Accepted Manuscript

PAVE: Personalized Academic Venue recommendation Exploiting co-publication networks

Shuo Yu, Jiaying Liu, Zhuo Yang, Zhen Chen, Huizhen Jiang, Amr Tolba, Feng Xia

PII: S1084-8045(17)30400-9

DOI: 10.1016/j.jnca.2017.12.004

Reference: YJNCA 2024

To appear in: *Journal of Network and Computer Applications*

Received Date: 29 November 2016

Revised Date: 23 November 2017

Accepted Date: 2 December 2017

Please cite this article as: Yu, S., Liu, J., Yang, Z., Chen, Z., Jiang, H., Tolba, A., Xia, F., PAVE: Personalized Academic Venue recommendation Exploiting co-publication networks, *Journal of Network and Computer Applications* (2018), doi: 10.1016/j.jnca.2017.12.004.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

PAVE: Personalized Academic Venue Recommendation Exploiting Co-publication Networks

Shuo Yu[a], Jiaying Liu[a], Zhuo Yang[a,*], Zhen Chen[a], Huizhen Jiang[a], Amr Tolba[b,c], Feng Xia[a]

[a] *School of Software, Dalian University of Technology, Dalian 116620, China*
[b] *Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia*
[c] *Mathematics Department, Faculty of Science, Menoufia University, Shebin-El-kom 32511, Egypt*

Abstract

Academic venues have risen beyond the imagination for the rapid development of information technology. It is necessary for researchers to acknowledge high quality and fruitful academic venues. However, the information overload problem in big scholarly data creates tremendous challenges for mining these venues and relevant information. In this work, we propose PAVE, a novel Personalized Academic Venue recommendation Exploiting co-publication networks. PAVE runs a random walk with restart model on a co-publication network which contains two kinds of associations, coauthor relations and author-venue relations. We define a transfer matrix with bias to drive the random walk by exploiting three academic factors, co-publication frequency, relation weight and researchers' academic level. PAVE is inspired from the fact that researchers are more likely to contact those who have high co-publication frequencies and similar academic levels. Additionally, in PAVE, we consider the difference of weights between two kinds of associations. Extensive experiments on DBLP data set demonstrate that, in comparison to relevant baseline approaches, PAVE performs better in terms of precision, recall, F1 and average venue quality.

Keywords: Big scholarly data, Recommender systems, Academic venue

*Corresponding author
Email address: yangzhuo@dlut.edu.cn (Zhuo Yang)

ACCEPTED MANUSCRIPT

recommendation, Random walk, Network science

1. Introduction

It is challenging to mine useful and effective information in big scholarly data due to information overload [38]. The number of researchers, publications, and academic venues have risen beyond the imagination for the rapid development of information technology. Recommender systems help researchers deal with the problem of rapid growth and complexity of information, and provide users with personalized information services. With the continuous growth in the size of research paper repository, recommendation technology for academic entities has been developed gradually [11]. Nowadays, academic recommender systems mainly focus on four aspects: collaborator recommendation, paper recommendation, citation recommendation and academic venue recommendation [41] [4]. Especially, the immense growth of academic venues makes it troublesome for researchers to choose the most relevant venue, which is witnessed by DBLP [18], a service that provides open bibliographic information on major computer science journals and proceedings. It has recorded 3,711 conferences and 1,391 journals (until 2015). Academic venues recommender systems have substantiated their necessity and importance because they provide researchers with personalized venues information pushing service.

In order to better recommend personalized venues for researchers, we consider the researchers' requirements for venues from the perspective of scientific research progress as followings. (1) Where can researchers obtain high quality venues? (2) What are the most relevant conferences researchers should participate? (3) Which venues are the most suitable for researchers to contribute papers? Firstly, researchers usually get inspirations from papers in high quality venues. When doing research, researchers would be better to follow high-quality conferences and journals, in which we can find more high-quality and relevant publications. Since researchers want to grow fast in certain domain by obtaining more specific knowledge and fresh ideas, they need to study more publi-

ACCEPTED MANUSCRIPT

cations that can inspire them. However, new researchers usually do not know which conferences or journals are better choices for their researches. This is because that there are big differences among different venues in research focus, research method, and writing style, etc. To avoid blindness and detours, it is extremely necessary to recommend researchers more conferences and journals of high quality [30]. Secondly, researchers participate in conferences to communicate with other researchers and promote scientific collaboration. As we all know, almost all of researchers participate in conferences every year. Academic conferences not only serve as the platforms to present research work, but also connect researchers in a domain to have a deep communication and boost the potential collaboration. Thus, researchers can benefit a lot and make progress together [26]. While, how to choose more relevant conferences to attend is a tedious task, especially for those new researchers due to the information overload. Finally, researchers are in need of submitting papers to the most suitable venues of high quality. With the rapid development of both the quantity and variety of publication venues in recent decades, it is difficult to decide where to submit papers. For those experts who have much publication experience, it might be a trivial task that select suitable conferences, journals, or scientific forums to publish their papers since they have already knew well about them. That is, they might have target venues in mind before they finish their papers. However, the junior researchers who have few or no publication records, may be not sure of which specific venue the work should be contributed to. Under the guidance of senior researchers, junior researchers may have some distinct or indistinct target venues to prepare their work. Nonetheless, since submissions may be rejected, researchers always need backup plans. Thus, choosing an appropriate venue will be very essential [35] [40].

In recent years, a variety of approaches relating to academic venue recommendation have been proposed [25, 40, 22, 8, 42, 20]. There are also some smart conference systems or solutions that help improve participation experience and solve the conference recommendation problems [34]. Although there are fruitful methods, systems, or solutions, some factors that may have an influ-

ACCEPTED MANUSCRIPT

ence on recommendation in practice are not roundly taken into consideration. Moreover, some of work recommend venues based on homogeneous network [5]. However, academic network is generally with a composition of authors, keywords, affiliations, and venues, which is heterogeneous indeed. In this work, we recommend venues for researchers based on a heterogeneous network and take the three aforementioned requirements into account as well. We propose a novel Personalized Academic Venue recommendation model Exploiting co-publication networks (PAVE). We firstly integrate the academic entities (i.e., authors, publications, and venues) into a co-publication network [17], which contains two kinds of nodes (author and venue) and two kinds of associations (co-author relations and author-venue relations). Figure 1 shows an example of the co-publication network. Alice, Bob, Cindy and David are four researchers whose papers have been published in the four venues. The links include co-author relations (e.g., the link between Alice and Bob) and author-venue relations (e.g., the links between venue A and Alice). Those links along with the nodes compose the co-publication networks. Furthermore, we introduce three metrics in PAVE: 1) co-publication frequency. It can reflect the occurring times of the relations; 2) relation weight. The two kinds of relations can make differences on the network edges; 3) academic level. Researchers are more likely to contact those who have similar academic levels (Researchers are more likely to approach those who have similar academic level rather than with high academic level, and the researchers with similar academic level performs similar characteristic in some ways. For example, in Figure 1, Alice and Bob are neighbors. If they are of similar academic level, then Venue C and Venue D, which Alice has not published any papers but Bob has already published his work, should be not only taken into account when we recommend venues for Alice, but also with more attention. This is because researchers are more likely to contact other researchers with similar academic levels and publish papers in a venue, which is most likely to accept their papers. Details are in Section 3.2). Based on these three hypotheses, we define a transfer matrix with bias to drive the random walk with restart model (RWR) [2, 12, 36] by introducing these three academic factors, co-publication frequency, relations

weight and researchers' academic level. Besides, we innovatively present a new metric called Ave-Quality to evaluate the performance of recommendation apart from precision, recall and F1 metrics. Ave-Quality can well show the quality of recommended venues. In our experiments, PAVE is proved to be effective in terms of leading a better academic venue recommendation.

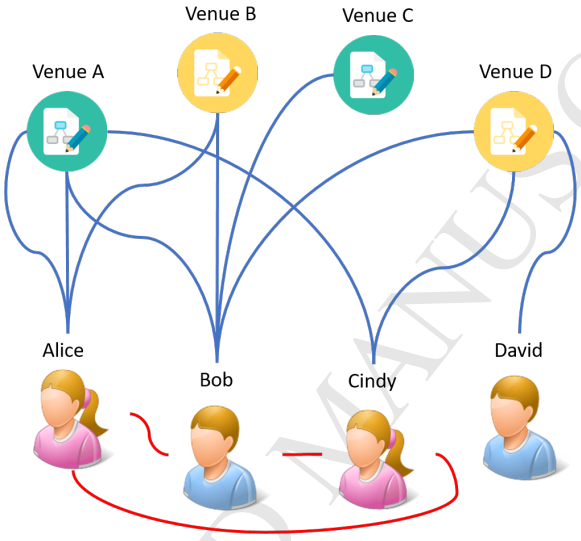

Figure 1: An example of co-publication networks.

In summary, we make the following contributions in this paper.

- We develop an innovative solution based on a random walk with restart model to deal with academic venue recommendation over big scholarly data. The proposed solution is more favourable in terms of achieving remarkable personalized academic venue recommendations.
- To reveal researchers' real intention of academic venues, we define a transfer matrix with bias by utilizing the aforementioned three academic factors, which can lead the random walk running on the co-publication network with preference.
- In addition to precision, recall and F1, We also propose a new metric to evaluate the performance. Extensive experiments on DBLP data set

ACCEPTED MANUSCRIPT

measure the basic RWR model, a topic-based model and a friends-based model for comparison and promising results are presented and analyzed.

The rest of the paper is organized as follows. Related work is discussed in the next section. Section 3 introduces PAVE model. Section 4 presents the performance evaluation results of PAVE, followed by a section dedicated to conclusion.

2. Related work

2.1. Academic Recommendation

Recommender system is proposed to deal with the issues of information overload and help people make decisions by providing accessible and high quality recommendations. Academic recommendation generally consists of academic collaborator recommendation, paper recommendation, citation recommendation, and academic venue recommendation. For the different recommendation, a variety of methods are basically divided into four types: Collaborative filtering (CF) recommendation, content-based recommendation, network-based recommendation, and hybrid recommendation. Details are as follows.

1. *Collaborative filtering recommendation.* CF is a popular and widely accepted approach for recommendation system, like user-based CF, item-based CF and Matrix Factorization (MF). There are some papers focus on academic recommendation by exploiting CF algorithm. For example, Yu et al. [43] present a prediction method based on collaborative filtering for personalized academic recommendation. Liang et al. [20] propose a new probabilistic approach that directly incorporates user exposure to items into collaborative filtering. They consider continuity feature of user's browsing content to help discover collaborative users.

2. *Content-based Recommendation.* Content-based recommendation mainly focuses on the profiles, the content of papers, and the context. It is widely

ACCEPTED MANUSCRIPT

used in academic paper recommendation [19] [**?**] and citation recommendation [13] [3] [7]. Sugiyama and Kan [31] examined the effect of modelling a researcher's past works in recommending scholarly papers to the researcher. The key part of this model is to enhance the profile derived directly from past works with information. The information comes from the past works' referenced papers and papers that cited the work. High quality papers can bring us shining ideas and also we can cite them in our papers. He et al. [13] present the initiative of building a context-aware citation recommendation system and implement a prototype system in CiteSeerX since it is challenging to obtain the relevant papers of high value. Caragea et al. [7] propose an application of Singular Value Decomposition to build a reliable citation recommendation system and to recommend the most relevant citations. Pan and Li [24] use topic model techniques to make topic analysis on research papers.

3. *Network-based Recommendation.* In academia, collaboration makes researchers more fruitful and productive. Friends-based model is a kind of neighborhood-based recommendation approach, which is simple and fundamental in social network-based recommendation methods. Lopes et al. [21] present an innovative approach to recommend collaborations on the context of academic social networks. Specifically, they introduce the architecture for such approach and the metrics involved in recommending collaborations and also present an initial case study to validate their approach. West et al. [33] propose a citation-based method that makes it possible to recommend multiple scales of relevance for different users by using the hierarchical structure of scientific knowledge. Xia et al. [37] consider features of different researchers and propose a novel recommendation method which results in better recommendations.

   In addition, based on the collaboration network, Random Walk model is frequently used to analyze the network. Fouss et al. [12] use a Markov-chain model of random walk to compute similarities between elements of

ACCEPTED MANUSCRIPT

a graph. Stokes et al. [29] use a biased random walk to estimate the expected time of finding a maximum degree node in a graph. Xia et al. [36] present MVCWalker (Random Walk-Based Most Valuable Collaborators Recommendation Exploiting Academic Factors), which takes three academic factors, i.e. coauthor order, latest collaboration time, and times of collaboration into consideration. They compare MVCWalker with the basic model of RWR and a common neighbor-based model in various aspects and achieve better performance. Extraordinary, researchers have already begun to study weights in random walk model using supervised learning algorithm. Lars and Jure [2] develop a method based on Supervised Random Walks that in a supervised way learns how to bias a PageRank-like random walk on the network so that it visits given nodes (i.e., positive training examples) more often than the others. Similarly to this goal, we propose the transfer matrix with bias by introducing three academic factors in this work.

4. *Hybrid-based Recommendation.* For the collaboration recommendation, with mined contents getting more and more, hybrid-based methods [16] [10] [32] come out gradually. Cohen and Ebel [10] focus on one particular flavor of context-based collaborator recommendation in a social network, given a set of keywords. However, collaborations among different domains broaden our researches a lot. Tang et al. [32] analyze the cross-domain collaboration data from research publications and propose the Cross-domain Topic Learning (CTL) model for ranking and recommending potential cross-domain collaborators. They considered a linear combination of the scores obtained by the Content-based and the CF methods.

Cold start issue is one of the most fundamental and intractable issues in recommender system. When tackling cold start problem, these above mentioned methods perform differently. CF recommendation methods rely on history collaboration relationships of other scholars, which results in poorer performance comparing with other kinds of methods. In contrast to CF, Content-based rec-

ACCEPTED MANUSCRIPT

ommendation methods mitigate this issue to some extent since these methods focus on researchers' own profiles. Nevertheless, content-based recommendation methods suffer from cold start challenge when new scholars do not have their own history profiles or collaboration relationships [1]. Network-based and hybrid-based recommendation methods perform better in solving cold start issues [27].

2.2. Academic Venue Recommendation

Plenty of studies have been done on academic collaborator recommendation, academic paper, and citation recommendation and conference session recommendation, while few focuses on the academic venue recommendation. The traditional way of recommending a venue to a researcher is by analyzing her/his papers and comparing it to the topics of different conferences using content-based analysis. However, this approach is not so precise due to mismatches caused by ambiguity in text comparisons. As a result, many researchers focus on social network based and CF methods. Additionally, some social aware approaches and hybrid methods have also been proposed for academic venue recommendation as mentioned above.

Previous studies have already done some work. Yang et al. [40] propose a memory-based neighborhood collaborative filtering model to recommend venues by incorporating both topic and writing-style information of papers. They assume that papers and venues are distinguishable by their writing styles [39]. Pham et al. [25] propose a clustering approach based on the social information of users to derive the academic recommendation. They utilize clustering techniques to improve the accuracy of collaborative filtering. However, this approach mainly involves predicting the publishing venue for a manuscript. Similarly, Luong et al. [22] propose a social network based approach to recommend publication venues by exploring author's network of related co-authors and other researchers in the same domain. In addition, Asabere et al. [42] propose a socially aware based approach to recommend presentation session (community) venues to participants based on high research interest similarity, strong social re-

ACCEPTED MANUSCRIPT

lations, and the matching of contextual information between the presenters and participants at the conference venue. Similarly, Xia et al. [35] propose a presentation session recommender for smart conference participants by utilizing social properties such as tie strength and degree centrality. Hornick et al. [14] provide a framework for extending preference-based recommender systems to deal with problems such as the conference recommendation problem. Huynh Hoang [15] proposed a collaborative knowledge model running on the collaborative network based on the combination of graph theory and probability theory, which aims at supporting publication venue recommendation. Besides, Wongchokprasitti et al. [34] present a design for a community-based conference navigator system collecting the wisdom of community to help conference participants examine the schedule of paper presentation and add the most interesting sessions.

Previous works have not recommended venues according to the associations with researchers. In our paper, we describe the academic publishing scene by a co-publication network including author-venue network and co-author network, and model the real publishing process by a RWR model based on graph theory and probability theory. Our academic venue recommendation model, PAVE, is extended from the basic RWR model. We propose the transfer matrix with bias by introducing three academic factors, i.e. co-publication frequency, relation weight, and researchers' academic levels, which ensures that the random walk performs better when making academic venue recommendations.

3. Design of PAVE

In this section, we describe the details of PAVE. Furthermore, we explain how to compute the link importance in the co-publication networks by considering three academic factors into consideration.

3.1. Overview of PAVE

We exploit PAVE to mine specific academic venues and make personalized recommendations for researchers. The model is inspired by the fact that, re-

ACCEPTED MANUSCRIPT

searchers usually desire to keep contact with suitable academic venues, i.e. acknowledging high-quality and fruitful academic venues, participating in most academic conferences which are closely related to their research, and contributing to suitable venues where it is possible for them to publish their research papers and achievements. Additionally, PAVE is the extension from our previous work [9], which proposes a random walk based academic venue recommendations and achieves good recommendation results. In this work, we regard the topic distribution of researchers' publications content and venues' publications content as feature vectors respectively, which are calculated by an LDA (Latent Dirichlet Allocation) model [6]. We define the K argument in LDA as value 10, which means we clustered 10 topics for each venue and researcher. Then, we consider more factors to evaluate the model. Most of all, the three academic factors we introduced, co-publication frequency, relation weight and researchers' academic level, aim at improving the recommendations by biasing the random walk, so that it traverses more easily to the positive nodes. However, in order to improve the academic level of researchers, the high academic level of venues needs to be guaranteed. Therefore, we define a new metric called Ave-Quality to evaluate the academic level of venues recommended. The detailed process of PAVE is described below. Also, the structure of our PAVE model is illustrated in Figure 2.

We model a co-publication network which consists in the author-venue network and co-author network. As shown in Figure 1, there are two kinds of nodes (venues and researchers) and two kinds of links (co-author relations and author-venue relations). Additionally, PAVE is the evolution from a basic RWR model, which has been proved to be suitable for calculating the similarity of nodes in networks. In PAVE, whether a venue should be recommended depends on its importance of the target researcher. The importance is defined by the rank score of the venue, which is determined by two factors, i.e. the number of neighbor nodes and the rank score of incident nodes. The theory seems like PageRank [23], a successful application of RWR, which provides us a suitable

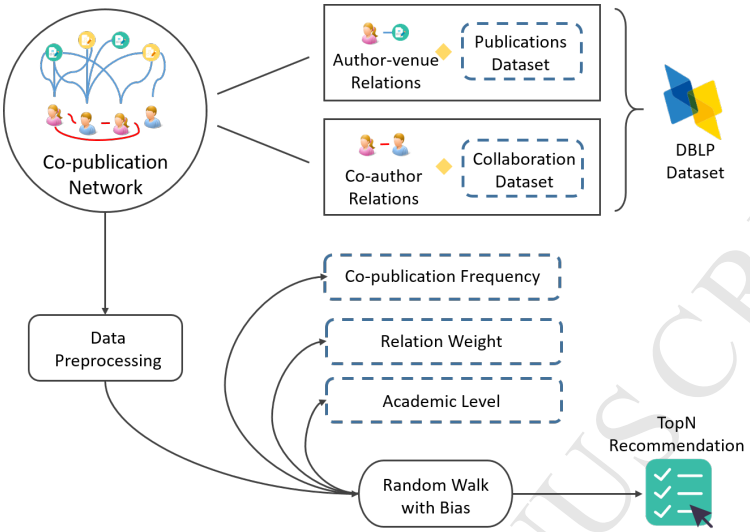

Figure 2: Structure of PAVE.

use for reference. Equation (1) is similar with PageRank in form.

$$AR_u = \frac{1-\alpha}{N} + \alpha \sum_{v \in I_u} AR_v P_{u,v} \tag{1}$$

$\boldsymbol{AR}$ represents the rank score vector. AR_u is the rank score (academic level) of node u. I_u is the set of nodes incident to node u. $P_{u,v}$ is the transition probability from node v to node u. α is the damping factor. N is the number of nodes in the network. PAVE compute the node ranking by driving an imaginary walker randomly walks in the network. The walker has two choices, i.e. with probability α, walking to next node v, which is one of u's direct neighbors ($v \in I_u$)), or with probability $1-\alpha$, returning to source vertex u. Equation (1) represents one step to get one rank score for node u. With respect to all nodes in the whole network, the approach is defined by Equation (2), which is an iterative process.

$$\boldsymbol{AR}^{(t+1)} = \alpha \mathbf{S} \cdot \boldsymbol{AR}^{(t)} + (1-\alpha)\boldsymbol{q} \tag{2}$$

$\boldsymbol{AR}^t$ is the rank score vector at step t. $\boldsymbol{q}$ is a row vector $(q_0, q_1 ..., q_u, ..., q_n)$. For the target node u, $q_u = 1$ and others equal 0. It should be noted that, $\boldsymbol{AR}^0 = \boldsymbol{q}$. $\mathbf{S}$ is the transfer matrix, representing the probability for each node

ACCEPTED MANUSCRIPT

to skip to the next node. For basic RWR model, the cell of matrix $\mathbf{S}$ (i.e. $P_{u,v}$ in Equation (1)) is defined as $\frac{1}{L_v}$, in which L_v is the number of node v's neighbors. It means that, the walker has the same probability to skip to next node. In PAVE, we do some guidance work by introducing three academic factors. The change of $P_{u,v}$ enables the walker to skip based on preference, which will be proved better in section 4 for academic venue recommendation.

With reference to Figure 2, the process of PAVE is described in detail as follows.

- *Step1.* The initial input data is a set of publications with authors' information and venues' information. PAVE firstly extracts the co-author relations and author-venue relations, and then, generates the co-publication networks. There is a link between two authors if they coauthored at least one paper, as well as a link between researcher and venue if the researcher published a paper in the venue.
- *Step2.* After initializing the rank score of nodes and weight of edges, PAVE runs on the network. During the random walk process, the walker skips to next node with a modified probability by considering the three academic factors. The walk will stop until the rank score approximate convergent or the iterations come to the upper limit.
- *Step3.* After getting the convergent rank score of each node, PAVE sorts the venue in accordance to their corresponding rank scores. Finally, remove the venues with which the target author has contacted, the TopN venues are recommended to the target author.

We then present details of how the transfer matrix with bias is computed by considering the three academic factors.

3.2. Transfer Matrix with Bias

A random walk in network is a transition from a node to another node. In the network, if the walker walks from node u, the probability that the walker

ACCEPTED MANUSCRIPT

walks to node v by the next step is only determined by the conditions of node u and node v. That means, the probability that the walker walks to node v is irrelevant to the step before node u. This process is called Markov process. The process of a random walk is actually a Markov process.

Let p_{uv} represent the probability that the walker walks from node u to node v, then p_{uv} can be represented in the following matrix form. This matrix is called transfer matrix.

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1m} \\ \cdots & \cdots & \cdots \\ p_{m1} & \cdots & p_{mm} \end{pmatrix}$$

Obviously, $0 \leqslant p_{uv} \leqslant 1$, $\sum_{v=1}^{m} p_{uv} = 1$.

Let $t_u(n)$ be the probability of the walker stops at node u after n times walk. $t_u(n)$ is called n steps state probability. Then the state vector

$$T(n) = (t_1(n), t_2(n), \cdots, t_m(n)) \tag{3}$$

Apparently, $\sum_{u=1}^{m} t_u(n) = 1$.

According to the total probability formula, we get

$$t_v(n+1) = \sum_{u=1}^{m} t_u(n) p_{uv} \qquad n = 0, 1, 2, \cdots \tag{4}$$

Then we get the general recursive formula

$$T(n) = T(0)P^n \tag{5}$$

It can be known from the Equation (5) that in order to improve the efficiency of the algorithm, we should reduce n to cut down the multiplication times of transition probability matrix. No matter what the final recommend rank in different algorithms is, transition probability matrix with bias can make the walker walk to the suitable venue faster than matrix without bias. This is because the walker walks on purpose under transition probability matrix with bias, which reduce the steps of walking. So both n and the multiplication times

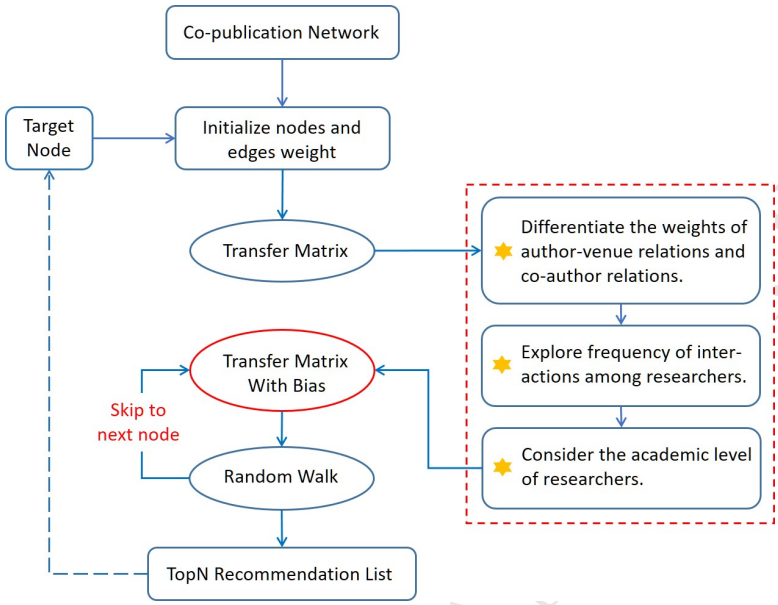

Figure 3: Process of random walk.

can be reduced to cut down running time of algorithm. That is, we should guide the walker to the nodes that are more proper by proposing a transfer matrix with bias instead of transfer matrix without bias. Therefore, we use a transfer matrix with bias in PAVE, of which each element represents the transition probability between two corresponding nodes.

According to the Figure 3, we can clearly see the process of random walk with the new transfer matrix. After initializing nodes and edges weight, we modify the transfer matrix by taking three steps as follows into consideration and get the transfer matrix with bias.

- Differentiate the weights of author-venue relations and co-author relations.
- Explore frequency of interactions among researchers.
- Take the academic level of researchers into consideration.

Referring to the example shown in Figure 1, there are eight academic entities. With respect to recommend venues to Alice, she has never contacted venues C

ACCEPTED MANUSCRIPT

and D. According to the characteristics of the RWR model, the walker can walk from Alice to venues C and D via Bob and Cindy respectively. After several times of iterative walking, venues C and D are recommended to Alice based on the sorted rank score. However, there are several academic factors that can be introduced to meet the real scene. We exploit three of them to redefine the transfer matrix in RWR.

Generally, researchers prefer contacting the academic entities (researchers and venues) which have high frequency of interaction with them, i.e. high publishing frequency in the venue or high collaborating frequency with the researchers. As shown in Figure 1, Alice prefers contacting Bob rather than Cindy because Alice collaborated with Bob twice and with Cindy once. Bob seems to be more important than Cindy for Alice. Furthermore, Alice prefers contacting venue A rather than B, since Alice published two papers in venue A. Based on this assumption, we define co-publication frequency as Equation (6).

$$F_{u,v} = \begin{cases} CP_{u,v} & u \in Author, \quad v \in Venues \\ CT_{u,v} & u, v \in Authors \end{cases} \tag{6}$$

wherein, $CP_{u,v}$ is the count of author u's publications in venue v. $CT_{u,v}$ is author u's collaboration times with author v.

In addition, there are two kinds of associations in co-publication networks, i.e., co-author relations and author-venue relations. In the case of basic random walk model, the difference between these two relations is ignored. Author-venue relations seems to be more important than co-author relations, because the event of publishing a paper in the venue is more preferable when profiling the researchers' interest. This proposition has been proved in subsequent experiments which can lead to better performance when making academic recommendation. We measure the relation weight using Equation (7) based on a ratio β.

$$W_{u,v} = \beta F_{u,v} \tag{7}$$

The ratio β is a variable empirical value which is used to regulate the importance of author-venue relations and co-author relations. The issue is how to set β

respectively for these two kinds of relations to achieve the best recommending performance. We conduct amount of experiments to determine the β. In PAVE, the settings of β is determined as 20 for author-venue relations and 1 for co-author relations, which verifies our hypothesis, the author-venue relations is more important than co-author relations when profiling the researchers' interest.

Finally, we propose an assumption: the interest features of academic entities can be more accurately reflected by similar level neighbors. In case of researchers, they are more likely to contact other researchers with similar academic levels and publish papers in a venue which is most likely to accept their papers. In other words, the relations between similar-level academic entities are more weighty. The walker should walk along these nodes with more probabilities in PAVE. In order to measure the similarity of academic entities, we define a simple metric as shown in equation 8.

$$LevSim_{u,v} = 1 - \frac{\|AR_u - AR_v\|}{\max_{x \in N_u}(\|AR_u - AR_x\|)} \tag{8}$$

The N_u is the neighbors set of node u. Equation (8) aims at discovering the neighbor with smallest rank score disparities based on a normalization method. If node v shows a maximal gap with node u comparing with u's other neighbors, the $LevSim_{u,v}$ will be zero. When computing the transfer probability $S_{u,v}$ from node u to node v, PAVE model adopts Equation (9). The walker can run on the network with a modified bias.

$$\mathbf{S}_{u,v} = \frac{W_{u,v}}{\sum_{x \in N_u} W_{u,x}} LevSim_{u,v} \tag{9}$$

4. Performance Evaluation

We conducted extensive experiments using data from DBLP [18], a computer science bibliography website hosted at University of Trier in Germany. In this section, we describe the statistics of the data set, the evaluation metrics and our experimental procedure for evaluating the performance of PAVE, as well as detailed analysis of the results.

ACCEPTED MANUSCRIPT

4.1. Experimental Settings

To measure the performance of PAVE, we implement three comparison approaches, i.e. the basic RWR model, a Topic-based model and a Friends-based model. The detailed settings are presented following. (1) RWR is a popular model widely used in recommender systems. Similar to popular random walk models, the details and verification method of RWR is resemble to PAVE, except the definition of transfer matrix with bias. The probabilities of skipping to next neighbor node are equal in RWR. (2) The Topic-based method is a content-based recommendation approach in the strict sense, which is also a kind of famous approach for content-based recommender system. The core of the approach is to compute the similarity between researchers and venues. In this implementation, we regard the topic distribution of researchers' publications content and venues' publications content as feature vectors respectively, which are calculated by an LDA model [6]. We define the K argument in LDA as value 10, which means we clustered 10 topics for each venue and researchers. The similarity of researchers and venues is defined by the Cosine Similarity based on these feature vectors. (3) The Friends-based model is a kind of neighborhood-based recommendation approach, which are widely used in social network-based recommendation. The basic idea of friends-based model is to recommend venues according to the number of neighbors who have relations with the venues. In this implementation, we treat the researcher's collaborators and "collaborators of collaborator" as neighbors. If there are many neighbors who contact a venue, the venue should be recommended to the researcher.

4.2. Data Set

DBLP indexes more than 3.35 million articles in computer science. In this experiments, the big scale of data makes it time consuming to process the data and run the PAVE model. To reduce training time, we use a subset of DBLP. This subset covers the field of data mining, involving 74 venues (36 journals and 38 conferences) and 70,326 researchers altogether. Researchers and venues are connected by 163,446 articles in this co-publication network. Covering most

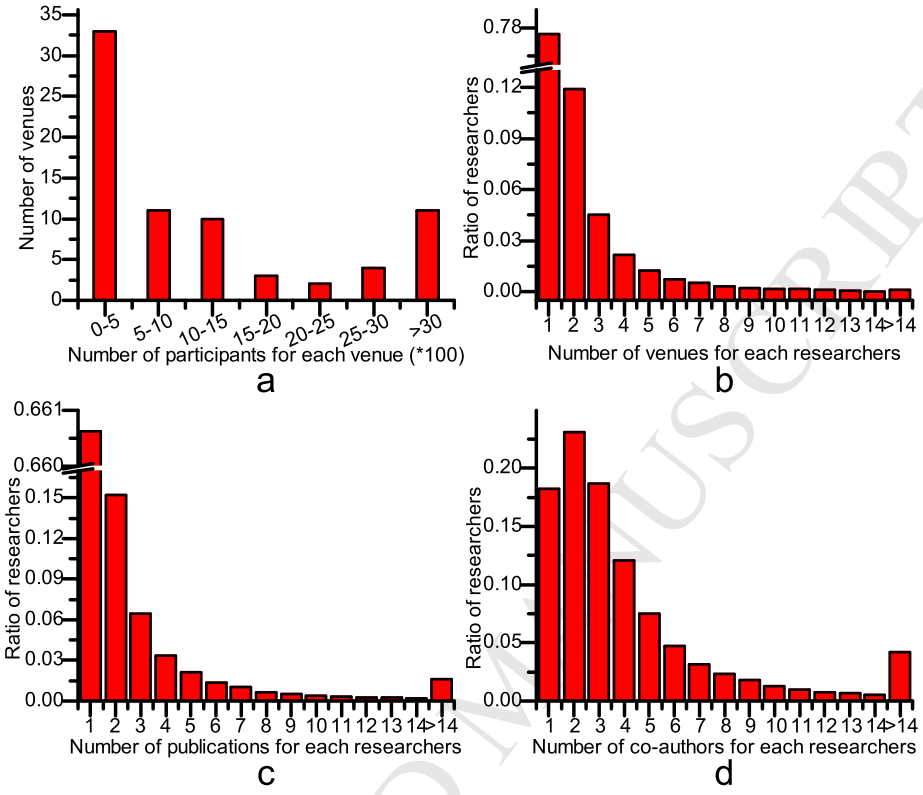

Figure 4: Detailed statistics of the data set from DBLP.

high-quality journals and conferences in the data mining area, the subset has been used by other related studies with no subjective bias [36]. The statistics pertaining to the data set is shown in Table 1. The data set is divided into two parts. The data before year 2011 are chosen as a training set, and the rest as a test set.

The detailed statistical characteristic of this co-publication network is shown in Figure 4. Figure 4(a) describes the scale of participants or contributors for each venue. Almost half of the venues keep no more than 500 researchers. The scale of 11 venues is so large that up to 3,000 researchers publish papers in them. We can also observe that from Figure 4(b), almost 94.09% of these 70,326 researchers contact not more than 3 venues (77.67% for 1 venue, 11.88% for 2 venues and 4.54% for 3 venues). However, there are also some excellent re-

ACCEPTED MANUSCRIPT

Table 1: Statistics of Data Set from DBLP.

Statistics	venues	researchers	articles
Number	74	70326	163446

searchers with high academic level (account for 0.13%) contributing more than 14 venues. Similarly, Figure 4(c) shows the same trend for the number of researchers' publications. Most of them published not more than five papers, but there were also 1.64% researchers publishing more than 14 papers. Figure 4(d) shows the number of co-authors for each researchers. In general, the distributions in Figure 4(b), Figure 4(c), and Figure 4(d) are in the line with long tail distribution, which correspond to the fact that fewer researchers contribute the most products or have the most co-authors. We can conclude that, the degrees (number of neighbors) of most researchers are under 14, which indicates that this data set is very sparse.

All experiments were performed on a 64-bit Linux-based operation system, Ubuntu 12.04 with a 4-duo and 3.2-Ghz Intel CPU, 8-G Bytes memory, and implemented with Python.

4.3. Metrics

In our previous work [9], we employ three popular metrics [28], precision, recall and F1 score, to evaluate the performance of recommendation. In this work, we propose a new metric Ave-Quality to enhance the performance of recommendation. For academic recommendation, we usually get a recommendation list as the output. There is also an accepted list for the target node. So, we can divide the result data into three parts, whose details are shown as follows.

- A: The recommended and collaborated nodes;
- B: The recommended and not collaborated nodes;

ACCEPTED MANUSCRIPT

- C: The collaborated and not recommended nodes.

The definition of precision is shown as below:

$$P = \frac{A}{A+B} \tag{10}$$

The metric recall is defined as:

$$R = \frac{A}{A+C} \tag{11}$$

To get an integrated metric over precision and recall, we can measure the model by F1 score, which is usually called F1 and the equation is:

$$F1 = \frac{2(P * R)}{P+R} \tag{12}$$

In recommender systems, the quality of recommended items is of great concern. The higher quality systems recommended, the better performance the recommendation achieved. It is worth noting that the recommendation could still be in high quality even if the authors paper was rejected. However, such data can hardly be obtained, which makes it difficult to consider rejected papers. As a consequence, in this work we regard a recommendation as high quality when the author's paper was accepted for publication in the test set. To evaluate the quality of the recommended venues generated by PAVE, we propose a metric Ave-Quality based on Google's h5-index[1]. h5-index is a famous and authoritative metric, which represents the venues academic level. A venue is with an h5-index refers that this venue has published h papers each of which has been cited in other papers at least h times in recent 5 years. The formalized definition is shown in equation 13. V is the set of recommended items. M is the length of recommendation list and $H5_v$ is the h5-index of venue v. If the average h5-index of recommended venues is high, that means the PAVE performs well in recommending high quality venues.

$$\textit{Ave-Quality} = \frac{\sum_{v \in V}^{M} H5_v}{M} \tag{13}$$

[1] https:\\scholar.google.comintlenscholarmetrics.html#metrics

In this work, we will use this four metrics to evaluate the performance of PAVE.

4.4. Results and Analysis

In this section, we initially implement several experiments for PAVE, basic RWR, topic-based and friends-based recommendation model on data set discussed above. We randomly choose 100 researchers as target nodes and run PAVE with different target nodes, then, average the value of metrics for the 100 times in the experiments. We repetitively implement such experiments with recommendation lists of different lengths to evaluate the influence of recommendation list on the result. Additionally, PAVE and RWR are implemented with a α of 0.8, which is proved to be appropriate in following experiments.

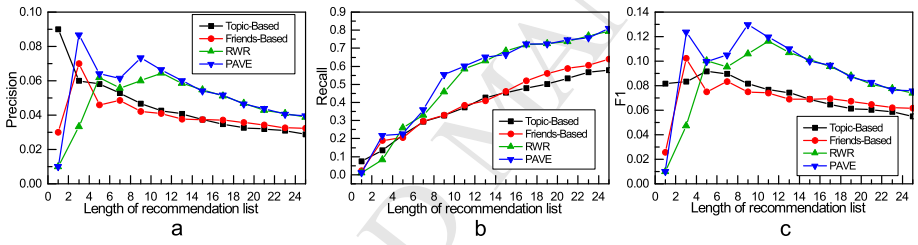

Figure 5: Performance of PAVE, basic RWR, topic-based and friends-based recommendation model.

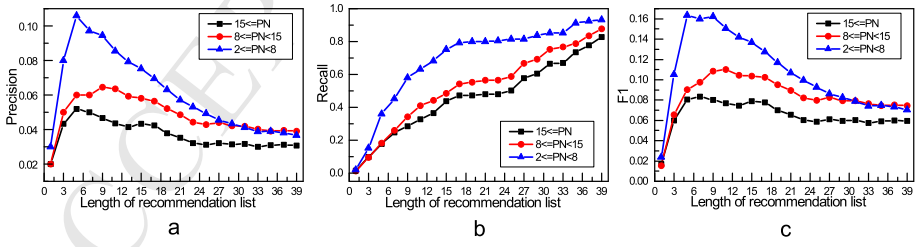

Figure 6: Impact of researchers' publications number (PN) on PAVE.

In recommendation models, higher efficiency generally refers to higher recommendation accuracy with shorter length of recommendation list. Figure 5 shows the performance of PAVE, basic RWR, topic-based and friends-based

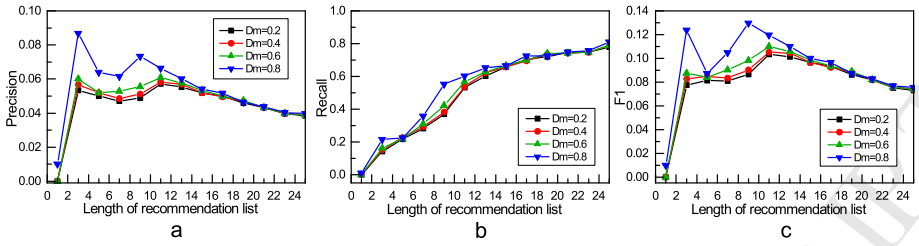

Figure 7: Impact of damping coefficient (Dm) on PAVE.

recommendation model. The x axis represents the length of recommendation list, which is in the range of 1-25. The y axis represents precision, recall and F1 score respectively. In Figure 5(a), topic-based model decreases with the length of recommendation list grows and the other three models decline with fluctuation when the length of recommendation list grows. Topic-based and friends-based recommendation models perform better in precision only when the length of recommendation list is 1. However, PAVE and basic RWR perform better in precision as a whole. A close view of range 1 to 11 on x axis, PAVE achieves higher precision, it comes to a peak value of 8.7% when recommending 3 venues. With the growth of recommendation list, the performance of the four recommendation approaches tend to be similar. In Figure 5(b), the lines rise. PAVE and basic RWR have no significant difference, but their recall perform better than that of topic-based and friends-based approach. With the number of recommended venues reaching the max of venues, the recall approximates to 1. According to Figure 5(c), the F1 score shows similar trend with precision. The F1 score of PAVE reaches the highest value of 12.95% when recommending 9 venues for each researcher. The upgrade rate ($\frac{F1(PAVE)-F1(RWR)}{F1(RWR)}$) is 11.3% in comparison to basic RWR. It is worth mentioned that, PAVE reaches its peak at point 9, while basic RWR achieves the highest F1 score at point 11. That means the recommendation efficiency of PAVE is higher.

These experimental results demonstrate that, the RWR based model can achieve more accurate academic venue recommendation than topic-based and friends-based approaches. Furthermore, our work on transfer matrix with bias

ACCEPTED MANUSCRIPT

improves the performance of PAVE, and makes the recommendation more efficient. Comparing with RWR, the proposed transfer matrix with bias in PAVE makes it possible for the walker walks along with preferred path rapidly and precisely. Based on the analysis of experiment data and the theory of PAVE model, it can be confirmed that PAVE model does improve the recommendation accuracy and the modification of transfer matrix with bias is quite proper.

We also made several extensive experiments to measure the performance of PAVE on different researchers. We mainly focused on the difference of researchers academic level, which is reflected by the number of publications. To some extent, the number of publications can reflect the researchers' contributions and activeness. Generally, in computer science domain, junior researchers show lower academic level with few publications, while senior professor show higher academic level with a lot of high-quality publications. We divide the researchers into three sets: (1) $C1$ contains researchers whose publications range from 2 to 8. This is to ensure the target researcher can appear in both training and testing data sets. Moreover, we ignore the researchers with only one publication; (2) $C2$ contains researchers with 8 to 15 publications; (3) $C3$ contains researchers with more than 15 publications. The experimental results are shown in Figure 6.

From Figure 6, we can see significant differences relating to the effect on different sets of researchers even similar trends are shown in precision, recall, and F1 score respectively. In Figure 6(c), the PAVE achieves the highest value of 16.37% for F1 score at point 5 when making academic venue recommendation for the researchers with 2 to 8 publications. The results mean that, PAVE can perform better at recommending academic venues for researchers with fewer publications, i.e., junior researchers, which meets our innovative intention that recommend academic venues for more effective research and collaboration.

We conduct experiments to show the impact of damping coefficient on PAVE as shown in Figure 7. For the damping coefficient is between 0 to 1, we test four different values of damping coefficient, 0.2, 0.4, 0.6, 0.8, respectively. We can see it also show the similar trends for the metrics precision, recall, and F1.

From Figure 7(a), we can see precision reaches the highest value of 8.7% when the damping coefficient is 0.8%. For recall, it shows an upward trend and also higher with the damping coefficient value of 0.8. Similar to precision, F1 gets the higher value when the damping coefficient is 0.8 as shown in Figure 7(c). All in all, PAVE shows the best performance when the damping coefficient is 0.8.

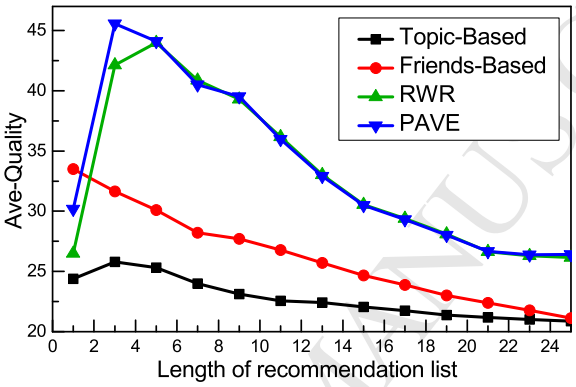

Figure 8: The New Metric Ave-Quality.

Furthermore, we explore the performance of the four models on Ave-Quality. The α is set as 0.8. In Figure 8, we can see PAVE shows the best performance for the Ave-Quality. In other words, PAVE recommends venues of higher academic level for researchers than other models. When the recommendation list is 3, Ave-Quality reaches the peak. With the increasing of recommendation list, Ave-Quality shows a downward trend, but the PAVE is still better than others. This phenomenon corresponds to the theory that random walk model can identify the high-level node with the biased transfer matrix, which means that the three academic factors we explored can lead the rank value transfer along high-level nodes. Therefore, PAVE model can rank the high-level node on the top of the recommending list and finally improve the quality of recommended venues. In conclusion, PAVE shows a better performance than the other baseline methods.

ACCEPTED MANUSCRIPT

5. Conclusion

In this paper, we have focused on academic venue recommendation for researchers based on the big scholarly data which is necessary in current academia. To this end, we have proposed a novel academic venue recommendation model called PAVE, which exploits three academic factors (i.e., co-publication frequency, relation weight and researchers' academic level) to define transfer matrix with bias which drives a random walk with restart model running on co-publication networks. We conduct extensive experiments on a subset of DBLP data set to evaluate the performance of PAVE in comparison to other state-of-the-art approaches: basic RWR, topic-based approaches, and friends-based approaches. The experimental results show that, PAVE outperforms the other approaches in terms of precision, recall, F1 score, and Ave-Quality. According to the extended experiment, PAVE performs better at recommending academic venues for researchers with fewer publications, i.e., junior researchers.

Nonetheless, there is still much work for future study in this direction. We only exploit three academic factors in co-publication networks. There are many other features such as citation relations that need to be explored in PAVE. As future work, more experiments will be performed on other academic data sets.

Acknowledgments

The authors extend their appreciation to the International Scientific Partnership Program ISPP at King Saud University for funding this research work through ISPP#0078.

References

[1] Adomavicius, G., Tuzhilin, A.. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 2005;17(6):734–749.

ACCEPTED MANUSCRIPT

[2] Backstrom, L., Leskovec, J.. Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the 4th ACM international conference on Web search and data mining. ACM; 2011. p. 635–644.

[3] Balog, K., Ramampiaro, H., Takhirov, N., Nørvåg, K.. Multi-step classification approaches to cumulative citation recommendation. In: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval. 2013. p. 121–128.

[4] Beel, J., Gipp, B., Langer, S., Breitinger, C.. Research-paper recommender systems: a literature survey. International Journal on Digital Libraries 2016;17(4):305–338.

[5] Beierle, F., Tan, J., Grunert, K.. Analyzing social relations for recommending academic conferences. In: Proceedings of the 8th ACM International Workshop on Hot Topics in Planet-scale mObile computing and online Social neTworking. ACM; 2016. p. 37–42.

[6] Blei, D.M., Ng, A.Y., Jordan, M.I.. Latent dirichlet allocation. The Journal of Machine Learning Research 2003;3:993–1022.

[7] Caragea, C., Silvescu, A., Mitra, P., Giles, C.L.. Can't see the forest for the trees?: a citation recommendation system. In: Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries. ACM; 2013. p. 111–114.

[8] Chen, J., Chen, G., Zhang, H., Huang, J., Zhao, G.. Social recommendation based on multi-relational analysis. In: IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. IEEE; volume 2; 2012. p. 471–477.

[9] Chen, Z., Xia, F., Jiang, H., Liu, H., Zhang, J.. Aver: Random walk based academic venue recommendation. In: Proceedings of the 24th

ACCEPTED MANUSCRIPT

International Conference on World Wide Web Companion. WWW; 2015. p. 579–584.

[10] Cohen, S., Ebel, L.. Recommending collaborators using keywords. In: Proceedings of the 22nd international conference on World Wide Web companion. WWW; 2013. p. 959–962.

[11] Dhanda, M., Verma, V.. Recommender system for academic literature with incremental dataset. Procedia Computer Science 2016;89:483–491.

[12] Fouss, F., Pirotte, A., Renders, J.M., Saerens, M.. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Transactions on Knowledge and Data Engineering 2007;19(3):355–369.

[13] He, Q., Pei, J., Kifer, D., Mitra, P., Giles, L.. Context-aware citation recommendation. In: Proceedings of the 19th international conference on World wide web. ACM; 2010. p. 421–430.

[14] Hornick, M.F., Tamayo, P.. Extending recommender systems for disjoint user/item sets: The conference recommendation problem. IEEE Transactions on Knowledge and Data Engineering 2012;24(8):1478–1490.

[15] Huynh, T., Hoang, K.. Modeling collaborative knowledge of publishing activities for research recommendation. In: International Conference on Computational Collective Intelligence Technologies and Applications. Springer; 2012. p. 41–50.

[16] Lee, D.H., Brusilovsky, P., Schleyer, T.. Recommending collaborators using social features and mesh terms. Proceedings of the American Society for Information Science and Technology 2011;48(1):1–10.

[17] Lemarchand, G.A.. The long-term dynamics of co-authorship scientific networks: Iberoamerican countries (1973–2010). Research Policy 2012;41(2):291–305.

ACCEPTED MANUSCRIPT

[18] Ley, M.. Dblp: some lessons learned. Proceedings of the VLDB Endowment 2009;2(2):1493–1500.

[19] Li, L., Chu, W., Langford, J., Wang, X.. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In: Proceedings of the fourth ACM international conference on Web search and data mining. ACM; 2011. p. 297–306.

[20] Liang, D., Charlin, L., McInerney, J., Blei, D.M.. Modeling user exposure in recommendation. In: Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee; 2016. p. 951–961.

[21] Lopes, G.R., Moro, M.M., Wives, L.K., De Oliveira, J.P.M.. Collaboration recommendation on academic social networks. In: Advances in Conceptual Modeling–Applications and Challenges. Springer; 2010. p. 190–199.

[22] Luong, H., Huynh, T., Gauch, S., Do, L., Hoang, K.. Publication venue recommendation using author network's publication history. In: Intelligent Information and Database Systems. Springer; 2012. p. 426–435.

[23] Page, L., Brin, S., Motwani, R., Winograd, T.. The pagerank citation ranking: bringing order to the web. Stanford Digital Libraries Working Paper 1999;9(1):1–14.

[24] Pan, C., Li, W.. Research paper recommendation with topic analysis. In: 2010 International Conference on Computer Design and Applications. IEEE; volume 4; 2010. p. V4–264–V4–268.

[25] Pham, M.C., Cao, Y., Klamma, R., Jarke, M.. A clustering approach for collaborative filtering recommendation using social network analysis. J UCS 2011;17(4):583–604.

[26] Pham, M.C., Kovachev, D., Cao, Y., Mbogos, G.M., Klamma, R.. Enhancing academic event participation with context-aware and social rec-

ACCEPTED MANUSCRIPT

ommendations. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE Computer Society; 2012. p. 464–471.

[27] Rohani, V.A., Kasirun, Z.M., Kumar, S., Shamshirband, S.. An effective recommender algorithm for cold start problem in academic social networks. Mathematical Problems in Engineering 2014;2014(2):505–519.

[28] Shani, G., Gunawardana, A.. Evaluating recommendation systems. In: Recommender systems handbook. Springer; 2011. p. 257–297.

[29] Stokes, J., Weber, S.. A markov chain model for the search time for max degree nodes in a graph using a biased random walk. In: 2016 Annual Conference on Information Science and Systems (CISS). IEEE; 2016. p. 448–453.

[30] Sugiyama, K., Kan, M.Y.. Scholarly paper recommendation via user's recent research interests. In: Proceedings of the 10th annual joint conference on Digital libraries. ACM; 2010. p. 29–38.

[31] Sugiyama, K., Kan, M.Y.. Towards higher relevance and serendipity in scholarly paper recommendation. ACM SIGWEB Newsletter 2015;(Winter):4.

[32] Tang, J., Wu, S., Sun, J., Su, H.. Cross-domain collaboration recommendation. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2012. p. 1285–1293.

[33] West, J.D., Wesley-Smith, I., Bergstrom, C.T.. A recommendation system based on hierarchical clustering of an article-level citation network. IEEE Transactions on Big Data 2016;2:113–123.

[34] Wongchokprasitti, C., Brusilovsky, P., Parra-Santander, D.. Conference navigator 2.0: community-based recommendation for academic conferences. In: Workshop on Social Reminder Systems. ACM; 2010. .

ACCEPTED MANUSCRIPT

[35] Xia, F., Asabere, N.Y., Rodrigues, J.J., Basso, F., Deonauth, N., Wang, W.. Socially-aware venue recommendation for conference participants. In: IEEE International Conference on Ubiquitous Intelligence and Computing. IEEE; 2013. p. 134–141.

[36] Xia, F., Chen, Z., Wang, W., Li, J., Yang, L.T.. Mvcwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors. IEEE Transactions on Emerging Topics in Computing 2014;2(3):364–375.

[37] Xia, F., Liu, H., Lee, I., Cao, L.. Scientific article recommendation: Exploiting common author relations and historical preferences. IEEE Transactions on Big Data 2016;2:101–112.

[38] Xia, F., Wang, W., Bekele, T.M., Liu, H.. Big scholarly data: A survey. IEEE Transactions on Big Data 2017;3(1):18–35.

[39] Yang, Z., Davison, B.D.. Distinguishing venues by writing styles. In: Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries. ACM; 2012. p. 371–372.

[40] Yang, Z., Davison, B.D.. Venue recommendation: Submitting your paper with style. In: International Conference on Machine Learning and Applications. IEEE; volume 1; 2012. p. 681–686.

[41] Yang, Z., Yin, D., Davison, B.D.. Recommendation in academia: A joint multi-relational model. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE; 2014. p. 566–571.

[42] Yaw Asabere, N., Xia, F., Wang, W., Rodrigues, J.J., Basso, F., Ma, J.. Improving smart conference participation through socially aware recommendation. IEEE Transactions on Human-Machine Systems 2014;44(5):689–700.

[43] Yu, J., Xie, K., Zhao, H., Liu, F.. Prediction of user interest based on collaborative filtering for personalized academic recommendation. In: 2nd

ACCEPTED MANUSCRIPT

International Conference on Computer Science and Network Technology. IEEE; 2012. p. 584–588.

ACCEPTED MANUSCRIPT