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

In academia, studies show that researchers are usually prolific by well collaboration with collaborators. However, due to the expansion of academic research in diverse domains, the problem of finding most relevant and potential collaborators from a large volume of big scholarly data has become cumbersome and time-consuming. In this work, we propose an academic collaborators recommendation model called C-CRec, an innovative model that combines publication contents with collaboration networks. A topic clustering model and a random walk model are adopted to help to effectively find the most potential collaborators 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 in precision, recall and $F1$ score, additionally 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

[*]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. 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 researchers tend to get acquainted with the most potential collaborators (MPCs, some influential scholars who are similar in interests and have never collaborator before), or the most valuable collaborators (MVCs, some influential scholars or colleagues who are active and valuable in adjacent circles). Considering the inherently requirements, 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 three aspects for academic collaboration recommendation, content-based, social network-based and hybrid recommendation. Some traditional content-based methods represent researchers by interests tags, and make collaborators recommendation by computing interests similarities. However, in general, a researchers shows bias on various academic domains. Moreover, interests tags are sometime not accurately to represent researchers' features. Thus, it would be imperative considering academic domains when recommending collaborators, particularly the cross-domain recommendation. Our previous work proposed a network-based model, ACRec [?], which solved the problem of recommending MVCs. ACRec make it easier for scientists to collaborate with colleagues in their social network. However, many scientists also initiate collaborations outside of their social networks. It is burdensome and fraught with risk to initiating collaboration with socially unconnected researchers. In additionally, considering the less value of already known collaborators, unconnected researchers are more deserve to be recommended to seek more MPCs. What's more, some excellent hybrid models are offered in recent years [?], which provides us many good references.

In this paper, we propose a novel hybrid model exploiting publication contents and collaboration networks for collaborators 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, C-CRec 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.

Many approaches have been presented to formalize academic collaboration recommendation as a link prediction problem. Link prediction is to predict new links that might be added to the social networks, which is useful for collaborators recommendation, but differs from it in that the latter is context-based. David et al. first defined link prediction: to predict the edges that will be added to the network in the future time t' at the time of t in a social network. Some of these approaches have been applied to large social networks and results show a good performance. Lichtenwalter et al. examined some important factors for link prediction and proposed a general framework.

For content-based method, the CollabSeer system recommends collaborators for a researcher, which a an open system introduced by Chen et al. considering research interests for collaborators recommendation. Lopes et al. considered researchers' publications area and the vector space model to make collaboration recommendation. Pavlov and Ichise extracted structural attributes from graph of past collaborations and uses them to train a set of predictors using supervised learning algorithms, then these predictors can be used to predict links between existing nodes in the graph. Tin Huynh et al. proposed a new approach that uses additional information as new features to make recommendations, i.e., the strength of the relationship between organizations, the importance rating, and the activity scores of researchers. We also propose a new method for evaluating the quality of collaborators recommendations.

Ma et al. analysed how social networks information can benefit recommender systems and proposed a method improving the performance of recommender systems by incorporating social network information. In addition, Lopes et al. present a novel method for recommending new collaborators and identifying existing collaborations. They took the semantic issues involved in the relationships among researchers in different areas. Barabasi et al. analysed the basic network properties of academic social networks in terms of degree distribution, average separation, clustering coefficient, average degree and so on. In another study, Newman researched a number of statistical properties of scientific collaboration networks like the number of papers, numbers of collaborators and degree of clustering in the networks, etc.

Lee et al. exploit how well content-based, social networks-based and hybrid recommendation algorithms predicted coauthor relationship, and the result show that a hybrid algorithm combining content and social networks information outperformed better. Actually, there is not much work for the combination of content and networks.

In summary, there are two deficiency not considered by traditional studies: (1) lack of a better integrated method combining content-based and networks-based, and (2) research fields of researchers are not considered when recommending collaborators.

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 a social-based

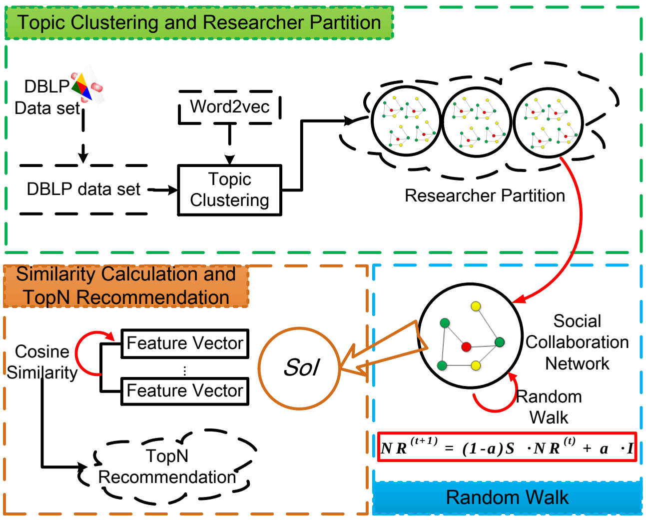


Figure 1: The architecture diagram of CCRec model

model, 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' strength of influence in specific domains can be well reflected by RWR. In this work, we first adopt a content-based method to acquire multiple domains of interests. and then using the social network-based method of RWR 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 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 the input and produces the word vectors as the output. The final 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 domains 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 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 clusters the words into the domains.

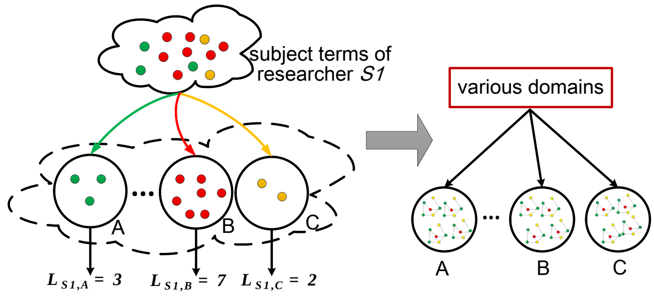


Figure 2: Researcher Partition

In addition, CCRec partitions researchers to specific domains with following method. 1, Extract mesh terms from a researcher's publications. 2, Traverse all the terms and check the word vector. The model tags the researcher to particular domains which contains these mesh terms. 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

As mentioned in section 2, in general, researchers devote to several adjacent domains. But in the case of attention and strength of influence in various domains, they are offering some biases. To measure the distribution of researchers' interests, we define the *SoI* (Strength of Influence) to denote the academic values (Rank Score) of researchers in different domains, which can be regarded as the elements of feature vector of researchers. Considering each of the domains, there are numerous researchers with similar research interests. Their co-author relationships can be modeled by a social network. Thus, there are many co-author networks corresponding to different domains. The *SoI* is measured by RWR model based on the co-author networks. The core equation of the RWR model is shown below:

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

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. RWR is an iterative process. After limited iterations, the vector R will be convergent. In this scenario, $SoI_s = R_{d,s}$. That is, the final value of the vector item $R_{d,s}$ is the SoI of researcher s.

In addition, with the help of RWR, The SoI in various domains is quantified for each researcher. To measure researchers academic feature, we define the vector F with SoI.

3.3 Collaboration Recommendation by Feature Vector Similarity

CCRec recommends collaborators for researchers based on their similarities. To measure the academic features similarities of researchers, we borrow a standard method, *cosine similarity* (CS). CS is employed to define the similarity between two users s_1 and s_2 based on their feature vectors F_{s_1} and F_{s_2}.

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

Finally, we consider that researchers with high similarities have common interests, they should be recommended to each other as potential academic collaborators. Hence, CCRec provides a TopN recommendation list for each researcher.

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.

We embarked on benchmarking experiments on CCRec. To evaluate the performance of CCRec model in a better way, we use three metrics which are widely used in the recommender systems, *Precision*, *Recall* and *F1* [?]. We compared CCRec with the two following approaches. ACRec: a random walk recommendation model based on collaboration networks [?]. CNRec: a common neighbors based recommendation model [?]. Four groups of experiments were conducted. 1, Find the most valuable collaborators, who may have known each other before, or be active in adjacent circles. 2, Recommend most potential collaborators, who have never cooperated with the target researcher before. 3, Evaluate how domains clustering impact the performance of CCRec. 4, Exploit the impact of clustered domains number on CCRec. For each experiment, we randomly chose 100 constant researchers who are at least somewhat active in academic activities, that is they have co-authored more than 30 person-time with others. We make collaborators recommendation for these 100 researchers, moreover compute the average of precision, recall and $F1$.

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 were implemented with Python.

4.1 Most Valuable Collaborators Recommendation

In our previous work [?], We proposed the ACRec model to make the most valuable collaborators recommendation for researchers. In this section, we explored the performance of CCRec and ACRec on making most valuable collaborators recommendation. The comparative results are shown in Fig. 2.

As shown in Fig. 2, The number of recommended collaborators has an obvious influence on the metrics with a clear trend. In the case of CCRec, as shown in Fig. 2(a), the precision drops when the number of recommended collaborators is increasing. At the same time, the recall in Fig. 2(b) rises with the increase of recommendation list, which approximates to 20% in the end. As for ACRec, it has the same trend with CCRec on precision and recall. Thus it can be seen, the precision and recall are a pair of contradictory metrics. Weighing the two metrics to maximum the profit, G. Shani et al. [?] adopt the metric $F1$. Fig. 2(c) describes the performance of CCRec and ACRec on $F1$. In case of CCRec model, the $F1$ generally increases until the number of recommended collaborators is over 15, and then decreases gradually. Since the point 15 is exactly the peak of $F1$. We can see that, CCRec performs best when recommend 15 collaborators to each researcher, and the $F1$ can reach 6.13%. However, in this scenario, ACRec gets its' highest $F1$ score 11.01% at point 30.

In terms of Fig. 2, It is much in evidence that ACRec model outperforms CCRec model on making most valuable collaborators recommendation. This is because, ACRec based on the link-importance guiding random walk, considering the walk distance and rank score, seeks the most valuable collaborators who may have known each other before, or active in adjacent circles. Thus, compared with ACRec, there is no obvious superiority for CCRec to find the most valuable collaborators in adjacent circles.

4.2 Most Potential Collaborators Recommendation

We define the Most Potential Collaborators as collaborators who are worthy of being recommended and have never cooperated with the target researcher. Making the most potential collaborators recommendation is of great significance as the new collaborators are more meaningful and practical in academia reality. In this section, we explored the performance of CCRec, ACRec and CCRec on making most valuable collaborators recommendation.

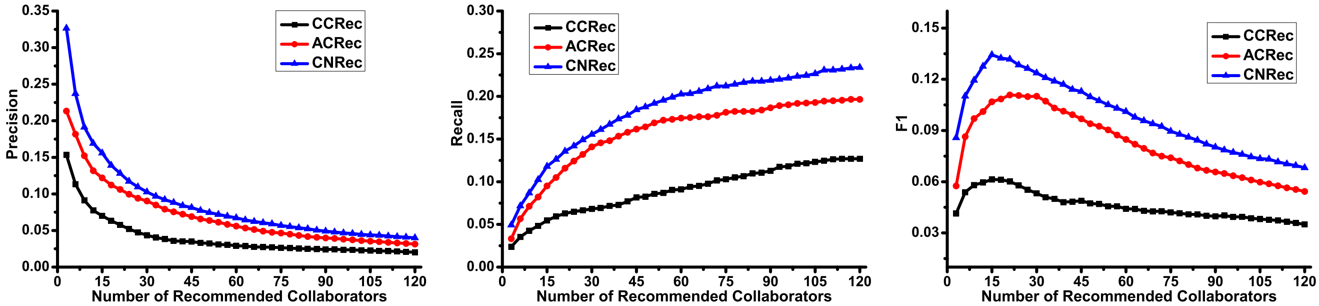
Figure 3 shows the performance of CCRec, ACRec and CNRec in terms of precision, recall and $F1$ with the number of recommended collaborators increasing. It can be observed that CCRec significantly outperforms ACRec and CNRec all the time on these three metrics. CCRec shows a downtrend for precision and an uptrend for recall rate. In the case of $F1$, it reaches the peak 4.18% when recommending 21 researchers. We also see the evidence that when making most potential collaborators recommendation, ACRec performs similarly to CNRec.

In a nutshell, CCRec outperforms ACRec and CNRec with higher precision, recall and $F1$ on making the most potential collaborators. We analysed the theory. Each researcher is represented by the feature vector, as well as CCRec model combines publications contents and collaboration networks to define the vector, which has distinct advantages (e.g. rich information, more accurately to represent researchers' feature) in recommending new collaborators.

4.3 Impact of clustered Domains number

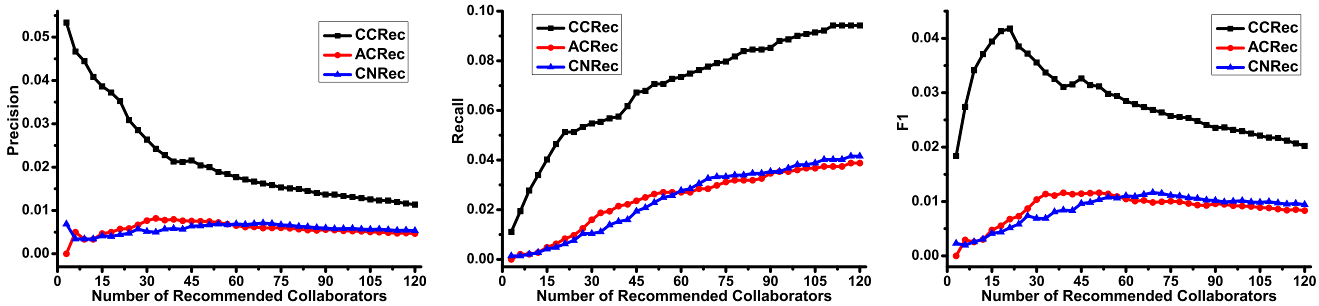
In this section, we exploit the impact of clustered domains number on the performance of CCRec. We adopted the following experiment settings. 1, Evaluating how the $F1$ changes with the number of collaborators recommended. 2, Make the most potential collaborators recommendation for those 100 researchers selected above. 3, Recommend 21 potential collaborators for each researcher. Fig. 4 shows the experimental results.

In terms of Fig. 4, the number of clustered domains does have certain effect on the performance of CCRec. If the



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

Figure 3: Performance of CCRec and ACRec on most valuable collaborators recommendation



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

Figure 4: Performance of CCRec, ACRec and CNRec on most potential collaborators recommendation

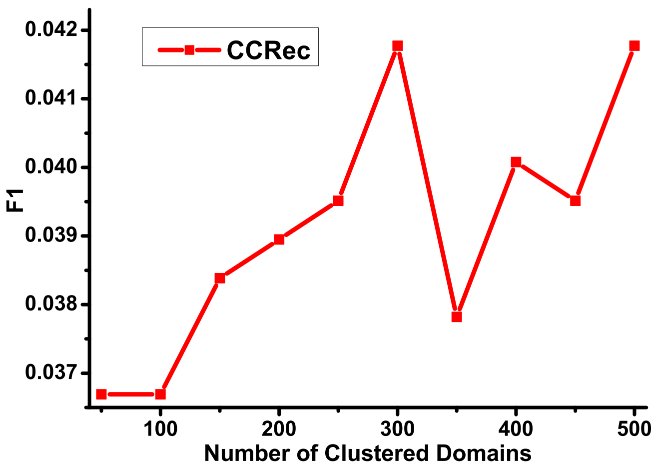


Figure 5: The impact of clustered domains number on CCRec

number of clustered domains is appropriate, the $F1$ score can get some enhancement. In this situation, when clustering the data mining academia into 300 or 500 domains, CCRec performs best for $F1$, which reaches 4.18%.

In summary, we can still claim that the model combining content-based method and social networks-based method is really effective. Moreover, in terms of precision, recall and $F1$, CCRec outperforms ACRec and CNRec on making most potential collaborators recommendation for academic researchers.

5. CONCLUSIONS

In this paper, we focused on how to find researchers' MPCs based on big scholarly data which is necessary in current academia. To this end, we proposed a novel model named CCRec, by combining the features of publications content and collaboration networks. A topic clustering model and A random walk model are adopted to represent the scholars features, and make MPCs recommendation for researchers. We conducted extensive experiments on a subset of DBLP data set to evaluate the performance of CCRec. We also conducted the ACRec and CNRec on the data set as comparisons. The experimental results show that, CCRec outperforms ACRec and CNRec in precision, recall and $F1$ score. With employing topic clustering model, the problem of topic drift has been solved to some extent.

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. Nonetheless, there is still room for future study in this direction. We extracted the titles of publications as the corpus of topic clustering model, which are not more comprehensive than the abstract and main body of publications. Besides, an exactly metric should be confirmed to evaluate the topic drift problem. As a future work, more experiments and studies should be conducted.

6. ACKNOWLEDGMENTS