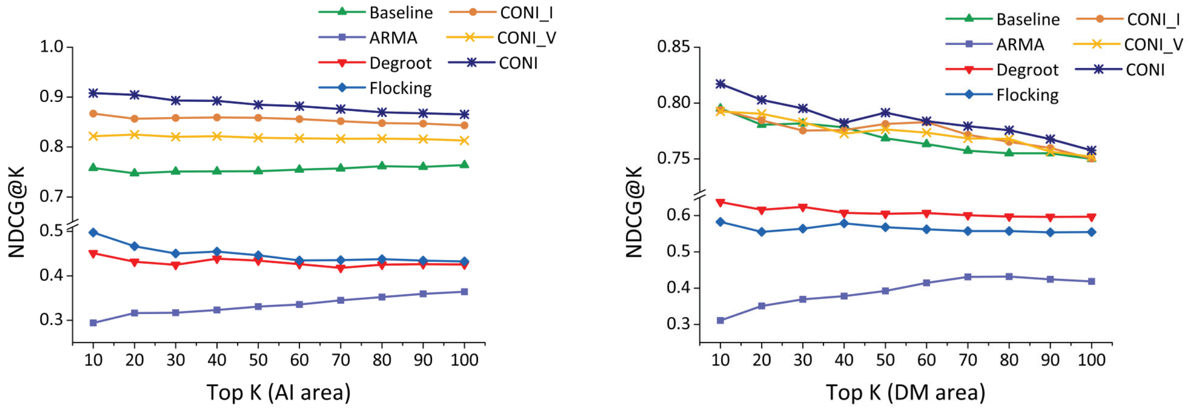

Figure 4: Performances on NDCG@K

Table 3: Predicted 2016 fast-rising trending topics

AAAI	generative, intelligent, spatial, translation, kernel, relation, function, extraction, robotic, reasoning
ACL	sense, discourse, dialogue, web, spoken, bayesian, interactive, temporal, induction, syntactic
CVPR	field, memory, recurrent, illumination, recovery, surface, training, shapes, bayesian, filtering
ICCV	natural, space, optimal, photometric, probabilistic, mixture, representations, group, knowledge, geometric
ICML	reinforcement, decision, active, kernel, online, prediction, temporal, structure, theory, recurrent
IJCAI	boosting, alignment,nonlinear, latent, markov, relational, regularization, active, bounds, tree
NIPS	regression, clustering, application, hierarchical, performance, representation, generalization, combine, knowledge, natural
KDD	feature, pattern, contextual, ranking, trust, ensemble, extraction, sequential, local, latent
CIKM	distributed, prediction, ensemble, selection, large, database, processing, set, support, application
ICDM	ensemble, pattern, contextual, graph, multiple, dynamic, retrieval, compression, similarity, stream
SDM	weighted, dimension, kernel, detection, ensemble, efficient, regression, knowledge, analysis, nearest

lications before 2015 to predict the trending topical words in 2016. Among all the words, we select 10 words that increase the most significantly from 2015 to 2016 in terms of their rankings. The results are presented in Table 3.

Since the publications in 2016 are not included in the dataset, we check the predicted rising trending topical words with the publication records on the conference website. For instance, 'reinforcement', which denotes reinforcement learning is predicted as the fastest rising trending topic in ICML in 2016. According to the publication records of ICML, the percentage of publications related to 'reinforcement' increases to 8/322 in 2016 compared with 3/270 in 2015. In 2017, the percentage of papers working on 'reinforcement' even increases up to 21/466. With the predicted rising topics, researchers could change to explore new promising research topics in their future studies.

Who is the most influential peer? With the proposed model, the scientific influence strengths among conferences are learned from the correlation of their research topics, and are denoted by the parameter $\boldsymbol{A}_{i*}$ for each conference c_i. We normalize the learned influence vector $\boldsymbol{A}_{i*}$ and plot it as a row in Figure 5, which represents the influence strength c_i receives from each of its peers.

From Figure 5, a lot of interesting findings about the conference community in AI and DM areas can be observed. For example, we find that the ACL conference which focuses on the natural language processing is greatly influenced by the conference CVPR in the computer vision area, which is consistent with the example shown in Figure 1. Meanwhile, ACL is among the top 3 influential peers for CVPR. It indicates that the research communities in the computer vision and natural language processing have strong mutual influence. Besides the above finding, more observations about the research community can be inferred from the figure, and we do not discuss in this paper due to the space limitations.

Related Work

Research Trend Analysis

In the camp of bibliometrics and scientometrics, researchers have been studying how to discover the emerging research topics for a long time. They treat the scientific literature as the citation networks and proposed different methods based on the citation network analysis (Shibata et al. 2008) or temporal citation patterns (Daim et al. 2006; Small 2006) to discover which research area will rapidly grows. In contrast to this line of work, our focus is not on the citations, but on predicting the research topics reflected from the text of the publications.

There is another branch of research that tries to understand the trending topics in scientific papers from the text information. Existing models of this branch are mainly based on the topic model, i.e., LDA (Blei, Ng, and Jordan 2003),

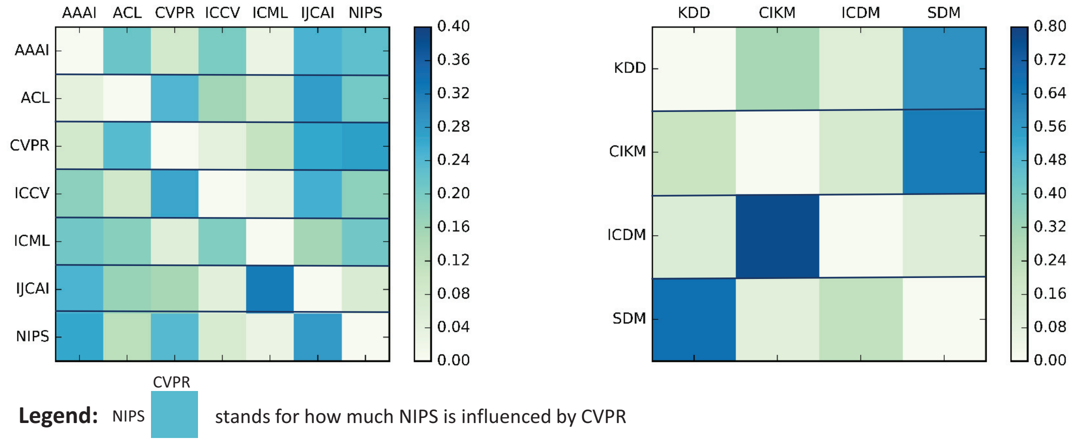

Figure 5: Who is the most influential peer?

which is initially applied to statically detect different research topics from corpus of scientific papers (Griffiths and Steyvers 2004; Steyvers et al. 2004). Considering the topic dynamics over time, these dynamic topic models are developed to discover when and where the topics evolve and capture the dependencies among different topics (Wang and McCallum 2006; Wang, Zhai, and Roth 2013). However, these models lack the ability of predicting future trending topics since they only focus on exploring how the content of a topic forms and changes over time. Recently, Prabhakaran et al., (2016) proposed that the rhetorical roles that authors ascribe to topics (as methods, as goals, as results,etc.) can be the clues of topic trends. They combined the topic model with rhetorical role analysis to predict the rise and the fall of each detected topic. Although the model is able to predict the trends, it cannot precisely produce the popularity of all topics by ignoring the correlation among different topics. Different from the above work, our proposed model globally tracks the evolution of all research topics and can be directly applied to trending topics prediction.

Influence Modeling

Measuring the scientific influence is very important to allocate the efforts and resources in science. Several indexes based on the citation records such as h-index (Hirsch 2005) or g-index (Egghe 2006) are proposed, and have been widely used to estimate the impact of an author or a paper over the whole research community. However, this macro view can hardly provide deep understanding on the detailed effect of scientific influence. Furthermore, several studies were developed to detect the scientific influence of articles from the citation data and textual content (Shen et al. 2016; Foulds and Smyth 2013). However, their studies did not attempt to explore the the future research trending topic prediction by employing the detected scientific influence, which is the focus of this paper.

On the other hand, the effects of influence in the opinion prediction has been well studied. Most popular work formulates each user's opinion as a digit and studies the correlation between a person's future opinion and her/his friends' opinions through learning the interpersonal influence. Several models with different influence assumptions are proposed (DeGroot 1974; Hegselmann and Krause 2002; Chen et al. 2016). Although these models achieve good performances on predicting personal opinion behaviors, the idea has never been applied to reveal the influence mechanisms among scientific conferences.

Conclusion and Future Work

In this paper, we propose a correlated neural influence model to study the intrinsic mechanism behind the research evolution of conferences and predict their trending research topics. It has the ability to capture the sequential properties of the research evolution and correlate different conferences under the cross-conference influence. We demonstrate the effectiveness of the proposed model by conducting experiments on a scientific dataset and our proposed model shows the best results compared with state-of-the-art methods. In the future, we will categorize the topical words with different aspects, e.g., algorithm, application, and capture the cross-conference influence in the aspect-level for providing deeper understandings on the research evolution.

Acknowledgements

The work in this paper was supported by Research Grants Council of Hong Kong (PolyU 5202/12E, PolyU 152094/14E), National Natural Science Foundation of China (61272291) and The Hong Kong Polytechnic University (4-BCB5, G-YBJP).

References

Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 961–971.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Brockwell, P. J., and Davis, R. A. 2013. *Time series: theory and methods*. Springer Science & Business Media.

Chen, C.; Wang, Z.; Lei, Y.; and Li, W. 2016. Content-based influence modeling for opinion behavior prediction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2207–2216.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cover, T. M., and Thomas, J. A. 2012. *Elements of information theory*. John Wiley & Sons.

Daim, T. U.; Rueda, G.; Martin, H.; and Gerdsri, P. 2006. Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change* 73(8):981–1012.

DeGroot, M. H. 1974. Reaching a consensus. *Journal of the American Statistical Association* 69(345):118–121.

Egghe, L. 2006. Theory and practise of the g-index. *Scientometrics* 69(1):131–152.

Elman, J. L. 1990. Finding structure in time. *Cognitive science* 14(2):179–211.

Foulds, J. R., and Smyth, P. 2013. Modeling scientific impact with topical influence regression. In *EMNLP*, 113–123.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* 101(suppl 1):5228–5235.

Hegselmann, R., and Krause, U. 2002. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation* 5(3).

Hirsch, J. E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America* 102(46):16569.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Jamali, H. R., and Nikzad, M. 2011. Article title type and its relation with the number of downloads and citations. *Scientometrics* 88(2):653–661.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mikolov, T.; Karafiát, M.; Burget, L.; Cernockỳ, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, 3.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Prabhakaran, V.; Hamilton, W. L.; McFarland, D.; and Jurafsky, D. 2016. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1170–1180.

Shen, J.; Song, Z.; Li, S.; Tan, Z.; Mao, Y.; Fu, L.; Song, L.; and Wang, X. 2016. Modeling topic-level academic influence in scientific literatures. In *AAAI Workshop: Scholarly Big Data*.

Shibata, N.; Kajikawa, Y.; Takeda, Y.; and Matsushima, K. 2008. Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation* 28(11):758–775.

Small, H. 2006. Tracking and predicting growth areas in science. *Scientometrics* 68(3):595–610.

Steyvers, M.; Smyth, P.; Rosen-Zvi, M.; and Griffiths, T. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 306–315. ACM.

Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L.; and Su, Z. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 990–998. ACM.

Vrettas, G., and Sanderson, M. 2015. Conferences versus journals in computer science. *Journal of the Association for Information Science and Technology* 66(12):2674–2684.

Wang, X., and McCallum, A. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 424–433. ACM.

Wang, X.; Zhai, C.; and 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*, 1115–1123. ACM.

Yogatama, D.; Heilman, M.; O'Connor, B.; Dyer, C.; Routledge, B. R.; and Smith, N. A. 2011. Predicting a scientific community's response to an article. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 594–604. Association for Computational Linguistics.