Combining Publication Contents and Collaboration Networks for Collaboration Recommendation

First Author
AuthorCo, Inc.
123 Author Ave.
Authortown, PA 54321 USA
author1@anotherco.com

Second Author
AuthorCo, Inc.
123 Author Ave.
Authortown, PA 54321 USA
author2@anotherco.com

Third Author
AuthorCo, Inc.
123 Author Ave.
Authortown, PA 54321 USA
author3@anotherco.com

Fourth Author
AuthorCo, Inc.
123 Author Ave.
Authortown, PA 54321 USA
author5@anotherco.com

Fifth Author
AuthorCo, Inc.
123 Author Ave.
Authortown, PA 54321 USA
author6@anotherco.com

Sixth Author
AuthorCo, Inc.
123 Author Ave.
Authortown, PA 54321 USA
author7@anotherco.com

Abstract
Due to the expansion of academic research in diverse fields, the problem of finding relevant and potential collaborators has become cumbersome. In this paper, we propose an academic collaboration recommendation model called CCRec. CCRec combines publication contents with collaboration networks to effectively generate academic collaboration recommendation for researchers. Using the DBLP data sets, we conduct benchmarking experiments to ascertain and evaluate the performance of CCRec. Our experimental results show that CCRec outperforms other state-of-the-art method especially in the scenario of topic drift problems.

Author Keywords
Collaboration recommendation, publication contents, collaboration networks, topic clustering, random walk.

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

Introduction
A collaboration network is a type of academic social networks formed by researchers and their collaborations. In academia, recommending collaborators to different researchers (groups) may help researchers build more

collaborations and become more prolific.

Some existing research studies have proposed the utilization of affiliations to exploit collaboration networks and profiles of researchers for academic collaboration recommendations [1]. However, one important factor that has been consistently ignored by researchers is that collaborations among researchers largely depend on the research field reflected from their publications. Consequently, improved academic collaboration can be achieved through the combination of publication contents and collaboration networks.

This paper proposes an academic collaboration recommendation model called CCRec. CCRec combines publication contents with collaboration networks to effectively generate academic collaboration recommendations for researchers. CCRec firstly uses topic clustering to partition the words from all the publications' titles into multiple domains. Then, CCRec computes the degree of interest (DoI) and the strength of influence (SoI) pertaining to each domain for each researcher. Finally, DoI and SoI are combined to form the feature vector for each researcher. By comparing the similarity of feature vector, CCRec provide a TopN collaboration recommending list.

PROPOSED SCHEME

Our proposed design scheme for CCRec is inspired by the reality and truth that a researcher usually desires to know other researchers who have similar research interests. As mentioned above, researchers often behave differently across multiple domains of interest. Such behaviors usually reveal academic features of different researchers in the network. In this work, we define the DoI and SoI for researchers in different domains. Furthermore, we use the feature vector combined by DoI and SoI to evaluate the similarity of researchers and then obtain the recommending list.

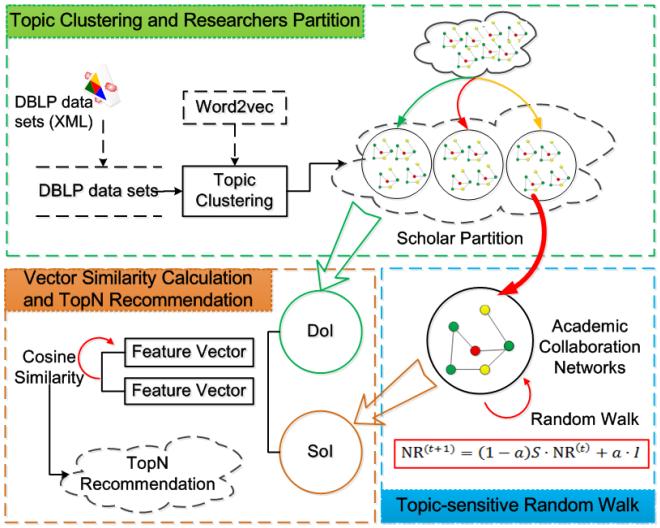


Figure 1: The architecture diagram of CCRec model

Figure 1 depicts the four components of CCRec: *topic clustering and researcher partition*, *Topic-sensitive random walk*, *vector similarity calculation and TopN recommendation*. Topic clustering and researcher partition distribute researchers according to multiple domains and acquire a DoI for each researcher. Topic-sensitive random walk calculates the SoI in each domain, and the TopN recommendation provides the recommending list.

Topic Clustering and researcher Partition

In CCRec, topic clustering and researcher partition generate various domains and map all researchers into these domains. Initially, CCRec extract keywords from titles of all the papers for each researcher, and filters out

some irrelevant words, e.g. "of","the", "and", etc. As core contents in a paper, preprocessed keywords in CCRec are signified as valuable and reliable data in a variety of academic topics. We use word2vec, a famous tool of NLP (Natural Language Processing) to cluster the keywords into various domains. Then, if some keywords of a researcher belong to a domain, we will partition the researcher to that particular domain. We emphasize that one researcher always belongs to several domains and there are also many researchers in one domain.

Feature Vector Calculation
To measure the distribution of researchers' interest, we define DoI as researcher's proportion of interest in one domain:

$$DoI_{s,d} = \frac{N_d}{\sum_{k=1}^{n} N_k} \tag{1}$$

Where N_d is the number of keywords of researcher s in domain d. It is a content-based method that utilizes the information on the titles of researchers' publications.

We define SoI as researcher's strength of influence in one domain, which is measured by a topic-sensitive random walk method based on collaboration networks. The core equation of the random walk method is shown below:

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

Where R_d represents the rank score vector of all researchers in domain d, q is the initial vector R^0, and α denotes the damping coefficient. Random walk is an iterative process. After limited iterations, the vector R will be convergent. The vector item in this scenario is defined as SoI. We therefore obtain SoI through $SoI_s = R_{d,s}$.

To be more specific, we define feature vector F by combining DoI and SoI, which measures the academic feature of researchers in various domains.

$$F_{s,d} = DoI_s * SoI_s \tag{3}$$

Collaboration Recommendation by Feature Vector Similarity
In CCRec, the academic features of researchers is measured by the feature vector F. We use a *cosine similarity* method to compute the similarity of these feature vectors, and further compute the similarity between researchers.

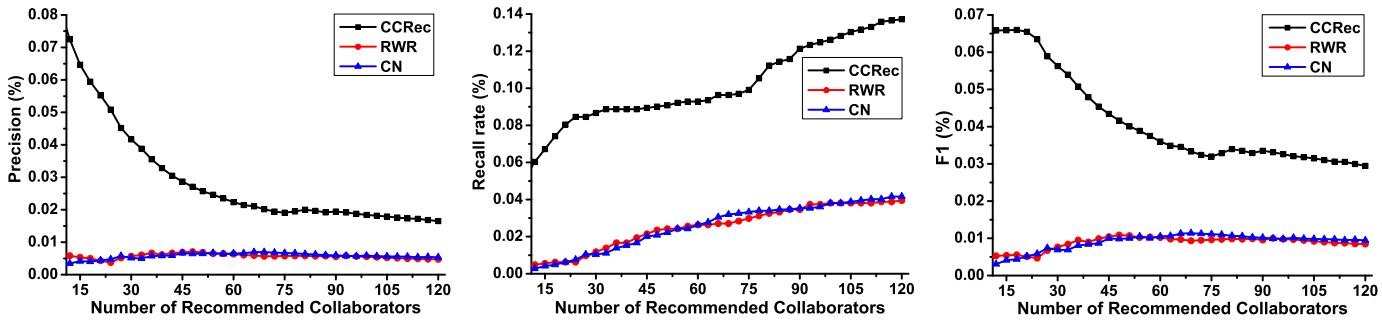
$$Sim(s_1, s_2) = \frac{\sum_{i=1}^{n}(F_{s_1,i} * F_{s_2,i})}{\sqrt{\sum_{i=1}^{n} F_{s_1,i}^2} * \sqrt{\sum_{i=1}^{n} F_{s_2,i}^2}} \tag{4}$$

Finally, CCRec recommends potential academic collaborators to researchers who have common interests and high similarities, by providing a Top N recommendation list for each researcher in the network.

Evaluation and Analysis

Using a subset of DBLP dataset relevant to data mining, we embarked on benchmarking experiments to evaluate the performance of CCRec. We took the year 2011 as the partition time of training and testing sets. To evaluate our model in a better way, we compared CCRec with two traditional approaches, namely: Random Walk with Restart (RWR) and Common Neighbors (CN) [2]. We adopted three metrics to evaluate the performance of CCRec, namely: precision, recall rate and F1. What we should illustrate is that we recommend the new collaborators who never cooperated with the target researcher, because the new collaborators are more meaningful and practical in academia.

Figure 2 shows the performance of CCRec, RWR and CN in terms of precision, recall rate and F1 with the number



(a) Precision (b) Recall Rate (c) F1

Figure 2: Performance of CCRec, RWR and CN

of recommended collaborators increasing. It can be observed that CCRec significantly outperforms RWR and CN all the time in these three metrics. CCRec shows a downtrend for precision and an uptrend for recall rate. In the case of F1, it reaches the peak 6.598% when recommending 18 researchers, which can be regarded as an encouraging and favorable situation point (scenario).

In a nutshell, CCRec outperforms RWR and CN with higher precision, recall rate and F1. This is because CCRec combines publication contents and collaboration networks which has a distinct advantage in recommending new collaborators.

Conclusion

The conclusions we reach are: 1) CCRec outperforms RWR and CN in precision, recall rate and F1 integrating the content of publications with academic collaboration network. 2) With topic clustering, the problem of topic drift has been well solved.

Our research on CCRec reveals that, the combination of information regarding publication contents and collaboration networks of researchers can improve the generation of effective academic collaborations.

References

[1] Brandão, M. A., Moro, M. M., Lopes, G. R., and Oliveira, J. P. Using link semantics to recommend collaborations in academic social networks. In *Proc. 22nd WWW* (2013), 833–840.

[2] Li, J., Xia, F., Wang, W., Chen, Z., Asabere, N. Y., and Jiang, H. Acrec: a co-authorship based random walk model for academic collaboration recommendation. In *Proc. 23rd WWW* (2014), 1209–1214.