Scientific Collaborator Recommendation in Heterogeneous Bibliographic Networks

Chen Yang
University of Science and Technology of China
City University of Hong Kong
yangc0201@gmail.com

Jianshan Sun
School of Management
Hefei University of Technology
sunjs9413@gmail.com

Jian Ma
Department of Information Systems
City University of Hong Kong
isjian@cityu.edu.hk

Shanshan Zhang
University of Science and Technology of China
City University of Hong Kong
zss_200890@126.com

Gang Wang
School of Management
Hefei University of Technology
wgedison@gmail.com

Zhongsheng Hua
School of Management
University of Science and Technology of China
zshua@ustc.edu.cn

Abstract

Most of the previous studies on scientific collaborator recommendation are based on social proximity analysis to suggest collaborators. However, the extracted homogeneous features cannot well represent the multiple factors which may implicitly affect the future scientific collaboration. In this paper we propose an approach based on the multiple heterogeneous network features, which has produced good results in our experiments based on a dataset of more than 30,000 ISI papers. This method can help solving the similar problems of people to people recommendation. It generates high quality expert's profiles via integrating research expertise, co-author network characteristics and researchers' institutional connectivity (local and global) through a SVM-Rank based information merging mechanism to perform intelligent matching. The generated comprehensive profiles alleviate information asymmetry and the multiple similarity measures overcome problems related to information overloading. The proposed method has been implemented in ScholarMate research network (www.scholarmate.com) which is a research 2.0 innovation, promoting research collaboration in virtual scientific community.

1. Introduction

In the past time, scientific researchers can only get acquainted with others who are close in the small coterie they involved in. Thanks to the rapid development of information techniques, social network empowered collaboration platforms have appeared and changed many people's life styles across continents, from developed countries to developing countries [1, 2]. A great deal of research scientists establish relationships and interact with one another in the scientific social network platforms such as ScholarMate, which is one of the largest scientific social network in China [3]. Millions of researchers are involved in the virtual communities, while it causes a big problem of information overloading. Thus, recommender systems are raised up in response to this issue [4, 5].

Most of the previous scientific collaborator recommendation studies are confined to the homogeneous co-authorship networks. These works usually limit the relied network on a homogeneous co-authorship network and only the network topology features are adapted [6]. The subsequent works have revealed that many latent features can also influence the researcher's intention to collaborate [7, 8]. Thus solely utilizing the network based features cannot achieve an optimal recommendation result. How to effectively and efficiently extract and make use of the multiple features in the heterogeneous bibliographic networks which may implicitly affect the future scientific collaboration is the key research question in this paper. For example, the research expertise of researchers, researcher affiliation network, the researchers' authority and the collaboration frequency are important features which may influence the performance of collaborator recommendation. In recent years, more and more researchers have realized the important impact of heterogeneous bibliographic

1530-1605/15 $31.00 © 2015 IEEE
DOI 10.1109/HICSS.2015.73

networks, thus heterogeneous network analysis methods become a mainstream trend and are playing a crucial role in the subsequent research in this area [8, 9].

Based on the above discussions, this paper wants to fill the existing gap by exploring and validating the latent features which can be leveraged to increase the accuracy of the collaborator recommender systems. In recent years, more and more researchers have realized the important impact of heterogeneous bibliographic networks. Though the integration of research topic similarity and co-author network analysis methods can benefit from both aspects, information including researcher-affiliation-paper network of scholars as well as the collaborator frequency preference of experts (i.e. the collaboration counts of two collaborators) has been largely ignored. For instance, some researchers prove that fusing the neighbourhood based network proximity measure and path based proximity measure can obtain a surprising improvement [10]. Some additional factors such as the collaboration frequency between two researchers would also affect the relationship between two researchers, as well as the author counts of a co-authored paper, namely exclusivity in [5]. Hence a collaboration frequency and paper exclusivity revised global path similarity is developed and combined with traditional local neighbourhood based method in the personal proximity module of this paper. Analogously, a collaborative filtering strategy boosted local institutional connectivity is proposed to enhance the previous global institutional connectivity in the institutional connectivity module [7, 11].

In this paper, we propose a hybrid approach combining five features from three heterogeneous networks: the research topic network, researcher collaboration network and the institution network. We use a language model based method to represent the expertise similarity in relevance aspect. For the co-authorship network, we build a novel distance feature which considers both the author number in a particular paper as well as the collaboration frequency in the shortest path between two nodes. We also propose a local institution preference indicator to enhance the previous global institutional connectivity measure [7].

The proposed method has been implemented in a recommendation system in the ScholarMate platform (www.scholarmate.com). Millions of Chinese researchers have registered and connected with each other in this platform. It provides various academic applications to the users such as research CV, collaborator recommendation, paper recommendation, and so on. To further validate the effectiveness of our approach we develop a scientific collaboration dataset and conduct experiments based on it. We compare it with several benchmarks, and the result shows the proposed approach achieves significant improvement both in the accuracy and efficacy measures. It suggests that by leveraging the heterogeneous networks information, our proposed approach which combines five individual features is a validated and robust solution for collaborator recommender systems.

This paper is organized as follows. Section 2 presents the related works of collaborator recommendation. The proposed approach which combines multiple features in the three-layered heterogeneous networks is presented in section 3. Section 4 illustrates the implementation of the proposed approach in a real recommender system and the experimental results comparing with several benchmark methods. Section 5 concludes the paper and lists a few directions in the future work.

2. Related work

Previous studies mainly regard the collaborator recommendation issue from two different aspects: the link predictions problem and general expert recommendation for collaboration.

In the previous studies, researchers and their co-author relationships are extracted to build a homogeneous co-author network. Several common network topology proximity based measures are compared and the Adamic-Adar (neighborhood based indicator) and several weighted global path based measures shows good performance in the conducted experiments [6]. A supervised random walks algorithm is proposed and validated with the co-authorship dataset in [12]. Recently, more features from other dimensions such as semantic similarity have been used to enhance the predictions. The papers' venues, research topics and previous papers of researchers are adopted into the heterogeneous networks, and a logistic regression model train and make predictions with the topology features derives from the networks [8]. Han et al. integrate the local neighborhood and global path measures and combine them with semantic similarities, and the experiment results demonstrate its effectiveness for collaborators recommendation [10].

The general collaborator recommendation studies focus on detecting and validating the features for efficient recommendation. For example, a liner hybrid of research topic and global network is raised to recommender collaborators [9]. A two-layered model combines social network and semantic similarities is introduced and it generate an improvement comparing to the individual features [3]. Yang et al. demonstrate the effect of a novel global institutional connectivity for collaborator recommendation in [7]. The structural proximity and textual similarity are integrated to retrieval proper experts for collaboration [13].

3. The proposed approach

As show in the figure below, we propose a hybrid approach combing five features to form three heterogeneous networks: the research topic network, researcher collaboration network and the institution network. A language model based method is introduced to represent the expertise similarity in relevance aspect. In the collaboration network we build a new distance feature which incorporates both the author counts in a particular paper as well as the collaboration frequency into the shortest path between two nodes. A local institution preference indicator is proposed to complement the global institutional connectivity measure in the institution network. Finally, the SVM-Rank method is leveraged to fuse the five features in the three-layer heterogeneous network.

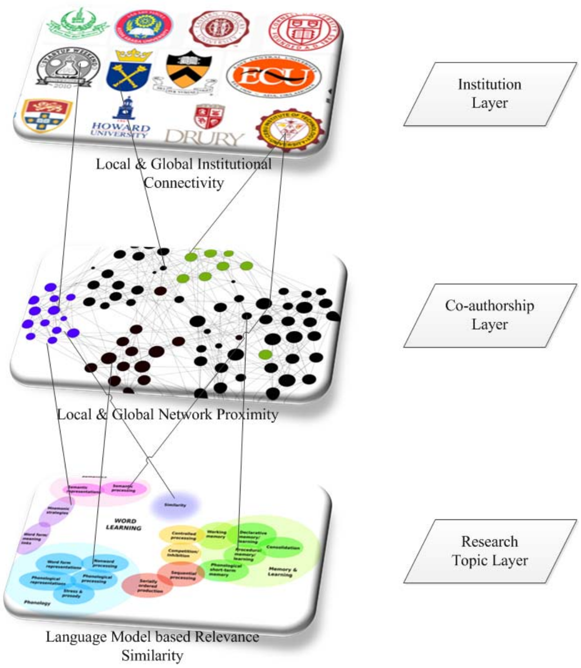

Institution Layer
Local & Global Institutional Connectivity
Co-authorship Layer
Local & Global Network Proximity
Research Topic Layer
Language Model based Relevance Similarity

Figure 1. The three-layer heterogeneous network

3.1. Research relevance similarity

The relevance similarity calculation aims at identifying candidate collaborators with similar expertise. In this article we employ a language model based method to address the expertise similarity calculation issue. Language model is widely used to measure the semantic similarity between a query and the candidate documents in information retrieval domain [14-16]. The common applied mechanism of language model is to estimate a particular language model for each of the documents in the corpus, and use the likelihood value of the query topics associated with each model to rank the candidate documents. In this paper, we incorporate the language model to formalize the expertise similarity calculation, because it can generate accurate semantic matching measure with a low computational complexity as efficient as traditional TF-IDF model [17]. Thus it is a good choice if we want to generate good result in short time periods, especially in the big data context. To demonstrate the effectiveness of the language model based method, we compare it with traditional BM25 approach in the evaluation stage [18].

After identifying expert profiles which can be viewed as the combination of past published papers as descripted in [19], the researchers' profiles are matched utilizing language models. We assume that a target researcher *r*'s profile is generated by a language model based on one of the candidate collaborators *c*, i.e. the one with the highest likelihood. Thus we can rank the candidate collaborators based on the probability of the candidate collaborator *c* being a domain expert given a researcher *r*'s profile. We can define the probability and apply Bayes' Theorem to derive this problem, we can obtain:

$$p(c\,|\,r) = \frac{p(r\,|\,c)\,p(c)}{p(r)} \propto p(r\,|\,c) \qquad (1)$$

where $p(c)$ is the probability of a candidate collaborator, here is assumed to be uniform as we regard that all the candidates in the set are eligible to be collaborators with the target researcher and not assign a prior to it. $p(r)$ is the probability of a researcher *r* and it is candidates independent constant. Thus the purpose of finding which candidate can achieve the highest score can be reduce to the calculation of $p(r\,|\,c)$.

In this stage, we treat the words in the researchers' profiles as unigram models, so the probability can be inferred as follows,

$$p(r\,|\,c) = \prod_i p(t_i\,|\,c) \qquad (2)$$

$$\log p(r\,|\,c) = \sum_{i:tf(t_i;c)>0} \log \frac{p_s(t_i\,|\,c)}{\alpha_c * p_u(t_i\,|\,C)} + n\log\alpha_c + \sum_i \log p(t_i\,|\,C) \qquad (3)$$

where α_c is a candidate dependent variable which represents the ratio of unseen words in the calculation, and $tf(t_i \mid c)$ is the occurrence count of term t_i in candidate *c*'s profile. We can infer from the first part, the weight of a matched term is proportional to the term frequency and inversely proportional to the collection frequency, which is fit with traditional assumption in TF-IDF models. As the target user's profile is usually a long query, we employ the Jelinek-Mercer method as the smoothing method [19],

$$p_\lambda(t \mid c) = (1-\lambda) p_{ml}(t \mid c) + \lambda p(t \mid C) \quad (4)$$

$$p_{ml}(t \mid c) = \frac{tf(t;c)}{|c|} \quad (5)$$

where $p_{ml}(t \mid c)$ represents the maximum likelihood estimation of the probability that the term *t* appeared in target researcher's profile can be generated by the language model of candidate collaborator *c*'s profile. The parameter λ which combines two language models is called Jelinek-Mercer smoothing parameter and it is introduced to serve as the role of α_c, thus it can avoid under estimation of probability of the unseen terms in the candidate profile *c*. It usually takes values in the range from 0.1 to 0.7, and for the long profile of target researcher, the optimal value is usually around 0.7 [17, 20]. In this article, we set the λ as 0.75.

3.2. Personal social network proximity

Social network proximity is used to measure the availability and connectivity of the recommended researchers towards effective collaboration with the target user. Arazy and Kumar argue that the personalized recommendation should be guided by some advice-taking theories (e.g., tie strength theory) in order to identify reliable connections [21]. For instance, the tie strength theory holds a view that a user's likelihood to accept the suggestions is determined by the relationship intensity between the user and the source. According to these theories, researchers with high topology proximity in their collaboration network tend to become trustable partners in the future. Many studies on collaborators recommendation revealed that network proximity is a key factor to consider [8, 22, 23].

Previous co-author link prediction studies usually adopt some local social network similarity measures, such as preferential attachment similarity, Jaccard similarity and Adamic-Adar similarity (most of which are neighborhood based) [6]. They usually combine local social network topology features with content based measures to recommender experts [6, 7, 19]. Recently some researchers reveal the importance of global network topology measures (most of which are path based), hence they incorporate the global network features (such as shortest path and Katz similarity) into the collaborator recommendation model and obtain better performance in the real world data tests [10].

In this paper, we also attempt to take advantages of both the local and global network proximity features in the personal social network proximity module. We choose the Adamic-Adar measure as the local proximity similarity, and it is a well-known neighborhood based indicator with proved robust performance in previous studies [6, 7]. In particular, a new global topology similarity is proposed on the basis of the collaboration frequency and the exclusivity in co-author networks [5]. It is an improved version based on the shortest path measure, and we demonstrate its capability to enhance the local network proximity in the subsequent experiments.

Research social network data extracted from the corpus of publications is used to construct the collaboration network. A node of this network represents one researcher and an edge represents the relationship as co-authors between the researchers. Once we construct the network, following Adamic-Adar similarity the local proximity between any two researchers is measured as descripted in [24]. As a local network proximity measure, Adamic-Adar similarity adapts the common friends set between two researchers in the collaboration network to perform calculation. The mechanism of it can be illustrated by the following equation:

$$Local\_Proximity(j,k) = \sum_{z \in \Gamma(j) \cap \Gamma(k)} \frac{1}{\log|\Gamma(z)|} \quad (6)$$

where *z* denotes the common neighbors of researcher *j* and *k* and $\Gamma(j)$ denotes the neighbors collection of researcher *j*. As a result, the researchers with lots of friends will be assigned a low score reflecting that very popular experts may not be able to contribute much for effective contribution.

The global network similarities usually take into account of path information from an overall perspective [10]. Most of these studies regard the weight of co-authorship links as uniform, and it ignores the frequency of co-authorship and the total number of co-authors in an article, which may also affect the similarity between two researchers [5]. The co-authorship frequency and exclusivity are defined and formulated in [5], hence in this paper we identify a novel global proximity measure based on shortest path similarity by incorporated the co-authorship frequency and exclusivity, which is given by

$$Global\_Proximity(j,k) = [1+ \sum_{(j;k)\in shortest\_path} \frac{1}{\sum_{i=1}^{co(j;k)} \frac{1}{f(a_i)}}]^{-1} \quad (7)$$

where $co(j;k)$ denotes the co-authored paper set of the researcher *j* and researcher *k*, $f(a_i)$ means the total number of authors in the paper. We can observe from this equation that the two crucial factors are implanted into the shortest path index to get a weighted path based matrices.

3.3. A comprehensive institutional connectivity

Furthermore, the relationship at institutional level will also affect the collaboration [7, 21, 25]. According to the trust theory, the impact of trust will influence the user's choice (when choosing a collaborator) in organizational advice networks [21, 26]. In other words, researchers tend to work with others who come from collaborated institutions because it suggests a higher trust level. A overall institutional connectivity is stated in [7], where the collaboration strength of potential collaborators at organizational level in terms of joint published publications is gathered and calculated. The ISI subject categories of the publications serve as the group label. Hence the global institutional level connectivity of researcher *k* and *j* at subject category *c* is denoted by $Col_{u_j,u_k,c}$ using Jaccard similarity, which can be generated as follows,

$$Col_{u_j,u_k,c} = \frac{N_{u_j,c} \cap N_{u_k,c}}{N_{u_j,c} \cup N_{u_k,c}} \quad (8)$$

where $N_{u_j,c}$ indicates the publications collection of institution u_j under the ISI subject category *c* [7]. Due to the fact that each researcher might be familiarized with one or more subject categories, the maximal value of organizational connectivity measures is to be selected as the global institutional connectivity between two researchers, that is:

$$Global\_InstC_{j,k} = \operatorname*{Max}_{c\in[c_j \cap c_k]} (Col_{u_j,u_k,c}) \quad (9)$$

where c_k denotes the subject categories of researcher *k*, c_j represents the subject categories of research *j* and *c* represents the common subject categories between researcher *k* and *j*.

Previous study has proved the significant performance of global institutional connectivity in collaborator recommendation study [7]. However, the global institutional connectivity is a general indicator which obtains the similarity from a whole institution level while ignoring the individual diversity of institutional preference. Especially for the researchers who have lots of publications as well as lots of cooperative institutions, it is improper to treat them at an overall level. In this paper, we propose a local institutional connectivity measure to fill in this gap. The local institutional connectivity consults the idea of traditional collaborative filtering methods [11, 27], that means the researchers' previous collaborated institutions will affect the final choice, the equation is as follows,

$$Ins_{j,u_i} = \frac{tf(j;u_i)}{|U_j|} \quad (10)$$

where $tf(j;u_i)$ denotes the total number of publications of researcher *j* which contains the affiliation u_i as collaborated institution. U_j means the publication number of research *j*. Analogously, each researcher might have several subject categories, the local institutional connectivity is given by

$$Local\_InstC_{j,k} = \operatorname*{Max}_{u\in[u_k]} (Ins_{j,u}) \quad (11)$$

where u_k denotes the institutions of researcher *k*.

3.4. Recommend research collaborators based on SVM-rank fusion strategy

Once individual similarity measures are determined it is necessary to integrate them in order to generate the lists of recommended experts. Thus we utilize a well-known learning to rank technique, SVM-Rank method, to obtain the optimize weight of each feature and aggregate the similarity measures [28, 29]. The SVM-Rank method is used in the proposed approach to learn the weights of the similarity scores obtained from different network layers and to calculate the final similarity score for collaborator recommendation application.

As an initial step, in order to guarantee fair comparison among recommended results it is necessary to normalize all the matching scores, here we use min–max normalization strategy [30], which is given by,

$$score_{norm} = \frac{score_{original} - score_{min}}{score_{max} - score_{min}} \quad (12)$$

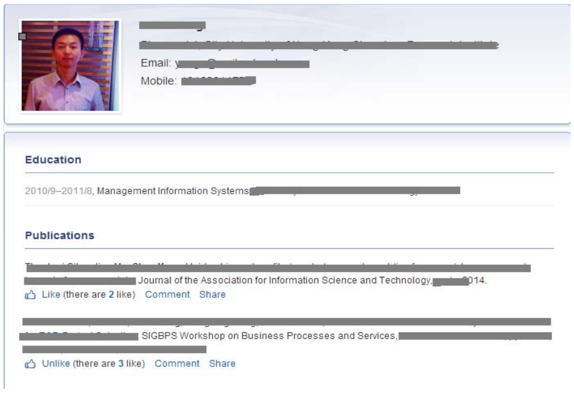
A list of normalized similarity values ranging from 0 to 1 can be generated at the end of the normalization process. SVM-Rank is a kind of Support Vector Machine (SVM) classification method [28, 31]. The learner of SVM-Rank algorithm will select a class of linear ranking functions from a family of ranking functions that maximizes the specific empirical variable [28]. As described in [31], it solves the optimization problem through learning a hyper-plane to separate the positive and negative results in the classification process[1].

In this paper, the positive label denotes the researcher pair are collaborators and the negative label indicates two researchers are not collaborators in the training set. Five lists of ranks according to three kinds of similarity measures are constructed. Let r_1, r_2, r_3, r_4 and r_5 denotes the score of the five lists, what respectively express relevance, global individual connectivity, local individual connectivity, global institutional connectivity and local institutional connectivity, with normalized similarity score. Thus the input to learn SVM-Rank weight is I = < r_1, r_2, r_3, r_4, r_5, label>, label is 0 or 1 based on whether the row data is positive. After training, we can establish the liner weight of the five lists, as W = <w1, w2, w3, w4, w5>. This method can automatically learn the parameters and is more convenient than other surprised fusion strategy.

4. System implementation & evaluation

4.1. Collaborator recommendation system

The proposed approach is implemented as one of the application services in ScholarMate, which is a well-known scientific community platform in China [3]. This service mainly focuses on recommending collaborators who can be served as research partners for academic writing. The researchers can expediently generate the research CV in ScholarMate platform. The generated CV is showed in the figure 2, the publications are displayed normatively. The screenshot in figure 3 represents collaborator recommendation interface. It reports the suggested collaborators based on the extracted information from the researcher's generated CV.



[redacted]

Email: [redacted]

Mobile: [redacted]

Education

2010/9–2011/8, Management Information Systems[redacted]

Publications

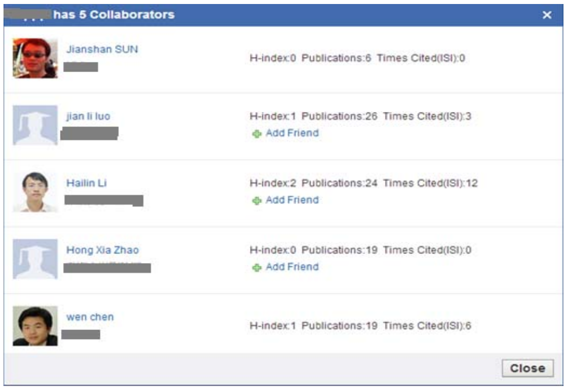
[redacted] Journal of the Association for Information Science and Technology,[redacted]014.

Like (there are 2 like) Comment Share

[redacted] SIGBPS Workshop on Business Processes and Services,[redacted]

Unlike (there are 3 like) Comment Share

Figure 2. Screenshot of the research CV



[redacted] has 5 Collaborators ×

Jianshan SUN [redacted] H-index:0 Publications:6 Times Cited(ISI):0

jian li luo [redacted] H-index:1 Publications:26 Times Cited(ISI):3 Add Friend

Hailin Li [redacted] H-index:2 Publications:24 Times Cited(ISI):12 Add Friend

Hong Xia Zhao [redacted] H-index:0 Publications:19 Times Cited(ISI):0 Add Friend

wen chen [redacted] H-index:1 Publications:19 Times Cited(ISI):6

Close

Figure 3. The recommender system interface

4.2. Experimental evaluation

In this section, we evaluate the proposed approach which utilizes heterogeneous network features in a real dataset. As there are no available dataset contains with the multiple dimensional information demanded by our approach, we crawled the data from ISI database and built a new dataset to validate the effectiveness of the proposed approach by comparing it with several benchmarks.

4.2.1. Data collection. The cataloging information of 36,424 research papers for was crawled in the dataset, which covered 5 ISI subject categories from 2008 to 2013 (which included information systems, artificial intelligence, software engineering and library science related works). We extracted 72,515 distinct authors and 9,731 distinct institutions as well as 314,166 co-author relationships from the corpus. For each of the papers we obtained and correlated its title, keywords, abstracts, authors, their affiliated institutions and publish year. The data from time period 2008-2011 was served as the source of training set and data from 2012-2013 was used as initial test set.

[1] The implementation of SVM-Rank can be found at http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html.

During the data cleaning phase, we discarded authors who do not has co-authored with others both in the training set and initial test set as it will disturb the recommendation mechanism and make noise as described in [9]. We further select only researchers who have co-authored with a researcher more than 2 times to refine the test set. After data cleansing, we got 563 test authors together with 964 links as the gold standard in the test set. For each of the test authors, we randomly selected 50 times researchers from training set as negative pairs. There are 17,453 related researchers with 72,549 co-author links remained in the training set. The data statistics are presented in the following table.

Table 1. Statistics of the experimental dataset

No. of Researchers		72515
No. of Publications		36424
Affiliations		9731
Training Set	Related Researchers	17453
	Links in the Network	72549
Test Set	Target Researchers	563
	Positive Links	964
	Negative Links	48200

4.2.2. Experimental design and evaluation matrices. In this stage, we conducted experiments to validate the effectiveness of the proposed methods by comparing with different types of methods in the individual layers. To demonstrate the advantages of the proposed heterogeneous network features enhanced approach versus methods using homogeneous network features, we chose and implemented six benchmark methods. Respectively, we implemented the language model based relevance method, Adamic-Adar method (local proximity method), revised shortest path measure (global proximity method), local institutional connectivity and global institutional connectivity, a hybrid approach combine local and global institutional connectivity and a hybrid combining the five individual methods from three layers as whole. All the test methods (5 individuals and 2 hybrids) were compared in the evaluation. In order to facilitate the reading, the related methods and their corresponding abbreviations are listed as follows:

Table 2. The abbreviations of the testing methods

Abbreviation	Methods
LM	Language model based relevance method,
AA	Adamic-Adar method (local proximity method)
RS	Revised shortest path measure (global proximity method)
LI	Local institutional connectivity
GI	Global institutional connectivity
HT	A hybrid combining local and global institutional connectivity
HF	A hybrid combining the five individual methods (the proposed method)

In this paper, SVM-Rank method serves as the fusion method as descripted in Section 3.4, thus a fraction of (one-third) data is randomly selected from training set and is used to tune the weight of the hybrid algorithms.

We employed two commonly used information retrieval metrics: F-measure and mean reciprocal rank (MRR) to evaluate our experimental results. F-measure, which is the harmonic average value of precision and recall, is adapted to measure the accuracy of the recommend approach. MRR is used to measure the rank of the first correct collaborator in the result list. The formulas of these metrics are listed as follows:

$$\text{Precision@N} = \frac{No._{correct}}{N} \tag{13}$$

$$\text{Recall@N} = \frac{No._{correct}}{Length_{gold\_standard}} \tag{14}$$

$$\text{F-measure=}2 * \frac{precision * recall}{precision + recall} \tag{15}$$

$$\text{MRR=}\frac{1}{|\text{U}|}\sum_{\text{i}=1}^{|\text{U}|}\frac{1}{rank_i} \tag{16}$$

where, N denotes the length of recommendation result list, in this study we set N= 5, 10, 15, 20 and 25. $No._{correct}$ denotes the number of hit researchers in the recommendation result which is verified through the gold standard. $Length_{gold\_standard}$ represents the number of researchers in the gold standard (i.e. true positive and false negatives). $|\text{U}|$ is the count of test users and $rank_i$ is the position of the first hit result in user *i*'s recommendation list.

4.2.3. Results & analysis. The following figure shows the comparison of F-measure of the testing methods at N = 5, 10, 15, 20 and 25. It is clear from the picture that the proposed method significantly outperformed the five individual methods and one hybrid methods in terms of F-measure metrics. It suggests that leveraging the heterogeneous network features could improve the performance of collaborator recommendation.

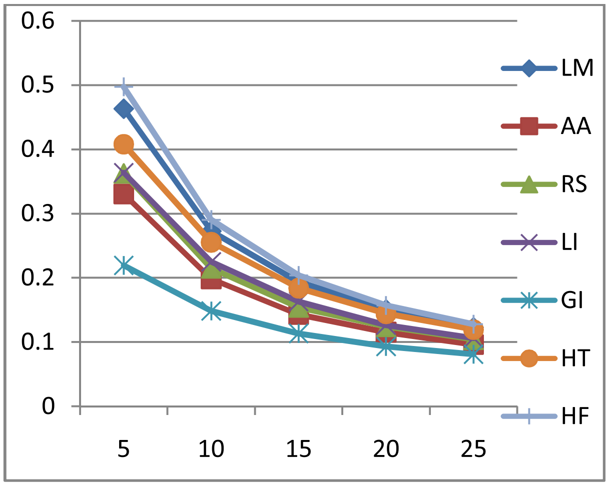


Figure 4. Comparison of F-measures of the proposed method (HF) and 6 baselines

The language model based method (LM) achieve a high accuracy, the reason behind is that the test dataset consisted of researchers from different disciplines, and the keywords used vary a lot from one researcher to another. If we narrowed down the selected disciplines scope, its effect would decrease. The two network proximity based measures, Adamic-Adar method (AA) and revised shortest path measure (RS), both generate proper result. And by incorporating the co-authorship frequency and exclusivity factor into the global path based approach, the latter is superior to the neighborhood based Adamic-Adar method. We can easily observe that the local institutional connectivity (LI) exceeds the global institutional connectivity (GI) distinctly. It is because the test users in our dataset have been refined in the data cleaning stage and only researchers with many collaborators (usually with many collaborated institutions) are chosen, thus the collaborative filtering strategy based LI get better performance. GI has its advantages in dealing with researchers with few collaborators and few collaborated institutions. Thus we can infer that the two methods can complement with each other well, which is demonstrated by the HT hybrid method. It generates a really good performance comparing with the LI and GI. Based on the achieved result of the proposed method (HF), we can conclude that our proposed method outperformed the 6 benchmark methods significantly and it shows the effectiveness by fusing the multiple features in the heterogeneous networks.

We further consider the rank of the recommended researchers in the final recommendation list to measure the relative importance of them via MRR. As Figure 5 shows, our proposed HF method gets the best performance and the MRR value is 0.9680.It reveals that the first hit collaborators at the top positions in the recommendation list and the user will be satisfied with the efficiency of the proposed approach.

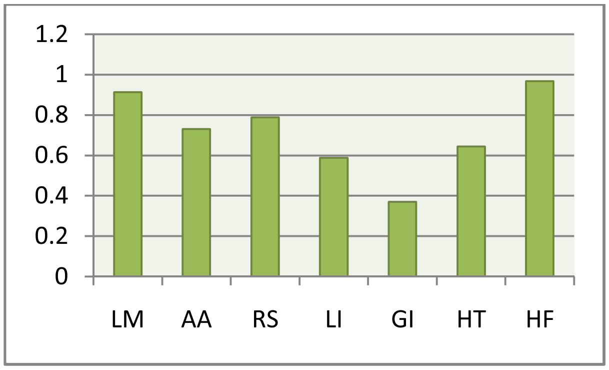


Figure 5. Comparison of MRR of benchmark methods and the proposed method

From this picture, we can also find that although the LI method archives higher score than network proximity measures (AA and RS) in the F-measure, neither the MRR scores of LI, GI, nor HT is higher than AA and RS. It reflects that the network proximity measure works well in the very front of the list.

5. Conclusions and future work

In this research work, a heterogeneous network features based hybrid approach is proposed for the research collaborator recommendation. The virtual scientific platforms have enabled the researchers (especially for those come from developing countries) to efficiently connect with peers and seek for collaboration opportunities. Our proposed solution for scientific collaborator suggestion consists of three main contributions. First we provide a new network proximity measure which incorporates the collaboration frequency and collaboration exclusivity factors into shortest path measure, and it is a complementary feature to the traditional neighborhood based methods. Second we proposed a local institutional connectivity measure, which derives from the collaborative filtering strategy, and it shows a surprising good performance than global institutional connectivity. Thirdly, we model all the five heterogeneous network features into a unified framework with a surprised SVM-Rank based method, and the experimental results demonstrate its robust performance.

There are also some research limitations in this study. For instance, to better demonstrate the performance of the proposed approach, we should test the experimental results on more data sets to obtain more convincing evidences in the next stage.

We have implemented the proposed approach into a scientific social network platform in China and it will help people to find collaborators and share knowledge with others. It is obviously that the proposed collaborator recommender system will help researchers to communicate with domain fellows and achieve information technology enhanced collaboration. From the experiment results we can find the combination of local and global institutional connectivity can achieve a good cross effect. In the future, we will further investigate how to balance the local and global institutional connectivity based on the researcher's previous publication and collaborators counts. Additionally the graph based algorithm such as random walk with restart and spreading activation algorithms can be leveraged to enhance the fusion process in the heterogeneous networks.

Acknowledgement

The authors would like to thank the mini-track chair and the anonymous reviewers for their valuable comments which have great benefits for modifying and enhancing our paper. This work was partially supported by National Natural Science Foundation of China under Grant (71101042, 71471054, 71171172 and 71090401/71090400), Specialized Research Fund for the Doctoral Program of Higher Education (20110111120014) and the China Postdoctoral Science Foundation (2011M501041, 2013T60611).

References

[1] Cheng, X., Nolan, T., and Macaulay, L., "Don't Give up the Community: A Viewpoint of Trust Development in Online Collaboration", Information Technology & People, Emerald, 26(3), 2013, pp. 298-318.

[2] Cheng, X., Li, Y., Sun, J., & Zhu, X., "Easy Collaboration Process Support System Design for Student Collaborative Group Work: A Case Study", Proceedings of 47th the Hawaii International Conference on System Sciences (HICSS), IEEE, 2014, pp. 453-462.

[3] Xu, Y., Hao, J., Lau, R.Y.K., Ma, J., Xu, W., and Zhao, D., "A Personalized Researcher Recommendation Approach in Academic Contexts: Combining Social Networks and Semantic Concepts Analysis", Proceedings of 2010 Pacific Asia Conference on Information Systems, AIS, 2010, pp. 1516-1527.

[4] Xiao, B., and Benbasat, I., "E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact", MIS Quarterly, Management Information Systems Research Center, 31(1), 2007, pp. 137-209.

[5] Liu, X., Bollen, J., Nelson, M.L., and Van De Sompel, H., "Co-Authorship Networks in the Digital Library Research Community", Information Processing & Management, Elsevier, 41(6), 2005, pp. 1462-1480.

[6] Liben-Nowell, D., and Kleinberg, J., "The Link-Prediction Problem for Social Networks", Journal of the American Society for Information Science and Technology, Wiley, 58(7), 2007, pp. 1019-1031.

[7] Yang, C., Ma, J., Silva, T., Liu, X., and Hua, Z., "A Multilevel Information Mining Approach for Expert Recommendation in Online Scientific Communities", The Computer Journal, Oxford, 2014, pp. bxu033.

[8] Sun, Y., Barber, R., Gupta, M., Aggarwal, C.C., and Han, J., "Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks", International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2011, pp. 121-128.

[9] Lee, D.H., Brusilovsky, P., and Schleyer, T., "Recommending Collaborators Using Social Features and Mesh Terms", Proceedings of the American Society for Information Science and Technology, Wiley, 48(1), 2011, pp. 1-10.

[10] Han, S., He, D., Brusilovsky, P., and Yue, Z., "Coauthor Prediction for Junior Researchers": Social Computing, Behavioral-Cultural Modeling and Prediction, Springer, 2013, pp. 274-283.

[11] Spaeth, A., and Desmarais, M.C., "Combining Collaborative Filtering and Text Similarity for Expert Profile Recommendations in Social Websites": User Modeling, Adaptation, and Personalization, Springer, 2013, pp. 178-189.

[12] Backstrom, L., and Leskovec, J., "Supervised Random Walks: Predicting and Recommending Links in Social Networks", Proceedings of the fourth ACM international conference on Web search and data mining, ACM, 2011, pp. 635-644.

[13] Cohen, S., and Ebel, L., "Recommending Collaborators Using Keywords", Proceedings of the 22nd international conference on World Wide Web companion, ACM, 2013, pp. 959-962.

[14] Mishne, G., Carmel, D., and Lempel, R., "Blocking Blog Spam with Language Model Disagreement", AIRWeb'05: Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web, ACM, 2005, pp. 1-6.

[15] Lau, R.Y., Lai, C., and Li, Y., "Leveraging the Web Context for Context-Sensitive Opinion Mining", 2nd IEEE International Conference on Computer Science and Information Technology, IEEE, 2009, pp. 467-471.

[16] Benczúr, A.A., Bíró, I., Csalogány, K., and Uher, M., "Detecting Nepotistic Links by Language Model Disagreement", 15th international conference on World Wide Web, ACM, 2006, pp. 939-940.

[17] Zhai, C., and Lafferty, J., "A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval", Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2001, pp. 334-342.

[18] Robertson, S., Zaragoza, H., and Taylor, M., "Simple Bm25 Extension to Multiple Weighted Fields", Proceedings of the thirteenth ACM international conference on Information and knowledge management, ACM, 2004, pp. 42-49.

[19] Balog, K., Azzopardi, L., and De Rijke, M., "Formal Models for Expert Finding in Enterprise Corpora", Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2006, pp. 43-50.

[20] Nie, J.-Y., Cao, G., and Bai, J., "Inferential Language Models for Information Retrieval", ACM Transactions on Asian Language Information Processing (TALIP), ACM, 5(4), 2006, pp. 296-322.

[21] Arazy, O., Kumar, N., and Shapira, B., "Improving Social Recommender Systems", IT professional, IEEE, 11(4), 2009, pp. 38-44.

[22] Ding, Y., "Scientific Collaboration and Endorsement: Network Analysis of Coauthorship and Citation Networks", Journal of Informetrics, Elsevier, 5(1), 2011, pp. 187-203.

[23] Mcdonald, D.W., "Recommending Collaboration with Social Networks: A Comparative Evaluation", Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, 2003, pp. 593-600.

[24] Adamic, L.A., and Adar, E., "Friends and Neighbors on the Web", Social networks, Elsevier, 25(3), 2003, pp. 211-230.

[25] Yan, E., and Guns, R., "Predicting and Recommending Collaborations: An Author-, Institution-, and Country-Level Analysis", Journal of Informetrics, Elsevier, 8(2), 2014, pp. 295-309.

[26] Levin, D.Z., and Cross, R., "The Strength of Weak Ties You Can Trust: The Mediating Role of Trust in Effective Knowledge Transfer", Management Science, INFORMS, 50(11), 2004, pp. 1477-1490.

[27] Zhong, J.A., and Li, X., "Unified Collaborative Filtering Model Based on Combination of Latent Features", Expert Systems with Applications, Elsevier, 37(8), 2010, pp. 5666-5672.

[28] Joachims, T., "Optimizing Search Engines Using Clickthrough Data", Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2002, pp. 133-142.

[29] Pera, M.S., and Ng, Y.-K., "Exploiting the Wisdom of Social Connections to Make Personalized Recommendations on Scholarly Articles", Journal of Intelligent Information Systems, Springer, 2013, pp. 1-21.

[30] Jain, A., Nandakumar, K., and Ross, A., "Score Normalization in Multimodal Biometric Systems", Pattern Recognition, Elsevier, 38(12), 2005, pp. 2270-2285.

[31] Joachims, T., "Training Linear Svms in Linear Time", Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, ACM, pp. 217-226.