Venue Topic Model enhanced Joint Graph Modelling for Citation Recommendation in Scholarly Big Data

WEI WANG, State Key Laboratory of Internet of Things for Smart City and Department of Computer and Information Science, University of Macau, People's Republic of China
ZHIGUO GONG, State Key Laboratory of Internet of Things for Smart City and Department of Computer and Information Science, University of Macau, People's Republic of China
JING REN, School of Software, Dalian University of Technology, China
FENG XIA*, School of Engineering, IT and Physical Sciences, Federation University Australia, Australia
ZHIHAN LV, School of Data Science and Software Engineering, Qingdao University, China
WEI WEI, School of Computer Science and Engineering, Xi'an University of Technology, China

Natural language processing technologies, such as topic models, have been proven to be effective for scholarly recommendation tasks with the ability to deal with content information. Recently, venue recommendation is becoming an increasingly important research task due to the unprecedented number of publication venues. However, traditional methods either focus on the author's local network or author-venue similarity, where the multiple relationships between scholars and venues are overlooked, especially the venue-venue interaction. To solve this problem, we propose an author topic model enhanced joint graph modeling approach which consists of venue topic modeling, venue-specific topic influence modeling, and scholar preference modeling. We first model the venue topic with Latent Dirichlet Allocation. Then, we model the venue-specific topic influence in an asymmetric and low-dimensional way by considering the topic similarity between venues, the top-influence of venues, and the top-susceptibility of venues. The top-influence characterizes venues' capacity of exerting topic influence on other venues. The top-susceptibility captures venues' propensity of being topically influenced by other venues. Extensive experiments on two real-world datasets show that our proposed joint graph modeling approach outperforms the state-of-the-art methods.

CCS Concepts: • **Information systems** → **Information retrieval**; **Learning to rank**; • **Computing methodologies** → **Artificial intelligence**.

*Corresponding Author

Authors' addresses: Wei Wang, State Key Laboratory of Internet of Things for Smart City and Department of Computer and Information Science, University of Macau, People's Republic of China, wwang@um.edu.mo; Zhiguo Gong, State Key Laboratory of Internet of Things for Smart City and Department of Computer and Information Science, University of Macau, People's Republic of China, fstzgg@um.edu.mo; Jing Ren, School of Software, Dalian University of Technology, China, ch.yum@outlook.com; Feng Xia, School of Engineering, IT and Physical Sciences, Federation University Australia, Australia, f.xia@acm.org; Zhihan Lv, School of Data Science and Software Engineering, Qingdao University, China, lvzhihan@gmail.com; Wei Wei, School of Computer Science and Engineering, Xi'an University of Technology, China, weiwei@xaut.edu.cn.

2375-4699/2020/7-ART000 $15.00
https://doi.org/10.1145/3404995

Additional Key Words and Phrases: Network embedding, academic information retrieval, scientific collaboration, natural language processing

ACM Reference Format:
Wei Wang, Zhiguo Gong, Jing Ren, Feng Xia, Zhihan Lv, and Wei Wei. 2020. Venue Topic Model enhanced Joint Graph Modelling for Citation Recommendation in Scholarly Big Data. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 00, 00, Article 000 (July 2020), 16 pages. https://doi.org/10.1145/3404995

1 INTRODUCTION

Due to the information overload, scholarly recommendation is becoming a research hot topic due to the increasing scholarly big data [26, 30]. Typical recommendation tasks include collaborator recommendation, paper recommendation, and publication venue recommendation. Among them, venue recommendation is an increasingly important research task due to the unprecedented number of publication venues. In the era of scholarly big data, the academia is producing unprecedented amounts of scholarly information. As a consequence, scholars are overwhelmed by numerous choices when selecting a suitable publication venue. This might negatively impact the achievement of academic success in the long run [9]. Despite the information overload issue resulted from increasing number of journals and conference, finding relevant publication venues is further complicated due to the increasing discipline overlap and interdisciplinary collaborations [2]. Publication venue recommendation, i.e., recommending for scholar relevant venues e.g., conferences and journals, has drawn extensive research interests from the fields of digital library as well as computer science. Fig. 1 illustrates the idea of publication venue recommendation. The goal of publication venue recommendation is to find a suitable venue for a given scholar based on his/her publication records in the scholarly big data.

When designing a venue recommendation system, there are mainly two entities to be considered, i.e., the target venues and the scholars. The ultimate goal of venue recommendation is to measure the interactions among venues and scholars properly. However, existing methods can not model these interactions comprehensively, which either focus on author's local network or author-venue similarity. In this paper, we aim to tackle this problem and propose a joint graph modelling approach for publication venue recommendation.

The nature of venue recommendation is to find an efficient matching strategy that connects scholars with venues. Existing venue recommendation methods mainly fall into two paradigms. The first kind of method utilizes the typical collaborative filtering (CF) method for venue recommendation. These methods treat the venue recommendation task as an item-based CF task, where the users are scholars and items are venues. However, it is difficult to gain the scholars' rating to venues due to the limitation of the scholarly datasets. To tackle this limitation, some scholars propose to generate venue rating with auxiliary information, i.e., topic similarity score [1, 27]. However, it is not clear what kind of auxiliary information is suitable and the venue rating matrix is sparse due to scholars' limited publication counts. The second kind of methods leverage the node similarity metrics in network science based on the constructed academic information network [4, 10, 11]. For these methods, the indicators can be direct node similarity, i.e., common friends [13], or indirect node similarity, i.e., random walk score [5, 28]. However, these methods can not capture the topic similarity between scholars due to the overlook of scholars' and venues' topics. Meanwhile, it is not clear which academic factor can bias the random walk best.

Besides the limitations addressed above, none of them take into account the relationships between venues [25]. The high variation of topic influence across venues has not been

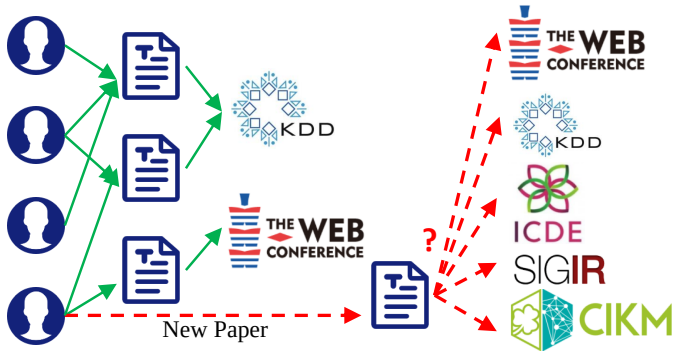

Fig. 1. Illustration of publication venue recommendation.

considered. It is intuitive that scholars are possible to contribute venues with similar topics. In this paper, we aim to tackle these problems.

To this end, we propose a joint graph modelling approach for publication venue recommendation, which consists of three parts, including venue topic modeling, venue-specific topic influence modeling, and scholar preference modeling. Instead of directly employing the topic similarity between venues, we exploit venue-specific topic influence for better recommendation. Specifically, we model the venue-specific topic influence based on three factors, including the topic similarity between venues, the top-influence of venues, and the top-susceptibility of venues. The top-influence characterizes venues' capacity of exerting topic influence to other venues. The top-susceptibility denotes venues' propensity of be topically influenced by other venues. Here, the top-influence of venues and the top-susceptibility are low-dimensional vectors.

The proposed venue-specific influence modelling has two unique advantages: (1) Topic influences between venues are measured asymmetric, enabling to capture the high variation of topic influence across venues. (2) It uses two low-dimensional vectors to represent venue topic influence, which can significantly reduce the number of free parameters so that the data sparsity problem can be better solved.

In summary, the contributions of this paper are as follows,

- We propose a venue-specific topic influence modeling method to measure the venue-venue relations in a asymmetric and low-dimensional way by considering the topic similarity between venues, the top-influence of venues, and the top-susceptibility of venues.
- We design a joint graph modelling approach for venue recommendation, which models the scholars' venue selection as a decision-making procedure, which considers both venue preference and scholar preference.
- We conduct extensive results on two real-word scholarly datasets to evaluate the performance of the proposed model by comparing with several state-of-the-art methods. The results indicates the superiority of the proposed method.

1.1 Organization of the Paper

The organization of this paper is as follows. Section 2 introduces some important notions and preliminaries. Our proposed method is presented in Section 3. Experimental setups and

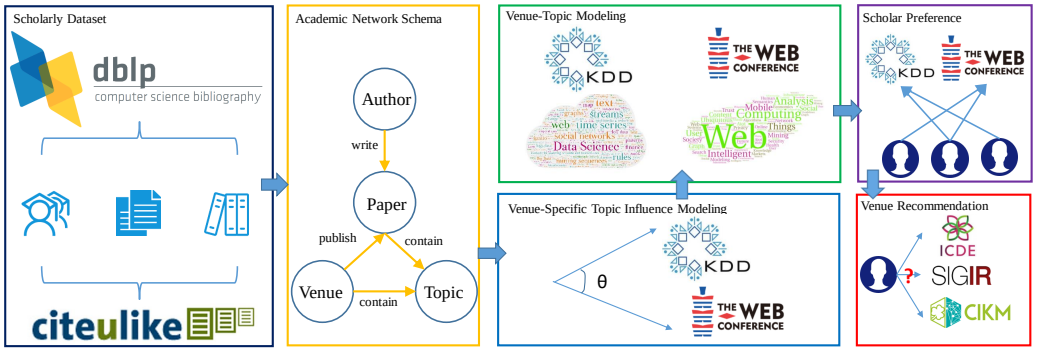

Fig. 2. Framework of proposed method.

results are presented in Sections 4 and 5. Related works are reviewed in Section 6. The paper is concluded in Section 7.

2 PRELIMINARIES AND PROBLEM FORMULATION

In this section, we briefly introduce some related preliminaries and define the research questions.

2.1 Preliminaries

Scholarly Big Data The background of this work is scholarly big data, which is the key foundation for important academic applications such as science and technology management and decision making, domain expert discovery, academic talent management, and innovative technology services. Scholarly big data includes millions of scholarly entities, essay information, institutional information and other academic entities, as well as other large-scale academic related data such as academic social networks, online libraries And academic search engines, etc.

Scholarly Recommendation With the continuous increase of scholarly big data, scholars are constantly facing the problem of information overload. It becomes increasingly difficult to find suitable academic information, i.e., collaborators, paper, and venues. Recommendation system is one of the effective ways to solve information overload, which has achieved effective results in the fields of product recommendation for e-commerce and friend recommendation in online social networks [22, 23]. Similarly, in view of the problem of scholarly information overload, many scholars have proposed several scholarly recommendation methods from different perspectives. Our work tries to recommend venues for authors [1, 3].

2.2 Problem Formulation

In the scenario of venue recommendation, U and V are denoted as the set of scholars and the set of venues, respectively. We represent each scholar $u \in U = I_u$ as a sequence of his/her venue interaction, where I_u denotes all venues that scholar u published based on his/her publication list. We denote each venue v as his/her all publications $P = \{p_1, p_2, p_3, ..., p_t\}$, where t is the total number of its paper collection. For each scholar u, we denote his/her preference as a vector $\overrightarrow{g_u}$. For each venue v, we denote its preference vector as $\overrightarrow{h_v}$, and denote its top-influence vector and top-susceptibility vector as $\overrightarrow{p_v}$ and $\overrightarrow{q_v}$, respectively. We denote w_{uv} as the number of publications that scholar u submits to a venue v. The

Table 1. Descriptions of key symbols.

Symbols	Descriptions
U, V	Sets of scholars, venues
$\overrightarrow{d_v}$	Topic distribution vector of venue v
sim_{ij}	Topic similarity between venues i and j
I_u	Set of venues that scholar u published
$\overrightarrow{g_u}$	Preference vector of scholar u
$\overrightarrow{h_v}$	Preference vector of venue v
$\overrightarrow{p_v}$	Top-influence vector of venue v
$\overrightarrow{q_v}$	Top-susceptibility vector of venue v
w_{uv}	Number of publications scholar u has on venue v
z_{ij}	Topic-specific influence from venue i to venue j

topic similarity between venue i and venue j is denoted as sim_{ij}. The venue-specific topic influence from venue i to venue j is demoted as z_{ij}. The main symbols are shown in Table 1.

Based on the above definitions, our research question can be formulated as follows: **Venue Recommendation**: Given a set of scholars U with publication list I and a set of publication venues V with publication records P, venue recommendation aims at recommending each target scholar $u \in U$ a list of venues $\{v|v \in V\}$ that target scholar is potentially interested in but has not published up to recommendation.

3 PROPOSED METHOD

Our proposed model consists of three parts, as shown in Fig. 2, including venue topic modeling, venue-specific topic influence modeling, and scholar preference modeling. The venue temporal topic modeling aims to capture the topic similarity among different venues. Our major innovation lies in the modeling of venue-specific topic influence based on scholar-venue interaction. For the scholar preference, to avoid bias caused by directly modeling the number of publishing frequency as a numeric quantity, we model a scholar publishing a paper on a specific venue as a process of selecting one target venue from all the potential venues. Next, we will introduce these three parts in details.

3.1 Venue-Topic Modeling

It is no doubt that measuring the topic similarity among venues play an important role in venue recommendation. Topic models, i.e., LDA (Latent Dirichlet Allocation), have been extensively studied for generating topic distributions of a given document [16, 29]. In our case, a venue is composed of a number of papers. To model the topic distribution of a venue i, we use the ATM (Author-Topic Model) [17] as the basic framework. In the venue-topic model, each venue i is regarded as a "document" symbolized in the ATM model, and each paper p_t in this venue is regarded as an "author" symbolized in the ATM model. In other words, all these papers "coauthor" this venue. Thus, we can directly apply the idea of ATM for venue-topic modeling. The graphic schema of the ATM is shown in Fig. 3.

Specifically, in LDA, the modeling process of document collection generation is generally divided into three steps. First, for each document, a topic distribution is sampled from the Dirichlet Distribution. Second, for each word in the document, we choose a separate

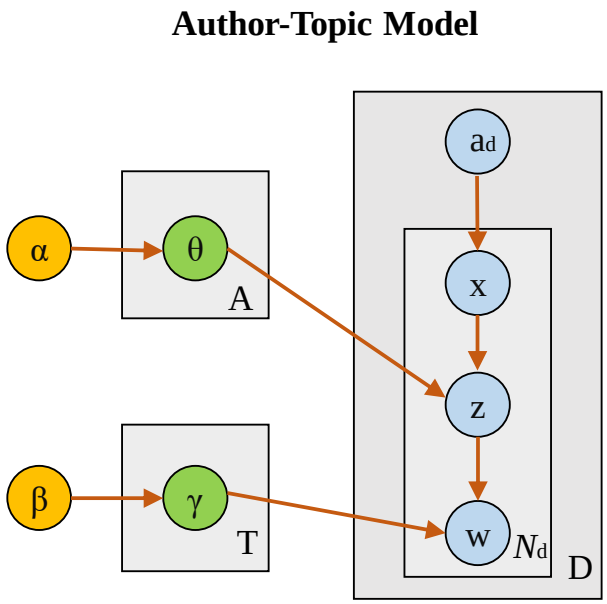

Fig. 3. Illustration of author topic model.

topic based on the topic distribution in the first step. Finally, each word is sampled from a polynomial distribution of words based on the topics sampled in the first and second steps.

This topic model does not explicitly provide information about the author's academic interest. That is, when the information in the document content is useful, the author may have written several articles as well as collaborators, so we do not know how these topics are used in these documents which may also be used to portray the author's academic preferences.

Therefore, we adopt a simple model to model the academic preferences of authors. Suppose there is a group of authors, a_d, ready to write this document D. For each word in the document, an author is randomly and uniformly selected. A word is selected from a word-based probability distribution for the selected author. The author-topic model combines the advantages of the above two models and is defined as follows: it uses a topic-based "performance". At the same time, it models the content of the document and the academic preference of the author . In the author-topic model, a group of authors is represented by ad, and a document is represented by d. Every word about the author in the document is randomly and uniformly selected. Then, in the topic model, a topic is selected from a topic distribution targeted to the author, and then the word is generated from the selected topic.

Based on the venue-topic model, we can gain the topic vector $\overrightarrow{d_v}$ of each venue. Therefore, the topic similarity sim_{ij} between two venues i and j can be calculated based on the cosine similarity as,

$$sim_{ij} = \frac{\overrightarrow{d_i} \cdot \overrightarrow{d_j}}{|\overrightarrow{d_i}||\overrightarrow{d_j}|}. \tag{1}$$

Table 2. Four types of topic similarity function.

Type	Definition
Linear Function	$f(x) = a * x$
Power-law Function	$f(x) = a * x^b$
Exponential Function	$f(x) = a * x^b * e^{cx}$
Hyperbolic Function	$f(x) = \frac{a}{x-b}$

3.2 Venue-Specific Topic Influence Modeling

Different scholars will publish papers on different venues and different venues will attract different scholars. In many recommendation scenarios, users prefer to select similar or neighboring items. Similarly, scholars are possible to publish papers on venues similar with their previous selections. However, different venues may have their own aims and scopes which can not be simply explained by topic similarity.

For a target venue j, we try to model the topic influence from each venue i in the publication list I_u of scholar u. To characterize the high variation of topic influence across venues, we propose a venue-specific topic influence modeling approach, which models the topic influence from venue i to venue j as follows,

$$z_{ij} = \overrightarrow{p_i}^T \overrightarrow{q_j} \times f(sim_{ij}). \tag{2}$$

Here, the vector $\overrightarrow{p_i}$ is the top-influence of venue i, which characterizes venues' capacity of exerting topic influence to other venues. In other words, it denotes the willing of venue i's authors selecting other venues. The vector $\overrightarrow{q_j}$ is the top-susceptibility of venue j, which is defined as the venue j's propensity of attracting authors from other venues. In other words, the willing of other venues' author selecting venue j. sim_{ij} denotes the topic similarity between venues i and j calculated based on Eq. (1). $f(\cdot)$ is the topic similarity function.

We believe Eq. (2) is rational because of:

- $f(sim_{ij})$ is designed to capture the possibility that two venues will be selected by same scholars based on their topic similarity sim_{ij}. In order to capture the fact that scholars are possible to publish papers with similar topic venues, $f(sim_{ij})$ will accordingly increases with the increase of of sim_{ij}. Specifically, we use four types of functions, including linear function, power-law function, exponential function, and hyperbolic function. The details are shown in Table 2.
- We use $\overrightarrow{p_i}^T \overrightarrow{q_j}$ to capture the impact of a scholar's previously selected venue i on the target venue j. Such process can asymmetrically capture the topic influence between two venues so that it can depict the high variation of topic influence across venues more flexible. Previous methods mainly model the venue topic influence by venue interaction matrix, which is time-consuming and not suitable for sparse datasets [28]. Our proposed model represents the topic influence across venues by two low-dimensional vectors so that the number of free parameters is reduced, increasing the ability to solve the data sparsity problem for venue recommendation.
- We model the venue-specific topic influence z_{ij} with a joint approach, which considers the effects from both topic similarity and the internal characteristics of two venues. Based on Eq. (2), given a target venue j, the topic similar and influencing venue will bring about a high z. A dissimilar but influencing venue or, to the opposite, a similar but less influencing venue will result in a relatively smaller z. Moreover, given a same

venue i, y is different for different target venue j because the score z is also related to the intrinsic top-susceptibility vector $\overrightarrow{q_j}$. This makes our proposed influence score venue-specific in terms of involved venues.

We consider the topic influence from each venue i in the publication list I_u of scholar u for a given target venue j. Therefore, given the set of venues I_u of scholar u and Eq. (2), the overall topic influence of I_u on the target venue j can be calculated as

$$\frac{1}{|I_u|}\sum_{i\in I_u} z_{ij} = \frac{1}{|I_u|}\sum_{i\in I_u} \overrightarrow{p_i}^T\overrightarrow{q_j} \times f(\frac{\overrightarrow{d_i}\cdot\overrightarrow{d_j}}{|\overrightarrow{d_i}||\overrightarrow{d_j}|}). \tag{3}$$

3.3 Scholar Preference Modeling

We take advantage of the impact from both scholar preference and venue-specific topic influence to measure a scholar's preference to a target venue. We use o_{uj} to denote scholar u's preference to venue j, which can be calculated as

$$\lambda_{uj} = \overrightarrow{g_u}^T\overrightarrow{h_j} + \frac{1}{|I_u|}\sum_{i\in I_u} z_{ij}, \tag{4}$$

where $\overrightarrow{g_u}$ is scholar u's preference and $\overrightarrow{h_j}$ is venue j's preference.

It is worth mentioning that scholars' publication list I_u show their publishing frequency over all venues, which can be regarded as a kind of implicit scholar preference. In most cases, previous recommendation methods directly employ w_{uv} to depict scholar preference for recommendation. In contrast to this, we model each venue publishing behavior as a process of selecting one target venues from all potential venues. Specifically, we denote the probability that scholar u selects venue j as ρ_{uj}, which can be calculated as,

$$\rho_{uj} = \frac{exp(\lambda_{uj})}{\sum_{n\in V} exp(\lambda_{un})}, \tag{5}$$

where V is the whole set of venues and $\sum_{n\in V} exp(\lambda_{un})$ is the normalization over all venues for a scholar u.

In this way, we can model the scholar u's selection of publication venue j as the result of a decision-making procedure, avoiding the bias caused by directly adopting the w_{uj} as a numeric quantity. The nature is to simulate the process of selecting one target venue from all potential venues. Therefore, a scholar's preference to venues can be regarded as samples extracted from the scholar's preference distribution $\{\rho_{uj}\}$. Therefore, we need to maximize the log-likelihood of scholars' venue preference as,

$$\begin{aligned} F &= \sum_{u\in U}\sum_{j\in V} w_{uv}\log\rho_{uj} = \sum_{u\in U}\sum_{j\in V} w_{uv}\frac{exp(\lambda_{uj})}{\sum_{n\in V} exp(\lambda_{un})} \\ &= \sum_{u\in U}\sum_{j\in V} w_{uv}\log(\frac{exp(\overrightarrow{g_u}^T\overrightarrow{h_j} + \frac{1}{|I_u|}\sum_{i\in I_u} z_{ij})}{\sum_{n\in V} exp(\overrightarrow{g_u}^T\overrightarrow{h_k} + \frac{1}{|I_u|}\sum_{i\in I_u} z_{ik})}). \end{aligned} \tag{6}$$

Note that for venues with which scholars do not publish any paper, w_{uv} is set as 0.

Finally, we recommend new venues to a given scholar u according to the probability ρ_{uj} that the scholar will select venue j. The top K recommendation list for user u is ranked based on the highest probability ρ_{uj} among all the new venues.

3.4 Optimization

In this section, we describe the optimization of the mentioned four latent factors, including $\overrightarrow{g_u}$, $\overrightarrow{h_v}$, $\overrightarrow{p_v}$, $\overrightarrow{q_v}$ and other parameters in venue-specific topic influence function.

Inspired by previous work on maximizing the log-likelihood function F [21], we employ the negative sampling approach [15] for optimization. To avoid noisy distribution, for each selected target venue, we randomly sample α negative venues. Moreover, we use $\mathcal{F}$ to substitute the likelihood function F, which is defined as,

$$\mathcal{F} = \sum_{u \in U} \sum_{j \in V} w_{uv} \sum_{f \in \{j\} \cup NEG(j)} \{\xi_{lf} \log[\eta(\lambda_{uf})] + (1 - \xi_{lf}) \log[1 - \eta(\lambda_{uf}]\}, \tag{7}$$

where $NEG(j)$ denotes the set of negative sampling venues to target venue j, $\eta(\cdot)$ is the sigmoid function, and ξ_{lj} is 1 if $l = j$ and 0 otherwise.

For the new substituted objective function $\mathcal{F}$, we employ the stochastic gradient ascent method to optimize it. During each iteration, a mini-batch of a pair venue is randomly sampled at a ratio γ for optimization. Meanwhile, the sampling probability is determined by the number of publications that scholar u has on venue v, i.e., w_{uv}. Specifically, if a pair of scholar and venue (u, j) is randomly sampled, the four latent vectors can be updated as,

$$\overrightarrow{g_u} = \overrightarrow{g_u} + \tau w_{uj} \sum_{f \in \{j\} \cup NEG(j)} [\xi_{lj} - \eta(\lambda_{uf})] \overrightarrow{h_f}, \tag{8}$$

$$\overrightarrow{h_f} = \overrightarrow{h_f} + \tau w_{uj} [\xi_{lj} - \eta(\lambda_{uf})] \overrightarrow{g_u}, \tag{9}$$

$$\overrightarrow{p_i} = \overrightarrow{p_i} + \tau w_{uj} \sum_{f \in \{j\} \cup NEG(j)} [\xi_{lj} - \eta(\lambda_{uf})] \frac{1}{|I_u|} f(sim_{ij}) \overrightarrow{q_j}, \tag{10}$$

$$\overrightarrow{q_f} = \overrightarrow{q_f} + \tau w_{uj} [\xi_{lj} - \eta(\lambda_{uf})] \frac{1}{|I_u|} \sum_{i \in I_u} f(sim_{if}) \overrightarrow{p_i}, \tag{11}$$

where τ is the learning rate.

4 EXPERIMENTAL SETUPS

In this section, we introduce the experimental setups including, datasets, evaluation metrics, and comparison methods.

4.1 Datasets

We use two real-world scholarly datasets for performance evaluation, including the Aminer dataset [20][1] in the field of Computer Science and CiteULike dataset [2]. For both datasets, the investigated year is from 2000 to 2010. We first do name disambiguation by the method in [18]. Then, we filter out those scholars who have an academic career less than 5 years or with less than 10 publications to screen out those scholars who are not active in academia. The venues sets are constructed based on these scholars' venue interaction. We further screen out those venues which exist less than 5 years or are selected with less than 20 scholars. We use the title and abstract content for venue topic generation for both datasets. The scale of venue topic distribution is set as 100. After preprocessing, we obtain 121,354 scholars over 2,028 venues in DBLP datasets and 63,471 scholar over 3,677 venues in CiteULike datasets.

[1] http://www.aminer.cn/citation

[2] http://www.CiteULike.org/faq/data.adp

For each scholar, we sort his/her publications chronologically. We use the early 80% as the training data and the rest as the testing data.

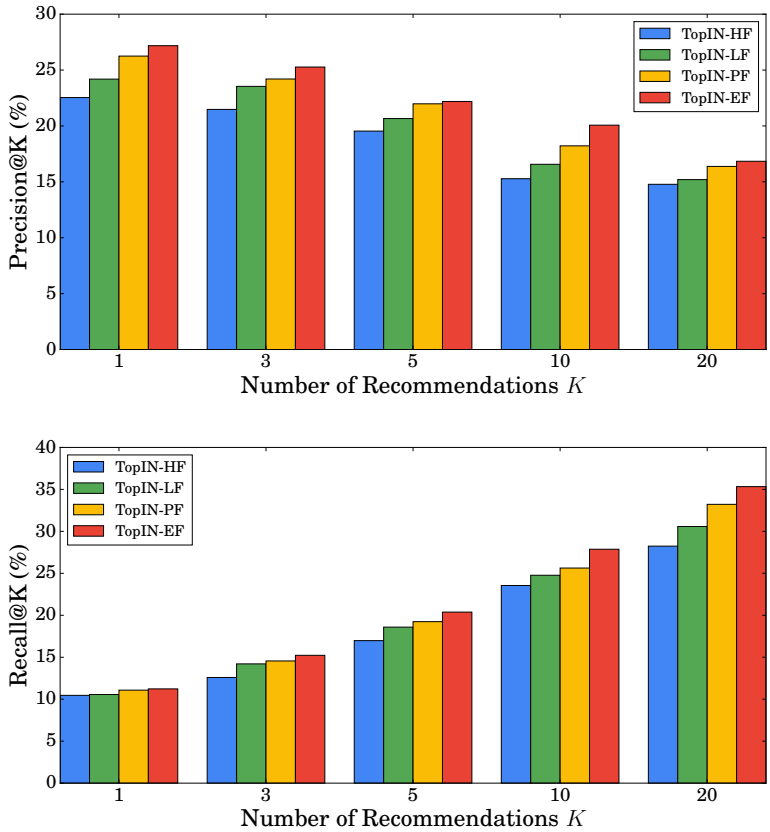

Fig. 4. Performance of TopIN with four types of topic similarity functions on the DBLP dataset.

4.2 Evaluation Metrics

We adopt three widely-used evaluation metrics in recommendation systems, including Precison@K, Recall@K, and F1@K, where K is the number of recommended venues. The Precision@K (P@K) is defined as,

$$P@N = \frac{\#\ of\ correct\ venue\ in\ the\ Top-K\ list}{\#\ of\ K}. \tag{12}$$

The Recall@K (R@K) is defined as,

$$R@N = \frac{\#\ of\ correct\ venues\ in\ the\ Top-K\ list}{total\ \#\ of\ new\ venues}. \tag{13}$$

The F1@K is a combination of precision and recall, which is defined as,

$$F1@K = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{14}$$

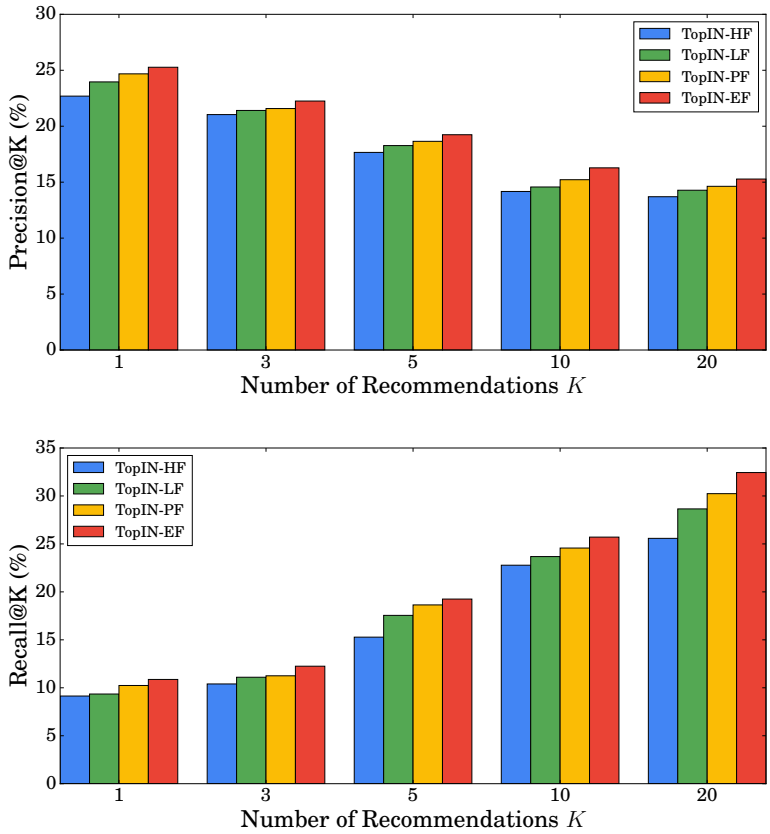

Fig. 5. Performance of TopIN with four types of topic similarity functions on the CiteULike dataset.

For the recommendation list K, we consider 5 values (i.e., 1, 3, 5, 10, 20) in the following experiments.

4.3 Compared Methods

We compare our proposed method with several state-of-the-art venue recommendation approaches, including:

- PAVE [28] It is a biased random walk with restart model which exploits three academic factors, including co-publication frequency, relation weight and researchers' academic in co-publication network.
- AVER [5] It is a personalized venue recommendation model by running a random walk on heterogeneous networks which contain two kinds of associations, coauthor relations and author-venue relations.
- PVR [1] It recommends relevant, specialized scholarly venues in terms of relevance to a given researcher's current scholarly pursuits and interests by calculating scholars' personal venue rating.
- ANPH [13] It recommends appropriate publication venues to scholars by exploring scholar's network of related co-authors and other researchers in the same domain.

Table 3. Recommendation performance comparisons on DBLP datasets in terms of Precision@N, Recall@N, and F1@N.

DBLP	P@K(%)					R@K(%)					F1@K(%)				
	1	3	5	10	20	1	3	5	10	20	1	3	5	10	20
AVER	25.33	24.08	21.39	17.25	15.28	10.84	12.98	17.09	24.59	29.66	15.18	16.87	19	20.28	20.17
PAVE	26.87	24.54	22.39	18.36	15.36	11.08	13.25	18.36	25.44	30.58	15.69	17.21	20.18	21.33	20.45
ANPH	23.68	22.08	19.36	15.29	13.11	8.54	11.54	15.08	21.28	26.25	12.55	15.16	16.95	17.79	14.44
PVR	25.18	23.18	20.56	15.77	14.87	9.38	12.45	16.28	23.58	28.88	13.67	16.2	18.12	18.9	19.63
CFVR	24.28	22.57	21.89	16.28	13.28	9.08	12.28	15.29	22.91	28.39	13.21	15.91	18.01	19.03	18.1
TopIN-EF	**27.18**	**25.27**	**22.19**	**20.07**	**16.84**	**11.23**	**15.23**	**20.38**	**27.87**	**35.33**	**15.89**	**19.01**	**21.25**	**23.34**	**22.8**

Table 4. Recommendation performance comparisons on Citeilike datasets in terms of Precision@N, Recall@N, and F1@N.

CiteULike	P@K(%)					R@K(%)					F1@K(%)				
	1	3	5	10	20	1	3	5	10	20	1	3	5	10	20
AVER	24.17	21.69	18.27	15.27	14.68	10.11	11.23	17.6	23.47	30.56	14.27	14.8	17.93	18.5	19.83
PAVE	24.56	22.07	18.98	15.77	15.08	10.28	11.98	18.25	24.17	31.58	14.49	15.53	18.61	19.09	20.41
ANPH	21.98	19.12	15.44	12.28	9.65	8.24	9.24	12.57	19.57	26.11	11.99	12.46	13.86	15.09	14.09
PVR	23.14	20.21	17.25	14.46	13.18	9.26	10.54	15.28	22.24	28.64	13.23	13.85	16.21	17.53	18.05
CFVR	22.5	19.28	16.54	13.82	12.01	9.09	10.27	14.58	20.65	26.39	12.95	13.4	15.5	16.56	16.51
TopIN-EF	**25.27**	**22.25**	**19.24**	**16.28**	**15.28**	**10.87**	**12.25**	**19.25**	**25.71**	**32.44**	**15.2**	**15.8**	**19.25**	**19.94**	**20.77**

- CFVR [27] It is a collaborative filtering based recommendation system which provides venue recommendations to scholars considering both topic and writing-style information.

5 EXPERIMENTAL RESULTS

In this section, we present the experimental results from the perspectives of topic similarity function selection and performance evaluation.

5.1 Topic Similarity Function Selection

For convenience, we name our proposed method as TopIN. After tuning, we set the sample number α as 10, the sampling ratio γ as 0.1, and the learning rate τ as 0.001 in each iteration. The dimension of the latent vectors is set as 50. The dimension of the venue topics is set as 100. To avoid over fitting, we adopts a L_2 regulation for latent vector optimization, where the regularization coefficient is set as 0.01.

In Table 2, we list four types of topic similarity function. Here, we investigate the influence of these functions and select suitable ones for the following experiments. We name the corresponding TopIN with different functions as TopIN-LF, TopIN-PF, TopIN-EF, and TopIN-HF, respectively. Figures 4 and 5 show the experimental results of TopIN with four types of topic similarity function on DBLP and CiteULike datasets, respectively. We can see from these two figures that TopIN-EF has the best performance on both datasets, which indicate that the exponential function is the best choice to measure the relationships between topic influence and topic similarity. Compared with other functions, exponential function has more parameters which may better capture the uncertain variation of topic influence.

5.2 Performance and Comparison

In this section, we employ the TopIN-EF as the representative of our proposed method for comparing with other state-of-the-art venue recommendation approaches over different

settings. Tables 3 and 4 show the comparison results on two datasets in terms of Precision@K, Recall@K, and F1@K, respectively.

We can observe from these two tables that our proposed TopIN-EF always outperforms baseline methods over all three evaluation metrics on two datasets. Notably, by comparison with the best baseline method PAVE, TopIN-EF achieves a 9.31% improvement in Precision@10, a 9.56% improvement in Recall@10, and a 9.42% improvement in F1@10 on DBLP dataset. Those improvements on CiteULike dataset are 3.23%, 6.37%, and 4.45%, respectively. The best performance of TopIN-EF indicates it is useful to model the venue-specific topic influence in designing venue recommendation systems.

Another interesting observation is that all the methods have a relative close performance when the recommendation number K is 1. But TopIN-EF has a faster performance improvement with the increase of K. This indicates that TopIN-EF is able to recommend more potential venues for a target scholars, which meets the requirements of a serendipitous recommendation [7].

Two random walk based venue recommendation methods, i.e., PAVE and AVER have the best performance over all the baseline methods. These two methods considers various academic factors for biasing random walk model on co-publication network or paper-venue networks. However, the relationships between venues are overlooked. Merely calculating the similarity between venues for recommendation is not sufficient. The ANPH method has the worst performance because it mainly uses target scholar and his/her direct preference for venue recommendation, where the similarity between venues is not considered. This indicates that topic similarity plays an important role in designing a venue recommendation system.

It can be clearly seen from two tables that with the increase of recommendation list K, Precison@K and F1@K decrease, and Recall@K increases accordingly. This is because of their definitions. Take Precison@K for example, with the increase of K, more papers are recommended, whereas the number of correct venue is stable.

By comparing the results on two datasets, we can see that almost all venue recommendation systems have a better performance on the DBLP dataset. One potential reason for this observation maybe the data sparsity. There are more venues in the CiteULike dataset and the average venue number is much larger than DBLP dataset. In fact, DBLP dataset focuses on the field of Computer Science, which CiteULike datasets consists of scholars from two disciplines, i.e., Computer Science and Physics.

In summary, our proposed jointly modeling approach can improve publication venue recommendation. The superiority of the proposed venue-specific topic influence model approach indicates that: (1) Not only topic similarity but also topic influence between venues should be considered for venue recommendation and the topic influence should be measure is uncertain which can not be measure by a simple function. Exponential or power law functions are two potential choices; (2) Modeling both scholar preference and venue preference for characterizing venue-specific topic influence is help for venue recommendation.

6 RELATED WORK

In recent years, with the rapid development of information and communication technology and the continuous development of economic globalization, the available data has become larger and more complex. Academic society is also entering the big data era. With the rapid development of science and technology, scientists all over the world continue to generate huge amounts of data in scientific research and discovery. Scholarly big data has emerged as a research direction [8]. Scholarly big data includes millions of scholarly entities, essay information, institutional information and other academic entities, as well as other large-scale

academic related data such as academic social networks, online libraries And academic search engines, etc.

Scholarly big data includes various academic related data, including journal articles, conference proceedings, dissertations, patent information, books, and academic report documents, etc. In the academic society, the development of scholarly big data has brought about new problems and challenges. Traditional data storage and data analysis methods can no longer meet the needs of scholarly big data acquisition, storage, management, processing, and analysis. Scholarly big data can help people improve our understanding of the academic society from a data perspective, promote the rationalization and efficiency of science and technology, help scholars discover the laws of scientific research, improve innovation capacity and scientific research efficiency, help the country to formulate scientific and technological development strategies, routes and guidelines to provide theoretical basis and method support. Due to information overload, scholarly recommendation becomes one of the key research directions in scholarly data mining, including collaboration recommendation, publication recommendation, and venue recommendation, etc [24, 26].

Publication venue recommendation models recommend suitable and new publication venues according to scholars' publication history extracted from large-scale scholarly datasets [26]. It is one of the most widely-studied research topics in academic recommendations where other typical recommendation tasks include collaborator recommendation [3, 12, 19], paper recommendation [14], and citation recommendation [6]. Most studies on publication venue recommendation to date mainly adopt the traditional recommendation techniques based on citation analysis and scholars' publication history for recommendation [1, 13, 27]. The most frequent techniques include social network analysis and collaborative filtering.

Social network based venue recommendation approaches mainly adopt the node similarity measurement in network science for recommendation. For example, Luong et al. [13] propose to explore authors' network of related co-authors and other scholars in the same domain for appropriate publication venue recommendation. Chen et al. [5] propose to run a random walk model on a heterogeneous network which contains coauthor relations and author-venue relations for publication venue recommendation. Yu et al. [28] improve this approach by exploiting three academic factors, co-publication frequency, relation weight and researchers' academic level to bias the random walk.

CF-based approaches mainly exploit the auxiliary information for venue rating matrix creation because most scholarly datasets do not contain the venue rating information. For example, Yang et al.[27] consider both scholar topic and writing-style information for recommendation. Alhoori and Furuta [1] propose an adaptive implicit rating technique for venue rating creation.

7 CONCLUSION

In this paper, we propose a joint graph modeling approach based on venue topic modeling for venue recommendation by measuring the venue-specific topic influence. We model the venue-specific topic influence considering three factors, including the topic similarity between venues, the top-influence of venues, and the top-susceptibility of venues. The venue topic similarity is calculated based on venue-topic model and the four types of functions are used to explore the impact of venue topic similarity on venue topic influence. The top-influence characterizes venues' capacity of exerting topic influence to other venues. The top-susceptibility denotes venues' propensity of be topically influenced by other venues. Experimental results on two real-word datasets demonstrates that our proposed model outperforms several state-of-the-art venue recommendation approaches. In future, we will

try to evaluate the performance of our method on other scholarly datasets to verify its scalability.

8 ACKNOWLEDGEMENT

This work is funded by National Key R&D Program of China (Grant No: 2019YFB1600700), The Science and Technology Development Fund, Macau SAR (SKL-IOTSC-2018-2020, FDCT/0045/2019/A1, FDCT/007/2016/AFJ), Guangzhou Science and Technology Innovation and Development Commission (EF005/FST-GZG/2019/GSTIC), Research Committee of University of Macau (MYRG2017-00212-FST, MYRG2018-00129-FST), National Natural Science Foundation of China (No. 61902203) and Key Research and Development Plan-Major Scientific, Technological Innovation Projects of ShanDong Province (2019JZZY020101), and China Postdoctoral Science Foundation (2019M651115).

REFERENCES

[1] Hamed Alhoori and Richard Furuta. 2017. Recommendation of scholarly venues based on dynamic user interests. *Journal of Informetrics* 11, 2 (2017), 553–563.

[2] Rebekah R Brown, Ana Deletic, and Tony HF Wong. 2015. Interdisciplinarity: How to catalyse collaboration. *Nature News* 525, 7569 (2015), 315.

[3] Hung-Hsuan Chen, Liang Gou, Xiaolong Zhang, and Clyde Lee Giles. 2011. Collabseer: a search engine for collaboration discovery. In *Proceedings of the 11th annual international ACM/IEEE joint conference on digital libraries*. ACM, 231–240.

[4] Jinyin Chen, Jian Zhang, Xuanheng Xu, Chenbo Fu, Dan Zhang, Qingpeng Zhang, and Qi Xuan. 2019. E-LSTM-D: A Deep Learning Framework for Dynamic Network Link Prediction. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* (2019). https://doi.org/10.1109/TSMC.2019.2932913

[5] Zhen Chen, Feng Xia, Huizhen Jiang, Haifeng Liu, and Jun Zhang. 2015. AVER: random walk based academic venue recommendation. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 579–584.

[6] Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and C. Lee Giles. 2015. A Neural Probabilistic Model for Context Based Citation Recommendation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*. AAAI Press, 2404–2410.

[7] Noriaki Kawamae. 2010. Serendipitous recommendations via innovators. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval*. ACM, 218–225.

[8] Samiya Khan, Xiufeng Liu, Kashish A Shakil, and Mansaf Alam. 2017. A survey on scholarly data: From big data perspective. *Information Processing & Management* 53, 4 (2017), 923–944.

[9] Xiangjie Kong, Jun Zhang, Da Zhang, Yi Bu, Ying Ding, and Feng Xia. 2020. The Gene of Scientific Success. *ACM Transactions on Knowledge Discovery from Data* 14, 4 (2020). https://doi.org/10.1145/3385530

[10] Jianxin Li, Taotao Cai, Ke Deng, Xinjue Wang, Timos Sellis, and Feng Xia. 2020. Community-diversified Influence Maximization in Social Networks. *Information Systems* 92, 101522 (2020). https://doi.org/10.1016/j.is.2020.101522

[11] Jiaying Liu, Feng Xia, Lei Wang, Bo Xu, Xiangjie Kong, Hanghang Tong, and Irwin King. 2019. Shifu2: A Network Representation Learning Based Model for Advisor-advisee Relationship Mining. *IEEE Transactions on Knowledge and Data Engineering* (2019). https://doi.org/10.1109/TKDE.2019.2946825

[12] Zheng Liu, Xing Xie, and Lei Chen. 2018. Context-aware Academic Collaborator Recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1870–1879.

[13] Hiep Luong, Tin Huynh, Susan Gauch, Loc Do, and Kiem Hoang. 2012. Publication venue recommendation using author network's publication history. In *Asian Conference on Intelligent Information and Database Systems*. Springer, 426–435.

[14] Fanqi Meng, Dehong Gao, Wenjie Li, Xu Sun, and Yuexian Hou. 2013. A unified graph model for personalized query-oriented reference paper recommendation. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 1509–1512.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[16] Welly Naptali, Masatoshi Tsuchiya, and Seiichi Nakagawa. 2010. Topic-dependent language model with voting on noun history. *ACM Transactions on Asian Language Information Processing (TALIP)* 9, 2 (2010), 1–31.

[17] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 487–494.

[18] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. 2016. Quantifying the evolution of individual scientific impact. *Science* 354, 6312 (2016), aaf5239.

[19] Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. 2012. Cross-domain collaboration recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1285–1293.

[20] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 990–998.

[21] Hao Wang, Huawei Shen, Wentao Ouyang, and Xueqi Cheng. 2018. Exploiting POI-Specific Geographical Influence for Point-of-Interest Recommendation.. In *IJCAI*. 3877–3883.

[22] Wei Wang, Junyang Chen, Jinzhong Wang, Junxin Chen, and Zhiguo Gong. 2019. Geography-Aware Inductive Matrix Completion for Personalized Point of Interest Recommendation in Smart Cities. *IEEE Internet of Things Journal* (2019). https://doi.org/10.1109/JIOT.2019.2950418

[23] Wei Wang, Junyang Chen, Jinzhong Wang, Junxin Chen, Jinquan Liu, and Zhiguo Gong. 2019. Trust-Enhanced Collaborative Filtering for Personalized Point of Interests Recommendation. *IEEE Transactions on Industrial Informatics* (2019). https://doi.org/10.1109/TII.2019.2958696

[24] Wei Wang, Jiaying Liu, Zhuo Yang, Xiangjie Kong, and Feng Xia. 2019. Sustainable Collaborator Recommendation Based on Conference Closure. *IEEE Transactions on Computational Social Systems* 6, 2 (2019), 311–322.

[25] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. 2019. Graph WaveNet for Deep Spatial-Temporal Graph Modeling. In *IJCAI*. 1907–1913.

[26] Feng Xia, Wei Wang, Teshome Megersa Bekele, and Huan Liu. 2017. Big scholarly data: A survey. *IEEE Transactions on Big Data* 3, 1 (2017), 18–35.

[27] Zaihan Yang and Brian D Davison. 2012. Venue recommendation: Submitting your paper with style. In *2012 11th International Conference on Machine Learning and Applications*, Vol. 1. IEEE, 681–686.

[28] Shuo Yu, Jiaying Liu, Zhuo Yang, Zhen Chen, Huizhen Jiang, Amr Tolba, and Feng Xia. 2018. PAVE: Personalized academic venue recommendation exploiting co-publication networks. *Journal of Network and Computer Applications* 104 (2018), 38–47.

[29] Yinggong Zhao, Shujian Huang, Xin-Yu Dai, and Jiajun Chen. 2016. Adaptation of Language Models for SMT Using Neural Networks with Topic Information. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 15, 3 (2016), 1–15.

[30] Xiaokang Zhou, Wei Liang, Kevin Wang, Runhe Huang, and Qun Jin. 2018. Academic Influence Aware and Multidimensional Network Analysis for Research Collaboration Navigation Based on Scholarly Big Data. *IEEE Transactions on Emerging Topics in Computing* (2018). https://doi.org/10.1109/TETC.2018.2860051