Towards Expert Finding by Leveraging Relevant Categories in Authority Ranking

Hengshu Zhu[1,2] Huanhuan Cao[2] Hui Xiong[3] Enhong Chen[1] Jilei Tian[2]
[1]University of Science and Technology of China [2]Nokia Research Center [3]Rutgers University
[1]{zhs, cheneh}@ustc.edu.cn [2]{happia.cao, jilei.tian}@nokia.com [3]hxiong@rutgers.edu

ABSTRACT

How to improve authority ranking is a crucial research problem for expert finding. In this paper, we propose a novel framework for expert finding based on the authority information in the target category as well as the relevant categories. First, we develop a scalable method for measuring the relevancy between categories through topic models. Then, we provide a link analysis approach for ranking user authority by considering the information in both the target category and the relevant categories. Finally, the extensive experiments on two large-scale real-world Q&A data sets clearly show that the proposed method outperforms the baseline methods with a significant margin.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.3.5 [**Online information services**]: Web-based services

General Terms

Algorithms, Experimentation

Keywords

Authority Ranking, Expert Finding, Category Relevancy, Topic Models

1. INTRODUCTION

A critical challenge in knowledge sharing social networks, such as online forums and Question Answering (Q&A) communities is how to find experts, i.e., a group of authoritative users with special skills or knowledge for a specific category. 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.

However, when performing authority ranking for expert finding, most of the state-of-the-art works only take the information in the target category into consideration. Indeed, every 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.

To this end, we propose to exploit the information in both target and relevant categories for improving the performance of authority ranking. The first task along this line is to measure category relevancies. In this paper, we propose to exploit topic models for representing categories as topic distributions and then measure the relevancies between categories by normalized Kullback Leibler (KL) divergence. In addition, we develop a link analysis approach, which is based on the Topical Random Surfer model [4], to collectively exploit the information in both target and relevant categories for authority ranking. 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.

2. PROBLEM STATEMENT

In this paper we propose a new framework for expert finding, namely, category relevancy based authority ranking. To be specific, here we first introduce the traditional authority ranking problem, and then formally define the problem of category relevancy based authority ranking.

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 the posts which are labeled with category c in S.
- $E_c = \{e_{ij}\}$ is a set of directed edges, where e_{ij} indicates that user u_j replied the posts which are labeled with category c and made by user u_i in S.
- $W_c = \{w^c_{ij}\}$ is a set of weights for the edges in E_c, where w^c_{ij} indicates the frequency that user u_j replied the posts which are labeled with category c and made 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. In this way, 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 first present some notations as follows.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

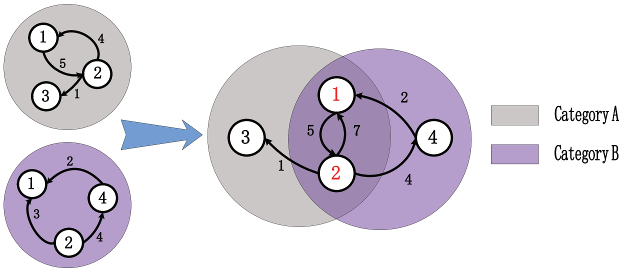


Figure 1: An example of extended category link graph in a Q&A community. A node denotes a user, and an edge from user u_i to user u_j denotes that u_j has answered a question posted by u_i.

Definition 1. (Extended Category Set, Extended Category Link Graph). *An **extended category set** $\Upsilon_c = \{c\} \cup R_c$, where R_c denotes the set of relevant categories of category c.*

*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 $V_{\Upsilon_c} = \bigcup_{c' \in \Upsilon_c} V_{c'}$, $E_{\Upsilon_c} = \bigcup_{c' \in \Upsilon_c} E_{c'}$, and $W_{\Upsilon_c} = \{w_{ij} | w_{ij} = \sum_{c' \in \Upsilon_c} w_{ij}^{c'}\}$ is the corresponding weight set.*

Figure 1 illustrates an example of extended category link graph in a Q&A community. With above notions, the problem of category relevancy based authority ranking is formally defined as follows.

Definition 2. (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}.*

Therefore, 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}. In the following sections, we present the technical details of our solutions for the two sub-problems, respectively.

3. INFERRING CATEGORY RELEVANCY THROUGH TOPIC MODELS

In this paper, we propose to leverage topic models for inferring category relevancies. The basic assumption is that two categories are relevant because their probabilities of belonging to the same latent topic are similar. For example, the categories "Singing", "Pop Music" and "Instruments" are related because they all belong to the latent topic *Music*.

3.1 Inferring Latent Topics by LDA

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. We first define a *user interactive log* consists of a set of category labels where the user made or replied the posts with these category labels and the corresponding frequencies. Then intuitively, if we take category labels as words, take user interactive logs 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) [2] in our approach. According to LDA, a user interactive log L_i is generated as follows. Firstly, a prior topic distribution θ_i is generated from a prior Dirichlet distribution α. Secondly, a prior category distribution ϕ_i is generated from a prior Dirichlet distribution β. Therefore, for the j-th category c_j in L_i, the model generates a topic $z_{i,j}$ from θ_i and then generates c_j from ϕ_i.

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 $P(z_i|c)$ by $P(z_i|c) = \frac{P(c|z_i)P(z_i)}{\sum_i P(c,z_i)}$, where $P(c|z_i)$ and $P(z_i)$ can be obtained from LDA. In this paper, we use Gibbs sampling method to estimate $P(c|z_i)$ and $P(z_i)$. After several rounds of Gibbs sampling, we can get the estimated value $\widetilde{P}(c|z_i)$ by $\widetilde{P}(c|z_i) = \frac{n_i^{(c)}+\beta}{n_i^{(\cdot)}+|C|\beta}$, 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 by $\widetilde{P}(z_i) = \frac{n_i^{(\cdot)}}{\sum_i n_i^{(\cdot)}}$

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 [1] to estimate Z.

3.2 Building Category Relevancy Matrix

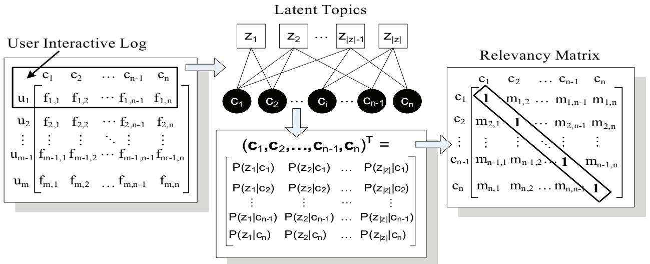

Latent Topics
z1 z2 … z|z|-1 z|z|
User Interactive Log
c1 c2 … cn-1 cn
u1 f1,1 f1,2 … f1,n-1 f1,n
u2 f2,1 f2,2 … f2,n-1 f2,n
um-1 fm-1,1 fm-1,2 … fm-1,n-1 fm-1,n
um fm,1 fm,2 … fm,n-1 fm,n
c1 c2 … ci … cn-1 cn
(c1,c2,…,cn-1,cn)T =
P(z1|c1) P(z2|c1) … P(z|z||c1)
P(z1|c2) P(z2|c2) … P(z|z||c2)
P(z1|cn-1) P(z2|cn-1) … P(z|z||cn-1)
P(z1|cn) P(z2|cn) … P(z|z||cn)
Relevancy Matrix
c1 c2 … cn-1 cn
c1 1 m1,2 … m1,n-1 m1,n
c2 m2,1 1 … m2,n-1 m2,n
cn-1 mn-1,1 mn-1,2 … 1 mn-1,n
cn mn,1 mn,2 … mn,n-1 1

Figure 2: Category relevancy matrix generation.

By utilizing LDA, each category c can be represented as a Z-dimension vector of topic distribution $P(z|c)$. Thus, the task of estimating category relevancy is converted to calculate the distance between vectors. In this paper, we propose to use normalized Kullback Leibler (KL) divergence, which is an asymmetric measure, for measuring category relevancies. The KL-divergence from category c_i to category c_j is computed by $KL(c_i||c_j) = \sum_z P(z|c_i)log\frac{P(z|c_i)}{P(z|c_j)}$.

Then we calculate the relevancy between categories c_i and c_j by $Rel(c_i||c_j) = 1 - \frac{KL(c_i||c_j)}{Max(KL(c_j))}$, where $Max(KL(c_j))$ denotes the maximum KL-divergence from other categories to category c_j. The bigger $Rel(c_i||c_j) \in [0,1]$, the more relevant c_i is for c_j.

After calculating the relevancies between each pair of categories, we can obtain the category relevancy matrix $M_C = \{m_{ij} = Rel(c_i||c_j)\}$, where $i,j \in [1,n]$. Figure 2 illustrates an example of generating the category relevancy matrix.

From M_C, we can easily find the relevant categories for a given category through a predefined relevancy threshold τ. In Section 6, we analyze the robustness of expert finding when given varying parameters τ and Z.

4. AUTHORITY RANKING THROUGH LINK ANALYSIS

By finding relevant categories from the category relevancy matrix, we can build the extended category link graph for a 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. To this end, we extend the Topical Random Surfer (TRS) model [4] to rank user authority in extended category link graphs for considering 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. Therefore, we have the following equations according to the TRS model.

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

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, $P(u',c'|u,c,*)$ 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 $*$, $D(u,u') = \frac{w_{u,u'}}{\sum_{u^\star:u\rightarrow u^\star} w_{u,u^\star}}$ and $\psi_{u,c} = P(c|u) = P(z|u)P(c|z)$ can be directly estimated by the LDA model trained in the stage of calculating category relevancies.

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' by

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

Therefore, 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 demonstrate the experimental results of 1) the performance comparison between our **C**ategory **R**elevancy based **A**uthority **R**anking (CRAR) approach and baselines, 2) robustness analysis of parameter setting.

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. [3]. 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 (`htpp://wenda.tianya.cn`). This data set contains 595 categories, more than 1.3 million questions, and 5.5 million answers posted by 274,896 users. In both data sets, all questions are resolved questions which contain a best answer voted by the question author. Moreover, each data set contains a predefined two-level category taxonomy. To avoid category overlap, we only use the leaf categories in the taxonomy in the experiments. In total, there are 94 leaf categories in the Yahoo! Answers data set and 595 leaf categories in the Tianya wenda data set.

Benchmark Methods. To evaluate the performance of the CRAR, we chose three baseline methods as follows. *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. *HITS* is an iterative approach which assigns two scores for each node in the category link graph, namely, hub score and authority score. *ExpertiseRank* [5] is extended from PageRank. *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.

Evaluation Metrics. 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. In the experiments, K is 10. *Mean Reciprocal Rank* (MRR) is the multiplicative inverse of the rank of the first mined authoritative user in each category. *Mean Average Precision* (MAP) is the mean of the average precision scores for each category.

Since both data sets have no principle benchmark for who are real authoritative users for a given category, we manually inspect the expert finding results. To be specific, firstly we carry out each measuring approach to find top K users as expert candidates for all target categories. Then, for each mined expert candidate u for category c, we ask three human evaluators to check whether u is a real expert for the category c by comprehensively considering the interactive history of u including the number of posted answers, the number of best answers, the voting from another users for the posted answers. Each identified authoritative user is voted by three evaluators with label **Yes** (the user is a real expert) or **No** (the user is not a real expert). It is worth noting that when the evaluators count 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 consider the answers with the category c.

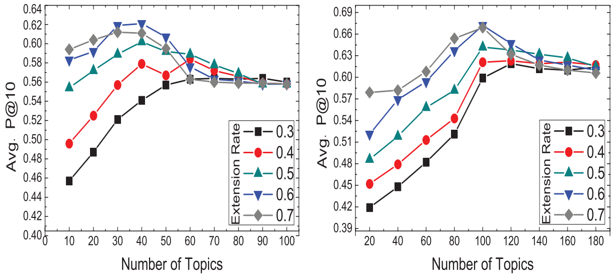
Table 1: The performance of expert finding.

Yahoo! Answers			
	Avg. P@10	**MRR**	**MAP**
Degree	0.434	0.853	0.642
HITS	0.547	0.885	0.699
ExpertiseRank	0.558	0.915	0.732
TRSO	0.569	0.917	0.753
CRAR	**0.619**	**0.953**	**0.808**

Tianya Wenda			
	Avg. P@10	**MRR**	**MAP**
Degree	0.523	0.883	0.687
HITS	0.586	0.916	0.724
ExpertiseRank	0.606	0.935	0.756
TRSO	0.625	0.942	0.771
CRAR	**0.669**	**0.973**	**0.828**

Overall Results of Expert Finding. According to the method introduced in [1], the numbers of topics Z are set to be 30 for the Yahoo! Answers data set and 100 for the Tianya Wenda data set. The two parameters α and β were empirically set to be $50/Z$ and 0.2. 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* [4]. Table 1 shows the average experimental results for all test categories with respect to different metrics. From this table we can see that our approach consistently outperforms other baselines with respect to varying metrics on both data sets. Moreover, we also observe that the topical analysis in original category link graphs can only slightly improve the performance of expert finding than ExpertiseRank. It is because that 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.

Robustness Analysis. The CRAR approach needs two parameters, namely, the latent topic number Z and the extension rate τ of categories. Figure 3 (a) and (b) show the Avg. P@10 of CRAR with varying topic numbers and extension rates for each data set, respectively. From these figures we can see that the setting of Z in this paper estimated by perplexity is reasonable. Moreover, we also can find that the performance of CRAR for expert finding is stable for extension rates with the large topic numbers. However, if a small topic number is used, the extension rate can dramatically impact the performance of CRAR. The phenomenon is reasonable because large topic numbers will cause stricter relevancy metrics while small topic numbers will make the relevancy metric weak. Then, a number of the irrelevant categories will be involved as noise information and will dramatically impact the performance of expert finding. In another case, if the relevancy metric which are strict enough the benefit from other relevant categories is very limited and the performance of CRAR is similar as the TRSO.



(a) (b)

Figure 3: The Avg. P@10 of expert finding versus varying numbers of topics and extension rates in (a) Yahoo! Answers, (b) Tianya Wenda.

6. CONCLUDING REMARKS

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 relevance by utilizing topic models and KL-divergence. Then, a multiple-category-based link analysis approach was extended from the TRS model for ranking user authority in extended category link graphs. Finally, we performed extensive experiments on two large-scale real-world Q&A data sets and results clearly show that our CRAR approach can significantly improve the performance of authority ranking for expert finding.

Acknowledgement. This work is supported by grants from Natural Science Foundation of China (No. 61073110), Key Program of National Natural Science Foundation of China (No. 60933013), National Major Special Science & Technology Projects (No. 2011ZX04016-071), Research Fund for the Doctoral Program of Higher Education of China (No. 20093402110017) and Nokia.

7. REFERENCES

[1] T. Bao, H. Cao, E. Chen, J. Tian, and H. Xiong. An unsupervised approach to modeling personalized contexts of mobile users. In *ICDM'10*.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Lantent dirichlet allocation. In *Journal of Machine Learning Research*.

[3] Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. In *SIGIR'08*.

[4] L. Nie, B. D. Davison, and X. Qi. Topical link analysis for web search. In *In SIGIR'06*.

[5] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *WWW'07*.