Exploiting Publication Contents and Collaboration Networks for Collaborator Recommendation

*

Ben Trovato[†]
Institute for Clarity in Documentation
1932 Wallamaloo Lane
Wallamaloo, New Zealand
trovato@corporation.com

G.K.M. Tobin[‡]
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
webmaster@marysville-ohio.com

Lars Thørväld[§]
The Thørväld Group
1 Thørväld Circle
Hekla, Iceland
larst@affiliation.org

Lawrence P. Leipuner
Brookhaven Laboratories
Brookhaven National Lab
P.O. Box 5000
lleipuner@researchlabs.org

Sean Fogarty
NASA Ames Research Center
Moffett Field
California 94035
fogartys@amesres.org

Charles Palmer
Palmer Research Laboratories
8600 Datapoint Drive
San Antonio, Texas 78229
cpalmer@prl.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 work, 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 examine the performance of CCRec. Our preliminary experimental results show that CCRec outperforms other state-of-the-art methods especially in addressing the topic drift problems.

Categories and Subject Descriptors

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

General Terms

Theory

---

[*]A full version of this paper is available as *Author's Guide to Preparing ACM SIG Proceedings Using LaTeX2_ε and BibTeX* at www.acm.org/eaddress.htm

[†]Dr. Trovato insisted his name be first.

[‡]The secretary disavows any knowledge of this author's actions.

[§]This author is the one who did all the really hard work.

Keywords

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

1. 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 recommendation. 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 recommendation can be achieved through the combination of publication contents and collaboration networks.

In this work we propose an academic collaboration recommendation model called CCRec. CCRec combines publication contents with collaboration networks to effectively generate academic collaboration recommendation for researchers. CCRec first 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 provides a TopN collaboration recommendation list.

2. PROPOSED SCHEME

Our proposed design scheme for CCRec is inspired by the reality and truth that a researcher usually desires to know
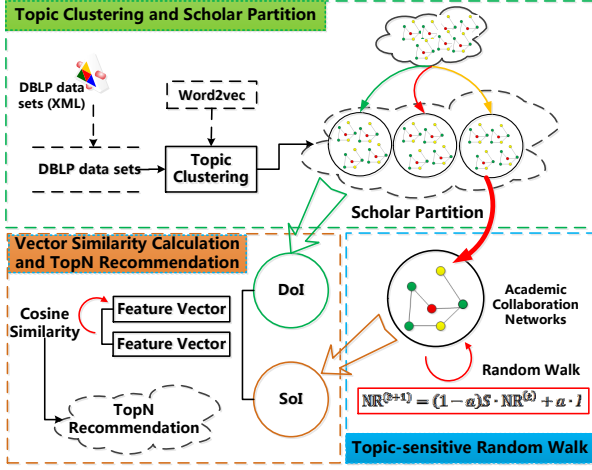
Figure 1: The architecture diagram of CCRec model

other researchers who have similar research interests. As mentioned above, researchers often behave differently across multiple domains of interests. 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 recommendation list.

Figure 1 depicts the three components of CCRec. 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 recommendation list.

2.1 Topic Clustering and Researcher Partition

In CCRec, topic clustering and researcher partition generate various domains and map all researchers into these domains. Initially, CCRec extracts 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.

2.2 Feature Vector Calculation

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

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

where N_d is the number of keywords of researcher s in domain d. It is a content-based method that utilizes the infor-

mation 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:

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

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.

$$F_{s,d} = DoI_s * SoI_s \qquad (3)$$

2.3 Collaboration Recommendation by Feature Vector Similarity

In CCRec, the academic features of researchers are 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.

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

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

3. 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 the two following approaches. RWRec: a random walk recommendation model based on collaboration networks. CNRec: a common neighbors based recommendation model [?]. We adopted three metrics to evaluate the performance of CCRec: precision, recall rate and F1. 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, RWRec and CNRec in terms of precision, recall rate and F1 with the number of recommended collaborators increasing. It can be observed that CCRec significantly outperforms RWRec and CNRec 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.
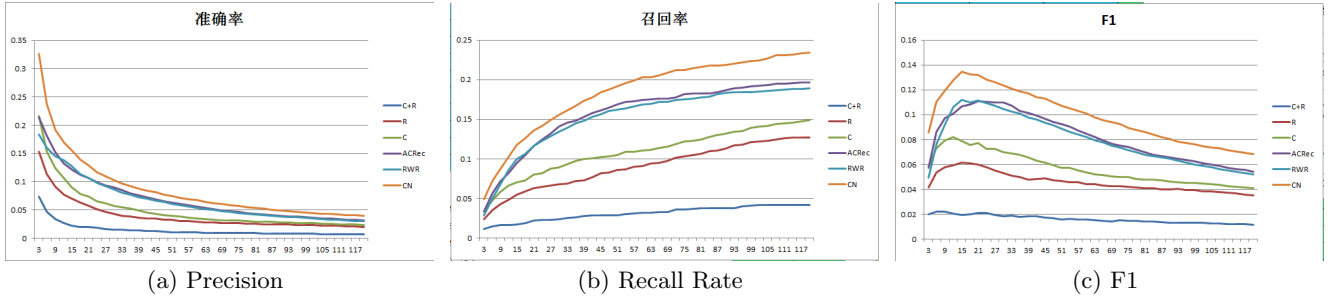
(a) Precision　　　　　　　(b) Recall Rate　　　　　　　(c) F1

Figure 2: Performance of CCRec, RWRec and CNRec

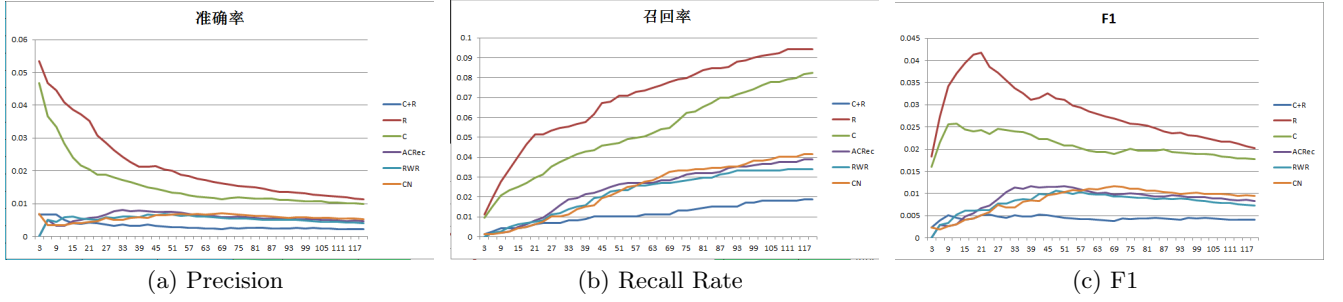(a) Precision　　　　　　　(b) Recall Rate　　　　　　　(c) F1

Figure 3: Performance of CCRec, RWRec and CNRec

In a nutshell, CCRec outperforms RWRec and CNRec 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.

4. CONCLUSIONS

The conclusions we reach are: 1) CCRec outperforms R-WRec and CNRec in precision, recall rate and F1 integrating publication contents with academic collaboration networks. 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.

5. ACKNOWLEDGMENTS