Exploiting Publication Contents and Collaboration Networks for Collaborator Recommendation

Elsevier[1]

Radarweg 29, Amsterdam

Elsevier Inc[a,b], *Global Customer Service*[b,*]

[a]*1600 John F Kennedy Boulevard, Philadelphia*
[b]*360 Park Avenue South, New York*

Abstract

In academia, studies show that researchers are usually prolific by well collaboration with collaborators. However, due to the expansion of academic researches 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 CCRec, an innovative model that combines content-based method with networks-based method. A topic clustering model and a random walk model are adopted to help effectively seek the most potential collaborators for researchers. Using 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.

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

[*] Corresponding author
Email address: `support@elsevier.com` (Global Customer Service)
URL: `www.elsevier.com` (Elsevier Inc)

[1] Since 1880.

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 [1]. This is to say, the collaborator is a considerable factor well connected with the productivity of a scholar. So researchers tend to discover the most potential collaborators (MPCs, some influential scholars who are similar in interests and have never collaborator before), or reinforce the collaboration with 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. [***references!***]Some traditional content-based methods represent researchers by interests tags, make collaborators recommendation by computing interests similarities.[***references!***] However, interests tags are sometime not accurately to represent researchers' features. Moreover, in general, a researchers shows bias on various academic domains. Such behaviors usually reveal academic features of researchers in different domains. Thus, it would be imperative considering academic domains when recommending collaborators. Our previous work proposed a network-based model, ACRec [2], 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 recommending 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 [3], which provides us many good references.

In this paper, we propose a novel hybrid model exploiting publication contents and collaboration networks for collaborators recommendation (CCRec). Utilizing a topic clustering model [4] [5] and a random walk model, CCRec integrates the features of publications content and collaboration networks. We extract the subject terms from all researchers' publications and cluster these terms into several topics, following distribute researchers to corresponding domains. To represent the feature vectors of each researchers, we run the random walk with restart model (RWR) on each domains, which have been proved to be competent for calculating the rank score of node in social networks. After that, the MPCs recommendation is provided by computing the similarities of researchers' feature vector.

In summary, we make the following contributions in this paper. 1) To make the most potential collaborators recommendation, we develop a model CCRec, which combines the content-based method and collaboration network-based method. That is more favourable to achieve remarkable personalized collaborators recommendation. 2) To reveal researchers' academic features in different domains, we present the feature vectors by utilizing a topic clustering model and a random walk model. 3) We conduct extensive experiments on a subset of DBLP data set to evaluate the performance of CCRec in various scenarios as compared against our previous model ACRec and the normal common neighbors-based model (CNRec). Promising results are presented and analyzed.

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 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

Collaboration plays an important role in academia research, a large body of works focuses on two key issues, reinforcing and discovering collaborators, which are respectively defined as MVCs recommendation and MPCs recommendation in this work. Lopes et al. [6] work on identifying new partners to execute joint research and enhancing the cooperation of current partners for researchers. Chen et al. [7] proposed that the purpose of friends recommendation is "Make new friends, but keep the old". Research on enterprise social networking [8] shows that users in a corporate context are interested in discovering valuable contacts not yet known to them, or connecting to weak ties, in addition to staying in touch with their close colleagues. Our previous work [2] focus on recommending MVCs for researchers and enhancing the cooperation with colleagues in their academic social networks. In this work, CCRec have an aptitude for discovering new collaborators with high similarity (i.e. MPCs recommendation).

In general, collaborators recommender systems are studied in three different perspectives according to the methodologies used to perform recommendation: *content-based*, *social network-based* and *hybrid approach*. The related works presented following are all correlation with these three types models.

content-based, which recommends items classified accordingly to the user profile and early choices considering semantic issues. Das G. et al. [9] propose models for computing the similarity between researchers based on expertise profiles extracted from their publications and academic homepages. Lopes et al. [6] considered researchers' publications area and the vector space model to make collaboration recommendation. Kim et al. [10] propose a collaborative filtering method to provide an enhanced recommendation quality derived from user-created tags. However, researchers often behave differently across multiple domains of interests, therefore, there may be topic drift problems in general recommendation systems [11].

social network-based, which recommend items considering the structural of social network or some social factors. Ma et al. [12] analysed how social networks information can benefit recommender systems and proposed a method improving the performance of recommender systems by incorporating social network information. T. Huynh et al. [13] proposed a method based on combination of probability theory and graph theory for modeling and analysing co-author network. They explore the similar vertices as potential candidates for collaboration recommendation, their main contribution is taking the trend information into considering similarity of vertices. Many other approaches have been presented to formalize academic collaboration recommendation as a link prediction problem [14] [15] in social network. Some of these approaches have been applied to large social networks and results show a good performance. Lichtenwalter et al. [16] examined some important factors for link prediction and proposed a general framework, in addition to our previous work [2].

hybrid, which combines the content-based and social network-based to take advantage of their benefits. Lee et al. [3] 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. Chen et al. [17] discuss CollabSeer, an open system to recommend potential research collaborators for scholars and scientists, which discovers collaborators based on the structure of the coauthor networks and the user's topic of research interests. Cohen et al. [18] also work on solving the collaborators recommendation problem, by combining traditional techniques for structural link prediction in social networks with textual relevancy and global importance metrics.

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

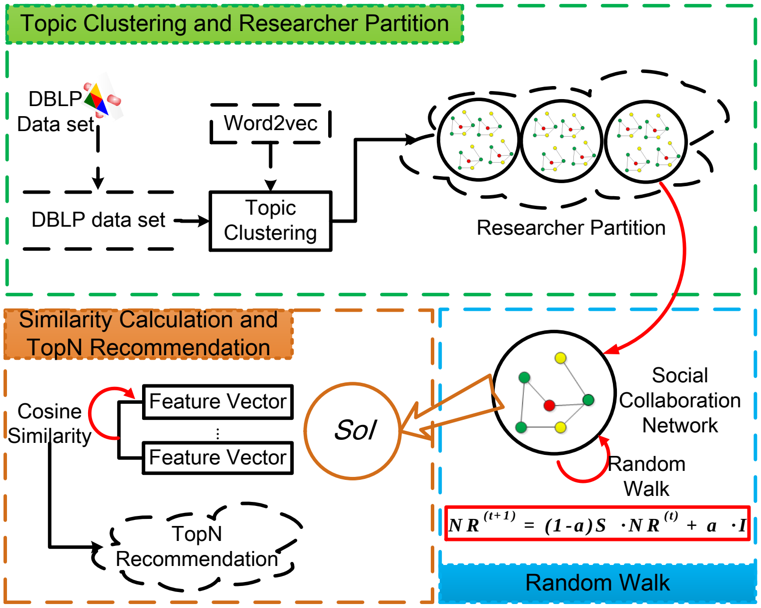

Topic Clustering and Researcher Partition
DBLP Data set
Word2vec
DBLP data set
Topic Clustering
Researcher Partition
Similarity Calculation and TopN Recommendation
Cosine Similarity
Feature Vector
⋮
Feature Vector
SoI
TopN Recommendation
Social Collaboration Network
Random Walk
$NR^{(t+1)} = (1-a)S \cdot NR^{(t)} + a \cdot I$
Random Walk

Figure 1: The architecture diagram of CCRec model

behaviors usually reveal academic features of researchers in different domains. Besides, as a social-based 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 [2], 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.

In addition, CCRec partitions researchers to specific domains with following method. 1, Extract subject terms from a researcher's publications. 2, Traverse all the terms and check the word vector. The model tags the researcher for particular domains which contains these subject terms. It should be emphasized that one researcher always belongs to several domains and there are also many researchers in one domain. Fig. 2 describe an example. Assuming that CCRec extracts 12 subject terms from the publications titles of researcher $S1$. After topic clustering, we can see that, three of these subject terms are assigned to domain A, and seven in B, two in C. Thus, researcher $S1$ is tagged for domains A, B and C. Through this method, each domains contain numerous researchers.

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'

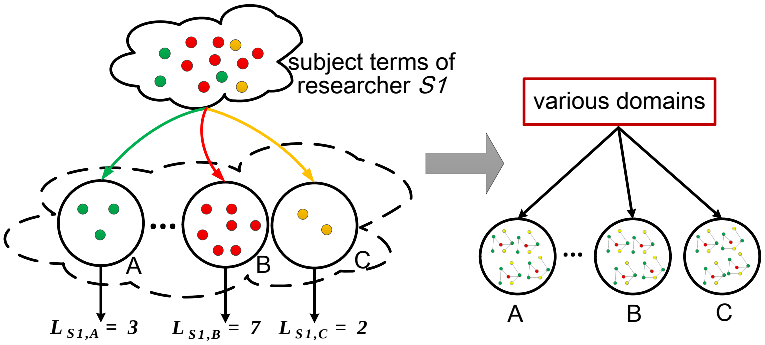


Figure 2: Researcher Partition

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}.

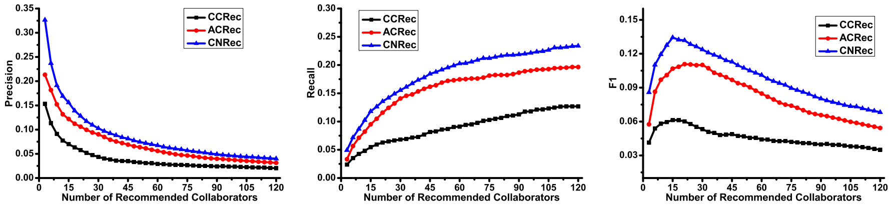
$$Sim(s_1, s_2) = \frac{\sum_{i=1}^{n}(F_{s_1,i} * F_{s_2,i})}{\sqrt{\sum_{i=1}^{n} F_{s_1,i}^2} * \sqrt{\sum_{i=1}^{n} F_{s_2,i}^2}} \tag{2}$$

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 [19], 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* [20]. We compared CCRec with the two following approaches. ACRec: a random walk recommendation model based on collaboration networks [2]. CNRec: a common neighbors based recommendation model [6]. 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



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

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

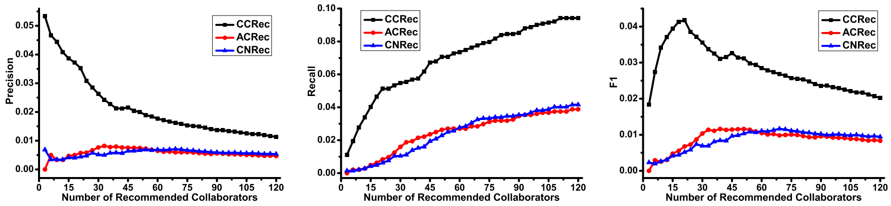
randomly chose 100 constant researchers who are at least somewhat active in academic activities, that is they have co-authored more than 30 person-time with others. We make collaborators recommendation for these 100 researchers, moreover compute the average of precision, recall and $F1$.

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 [2], 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. [20] 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



(a) Precision

(b) Recall

(c) F1

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

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 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.

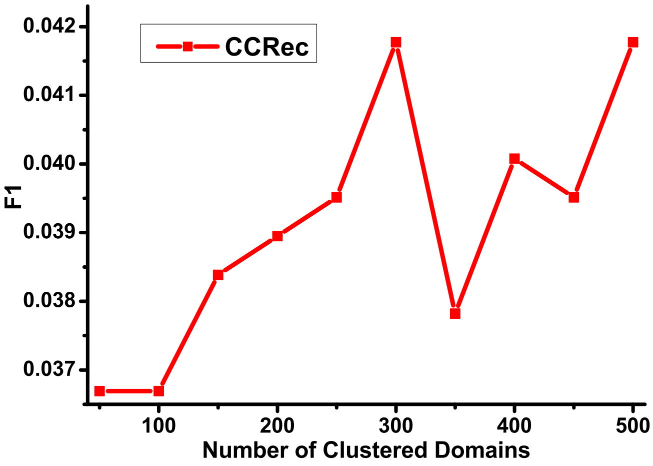


Figure 5: The impact of clustered domains number on CCRec

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 obtain 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 content-based method and networks-based method 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 future work, more experiments and studies should be conducted.

References

[1] S. Lee, B. Bozeman, The impact of research collaboration on scientific productivity, Social studies of science 35 (5) (2005) 673–702.

[2] J. Li, F. Xia, W. Wang, Z. Chen, N. Y. Asabere, 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, International World Wide Web Conferences Steering Committee, 2014, pp. 1209–1214.

[3] D. H. Lee, P. Brusilovsky, T. Schleyer, Recommending collaborators using social features and mesh terms, Proceedings of the American Society for Information Science and Technology 48 (1) (2011) 1–10.

[4] C. Pan, W. Li, Research paper recommendation with topic analysis, in: Computer Design and Applications (ICCDA), 2010 International Conference on, Vol. 4, IEEE, 2010, pp. V4–264.

[5] M. C. Pham, Y. Cao, R. Klamma, M. Jarke, A clustering approach for collaborative filtering recommendation using social network analysis., J. UCS 17 (4) (2011) 583–604.

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

[7] J. Chen, W. Geyer, C. Dugan, M. Muller, I. Guy, Make new friends, but keep the old: recommending people on social networking sites, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2009, pp. 201–210.

[8] J. DiMicco, D. R. Millen, W. Geyer, C. Dugan, B. Brownholtz, M. Muller, Motivations for social networking at work, in: Proceedings of the 2008 ACM conference on Computer supported cooperative work, ACM, 2008, pp. 711–720.

[9] S. D. Gollapalli, P. Mitra, C. L. Giles, Similar researcher search in academic environments, in: Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, ACM, 2012, pp. 167–170.

[10] H.-N. Kim, A.-T. Ji, I. Ha, G.-S. Jo, Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation, Electronic Commerce Research and Applications 9 (1) (2010) 73–83.

[11] J. Tang, S. Wu, J. Sun, H. Su, Cross-domain collaboration recommendation, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2012, pp. 1285–1293.

[12] H. Ma, D. Zhou, C. Liu, M. R. Lyu, I. King, Recommender systems with social regularization, in: Proceedings of the fourth ACM international conference on Web search and data mining, ACM, 2011, pp. 287–296.

[13] T. Huynh, K. Hoang, D. Lam, Trend based vertex similarity for academic collaboration recommendation, in: Computational Collective Intelligence. Technologies and Applications, Springer, 2013, pp. 11–20.

[14] H.-H. Chen, L. Gou, X. L. Zhang, C. L. Giles, Discovering missing links in networks using vertex similarity measures, in: Proceedings of the 27th Annual ACM Symposium on Applied Computing, ACM, 2012, pp. 138–143.

[15] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, J. Han, Co-author relationship prediction in heterogeneous bibliographic networks, in: Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on, IEEE, 2011, pp. 121–128.

[16] R. N. Lichtenwalter, J. T. Lussier, N. V. Chawla, New perspectives and methods in link prediction, in: Proceedings of the 16th ACM SIGKDD

international conference on Knowledge discovery and data mining, ACM, 2010, pp. 243–252.

[17] H.-H. Chen, L. Gou, X. Zhang, C. L. Giles, Collabseer: a search engine for collaboration discovery, in: Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, ACM, 2011, pp. 231–240.

[18] S. Cohen, L. Ebel, Recommending collaborators using keywords, in: Proceedings of the 22nd international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee, 2013, pp. 959–962.

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

[20] G. Shani, A. Gunawardana, Evaluating recommendation systems, in: Recommender systems handbook, Springer, 2011, pp. 257–297.