Cold-Start Expert Finding in Community Question Answering via Graph Regularization

Zhou Zhao[1**], Furu Wei[2], Ming Zhou[2], and Wilfred Ng[1]

[1] The Hong Kong University of Science and Technology,
[1]{zhaozhou,wilfred}@cse.ust.hk
[2] Microsoft Research, Beijing, China,
[2]{fuwei,mingzhou}@microsoft.com

Abstract. Expert finding for question answering is a challenging problem in Community-based Question Answering (CQA) systems such as Quora. The success of expert finding is important to many real applications such as question routing and identification of best answers. Currently, many approaches of expert findings rely heavily on the past question-answering activities of the users in order to build user models. However, the past question-answering activities of most users in real CQA systems are rather limited. We call the users who have only answered a small number of questions the cold-start users. Using the existing approaches, we find that it is difficult to address the cold-start issue in finding the experts.

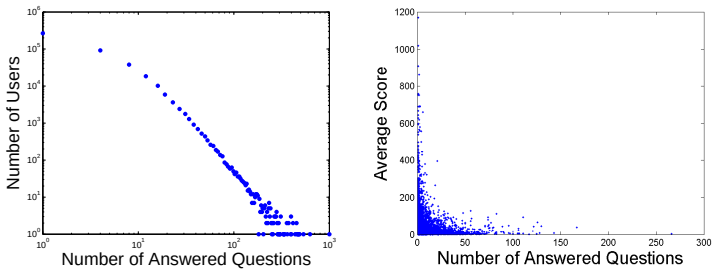
In this paper, we formulate a new problem of cold-start expert finding in CQA systems. We first utilize the "following relations" between the users and topical interests to build the user-to-user graph in CQA systems. Next, we propose the *Graph Regularized Latent Model* (GRLM) to infer the expertise of users based on both past question-answering activities and an inferred user-to-user graph. We then devise an iterative variational method for inferring the GRLM model. We evaluate our method on a well-known question-answering system called Quora. Our empirical study shows encouraging results of the proposed algorithm in comparison to the state-of-the-art expert finding algorithms.

1 Introduction

Expert finding is an essential problem in CQA systems [4, 25], which arises in many applications such as question routing [28] and identification of best answers [2]. The existing approaches [39, 27, 34, 2, 28] build a user model from their past question-answering activities, and then use the model to find the right experts for answering the questions. However, the past question-answering activities of most users in real CQA systems are rather limited. We call the users who have only answered a small number of questions the *cold-start users*. The existing approaches work well if the users have sufficient question-answering activities, while they may not provide satisfactory results for the cold-start users.

In fact, a vast majority of existing users in real CQA systems, including many that have joined the system for a relatively long period of time, do not have sufficient activities. To illustrate this fact, we summarize the question-answering activities of the

** The work was done when the first author was visiting Microsoft Research, Beijing, China.



(a) Cold-Start Property (b) Performance

Fig. 1. Cold-Start Users in Quora

users in Quora in Figure 1(a). From the figure, we can see that the participation of most users in question-answering activities falls into the *long tail* part of the power-law curve. This indicates that the majority of the users only answered very few questions. Thus, it is difficult to build an effective user model for cold-start users by using existing methods. Let us call the problem of expert finding with the presence of many cold-start users in a CQA system *the cold-start expert finding problem*. Interestingly, we observe that CQA systems enjoy great benefits contributed by the cold-start users. To show this point, we summarize the performance of the users in Quora in Figure 1(b). Consider the thumbs-up/downs voted by the community as quality score for users on answering questions [28]. We can see that a significant number of cold-start users obtain high quality scores.

To address the cold-start expert finding problem, we incorporate the user-to-user graph in CQA systems to build a regularized user model. Currently, CQA systems such as Quora define thousands of topical interests, which are represented by keywords such as "startups" and "computer programming". The users may follow these keywords when they have the topical interests. If two users follow some common topical interests, we consider that there is a user-to-user relation (i.e. an edge) between them. The works [13, 17] show that a user-to-user relation between two users provides a strong evidence for them to have common interests or preferences. Thus, we attempt to integrate both user-to-user graph and question-answering activities into a seamless framework that tackles the cold-start expert finding problem.

The main contributions of our work are summarized as follows:

- We illustrate that the question-answering activities of most users in real CQA systems are rather few and formally propose a new problem of cold-start expert finding in CQA systems.
- We explore the "following relations" between users and topical interests to build the user-to-user graph. We then propose the graph regularized latent model by incorporating with the user-to-user graph and devise a variational method for inferring the model.
- We conduct extensive experiments on our proposed method. We demonstrate that, by incorporating with user-to-user graph, our method significantly outperforms other state-of-the-art expert finding techniques.

Fig. 2. An Illustration of User's Topical Interests

There exists some work addressing the cold-start problem in user-item recommendation systems [20, 13, 21, 14, 36, 29]. However, most of them are not applicable in addressing the problem of cold-start expert finding in CQA systems. Even though finding an expert for a question seems to be analogous to recommending an item to a user, there are some subtle differences between them. First, the existing work incorporates with the social relations of users to improve the performance of recommending an item to a user. In the context of our work, there is no relation between the questions and thus the existing cold-start recommendation techniques cannot be applied to our problem. Second, the goal of expert finding is fundamental different from that of recommendation. The existing recommendation techniques focus on recommending existing items to the users while expert finding aims to select the right users to answer some new questions.

The rest of the paper is organized as follows. Section 2 introduces some notations and formulates the problem. We then propose a graph regularized latent model for cold-start expert finding in Section 3. We report the experimental results in Section 4. Section 5 surveys the related work. We conclude the paper in Section 6.

2 Background

In this section, we first introduce some notation of community-based question answering used in our subsequent discussion. The notation includes a data matrix of questions $\mathbf{Q}$, a data matrix of users $\mathbf{U}$, a question-answering activity set $\mathbf{\Omega}$ and an observed quality score matrix $\mathbf{S}$. Then, we formulate the problem of cold-start expert finding. The summary of the notation is given in Table 1.

We represent the feature of questions by *bag of words*, which has been shown to be successful in many question answering applications [5, 35, 37]. Therefore, the feature of each question $\mathbf{q}_i$ is denoted by d-dimensional word vector. We then denote the collection of questions by $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_m] \in R^{d \times m}$ where m is the total number of the questions.

We denote by $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n] \in R^{d \times n}$ the collection of users in CQA systems, where n is the total number of the users. The parameter $\mathbf{u}_j$ represents a d-dimensional vector for modeling the j-th user. The terms in $\mathbf{u}_j$ indicate the strengths and weakness of the j-th user on the latent space of the questions.

Table 1. Summary of Notation

Group	Notation	Notation Description
Data	$\mathbf{Q}$	data matrix of questions
	$\mathbf{U}$	a data matrix of users
	$\mathbf{S}$	an observed quality score matrix
	$\mathbf{F}$	a set of topical interests
	$\mathbf{W}$	a similarity matrix of users
	$\mathbf{\Omega}$	a set of existing question-answering activities
	$\mathbf{I}_{\Omega}$	an indicator matrix for existing activities
	$\mathbf{\Theta}$	a topic matrix of questions
	$\mathbf{Z}$	a topic assignment matrix of words
Model	$Mult(\cdot)$	a multinomial distribution
	$Dir(\cdot)$	a dirichlet distribution
	$Norm_{\delta}(\cdot)$	a normal distribution with standard deviation δ
	λ	a graph regularization term
	K	a dimension of latent space

We denote by score matrix $\mathbf{S} \in R^{m \times n}$ the quality of all users on answering the questions. The thumb-ups/downs value in $\mathbf{S}$ is voted by the users in a CQA community. The voting result indicates the community's long term view for the quality of users on answering the questions. Let $\mathbf{\Omega}$ be the set of existing question-answering activities of users. The quality score S_{ij} exists in matrix $\mathbf{S}$ if activity $(i, j) \in \mathbf{\Omega}$.

We observe that many users in CQA systems follow some topical interests. Figure 2 shows the set of topical interests followed by a Quora user. In this example, the user Adam (one of Quora co-founders) follows four topical interests, which are "startups" , "google" , "computer programming" and "major Internet companies". Let $\mathbf{F}_i$ be the set of topical interests followed by the i-th user. Consider the topical interests of the i-th user and the j-th user, $\mathbf{F}_i$ and $\mathbf{F}_j$. We use the *Jaccard Distance* to model the similarity between them, which is denoted by $W_{ij} = \frac{|\mathbf{F}_i \bigcap \mathbf{F}_j|}{|\mathbf{F}_i \bigcup \mathbf{F}_j|}$. The $\mathbf{F}_i \bigcap \mathbf{F}_j$ is the set of two users' common following topical interests and $\mathbf{F}_i \bigcup \mathbf{F}_j$ is the set of two users' total following topical interests. We note that the similarity value in matrix $\mathbf{W}$ is within the range [0, 1]. We therefore model the user-to-user graph based on the similarity between users by $\mathbf{W} \in R^{n \times n}$.

Using the notation given in Table 1, we now define the problem of cold-start expert finding with respect to a CQA system as follows.

Consider a data matrix of questions $\mathbf{Q}$, a quality score matrix $\mathbf{S}$ and a similarity matrix of users $\mathbf{W}$. Given a new question $\mathbf{q}$, we aim to choose the users with high predicted quality score for answering the question.

3 Cold-Start Expert Finding Algorithm

In this section, we present our algorithm for tackling the problem of cold-start expert finding in CQA systems. We first introduce the basic latent model, which has been widely used for addressing the problem of expert finding in [39, 27, 34, 28]. Next, we

propose our graph regularized latent model (GRLM). The graphical representation of GRLM is illustrated in Figure 3. We then devise a variational method for solving the optimization problem in GRLM. Finally, we present the expert finding algorithm based on GRLM.

3.1 Basic Latent Model

The basic latent model tackles the problem of expert finding based on the past question-answering activities and quality score matrix. The latent topic model is first utilized to extract the feature of the questions. Then the user model is inferred from question features and a quality score matrix. The main procedure of basic latent model can be summarized as follows:

Question Feature Extraction. The topic modelling technique [23] has been widely used for question feature extraction in many recent work concerning the problem of expert finding [39, 27, 34, 28]. Topic models provide an interpretable low-dimensional representation of the questions. In this work, we employ the famous latent dirichlet allocation model (LDA) [1] to extract the feature of the questions, which has been shown to be successful in [39, 34, 28]. The graphical representation of the LDA model is illustrated in the left box in Figure 3. Given K topics, the generative process of LDA is given as follows:

For each question $\mathbf{q}_i$:

- Draw topic proportions $\theta_i \sim Dir(\alpha)$
  - For the j-th word in $\mathbf{q}_i$
    - Draw a topic assignment of j-th word $z_{ij} \sim Mult(\theta_i)$
    - Draw a word $q_{ij} \sim Mult(\beta_{z_{ij}})$

Therefore, the latent topic proportion θ_i is inferred for the feature of the i-th question. We denote the feature of the existing questions in CQA systems by $\mathbf{\Theta}$.

User Model Inference. Given latent feature of questions $\mathbf{\Theta}$ and quality score matrix $\mathbf{S}$, we infer the latent feature of users $\mathbf{U}$. We assume that the quality score matrix $\mathbf{S}$ is generated by

$$\mathbf{S} \sim Norm_{\lambda_S^{-1}}(\mathbf{\Theta}^T\mathbf{U}) \tag{1}$$

where $Norm(\cdot)$ is a normal distribution with mean $\mathbf{\Theta}^T\mathbf{U}$ and standard deviation λ_S^{-1}. The graphical representation of this quality score generative model is illustrated in the upper box in Figure 3. For each question-answering activity (i, j), its quality score is generated by

$$S_{ij} \sim Norm_{\lambda_S^{-1}}(\mathbf{q}_i^T\mathbf{u}_j) = Norm_{\lambda_S^{-1}}(\sum_{k=1}^{K} q_{ik}u_{jk}). \tag{2}$$

The underlying idea of the quality score generative model is as follows. The quality score value is proportional to the dot-product of the question feature and user feature.

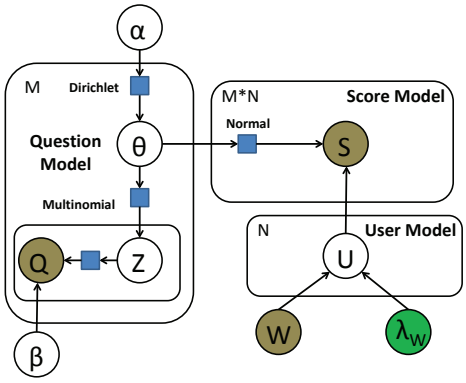


Fig. 3. Graphical Representation of the GRLM Model

We consider that the feature $\mathbf{U}$ represents the strongness and weakness of users on a specified topic.

Assume that the standard variance λ_S^{-1} is independent on and identical for different question-answering activities, we the problem of maximum likelihood inference for user feature $\mathbf{U}$ can be given by

$$\max_{\mathbf{U}} -||I_\Omega \otimes (\mathbf{S} - \mathbf{\Theta}^T\mathbf{U})||_F^2 \tag{3}$$

where $||\cdot||_F^2$ denotes the Frobenius norm, and $\otimes$ represents the Hadamard element-wise product. I_Ω is an indicator matrix with ones for the existing question-answering activities, and zeros for the missing ones.

Therefore, the user feature $\mathbf{U}$ can be inferred by solving the expression in Formula 3. Then, we can predict the quality score for new questions by Equation 1, and then choose those users who have high predicted scores for answering the questions.

Although a latent model is feasible for tackling the problem of expert finding, it may not be able to solve the cold-start problem well. In cold-start expert finding, there may be a number of users having only few question-answering activities. Under the framework of latent model, the inference for user feature may not be accurate, since there are many missing values in matrix $\mathbf{S}$. Thus, we propose to make use of a user-to-user graph to tackle the cold-start problem.

3.2 Graph Regularized Latent Model

In this section, we present our *Graph Regularized Latent Model* (referred to as GRLM) to tackle the problem of cold-start expert finding. First, we introduce the general idea of our model. Then, we present the detail of the generative process.

Consider the similarity matrix of users $\mathbf{W}$ which is inferred from the following relation between users and topical interests. Based on the property of the user-to-user relation, it is natural to require the similar users in matrix $\mathbf{W}$ have similarity user feature, that is, $W_{ij}(\mathbf{u}_i - \mathbf{u}_j)^2$. Thus, the generation process for data matrix of users $\mathbf{U}$

Algorithm 1 Generate Observed Question-Answering Activities

Input: a set of users, indicator matrix for existing activities I_Ω
Output: a data matrix of questions $\mathbf{Q}$, a quality score matrix $\mathbf{S}$

1: **for** each question $\mathbf{q}_i \in \mathbf{Q}$ **do**
2: Draw topic proportions $\theta_i \sim Dir(\alpha)$.
3: **for** each word $\mathbf{q}_{ij}$ **do**
4: (a) Draw a topic assignment $z_{ij} \sim Mult(\theta_i)$
5: (b) Draw a word $q_{ij} \sim Mult(\beta_{z_{ij}})$.
6: **end for**
7: **end for**
8: Draw a data matrix of users $\mathbf{U}$ by Equation 4.
9: Draw a quality score matrix $\mathbf{S}$ by Equation 1.

with graph regularization can be achieved by

$$p(\mathbf{U}) = -\sum_i \mathbf{u}_i^T \mathbf{u}_i - \lambda_W \sum_{i,j} W_{ij}(\mathbf{u}_i - \mathbf{u}_j)^2 \tag{4}$$

We denote by $\lambda_\mathbf{U}$ the collection of standard deviations for generating the data matrix of users $\mathbf{U}$. Thus, the prior distribution of the data matrix of users $\mathbf{U}$ is given by a product of normal distributions. Note that we set the standard deviation inversely proportional to the similarity of users with constant parameter λ_W. We illustrate the impact of parameter λ_W in the experimental study.

We denote a set of parameters α, β and λ_W as hyper parameters of our model. Referring to Figure 3, the whole generative procedure of our model is outlined in Algorithm 1. We then present the objective function for our graph regularized latent model below:

We observe that the joint distribution for generating quality score matrix $\mathbf{S}$, latent topics of the questions $\mathbf{\Theta}$, a data matrix of questions $\mathbf{Q}$ and a data matrix of users $\mathbf{U}$ can be factorized. Thus, we give the posterior distribution based on hyper parameters by

$$\begin{aligned} & p(\mathbf{S}, \mathbf{\Theta}, \mathbf{Q}, \mathbf{U}, \mathbf{Z}, \mathbf{Q} | \lambda_S^{-1}, \lambda_Q^{-1}, \lambda_U^{-1}, \alpha, \beta, \Omega) \\ = & \, p(\mathbf{\Theta}|\alpha) p(\mathbf{Z}|\Theta) p(\mathbf{Q}|\mathbf{Z}, \beta) \\ \times & \, p(\mathbf{U}|\lambda_\mathbf{U}^{-1}) p(\mathbf{S}|\mathbf{Q}^T\mathbf{U}, \Omega) \end{aligned} \tag{5}$$

where the generation for a data matrix of question is

$$\begin{aligned} p(\mathbf{\Theta}|\alpha) &= \prod_{\theta_i \in \mathbf{\Theta}} Dir(\alpha) \\ p(\mathbf{Z}|\mathbf{\Theta}) &= \prod_{z_{i,j} \in \mathbf{Z}} Mult(\theta_i) \\ p(\mathbf{Q}|\mathbf{Z}, \beta) &= \prod_{\mathbf{q}_i \in \mathbf{Q}} \beta_{\mathbf{z}_i, \mathbf{q}_i} \end{aligned}$$

and the generation for a data matrix of users and a quality score are

$$p(\mathbf{U}|\lambda_{\mathbf{U}}) = Norm_1(\mathbf{u}_i) \prod_{i,j} Norm_{\delta_{ij}}(\mathbf{u}_i - \mathbf{u}_j)$$
$$p(\mathbf{S}|\mathbf{Q}, \mathbf{U}, \Omega) = \prod_{(i,j)\in\Omega} Norm_{\lambda_S^{-1}}(S_{i,j}|\mathbf{q}_i^T \mathbf{u}_j).$$

We solve the probabilistic inference problem for GRLM by finding a maximum a posterior (MAP) configuration of the data matrix of questions $\mathbf{Q}$ and data matrix of users $\mathbf{U}$. The MAP is an objective function conditioning on the quality score matrix $\mathbf{S}$ and data matrix of questions $\mathbf{Q}$. That is, we aim to find

$$(\mathbf{Q}^*, \mathbf{U}^*) = \arg\max_{\mathbf{Q},\mathbf{U}} p(\mathbf{Q}, \mathbf{U}|\mathbf{S}, \mathbf{Q}, \lambda_W, \alpha, \beta, \Omega).$$

We observe that maximization a posterior configuration is equivalent to maximizing the complete log likelihood of matrix $\mathbf{Q}$ and $\mathbf{U}$. Thus, we give the complete log likelihood by

$$\begin{aligned}
\mathcal{L} = &-\frac{\lambda_{\mathbf{U}}}{2} \sum_{\mathbf{u}_i \in \mathbf{U}} {\mathbf{u}_i}^T \mathbf{u}_i - \frac{\lambda_{\mathbf{Q}}}{2} \sum_{\mathbf{q}_j \in \mathbf{Q}} (\mathbf{q}_j - \theta_j)^T (\mathbf{q}_j - \theta_j) \\
&+ \sum_{\mathbf{q}_i \in \mathbf{Q}} \sum_{q_{ij} \in \mathbf{q}_i} \log(\sum_{k=1}^{K} \theta_{i,k} \beta_{z_{i,j}=k, q_{i,j}}) \\
&- \frac{\lambda_W}{2} \sum_{\mathbf{u}_i, \mathbf{u}_j \in \mathbf{U}} \delta_{ij} (\mathbf{u}_i - \mathbf{u}_j)^T (\mathbf{u}_i - \mathbf{u}_j) \\
&- \frac{\lambda_S}{2} \sum_{(i,j)\in\Omega} (S_{i,j} - \mathbf{q}_i^T \mathbf{u}_j)
\end{aligned} \tag{6}$$

where $\lambda_W \geq 0$ and $\lambda_S \geq 0$ are trade-off parameters. We assume that the prior latent topic distribution for questions α is a uniform distribution. We adopt this assumption from the topic-model based performance prediction work [26, 24].

3.3 The Optimization Method

In this section, we propose an optimization method for solving Problem (6). We take the partial derivative for parameters $\mathbf{Q}$, $\mathbf{U}$, $\mathbf{\Theta}$ and β in the complete log likelihood $\mathcal{L}$ in Problem (6) and set them to zero.

We first report the optimization result for data matrix of questions $\mathbf{Q}$ and data matrix of users $\mathbf{U}$ by

$$\mathbf{q}_i \leftarrow (\lambda_S \mathbf{U}^T \mathbf{U} + \lambda_Q I_K)^{-1} (\lambda_S \mathbf{U} \mathbf{S}_i^q + \lambda_Q \theta_{\mathbf{q}_i}) \tag{7}$$

$$\begin{aligned}
\mathbf{u}_j \leftarrow &(\lambda_S \mathbf{Q}^T \mathbf{Q} + \lambda_W \sum_{\mathbf{u}_k \in \mathbf{U}} \delta_{jk} + \lambda_U I_K)^{-1} \\
&\times (\lambda_S \mathbf{Q} \mathbf{S}_i^u + \lambda_W \sum_{\mathbf{u}_k \in \mathbf{U}} \delta_{jk} \mathbf{u}_k)
\end{aligned} \tag{8}$$

where $\mathbf{S}_j^q$ and $\mathbf{S}_i^u$ are the diagonal quality score matrices for j-th question and i-th uesr, respectively.

We then present the optimization result for the latent topic of questions $\mathbf{\Theta}$ and β, respectively. We first find that it is difficult to directly take the derivative for the complete log likelihood problem with respect to parameter $\mathbf{\Theta}$. This is due to the decoupling between β and topic assignment matrix of words $\mathbf{Z}$ [1]. Thus, we introduce a new variational parameter $\mathbf{\Phi}$ for topic assignment matrix of words $\mathbf{Z}$ to derive a lower bound for the complete log likelihood, denoted by $\mathcal{L}'$. Consider the term $\sum_{q_{ij}\in\mathbf{q}_i}\log(\sum_{k=1}^{K}\theta_{i,k}\beta_{z_{ij},q_{ij}})$ in $\mathcal{L}$. We derive its lower bound such that the lower bound of $\mathcal{L}$ can also be obtained. The derivation is based on Jensen's Inequality.

By introducing the new variational parameter Φ, the lower bound of $\sum_{q_{ij}\in\mathbf{q}_i}\log(\sum_{k=1}^{K}\theta_{i,k}\beta_{z_{ij},q_{ij}})$ in $\mathcal{L}$ is given by

$$\begin{aligned}
&\sum_{q_{ij}\in\mathbf{q}_i}\log(\sum_{k=1}^{K}\theta_{i,k}\beta_{z_{ij},q_{ij}})\\
=&\sum_{q_{ij}\in\mathbf{q}_i}\log(\sum_{k=1}^{K}\frac{\theta_{i,k}\beta_{z_{ij},q_{ij}}\phi_{(q_i,j),k}}{\phi_{(q_i,j),k}})\\
\geq&-\sum_{q_{ij}\in\mathbf{q}_i}\sum_{k=1}^{K}\phi_{q_{ij},k}\log\phi_{q_{ij},k}\\
&+\sum_{q_{ij}\in\mathbf{q}_i}\sum_{k=1}^{K}\phi_{q_{ij},k}\log(\theta_{i,k})\beta_{z_{ij}=k,q_{ij}}.
\end{aligned}$$

Thus, we can iteratively estimate the latent topic of questions $\mathbf{\Theta}$ and β on $\mathcal{L}'$.

We then estimate the parameters Θ, Φ and β iteratively on $\mathcal{L}'$. We first report the optimization results for parameters Φ and β by

$$\phi_{q_{ij},k}\propto\theta_{i,k}\beta_{k,q_{ij}} \tag{9}$$

$$\beta_{k,q_{ij}}\propto\sum_{\mathbf{q}_i\in\mathbf{Q}}\sum_{q_{ij}\in\mathbf{q}_i}\phi_{q_{ij},k}. \tag{10}$$

We then estimate the latent topic of the questions $\mathbf{\Theta}$ by using the root fining algorithm in numerical optimization tool in [7].

3.4 The Expert Finding Algorithm

We now present a cold-start expert finding algorithm based on our proposed model GRLM in Algorithm 2.

Given an i-th new question $\mathbf{q}_i$, Algorithm 2 aims to choose the users with highly predicted score for answering this question. The main process of our algorithm can be divided into two parts. First, the algorithm estimates the data vector of the i-th question, denoted by $\mathbf{q}_i$. Second, the algorithm ranks the users for answering the i-th question based on both data matrix of users $\mathbf{U}$ and data matrix of questions $\mathbf{q}_i$ by Equation 1.

Algorithm 2 The Expert Finding Algorithm

Input: An i-th new question $\mathbf{q}$, data matrix of users $\mathbf{U}$ and β
Output: A ranked list of users $R(\mathbf{U})$

1: Set latent topic $\theta_i \propto$ Uniform distribution
2: **for** $t : 1 \rightarrow \tau_{max}$ **do**
3: **for** each word $q_{ij} \in \mathbf{q}_i$ **do**
4: **for** $k : 1 \rightarrow K$ **do**
5: Compute variational parameter $\phi_{q_{ij},k} \propto \theta_i \beta_{k,q_{ij}}$
6: **end for**
7: **end for**
8: Sample $\theta_i \propto \prod_{j=1}^{I} \sum_{k=1}^{K} \theta_{i,k} \phi_{q_{ij},k}$
9: **end for**
10: Rank users by Equation 1
11: **return** A ranked list of users $R(\mathbf{U})$

We now give the details of our expert finding algorithm as follows. First, Algorithm 2 iteratively estimates the latent topic of the i-th new question θ_i and variational parameter ϕ_i from Lines 1 to 9. Then, the algorithm ranks the users for question $\mathbf{q}_i$ in Line 11 and returns a ranked list.

4 Experimental Study

In this section, we conduct several experiments on the question-answering platform, Quora, and the social network, Twitter. The experiments are conducted by using Java, tested on machines with Linux OS Intel(R) Core(TM2) Quad CPU 2.66Hz, and 32GB RAM. The objectives of the study is to show the effectiveness of our proposed model GRLM for the problem of expert finding in CQA.

4.1 Datasets

We collect the data from Quora. Quora is a popular question-and-answer website, in which questions are posted and then answered by the community of its users. Quora was launched to the public in June, 2010 and it becomes very successful in terms of the number of users since then. We first crawl the questions posted between September 2012 and August 2013 and then crawl all the users who answered these questions. In total, we collect 444,138 questions, 95,915 users, 887,771 answers and 32,231 topical interests. In the following experiments, we evaluate our model GRLM on Quora instead of Yahoo Answer, since Quora provides the user specified topical interests.

We first sort the resolved questions by their posted timestamp and then split the resolved questions in Quora into the training dataset (i.e. first half of the questions) and the testing dataset (i.e. second half of the questions). Based on the number of the collected answers running from 1 to 6, we split the resolved questions into six groups denoted by $\mathbf{Q}_1$, $\mathbf{Q}_2$, ..., $\mathbf{Q}_6$. For example, group $\mathbf{Q}_1$ contains the questions with at least one answer. For each group $\mathbf{Q}_i$, we randomly sample 100 questions as the testing

dataset, denoted by $\mathbf{Q}'_i$. In total, we have 600 testing questions. We then keep other questions in groups $\mathbf{Q}_1$, $\mathbf{Q}_2$, ..., $\mathbf{Q}_6$ as the training dataset. Therefore, we generate a pair of training and testing datasets. In this study, we generate ten pairs of training and testing datasets to evaluate the performance of the algorithms. We take the average of the experimental results of these algorithms on ten pairs of datasets. The summary of the datasets is given in Table 2. The dataset will be provided later.

Table 2. Summary of Datasets

Dataset	#Questions	Average #Answers
$\mathbf{Q}_1$	444k	2
$\mathbf{Q}_2$	178k	3.5
$\mathbf{Q}_3$	86k	5.0
$\mathbf{Q}_4$	48k	6.7
$\mathbf{Q}_5$	30k	8.3
$\mathbf{Q}_6$	20k	10.0

4.2 Evaluation Criteria

We now discuss how to evaluate our algorithm. The performance of the expert finding algorithm can be gauged by the following three metrics: **Precision**, **Recall** and **Cold-Start Rate**.

Precision. We evaluate the ranking quality of different algorithms for the users who answered the questions by the two measurements of $Accu$ and $Precision@1$. Given a question, we consider the user whose answer receives the highest number of thumb-ups as the best answerer. Both $Accu$ and $Precision@1$ evaluate the ranking of the best answerer by different algorithms (i.e. whether the best answerer can be ranked on top). These measurements are widely used in existing work [27, 34] to evaluate the performance of the expert finding algorithms in CQA systems.

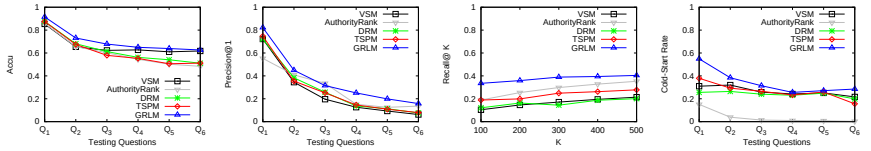
Given a question $\mathbf{q}_i$, we denote by $R(\mathbf{U})^i$ the ranking of the users who answered this question. We denote by $|R(\mathbf{U})^i|$ the number of the users in the ranking $R(\mathbf{U})^i$. We denote by R^i_{best} the rank of the best answerer for question $\mathbf{q}_i$ by different algorithms. The formula of $Accu$ is given by

$$Accu = \sum_{\mathbf{q}_i \in \mathbf{Q}'} \frac{|R(\mathbf{U})^i| - R^i_{best}}{|R(\mathbf{U})^i||\mathbf{Q}'|},$$

where $\mathbf{Q}'$ is the set of the testing questions. The $Accu$ illustrates the ranking percentage of the best answerer by different algorithms.

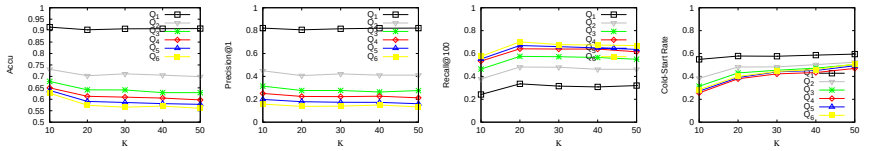
We now evaluate the precision of the experts ranked on top by different algorithms. We use $Precision@1$ to validate whether the expert ranked on top is the best answerer by different algorithms. The formula of $Precision@1$ is given by

$$Precision@1 = \frac{|\{\mathbf{q}_i \in \mathbf{Q}' | R^i_{best} \leq 1\}|}{|\mathbf{Q}'|}.$$



(a) $Accu$ (b) $Precision@1$ (c) $Recall@K$ (d) $Cold\text{-}Start\ Rate$

Fig. 4. Performance Comparison of the Algorithms



(a) $Accu$ (b) $Precision@1$ (c) $Recall@K$ (d) $Cold\text{-}Start\ Rate$

Fig. 5. Effect on Dimension of Latent Space K

Recall. We employ the measurement $Recall@K$ to evaluate the ranking quality for all users in CQA systems by different algorithms. Given the i-th new question $\mathbf{q}_i$, we denote by R^i_{TopK} the set of users ranked on $TopK$ by the algorithms. The formula of $Recall@K$ is given by

$$Recall@K = \frac{|\{\mathbf{q}_i \in \mathbf{Q}' | j \in R^i_{TopK} \text{ and } (i,j) \in \Omega\}|}{|\mathbf{Q}'|}.$$

The $Recall@K$ aims to choose the right experts from all the users in CQA systems.

Cold-Start Rate. We also investigate the types of the experts found by different algorithms (i.e. cold-start users or warm-start users). In this experimental study, we consider the users who answered less than τ questions as cold-start users, where τ is the threshold for cold-start users. We propose the measurement $Cold\text{-}Start\ Rate$ to illustrate the type of the experts ranked on top (i.e. Top1), which is given by

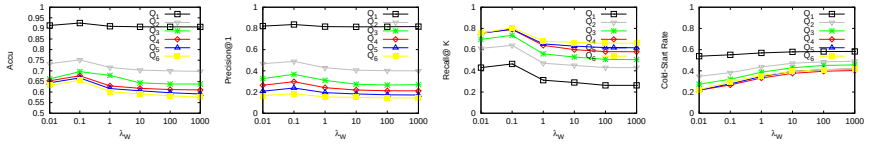
$$\text{Cold-Start Rate} = \frac{|\{\mathbf{q}_i \in \mathbf{Q}' | j \in R^i_{Top1} \text{ and } 1_{(i,j)} \leq \tau\}|}{|\mathbf{Q}'|},$$

where R^i_{Top1} is the set containing the found expert ranked the top.

We compute the Precision, Recall and Cold-Start Rate of all the algorithms on $\mathbf{Q}'_1$, $\mathbf{Q}'_2, \ldots, \mathbf{Q}'_6$.

4.3 Performance Evaluation

We compare our model GRLM with the following state-of-the-art expert finding algorithms: Vector Space Model (VSM) [27], AuthorityRank [2], Dual Role Model (DRM) [27] and Topic Sensitive Probabilistic Model (TSPM) [34]. The underlying idea of using these algorithms for expert finding in CQA systems are highlighted below:



(a) $Accu$ (b) $Precision@1$ (c) $Recall@K$ (d) $Cold\text{-}Start\ Rate$

Fig. 6. Effect on Regularization Term λ_W

– **VSM.** The VSM constructs the feature of the users based on the past question-answering activities in a word level. Consider the word vector of the i-th question as $\mathbf{q}_i$. The word vector of the j-th user is constructed from the word vector of the answered questions, denoted by $\mathbf{u}_j$. Given a new question q, VSM ranks the relevance of the users based on the dot product of the word vectors of the i-th question and the j-th user u by $\hat{S_{ij}} = \mathbf{q}_i^T \mathbf{u}_j$.
– **AuthorityRank.** The AuthorityRank computes the expertise authority of the users based on the number of provided best answerers, which is an in-degree method. Given a new question, AuthorityRank ranks the users based on their expertise authority.
– **DRM.** The DRM discovers the latent expertise of the users from their past question-answering activities, which is based on the famous topic modeling technique called *probabilistic latent semantic analysis* (PLSA) [10]. Given a new question, DRM ranks the users based on their latent expertise.
– **TSPM.** The TSPM discovers the latent expertise of the users based on another famous topic modeling technique called *latent Dirichlet allocation* (LDA) [1] and ranks the users based on their latent expertise.

Figures 4(a) to 4(d) show the evaluation results based on $Accu$, $Precison@1$, $Recall@K$ and $Cold\text{-}Start\ Rate$, respectively. The evaluation were conducted with different types of the questions. For each dataset, we report the performance of all methods.

The AuthorityRank method is based on the link analysis of the question-answering activities of users while DRM and TPSM models are based on topic-oriented probabilistic model. These experiments reveal a number of interesting points as follows:

– The topic-oriented probabilistic models DRM and TPSM outperform the authority-based model. This findings suggests that using the latent user model for tackling the problem of expert finding in CQA systems is effective.
– Our GRLM model achieves the best performance, which indicates that leveraging the user-to-user graph can further improve the performance of expert finding in CQA systems.
– The experimental study on $Recall@K$ indicates that our method can find the right experts where the candidate experts are all the users in CQA systems. We notice that our model GRLM is able to find the best answerer in the top 100 ranked users with the probability of 0.37 as shown in Figure 4(c).

There are two essential parameters in our model, which are the dimension of latent space K, and the graph regularization parameter λ_W. The parameter K represents the latent feature size of latent user model and latent topic space of questions. The parameter λ_W shows the obtained benefits of our method from the inferred user-to-user graph.

We first study the impact of parameter K by varying its value from 10 to 50, and present the experimental results in Figures 5(a) to 5(d). Figure 5(a) shows that the Cold-Start Rate of the experts found by GRLM increases and then becomes convergent with respect to the dimension of latent space. Figure 5(d) shows that the recall of the expert finding has 10% improvement by varying the parameter K. By transferring the knowledge to the cold-start users, both cold-start users and warm-start users can be selected such that the recall is improved. Figures 5(b) and 5(c) illustrate that the accuracy doesn't vary for the parameter K. From these results, we conclude that the setting $K = 10$ is good enough to represent the latent features of both users and questions in CQA systems.

We then study the impact of the regularization term λ_W on the performance of GRLM, which is illustrated in Figures 6(a) to 6(d). We vary the value of the regularization term λ_W from 0.01 to 1000. The success of graph regularized latent model for expert finding relies on the assumption that two neighboring users share the similar user model. When the value of λ_W becomes small, our model can be considered as the previous topic-oriented expert finding methods, which are only based on the past question-answering activities. We vary parameter λ_W to investigate the benefits of our methods from the idea of graph regularized latent model for the problem of expert finding. We notice that our method consistently performs on most of the varied values of parameter λ_W. To balance the inference of latent user model from both past question-answering activities and user-to-user graph, we set the value of parameter λ_W as a new regularization term. We report that the overall performance of GRLM with the new regularization term can also be improved by 3%, 3% and 10% on $Accu$, $Precision@1$ and $Recall@100$, respectively.

5 Related Work

In this section, we briefly review some related work on the problem of expert finding, cold-start recommendation in the literature.

Expert Finding. The problem of expert finding in CQA systems has attracted a lot of attention recently. Roughly speaking, the main approaches for expert finding can be categorized into two groups: the authority-oriented approach [2, 39, 11, 30, 15] and the topic-oriented approach [28, 19, 27, 18, 34, 9, 16, 31, 33, 32, 22].

The authority-oriented expert finding methods are based on link analysis of the past question-answering activities of the users in CQA systems. Bouguessa et al. [2] discover the experts based on the number of best answers provided by users, which is an in-degree-based method. Zhu et al. [39, 38] select experts based on the authority of the users on the relevant categories of the questions. Jurczyk et al. [11] propose a HITS [12] based method to estimate the ranking score of the users based on question-answering activity graphs. Zhang et al. [30] propose an expertise ranking method and

evaluated link algorithms for specific domains. Jing et al. [15] propose a competition model to estimate the user expertise score based on question-answering activity graphs.

The topic-oriented expert finding methods are based on latent topic modeling techniques. Deng et al. [3] and Hashemi et al. [9] tackle the problem of expert finding in bibliographic networks. Using the generative topic model, Xu et al. [27] propose a dual role model that jointly represents the roles of answerers and askers. Liu et al. [16] propose a language model to predict the best answerer. Guo et al. [8] and Zhou et al. [34] devise the topic sensitive model to build the latent user model for expert finding. Liu et al. [28] model both topics and expertise of the users in CQA for expert finding. Saptarshi et al. [6] utilize the crowdsourcing techniques to find the topic experts in microblogs. Fatemeh et al. [22] incorporate the topic modeling techniques to estimate the expertise of the users.

In contrast to the above-mentioned work, our emphasis is on cold-start expert finding in CQA systems. We suggests exploiting the "following relation" between users and topical interests to resolve the problem. The existing work mainly focus on the problem of expert finding based on the past question-answering activities of users.

Cold-Start Recommendation. Recently, the cold-start problem in user-item recommendation has attracted a lot of attention and several approaches [20, 21, 14, 36, 29] are proposed to solve this problem. Park et al. [20] propose a latent regression model that leverages the available attributes of items and users to enrich the information. Zhou et al. [36] devise an interview process that iteratively enriches the profile of the new users. Yin et al. [29] propose a random walk based method to choose the right cold-start items for users. Purushotham et al. [21] utilize both the text information of items and social relations of users to user-item recommendation. Zhu et al. [14] extract the information of items from Twitter to overcome the difficulty of cold-start recommendation. However, the cold-start recommendation techniques cannot be applied to the context of this work.

6 Conclusion

We formulate the problem of cold-start expert finding and explore the user-to-user graph in CQA systems. We propose a novel method called graph regularized latent model. We consider the latent user model based on the topic of the questions which can be inferred from question feature and the quality score matrix. Our approach integrates the inferred user-to-user graph and past question-answering activities seamlessly into a common framework for tackling the problem of cold-start expert finding in CQA systems. In this way, our approach improves the performance of expert finding in the cold-start environment. We devise a simple but efficient variational method to solve the optimization problem for our model. We conduct several experiments on the data collected from the famous question-answering system, Quora. The experimental results demonstrate the advantage of our GRLM model over the state-of-the-art expert finding methods.

ACKNOWLEDGEMENTS This work is partially supported by GRF under grant number HKUST FSGRF13EG22 and HKUST FSGRF14EG31

References

1. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
2. M. Bouguessa, B. Dumoulin, and S. Wang. Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In *SIGKDD*, pages 866–874. ACM, 2008.
3. H. Deng, I. King, and M. R. Lyu. Formal models for expert finding on dblp bibliography data. In *ICDM*, pages 163–172. IEEE, 2008.
4. G. Dror, Y. Koren, Y. Maarek, and I. Szpektor. I want to answer; who has a question?: Yahoo! answers recommender system. In *Proceedings of SIGKDD*, pages 1109–1117. ACM, 2011.
5. A. Figueroa and G. Neumann. Learning to rank effective paraphrases from query logs for community question answering. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.
6. S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 575–590. ACM, 2012.
7. GSL. `https://www.gnu.org/software/gsl/`.
8. J. Guo, S. Xu, S. Bao, and Y. Yu. Tapping on the potential of q&a community by recommending answer providers. In *CIKM*, pages 921–930. ACM, 2008.
9. S. H. Hashemi, M. Neshati, and H. Beigy. Expertise retrieval in bibliographic network: a topic dominance learning approach. In *CIKM*, pages 1117–1126. ACM, 2013.
10. T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57. ACM, 1999.
11. P. Jurczyk and E. Agichtein. Discovering authorities in question answer communities by using link analysis. In *CIKM*, pages 919–922. ACM, 2007.
12. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.
13. W.-J. Li and D.-Y. Yeung. Relation regularized matrix factorization. In *IJCAI*, pages 1126–1131, 2009.
14. J. Lin, K. Sugiyama, M.-Y. Kan, and T.-S. Chua. Addressing cold-start in app recommendation: Latent user models constructed from twitter followers. 2013.
15. J. Liu, Y.-I. Song, and C.-Y. Lin. Competition-based user expertise score estimation. In *SIGIR*, pages 425–434. ACM, 2011.
16. X. Liu, W. B. Croft, and M. Koll. Finding experts in community-based question-answering services. In *CIKM*, pages 315–316. ACM, 2005.
17. H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *WSDM*, pages 287–296. ACM, 2011.
18. G. Miao, L. E. Moser, X. Yan, S. Tao, Y. Chen, and N. Anerousis. Generative models for ticket resolution in expert networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 733–742. ACM, 2010.
19. D. Mimno and A. McCallum. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 500–509. ACM, 2007.
20. S.-T. Park and W. Chu. Pairwise preference regression for cold-start recommendation. In *RecSys*, pages 21–28. ACM, 2009.
21. S. Purushotham, Y. Liu, and C.-C. J. Kuo. Collaborative topic regression with social matrix factorization for recommendation systems. *arXiv preprint arXiv:1206.4684*, 2012.
22. F. Riahi, Z. Zolaktaf, M. Shafiei, and E. Milios. Finding expert users in community question answering. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 791–798. ACM, 2012.

23. A. N. Srivastava and M. Sahami. *Text mining: Classification, clustering, and applications*. CRC Press, 2009.
24. C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456. ACM, 2011.
25. G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao. Wisdom in the social crowd: an analysis of quora. In *Proceedings of WWW*, pages 1341–1352. International World Wide Web Conferences Steering Committee, 2013.
26. H. Wang, B. Chen, and W.-J. Li. Collaborative topic regression with social regularization for tag recommendation. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2719–2725. AAAI Press, 2013.
27. F. Xu, Z. Ji, and B. Wang. Dual role model for question recommendation in community question answering. In *Proceedings of SIGIR*, pages 771–780. ACM, 2012.
28. L. Yang, M. Qiu, S. Gottipati, F. Zhu, J. Jiang, H. Sun, and Z. Chen. Cqarank: jointly model topics and expertise in community question answering. In *CIKM*, pages 99–108. ACM, 2013.
29. H. Yin, B. Cui, J. Li, J. Yao, and C. Chen. Challenging the long tail recommendation. *VLDB*, 5(9):896–907, 2012.
30. J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *WWW*, pages 221–230. ACM, 2007.
31. Z. Zhao, J. Cheng, F. Wei, M. Zhou, W. Ng, and Y. Wu. Socialtransfer: Transferring social knowledge for cold-start cowdsourcing. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 779–788. ACM, 2014.
32. Z. Zhao, W. Ng, and Z. Zhang. Crowdseed: query processing on microblogs. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 729–732. ACM, 2013.
33. Z. Zhao, D. Yan, W. Ng, and S. Gao. A transfer learning based framework of crowd-selection on twitter. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1514–1517. ACM, 2013.
34. G. Zhou, S. Lai, K. Liu, and J. Zhao. Topic-sensitive probabilistic model for expert finding in question answer communities. In *Proceedings of CIKM*, pages 1662–1666. ACM, 2012.
35. G. Zhou, Y. Liu, F. Liu, D. Zeng, and J. Zhao. Improving question retrieval in community question answering using world knowledge. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2239–2245. AAAI Press, 2013.
36. K. Zhou, S.-H. Yang, and H. Zha. Functional matrix factorizations for cold-start recommendation. In *SIGIR*, pages 315–324. ACM, 2011.
37. T. C. Zhou, X. Si, E. Y. Chang, I. King, and M. R. Lyu. A data-driven approach to question subjectivity identification in community question answering. In *AAAI*, 2012.
38. H. Zhu, H. Cao, H. Xiong, E. Chen, and J. Tian. Towards expert finding by leveraging relevant categories in authority ranking. In *CIKM*, pages 2221–2224. ACM, 2011.
39. H. Zhu, E. Chen, H. Xiong, H. Cao, and J. Tian. Ranking user authority with relevant knowledge categories for expert finding. *World Wide Web*, pages 1–27, 2013.