# CCRec: Content of Publications and Collaboration Network Combined Academic 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

With the academic research filed expanding, the problem of finding proper potential collaborators is really cumbersome. In this paper, we proposed a content of publications and collaboration network combined academic Collaboration recommendation model (CCRec). Compared to traditional approaches, CCRec is more effective because it recommends collaborators combining the content of publications and collaboration network in different topics. Experiments based on DBLP data sets show that CCRec significantly outperforms traditional approaches, with the topic drift problem well solved.

## Author Keywords

Guides, instructions, author's kit, conference publications Mandatory section to be included in your final version.

## ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous. See: http://www.acm.org/about/class/1998/ for help using the ACM Classification system. Mandatory section to be included in your final version.

## Introduction

Collaboration network is one kind of academic social networks formed by researchers and their collaborations.

In the academic field, recommending collaborators to researchers (groups) may help researchers build more collaborations and become more prolific.

Some Studies proposed to recommend academic collaborators by exploiting the collaboration network and the profiles of researchers such as affiliation[]. However, the fact is always ignored that collaborations among researchers largely depend on the research field reflected from their publications. Thus it may have a superior performance compare the similarity of researchers by combining the content of publications and collaboration networks.

This paper proposed a content of publications and collaboration network combined academic collaboration recommendation model (CCRec). CCRec firstly uses the topic clustering (sensitive) 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 SCHEMA

CCRec is inspired by the truth that researchers usually desire to know people who have high similarity with them. As mentioned above, researchers often behave differently across multiple domains of interest, which can reveal researchers academic feature. 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 get the recommending list.

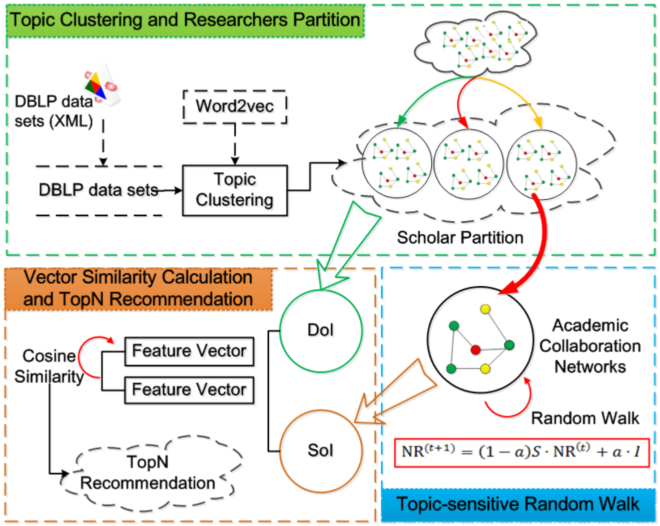

Topic Clustering and Researchers Partition
DBLP data sets (XML)
Word2vec
DBLP data sets
Topic Clustering
Scholar Partition
Vector Similarity Calculation and TopN Recommendation
Cosine Similarity
Feature Vector
Feature Vector
TopN Recommendation
DoI
SoI
Academic Collaboration Networks
Random Walk
$\mathrm{NR}^{(t+1)} = (1-a)S \cdot \mathrm{NR}^{(t)} + a \cdot I$
Topic-sensitive Random Walk

**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 get a DoI for each researcher. Topic-sensitive random walk calculates the SoI in each domain. And the TopN recommendation gives 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, meanwhile filtering out some meaningless words, e.g. "of","the", "and", etc. As a core text for a paper, these keywords preprocessed are rich 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 this domain. What we should emphasis is that one researcher always belongs to several domains and there are also a great 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 key words of researcher $s$ in domain $d$. It is a content-based method by utilizing 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 follow:

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

Where $R_d$ represent the rank score vector of all researchers in domain $d$, $q$ is the initial vector of $R$, $\alpha$ denotes the damping coefficient. Random walk is a iterative process. After limited iterations, the vector $R$ will be convergent. The vector item is $SoI$ here. We can get $SoI_s = R_{d,s}$.

To be more accurate, We define feature vector $F$ by combining $DoI$ and $SoI$, which measures the academic feature of researchers on various domains.

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

*Collaboration Recommendation by Feature Vector Similarity*
In CCRec, The academic feature 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 denoting the similarity between researchers.

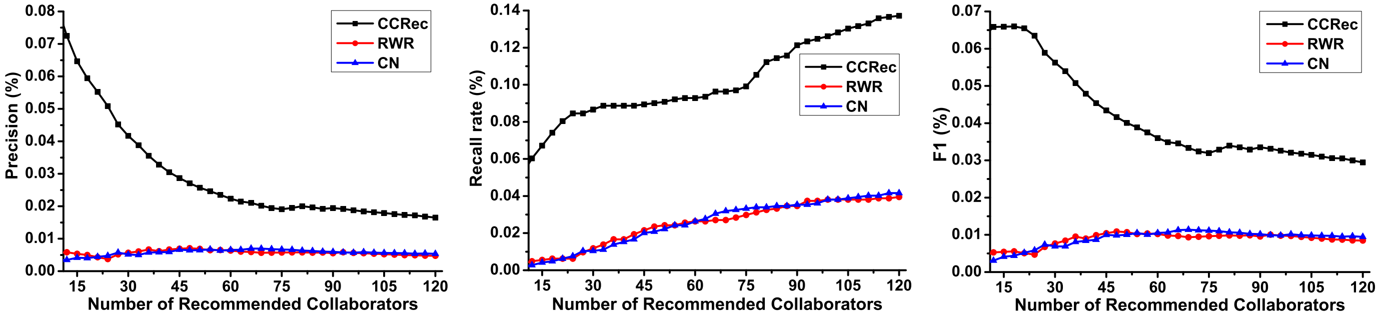
$$Sim(s_1, s_2) = \frac{\sum_{i=1}^{n}(F_{s_1,i} * F_{s_2,i})}{\sqrt{\sum_{i=1}^{n} F_{s_1,i}^2} * \sqrt{\sum_{i=1}^{n} F_{s_2,i}^2}} \tag{4}$$

Finally, CCRec recommends to researchers those potential collaborators who have high similarity with them, and provide a TopN collaborators recommendation list.

## Evaluation and Analysis

We have conducted experiments based on a subset of DBLP data, which relevant to data mining. We take the year 2011 as the partition time of training set and testing set. To better evaluate our model, we compared CCRec with two traditional approaches Random Walk with Restart (RWR) and Common Neighbors (CN) [1]. We adopt three metrics to evaluate the performance, 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 real academic filed.

Figure 2 shows the performance of CCRec, RWR and CN in precision, recall rate and F1 with the number of recommended collaborators increasing. It can be observed that CCRec significantly outperforms RWR and CN. For these three metrics, CCRec exceeds RWR and CN all the time. CCRec shows a downtrend for precision and an uptrend for recall rate. For the F1, it reaches the peak



(a) Precision

(b) Recall Rate

(c) F1

**Figure 2:** Performance of CCRec, RWR and FOF

6.598% when recommending 18 researchers, which can be regarded as an best point.

Figure 2-a reveals that the precision of CCRec is always higher than that of RWR and CN. It shows an upward tendency for the recall rate of CCRec, which is obviously superior to RWR and CN. For the F1, CCRec exceeds RWR and CN all the time, and

In short, CCRec outperforms RWR and CN with higher precision, recall rate and F1. This is because CCRec with content of publications and collaboration network combined 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, combining the information of publications' content and collaboration networks can do help to recommend collaborators better and make collaboration recommendation be more specific and effective.

## References

[1] 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 *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, International World Wide Web Conferences Steering Committee (2014), 1209–1214.