

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 C-CCRec. C-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 C-CCRec. Our preliminary experimental results show that C-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 L^AT_EX₂ 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

Nowadays, with rapid development of Internet technology, the scale of Internet is beyond the imagination of people and Internet gradually becomes the main carrier of sharing information. Thus, how to obtain the useful one from vast information has become a complex task with the problem of information overload phenomena occurring. Therefore, recommender systems and techniques immensely help people by providing easier access to the specific resources they really need.

In academia, cooperation among researchers is of vital necessary. Studies show that researchers are usually prolific by well collaboration with collaborators. This is to say, the collaborator is a considerable factor well connected with the productivity of a scholar. So scholars tend to get acquainted with potential collaborators, especially the most valuable collaborators. In academic social networks, there has been a variety of methods proposed recommending collaborators who they have ever collaborated with or never.

In this context, previous studies have exploited mainly two aspects for academic collaboration recommendation, that is content-based and collaboration network-based methods respectively. Plenty of researches show that the traditional content-based method mostly focuses on the tags made by people, which can not completely stand for the research topic of the scholar. Thus, it would be imperative taking research field into account when recommending collaborators, particularly the personalized recommendation on account of various research fields. For the collaboration network-based method, there has been a suite of methods such as Random Walk algorithm, but rarely methods have ever integrated the content-based and collaboration network-based method. What's more, in view of the burdensome and embarrassing

initial collaboration with unconnected researchers ever and the less value of already known collaborators, the unconnected researchers are more deserve to be recommended to seek more potential and valuable collaborators.

In this paper, we propose a novel hybrid model exploiting publication contents and collaboration networks for collaborator recommendation called CCRec. CCRec combines the content-based method and collaboration network-based method by synthesizing publication contents and collaboration networks. Because the publication contents are largely on behalf of the research field of a scholar, we can more precisely achieve the personalized recommendation according to their different kinds of research fields. We extract key words from publication contents of all researchers and divide these key words into several topics, then get a number of research fields corresponding to different researchers. For researchers tend to collaborate with those in the similar research field, thus in each topic there is a big collaboration network. Therefore, we explore all the collaboration networks in each research topic. With the publication contents and collaboration networks combined, CCRec greatly outperforms other traditional methods recommending totally new collaborators.

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.

The remainder of the paper is structured as follows. Section 2 briefly presents the related work. We discuss the details of our model in section 3, which highlights the problem statement and the structure of the model. In section 4, we conduct a mass of experiments and analyzes the results. At last, section 5 concludes the paper.

2. RELATED WORK

Social networks have been researched for decades to exploit the relationships and interactions between people. There is also a sample system helping people find experts on a certain topic. While summarizing the preterit work, researches mainly focus on three parts including content-based method, social networks-based method and minority hybrid method. For the first two methods, many practical examples have confirmed it, but fewer researches have done for the hybrid method.

3. DESIGN OF CCRec

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 and high academic value. As mentioned above, researchers often behave differently across multiple domains of interests. Such behaviors usually reveal academic features of different researchers in different domains. Besides, as the RWR model has been proved to be competent for calculating the rank score of node in social networks derived from the co-authorship, researchers' academic value can be well reflected

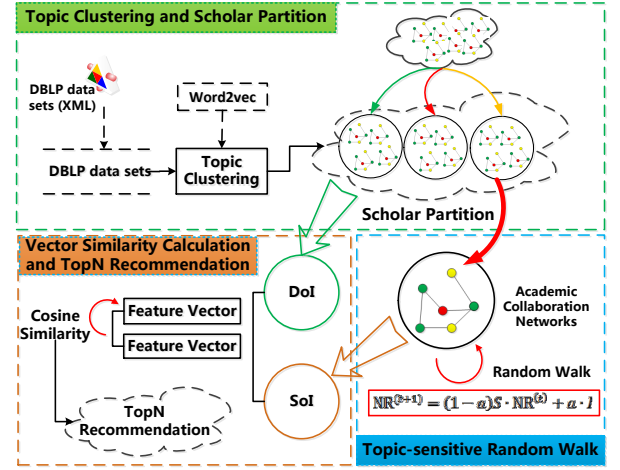


Figure 1: The architecture diagram of CCRec model

by RWR. In this work, we proposed a content-based method to acquire multiple domains of interests. Furthermore, we define the *** to represent the academic values (Rank Score) of researchers in different domains, which can be regarded as the elements of feature vector of researchers. Finally, we use the feature vector to evaluate the similarity of researchers and then obtain the recommendation list.

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.

3.1 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.

3.2 Feature Vector Calculation

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

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

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:

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

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.

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

3.3 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.

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

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.

4. EVALUATION AND ANALYSIS

We conduct various experiments using data from DBLP [?], a computer science bibliography website hosted at University Trier. We extracted the subsets of the entire data using the required information, which are all in the field of data mining involving 34 journals and 49 conferences. The data was modeled by an academic social network, which contains 59659 nodes (authors) and 90282 edges (coauthor relations). We divided the data set into two parts: the data before year 2011 as a training set, and others as a testing set. To evaluate the performance of CCRec model, we use three metrics widely used in the recommender systems, *Precision*, *Recall* and *F1* [?].

All experiments were performed on a 64-bit Linux-based operation system, Ubuntu 12.04 with a 4-core and 32GHz Intel

CPU, 4-G Bytes memory. All the programs are implemented with Python.

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 downward trend for precision and an upward trend for recall rate. In the case of F1, it reaches the peak 6.598% when recommending 18 researchers.

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.

5. CONCLUSIONS

The conclusions we reach are: 1) CCRec outperforms RWRec 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.

6. ACKNOWLEDGMENTS

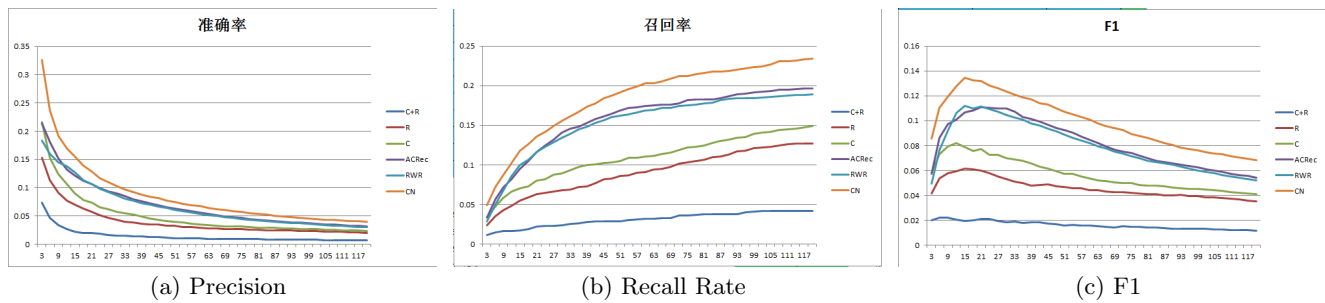


Figure 2: Performance of CCRec, RWRec and CNRec

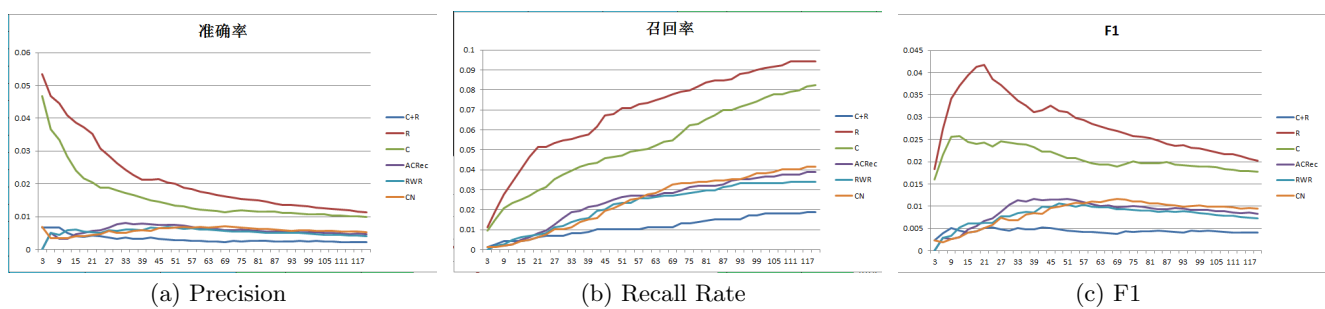


Figure 3: Performance of CCRec, RWRec and CNRec