AVER: Random Walk Based Academic Venue Recommendation

Zhen Chen, Huizhen Jiang, Haifeng Liu, Jun Zhang, Feng Xia
School of Software, Dalian University of Technology, Dalian 116620, China
f.xia@acm.org

ABSTRACT

Academic venues act as the main platform of communities in academia and the bridge of connecting researchers, which have rapidly developed in recent years. However, information overload in big scholarly data creates tremendous challenges for mining useful and effective information in order to recommend researchers to acknowledge high quality and fruitful academic venues, thereby enabling them to participate in relevant academic conferences as well as contributing to important/influential journals. In this work, we propose AVER, a novel random walk based Academic VEnue Recommendation model. AVER runs a random walk with restart model on a co-publication network which contains two kinds of associations, coauthor relations and author-venue relations. Moreover, we define a transfer matrix with bias to drive the random walk by exploiting three academic factors, co-publication frequency, weight of relations and researchers' academic level. AVER is inspired from the fact that researchers are more likely to contact those who have high co-publication frequency and similar academic levels. Additionally, in AVER, we consider the difference of weights between two kinds of associations. We conduct extensive experiments on DBLP data set in order to evaluate the performance of AVER. The results demonstrate that, in comparison to relevant baseline approaches, AVER performs better in terms of precision, recall and F1.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

Academic venue recommendation, Big scholarly data, Random walk, Co-publication network

1. INTRODUCTION

Nowadays, the number of the researchers, articles and academic venues has risen beyond the imagination of various research communities due to rapid development of Information Technology (IT). However, the task of mining useful and effective information in big scholarly data is more complex and challenging due to information overload. Academic recommender systems have substantiated their necessity and importance because they objectively provide users with personalized information services. Most academic recommender systems focus on these four problems: collaborator recommendation, paper recommendation, citation recommendation and academic venue recommendation [1].

The immense growth of academic venues makes it troublesome for researchers to choose the most suitable venue, which is witnessed by DBLP, a service that provides open bibliographic information on major computer science journals and proceedings[1]. It has recorded 3711 conferences and 1391 journals. Researchers usually desire to contact suitable academic venues, i.e. acknowledging high-quality and fruitful academic venues, participating in academic conferences or workshops which are closely related to their research, and publishing their papers and research achievements in important and relevant journals. Let's verify these two scenarios. 1) An industrious researcher has made a breakthrough in his research area. Consequently, in order to share his work with other relevant researchers, such an industrious researcher has to find a suitable academic venue (conference). The question is, how can he find the relevant one with significant effects. 2) A junior researcher, i.e. a researcher who is at the initial stage of his research and has few publications, intends to extend his research. But the lack of appropriate academic venues' information is a challenge for him to find a relevant venue to consider and to publish his manuscript. Additionally, although a veteran researcher knows his research area well, he may need a solution relating to cross field venue recommendation.

Considering the inherent requirements, a variety of approaches relating to academic venue recommendation have been proposed [2, 3, 4, 5, 6]. There are also some smart conferences systems or solutions that help improve participation experience and solve the conference recommendation problems [7]. However, most of the researches didn't take the aforementioned problems into consideration. In this work, we propose a novel random walk based Academic VEnue Recommendation model (AVER). We firstly integrate the

WWW '15 Companion, May 18-22, 2015, Florence, Italy

[1]http://dblp.uni-trier.de/db/

academic entities (i.e. authors, publications and venues) into a co-publication network [8], which contains two kinds of nodes (author and venue) and two kinds of associations (co-author relations and author-venue relations). Furthermore, we propose three notable hypotheses, 1) the co-publication frequency can reflect the weight of the relations, 2) the two kinds of relations show difference in importance for researchers, 3) researchers are more likely to contact those who have similar academic levels. Based on these three hypotheses, we define a transfer matrix with bias by introducing three academic factors, co-publication frequency, weight of relations and researchers' academic level. The transfer matrix with bias, which is utilized to drive the random walk with restart model (RWR), have been proved to be effective in terms of leading a better academic venue recommendation.

In summary, we make the following contributions in this paper. 1) To deal with academic venue recommendation based on big scholarly data, we develop AVER based on a random walk with restart model. AVER is more favourable in terms of achieving remarkable personalized academic venue recommendations. 2) 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. 3) We conduct extensive experiments on a subset of DBLP data set to evaluate the performance of AVER. Moreover, we also measure the basic RWR model, a topic-based model and a friends-based model for comparison. Promising results are presented and analyzed.

2. RELATED WORK

Quite a number of recommender systems and algorithms involving the academic venue recommendation have been presented and discussed by various researchers in recent years. These can be classified as content-based, social network-based, hybrid-based and social aware based approaches according to the suggestions of Adomavicius and Tuzhilin [9].

The traditional way of recommending a venue to a researcher is by analyzing her/his paper and comparing it to the topics of different conferences using content-based analysis. However, this approach introduces errors due to mismatches caused by ambiguity in text comparisons. As a consequence, most researchers focus on social network based [4, 5] and collaborative filtering based [2, 3] methods. Additionally, some social aware approaches have also been proposed for academic venue recommendation [6, 10, 11].

Yang et al. [3] proposed an extended version of the neighborhood collaborative filtering model to solve this problem by incorporating style metric features of papers. They assumed papers and venues are distinguishable by their writing styles [12]. Pham et al. [2] proposed a clustering approach based on the social information of users to derive the academic recommendation. They utilized 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. [4] proposed 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. [6] proposed a socially aware based approach to recommend presentation session (community) venues to participants based on high research interest similarity, strong social relations, and the matching of contextual information between the presenters and participants at the conference venue. Similarly, Xia et al. [10] proposed a presentation session recommender for smart conference participants by utilizing social properties such as tie strength and degree centrality. Hornick et al. [11] recommended items from a new disjoint set to users. Their proposed recommender system requires no item ratings, but operates on observed user behavior such as past conference session attendance.

In our work, we describe the academic publishing scene by a co-publication network, and model the real publishing process by a random walk with restart model based on graph theory and probability theory. Similarly, Tin Huynh and Kiem Hoang [13] proposed a collaborative knowledge model running on the collaborative network based on the combination of graph theory and probability theory, which aimed at supporting publication venue recommendation. Chen et al. [5] proposed a recommendation method based on multi relational analysis by combining different relation networks based on optimal linear regression analysis. In our previous works, we proposed a modified random walk with restart model to compute the most valuable academic collaborators recommendation by introducing some academic factors [14], which demonstrates that the RWR model works well in academic social networks. In this paper, our academic venue recommendation model, AVER, is extended from the basic RWR model. We propose the transfer matrix with bias by introducing three academic factors, i.e. co-publication frequency, weight of relations and researchers' academic level, which ensures that the random walk performs better when making academic venue recommendations.

3. DESIGN OF AVER

AVER is designed to mine specific academic venues and make personalized recommendation for researchers. The model is inspired by the fact that, researchers usually desire to keep contact with suitable academic venues, i.e. acknowledging high-quality and fruitful academic venues, participating in academic conferences which are closely related to their research, and contributing to some venues where it is possible for them to publish their research papers and achievements. Additionally, AVER is the evolution from a basic RWR model which has been proved to be suitable for calculating the similarity of nodes in networks. Most of all, the three academic factors we introduced, co-publication frequency, weight of relations and researchers' academic level, aim at biasing the random walk, so that it traverses more easily to the positive nodes. The detailed process of AVER is described below. Additionally, the structure of our AVER model is illustrated in Figure 1.

3.1 Overview of AVER

In this work, we model a kind of co-publication network which are characterized by researchers and academic venues. Figure 2 shows an example of the network. The colorized nodes represent venues A, B, C and D. The three researchers Bob, David and Alice collaborate to write five papers which are published in the four venues respectively (note that Bob

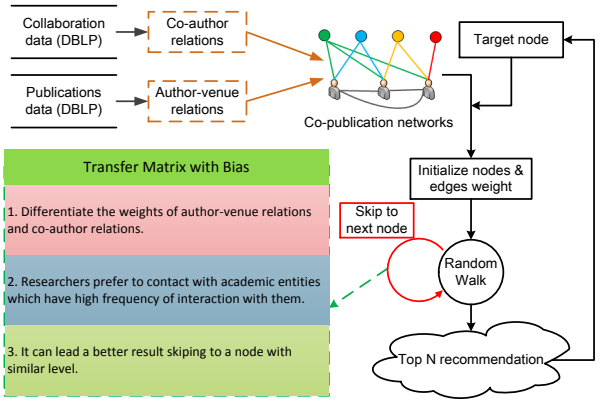

Figure 1: The structure of AVER.

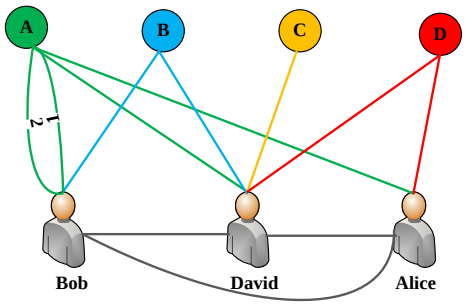

Figure 2: An example of co-publication network

publishes two papers in venue A). The nodes (venues and researchers) along with links (co-author relations and author-venue relations) form the co-publication networks. We define two kinds of node sets, *Venues* and *Authors*.

In AVER, whether a venue should be recommended depends on its importance to 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. Equation (1) describes this theory.

$$AR(p_i) = \frac{1-\alpha}{N} + \alpha \sum_{p_j \in A(p_i)} AR(p_j)P(p_j, p_i) \qquad (1)$$

AR represents the rank score vector. $AR(p_i)$ is the rank score of node p_i. $A(p_i)$ is the set of nodes incident to node p_i. $P(p_j, p_i)$ is the transition probability from node p_j to node p_i. α is the damping factor. When AVER is running on the network to compute node ranking, starting from source node p_0, an imaginary walker randomly walks in the network. The walker has two choices, i.e. with probability α, walking to next node p_x, which is one of p_0's direct neighbors ($p_x \in A(p_0)$), or with probability $1-\alpha$, returning to source vertex p_0. Equation (1) represents one step to getting one rank score for node p_i. With respect to all nodes in the whole network, the approach is defined by Equation (2), which is an iterative process.

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

AR^t is the rank score vector at step t. q is a row vector $(0, ..., 1, ..., 0)$. It should be noted that, $AR_0 = q$. The rank score of the target node is 1, while others' are 0. $\mathbf{S}$ is the transfer matrix, representing the probability for each node to skip to the next node. For basic RWR model, the cell of matrix $\mathbf{S}$ (i.e. $P(p_j, p_i)$ in Equation (1) is defined as $\frac{1}{L(p_i)}$. ($L(p_i)$ is the number of node p_i's neighbors). It means that, the walker has the same probability to skip to next node. In AVER, we do some guidance work by introducing three academic factors. The change of $P(p_j, p_i)$ enables the walker to skip based on preference, which will be proved to be better in section 4 for academic venue recommendation.

With reference to Figure 1, the detailed process of AVER is described below.

- *Step1.* The initial input data is a set of publications with authors' information and venues' information. AVER 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, AVER 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 is approximately convergent or the iterations come to the upper limit.
- *Step3.* After getting the convergent rank score of each node, AVER sorts the venue in accordance to their corresponding rank scores. Finally, removing the venues with which the target author has contacted, the Top-N venues are recommended to the target author.

Below, we present details of how the transfer matrix with bias is computed by considering the three academic factors.

3.2 Transfer Matrix with Bias

Referring to the example shown in Figure 2, there are seven academic entities. With respect to recommending venues to Bob, he has never contacted venues C and D. According to the characteristics of the random walk with restart model, the walker can walk from Bob to C and D via David and Alice respectively. After several times of iterative walking, venues C and D are recommended to Bob 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 random walk with restart model.

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 2, Bob prefers contacting David rather than Alice because Bob collaborated with David twice and Alice once. David seems to be more important than Alice for

Bob. Furthermore, Bob prefers contacting venue A rather than B, since Bob published two papers in venue A. Based on this assumption, we define co-publication frequency as Equation (3) which is a part of the links' weight.

$$F_{i,j} = \begin{cases} cp_{i,j} & i \in Author, \quad j \in Venues \\ ct_{i,j} & i,j \in Authors \end{cases} \tag{3}$$

Where $cp_{i,j}$ is the count of author i's publications in venue j. $ct_{i,j}$ is author i's collaborating times with author j.

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 weight of relations using Equation (4) based on a ratio β.

$$W_{i,j} = \beta F_{i,j} \tag{4}$$

The ratio β is a variable empirical value. In our experiments, β is set as 20 for author-venue relations and 1 for co-author relations.

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 prefer contacting other researchers with similar academic levels and publishing 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 probability in AVER. In order to measure the similarity of academic entities, we define a simple metric as shown in equation 5.

$$LevSim_{i,j} = 1 - \frac{\|AR_i - AR_j\|}{\max_{x \in L(i)}(\|AR_i - AR_x\|)} \tag{5}$$

Equation (5) aims at discovering the neighbor with smallest rank score disparities based on a normalization method. When computing the transfer probability $S_{i,j}$ from node i to node j, our AVER model adopts Equation (6). In Equation (6), the walker can run on the network with a modified bias.

$$\mathbf{S}_{i,j} = \frac{W_{i,j}}{\sum_{x \in L(i)} W_{i,x}} LevSim_{i,j} \tag{6}$$

4. EVALUATION AND ANALYSIS

We conducted extensive experiments using data from DBLP [15], a computer science bibliography website hosted at University of Trier in Germany. In this section, we describe three academic venue recommendation approaches for comparison, statistics of the data set, the evaluation metrics and our experimental procedure for evaluating the performance of AVER, as well as detailed analysis of the results.

4.1 Three Comparison Approaches

To measure the performance of AVER, we performed three comparison approaches, i.e. the basic random walk with restart model (RWR), a topic-based model and a friends-based model.

Similar to popular random walk models, the details and verification method of RWR is just like AVER, except that the definition of transfer matrix with bias. The topic-based method is a content-based recommendation approach in the strict sense. 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's publications content as feature vectors respectively, which are calculated by an LDA(Latent Dirichlet Allocation) model [16]. The similarity of researchers and venues is defined by the Cosine Similarity based on these feature vectors. The friends-based model is a kind of collaborative filtering recommendation approach. Its basis of recommending venues is the number of neighbors who have relations with the venues. In this implementation, we treat 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.

Table 1: Statistics of Data Set from DBLP

Statistics	venues	researchers	articles
Number	74	70326	163446

4.2 Data Set and Metrics

DBLP indexes more than 2.3 million articles in computer science. In our experiments, we use a subset of DBLP. The subset data are all in the field of data mining involving 34 journals and 38 conferences altogether. The statistics pertaining to the data set is shown in Table 1. The data set contains 74 venues and 70326 researches. Researchers and venues are connected by 163446 articles in this co-publication network. We divided the data set into two parts: the data before year 2011 as a training set, and others as a testing set.

The detailed statistical characteristic of this co-publication network is shown in Figure 3. Figure 3(a) describes the scale of participants or contributors for each venue. Almost half of the venues keep not more than 500 researchers. The scale of 11 venues is so large that up to 3000 researchers publish papers in them. We can also observe that from Figure 3(b), almost 94% of these 70326 researchers contact not more than 3 venues. However, there are also some "academic stars" (account for 0.13%) contributing more than 14 venues. Similarly, Figure 3(c) shows the same trend for the number of researchers' publications. Most of them published not more than five papers, but there were also many researchers publishing more than 14 papers. Figure 3(d) shows the number of co-authors for each researchers. We can conclude that, the degrees of most researchers are under 14, which indicates that this data set is very sparse.

We employed three popular metrics, precision, recall and F1 score, to evaluate the performance of AVER. Detailed information about these metrics has been discussed in [14]. All experiments were performed on a 64-bit Linux-based operation system, Ubuntu 12.04 with a 4-duo and 3.2-Ghz Intel

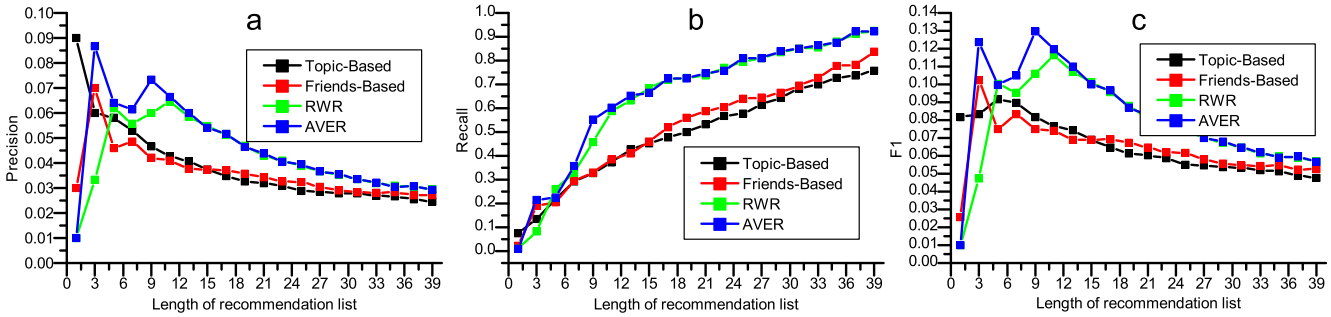

Figure 4: Performance of AVER, basic RWR, topic-based and friends-based recommendation model

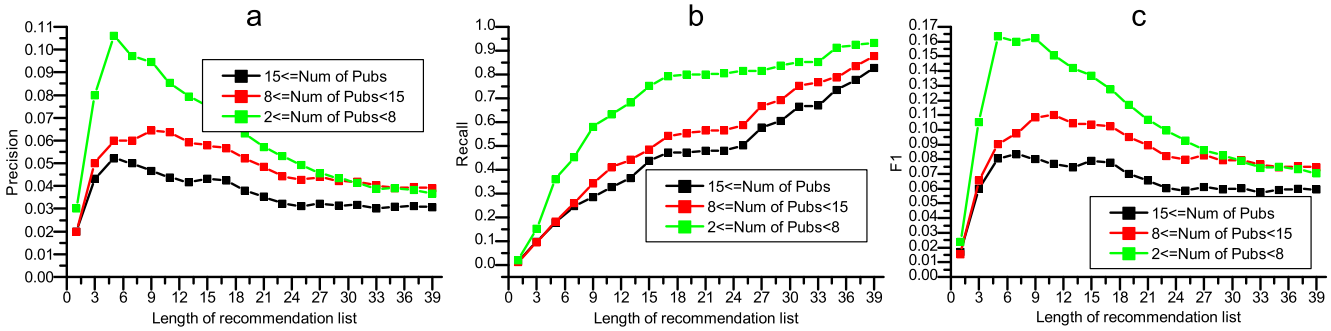

Figure 5: The impact of researchers' publications number on AVER

CPU, 8-G Bytes memory, and implemented with Python.

4.3 Results and Analysis

In this section, we initially performed several experiments for AVER, basic RWR, topic-based and friends-based recommendation model on data set discussed above. Secondly, we measured the performance of AVER when recommending academic venues for researchers at different levels. We randomly chose 100 researchers as target nodes. Additionally, AVER and RWR were run with a damping factor of 0.8.

Figure 4 shows the performance of AVER, basic RWR, topic-based and friends-based recommendation model. The x axis represents the length of recommendation list, which is in the range of 1-39. The y axis represents precision, recall and F1 score respectively. In Figure 4(a), all lines roughly show a coincident downtrend. However, AVER and basic RWR performed better in precision as a whole. A close view of range 1 to 11 on x axis, AVER gets higher precision, it comes to a peak value of 8.7% at when recommending 3 venues. With the growth of recommendation list, the performance of the four recommendation approach tends to be similar. In Figure 4(b), the lines rise. AVER and basic RWR have no significant difference, but their recall performed better than that of topic-based and friends-based approach. With the number of recommended venues reaching the sum of venues, the recall approximates to 1. According to Figure 4(c), the F1 score shows similar trend with precision. The F1 score of AVER reaches the highest value of 12.95% when recommending 9 venues for each researcher. The upgrade rate $(\frac{F1(AVER)-F1(RWR)}{F1(RWR)})$ is 11.3% in comparison to basic RWR. It is worth mentioning that, AVER reaches its peak at point 9, while basic RWR achieves the highest F1 score at point 11. That means the recommendation efficiency of AVER is higher.

These experiments demonstrated that, the random walk with restart based model can achieve more accurate academic venue recommendation than topic-based and friends-based approaches. Furthermore, our work on transfer matrix with bias improves the performance of AVER, and makes the recommendation more efficient.

We also made several extensive experiments to measure the performance of AVER on different researchers. We mainly focused on the difference of researchers academic level, which is reflected by the number of publications. Generally, junior researchers show lower academic level with few publications, while a famous professor shows high academic level with a lot of high-quality publications. We divided the researchers into three sets, i.e. $C1$ contains researchers whose publications range from 2 to 8, $C2$ contains researchers with 8 to 15 publications and $C3$ contains researchers with more than 15 publications. The experimental results are show in Figure 5.

From Figure 5, we can see significant differences relating to the effect on different sets of researchers even though they show a similar trends in precision, recall and F1 score respectively. In Figure 5(c), the AVER achieves the highest

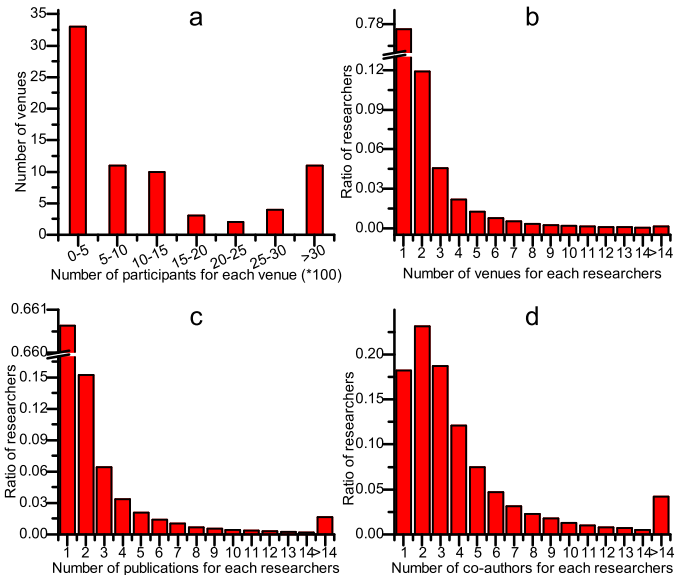

Figure 3: Detailed statistics of the data set from DBLP

value of 16.24% for F1 score at point 9 when making academic venue recommendation for the researchers with 2 to 8 publications. The results mean that, AVER can perform better at recommending academic venues for researchers with fewer publications, i.e. junior researchers, which meets our innovative intention of recommending academic venues for effective research and collaboration.

5. CONCLUSION

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

Nonetheless, there is still room for future study in this direction. We only exploited three academic factors in co-publication network. There are many other features such as citation relations that need to be explored in AVER. As a future work, more experiments should be performed on other academic data sets.

6. REFERENCES

[1] Zaihan Yang, Dawei Yin, and Brian D Davison. Recommendation in academia: A joint multi-relational model. In *ASONAM*, pages 566–571. IEEE, 2014.

[2] Manh Cuong Pham, Yiwei Cao, Ralf Klamma, and Matthias Jarke. A clustering approach for collaborative filtering recommendation using social network analysis. *J. UCS*, 17(4):583–604, 2011.

[3] Zaihan Yang and Brian D Davison. Venue recommendation: Submitting your paper with style. In *ICMLA*, volume 1, pages 681–686. IEEE, 2012.

[4] Hiep Luong, Tin Huynh, Susan Gauch, Loc Do, and Kiem Hoang. Publication venue recommendation using author networkąŕs publication history. In *Intelligent Information and Database Systems*, pages 426–435. Springer, 2012.

[5] Jian Chen, Guanliang Chen, Haolan Zhang, Jin Huang, and Gansen Zhao. Social recommendation based on multi-relational analysis. In *WI-IAT*, volume 2, pages 471–477. IEEE, 2012.

[6] Nana Yaw Asabere, Feng Xia, Wei Wang, Joel JPC Rodrigues, Filippo Basso, and Jianhua Ma. Improving smart conference participation through socially aware recommendation. *IEEE Trans Hum Mach Syst*, 44:689–700, 2014.

[7] Chirayu Wongchokprasitti, Peter Brusilovsky, and Denis Parra-Santander. Conference navigator 2.0: community-based recommendation for academic conferences. In *Proc. SRS*. ACM, 2010.

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

[9] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans Knowl Data Eng*, 17(6):734–749, 2005.

[10] Feng Xia, Nana Yaw Asabere, Joel JPC Rodrigues, Filippo Basso, Nakema Deonauth, and Wei Wang. Socially-aware venue recommendation for conference participants. In *UIC/ATC*, pages 134–141. IEEE, 2013.

[11] Mark F Hornick and Pablo Tamayo. Extending recommender systems for disjoint user/item sets: The conference recommendation problem. *IEEE Trans Knowl Data Eng*, 24(8):1478–1490, 2012.

[12] Zaihan Yang and Brian D Davison. Distinguishing venues by writing styles. In *Proc. JCDL*, pages 371–372. ACM, 2012.

[13] Tin Huynh and Kiem Hoang. Modeling collaborative knowledge of publishing activities for research recommendation. In *ICCCI*, pages 41–50. Springer, 2012.

[14] Feng Xia, Zhen Chen, Wei Wang, Jing Li, and Laurence T Yang. Mvcwalker: Random walk based most valuable collaborators recommendation exploiting academic factors. *IEEE Trans Emerg Top Comput*, 2:364–375, 2014.

[15] Michael Ley. Dblp: some lessons learned. *Proceedings of the VLDB Endowment*, 2(2):1493–1500, 2009.

[16] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.