AVERec: A Random Walk Based Academic Venues Recommendation

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

ABSTRACT

In academia, venue paly the main platforms of academic community and the bridge of connecting researchers, which have rapidly developed in recent years. Consequently, It becomes complex as a result of information overload in the big scholarly data for researchers to keep concerns on high-quality and fruitful academic venues, participate in relevant academic conferences and contribute to influential journals. In this work, we propose AVERec, a novel random walk based academic venue recommendation model. AVERec builds the co-publishing network with two kinds of associations, i.e. co-author relations and author-venue relations. We exploit three academic factors to define a transfer matrix with bias which drives a random walk with restart model running on the co-publishing network. The three academic factors, i.e. co-publishing frequency, weight of relations and similar-level preferred, are inspired by that, researchers are more likely to contact with those who have high co-publishing frequency and similar academic level with them, as well as the weight of the two kinds of associations should be differentiated. We conduct extensive experiments on DBLP data set in order to measure AVERec. The results demonstrate that AVERec significantly improves the performance on precision, recall and F1 when compared to the baseline approaches.

Categories and Subject Descriptors

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

General Terms

Theory

Keywords

Academic venues recommendation, Big scholarly data, Random walk, Co-publishing network

1. INTRODUCTION

In academia, the scale of the researchers, articles and academic venues has risen beyond the imagination of people due to its rapid development. Consequently, it has become a complex task mining the useful and effective information in the big scholarly data because of information overload. The 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: the collaborators recommendation, paper recommendation, citation recommendation and academic venue recommendation [1].

As the main platform of academic community and the bridge of connecting researchers, academic venues have got rapidly developed. DBLP, a service provides open bibliographic information on major computer science journals and proceedings, witnessing the growth of academic venues[1]. It has recorded 3711 conferences and 1391 journals. The scale, dispersion and diversity of academic venues make it troublesome for researchers to choose. Researchers usually desire to contact with suitable academic venues, i.e. keep concerns on high-quality and fruitful academic venues, participating in academic conferences or workshops which are closely related to their research, and contributing to some venues where they are possible to publish their research achievements. Let's think these two cases. 1) An industrious researcher has made a breakthrough in his research area. Hence he want to find an academic venue where he can participate and publish his achievement to share his work. The question is, how can he find the relevant one with enough of effects. 2) An academic novice (i.e. the researcher who is in initial stage of his research and has few publications) intends to extend his research. But the lack of academic venues' information makes him puzzled as a result of that he can not accurately find a relevant venue to be concerned on and to publish his draft. Additionally, although a veteran researcher knows his research area well, he may need a solution to be recommended cross field venues.

Considering the inherent requirements, a variety of approaches relating to academic venues recommendation have been proposed. As Adomavicius and Tuzhilin suggested, these approaches can be classified as content-based, social network-based, hybrid-based and social aware based approaches [2]. There are also some smart conferences systems or solution-

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

s helping improve participate experience and incidentally solve the conferences recommendation problems. However, most of the researches can not complete the aforesaid problems. In this work, we integrated the academic entities (i.e. authors, publications and venues) into a co-publishing network, which contains two kinds of nodes (author and venue) and two kinds of associations (co-author relations and author-venue relations). We proposed three commonsense hypotheses, i.e. the co-publishing frequency can reflect the weight of the relations, the two kinds of relations show difference in importance for researchers, as well as researchers more likely contact with who are in similar academic level with them. AVERec, a novel random walk based academic venue recommendation model was proposed with introducing the three academic factors, co-publishing frequency, weight of relations and similar-level preferred, into a random walk with restart model. In summary, we make the following contributions in this paper. 1) To deal with academic venues recommendation based on big scholarly data, we develop AVERec based on a random walk with restart. AVERec is more favourable in terms of achieving remarkable personalized academic venues recommendation. 2) To reveal researchers' real intention of academic venues, we defined a transfer metrics with bias by utilizing the aforementioned three academic factors. which can lead the random walk running on the co-publishing network with preference. 3) Extensive experiments on a subset of DBLP data set evaluated the performance of AVERec. Moreover, we also measured the basic RWR model, a topic-based model and a friend-based model for comparison. Promising results are presented and analyzed.

2. RELATED WORK

Quite a number of recommender systems and algorithms involving the academic venues recommendation have been presented and discussed by various researchers in recent years.

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 can make errors due to mismatches caused by ambiguity in text comparisons. As a consequence, most researchers focus on social network based [3, 4] and collaborative filtering based [5, 6] methods. Additionally, some social aware approaches are also proposed for academic venues recommendation [7, 8, 9].

Yang et al. [6] proposed an extended version of the neighborhood collaborative filtering model to solve this problem recently by incorporating style metric features of papers. They think papers and venues are distinguishable by their writing styles [10]. Pham et al. [5] proposed a clustering approach based on the social information of users to derive the academic recommendation. They concerned on the clustering techniques to improve the accuracy of collaborative filtering. However, this approach mainly involves predicting the publishing venue for a given draft. And the same case with Luong et al. [3], they 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. [7] proposed a socially aware based approach to recommend presentation sessions (communities) 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. As well as in their another work [8], they proposed good solutions to the problems of venues (sessions/conference) recommendation. Hornick et al. [9] recommended items from a new disjoint set to users. It 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 co-publishing 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 [11] 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. [4] proposed a recommendation method based on multi relational analysis by combining different relation networks based on optimal linear regression analysis. We used to proposed a modified random walk with restart model to make most valuable academic collaborators recommendation by introducing some academic factors [12, 13], which demonstrate the RWR model works well in academic social networks. In this paper, our academic venues recommendation model, AVERec is extended from the basic RWR model. we proposed the transfer matrix with bias by introducing three academic factors, i.e. co-publishing frequency, weight of relations and similar-level preferred, which used to lead the random walk performing better when making academic venues recommendation.

3. DESIGN OF AVEREC

VERec is designed to mine specific academic venues and make personalized recommendation for researchers. The model is inspired by the truth that, researchers usually desire to keep contact with suitable academic venues, i.e. keeping concern on 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 research achievements. Besides, AVERec is the evolution from a basic RWR model which has been proved to be competent for calculating the similarity of nodes in networks. Most of all, the three academic factors we introduced, co-publishing frequency, weight of relations and similar-level preferred, aim at biasing the random walk, such that it will more easily traverse to the positive nodes. The detailed process of AVERec is described below. Additionally, the structure is illustrated in Figure 1.

3.1 Overview of AVERec

In this work, we model a kind of co-publishing networks 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. As well as the three researchers Bob, David and Alice collaborate to write five papers which are published in the four venues respectively (note that Bob publish two papers in venue A). The nodes (venues and researchers) along with links (co-author relations and author-venue relations) form the co-publishing

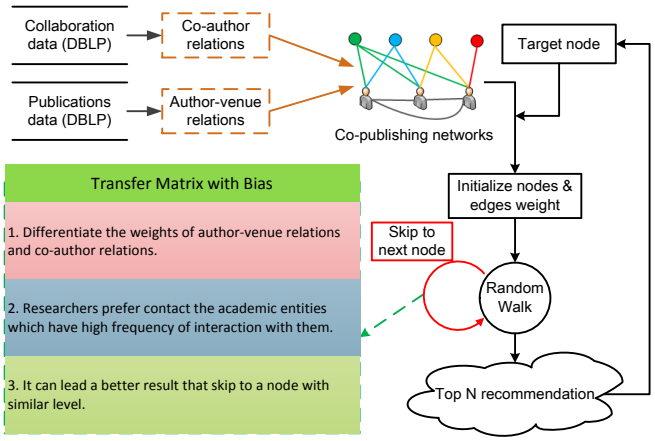

Figure 1: The structure of AVERec.

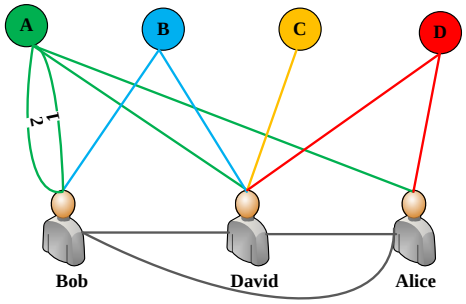

Figure 2: An example of co-publishing network

networks. We define two kinds of node sets, $Venues$ and $Authors$.

In AVERec, 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 the AVERec is run on the network to compute node ranking, starting from source node p_0, and an imaginary walker randomly walk in the network. The walker has two choice, 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 get 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 noticed that, $AR_0 = q$. The rank score of target node is 1, while others' are 0. $\mathbf{S}$ is the transfer matrix, representing the probability for each node to skip to 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 AVERec, we do some guidance work by introducing three academic factors. The change of $P(p_j, p_i)$ can lead the walker skips with preference, which will be proved to be better in section 4 for academic venues recommendation.

The detailed process of AVERec is described below corresponding with the structure in Figure 1.

- *Step1.* The initial input data is a set of publications with authors' information and venues' information. AVERec firstly extracts the co-author relations and author-venue relations. Then, generates the co-publishing 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.* Following initializing the rank score of nodes and weight of edges, AVERec run 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, AVERec 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.

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

3.2 Transfer Matrix with Bias

As the example shown in Figure 2, there are seven academic entities. With respect to recommending venues to Bob, he has never contacted with venue 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 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, we think Bob prefer contacting David rather than Alice, because Bob collaborated with David twice while Alice once. David looks more important than Alice for Bob. As well as Bob prefer contacting venue A

rather than B, since that Bob published two papers in venue A. Based on this assumption, We define co-publishing 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 i's collaborating times with author j.

In addition, there are two kinds of associations in co-publishing networks, i.e. co-author relations and author-venue relations. In case of basic random walk model, the difference between these two relations is ignored. Author-venue relations seems more important than co-author relations, because the event of publishing a paper in the venue is more ponderable when profiling the researchers' interest. This proposition has been proved in subsequent experiments which can lead better performance when making academic recommendation. We measure the weight of relations by 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 proposed 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 at similar academic levels and publishing papers in the venue which are more likely to accept the 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 AVERec. In order to measure the similarity of academic entities, we define a simple metric as shown in equation 5.

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

This 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 AVERec model adopts equation 6. With equation 6, the walker can run on the network with 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 [14], a computer science bibliography website hosted at University Trier. In this section, we describe three academic venue recommendation approaches as comparison, the statistics of data set, the evaluation metrics and our experimental procedure for evaluating the performance of AVERec, as well as detailed analysis of the results.

4.1 Three Comparison Approaches

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

Table 1: Statistics of Data Set from DBLP

Statistics	venues	researchers	articles
Number	74	70326	163446

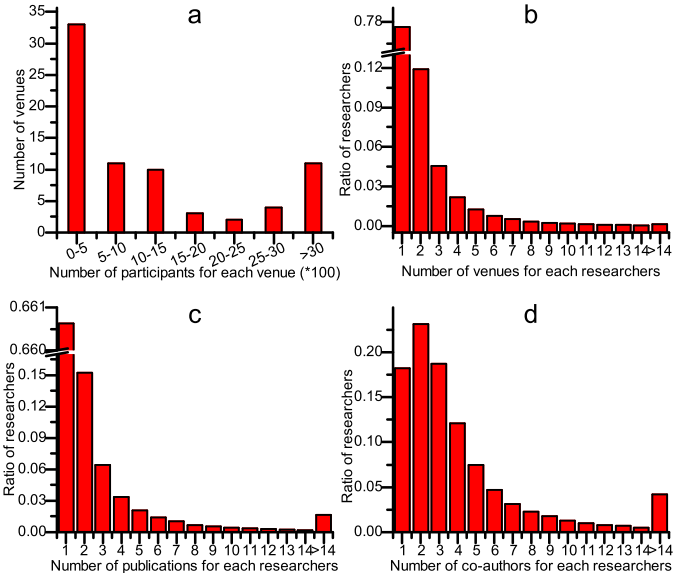

Figure 3: Detailed statistics of the data set from DBLP

Similar to popular random walk models, the details and verification method of RWR is just like AVERec, except for 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 vector respectively, which are calculated by LDA(Latent Dirichlet Allocation) model [15]. 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 contact a venue, the venue should be recommended to the researcher.

4.2 Data Set and Metrics

DBLP indexes more than 2.3 million articles on 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 about the data sets are shown in Table 1. The data set contains 74 venues and 70326 researches. 163446 articles connect researchers and venues, as well as come into being the co-publishing networks. We divided the data set into two parts: the data before year 2011 as a training set, and others as a testing set.

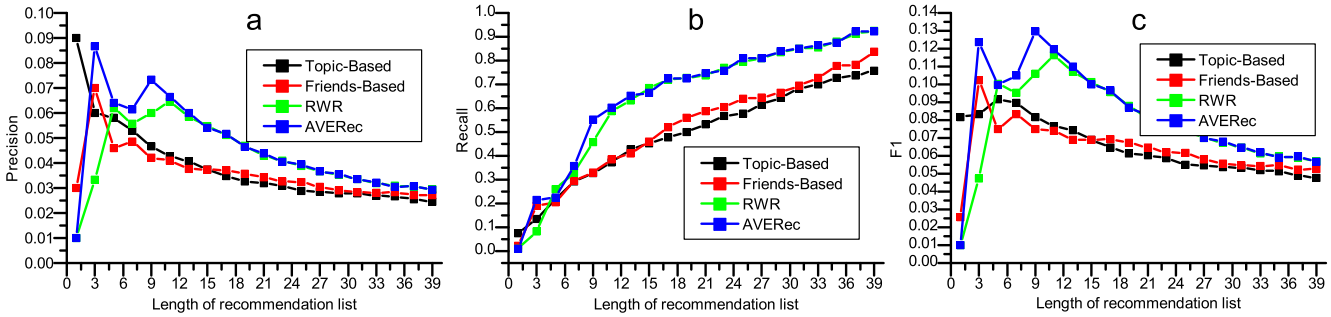

Figure 4: Performance of AVERec, basic RWR, topic-based and friends-based recommendation model

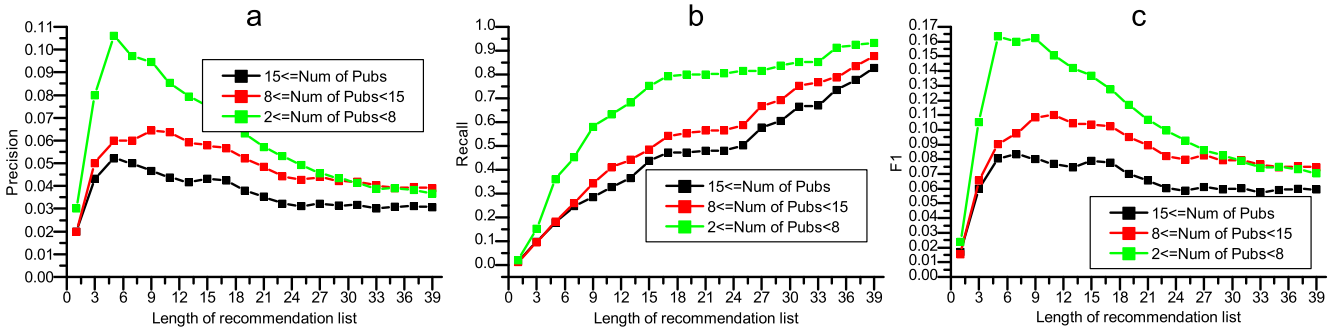

Figure 5: The impact of researchers' publications number on AVERec

The detailed statistical characteristic of this co-publishing 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 on them. We can also get that from Figure 3(b), almost 94 percentage of these 70326 researchers contact not more than 3 venues. However, there are also some "academic stars" contributing more than 14 venues, which account for 0.13%. 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 use three popular metrics, precision, recall and F1 score, to evaluate the performance of AVERec. Detailed information about these metrics has been discussed in [12]. 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. All the programs were implemented with Python.

4.3 Results and Analysis

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

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

These experiments demonstrated that, the random walk with restart based model can get more accurate academic venue recommendation than topic-based and friends-based approaches. What's more, our work on transfer matrix with bias does improve the performance of AVERec, and makes the recommendation more efficient.

We also made several extensive experiments to measure the performance of AVERec on different researchers. We mainly focused on the difference of researchers academic level, which is reflected by the number of publications. Generally, the newcomer shows lower academic level with few publications, while a fruitful professor shows high academic level with a plenty of high-quality publications. We divided the researchers into three sets, i.e. $C1$ contains researchers whose publications number 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 difference of the effect on different sets of researchers even though they show a similar trend in precision, recall and F1 score respectively. In Figure 5(c), Especially the AVERec can get highest value 16.24% for F1 score at point 9 when making academic venues recommendation for the researchers with 2 to 8 publications. The results mean that, AVERec can do better at recommending academic venues for researchers with fewer publications, i.e. the relatively newcomer, which meets our original intention of recommending academic venues for newcomer well.

5. CONCLUSIONS

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

Nonetheless, there is still room for future study in this direction. We only exploited three academic factors in co-publishing network. There are also many other features. For instance, citation relations should be explored which has been incorporated in venue recommendation by Onur [16]. As a future work, more experiments should be conducted on other academic data set.

6. ACKNOWLEDGMENTS

7. 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] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] Jing Li, Feng Xia, Wei Wang, Zhen Chen, Nana Yaw Asabere, and Huizhen Jiang. Acrec: a co-authorship based random walk model for academic collaboration recommendation. In *Proc. WWW*, pages 1209–1214, 2014.

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

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

[16] Onur Küçüktunç, Erik Saule, Kamer Kaya, and Ümit V Çatalyürek. Recommendation on academic networks using direction aware citation analysis. *CoRR*, abs/1205.1143, 2012.