A WEIGHTED TOPIC MODEL ENHANCED APPROACH FOR COMPLEMENTARY COLLABORATOR RECOMMENDATION

Chen Yang, School of Management, University of Science and Technology of China, Hefei, Anhui, PR China; Department of Information Systems, City University of Hong Kong, Kowloon, Hong Kong, PR China, yangc0201@gmail.com

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

Jianshan Sun, Department of Information Systems, City University of Hong Kong, Kowloon, Hong Kong, PR China, sunjs9413@gmail.com

Thushari Silva, Department of Information Systems, City University of Hong Kong, Kowloon, Hong Kong, PR China, thushi.silva@my.cityu.edu.hk

Xiaoyan Liu, Advanced Analytics Institute, University of Technology, Sydney, Australia, xiaoyan.liu@uts.edu.au

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

Abstract

Collaborations among interdisciplinary scientists are playing an increasingly important role in science innovations. As it is very difficult for a researcher to master the full knowledge of his/her targeted research areas, how to find suitable collaborators of complementary expertise has turned to be a key factor for researchers to succeed. With the expansion of the Web, the availability of sheer volume of information has resulted in information overload issue and posed significant challenges on determining appropriate scientists to collaborate with effectively for research opportunities. However, current studies on collaborator recommendation ignored this phenomenon and particularly overlooked the complementarity of their expertise within a restrictive context, i.e. for a given funding proposal or a research manuscript draft. In this study we propose a complementary expertise analysis enhanced approach to retrieval experts for research collaboration. It produces recommendation list using a heuristic greedy algorithm based on probabilistic topic model, and generates experts who ought to be complemented in expertise as well as to have good ability. The proposed method has been implemented in ScholarMate research community (www.scholarmate.com). We have conducted a user study to verify the effectiveness of the proposed approach and the preliminary results show its good performance comparing to the benchmarks.

Keywords: recommendation systems, researcher profiling, topic models, complementarity, latent dirichlet allocation
1 INTRODUCTION

In the era of Web 2.0, many traditional patterns of research activities have been dramatically transformed. Under this trend, finding an academic collaborator, which is a common activity among scientists, has changed its conventional forms to a new problem similar to expert retrieval through various online platforms (Fazel-Zarandi et al. 2011; Dráždílová et al. 2012). It is more and more difficult for a researcher to master all the demanded areas of a funding application proposal or a research paper. As a consequence finding suitable collaborators is very important for the researchers to win funding supports or publish good research outputs. Previous studies usually concentrate on how to seek collaborators who have a similar and relevant expertise or have a close social proximity to the target user while ignoring the complementarity of their expertise within a restrictive context.

There are a lot of researches studied the collaborator recommendation in online social network settings (Lee et al. 2011; Chen et al. 2012; Cohen & Ebel 2013). Most of these studies recommend collaborators without considering the matching of complementary expertise to the related contexts. However, in real world researchers often seek collaborators who are capable of complementary expertise in a specific context. For instance, in a R&D funding application scenario, funding agencies often launch cross-disciplinary tasks for applicants, thus it is a mainstream trend for researchers to seek for complementary collaborators out of their own laboratories. More and more interdisciplinary collaborations have emerged and observably influenced the patterns of knowledge sharing (Dodson et al. 2010). As a consequence, researchers need to find people who are suitable to collaborate with regarding to the restrictive context: the funding agency’s call for proposal which lists specific expertise requirements. Many studies on bibliometrics indicate that the complementarity of expertise serves as an important factor for researchers to establish connections and collaborate (Hara et al. 2003). We can easily find some other scenarios, e.g., a researcher of information retrieval domain comes up with a manuscript draft and wants to find some collaborators who are skilled in machine learning area. Although there are great prospects and potentials to recommend collaborators with the context constrained, this problem has yet been studied.

The information overload problem and terminology difference issue are the two common challenges to find collaborators in a new area. The information overload problem – the availability of sheer volume of information through the expanding of the Web – has hindered the activity of finding a suitable expert. It can be solved by leveraging information retrieval algorithms. Topic modeling, which is widely used for the expertise modeling and document clustering area, can be effectively adapted to address the terminology normalization problem encountered by interdisciplinary researchers. In addition, seeking for collaborators who are capable of complementary expertise within a restrictive context is a new trend and put forward more serious challenges.

Hence in this paper, we will propose a novel context-aware method to address this issue: a heuristic greedy algorithm based on weighted probabilistic topic model is proposed to recommend experts who ought to be complemented in expertise regard to the specific context as well as to have good ability. Previous topic models usually learn the model with the assumption that each document or profile holds the same importance, while the quality of the one’s expertise is ignored. In our proposed weighted topic model, we incorporate an expertise quality parameter as a prior probability to express the importance of the experts’ quality. The preliminary experiment results show the effectiveness of the quality prior enhanced model over traditional topic model such as Latent Dirichlet Allocation (LDA).

The main contributions of this research are threefold. First, we build a novel weighted generate topic model by incorporating a quality prior into the traditional LDA model and obtain the experts expertise distribution over subtopics with regarding to both the relevance and the quality of experts. Second, we provide an effective solution to address the problem of complementary collaborator recommendation within a theme-specific context, and the greedy strategy of our approach ensured the optimal complementarity of research expertise. Third, our proposed complementary collaborator
recommendation mechanism has been implemented into an online research community, and the conducted user study shows its strength to obtain the best experts to collaborate with. Meanwhile the proposed approach can be extended to solve some similar problems such as expert retrieval and team formation.

The rest of this paper is as follows: in Section 2, we discuss related work of expert retrieval and collaborator recommendation, including semantic content based methods and social proximity enhanced methods. In Section 3, we propose a complementary collaborator recommendation method in the restrictive context. We introduce the recommender system which has been implemented in an online platform and the preliminary experimental results in Section 4. We conclude our research and outlook the future directions in Section 5.

2 RELATED WORK

The collaborator finding task is a well-studied research problem, in which the expert’s expertise is usually characterized by published papers in the pre-collected document set (McDonald & Ackerman 2000). Many studies on collaborator finding leverage various features obtained from researchers’ profiles and their social networks. For instance, Li et al. proposed a new approach which takes a comprehensive consideration of the aspects of the expertise sharing such as semantic similarity between experts’ profiles, social proximity and some social network-based models for seeking experts for collaboration in online communities (Li et al. 2012). A hybrid method which combines content-based and social proximity features is proposed to recommend academic collaborators in biomedical domain (Lee et al. 2011). Some researchers built new relation strength similarity measure on the collaborator social network and verified its effectiveness (Chen et al. 2011). Cohen and Ebel investigated the collaborator recommendation problem from a comprehensive perspective with structural proximity, relevance and importance features (Cohen & Ebel 2013). Xu et al. employed several conceptual features and social network analysis measures to make up a unified framework for collaborator recommendation (Xu et al. 2010; Xu et al. 2012).

Existing collaborators finding researches concentrate on search experts for a given topic by incorporating various semantic features, while the quality of the experts is less considered. Some researchers have taken actions in this direction. Deng et al. introduced the document citation weight into the language model and showed its strength over traditional ones (Deng et al. 2008). Recently, Duan et al. incorporated two kinds of link based importance factors into pLSI topic model (Duan et al. 2014). This suggests the potential advantages of incorporating expert’s quality into expertise profiling based on topic models.

In particular, there are few papers recommending collaborators not only for a given topic, but also considering the complementarity of their expertise in this context. The most related works lie in some reviewer assignment researches which take advantage of both the expertise relevance and the expertise coverage. Karimzadehgan et al. designed three methods to assign reviewers based on multiple subtopics matching and ensured the assigned reviewers to cover the subtopics of the proposal well (Karimzadehgan et al. 2008). An improved version of their study added more constraints and casted the problem as a optimization issue (Karimzadehgan & Zhai 2009). Tang et al. have designed a collaborator recommendation system for interdisciplinary collaborations with cross-domain topic learning model (Tang et al. 2012).

However, most of the previous studies obtain candidate collaborators from an overall view and have neither considered the context background nor the complementarity of their expertise of potential collaborators. In this paper, we will explore the complementary collaborator recommendation within a restrictive context.
3 PROPOSED APPROACH

In this paper, we will study how to perform academic collaborators seeking tasks from a new perspective: complementary expertise matching within a restrictive context. A novel expert quality weighted topic model is presented and final recommendation is generated through a heuristic greedy algorithm based on the SKL divergence over the proposed topic model. Figure 1 indicates the behind mechanism of our recommender system. We will illustrate the details through three subsections respectively: the expert quality identification module, the weighted topic model generation module and the greedy seeking strategy module. We first interpret how to measure the expert’s quality through the online communities, and then we illustrate the weighted topic model based on LDA and adapt it into a greedy algorithm to recommend collaborators within a constrained context.

Figure 1. The conceptual framework of the collaborator recommender system

3.1 Expert Quality Identification

Expert quality measures the research performance of experts, and it serves as an important factor to recommend collaborators with high level expertise. Specifically, we measure the research quality of a potential collaborator in terms of the number of research papers, quality of the papers, citation impact in the past five years and the academic titles.

Sun et al. pointed that the journal rank could reflect the quality of the articles published in that journal because it is a popular measure which has been adapted in many research performance identification activities for the merit increasing and the allocation of research funding in university settings (Turban et al. 2004; Sun et al. 2008). Following Sun et al., we employ a liner weighted function to generate the research quality as a measure of overall contribution of a researcher to the domain (Sun et al. 2008).

Let \( q_i \) be researcher \( j \)'s total number of publications in rank \( i \) level journals, where \( i = A, B, C \). The publication score of researcher \( j \) can be represented as:

\[
\text{pub\_score}_j = w_A \cdot q_{A_j} + w_B \cdot q_{B_j} + w_C \cdot q_{C_j}
\] (1)
where \( w_h > w_p > w_e \), indicating the emphasis on high quality works. There are different ways to define the weights. In this paper, we utilize the average impact factors for all the journals classified at the same level to define the corresponding weight.

Professional titles (e.g. professor, associate professor and assistant professor) and H-index can also be taken into consideration for expert quality identification. Let \( \text{acad}_j \) and \( H_j \) be potential expert \( j \)'s rank score and H-index, respectively. An integrated research quality measure can be obtained as follows:

\[
e_j = u \cdot \text{pub} \_\text{score}_j + s \cdot \text{acad}_j + t \cdot H_j
\]

where \( u + s + t = 1 \), and the three corresponding weights are determined by the experts in related fields.

### 3.2 The Weighted Topic Model Based on LDA

In this subsection, we will briefly introduce the Latent Dirichlet Allocation (LDA) model, and detail how we leverage the expert quality weight to improve the basic LDA model, hence we can use the weighted LDA model to characterize the expert and the restrictive context.

#### 3.2.1 The Weighted LDA model for expertise identification

The LDA model, known as a document generate model, is widely used in many related studies (Blei et al. 2003; Wei & Croft 2006). It assumes that each document has a multinomial distribution over a set of topics, and each topic has a multinomial distribution over the full vocabulary. In regard to the situation in this paper, we cast each expert’s publications as an independent document. Thus we can build the experts’ profiles in forms of a distribution on the learned topics.

LDA model has shown its strength in semantic representing on multiple topics, and it has applications in many domains. However, it does not mean that LDA is appropriate to any cases. For the experts’ profile corpus, LDA assumes that each profile holds the same importance to the generated topic distributions as well as word distributions. Intuitively, the expert’s quality will affect the probability weight of each profile. Deng et al. studied the expertise modeling with language model and topic model. They incorporated the document citation weight into the language model and demonstrated its advantages (Deng et al. 2008). Analogously, we can infer that it is necessary to introduce the expert quality into the LDA model.

We can assume that there is a corpus with \( N \) profiles, \( T \) topics and \( W \) unique words. We can express the weighted LDA model with the graphical model showed in Figure 2. The two shaded nodes denote the observed word \( w \) and the observed expert quality \( q \) of profile \( p \). \( \alpha \) and \( \beta \) are the hyper-parameters of the Dirichlet priors on \( \theta \) (topics distribution of each profile) and \( \phi \) (words distribution of each topic).

![Graphical model of the proposed quality weighted LDA model](image)
The expert quality weight which represents the expert’s importance to the topic learning process can be derived from the following equation, where \( e_p \) denotes the expert quality of profile \( p \) as introduced in Section 3.

\[
q_p = \frac{N * e_p}{\sum_{p \in N} e_p}
\]

(3)

### 3.2.2 Weighted LDA Model Building and Parameter Estimation

The Gibbs sampling described in (Griffiths & Steyvers 2004) is adapted to obtain the approximation of \( \theta \) and \( \phi \). Following with this approach, we leverage a Markov chain Monte Carlo method to estimate the parameters, and the probability of \( z_i = j \) can be computed as follows,

\[
P(z_i = j | z_{-i}, w, \theta) \propto \sum_{p \in N} q_p * n_{i,j,p} + \beta \frac{q_p * n_{i,j,p} + \alpha}{\sum_{p \in N} q_p * n_{i,j,p} + W \cdot \beta} q_p * n_{i,j,p} + T \cdot \alpha
\]

(4)

In the above equation, \( n_{i,j,p} \) indicates a count that excludes the current assignment, which records the assign times of word \( w_i \) to topic \( j \) in profile \( p \). The hyper parameters \( \alpha \) and \( \beta \) are set by the users in regard to the application field. The approximation of \( \theta \) and \( \phi \) can be gained with the formulas below:

\[
\hat{\phi}_j^{(w)} = \frac{\sum_{p \in N} q_p * n_{i,j,p}^{(w)} + \beta}{\sum_{p \in N} q_p * n_{i,j,p} + W \cdot \beta}
\]

(5)

\[
\hat{\phi}_j^{(p)} = \frac{q_p * n_{j,p}^{(p)} + \alpha}{q_p * n_{j,p} + T \cdot \alpha}
\]

(6)

### 3.3 A greedy algorithm to seek collaborators with complementary expertise in a restrictive context

In this subsection we illustrate a greedy strategy to optimize the generated collaborators list for both expertise matching and expertise coverage within the given context. In other words, it is a general method to pursue the expertise complementarity as well as the expertise quality in a particular context. After the inference of the weighted LDA model on experts’ profiles corpus, we can generate a new representation of expertise over a set of topics. In the same way, we can project a given context on the learned topics set via the word distribution over topics. We denote the topic distribution of the restrictive context as \( \theta_C \). We use the symmetric Kullback-Leibler (SKL) divergence to calculate the distance between the context and candidate collaborators, the SKL divergence is detailed in (Rosen-Zvi et al. 2004). We utilize a similar strategy to match the multiple subtopics under a restrictive background by (Karimzadehgan et al. 2008).

\[
D_{SKL}(\theta_C \mid \theta_{u,p_1, \ldots, p_k}) = \sum_t (\theta_t \ln \frac{\theta_t}{\theta_{u,p_1, \ldots, p_k}} + \theta_{u,p_1, \ldots, p_k} \ln \frac{\theta_{u,p_1, \ldots, p_k}}{\theta_t})
\]

(7)

\( \theta_{u,p_1, \ldots, p_k} \) is a quality weighted average of the topic distributions of the target user and a list of his/her \( k \) candidate collaborators, and it can be obtained by the equation below, where \( e_p \) indicates the quality of expert \( p \).
\[ P(t \mid \theta_{u,p_1}, \ldots, p_k) = \frac{1}{k+1} \left\{ \sum_{p=1}^{k} \left[ \frac{e_p}{e_u + \sum_{p=1}^{k} e_p} \right] + \frac{e_u}{e_u + \sum_{p=1}^{k} e_p} \right\} \]  

(8)

We gradually select a candidate collaborator within the context which holds the minimal SKL divergence with a greedy strategy. The next round of calculation will take use of the profiles of previous k-1 generated collaborators as well as the target user. Hence we could obtain the fusing result of \( \theta_{u,p_1, \ldots, p_k} \). It is said the final recommended collaborators list will cover the context with complementary expertise and ensure they have good ability in the domain.

4 EVALUATION

4.1 System Implementation

ScholarMate (http://www.scholarmate.com) is a research social network which connects people with common interests and skills to research smarter. It automatically integrates scientific information from multiple databases including CNKI, ISI and Scopus, verifies research outputs and intelligently resolves authors’ ambiguity and effectively identifies research outputs of the experts. In addition to its important function of connecting people with similar interests, ScholarMate generates visualized research CV for various purposes based on the method discuss above. The following left-hand screenshot shows this function (see Figure 3). Leverage the research CV, expertise areas and expert quality can be identified by the objective information such as professional titles and bibliometrics indices such as H-Index, number of publications, SCI/SSCI citations as shown in the figure. The right column of Figure 3 reflects the interface of the complementary collaborator recommender system which has been implemented in ScholarMate platform. Researchers can input a short abstract, with which the system will automatically recommend collaborators in this restrictive context and generate the candidate collaborator list in the lower part.

Figure 3. Screenshots of the generate Visual Research CV and the recommendation system

4.2 Preliminary Experiment

We have conducted a user study to verify the effectiveness of our proposed recommender mechanism. Previous collaborator recommendation studies usually recommend collaborators by leveraging various content based features to obtain most relevant experts in expertise. Thus we compared our weighed LDA (WLDA) enhanced complementary expertise matching approach with two content based
approaches: BM25 method (Macdonald & Ounis 2006) and LDA based method (Kongthon et al. 2009). Due to the disturbance deviation to evaluate subjects in a small group, we performed a subjective user study that each user will face with the integrated results from three compared approaches. We invited 8 subjects majored in information systems to participate in our user study, and each of the subjects was required to provide one abstract from his/her latest research manuscripts as the restrictive context. Meanwhile we built a candidate experts set, which contains 500 researchers and their publications from information systems domain. We generated three lists of collaborators based on the subjects’ publications and the given context. Then all the results are mixed randomly. The participants were asked to score each candidate collaborators in the mixed results from 1-5. The high score suggests good recommendation.

We employed the Average Rating score and Normalized Discounted Cumulative Gain (NDCG) as evaluation metrics and calculated top 5 and top 10 lists respectively (Liu et al. 2009; Derhami et al. 2013). The experiment results are listed as follows:

<table>
<thead>
<tr>
<th></th>
<th>AR</th>
<th>NDCG</th>
<th></th>
<th>AR</th>
<th>NDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>3.23</td>
<td>0.94</td>
<td>BM25</td>
<td>3.14</td>
<td>0.91</td>
</tr>
<tr>
<td>LDA</td>
<td>3.53</td>
<td>0.96</td>
<td>LDA</td>
<td>3.25</td>
<td>0.96</td>
</tr>
<tr>
<td>WLDA</td>
<td>4.10</td>
<td>0.97</td>
<td>WLDA</td>
<td>3.68</td>
<td>0.98</td>
</tr>
</tbody>
</table>

*Table 1. Experimental results of the proposed method and two benchmarks*

We can obtain from the result that the AR score of the proposed method has 16% and 13% improvement over the best benchmark in the two intervals respectively. The results indicate that our proposed complementary expertise matching method based on WLDA models exceeds the two content based methods in both measures, as it achieves the highest AR and NDCG scores. We can also conclude from the table that topic model based approaches outperform information retrieval methods, because BM25 gets the worst performance as shown in the table.

5 CONCLUSIONS AND FUTURE WORK

In this paper, a quality weighted LDA model is proposed to generate experts profiles, based on which we further apply a greedy method to recommend collaborators who are complemented in expertise as well as to have good ability in a restrictive context. The recommender system has been implemented in ScholarMate. The preliminary result of the conducted user study demonstrates the strength of our proposed approach and it shows a significant advantage over benchmarks.

The main contributions of our research are threefold. First, we build a novel quality weighted LDA model to obtain the experts expertise distribution over subtopics. Second, we effectively fill in the gap of finding suitable collaborators within a theme-specific context, as the greedy strategy of our approach guarantees the coverage of research expertise under a particular background. Third, our proposed greedy approach has been implemented in a real platform and showed its strength to recommend suitable collaborators in the user study. In the future we will extend this approach to solve some similar problems such as expert retrieval and research team formation. We will build a large data set to conduct offline experiments to test the validity, reliability and precision of our proposed approach. Meanwhile, we would like to expand the sample in the user study to measure the statistical improvement of our approach over benchmarks with considerable amount of data.

Acknowledgments

This work was supported by the General Research Fund of Hong Kong Research Grant Council (CityU 119611 and CityU 148012), City University of Hong Kong Teaching Development Grant (6000201) and National Natural Science Foundation of China (71090401/71090400, 71171172).
References


Liu, Y., T. Mei and X.-S. Hua (2009). CrowdReranking: exploring multiple search engines for visual search reranking. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, Boston, Massachusetts, USA, ACM.


