Finding Experts in Tag Based Knowledge Sharing Communities

Hengshu Zhu[1,2], Enhong Chen[1], and Huanhuan Cao[2]

[1] University of Science and Technology of China
[2] Nokia Research Center
{zhs,cheneh}@ustc.edu.cn, happia.cao@nokia.com

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 under-explored. To this end, in this paper, we develop 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. Then, a topic model based approach is applied for capturing the semantic relationship between tags and then taking advantage of them for ranking user authority. We evaluate the proposed framework for expert finding on a large-scale real-world data set collected from a tag based Chinese commercial Q&A web site. Experimental results clearly show that the proposed method outperforms several baseline methods with a significant margin.

Keywords: Expert finding, knowledge sharing communities, question answering, user-generated tags, topic models.

1 Introduction

Recent years 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 and more people would like to use these KSCs as platforms for problem resolving. Researchers have found that Q&A content is usually the largest part of content in KSCs [9,10]. 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 [5,18].

H. Xiong and W.B. Lee (Eds.): KSEM 2011, LNAI 7091, pp. 183–195, 2011.

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. Moreover, some of the novel KSCs are based on multimedia content and the textual information contained in questions are often not rich enough for building language models [23]. Therefore, most of the state-of-the-art works leverage question categories as query inputs to find experts [8,15,16]. 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 N8?". As a result, the conventional category based expert finding approaches may have poor performances for a multiple-category question.

A recent development trend of KSCs is to allow users to add text tags for their questions, such as Tianya Wenda[1] and Douban[2]. 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 "N8", "Mobile Market" and "Where" for 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 challenge which is still under-explored.

* Example 1. *Joy posts a question about Sony game player "Play Station" and adds tag "PS" to annotate the question. However, the tag "PS" may be referred to the Adobe software "Photo Shop".*

* Example 2. *Kate wants to buy a mobile phone and posts a question with tags "Mobile Phone", "Market". However, the latent information needs for Kate is actually the "Discount" and "Trustable store".*

* Example 3. *Joy posts a question about computer with tag "PC", and Kate may add tag "Laptop" for the same question. The different tags may represent same meaning.*

Inspired by the above observations, in this paper, we propose a novel tag based approach for expert finding by inferring the users' latent information needs and uncovering the semantic relationship between tags. Specifically, we first introduce an effective tag expansion method which enrich the original question tags with the question creator's latent information needs. The latent information needs are modeled by the tag concepts, which are referred as ***context*** for the original

[1] `http://wenda.tianya.cn`

[2] `http://www.douban.com`

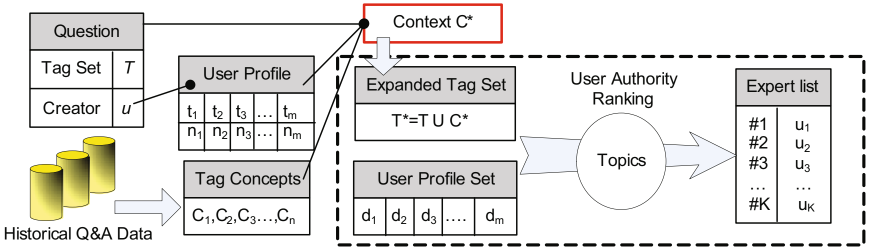


Fig. 1. The framework of our tag based expert finding approach

question tags, and the concepts are extracted from the users' historical Q&A data and the question creator's profile. Then, we propose a topic based probabilistic model to rank user authority in KSCs, which can model tags as topic distributions and thus can capture the latent semantic relationship between tags. The framework of our approach is demonstrated in Figure 1. Finally, we perform extensive experiments on a large-scale real-world data set collected from a major Chinese commercial Q&A web sites. Experimental results clearly show that the proposed method outperforms several baseline methods with a significant margin.

Overview. The remainder of this paper is organized as follows. Section 2 provides a brief overview of related work. Section 3 and Section 4 show the details of the tag based expert finding approach. In Section 5, we give the experimental results. Finally, Section 6 concludes the work.

2 Related Work

The problem of expert finding has been well studied for years. In general, the previous works of expert finding can be grouped into two categories.

In the first category, researchers utilize content based approaches to finding experts for answer-lack questions. For example, Balog *et al.* [5] used conventional language models for finding experts in enterprise corpora. Liu *et al.* [18] have investigated finding experts in community based Q&A services by leveraging user profiles into language models. Zhang et al. [26] proposed a mixture model based on Probabilistic Latent Semantic Analysis (PLSA), which can discover semantically related experts for a given question. Although these approach 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 of expert finding are focus on finding the most authoritative users for a specific question category. These category based approaches often leverage link analysis algorithms on the 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 *et al.* [15] formulated a graph structure in Q&A communities and proposed a variation of the HITS [17] algorithm for predicting authoritative users in Yahoo! Answers. Zhang *et al.* [24] investigated various authority ranking algorithms in the Java forum and also proposed a PageRank [21] like algorithm named "ExpertiseRank" to find experts. Zhang *et al.* [25] proposed a propagation-based approach for finding experts in co-author social networks which take into account user profiles and Lu *et al.* [19] extended it with latent link analysis and language model. However, these category based approaches have poor performance when 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 topic models for ranking user authority by taking into account the latent semantic relationships between tags. Indeed, topic models are widely used in text retrieval and information extraction. Typical topic models include the Mixture Unigram (MU) [20], the Probabilistic Latent Semantic Indexing (PLSI) [14], and the Latent Dirichlet Allocation (LDA) [7]. 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 Expansion

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 question content. 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 latent information needs of question creator, we take into account of the context of question tags by taking advantage of historical Q&A pairs to expand the original tags. To be specific, we firstly summarize user tags by concepts for identifying the latent context of tags. Then we expand 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 often appear in the same context if they usually co-occur in same tag sets. A set of tags which often co-occurs in the same context is referred as a ***tag concept***. Intuitively, given a user-generated tag, the tags in the same tag concept can be used as candidate expansions 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 he like play games, thus we can expand tags "Sony", "Game" with the original tag "PS".

3.1 Identifying Tag Concepts

To build tag concepts and infer latent information need of question creator, in this paper, we take advantage of the frequent pattern mining approach to capturing tag co-occurrences. When given a transaction database TDB, where

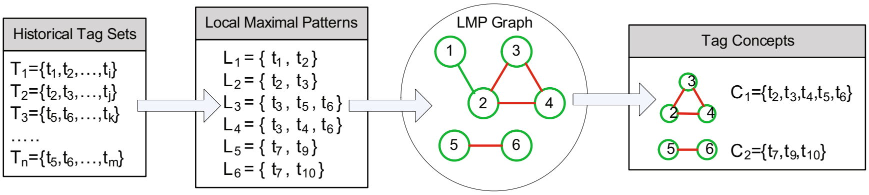


Fig. 2. The generation of local maximal pattern graph

is a set of transactions $\{T\}$, 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 firstly 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 [3], FP Growth [12] and PrefixSpan [22]. 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.

We define a frequent pattern with no 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. 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 subgraph 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 subgraph with more than two nodes or an isolated connected subgraph with two nodes. And there is no other concept $c^{'}$ makes $c^{'} \subset c_i$.*

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

3.2 Tag Expansion by Concepts

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

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 based 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:

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

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 $t \in c$.

```
Input: concept set $C$, original question tag set $T$ and creator's profile $d$
Output: the expanded tag set $T^*$
for $c \in C$ do
    if $c \cap T \cap d == \emptyset$ then
        $Rel(c, d, T) = 0$;
    end
    else
        Compute $Rel(c, d, T)$ for $c$;
    end
end
descending rank $c \in C$ according to $Rel(c, d, T)$;
for $1 \leq i \leq K$ do
    if $Rel(c, d, T) \neq 0$ then
        $T^* = T \cup c_i$;
    end
end
return $T^*$;
```

Algorithm 1. Context-aware tag expansion

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

4 Topic Model 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 expanded tags. A challenge here is how to capture the latent semantic relationship between tags, such as "PC", "iPad" and "Laptop". To this end, we propose to leverage topic models to rank user authority, which can capture the semantic relationship between tags.

To formalize the authority ranking task, we use $P(u|q)$ to denote the probability of an user u being an expert for the given question q. Using Bayes formula we have the following equations:

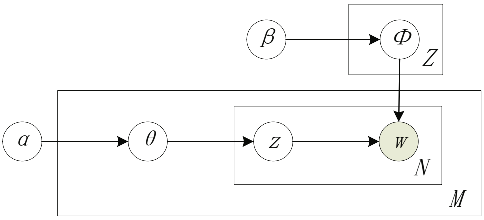


Fig. 3. The graphical model of LDA

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

According to the equations above, the main procedure for expert ranking is to calculate the conditional probability $P(q|u)$ 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 independent given a user and using tags $T = \{t_1, t_2, ..t_n\}$ to represent question q, we have the following equation:

$$P(u|q) \propto P(u) \prod_{t_i \in q} P(t_i|u). \tag{3}$$

Here if we calculate the probability $P(t_i|u)$ without taking into account the latent semantic relationship between tags, it will be equal to zero if none of the tags appears in the user profile of an candidate expert u. Therefore, here we leverage the topic models to calculate $P(t_i|u)$.

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

$$P(u|q) \propto P(u) \sum_{z_j \in \theta} \prod_{t_i \in q} P(t_i|z_j)P(z_j|u). \tag{4}$$

Among several existing topic models, we use the widely used Latent Dirichlet Allocation model (LDA) [7] in our approach. According to LDA model as shown in Figure 3, a user profile d_i is generated as follows. Firstly, a prior topic distribution θ is generated from a prior Dirichlet distribution α. Secondly, 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 $P(z_j|u)$ and $P(t_i|z_j)$, which can directly obtained from LDA. In this paper, we use Gibbs sampling method [11] to estimate the two probabilities. After several rounds of Gibbs sampling, we can get the estimated value $\widetilde{P}(t_i|z_j)$ and $\widetilde{P}(z_j|u)$ as follows.

$$\widetilde{P}(t_i|z_j) = \frac{n_j^{(t_i)} + \beta}{n_j^{(\cdot)} + |T|\beta}, \qquad \widetilde{P}(z_j|u) = \frac{n_j^{(u)} + \alpha}{n_\cdot^{(u)} + |Z|\alpha}, \tag{5}$$

where the $n_j^{(t_i)}$ is the number of times tag t_i has been assigned to topic z_j, while $n_j^{(u)}$ is the number of times a tag from user u that has been assigned to topic z_j. The $|T|$ is the number of tags from u, and $|Z|$ is the number of latent topics.

LDA model needs a predefined parameter Z to indicate the number of latent topics. How to select an appropriate Z 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 [6] to estimate Z according to the performance of perplexity [4,7].

After estimating the probabilities $\widetilde{P}(z_j|u)$ and $\widetilde{P}(t_i|z_j)$ we can rank user authority according to Equation 4. Then, the top K ranked authoritative users will be regarded as the experts for the given question with tags.

5 Experiments

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

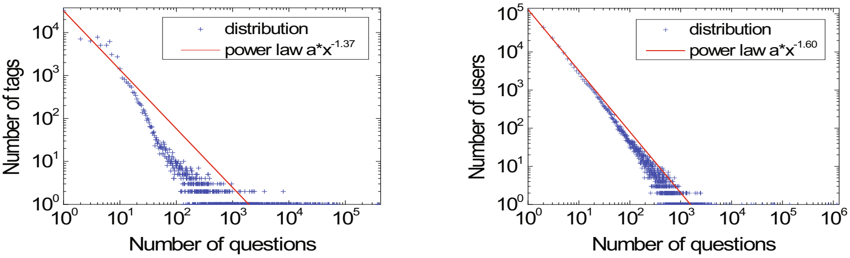
5.1 Experimental Data

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 [1,2] 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, these 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 training data set. The 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.

Table 1. Details of Experimental Data

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



(a) (b)

Fig. 4. Distribution of (a) tags, (b) users 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.

5.2 Concept Clustering and Tag Expansion

We extract concepts from using the training data with minimal support equals 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 expanding original tags. Figure 5 demonstrates the average increased tag number in the test data set with respect to varying expansion thresholds K and *min_sup*.

Table 2. Results of Concept Mining

	$min\_sup = 5$	$min\_sup = 10$	$min\_sup = 20$
Number of Frequent Patterns	73,613	26,617	13,087
Number of LMPs	49,239	18,281	9,126
Number of Concepts	9,186	3,286	1,236
Avg. Length of Concepts	5.23	3.51	2.33

How to select the proper *min_sup* for mining concepts and the number of K for expanding original tags is an open question, thus we empirically study the performance of expert finding when given different setting of parameters in Section 5.4.

5.3 Baseline Methods and Evaluation Metrics

To evaluate our **c**ontext and **t**opic based **a**uthority **r**anking (CTAR) approach, we select several baseline methods as follows. The first method is the basic language model without tag expansion and topical analysis (LM), and the authority ranking is based on Equation 3. The second method is context-aware language

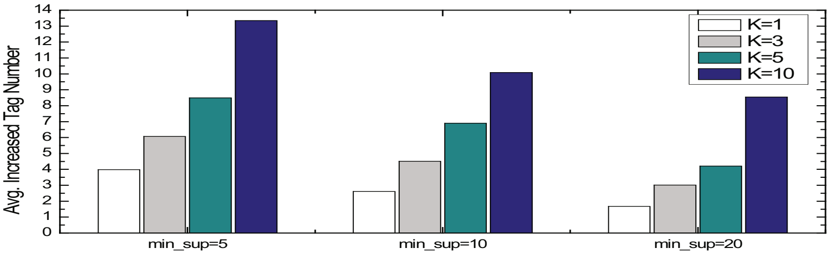


Fig. 5. The average increased tag number in the test data set

model (CLM), which expands original tags in LM approach. The third method is LDA topic model (LDA) without tag expansion and authority ranking based on Equation 4. Besides these tag based methods, we also compare our approach with two well-known category based authority ranking approaches introduced in [24], which are ExpertiseRank and HITS.

In this paper, we choose three metrics to evaluate the performance of expert finding. The first is **Avg. P@10**, which means the average precision of top 10 expert finding results. To be specific, given a testing data set TS, $Avg.P@10 = \frac{\sum_{q\in TS} f(q,10)}{|TS|}$, 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. The second metric is **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. The last metric is **Mean Reciprocal Rank (MRR)**, it is computed by $\frac{1}{|TS|}\sum_{q\in TS}\frac{1}{rank_i}$, 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 $\frac{1}{rank_i} = 0$.

5.4 Performance Comparison

We test all 100,000 questions in test data set, we empirically study the performance of expert finding when set $min\_sup = 5, 10, 20$ and expansion parameter $K = 1, 3, 5, 10$ in our experiments. In addition, according to the perplexity introduced in [6], the number of topics Z is set to be 100 for the data set, the two parameters α and β in LDA model are empirically set to be $50/Z$ and 0.2 according to [13].

We first test two context-aware approaches, namely CTAR and CLM, with respect to different metrics and varying parameters min_sup and K. From the results showed in Table 3, we observe that when given K 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 K and a big value of min_sup, the performance of expert finding will be limited, because

Table 3. The performance of expert finding by CTAR and CLM

min_sup = 5	Avg. P@10		Avg. B@10		MRR	
	CTAR	CLM	CTAR	CLM	CTAR	CLM
K=1	0.6423	0.5443	0.3456	0.2376	0.4223	0.3398
K=3	**0.6791**	0.5775	**0.3747**	0.2598	**0.4433**	0.3379
K=5	0.6623	0.5893	0.3596	0.2632	0.4363	0.3619
K=10	0.6112	0.5124	0.2893	0.2247	0.3899	0.3248
min_sup = 10	Avg. P@10		Avg. B@10		MRR	
	CTAR	CLM	CTAR	CLM	CTAR	CLM
K=1	0.6798	0.5621	0.3908	0.2493	0.4392	0.3477
K=3	0.7034	0.5977	0.3955	0.2646	0.4518	0.3555
K=5	**0.7191**	0.6055	**0.3997**	0.2715	**0.4635**	0.3647
K=10	0.6311	0.5294	0.3122	0.2374	0.4218	0.3396
min_sup = 20	Avg. P@10		Avg. B@10		MRR	
	CTAR	CLM	CTAR	CLM	CTAR	CLM
K=1	0.6232	0.5246	0.3029	0.2316	0.4013	0.3436
K=3	0.6556	0.5646	0.3529	0.2519	0.4353	0.3292
K=5	**0.6716**	0.5961	**0.3674**	0.2637	**0.4416**	0.3592
K=10	0.6393	0.5371	0.3236	0.2446	0.4292	0.3436

Table 4. The performance comparison of expert finding

	Avg. P@10	Avg. B@10	MRR
CTAR-B	0.7191	0.3997	0.4635
CTAR-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
LDA	0.6073	0.2749	0.3696
ExpertiseRank	0.4192	0.1833	0.2547
HITS	0.3924	0.1724	0.2396

there are only few of the concepts will be used for tag expansion and the two approaches will be similar with LDA and LM.

Table 4 shows the average performance of expert finding by each baseline method with respect to different metrics. Specially, the CTAR-B and CLM-B are the best performance of CTAR and CLM in Table 3, the CTAR-W and CLM-W are the corresponding worst performance of CTAR and CLM. From this table we can see that our approach CTAR consistently outperforms other baselines with respect to varying metrics on test data set. Moreover, we observe that the context-aware tag expansion and topic based authority ranking can both improve the performance of basic LM method. We also observe 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.

6 Concluding Remarks

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 firstly 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 expansion. Then, we developed a topic model based approach for uncovering the latent relationship between tags and authoritative users, and thus could rank user authority more accurately. 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 expansion is combined with topic model based authority ranking method, the tag based expert finding approach can achieve the best performance.

Acknowledgement. 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), Research Fund for the Doctoral Program of Higher Education of China (20093402110017) and Nokia.

References

1. http://wenda.google.com.hk/
2. http://wenda.tianya.cn/
3. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: VLDB 1994 (1994)
4. Azzopardi, L., Girolami, M., Risjbergen, K.V.: Investigating the relationship between language model perplexity and ir precision-recall measures. In: SIGIR 2003 (2003)
5. Balog, K., Azzopardi, L., Rijke, M.D.: Formal models for expert finding in enterprise corpora. Research and Development in Information Retrieval
6. Bao, T., Cao, H., Chen, E., Tian, J., Xiong, H.: An unsupervised approach to modeling personalized contexts of mobile users. In: ICDM 2010 (2010)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Lantent dirichlet allocation. Journal of Machine Learning Research
8. Bouguessa, M., Dumoulin, B., Wang, S.: Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In: KDD 2008 (2008)
9. Cong, G., Wang, L., Lin, C.-Y., Song, Y.-I., Sun, Y.: Finding question-answer pairs from online forums. In: SIGIR 2008 (2008)
10. Feng, D., Shaw, E., Hovy, E.: Mining and assessing discussions on the web through speech act analysis. In: ISWC 2006 Workshop on WCMHLT (2006)

11. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of National Academy of Science of the USA
12. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. SIGMOD Rec.
13. Heinrich, G.: Parameter estimation for text analysis. Technical report, University of Lipzig
14. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR 1999 (1999)
15. Jurczyk, P., Agichtein, E.: Discovering authorities in question answer communities by using link analysis. In: CIKM 2007 (2007)
16. Kao, W.-C., Liu, D.-R., Wang, S.-W.: Expert finding in question-answering websites: a novel hybrid approach. In: SAC 2010 (2010)
17. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM
18. Liu, X., Croft, W.B., Koll, M.: Finding experts in community-based question-answering services. In: CIKM 2005 (2005)
19. Lu, Y., Quan, X., Ni, X., Liu, W., Xu, Y.: Latent link analysis for expert finding in user-interactive question answering services. In: SKG 2009 (2009)
20. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using em. In: Machine Learning
21. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Stanford Digital Library Technical Report
22. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Chun Hsu, M.: Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: ICDE 2001 (2001)
23. Yeh, T., Darrell, T.: Multimodal question answering for mobile devices. In: IUI 2008 (2008)
24. Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: WWW 2007 (2007)
25. Zhang, J., Tang, J., Li, J.: Expert Finding in a Social Network. In: Kotagiri, R., Radha Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, vol. 4443, pp. 1066–1069. Springer, Heidelberg (2007)
26. Zhang, J., Tang, J., Liu, L., Li, J.: A Mixture Model for Expert Finding. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 466–478. Springer, Heidelberg (2008)