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-CRec. 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

Keywords

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

1. INTRODUCTION

With the 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 is really tough and annoying with the problem of information overload occurring. Therefore, recommender systems and techniques immensely help people by providing easier access to the resources they need.

initiating collaboration with unconnected researchers is burdensome and fraught with risk, despite potentially relevant expertise.

Some existing research studies [1] [2] [3] 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

[*] A full version of this paper is available as *Author's Guide to Preparing ACM SIG Proceedings Using $\LaTeX2_\epsilon$ 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.

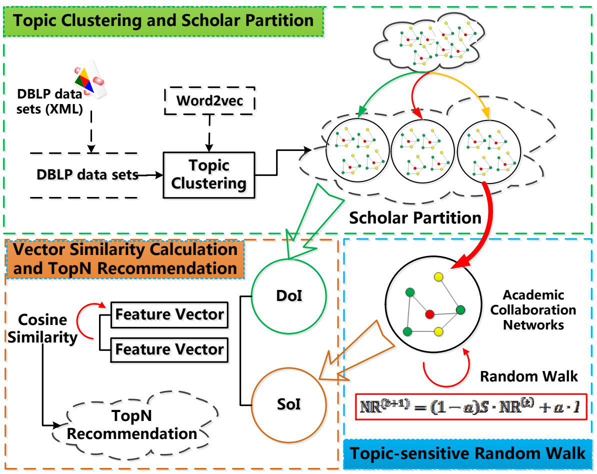

Topic Clustering and Scholar Partition
DBLP data sets (XML)
Word2vec
DBLP data sets
Topic Clustering
Scholar Partition
Vector Similarity Calculation and TopN Recommendation
Cosine Similarity
Feature Vector
Feature Vector
TopN Recommendation
DoI
SoI
Academic Collaboration Networks
Random Walk
$NR^{(t+1)}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

Figure 1: The architecture diagram of CCRec model

the similarity of feature vector, CCRec provides a TopN collaboration recommendation list.

2. RELATED WORK

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 strong influence in academia. 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 coauthorship [5], researchers' strength of influence in specific domains can be well reflected by RWR. In this work, we adopt a content-based method to acquire multiple domains of interests. Besides, as a social-based model, RWR is used to measure the researchers' strength of influence in different domains. After that, We use the feature vector to evaluate the similarity of researchers and then obtain the recommendation list. The detailed process is described bellow and the corresponding pseudo-code is illustrated in Algorithm 1. Figure 1 depicts the three components of CCRec.

3.1 Topic Clustering and Researcher Partition

It is a content-based method for topic clustering and researcher partition, which generates various domains and maps all researchers into these domains. In this work, we use a famous tool of NLP (Natural Language Processing), word2vec, which provides an efficient implementation of the continuous of *bag-of-words* and *skip-gram* architectures for computing vector representations of words. It takes a text corpus as input and produces the word vectors as output. The resulting word vector file can be used as features in many natural language processing and machine learning applications. The word vectors can be also used for deriving word classes from huge data sets. This is achieved by performing K-means clustering on top of the word vectors. The output is a vocabulary file with words and their corresponding class IDs. In case of CCRec model, the input data is a set of titles from all the papers created by each researcher. The titles are split in many sequential words. In addition, there is a necessary to filter out some irrelevant words, e.g. "of","the", "and", etc. For extracting from titles, this set of preprocessed words can outline the core contents of papers, which are signified as valuable and reliable corpus to denote a variety of academic topics. With this English corpus,word2vec obtains various domains and cluster the word into the domains.

In addition, CCRec partitions researchers to particular domains with follows method. 1, Extract mesh terms from a researcher's publications. 2, Traverse all the terms and check the word vector. The model will tag the researcher to the particular domains which contains these domains. It should be emphasized 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:

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

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.

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

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.

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

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 [4], 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* [7].

All experiments were performed on a 64-bit Linux-based operation system, Ubuntu 12.04 with a 4-duo 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 [6]. 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.

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 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.

6. ACKNOWLEDGMENTS

7. ADDITIONAL AUTHORS

Additional authors: John Smith (The Thørväld Group, email: jsmith@affiliation.org) and Julius P. Kumquat (The Kumquat Consortium, email: jpkumquat@consortium.net).

8. REFERENCES

[1] N. Benchettara, R. Kanawati, and C. Rouveirol. A supervised machine learning link prediction approach for academic collaboration recommendation. In *Proc. ACM RecSys*, pages 253–256, 2010.

[2] M. A. Brandão and M. M. Moro. Affiliation influence on recommendation in academic social networks. In *Proc. AMW*, pages 230–234, 2012.

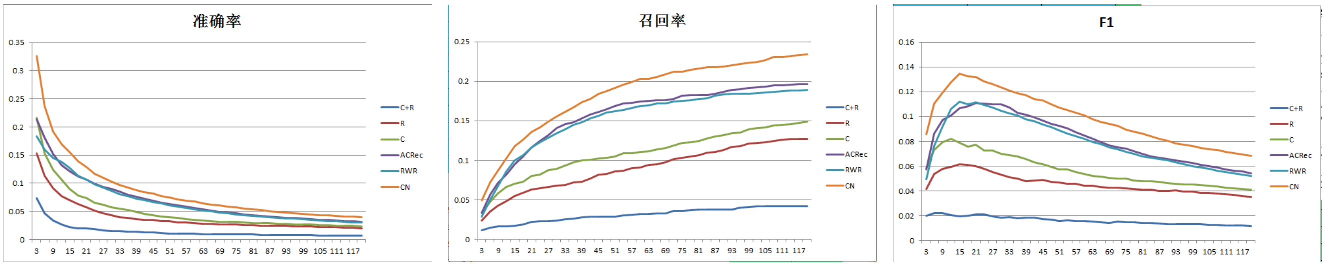
[3] M. A. Brandão, M. M. Moro, G. R. Lopes, and J. P. Oliveira. Using link semantics to recommend collaborations in academic social networks. In *Proc. WWW*, pages 833–840, 2013.

[4] M. Ley. Dblp: some lessons learned. *Proceedings of the VLDB Endowment*, 2(2):1493–1500, 2009.

[5] J. Li, F. Xia, W. Wang, Z. Chen, N. Y. Asabere, and H. Jiang. 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*, pages 1209–1214. International World Wide Web Conferences Steering Committee, 2014.

[6] G. R. Lopes, M. M. Moro, L. K. Wives, and J. P. M. De Oliveira. Collaboration recommendation on academic social networks. In *Advances in Conceptual Modeling–Applications and Challenges*, pages 190–199. Springer, 2010.

[7] G. Shani and A. Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.

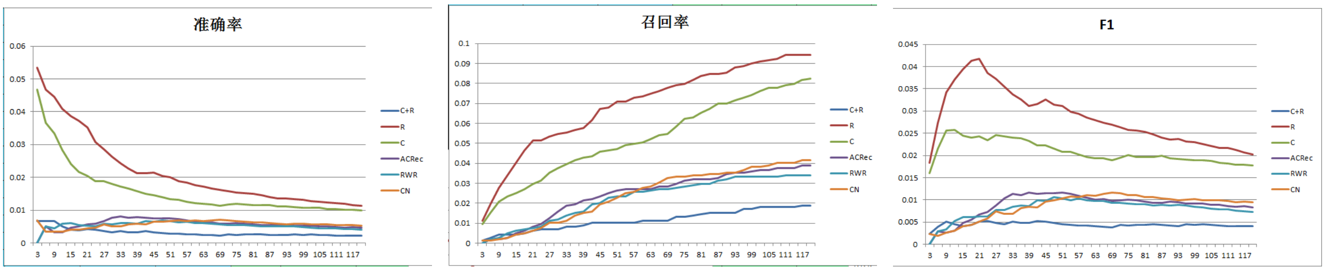


(a) Precision

(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