AVERec: A random walk based academic venue recommendation model

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

ABSTRACT

In this work, we propose AVERec, a novel random walk based academic venue recommendation model. AVERec models 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 academic entities with high co-publishing frequency and similar academic level, 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 method.

Categories and Subject Descriptors

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

General Terms

Theory

Keywords

ACM proceedings, LaTeX, text tagging

1. INTRODUCTION

Introduction...

2. RELATED WORK

Related work...

3. ACADEMIC VENUE RECOMMENDATION MODEL

The AVERec academic venue recommendation model 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 contact with suitable academic venues, i.e. keep concern on high-quality and fruitful academic venues, participate in academic conferences which are closely related to their research, contribute to some venues in where they are possible to publish research achievements. Besides, AVERec is a 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.

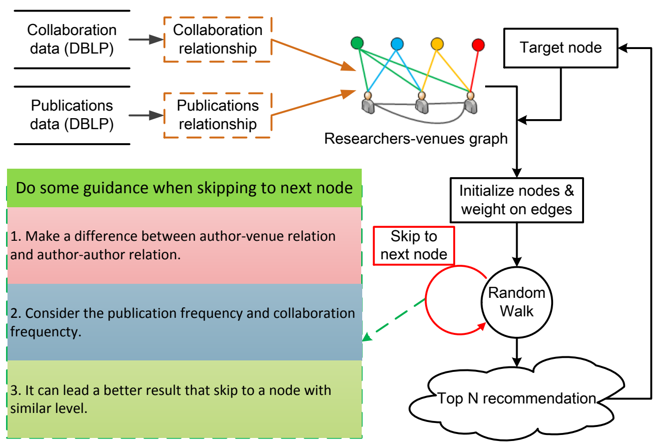

Figure 1: The structure of AVERec.

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 networks. We define two kinds of node sets, $Venues$ and $Authors$. In AVERec academic venue recom-

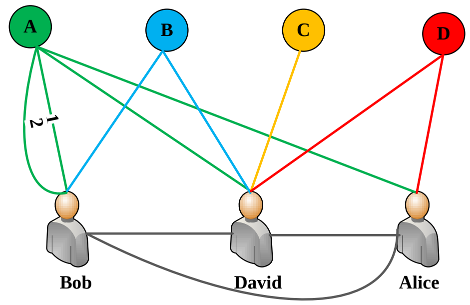

Figure 2: An example of co-publishing network.

mendation model, 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.

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

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, an imaginary walker randomly walk in the network. The walker has two choice, i.e. with probability α, walk to next node p_x, which is one of p_0's direct neighbors ($p_x \in A(p_0)$), with probability $1-\alpha$, return to source vertex p_0. Equation 1 represents one step to get rank score for node p_i. With respect to all node in the whole network, the approach is defined by equation 2, which is an iterative process.

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

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 have 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 venue 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 skip 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 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 have contacted, the Top-N venues are recommended to the target author.

We present below the details 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 contact 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.

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

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 terms 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 experiment 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 terms of researchers, they prefer contacting other researchers at similar academic levels and publishing papers in the venue which have potential to accept. 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 Equation 5.

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

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

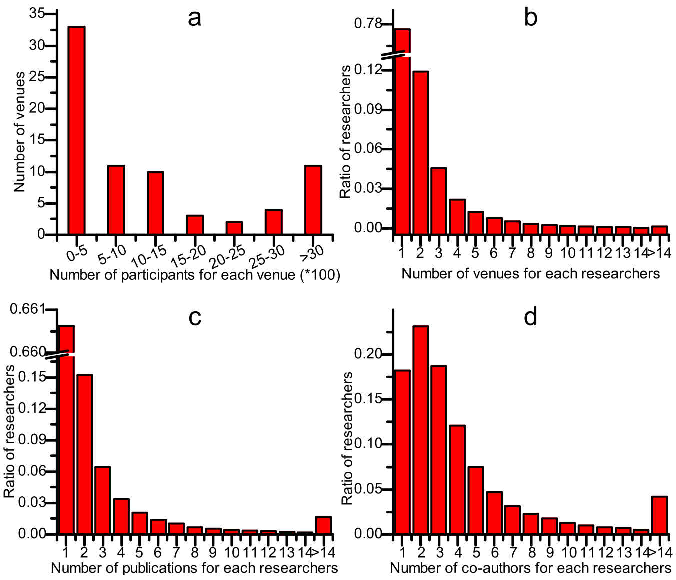

Figure 3: A sample black and white graphic (.eps format).

4. EVALUATION AND ANALYSIS

5. CONCLUSIONS

6. ACKNOWLEDGMENTS

7. REFERENCES

APPENDIX

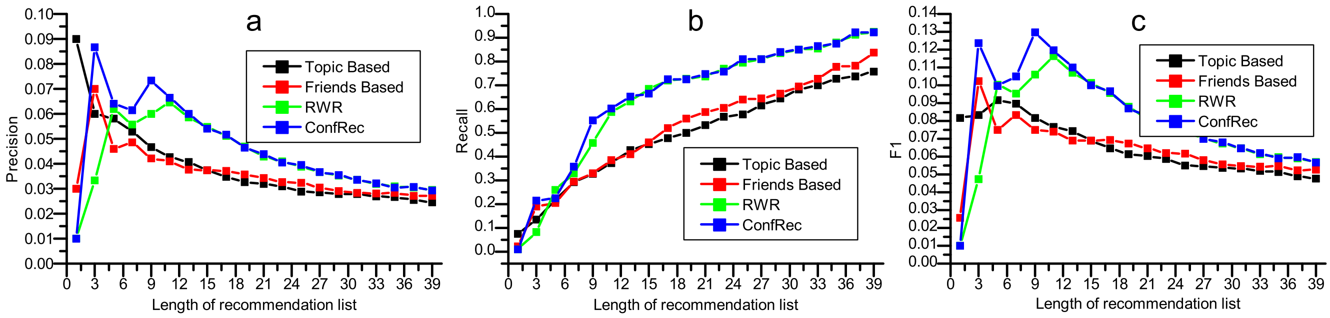

Figure 4: A sample black and white graphic (.eps format).

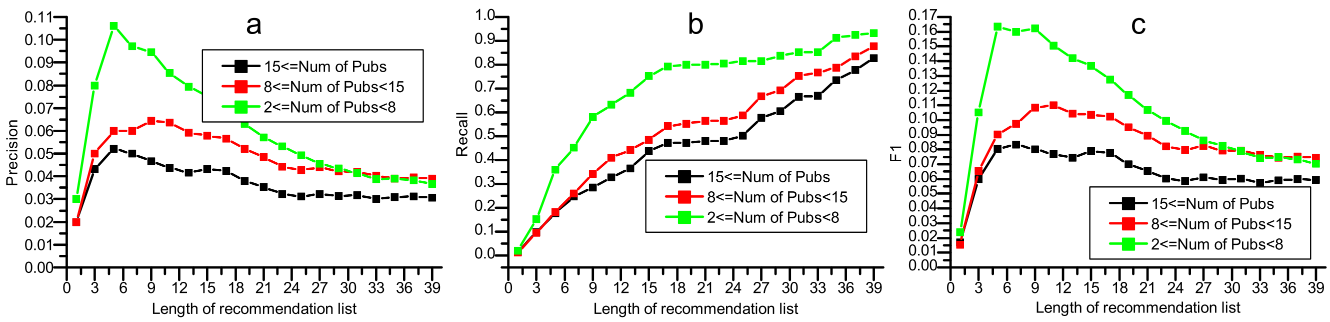

Figure 5: A sample black and white graphic (.eps format).