World Wide Web (2014) 17:1081–1107
DOI 10.1007/s11280-013-0217-5

Ranking user authority with relevant knowledge categories for expert finding

Hengshu Zhu · Enhong Chen · Hui Xiong · Huanhuan Cao · Jilei Tian

Received: 6 September 2012 / Revised: 7 March 2013 /
Accepted: 17 April 2013 / Published online: 28 April 2013

Abstract The problem of expert finding targets on identifying experts with special skills or knowledge for some particular knowledge categories, i.e. knowledge domains, by ranking user authority. In recent years, this problem has become increasingly important with the popularity of knowledge sharing social networks. While many previous studies have examined authority ranking for expert finding, they have a focus on leveraging only the information in the target category for expert finding. It is not clear how to exploit the information in the relevant categories of a target category for improving the quality of authority ranking. To that end, in this paper, we propose an expert finding framework based on the authority information in the target category as well as the relevant categories. Along this line, we develop a scalable method for measuring the relevancies between categories through topic models,

This is a substantially extended and revised version of [35], which appears in Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM2011).

H. Zhu · E. Chen (✉)
School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230026, China
e-mail: cheneh@ustc.edu.cn

H. Zhu
e-mail: zhs@mail.ustc.edu.cn

H. Cao · J. Tian
Nokia, Beijing 100010, China

H. Cao
e-mail: happia.cao@nokia.com

J. Tian
e-mail: jilei.tian@nokia.com

H. Xiong
Management Science and Information Systems Department, Rutgers Business School, Rutgers University, Newark, NJ 07102, USA
e-mail: hxiong@rutgers.edu

which takes consideration of both content and user interaction based category similarities. Also, we provide a topical link analysis approach, which is multiple-category-sensitive, for ranking user authority by considering the information in both the target category and the relevant categories. Finally, in terms of validation, we evaluate the proposed expert finding framework in two large-scale real-world data sets collected from two major commercial Question Answering (Q&A) web sites. The results show that the proposed method outperforms the baseline methods with a significant margin.

Keywords Authority ranking · Expert finding · Category relevancy · Link analysis · Question answering

1 Introduction

Recent years have witnessed increasing interests in research on facilitating knowledge propagation and putting crowd wisdom to work through knowledge sharing social networks, such as online forums and Question Answering (Q&A) communities. A critical challenge along this line is how to find experts—a group of authoritative users with special skills or knowledge for a specific ***knowledge category***, i.e. knowledge domain. Indeed, the problem of expert finding has attracted a lot of attention in the literature and a central issue of expert finding is how to perform effective authority ranking.

Some traditional works for expert finding are based on language models [2, 3, 20, 36, 37]. In these works, researchers can leverage discriminative or generative models to rank user authority through the content distributions in users' historical records. However, the language model based authority ranking approaches do not take into account the user relationships and their interactions, which become increasingly important in social networks. Moreover, a recent trend in knowledge sharing social networks is to allow users to share multimedia based knowledge, such as Youtube[1] and Flickr,[2] where the textual information is not rich enough for building language models. Therefore, these years, most of the state-of-the-art techniques of authority ranking for expert finding take advantage of link analysis methods, such as HITS [14, 16], PageRank [27], ExpertiseRank [33] and some social network based propagation algorithms [22, 34]. These works can rank user authority based on the link graphs which consist of users according to their social interactions and relationships. However, when performing authority ranking for expert finding, these existing works only take the information in the target category into consideration. Indeed, a target category usually has some very relevant categories. The information in these relevant categories might be exploitable for improving the performance of authority ranking for the target category.

A motivating example Suppose that Kate is a junior student and she posts a question "How to learn data mining related algorithms?" with a category label "Data Mining"

[1] http://www.youtube.com

[2] http://www.flickr.com

in a Q&A web site. Meanwhile, Sam is an authoritative user of the web site on data mining related knowledge and thus he is a good candidate to answer Kate's question. However, since Sam typically answered data mining related questions in the relevant categories "Database" and "Artificial Intelligence" and answered relatively few questions in the "Data Mining" category, his authority would be ranked lower than the users who typically answer questions in the "Data Mining" category. As a result, Sam would not be recommended to Kate by most existing expert finding approaches as an expert for the "Data Mining" category. In this case, it would be inappropriate to neglect Sam's expertise in the relevant categories "Database" and "Artificial Intelligence" when ranking the authority in the "Data Mining" category due to the high relevancy between these categories.

To address the above challenge, in this paper, we propose to exploit the information in both target and relevant knowledge categories for improving the performance of link analysis based authority ranking. The first task along this line is to accurately measure category relevancies. To this end, we propose to exploit both content based category similarity and user interaction based category similarity for inferring the relevancies between different categories. To be specific, we first leverage topic models to build latent topic space for each category and measure their similarities by normalized Kullback Leibler (KL) divergence. Then, based on the user historical interaction logs, we also propose to measure the user interaction based category similarity through topic models. At last, we combine both content and user interaction based similarities to measure the final category relevancy.

In addition, we develop a topical link analysis approach, which is based on the Topical Random Surfer model [24, 25] and multiple-category-sensitive, to collectively exploit the information in both target and relevant categories for authority ranking. For example, in the motivating example, we first identify the relevant categories for "Data Mining", such as "Database" and "Artificial Intelligence". Then, a link graph consisting of users appearing in both "Data Mining" and other relevant categories will be built. Therefore, the user authority can be better ranked in the resulted link graph. The example in Figure 1 illustrates how the proposed approach differs from the traditional authority ranking approaches.

Finally, we perform extensive experiments on two large-scale real-world data sets collected from two major commercial Q&A web sites. The results demonstrate the efficiency and effectiveness of the proposed approach. Specifically, our contributions in this work can be summarized as follows:

- We propose to leverage relevant categories in authority ranking for expert finding and establish a corresponding novel framework.
- We propose an unsupervised method for inferring category relevancy based on both content similarity and user interaction similarity.
- We develop an extended category link graph and a topical link analysis approach for ranking user authority in the proposed framework.
- We demonstrate an overall experimental comparison between our approach and lots of well-known benchmarks on two large-scale real-world date sets, which indicates some inspiring conclusions.

Overview The remainder of this paper is organized as follows. In Section 2, we formulate the problem of category relevancy based authority ranking. Then, Section 3 shows how to measure the relevance between the target category and

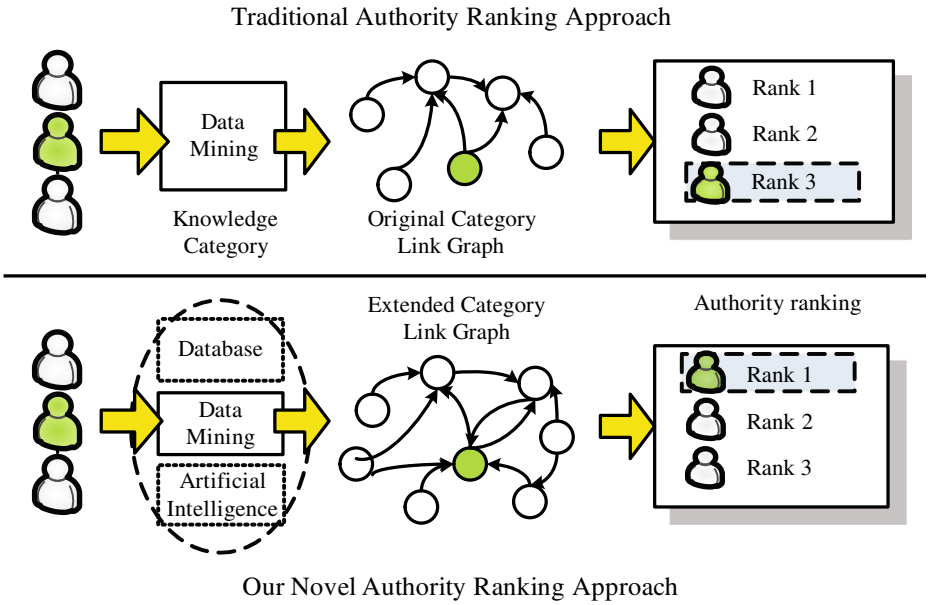

Figure 1 The traditional authority ranking approach versus our novel approach in the motivating example, where "Data Mining" is the target category while "Database" and "Artificial Intelligence" are the relevant categories

the relevant categories through topic models. In Section 4, we introduce a topical link analysis approach, which is multiple-category-sensitive, for authority ranking using the information in both the target category and relevant categories. Section 5 presents the experimental results. Section 6 provides a brief overview of related works. Finally, in Section 7, we draw the conclusions and future work.

2 Problem statement

In this section, we first introduce the traditional authority ranking problem, and then formally define the problem of category relevancy based authority ranking. In this paper, we define that a ***category*** is a label to represent a specific knowledge domain. The category based classification is wildly used in online knowledge sharing social networks.

Traditional authority ranking Given a category set $C = \{c_1, c_2, ..., c_n\}$ and a user set $U = \{u_1, u_2, ..., u_m\}$, the category link graph G_c ($c \in C$) for a given knowledge sharing social network S is denoted as $G_c = (V_c, E_c, W_c)$, where

- $V_c = \{u_i\}$ is a set of user nodes, where each user in V_c made or replied[3] the content which is labeled with category c in S.

[3]The way of replying is different in different types of knowledge sharing social networks. For example, in online forums, it refers to commenting the threads posted by other users, while in Q&A communities, it refers to answering the posted questions by other users.

- $E_c = \{e_{ij}\}$ is a set of directed edges, where e_{ij} indicates that user u_j replied the content which is labeled with category c and made by user u_i in S.
- $W_c = \{w_{ij}^c\}$ is a set of weights for the edges in E_c, where w_{ij}^c indicates the frequency that user u_j replied the content which are labeled with category c and posted by user u_i in S.

Given a knowledge sharing social network S, the task of the traditional authority ranking for category c is to find top K authoritative users from G_c. Therefore, only the information in target categories are taken into account. In contrast, we introduce a new approach for authority ranking by exploiting the information in both target and relevant categories. Next, we present some notations as follows.

Definition 1 (Extended category set) An **extended category set** $\Upsilon_c = \{c\} \cup R_c$, where R_c denotes the set of relevant categories of category c. For each $c^{'} \in R_c$, we have category relevancy $Rel(c^{'}||c) > \tau$.

According to the above definition, we further propose an ***extended category link graph*** built by both relevant and target categories as follows.

Definition 2 (Extended category link graph) An **extended category link graph** $G_{\Upsilon_c} = (V_{\Upsilon_c}, E_{\Upsilon_c}, W_{\Upsilon_c})$ is the extension of the category link graph G_c, where

- $V_{\Upsilon_c} = \bigcup_{c' \in \Upsilon_c} V_{c'}$ *is the corresponding user node set.*
- $E_{\Upsilon_c} = \bigcup_{c' \in \Upsilon_c} E_{c'}$ *is the corresponding edge set.*
- $W_{\Upsilon_c} = \{w_{ij} | w_{ij} = \sum_{c' \in \Upsilon_c} (w_{ij}^{c'} \cdot Rel(c^{'}||c))\}$ *is the corresponding weight set.*

Figure 2 illustrates an example of extended category link graph in a Q&A community. In the figure, we extend the category link graph of category A through the category link graph of its relevant category B. The common nodes and edges between the two category link graphs are merged, and corresponding weights are also summed with respect to category relevancy.

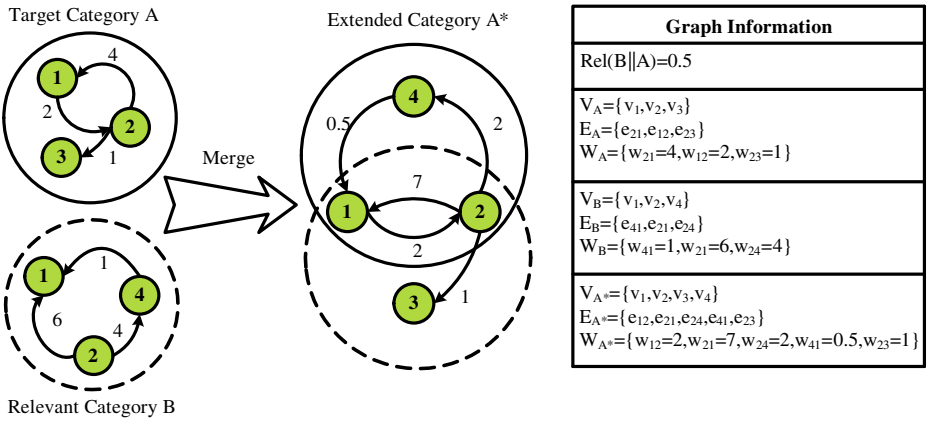

Figure 2 An example of extended category link graph in a Q&A community. Node denote users, and an edge from user u_i to user u_j denotes that u_j has answered a question posted by u_i

Indeed, some state-of-the-arts works on topical influence analysis (e.g., [19, 29, 32]) can be leveraged to rank user authority for all categories in the full user graph with respect to users' static topical distributions of category. However, these works are not suitable for our expert finding work, which is mainly focus on finding experts for a specific target category, because of two reasons. First, topical influence analysis should run in full user graph, thus the computational cost in both time and memory are very expensive when given large-scale data sets. Second, the full user graph contains a number of irrelevant users for the target category, thus the ranking process will be impacted by the noise information. In the experimental results in Section 5.4.4, it is clearly showed that the expert finding performance will be impacted dramatically when given more users of irrelevant categories.

Based on the discussion above, in this paper we propose the new framework for expert finding, namely, category relevancy based authority ranking. With above notions, the problem of category relevancy based authority ranking is formally defined as follows.

Definition 3 (Category relevancy based authority ranking) Given a category c, the task of **category relevancy based authority ranking** is to build the extended category link graph G_{Υ_c} and then find top K authoritative users for category c in G_{Υ_c}.

To be specific, the problem of category relevancy based authority ranking can be divided into two sub-problems as follows. The first problem is how to find the relevant category set R_c to extend the original category link graph G_c. The second problem is how to rank user authority for category c in the extended category link graph G_{Υ_c}. For the first problem, we propose to exploit topic models for measuring category relevancy based on both content and user interaction similarities. For the second problem, we develop a novel topical link analysis approach for authority ranking by extending the Topical Random Surfer model [24] for leveraging the information in both target and relevant categories. In the following sections, we present the technical details of our solutions for the two sub-problems, respectively.

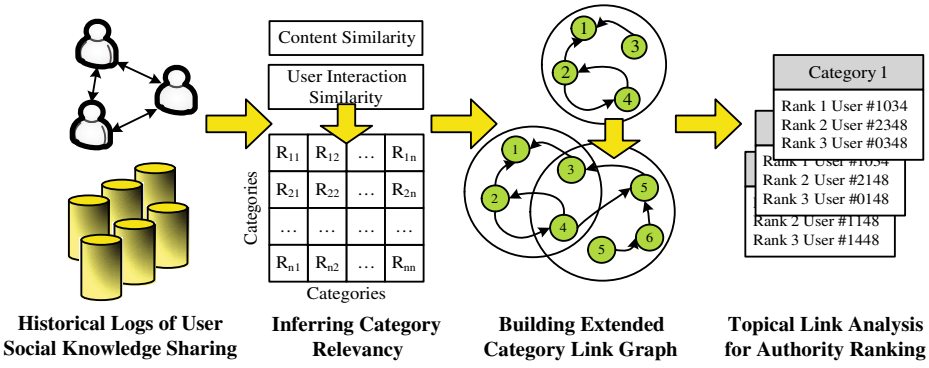

Figure 3 The framework of our category relevancy based authority ranking approach, which contains two main steps: (1) Inferring category relevancy and building extended category link graph, and (2) topical link analysis based authority ranking in extended category link graphs for expert finding

The Figure 3 shows the framework of our category relevancy based authority ranking approach.

3 Inferring category relevancy

In this section, we introduce how to infer the category relevancy for building extended category link graphs. Indeed, an intuitive method to find the relevant categories of a given category is directly utilizing the category taxonomies [13] used by knowledge sharing social networks. However, some of the knowledge sharing web sites do not have effective taxonomies, such as most of the online forums. Furthermore, it is difficult to calculate precise category relevancy from the predefined category taxonomies. For example, we cannot evaluate which category is more relevant to "Information Technology" between categories "Computer Science" and "Internet" if they all belong to the same parent category "Technology". As a results, it motivates us to develop a novel approach to precisely estimating category relevancy.

In this paper, we propose to leverage both content based similarity and user interaction based similarities for measuring category relevancy. To be specific, it is on the basis of two intuitive principles. First, we argue that if the content posted in some of the categories are similar, these categories are mutually relevant. Second, we argue that if many users have interactions (e.g., post/answer questions in Q&A web sites.) in some specific knowledge categories, these common categories are mutually relevant. Since these categories may have latent semantic relationships. Based on these two principles, we calculate the relevancy between categories c_i and c_j as follows.

$$Rel(c_i||c_j) = ContSim(c_i||c_j) \cdot InterSim(c_i||c_j), \tag{1}$$

where $ContSim(c_i||c_j)$ is the content based similarity between categories c_i and c_j, and $InterSim(c_i||c_j)$ is the user interaction based similarity. Therefore, in the following sections, we will present how to estimate the two similarities $ContSim(c_i||c_j)$ and $InterSim(c_i||c_j)$ for measuring final category relevancy.

3.1 Inferring content based category similarity

We argue that if the content posted in some of the categories are similar, these categories are mutually relevant. It is very intuitive, since we find that the relevant knowledge categories will cover many common knowledge. For example, in a Q&A web site, the two categories "Data Mining" and "Machine Learning" contain many common questions, such as questions about "Classification".

An intuitive approach to measuring content based similarity is based on vector space model (VSM) [28]. To be specific, we can first integrate all the content posted with category label c as the *category profile* d_c, and then build a normalized words vector $\overrightarrow{w_c} = dim[n]$ for each category c, where n indicates the number of all unique normalized words in all category profiles. Finally we can calculate the content similarity through the Cosine distance between two category vectors.

Although the explicit feedback of VSM can capture the content similarity between two different categories in terms of the occurrences of words, it does not take

advantage of the latent semantic meaning behind words and may not work well in some cases. For example, in VSM, the following words "Game", "Play" and "Funny" are treated as totally different measures to calculate the distance between word vectors. However, these words indeed have latent semantic relationships because they can be categorized into the same semantic topic "Entertainment". According to some previous studies [30], the latent semantic topics can improve the performance of measuring content based similarity. Therefore, instead of using VSM based approach, we propose to leverage topic models for inferring content based category similarity. The basic assumption is that two categories are relevant because their probabilities of belonging to the same latent topic are similar.

Topic models assume that there are several latent topics 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. Intuitively, if we take category profile as documents we can directly take advantage of topic models for inferring latent topics of categories. Then we can represent each category c as a conditional probabilistic distribution $P(z|c)$ which denotes the probability of category c being labeled with topic z.

Among several existing topic models, we use the Latent Dirichlet Allocation model (LDA) [5] in our approach. According to LDA, a category profile d_i is generated as follows. Firstly, a prior topic distribution θ_i is generated from a prior Dirichlet distribution α. Secondly, a prior word distribution ϕ_i is generated from a prior Dirichlet distribution β. Therefore, for the j-th word w_j in d_i, the model generates a topic $z_{i,j}$ from θ_i and then generates w_j from ϕ_i. Figure 4 shows the graphical model of LDA.

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., $P(d|\alpha, \beta, \theta, \phi)$. In this paper, we choose a Markov chain Monte Carlo method, namely Gibbs sampling introduced in [9] to provide a relatively efficient process for training LDA model. This method begins with a random assignment of words to topics for initializing the state of Markov chain. In the following each iteration of the chain, the method will re-estimate the conditional probability of assigning a word to each topic, which is conditioned on the assignment of all other words. Then a new assignment of words 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 word is assigned a stable topic. After the model training, we can get the estimated value $\widetilde{P}(z_i|c)$ by

$$\widetilde{P}(z_i|c) = P(z_i|d_c) = \frac{n_i^{(d_c)} + \alpha}{n_{\cdot}^{(d_c)} + |K|\alpha}, \tag{2}$$

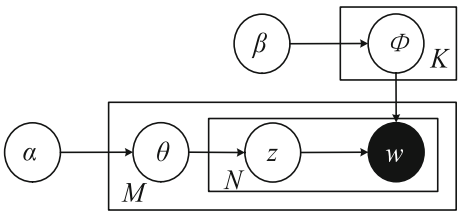

Figure 4 The graphical model of LDA, where M is the number of category profiles, N is the number of words, K is the number of latent topics, α and β are the hyper parameters

where the $n_i^{(d_c)}$ is the number of times a word from category profile d_c that has been assigned to topic z_i. The $|K|$ is the number of latent topics.

By utilizing LDA, each category c can be represented as a K-dimension vector of topic distribution $P(z|c)$. Thus, the task of estimating content based category similarity is converted to calculate the distance between vectors. There are a lot of methods which can be used to calculate the distance between two vectors, such as Euclid distance and Cosine distance [12]. Most of these methods are all symmetric measures, which means $Distance(\overrightarrow{A}, \overrightarrow{B}) = Distance(\overrightarrow{B}, \overrightarrow{A})$ where both $\overrightarrow{A}$ and $\overrightarrow{B}$ denote vectors. However, we observe that the similarity between two categories are often asymmetric. For example, it is more possible that a user who has expertise in the "Computer science" category will also know about the "Mathematics" category well than the opposite situation. To this end, in this paper, we propose to use normalized Kullback Leibler (KL) divergence [17], which is an asymmetric measure, for measuring content based category similarity. The KL-divergence from category c_i to category c_j is computed by

$$KL(c_i||c_j) = \sum_z P(z|c_i) log \frac{P(z|c_i)}{P(z|c_j)}. \tag{3}$$

It is worth noting that $\forall_{i,j} KL(c_i||c_j) \geqq 0$ according to the Gibbs' inequality [17]. Then we calculate the content based similarity between categories c_i and c_j by the following normalized formula,

$$ContSim(c_i||c_j) = 1 - \frac{KL(c_i||c_j)}{Max(KL(c_j))}, \tag{4}$$

where $Max(KL(c_j))$ denotes the maximum KL-divergence from other categories to category c_j.

Specifically, in some of the knowledge sharing social networks, which do not have rich word base content information (e.g., multimedia sharing web sites), we can leverage the meta data (e.g., tags) for building category vector space.

3.2 Inferring user interaction based category similarity

We argue that if many users have interactions (e.g., post/answer questions in Q&A web sites.) in some specific knowledge categories, these common categories are mutually relevant. An intuitive method to capture user interaction based category similarity is utilizing the category co-occurrence based method. Specifically, we can calculate the similarity between two categories by considering their co-occurrences in each user's *interactive log*. A user interactive log consists of a set of category labels where the user made or replied the content with these category labels and the corresponding frequencies. With interactive logs, we can directly calculate the distance between two categories by representing each category as a vector of users. However, this method does not consider the latent relationships between categories thus cannot estimate accurate similarity. For example, if both category A and category B have lots of common users with category C but they only have few common users between themselves, they will be given a low relevant score in this method. However, intuitively, we can think that they may have a latent relationship through the category C.

Instead, in this section we also propose to leverage topic models for inferring user interaction based category similarity. The basic assumption is that two categories often attract similar users because they all belong to similar semantic category topics. For example, many users are interested in the categories "Singing", "Pop Music" and "Instruments" because they all belong to the latent topic "Music". Intuitively, if we take category labels as words, take user interactive logs L as documents we can directly take advantage of LDA topic model for inferring latent topics of categories. Then we can represent each category c as a conditional probabilistic distribution $P(z|c)$ which denotes the probability of category c being labeled with topic z.

The main requirement for our approach is to estimate the probability $P(z_i|c)$, which cannot be obtained directly from LDA. However, according to the Bayes formula we can calculate $P(z_i|c)$ as follows.

$$P(z_i|c) = \frac{P(c|z_i)P(z_i)}{\sum_i P(c, z_i)}, \tag{5}$$

where $P(c|z_i)$ and $P(z_i)$ can be obtained from LDA model. We choose the Gibbs sampling to train LDA model. After the model training, we can get the estimated value $\widetilde{P}(c|z_i)$ as follows.

$$\widetilde{P}(c|z_i) = \frac{n_i^{(c)} + \beta}{n_i^{(\cdot)} + |C|\beta}, \tag{6}$$

where $n_i^{(c)}$ indicates the frequency that category c has been assigned to topic z_i, $n_i^{(\cdot)}$ indicates the frequency that any category is assigned to topic z_i, and $|C|$ indicates the total number of unique categories. Similarly, the estimated value $\widetilde{P}(z_i)$ can be calculated as follows.

$$\widetilde{P}(z_i) = \frac{n_i^{(\cdot)}}{\sum_i n_i^{(\cdot)}}. \tag{7}$$

After the training process, we also leverage the KL divergence to measure the user interaction based similarity. The calculation is similar as (3) and (4).

According to both content and user interaction based category similarity, we can calculate the relevancies between each pair of categories according to (1), we can obtain the category relevancy matrix $M_C = \{m_{ij} = Rel(c_i||c_j)\}$, where $i, j \in [1, n]$. Figure 5 illustrates an example of generating the category relevancy matrix. From M_C, we can easily find the relevant categories R_c for a given category c through a predefined relevancy threshold τ. In Section 6, we analyze the robustness of expert finding given varying parameters τ.

3.3 Selecting the number of latent topics

In both processes of measuring content and user interaction based category similarity, we leverage the LDA topic model. However, 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. [4] to estimate Z. To be specific, take the calculation of user interaction based similarity as an example, we first empirically define a topic number range $KR = [K_{\min}, K_{\max}]$ and then select

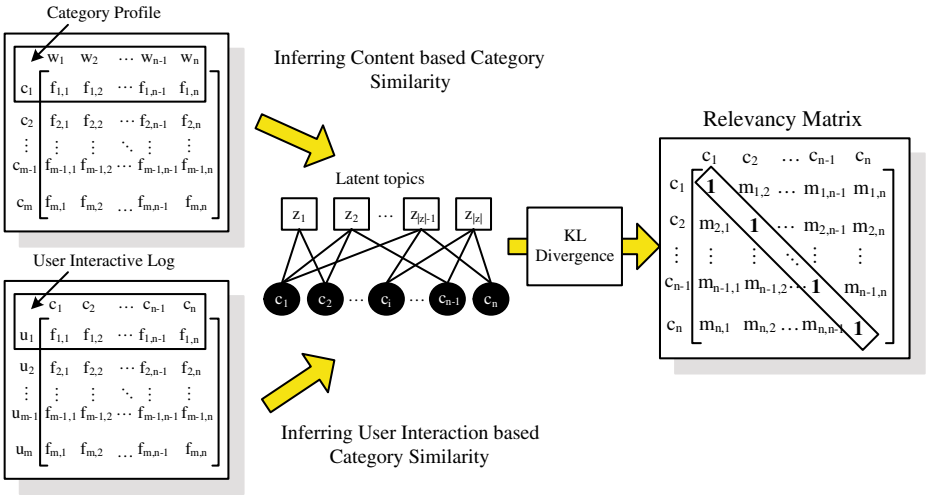

Figure 5 The generation of category relevancy matrix by LDA, which contains two main steps: (1) Inferring content based category similarity from category profiles, and (2) Inferring user interaction based category similarity from user interaction logs

two groups of L_i as training data set S_a and test set S_b. After training from S_a with a given $K_p \in KR$, we obtain K_p topics $z_1, z_2, ..., z_{K_p}$. At last we use perplexity [1, 5] to determine topic number Z. The perplexity is defined by the following equation,

$$Perplexity(S_b) = \text{Exp}\left\{-\frac{\sum_{L_i \in S_b} \log P(L_i|S_a)}{\sum_{L_i \in S_b} N_{L_i}}\right\}, \tag{8}$$

where N_{L_i} is the number of categories in L_i and $P(L_i|S_a)$ represents the conditional probability that L_i appears in S_a and it can be calculated as follows:

$$\begin{aligned} P(L_i|S_a) &= \prod_{c \in L_i} P(c|S_a) = \prod_{c \in L_i} \sum_{j=1}^{K_p} P(c, z_j|S_a) \\ &= \prod_{c \in L_i} \sum_{j=1}^{K_p} P(c|z_j) P(z_j|S_a), \end{aligned} \tag{9}$$

where $P(c|z_j)$ can be obtained by (6). $P(z_j|S_a)$ can be computed by $\frac{n_{L_i,j}+\alpha}{\sum_{q=1}^{K_p} n_{L_i,q}+\alpha}$, where $n_{L_i,j}$ denotes the number of categories labeled by z_j in L_i.

The perplexity has the inversely-proportional relationship to the performance of topic estimation. However, to avoid the model over-fitting, which will result the perplexity of test set always drops with the increasing of K_p, we cannot only use the minimum perplexity as the metric for topic estimation [1, 5]. An alternative method is to define a decline rate ζ of perplexity as best condition, if the decline rate of perplexity is less than ζ we choose the current K_p to infer topics. In our experiments, we set ζ to be 10 % according to [4].

4 Authority ranking through link analysis

By finding relevant categories from the category relevancy matrix, we can build the extended category link graph for each target category. Compared with a normal category link graph, in an extended category link graph the authority propagation in the target category between two users may be impacted by their different original expertise in the target category before authority propagation. However, most state-of-the-art methods of link analysis, such as PageRank [27], do not take into account the user nodes with multiple category labels and different original expertise in these categories. For example, if user u_1 mainly contribute knowledge to category "Mathematics" and rarely to category "Physics", while user u_2 only dedicates to category "Physics", the traditional link analysis approaches may treat them as experts with same contribution for authority propagation in the extended category link graph of "Physics". Therefore, these approaches are not proper to be applied to ranking authority in extended category link graphs because the users with little expertise in the target category may be overestimated in authority propagation. Intuitively, different users may have different expertise in different categories, and a user who has little expertise in the target category should have little contributions in authority propagation. To this end, we extend the Topical Random Surfer (TRS) model [24, 25], which is multiple-category-sensitive, to rank user authority in extended category link graphs for taking into account their different original expertise in the target category.

The TRS model is originally proposed for web page ranking. Its basic idea is similar to the "random surfer" process described in PageRank model and the special property is that the "random surfer" is sensitive to different topics of web pages. Specifically, in the TRS model, there are two possible ways to move to another web page v' for a web surfer who is browsing a web page v for the interesting topic z. The first is with probability $(1 - d)$ to follow a outgoing link on the current page v (e.g., clicking a hyper-link). Another is with probability d the surfer will jump to a random page from the entire web W (e.g., directly typing an url in the address field). Moreover, for each new page v', the surfer will browse it either because of the same interesting topic z with probability $\psi_{v,z}$ or any other interesting topic z' with probability $(1 - \psi_{v,z})$. Therefore, there are total three reasons for the web surfer to browse a new web page v', namely, 1) following a link for the same interesting topic z, 2) following a link for any other interesting topic ($z' \neq z$) and 3) jumping to another page for any interesting topic z'. To facilitate expression, TRS model names these three reasons as "F_S", "F_J" and "J_J", respectively.

To utilize TRS model for our authority ranking problem, we take the extended category link graph G_{Υ_c} as a web page link graph G, let each $u \in G_{\Upsilon_c}$ correspond to a web page v and let the original expertise of each user in different categories (without considering the authority propagation) correspond to different topics of a web page. Moreover, in our problem "F_S", "F_J" and "J_J" denote 1) following a link to select the next user as the authoritative user for the same category c, 2) following a link to select the next user as the authoritative user for any other interesting category c' ($c' \neq c$), and 3) randomly select a user as the authoritative user for any category c', respectively. The Figure 6 demonstrates the example of three steps in TRS for authority ranking. In this example, the category "Visual Arts" is extended

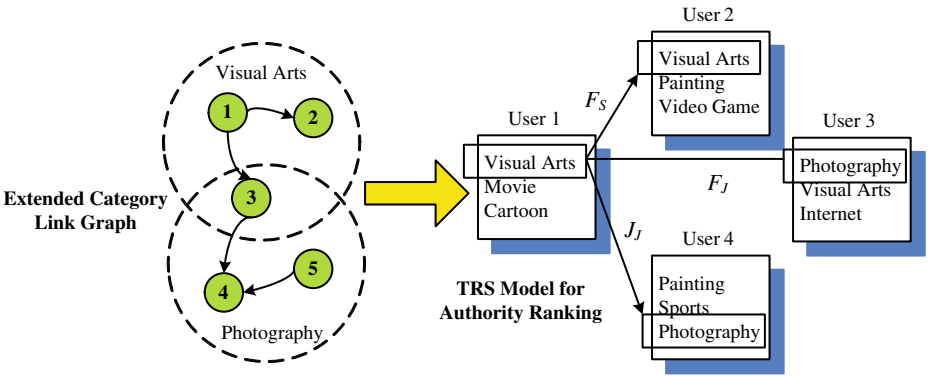

Figure 6 The example of TSR for authority ranking, where category Photography is leveraged to extend the original category Visual Arts. Through User 1, we can find other authoritative users with different reasons, namely F_S, F_J and J_J

by category "Photography". Specifically, from user 1 we can find users 2, 3 and 4 according to "F_S", "F_J" and "J_J" respectively. Therefore, we have the following equations according to the TRS model.

$$\begin{cases} P(F_S|u,c) = (1-d)\psi_{u,c} \\ P(F_J|u,c) = (1-d)(1-\psi_{u,c}) \\ P(J_J|u,c) = d \end{cases}, \tag{10}$$

where $P(*|u,c)$ denotes the conditional probability of next choice of the surfer denoted as $*$ given that the surfer has selected u as the authoritative user for category c. $\psi_{u,c} = P(c|u) = P(z|u)P(c|z)$ can be directly estimated by the LDA model trained in the stage of calculating user interaction based category similarity. Accordingly, we can have the following equations,

$$\begin{cases} P(u',c'|u,c',F_S) = D(u,u') \\ P(u',c'|u,c,F_J) = D(u,u')\psi_{u,c'} \\ P(u',c'|u,c,J_J) = \dfrac{1}{|V_{\Upsilon_c}|}\psi_{u',c'} \end{cases}, \tag{11}$$

where $P(u',c'|u,c,*)$ denotes the conditional probability of selecting u' as the authoritative user for category c given that the surfer selected u as the authoritative user for category c previously and then selected the choice $*$, and we have

$$D(u,u') = \frac{w_{u,u'}}{\sum_{u^\star:u\to u^\star} w_{u,u^\star}}. \tag{12}$$

According to above equations, we can calculate the joint probability $P(u', c')$ which denotes the probability that the surfer is selecting user u' as an authoritative user for category c' as follows.

$$\begin{aligned} P(u', c') &= f(F_S, F_J, J_J) \\ &= \sum_{u:u\to u'} P(u', c'|u, c', F_S) P(F_S|u, c') P(u, c') \\ &\quad + \sum_{u:u\to u'} \sum_{c} P(u', c'|u, c, F_J) P(F_J|u, c) P(u, c) \\ &\quad + \sum_{u\neq u'} \sum_{c} P(u', c'|u, c, J_J) P(J_J|u, c) P(u, c). \end{aligned}$$

According to (10) and (11), we can obtain the final estimation of user authority by

$$\begin{aligned} P(u', c') &= f(F_S, F_J, J_J) \\ &= \sum_{u:u\to u'} D(u, u') P(u, c')(1-d)\psi_{u,c'} \\ &\quad + \sum_{u:u\to u'} \sum_{c} D(u, u')\psi_{u,c'}(1-d)(1-\psi_{u,c}) P(u, c) \\ &\quad + \sum_{u\neq u'} \sum_{c} \frac{d}{|V_{\Upsilon_c}|}\psi_{u',c'} P(u, c). \end{aligned}$$

With the above equation, we can iteratively calculate $P(u, c)$ for each user u for the target category c. In the first round of propagation, we let $P(u', c') = \frac{1}{|V_{\Upsilon_c}|\times|C|}$. Then the result will converge after several rounds of propagation. Therefore, we can rank all users' authority in G_{Υ_c} for category c by $P(u, c)$.

5 Experimental results

In this section, we evaluate the performance of our **C**ategory **R**elevancy based **A**uthority **R**anking (CRAR) approach. Specifically, we compare CRAR with several benchmark link analysis methods for expert finding. All the experiments implemented by standard C++ and conducted on a 2.8 Ghz × 2 cores CPU, 4 G main memory PC.

5.1 Data sets

The data sets used in the experiments are collected from two major commercial Q&A web sites. The first one is a public data set collected from Yahoo! Answers (http://answers.yahoo.com/) by Liu et al. [21]. There are 100 categories, 216,563 questions, and more than 1.9 million answers posted by 171,266 users in this data set. Another data set was collected from a major Chinese Q&A service web site named Tianya Wenda (http://wenda.google.com.hk/; http://wenda.tianya.cn/) from 15 August 2008 to 20 June 2010. This data set contains more than 1.3 million questions, 5.5 million answers, and 595 categories. The collected questions and

Table 1 Statistics of our two real-world data set in experiments

	Yahoo! Answers	Tianya Wenda
Number of questions	216,563	1,311,907
Number of answers	1,982,006	5,520,303
Number of users	171,266	274,896
Number of categories	100	595

answers were posted by 274,896 users. In both data sets, all questions are resolved questions which contain a best answer voted by the question author. Therefore, these data sets contain few noise data such as questions posted by robots. The detailed data statistics are listed in Table 1.

Figure 7a and b show the distributions of users with respect to the number of unique categories appearing in their interactive logs for two data sets. In these figures, we can observe that both distributions roughly follow the power law distribution. Thus, it is common for users to answer questions in multiple categories which may include relevant categories, which implies the need to consider the information in relevant categories for authority ranking.

5.2 Estimation of category relevancy

In this section, we study the effectiveness of our method of finding relevant categories. First, we build category profiles and user interaction logs, and then apply LDA for inferring both latent word and category topics. According to the perplexity introduced in Section 3, the numbers of word topics and category topics are set to be 50 and 30 for the Yahoo! Answers data set, while 150 and 100 for the Tianya Wenda data set. The two parameters α and β are empirically set to be $50/Z$ and 0.2 according to [10].

Since the effectiveness of mining word topics are validated by many previous works (e.g., [5, 30]), here we focus on the performance of mining category topics. In Tables 2 and 3, we randomly select and show four mined category topics for each data set. For each topic, the top 10 categories with respect to generation probability are listed. From the table we can observe that the mined latent category topics are

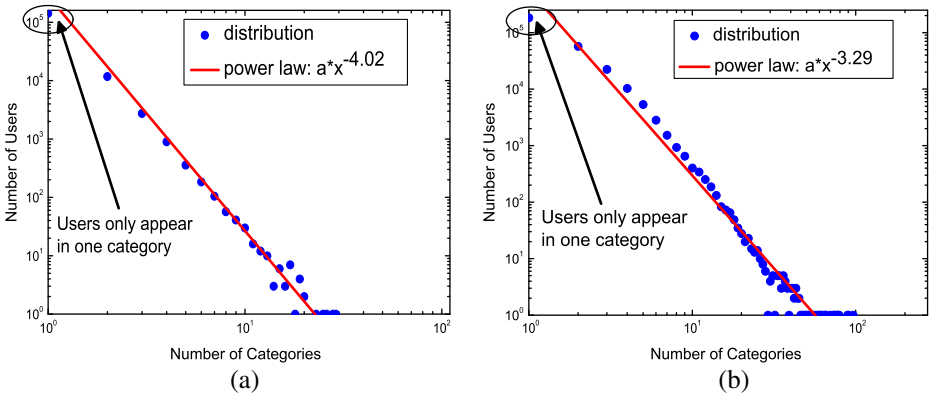

Figure 7 The distributions of users with respect to the number of unique categories in **a** Yahoo! Answers, and **b** Tianya Wenda

Table 2 Examples of latent category topics learnt from Yahoo! Answers data set by LDA

Yahoo! Answers	
Topic 5	Basketball, baseball, fantasy sports, NASCAR, golf, dancing horse, racing, injuries, standards & testing, studying abroad
Topic 13	Chemistry, financial aid, engineering, homework help, cancer, injuries, infectious diseases, environment, botany, teaching
Topic 17	History, home schooling, higher education, teaching, poetry, studying abroad, painting, quotations, homework help, physics
Topic 24	Men's premiere leagues, 2006 FIFA World Cup, Scottish football, Mexi.Soccer, olympics, baseball, injuries, genealogy, dancing, rugby

reasonable. For example, the first category topic learnt from the Yahoo! Answers data set is about "*Sports*" since it contains many categories which are relevant to sports.

After inferring latent topics, we use KL-divergence to calculate the both content and user interaction based similarity. Finally, we integrate both category similarities into (1) for building category relevancy matrix. Since all the categories have been mapped into low dimensional representations of latent topics, the process of estimating pairwise similarity between categories is very fast (i.e., less than 50 millisecond in measuring each pair of categories in both data sets). Since the category relevancy matrix of Tianya Wenda with 595 categories needs a lot of space to show, we only show the visualization of category relevancy matrix in the Yahoo! Answers data set in Figure 8.

Tables 4 and 5 show some categories and the corresponding Top 5 relevant categories for both data sets. From the table we can see that our method can effectively find relevant categories for many categories. In our experiments, only the categories with relevancy score (extension rate) higher than $\tau = 0.5$ are selected to extend the original category link graphs, moreover, we will also show the validation of this setting by analyzing the robustness of expert finding in the following sections. Given the above settings, the average numbers of relevant categories used for graph extension of each given category are 3.9 in Yahoo! Answers and 5.4 in Tianya Wenda.

Table 3 Examples of latent category topics learnt from Tianya Wenda data set by LDA

Tianya Wenda[a]	
Topic 18	Internet, software, E-business, programming, windows, QQ communication, higher education, movies, computer games
Topic 34	Entertainment, asian stars, star shows, sports stars, music movies, instruments, computer games, hobbies, internet
Topic 56	Economics, markets, corporate management, politics, stock lottery, E-business, ERP, investment, higher education
Topic 77	Science, examination, philosophy, physics, studying abroad homework, mathematics, technology, engineering, internet

[a]The original labels of categories in the Tianya Wenda data set are in Chinese, thus we manually translate them into English

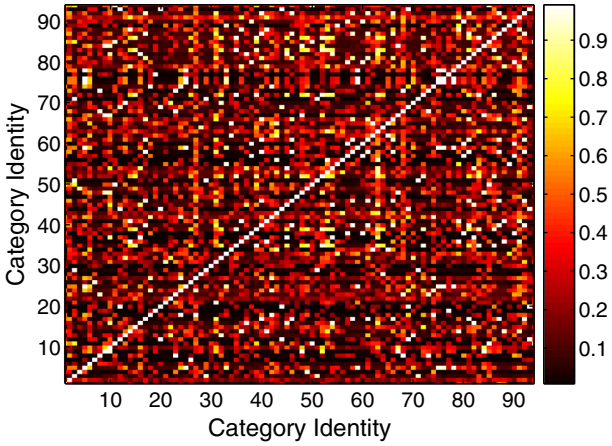

Figure 8 The pair-wise category relevancy measured by our approach between each pair of the 100 categories in the Yahoo! Answers data set

5.3 Benchmark methods

To the best of our knowledge, there is no previous work has been reported that can leverage relevant categories for user authority ranking. Thus to evaluate our novel approach, we first choose one intuitive authority ranking approaches and two state-of-the-arts authority ranking approach which are based on original category link graph as baselines.

Degree is a simple statistical measure which ranks user authority in the order of the in-degrees of the according user node in the category link graph. This approach is intuitive and widely used in related works.

HITS [16] 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 good hub users. In authority ranking, users are ranked in the descending order of authority scores.

ExpertiseRank [33] is extended from PageRank. This algorithm does not only considers how many other users one helped, but also whom he/she helped. To be

Table 4 Examples of relevant categories learnt for 4 target categories in Yahoo! Answers data set by our approach

Yahoo! Answers				
	Earth science & geology		Drawing & illustration	
#1	Geography	0.714	Painting	0.770
#2	Physics	0.684	Photography	0.707
#3	Environment	0.618	Home schooling	0.632
#4	Higher education	0.506	Books & Authors	0.563
#5	Biology	0.309	Teaching	0.381
	Performing arts		Diet & fitness	
#1	Theater & acting	0.815	Volleyball	0.595
#2	Sculpture	0.771	Women's health	0.589
#3	Martial arts	0.601	Hunting	0.576
#4	Special education	0.370	Diabetes	0.531
#5	Photography	0.338	Optical	0.482

Table 5 Examples of relevant categories learnt for 4 target categories in Tianya Wenda data set by our approach

Tianya Wenda				
	Computer fundamental		Higher education	
#1	Internet	0.802	Examination	0.823
#2	Windows	0.775	School	0.723
#3	Higher education	0.630	English	0.610
#4	Software	0.590	Homework help	0.507
#5	Computer games	0.534	Computer	0.432
	Music		History	
#1	Pop music	0.732	Polity	0.764
#2	Performing arts	0.623	Higher education	0.607
#3	Asian stars	0.582	Philosophy	0.555
#4	Hobbies	0.523	Examination	0.531
#5	Instruments	0.387	Teaching	0.506

specific, a user who help more authoritative users will be assigned a higher authority score.

Besides the above approaches, we also propose an authority ranking approach extended from our CRAR approach as baselines.

TRSO stands for TRS for original category link graph. It is an topical link analysis approach by leveraging TRS model in the original category link graphs but not the extended category link graphs.

To further validate the effectiveness category relevancy in authority ranking, we propose two extended category link graph based authority ranking approach as baselines.

HITS-E is extended from HITS, which is based on the extended category link graph introduced in Section 2.

ExpertiseRank-E is extended from ExpertiseRank, which is based on the extended category link graph introduced in Section 2.

5.4 Empirical evaluation of authority ranking

How to evaluate the authority ranking performance is not a trivial problem, since both data sets have no principle benchmark for who are real authoritative users for a given category. According to the discussion in previous related works [6, 14, 34], in this paper we leverage two different benchmarks to empirically evaluate each authority ranking, namely, human judgement and golden standard.

5.4.1 Human judgement based evaluation

This benchmark is on the basis of the quality of content posted by each expert candidate. Intuitively, the more authoritative users will post higher quality content in corresponding knowledge categories they have expertise in. To accurately evaluate the semantic meanings of content, we manually inspect the results. To be specific, firstly we carry out each measuring approach to find top K (i.e. $K = 10$ in our experiments) users as expert candidates for all target categories. Actually, different approaches may return same users in the results, and the interactions in some categories are very sparse. Therefore, the average number of returned unique users by all approaches is 42.3 in Yahoo! Answers data set and 48.7 in Tianya Wenda

data set. Moreover, the average number of the answers of each candidate is 21.5 in Yahoo! Answers data set and 30.9 in Tianya Wenda data set. Then, for each mined expert candidate u for category c, we guarantee that there are three different human evaluators to check whether u is a real expert for the category c by comprehensively considering the quality of content (i.e., answers) u has posted which is related to category c. Each identified authoritative user is voted by three different evaluators with label **Yes** (the user is a real expert) or **No** (the user is not a real expert). To achieve above settings, we totally invited twenty human evaluators who are students major in computer science for labeling experimental results. It is worth noting that when the evaluators judge the answers of a user for category c, they are asked to manually check each answer in the history of the user whether it is relevant to category c other than only take into account the answers with the category label c. Therefore, it is less likely to miss the answers which are indeed relevant to category c but labeled with other relevant category labels.

To evaluate the performance for expert finding, we used three widely-used metrics as follows.

Average Precision@K (**Avg. P@K**) denotes the average ratio of real experts in top K identified authoritative users for each category.

Mean Reciprocal Rank (**MRR**) equals to $\frac{1}{|C|}\sum_{c_i \in C}\frac{1}{rank_i}$, where C is the category set for retrial and $rank_i$ is the rank of the first found real authoritative user, i.e., an expert, for category c_i in top K results. If there is no authoritative user has been found, we let $\frac{1}{rank_i} = 0$.

Mean Average Precision (**MAP**) equals to $\frac{1}{|C|}\sum_{c_i \in C}\frac{\sum_{r=1}^{K}(P_i(r)\times rel_i(r))}{|R_i|}$, where C is the category set for retrial and $|R_i|$ is the number of found authoritative users for category c_i. r is a given cut-off rank, $P_i(r)$ is the precision of c_i at a given cut-off rank r_i and $rel_i()$ is the binary function on the relevance of results.

Because there are 595 categories in the Tianya Wenda data set, it is expensive to test the performance of expert finding for all these categories. Alternatively, we randomly select 100 categories in Tianya Wenda to test the overall performance of our approach and other baselines for expert finding. For the Yahoo! Answers data set, we evaluate the performance for all categories. In addition, as PageRank usually does, d is set as 0.15 here [24]. Note that, CRAR and all the baselines except Degree need to compute user authority through iteration algorithms. Indeed, the time cost of each iteration and the number of iterations for converging of each approach highly depends on the scale of the input category link graphs. Particularly, in our data sets, all these approach can converge in less than 50 iterations for each given category link graph. Table 6 shows the average computational cost in each iteration of different approaches in both data sets. From this table we can find that the time cost of CRAR is higher than that of other approaches. It is reasonable since CRAR is a topical link analysis approach based on extended category link graph, which may involve more user nodes and have more random walk processes than other approaches.

Table 6 The average computational cost in each iteration of different approaches

	HITS	ExpertiseRank	TRSO	HITS-E	ExpertiseRank-E	CRAR
Yahoo! Answers	3.17s	3.32s	3.64s	3.88s	4.08s	**4.97s**
Tianya Wenda	3.24s	3.51s	3.82s	4.01s	4.34s	**5.48s**

Figure 9 shows the average experimental results for all test categories with respect to different metrics. First, from this table we can see that our approach CRAR consistently outperforms other baselines with respect to varying metrics on both data sets. Particularly, CRAR outperforms HITS-E and ExpertiseRank-E, which also rank user authority in extended category link graphs. It is because that TRS model can take into account users' different original expertise in the target category during authority propagation, which clearly validates the discussion in Section 4. Second, HITS-E and ExpertiseRank-E outperform other baselines, which indicates that the knowledge in relevant categories can improve the performance of expert finding, even for the traditional link analysis approaches which do not distinguish the users' different original expertise in the target category during authority propagation. Third, we also observe that the topical analysis in original category link graphs, i.e., TRSO, can only slightly improve the performance of expert finding than ExpertiseRank and HITS. It is because that although TRSO can capture users' different original expertise in the target category, the number of relevant users to the target category are limited in the original category link graphs, thus the topical related information from other users cannot be fully taken advantage of. Finally, Degree has the worst performance of expert finding, which may imply that the in-degrees of each user nodes are not sufficient for estimating user authority.

5.4.2 Golden Standard based evaluation

This benchmark is based on the Golden Standard introduced in [6, 14, 34]. To be specific, we argue that the more authoritative user of knowledge category c will post more high quality answers for questions in category c, and more answers will be selected as best answers. What should be noticed here is that in both data sets, the best answer is selected by question creators, and each answer in Yahoo! Answers has the user voting generated by other users who had browsed the corresponding Q&A thread if they believed the answer is reasonable. However, different from the Yahoo! Answers, the answers in Tianya Wenda do not contain the user voting data. Therefore, our three human evaluators manually checked each posted answer for each expert candidate and assign voting if the quality of the answer is good.

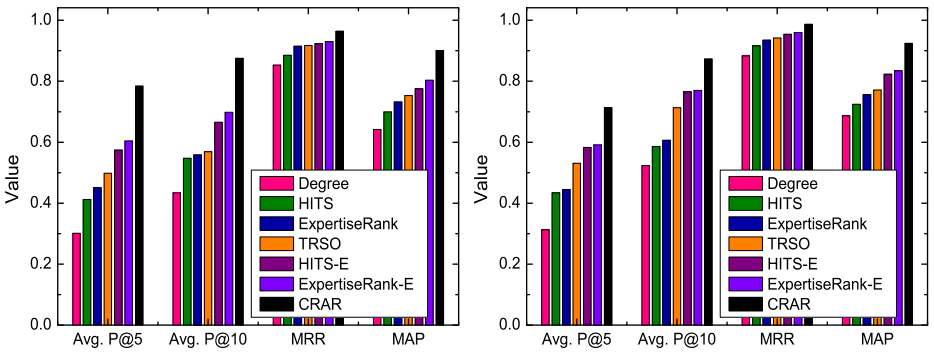

Figure 9 The authority ranking performance of each approach based on human rate in **a** Yahoo! Answers, and **b** Tianya Wenda

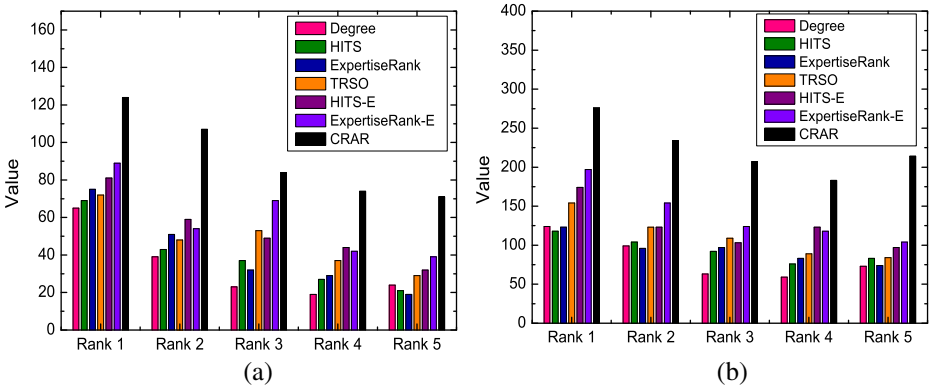

Figure 10 The authority ranking performance of each approach based on golden standards in **a** Yahoo! Answers, and **b** Tianya Wenda

Finally, we propose to leverage the following standard for evaluate authority ranking performance,

$$Score(u, c) = w_1 \times N_{\text{Best}} + w_2 \times N_{\text{Voting}} + w_3 \times N_{\text{Answer}}, \tag{13}$$

where N_{Answer} is the number of answers posted by user u for questions with category label c, N_{Best} is the number of selected best answers in these answers, N_{Voting} is the number of user voting in these answers, and weight parameter w_i is used to highlight the importance of each value. Particularly, in our experiments we empirically set $w_1 = 3$, $w_2 = 2$ and $w_1 = 1$.

Figure 10 shows the results of average *Score* of each top 5 ranked expert candidates found by each authority ranking approach for all tested categories. From this figure we can observe that 1) CRAR consistently outperforms other baselines with respect to different ranking results, and 2) HITS-E and ExpertiseRank-E outperforms other baselines, which indicates that the effectiveness of extended category link graph.

In addition to the studies on the overall performance of CRAR, we also study the cases in which CRAR outperforms the baselines. For example, Table 7 shows the top 5 expert candidates found by CRAR and ExpertiseRank from Yahoo! Answers for the category "Performance Arts" and the corresponding manually

Table 7 Top 5 experts found from Yahoo! Answers for the *Performance Arts* category

	User 1	User 2	User 3	User 4	User 5
CRAR					
# Answers	41	35	37	29	32
# Best answers	14	10	19	5	13
# User voting	47	42	49	23	28
Human judgement	Yes	Yes	Yes	No	Yes
ExpertiseRank					
# Answers	26	35	31	22	15
# Best answers	3	10	14	2	4
# User voting	22	42	25	12	8
Human judgement	No	Yes	Yes	No	No

Table 8 Top 5 experts found from Tianya Wenda for the *Music* category

	User 1	User 2	User 3	User 4	User 5
CRAR					
# Answers	73	89	84	73	64
# Best answers	25	22	17	11	16
# User voting	106	104	68	51	79
Human judgement	Yes	Yes	Yes	Yes	Yes
ExpertiseRank					
# Answers	54	84	62	73	42
# Best answers	6	17	8	11	6
# User voting	21	68	25	51	17
Human judgement	No	Yes	No	Yes	No

counted statistics. To be specific, both the numbers of answers and best answers of each expert for the category "Performance Arts" are manually counted and listed in the table. Moreover, the numbers of user voting for the posted answers in category "Performance Arts" are also listed. In the table, we can see CRAR finds more real experts for the category "Performance Arts" than ExpertiseRank. Moreover, the real experts are ranked higher in the rank list of CRAR. For another example, Table 8 shows the top 10 experts found by CRAR and ExpertiseRank from Tianya Wenda for the category "Music" and the corresponding manually counted statistics. From the table, we can see that CRAR outperforms ExpertiseRank in terms of finding experts for the "Music" category.

5.4.3 Result analysis

As mentioned in Section 3, in traditional authority ranking methods which do not take consider of relevant categories, the users with high in-degrees in the original category link graph will be assigned with high authority scores. However, as we observed, some experts for a given category may often appear in other relevant categories and have low in-degrees in the original category link graph, which results

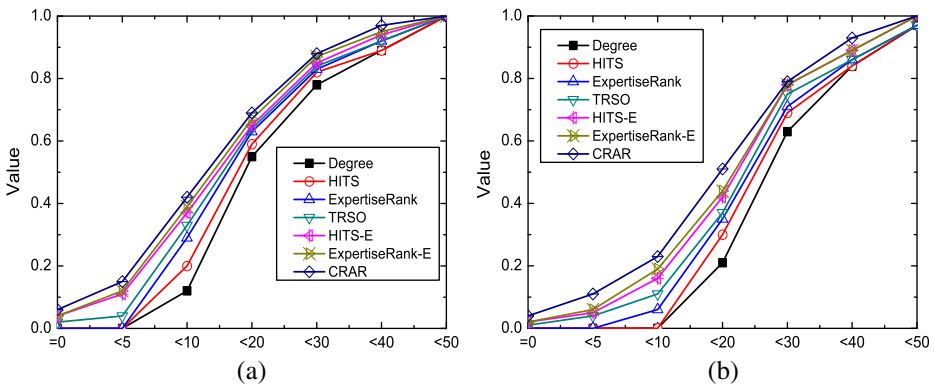

Figure 11 The distributions of user coverage versus the minimum in-degrees of user nodes in original category link graphs for **a** Yahoo! Answers, and **b** Tianya Wenda

that they are assigned relatively low authority scores by traditional authority ranking methods. The advantage of CRAR is the ability of comprehensively taking into account a user's history in the given category and that in relevant categories when determining the authority of the user. Figure 11 illustrates the coverage of top 100 authoritative users found by CRAR and baselines with respect to the minimum in-degree in the original category link graph. From this figure we can see that our approach can find more authoritative users with low in-degrees in the original category link graph than other approaches. It implies that the authoritative users who contribute more in relevant categories but relatively less in the original category are fairly regarded by CRAR.

5.4.4 Robustness analysis

The CRAR approach needs one parameter, namely, the extension rate τ of categories to determine extended category link graph. To evaluate the impact of this parameter, we test our CRAR approach with varying settings of τ. Figure 12a and b show the Avg. P@10 and MAP of CRAR with varying extension rates for each data set, respectively. From these figures we can observe that the performance of CRAR for expert finding roughly satisfies that first increases with respect to the increase of extension rate, and then after a certain extension rate it decreases with the increase of τ. The phenomenon is reasonable, since we find that while the extension rate is small, many of the categories are regarded as relevant to the target category and used to extend the original category link graph. In this case, a number of the irrelevant users will be involved as noise information and will dramatically impact the performance of expert finding, which is also one of the important reasons why we cannot rank user authority in full user graphs. In another case that when extension rate is big, only few of the categories are worth being used for extending the original category link graph. Therefore, the benefit from other relevant categories is very limited and the performance of CRAR is similar as the TRSO.

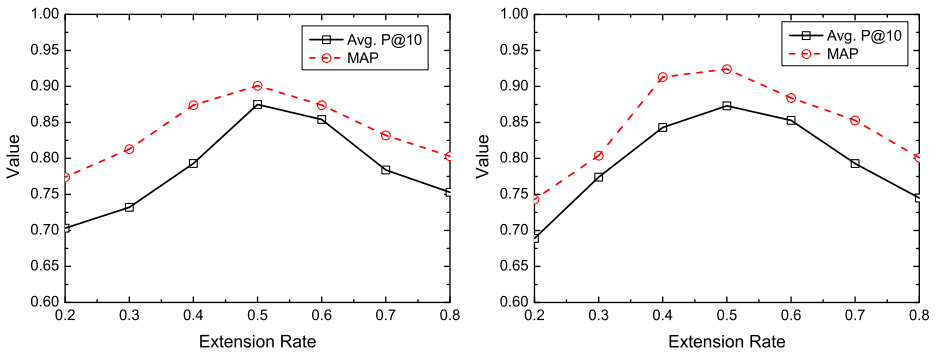

Figure 12 The Avg. P@10 and MAP performance of CRAR with respect to varying numbers extension rates in **a** Yahoo! Answers, and **b** Tianya Wenda

6 Related work

Most of the state-of-the-arts works of authority ranking for expert finding are based on link graphs where the nodes represent the interactive users and the edges represent their relationships. The graph representation allows researchers to apply link analysis techniques and graph based ranking algorithms to find authoritative users. In general, these works can be grouped into two categories.

In the first category, the studies on user authority ranking are based on conventional web page ranking algorithms and their variations. For example, Jurczyk et al. [14] formulated a graph structure in Q&A communities and proposed a variation of the HITS [16] algorithm for predicting authoritative users in Yahoo! Answers. Zhang et al. [33] investigated various authority ranking algorithms in the Java forum and also proposed a PageRank [27] like algorithm named "ExpertiseRank" to find experts. Bouguessa et al. [6] proposed a method for automatic identification of authoritative users without pre-specifying an expert number K in Yahoo! Answers based on PageRank and HITS algorithms. Kao et al. [15] proposed a novel hybrid approach based on basic link analysis and language model to effectively find experts for the category of the target question in question answering web sites. Finally, Campbell et al. [7] and Dom et al. [8] analyzed the link structure of email networks and studied the performances of various link analysis algorithms.

Although the conventional web page ranking algorithms are appealing for the authority ranking problem, there are still some specific needs for expert finding in knowledge sharing social networks, such as the needs of considering social relationships and social topic distributions [31]. Therefore, in the second category, researchers proposed some novel link analysis approaches according to these requirements. For example, Zhang et al. [34] proposed a propagation-based approach for finding experts in co-author social networks which take into account user profiles and Lu et al. [22] extended it with latent link analysis and language model. McCallum et al. [23] proposed an Author-Recipient-Topic model for social network analysis and expert finding which take into account the topic distribution in the content posted by authors. Lappas et al. [18] studied the team formation problem by graph based algorithms, which can find a team of experts with consideration of their communication cost. Tang et al. [29] proposed a topic-level model to analyze the social influence in large-scale networks which can be used for expert finding. Weng et al. [32] investigated how to find topic-sensitive influential twitterers in the twitter social network by leveraging topic models.

While the above methods can achieve a reasonable performance in many situations, none of these studies can efficiently and properly exploit the information in relevant categories for authority ranking. Instead, in this paper, we first find the relevant categories for the target category and then utilize a novel link analysis approach to rank user authority by leveraging the information in both target and relevant categories.

In addition, the proposed approach in this paper exploits topic models for measuring category relevancies. Topic model is an effective way for dimensionality reduction and latent similarity inferring of data in the field of text retrieval and information extraction. In topic model, each category is mapped into a latent topic layer. Thus the probability of distribution between each category and latent topics can be obtained. Typical topic models include the Mixture Unigram (MU) [26],

the Probabilistic Latent Semantic Indexing (PLSI) [11], and the Latent Dirichlet Allocation (LDA) [5]. MU is a single-topic-based topic model and others are multiple-topic-based topic models. PLSI which is also known as Probabilistic Latent Semantic Analysis (PLSA) is a statistical technique for the analysis of two-mode and co-occurrence date. LDA topic model is evolved from PLSI with assuming that the topic distribution has a Dirichlet prior. The original parameter estimating method in reference [5] is Expectation-Maximization (EM) algorithm. To improve the performance and apply LDA into large-scale data set, Griffiths et al. [9] proposed to use Gibbs sampling method to estimate parameters. 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.

7 Concluding remarks and future work

In this paper, we investigated how to exploit the information in both target and relevant categories for enhancing authority ranking in expert finding. Specifically, we first provided a method for measuring category relevancy by taking consideration of both content based and user interaction based category similarities. Then, a multiple-category-sensitive topical link analysis approach was extended from the TRS model for ranking user authority in extended category link graphs which were built from both target and relevant categories. Finally, we performed extensive experiments on two large-scale real-world Q&A data sets. The results clearly show that the information in relevant categories, if properly used, can significantly improve the performance of authority ranking for expert finding.

The main limitation of our CRAR lies in its parameter setting, i.e. latent topic number Z and extension rate τ. Although we can set them empirically and achieve the relatively good performance of expert finding, sometimes we need more accurate and stable method for parameter selection. In the future, we plan to investigate more novel approaches to overcome this problem and extend it to go beyond the usual expert finding problem. Moreover, we also plan to apply our approach into some related areas, such as social influence analysis and key person mining etc.

Acknowledgements This work was supported in part by grants from Natural Science Foundation of China (NSFC, Grant Numbers 61073110, 71028002), Research Fund for the Doctoral Program of Higher Education of China (20113402110024), the Key Program of National Natural Science Foundation of China (Grant No. 60933013), and Nokia. It was also partially supported by grants from National Science Foundation (NSF) via grant numbers CCF-1018151, IIS-1256016.

References

1. Azzopardi, L., Girolami, M., Risjbergen, K.V.: 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 (SIGIR'03), pp. 369–370 (2003)
2. Balog, K., Azzopardi, L., de Rijke, M.: A language modeling framework for expert finding. Inf. Process. Manag. **45**, 1–19 (2009)
3. Balog, K., Azzopardi, L., Rijke, M.D.: Formal models for expert finding in enterprise corpora. In: Research and Development in Information Retrieval, pp. 43–50 (2006)

4. Bao, T., Cao, H., Chen, E., Tian, J., Xiong, H.: An unsupervised approach to modeling personalized contexts of mobile users. In: ICDM'10, pp. 38–47 (2010)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Lantent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
6. Bouguessa, M., Dumoulin, B., Wang, S.: Identifying authoritative actors in question-answering forums: the case of Yahoo! answers. In: Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, pp. 866–874. ACM, New York (2008)
7. Campbell, C.S., Maglio, P.P., Cozzi, A., Dom, B.: Expertise identification using email communications. In: Proceedings of the 12th International Conference on Information and Knowledge Management, CIKM '03, pp. 528–531. ACM, New York (2003)
8. Dom, B., Eiron, I., Cozzi, A., Zhang, Y.: Graph-based ranking algorithms for e-mail expertise analysis. In: Proceedings of the 8th ACM SIGMOD Sorkshop on Research Issues in Data Mining and Knowledge Discovery, DMKD '03, pp. 42–48. ACM, New York (2003)
9. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. USA **101**, 5228–5235 (2004)
10. Heinrich, G.: Parameter estimation for text analysis. Technical report, University of Lipzig (2009)
11. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, pp. 50–57. ACM, New York (1999)
12. Huang, A.: Similarity measures for text document clustering. In: Proceedings of the 6th New Zealand Computer Science Research Student Conference (NZCSRSC2008), pp. 49–56. Christchurch, New Zealand (2008)
13. Jiang, J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: In ROCLING X, pp. 19–33 (1997)
14. Jurczyk, P., Agichtein, E.: Discovering authorities in question answer communities by using link analysis. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management, CIKM '07, pp. 919–922. ACM, New York (2007)
15. Kao, W.C., Liu, D.R., Wang, S.W.: Expert finding in question-answering websites: a novel hybrid approach. In: Proceedings of the 2010 ACM Symposium on Applied Computing, SAC '10, pp. 867–871. ACM, New York (2010)
16. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)
17. Kullback, S., Leibler, R.A.: On Information and Sufficiency, pp. 79–86 (1951)
18. Lappas, T., Liu, K., Terzi, E.: Finding a team of experts in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, pp. 467–476. ACM, New York (2009)
19. Liu, L., Tang, J., Han, J., Jiang, M., Yang, S.: Mining topic-level influence in heterogeneous networks. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10, pp. 199–208. ACM, New York (2010)
20. Liu, X., Croft, W.B., Koll, M.: Finding experts in community-based question-answering services. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM '05, pp. 315–316. ACM, New York (2005)
21. Liu, Y., Bian, J., Agichtein, E.: Predicting information seeker satisfaction in community question answering. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, pp. 483–490. ACM, New York (2008)
22. Lu, Y., Quan, X., Ni, X., Liu, W., Xu, Y.: Latent link analysis for expert finding in user-interactive question answering services. In: Proceedings of the 5th International Conference on Semantics, Knowledge and Grid, SKG '09, pp. 54–59. IEEE (2009)
23. McCallum, A., Corrada-Emmanuel, A., Wang, X.: Topic and role discovery in social networks. In: Proceedings of the 16th International Joint Conferences on Artificial Intelligence, IJCAI '05, pp. 786–791 (2005)
24. Nie, L., Davison, B.D., Qi, X.: Topical link analysis for web search. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, pp. 91–98. ACM, New York (2006)
25. Nie, L., Davison, B.D., Wu, B.: From whence does your authority come? Utilizing community relevance in ranking. In: Proceedings of the 22nd National Conference on Artificial Intelligence, AAAI '07, vol. 2, pp. 1421–1426. AAAI Press (2007)

26. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using em. Mach. Learn. **39**, 103–134 (2000)
27. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web. In: Stanford Digital Library Technical Report (1998)
28. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**, 613–620 (1975)
29. Tang, J., Sun, J., Wang, C., Yang, Z.: Social influence analysis in large-scale networks. In: Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, pp. 807–816. ACM, New York (2009)
30. tau Yih, W., Toutanova, K., Platt, J., Meek, C.: Learning discriminative projections for text similarity measures. In: Proceedings of the 15th Conference on Computational Natural Language Learning, CoNLL'11 (2013)
31. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press (2002)
32. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, WSDM '10, pp. 261–270. ACM, New York (2010)
33. Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th International Conference on World Wide Web, WWW '07, pp. 221–230. ACM, New York (2007)
34. Zhang, J., Tang, J., Li, J.: Expert finding in a social network. In: Proceedings of the 12th International Conference on Database Systems for Advanced Applications, DASFAA '07, pp. 1066–1069. Springer (2007)
35. Zhu, H., Cao, H., Xiong, H., Chen, E., Tian, J.: Towards expert finding by leveraging relevant categories in authority ranking. In: Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM '11 (2011)
36. Zhu, J., Huang, X., Song, D., Ruger, S.: Integrating multiple document features in language models for expert finding. Knowl. Inf. Syst. **23**, 29–54 (2010)
37. Zhu, H., Chen, E., Cao, H.: Finding experts in tag based knowledge sharing communities. In: Proceedings of the 5th International Conference on Knowledge Science, Engineering and Management, KSEM'11, pp. 183–195. Springer, Berlin (2011). doi:10.1007/978-3-642-25975-3_17