Context-Aware Expert Finding in Tag Based Knowledge Sharing Communities

Hengshu Zhu, University of Science and Technology of China and Nokia Research Center, China

Enhong Chen, University of Science and Technology of China, China

Huanhuan Cao, Nokia Research Center, China

Jilei Tian, Nokia Research Center, China

ABSTRACT

With the rapid development of online Knowledge Sharing Communities (KSCs), the problem of finding experts becomes increasingly important for knowledge propagation and putting crowd wisdom to work. A recent development trend of KSCs is to allow users to add text tags for annotating their posts, which are more accurate than traditional category information. However, how to leverage these user-generated tags for finding experts is still underdeveloped. To this end, this paper develops a novel approach for finding experts in tag based KSCs by leveraging tag context and the semantic relationship between tags. Specifically, the extracted prior knowledge and user profiles are first used for enriching the query tags to infer tag context, which represents the user's latent information needs. Specifically, two different approaches for addressing the problem of tag sparseness in authority ranking are proposed. The first is a memory-based collaborative filtering approach, which leverages non-negative matrix factorization (NMF) to find similar users for alleviating tag sparseness. The second approach is based on Latent Dirichlet Allocation (LDA) topic model, which can further capture the latent semantic relationship between tags. A large-scale real-world data set is collected from a tag based Chinese commercial Q&A web site. Experimental results show that the proposed method outperforms several baseline methods with a significant margin.

Keywords: *Collaborative Filtering, Expert Finding, Knowledge Sharing Communities, Question Answering, Topic Models, User-Generated Tags*

1. INTRODUCTION

In recent years, researchers have witnessed the rapid development of online Knowledge Sharing Communities (KSCs), such as blogs, discussion boards, and question answering (Q&A) communities. Users can share experiences and exchange ideas with others in such KSCs. Their sharing activities generate a large amount of knowledge and also attract many expert users of each domain to participate. As a result, more

DOI: 10.4018/jkss.2012010104

and more people would like to use KSCs for problem solving. Some researchers have found that Q&A content is usually the largest part of content in KSCs (Cong et al., 2008; Feng et al., 2006). However, comparing with the large number of questions, the expert users are still scarce resources in KSCs. As a result, there are a lot of questions without satisfactory answers due to the lack of relevant experts. Thus, how to find the experts for an answer-lacking question becomes an important problem to be addressed.

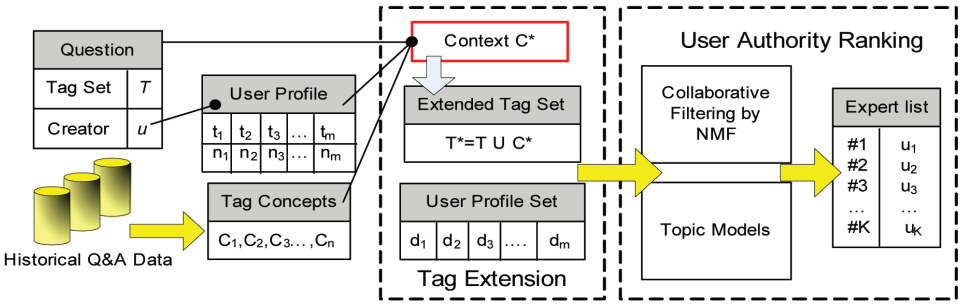
The problem of expert finding for answer-lacking questions has been well studied. Some of the traditional works leverage content based approaches (e.g., Balog et al., 2006; Liu et al., 2005). In these works, researchers can utilize language models to rank user authority through the question textual distribution in each of the users' historical records. However, these approaches are usually computationally intensive and are hardly applicable to large-scale data sets. Likewise, some novel KSCs are based on multimedia content and the textual information contained in questions are often not rich enough for building language models (Yeh & Darrell, 2008). Therefore, most of the state-of-the-art works leverage question categories as query inputs to find experts (e.g., Bouguessa et al., 2008; Jurczyk & Agichtein, 2007; Kao et al., 2010). A drawback of these works is that each question can only be classified into one category by them. Actually it is usually difficult to select the best category for a question because a question is usually related to multiple categories. For example, an inexperienced user cannot easily select a better category between "Mobile Device" and "Market" for the question "Where can I buy the Nokia new mobile phone N9?" As a result, the conventional category based expert finding approaches may have poor performances for a multiple-category question.

A recent trend of KSCs is to allow users to add text tags for their questions, such as Tianya Wenda (http://wenda.tianya.cn) and Douban (http://www.douban.com). In these web sites, users can use tags as descriptive labels to annotate the contents they post. To be specific, user can add tags like "N9", "Mobile Market" and "Where" for the above question. Expert users can check the tags of a given question to decide whether to answer it. Compared with the textual information of question content, user-generated tags are simplified as query inputs and can be utilized on large-scale data sets for expert finding. Moreover, user-generated tags contain richer information of the user needs than category information and can be used for facilitating experts finding. However, because the tags are generated by users but not system, they are usually ambiguous and not regular. Therefore, how to leverage these user-generated tags for expert finding becomes a great challenge. The following motivating examples intuitively illustrate the challenges of using tags for expert finding.

- **Motivating Example 1.** *Joy posts a question about the Sony video game console "Play Station" and adds a tag "PS" to annotate the question. However, in many contexts, the tag "PS" may be also referred to the Adobe software "Photo Shop".*
- **Motivating Example 2.** *Kate wants to buy a new mobile phone and posts a question with tags "Mobile Phone", "Market". However, the latent information needs for Kate are about "Discount" and "Trustable store".*
- **Motivating Example 3.** *Joy posts a question about computer devices with a tag "PC", however, Kate may add a tag "Laptop" for the same question. Actually, the different tags may represent similar meanings.*

Inspired by above observations, in this paper, we propose a novel tag based framework for expert finding by inferring the users' latent information needs and uncovering the semantic relationship between tags. Specifically, we first introduce an effective tag extension method which enriches the original question tags with the question creator's latent information needs. The latent information needs are modeled by tag concepts extracted from the users' historical Q&A data and the question creator's profile,

Figure 1. The overview of our tag based expert finding approach



which are referred to as the ***tag contexts*** of original question tags. Then, we propose a probabilistic framework for ranking user authority to find experts. To be specific, we develop two different approaches for addressing the problem of tag sparseness in authority ranking. The first is a memory-based collaborative filtering approach, which leverages non-negative matrix factorization (NMF) (Lee & Seung, 1999) to find similar users for alleviating tag sparseness. The second approach is based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic model, which models tags as topic distributions and thus can further capture the latent semantic relationship between tags. The overview of our approach is illustrated in Figure 1. Finally, we conduct extensive experiments on a large-scale real-world data set collected from a major Chinese commercial Q&A web site. Experimental results clearly show that the proposed method outperforms several baseline methods with a significant margin.

1.1. Overview

The remainder of this paper is organized as follows. Section 2 provides a brief overview of related works. Section 3 shows the novel approach of context-aware tag extension, and Section 4 shows the details of our NMF collaborative filtering and topic model based authority ranking approaches for finding experts. In Section 5, we present the experiment results. Finally, Section 6 concludes the work.

2. RELATED WORK

With the rapid development of online Knowledge Sharing Communities (KSCs) in these years, expert finding becomes one of the most important problems with great application potentials in the research domains of knowledge management and social networks. Indeed, the problem of expert finding has been well studied by many researchers. Generally, the previous works of expert finding can be grouped into two categories, which are content based and category based approaches, respectively.

In the first category, researchers utilize content based approaches to finding experts for answer-lack questions. For example, Balog et al. (2006) used conventional language models for finding experts in enterprise corpora. As a further research, Balog et al. (2009) also proposed a generative probabilistic framework of leveraging language models for expert finding. Based on this model, they also proposed two basic models for implementing various expertise search strategies in experiments. Liu et al. (2005) have investigated finding experts in community based Q&A services by leveraging user profiles into language models. Zhang et al. (2007) proposed a mixture model based on Probabilistic Latent Semantic Analysis (PLSA), which can discover semantically related experts for a given question. Although these approaches can estimate the similarity between question content with experts directly, they cannot be utilized into multimedia based KSCs and are

usually computationally intensive when applied to a large-scale data set.

In the second category, most of the state-of-the-art works on expert finding focus on ascertaining the most authoritative users for a specific question category. These category based approaches often leverage link analysis algorithms on category link graphs where the nodes represent the interactive users and the edges represent their Q&A relationships on the given category. For example (Jurczyk & Agichtein, 2007) formulated a graph structure in Q&A communities and proposed a variation of the HITS (Kleinberg, 1999) algorithm for predicting authoritative users in Yahoo! Answers. Zhang et al. (2007) investigated various authority ranking algorithms in the Java forum and also proposed a PageRank (Page et al., 1999) like algorithm named "ExpertiseRank" to find experts. Zhang et al. (2008) proposed a propagation-based approach for finding experts in co-author social networks which take into account user profiles and Lu et al. (2009) extended it with latent link analysis and language model. However, these category based approaches have poor performance given a multiple categories question. Therefore, alternatively, in this paper we exploit user-generated tags but not question category for expert finding.

In addition, the proposed approach in this paper exploits non-negative matrix factorization (NMF) and topic models for ranking user authority. Both of the two approaches are widely used in the areas of information extraction and text retrieval. Specifically, the NMF is proposed by Lee and Seung (1999), which leverages the non-negativity constraints in the process of matrix factorization. Therefore, the proposed approach can be naturally used in many real-world scenarios, such as image recognition and recommender systems. Topic model is one of the most widely used models for discovering latent semantic relationships between objects. Typical topic models include the Mixture Unigram (MU) (Nigam et al., 1999), the Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999) and the Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Most of other topic models are extended from the above ones for satisfying some specific requirements. In our approach, we exploit the widely used LDA model.

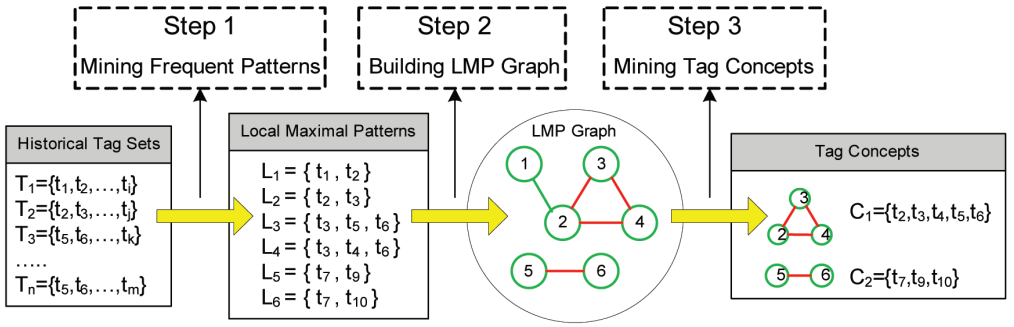
3. CONTEXT-AWARE TAG EXTENSION

In this section, we first introduce two related notions that will be used in this paper. A *question and answer pair* (Q&A pair) p_i contains a question q_i and all of its answers $A_i = \{a_{i1}, a_{i2}, ..., a_{ir}\}$. There exists and only exists one creator for q_i and each answer a_{ij}. Every question q_i contains a tag set $T_i = \{t_{i1}, t_{i2}, ..., t_{ik}\}$ generated by its creator to describe the question. The *user profile* d_i is a set of tags where the user u_i created or replied the questions with these tags and the corresponding frequencies.

To address the problem of inferring the latent information needs of a question creator, we take into account the context of question tags by taking advantage of historical Q&A pairs to extend the original tags. To be specific, we first summarize user tags by concepts for capturing the semantic information of tags. Then we extend the tags of a given question by leveraging the concepts which are most relevant to the question creator's user profile and the original tags.

We assume two tags belong to a same concept if they usually co-occur in same question tag sets. A set of tags which often co-occur in the same question tag sets is referred as a *tag concept*. Intuitively, given a user-generated tag, the tags in the same tag concept can be used as candidate extensions for reflecting the context information. The selection of tag concepts can be based on (1) the relevance between the candidate tag concepts and the original tags, and (2) the frequencies of their contained tags in the question creator's user profile. For example, if from Joy's profile we can find she likes playing games, thus we can extend tags "Sony", "Game" with the original tag "PS". The selected tag concepts are referred to as the tag context of the original question tags.

Figure 2. The generation process of local maximal pattern graph and tag concepts



3.1. Identifying Tag Concepts

To build tag concepts, in this paper, we take advantage of the frequent pattern mining approach for capturing tag co-occurrences. Given a transaction database $TDB = T$ and a minimal support threshold $min\_sup = \sigma$, a set c of items is a frequent pattern if $Count(T : T \in TDB, c \subseteq T) \geq \sigma$.

To be specific, we first mine the subset of tags which co-occur frequently in user-generated tag sets of historical questions. There are several successful frequent pattern mining algorithms, such as Apriori (Azzopardi & Srikant, 1994), FP Growth (Han et al., 2004) and PrefixSpan (Pei et al., 2001). All of these algorithms can be leveraged in our approach for mining tag concepts. In our experiments we utilize the widely used FP-Growth algorithm. Specifically, this algorithm will first build a FP-tree by scanning the *TDB* once, and then find frequent patterns according to this FP-tree. Therefore, this algorithm can find large item sets without candidate generation, which guarantees the computational cost in large-scale data set (e.g., the data set used in our experiments).

In this paper, we define a frequent pattern without super patterns as *Local Maximal Pattern* (LMP). That is, there are no frequent tags can be used for further pattern growth. Furthermore, we define a *LMP graph* where each node denotes the local maximal pattern, and there is an undirected edge between two nodes if and only if they have common tags. The step 2 in Figure 2 shows an example of the generation of LMP graph.

Based on the LMP graph, we can cluster tag patterns into concepts based on an intuitive assumption that the patterns in a connected sub-graph is likely to be appeared in the same context. Therefore, we define the tag concept in our approach as follows.

Definition 1 (Tag Concept)

Tag concept c_i is a union of all local maximal patterns l_j in a maximal complete connected sub-graph with more than two nodes or an isolated connected sub-graph with two nodes. And there is no other concept c_i makes $c' \subset c_i$.

Let us take the LMP graph in Figure 2 for example, there is only one maximal complete connected sub-graph with more than two nodes, which is $\{n_2, n_3, n_4\}$. Moreover, there is also an isolated connected sub-graph with two LMP nodes, which is $\{n_5, n_6\}$. Therefore, there exists two concepts $c_1 = l_2 \cup l_3 \cup l_4 = \{t_2, t_3, t_4, t_5, t_6\}$ and $c_2 = l_5 \cup l_6 = \{t_7, t_9, t_{10}\}$.

3.2. Tag Extension by Concepts

For each user question, we select at most top R relevant tag concepts, which are namely the question context, for extending the original tags by taking into account the user profile. The extension process is shown in Algorithm 1.

Algorithm 1. Context-aware tag extension

Input: concept set C, original question tag set T and creator's profile d
Output: the extended tag set T^*
1 **For** $c \in C$ **do**
2 **If** $c \cap T \cap d == \varnothing$ **then**
33 $Rel(c,d,T) = 0$;
4 **End**
5 **Else**
6 Compute $Rel(c,d,T)$ for c;
7 **End**
8 **End**
9 Descending rank $c \in C$ according to $Rel(c,d,T)$;
10 **For** $1 \le i \le R$ **do**
11 **If** $Rel(c,d,T) \neq 0$ **then**
12 $T^* = T \cup \{c_i\}$;
13 **End**
14 **End**
15 **Return** T^*

Specially, we rank the relevant concepts according to a relevant function $Rel(c,d,T)$ obtained in Step 6, where c is a given tag concept. The definition of the relevant function $Rel(c,d,T)$ should base on two basic principles, (1) the concept contains more frequent tags in question creator's profile will be ranked higher, and (2) the concept contains more common tags with original question tags will be ranked higher. With respect to the above basic principles, we define the function $Rel(c,d,T)$ as follows:

$$\mathrm{Re}\,l(c,d,T) = \frac{|\,c \cap T\,|}{|\,T\,|} \times \sum_{t \in d} \frac{f(t)}{rank(t)}, \tag{1}$$

where $rank(t)$ denotes the rank of tag t in question creator's profile d according to its frequency, and the binary function $f(t) = 0$ if $t \notin c$ and $f(t) = 1$ if $t \in c$.

How to select a proper number of relevant concepts (i.e., R) for extending original tags is an open question. Intuitively, a smaller R will limit the performance of inferring latent requirements for the given question, and a bigger R will also impact the performance of expert finding due to the false extension with irrelevant tags. In our experiments, we test different R for expert finding and the results justified our discussion.

4. TAG BASED USER AUTHORITY RANKING

After generating richer tags for reflecting the context information of questions, the remaining task is to rank user authority according to the extended tags. To formalize the authority ranking task, we use $P(u \mid q)$ to denote the probability of a user u being an expert for the given question q. Therefore, we can rank user authority according to this conditional probability, and select the most authoritative users as candidate experts for the given question. Using Bayes formula we have the following equations:

$$P(u \mid q) = \frac{P(q \mid u)P(u)}{P(q)} \propto P(q \mid u)P(u). \tag{2}$$

According to the equations above, the main procedure for expert ranking is to calculate the

conditional probability $P(q \mid u)$ and $P(u)$. The probability $P(u)$ can be estimated by the frequency that u appears in all Q&A pairs divided by the total number of Q&A pairs. Assuming that the probabilities of generating different tags are conditionally independent given a user and using tags $T = \{t_1, t_2, ..t_n\}$ to represent the question q, we have the following equation:

$$P(u \mid q) \propto P(u) \prod_{t_i \in q} P(t_i \mid u), \quad (3)$$

where the probability $P(t_i \mid u)$ equals to the relative frequency of tag t_i appearing in user profile of u. Therefore, if one of the tags in the given question does not appear in the user's profile, the probability of being a candidate expert for this user will be equal to zero. However, this situation is not reasonable in practice because the tags in individual users' profiles are often very sparse. For example, when given a question with tags "Mobile" and "N9" for finding experts, many users do not contain both of the two tags in their profiles. However, we cannot neglect the users with only tag "Mobile" or "N9" in their profiles, because they also have expertise in the relevant domains for resolving the given questions. To address this problem, we propose two different approaches for ranking user authority. To be specific, the first is a memory-based collaborative filtering approach, which leverages NMF to find similar users for enriching the tags in individual users' profiles. However, a further problem is tags often have latent semantic relationships, which cannot be resolved by the first approach. Therefore, we propose the second approach based on topic models, which can resolve the problem effectively. In the following sub-sections, we introduce the details of the two approaches, respectively.

4.1. Authority Ranking by NMF Collaborative Filtering

An intuitive strategy for resolving tag sparseness in individual users' profiles is leveraging collaborative filtering approaches, which are widely used in recommender systems for resolving the problem of item sparseness. Therefore, in this paper we propose to use a memory-based collaborative filtering approach for enriching tags in user profiles. To be specific, in this approach we first find similar users for each candidate expert u, and then use tags in these similar users' profiles to enrich the profile of u. However, according to the analysis in previous works of recommender systems (e.g., Resnick et al., 1994), it is very hard to estimate the user similarities in the very sparse tag space. To this end, we first proposed to leverage the widely used matrix factorization approach to map tags into low dimensional space for alleviating tag sparseness. To be specific, we first represent all user profiles as an observed matrix B_{MN}, where $b_{ij} \in B_{MN}$ is the number of times tag t_j has appeared in user profile of u_i, M is the number of unique users and N is the number unique tags. Then we let:

$$B_{MN} = W_{MK} \times H_{KN}, \quad (4)$$

where W_{MK} and H_{KN} are two new low-rank matrixes, which map users and tags into a K dimension space $\{z_{1,...,K}\}$. To be specific, here we have $K << M, N$.

To facilitate the process of leveraging factorization results for resolving our problem, in this paper we propose to use NMF (Lee & Seung, 1999) for obtaining matrix W and H. Specifically, in the matrix factorization process, we add non-negativity constraints in two matrixes W and H, which mean each value in these two matrix should be non-negative. It is a natural idea because for each value $w_{ij} \in W$ and $h_{jk} \in H$ implies the frequency of user u_i appears in the latent space z_j and the frequency of tag t_k appears in the latent space z_j, respectively. Therefore, both of these values should be non-negative. To efficiently obtain the two matrix W and H, we leverage the iteration algorithm proposed in Lee and Seung (2001) for matrix factorization. In this algorithm, the objective is to minimize the Euclid-

ean distance $|| M - WH ||^2$ with constraints $W, H \geq 0$. In the first round of iteration, we randomly assign non-negative values for both matrix M and H. Then in each further round of iterations, we update value for W and H by:

$$h_{ij} \leftarrow h_{ij} \frac{(W^T B)_{ij}}{(W^T WH)_{ij}},$$

$$w_{ij} \leftarrow w_{ij} \frac{(BH^T)_{ij}}{(WHH^T)_{ij}}. \quad (5)$$

After several rounds of iteration, the two matrix W and H will converge, then we can use matrix W to estimate the similarity between two users. To be specific, according to Resnick et al. (1994), given two user u_i and u_j, we can calculate their similarity by Pearson correlation coefficient:

$$Sim(u_i, u_j) = \frac{\sum_{t_k \in S}(w_{ik} - \overline{w}_i)(w_{jk} - \overline{w}_j)}{\sqrt{\sum_{t_k \in S}(w_{ik} - \overline{w}_i)^2}\sqrt{\sum_{t_k \in S}(w_{jk} - \overline{w}_j)^2}}, \quad (6)$$

where S is the set of common tags in both u_i and u_j's profiles and $\overline{w}_i = \frac{\sum_{t_k \in S} w_{ik}}{|S|}$.

Therefore, given a user u, we can enrich his/her profile by other users' profiles with respect to different similarities. To be specific, according to Resnick et al. (1994), the revised frequency of tag t_i in u's profile, denoted as n_i^u, can be calculated by:

$$n_i^u = \overline{n}^u + \frac{\sum_{u' \in NS_u} Sim(u, u')(n_i^{u'} - \overline{n}^{u'})}{\sum_{u' \in NS_u} |Sim(u, u')|}, \quad (7)$$

where NS_u denotes the set of most similar users of u, which is set to contain top 10% similar users in our experiments. Moreover, $\overline{n}^u$ is the average frequency of all tags in u's profile. After this process, we can estimate $P(t \mid u) = \frac{n_t^u}{\sum_t n_t^u}$ and use Equation 3 for user authority ranking.

Another open question is to find a proper dimension K for the latent space. In this paper, we leverage the Chib's method (Schmidt et al., 2009), which is performed by evaluating the marginal likelihood $P(B)$, for inferring the proper K.

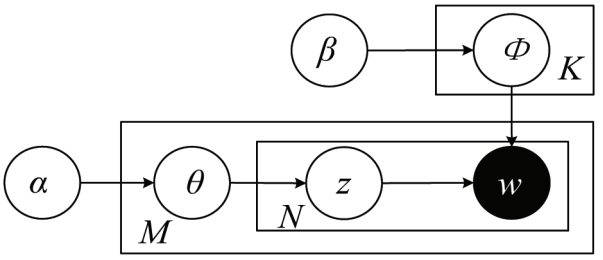
4.2. Authority Ranking by Topic Models

Although the collaborative filtering approach can efficiently resolve the problem of tag sparseness in individual users' profiles, another challenge of leveraging tags for authority ranking is tags often have latent semantic relationships. To be specific, some different tags may have same latent semantic meanings. For example, tags "PC", "iPad" and "Laptop" may all represent semantic meanings when user want to post questions about personal computers. Indeed, in the NMF process of our collaborative filtering approach, we can obtain a latent low dimensional space for tags. However, there is no prior knowledge in NMF process to explicitly model the latent space as semantic space for tags and thus the performance of finding experts may be impacted.

To capture these latent semantic relationships between tags, we propose to leverage topic models to rank user authority, which can capture the semantic relationship between tags. Topic models are widely used for text retrieval, which assume that there are several latent topics z for a corpus D and a document d in D can be represented as a bag of words $\{w_{d,i}\}$ which are generated by these latent topics. To be specific, although the tags "iPad", "PC" and "Laptop" are different words, they may all belong to the topic "Computer" and we can find the topic related experts.

Intuitively, if we take tags as words, take user profiles as documents we can directly

Figure 3. The graphical model of LDA



take advantage of topic models for inferring latent topics of tags. Thus, the Equation 3 can be calculated by:

$$P(u \mid q) \propto P(u)\sum_{z_j \in \theta}\prod_{t_i \in q} P(t_i \mid z_j)P(z_j \mid u). \tag{8}$$

Among several existing topic models, we use the widely used Latent Dirichlet Allocation model (LDA) (Blei et al., 2003) in our approach. According to LDA model as shown in Figure 3, a user profile d_i is generated as follows. First, a prior topic distribution θ is generated from a prior Dirichlet distribution α. Second, a prior category distribution ϕ is generated from a prior Dirichlet distribution β. Therefore, for the i-th tag t_i in u, the model first generates a topic z_j from θ_u and then generates t_i from ϕ_{z_j}.

The main requirement for our approach is to estimate the probability $P(z_j \mid u)$ and $P(t_i \mid z_j)$, which can directly obtained from LDA model training. The process of LDA model training is to learn the proper latent variables θ and ϕ to maximize the posterior distribution of the observed categories, i.e., $P(U \mid \alpha, \beta, \theta, \phi)$. In this paper, we choose a Markov chain Monte Carlo method, namely Gibbs sampling (Griffiths & Steyvers, 2004) to provide a relatively efficient process for training LDA model. This method begins with a random assignment of tags to topics for initializing the state of Markov chain. In the each following iteration of the chain, the method will re-estimate the conditional probability of assigning a tag to each topic, which is conditioned on the assignment of all other tags. Then a new assignment of tags to topics according to those conditional probabilities will be scored as a new state of Markov chain. Finally, after enough rounds of iteration, the assignment will converge, which means every tag is assigned a stable topic. After the model training, we can get the estimated value $\tilde{P}(t_i \mid z_j)$ and $\tilde{P}(z_j \mid u)$ as follows.

$$\tilde{P}(t_i \mid z_j) = \frac{n_j^{(t_i)} + \beta}{n_j^{(\cdot)} + \mid T \mid \beta}, \tag{9}$$

$$\tilde{P}(z_j \mid u) = \tilde{P}(z_j \mid d_u) = \frac{n_j^{(d_u)} + \alpha}{n_{\cdot}^{(d_u)} + K\alpha}, \tag{10}$$

where the $n_j^{(t_i)}$ indicates the number of times tag t_i is assigned to topic z_j, while $n_j^{(d_u)}$ indicates the number of times a tag from user profile d_u is assigned to topic z_j. $\mid T \mid$ indicates the number of unique tags, and K indicates the number of latent topics.

LDA model needs a predefined parameter K to indicate the number of latent topics. How to select an appropriate K for LDA is an open question. In terms of guaranteeing the performance of expert finding, in this paper we utilize the method proposed by Bao et al. (2010) to estimate K according to the performance of perplexity (Azzopardi et al., 2003; Blei et al., 2003).

Table 1. Details of experimental data

	Training Data	Test Data	Total Data
Num. of Q&A Paris	1,211,907	100,000	1,311,907
Num. of Answers	5,039,264	481,039	5,520,303
Num. of Unique Tags	111,925	13,486	115,925
Num. of Unique Users	263,236	44,384	274,896

5. EXPERIMENTS

In this section we provide an empirical evaluation for the performance of our tag-based expert finding approach on a large-scale real-world data set.

5.1. Experimental Data

We collected a large-scale real-world data set of Q&A pairs from a tag-based Chinese commercial Q&A service web site named Tianya Wenda (http://wenda.tianya.cn, http://wenda.google.com.hk) from Aug. 15, 2008 to Jun. 20, 2010. This data set contains more than 1.3 million Q&A pairs, 5.5 million answers, 4.3 million tagging records, which contains 115,925 unique tags, and 595 predefined question categories. The collected questions and answers were posted by 274,896 users. In the data set, all questions are resolved questions which contain a best answer voted by the question creator. Therefore, this data set contains few noise data such as questions posted by robots.

To evaluate our approach, we randomly select 100,000 Q&A pairs as the test data set and others as the training data set. Table 1 shows some details of our experimental data. Figure 4 (a) shows the distribution of tag number respect to the corresponding frequency in questions and Figure 4 (b) shows the distribution of user number respect to the number of answered questions in our data set. Both distributions roughly follow power law. Thus, we find that the uneven distribution of tags and users in KSCs is common. The long tail distributions of users also implicate the high rate of under-exploited expert users. Moreover, Figure 4 (c) also shows the distribution of user number with respect to the number of different tags in their profiles. From the figure we can observe that only a few users' profiles contain lots of unique tags while most of the users' profiles only contain few unique tags. This result also indicates the problem of tag sparseness in individual users' profiles.

5.2. Concept Clustering and Tag Extension

We extract concepts from the training data by letting the minimal support equal to 5, 10, and 20, respectively. Table 2 shows the results of concept clustering process. From the table we can find that with the increasing of min_sup the number of concepts will decrease dramatically. With these concepts, we can utilize Algorithm 1 for extending original tags. Figure 5 demonstrates the average increased tag number in the test data set with respect to varying extension thresholds R and min_sup.

How to select a proper min_sup for mining concepts and the extension number of R for extending original tags is an open question, thus we empirically study the performance of expert finding given different settings of parameters in Section 5.4.

5.3. Baseline Methods and Evaluation Metrics

To evaluate the performance of our two authority ranking approaches for expert finding, which namely **CNMF** (Context-aware Non-negative Matrix Factorization) and CLDA (Context-aware LDA), we select several baseline methods as follows.

Figure 4. Distribution of number of (a) tags in questions, (b) users in questions, and (c) user profiles with respect to tag number in our data set

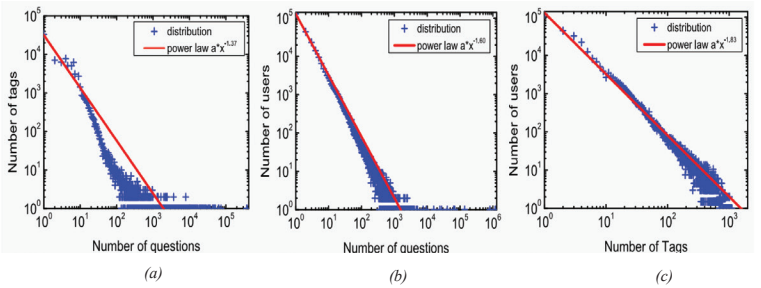
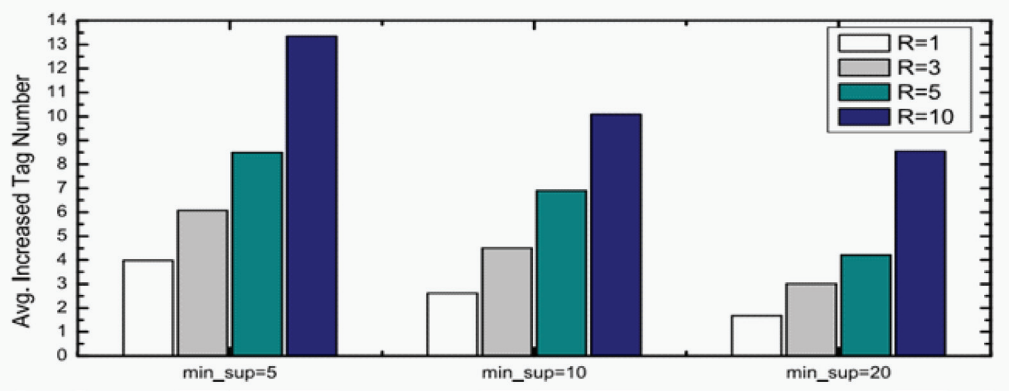


Table 2. Results of concept clustering

	min_sup=5	*min_sup=10*	*min_sup=20*
Num. of FPs	73,613	26,617	13,087
Num. of LMPs	49,239	18,281	9,126
Num. of Concepts	9,186	3,286	1,236
Avg. Length of Concepts	5.23	3.51	2.33

Figure 5. The average increased tag number in the test data set



- **LM**, which stands for language model without tag extension and topical analysis, and the authority ranking is based on Equation 3.
- **CLM**, which stands for context-aware language model, which extends original tags in LM approach.
- **NMF,** which stands for non-negative matrix factorization without tag extension and the authority ranking is based on Equations 3 ~ 7.
- **LDA,** which stands for LDA without tag extension and the authority ranking is based on Equations 8 ~ 10.

Besides these tag-based methods, we also compare our approach with category-based

methods. To be specific, in our data set, each question also contains a system-predefined question category, thus we can rank user authority for each category and find top K authoritative users for all the questions with corresponding categories. In our experiments, we select two well-known category based authority ranking approaches introduced in Zhang et al. (2007), which named ExertiseRank and HITS. Both of these two approaches are based on category link graph $G_c = \{V_c, E_c\}$, where node set V_c denotes the users have appeared in category c, and $e_{ij} \in E_c$ is an edge if user u_j has answered a question with category c for user u_i.

- **ExpertiseRank**, which is extended from PageRank. This algorithm does not only consider how many other users one helped, but also whom he/she helped. To be specific, a user who helps more authoritative users will be assigned a higher authority score.
- **HITS**, is an iterative approach which assigns two scores for each node in the category link graph, namely, hub score and authority score. A user with a higher hub score may be helped by more authoritative users and a user with a higher authority score may help more hub users.

In this paper, we use three metrics to evaluate the performance of expert finding.

- **Avg. P@10**, which means the average precision of top 10 expert finding results. To be specific, given a testing data set TS. $Avg.P@10 = \sum_{q \in TS} f(q,10) / |TS|$, where $f(q,10)$ is a binary function and it equals to 1 if one of the top 10 mined experts really answered the question q, and otherwise it equals to 0.
- **Avg. B@10**, which means the average precision of best answer. The calculation of Avg. B@10 is like Avg. P@10, but the binary function $f(q,10)$ equals to 1 if one of the top 10 mined experts really post a best answer for question q.
- **Mean Reciprocal Rank (MRR)**, which is computed by $\frac{1}{|TS|}\sum_{q \in TS} rank^{-1}{}_i$, where TS is the test data set and $rank_i$ is the rank of the first found expert in top 10 results who really answered the question q. If there is no such user has been found in top 10 results, we let $rank^{-1}{}_i = 0$.

5.4. Performance Comparison

We test all 100,000 questions in the test data set and empirically study the performance of expert finding when setting $min\_sup = 5, 10, 20$ and extension parameter $R = 1, 3, 5, 10$ in our experiments. In addition, the dimension number for non-negative matrix factorization and topic number for LDA are both set to be 100 in the training process, the two parameters α and β in LDA model are empirically set to be $50 / Z$ and 0.2 according to Heinrich (2004). Both our approaches and the baselines are implemented by C++ and the experiments are conducted on a $2.8GHZ \times 2$ Dub-Core CPU, 2G main memory PC.

We first test three context-aware approaches, namely CNMF, CLDA and CLM, with respect to different metrics and varying parameters min_sup and extension number R. From the results showed in Table 3, we observe that when given R with a big value and min_sup with a small value, the performance of expert finding will be impacted dramatically. It is because these settings will introduce more irrelevant tags as noise data in authority ranking. Moreover, with a small value of R and a big value of min_sup, the performance of expert finding will be limited, because there are only few of the concepts will be used for tag extension and the two approaches will be similar with NMF, LDA and LM.

Table 4 shows the average performance of expert finding by each baseline method with

Table 3. The performance of expert finding by CNMF, CLDA and CLM with varying parameters

min_sup=5	Avg. P@10			Avg. B@10			MRR		
	CNMF	CLDA	CLM	CNMF	CLDA	CLM	CNMF	CLDA	CLM
R=1	0.6104	0.6423	0.5443	0.3014	0.3456	0.2376	0.3962	0.4223	0.3398
R=3	0.6423	**0.6791**	0.5775	0.3321	**0.3747**	0.2598	0.4098	**0.4433**	0.3379
R=5	0.6301	0.6623	0.5893	0.3143	0.3596	0.2632	0.4003	0.4363	0.3619
R=10	0.5702	0.6112	0.5124	0.2323	0.2893	0.2247	0.3424	0.3899	0.3248
min_sup=10	Avg. P@10			Avg. B@10			MRR		
	CNMF	CLDA	CLM	CNMF	CLDA	CLM	CNMF	CLDA	CLM
R=1	0.6302	0.6798	0.5621	0.3503	0.3908	0.2493	0.3923	0.4392	0.3477
R=3	0.6623	0.7034	0.5977	0.3593	0.3955	0.2646	0.4104	0.4518	0.3555
R=5	0.6798	**0.7191**	0.6055	0.3621	**0.3997**	0.2715	0.4238	**0.4635**	0.3647
R=10	0.5943	0.6311	0.5294	0.2634	0.3122	0.2374	0.3742	0.4218	0.3396
min_sup=20	Avg. P@10			Avg. B@10			MRR		
	CNMF	CLDA	CLM	CNMF	CLDA	CLM	CNMF	CLDA	CLM
R=1	0.5823	0.6232	0.5246	0.2723	0.3029	0.2316	0.3743	0.4013	0.3436
R=3	0.6192	0.6556	0.5646	0.3198	0.3529	0.2519	0.4062	0.4353	0.3292
R=5	0.6423	**0.6716**	0.5961	0.3258	**0.3674**	0.2637	0.4192	**0.4416**	0.3592
R=10	0.6013	0.6393	0.5371	0.2972	0.3236	0.2446	0.3904	0.4292	0.3436

respect to different metrics. Specially, the CNMF-B, CLDA-B and CLM-B are the best performance of CNMF, CLDA and CLM in Table 3, the CNMF-W, CLDA-W and CLM-W are the corresponding worst performance of CNMF, CLDA and CLM. From this table we can see that our approaches CNMF and CLDA consistently outperform other baselines with respect to varying metrics on test data set, and CLDA outperforms CNMF slightly, which may because LDA can address the problem of semantic tags. Moreover, we observe that the context-aware tag extension and topic based authority ranking can both improve the performance of basic LM method. We also observed that the performance of tag based methods consistently outperform the category based methods, which implies the user-generated tags are more proper for expert finding than question categories.

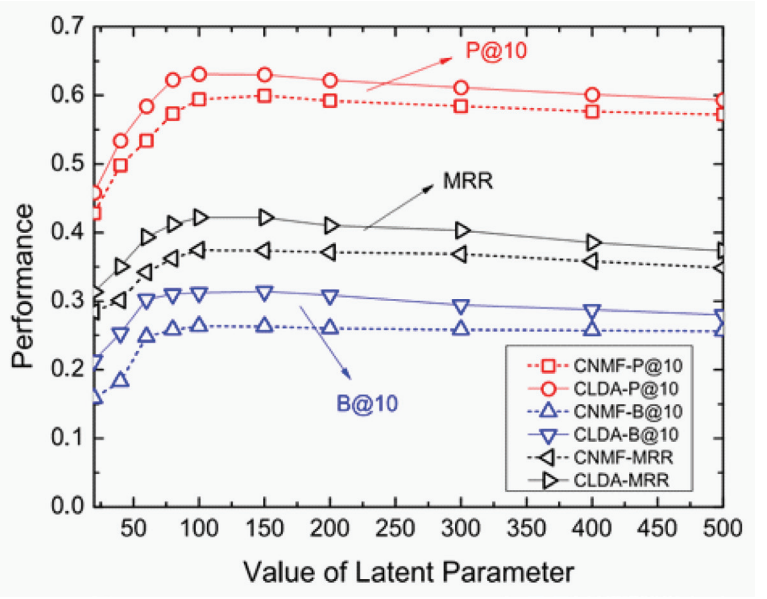
5.3. Robustness Analysis

Both of NMF and LDA training need a predefined parameter K to decide the latent dimension number or latent topic number. Indeed we can learn a proper value for this parameter according to some specific approaches, such as marginal likelihood or perplexity introduced in Section 4. However, we still need to analyze the robustness of expert finding when given varying parameters. Figure 6 shows the *Avg. P@10*, *Avg. B@10* and *MRR* of CNMF and CLDA with respect to 10 different parameter K (K =*20, 40, 60, 80, 100, 150, 200, 300, 400, 500*). To be specific, in this experiment, we set both min_sup and extension number equal to 10. From the figure we can observe that the expert finding performance will be impacted dramatically when given a small K and the performance becomes stable with the increasing of K. It may be because the small K may indicate strong relationships between

Table 4. The performance comparison of expert finding

	Avg. P@10	Avg. B@10	MRR
CNMF-B	0.6798	0.3621	0.4192
CNMF-W	0.5702	0.2323	0.3424
CLDA -B	**0.7191**	**0.3997**	**0.4635**
CLDA -W	0.6112	0.2893	0.3899
CLM-B	0.6055	0.2715	0.3647
CLM-W	0.5124	0.2247	0.3248
LM	0.5012	0.2195	0.3174
NMF	0.5623	0.2033	0.3253
LDA	0.6073	0.2749	0.3696
ExpertiseRank	0.4192	0.1833	0.2547
HITS	0.3924	0.1724	0.2396

Figure 6. The robustness analysis with different parameter K



tags, which may introduce more noise information in authority ranking.

6. CONCLUSION

In this paper, we studied the problem of expert finding by taking advantage of user-generated tags. In our approach, we exploit context information of question tags to infer latent information needs of question creator and leveraging the topic distribution of tags to rank user authority. Specifically, we first extracted tag concepts from historical Q&A pairs to capture the context of tags and select the most relevant concepts by

user profile for tag extension. Then, we proposed a probabilistic framework for ranking user authority. Based on this framework, to address the problem of tag sparseness, we developed two different authority ranking approaches, which leverage NMF collaborative filtering approach and LDA topic model, respectively. Finally, we showed the effectiveness of the proposed approach with multiple baseline methods by the experiments on a large-scale real-world Q&A data set. The results clearly indicate that when the context-aware tag extension is combined with our proposed authority ranking methods, the tag based expert finding approach can achieve the best performance.

7. ACKNOWLEDGMENTS

The work described in this paper was supported by grants from Natural Science Foundation of China (Grant No. 61073110), Key Program of National Natural Science Foundation of China (Grant No. 60933013), National Major Special Science & Technology Projects (Grant No. 2011ZX04016-071), Natural Science Foundation of China major program (Grant No. 71090401/71090400) and Nokia.

REFERENCES

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 487-499).

Azzopardi, L., Girolami, M., & Risjbergen, K. V. (2003). Investigating the relationship between language model perplexity and IR precision-recall measures. In *Proceedings of the 26th International Conference on Research and Development in Information Retrieval* (pp. 369-370).

Balog, K., Azzopardi, L., & Rijke, M. D. (2006). Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th International Conference on Research and Development in Information Retrieval* (pp. 43-50).

Balog, K., Azzopardi, L., & Rijke, M. D. (2009). A language modeling framework for expert ðnding. *Journal of Information Processing and Management*, *45*, 1–19. doi:10.1016/j.ipm.2008.06.003

Bao, T., Cao, H., Chen, E., Tian, J., & Xiong, H. (2010). An unsupervised approach to modeling personalized contexts of mobile users. In *Proceedings of the 10th International Conference on Data Mining* (pp. 38-47).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Lantent dirichlet allocation. *Journal of Machine Learning Research*, 993–1022.

Bouguessa, M., Dumoulin, B., & Wang, S. (2008). Identifying authoritative actors in question-answering forums: the case of Yahoo! Answers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 866-874).

Cong, G., Wang, L., Lin, C. Y., Song, Y. L., & Sun, Y. (2008). Finding question-answer pairs from online forums. In *Proceedings of the 31st International Conference on Research and Development in Information Retrie*val (pp. 467-474).

Feng, D., Shaw, E., & Hovy, E. (2006). Mining and assessing discussions on the web through speech act analysis. In *Proceedings of the IWorkshop on Web Content Mining with Human Language Technologies.*

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 5228–5235. doi:10.1073/pnas.0307752101

Han, J., Pei, J., & Yin, Y. (2004). Mining frequent patterns without candidate generation. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 1-12).

Heinrich, G. (2004). *Parameter estimation for text analysis*. Leipzig, Germany: University of Leipzig.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval* (pp. 50-57).

Jurczyk, P., & Agichtein, E. (2007). Discovering authorities in question answer communities by using link analysis. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management* (pp. 919-922).

Kao, W. C., Liu, D. R., & Wang, S. W. (2010). Expert finding in question-answering websites: a novel hybrid approach. In *Proceedings of the ACM Symposium on Applied Computing* (pp. 867-871).

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 668–677.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791. doi:10.1038/44565

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Proceedings of the Advances in Neural Information Processing Systems* (pp. 556-562).

Liu, X., Croft, W. B., & Koll, M. (2005). Finding experts in community-based question-answering services. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (pp. 315-316).

Lu, Y., Quan, X., Ni, X., Liu, W., & Xu, Y. (2009). Latent link analysis for expert finding in user-interactive question answering services. In *Proceedings of the 5th International Conference on Semantics, Knowledge and Grid* (pp. 54-59).

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (1999). Text classification from labeled and unlabeled documents using em. *Journal of Machine Learning*, *39*, 103–134. doi:10.1023/A:1007692713085

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web (Tech. Rep.)*. Stanford, CA: Stanford University.

Pei, J., Han, J., Mortazavi-asl, B., Pinto, H., Chen, Q., Dayal, U., & Hsu, M. C. (2001). Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings of the 17th International Conference on Data Engineering* (pp. 215-214).

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the Conference on Computer Supported Cooperative Work* (pp. 175-186).

Schmidt, M. N., Winther, O., & Hansen, L. K. (2009). Bayesian non-negative matrix factorization. In *Proceedings of the International Conference on Independent Component Analysis and Signal Separation* (pp. 540-547).

Yeh, T., & Darrell, T. (2008). Multimodal question answering for mobile devices. In *Proceedings of the 13th International Conference on Intelligent User Interfaces* (pp. 405-408).

Zhang, J., Ackerman, M. S., & Adamic, L. (2007). Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 221-230).

Zhang, J., Tang, J., & Li, J. (2007). Expert finding in a social network. In *Proceedings of the 12th International Conference on Database Systems for Advanced Applications* (pp. 1066-1069).

Zhang, J., Tang, J., Liu, L., & Li, J. (2008). A mixture model for expert finding. In *Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 466-478).