Topic-Sensitive Probabilistic Model for Expert Finding in Question Answer Communities

Guangyou Zhou, Siwei Lai, Kang Liu, and Jun Zhao
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing 100190, China
{gyzhou, swlai, kliu, jzhao}@nlpr.ia.ac.cn

ABSTRACT

In this paper, we address the problem of expert finding in community question answering (CQA). Most of the existing approaches attempt to find experts in CQA by means of link analysis techniques. However, these traditional techniques only consider the link structure while ignore the topical similarity among users (askers and answerers) and user expertise and user reputation. In this study, we propose a topic-sensitive probabilistic model, which is an extension of PageRank algorithm to find experts in CQA. Compared to the traditional link analysis techniques, our proposed method is more effective because it finds the experts by taking into account both the link structure and the topical similarity among users. We conduct experiments on real world data set from Yahoo! Answers. Experimental results show that our proposed method significantly outperforms the traditional link analysis techniques and achieves the state-of-the-art performance for expert finding in CQA.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval - *information filtering , selection process*; H.3.5 [**Information Systems and Applications:**]: Web-based services

General Terms

Algorithms, Experimentation, Performance

Keywords

Expert Finding, PageRank, Yahoo! Answers

1. INTRODUCTION

Community question answering (CQA) is a particular form of online service for leveraging user-generated content, which has gained increasing popularity in recent years. These online services, such as Yahoo! Answers[1] and Live QnA[2], provide a platform for users to ask and answer questions.

[1]http://answers.yahoo.com/
[2]http://qna.live.com/

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

Unfortunately, the quality of answers has high variance: ranging from very high to low quality, sometimes abusive content or even spam [1]. Although CQA provides many mechanisms for community feedback ("thumbs up" and "thumbs down" votes), such community feedback requires some time to accumulate, and often remains sparse for obscure or unpopular topics. We analyze a large sample of Yahoo! Answers data, fewer than 20% of all questions have any user votes cast for any of the answers. Therefore, it is desirable to automatically find experts in CQA, so as to route the newly posted questions to the appropriate experts, who can provide high quality answers to these questions [3, 7, 14]. Finally, the overall answer quality can be substantially improved.

Expert finding in CQA is the task of finding users who can provide a large number of high quality, complete, and reliable answers [18], which has recently gained a wide interest in NLP and IR communities [3, 7, 8, 24]. These existing approaches identify the experts by means of link analysis techniques such as PageRank [16] and HITS [9], or their variants. However, the traditional link analysis techniques only consider the link structure while *ignore the topical similarity between askers and answerers*.

In this paper, we propose a topic-sensitive probabilistic model for expert finding in CQA. Given a set of users in CQA, we first automatically distill the topics that users are interested in by analyzing the content of their asked questions and answered questions. Based on the topics distilled, topic-sensitive question-answer relationships between askers and answerers are constructed. Then, we measure the expert saliency score by taking into account both the link structure and the topical similarity between askers and answerers.

To the best of our knowledge, it is the first extensive and empirical study of expert finding in CQA by taking into account both the link structure and the topical similarity between askers and answerers. Although topical similarity information has been proved very effective for web search [6, 15, 23], our goal is to capture the topical similarity between askers and answerers rather than to calculate the topical similarity of web contents. We focus on a substantially different task and model formulation. To date, little work has been made regarding topical similarity among users in studies of expert finding in CQA, which remains an under-explored research area. This paper is thus designed to fill the gap. Specifically, we make the following contributions:

- We automatically distill the topics that users are interested in by analyzing the content of their answered questions (in subsection 2.1).
- We propose a topic-sensitive expert finding method by taking into account both the link structure and the topical similarity between askers and answerers (in subsection 2.4).
- Finally, we conduct experiments on CQA data set. The re-

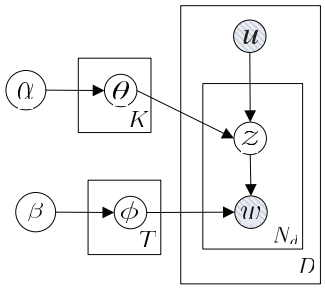

Figure 1: The graphical model for the user-topic model using plate notation.

sults show that our proposed method significantly outperforms the traditional methods (in section 3).

The rest of this paper is organized as follows. Section 2 presents our proposed method. Experimental results are presented in Section 3. Finally, we conclude with ideas for future work in Section 4.

2. OUR METHOD

2.1 Topic Distillation

Topic distillation aims to automatically identify the topics that users (askers and answerers) are interested in based on the user profiles.[3] Because our data set is large, it is only feasible to use fully unsupervised or weakly supervised methods to automatically discover topics. In this paper, we use the widely studied topic model−−Latent Dirichlet Allocation (LDA) [2] to identify the latent topic information from the large scale question-answer collection.

Although we could also apply LDA to distill the topics from questions by treating each question as a single document, this direct application would most likely not work well because questions are very short (average 11.2 words for each question), often containing only a single sentence [25]. To overcome this difficulty, we propose a user-topic model shown in next subsection.

2.1.1 User-Topic Model

Figure 1 illustrates the generative process with a graphical representation of user-topic model. For readers not familiar with plate notation, shaded and unshaded variables indicate observed and latent variables, respectively. An arrow indicates a conditional dependency between variables and plates (the boxes in the figure) indicate repeated sampling with the number of repetitions given by the variable in the bottom. In our user-topic model, D in the figure refers to all user profile, and N_d refers to the number of words in a specific user profile. From the figure, we can see that the main characteristic of user-topic model is that it is user-centric in a general generative manner. Each user is considered as a pseudo-document which represents the user profile. Each user is associated with a multinomial distribution over topics, represented by θ. Each topic is associated with a multinomial distribution over words, represented by ϕ. The multinomial distribution θ and ϕ have a symmetric Dirichlet prior with hyperparameters α and β.

It is interesting to note the difference between our user-topic model and the previous author-topic model in [20, 22]. In author-topic model, the authorship of an arbitrary word in the multi-author document is not known so that author-topic model assumes a uniform contribution of all documents authors. However, in our problem setting, the user of each user profile in CAQ is explicitly represented. In fact, the user information of word is important to precisely identify user interests.

[3]Here, a user profile refers to the questions asked and answered by the user.

2.1.2 Model Inference

The user-topic model includes two sets of parameters −− the K user-topic distributions θ, and the T topic distribution ϕ −− as well as the latent variables corresponding to the assignments of individual words to topics z and user u. The Expectation-Maximization (EM) algorithm [4] is a standard technique for estimating parameters. However, this method is susceptible to local maxima and computationally inefficient [2]. We use an alternative parameter estimation strategy, proposed by Griffiths and Steyvers [5], using Gibbs sampling. Instead of estimating the parameters directly, we evaluate the posterior distribution on just u and z and then use the result to infer θ and ϕ. For each word, the topic and user assignment are sampled from:

$$p(z_i = j, u_i = k | w_i = w, \mathbf{z}_{-i}, \mathbf{u}_{-i}) \propto \frac{C^{WT}_{wj} + \beta}{\sum_{w'} C^{WT}_{w'j} + \beta|\mathcal{V}|} \cdot \frac{C^{UT}_{kj} + \alpha}{\sum_{j'} C^{UT}_{kj'} + \alpha T} \quad (1)$$

where $z_i = j$ and $u_i = k$ represent the assignments of the ith word in the user profile to topic j and user k respectively, $w_i = w$ represents the observation that the ith word is the ith word in the lexicon, and $\mathbf{z}_{-i}$, $\mathbf{u}_{-i}$ represent all topic and user assignments not including the ith word. Furthermore, C^{WT}_{wj} is the number of times word w is assigned to topic j, not including the current user profile, and C^{UT}_{kj} is the number of times user k is assigned to topic j, not including the current user profile, and $|\mathcal{V}|$ is the size of the lexicon.

After parameter estimation, the algorithm only needs to keep track of a $|\mathcal{V}| \times T$ (word by topic) count matrix, and a $K \times T$ (user by topic) count matrix, both of which can be represented efficiently in sparse format. From these count matrices, we can easily estimate the word-topic distributions ϕ and user-topic distribution θ as follows:

$$\phi_{wj} = \frac{C^{WT}_{wj} + \beta}{\sum_{w'} C^{WT}_{w'j} + \beta|\mathcal{V}|} \quad (2)$$

$$\theta_{kj} = \frac{C^{UT}_{kj} + \alpha}{\sum_{j'} C^{UT}_{kj'} + \alpha T} \quad (3)$$

where ϕ_{wj} is the probability of using word w in topic j, and θ_{kj} is the probability of using topic j by user k. These values corresponds to the predictive distributions over new words w and new topics z conditioned on w and z.

In these two matrices, we can row normalize user-topic matrix as $\theta^{'}$ such that $\| \theta^{'}_k \|_1 = 1$ for each row $\theta^{'}_k$. Each row of matrix $\theta^{'}$ is the probability distribution of k's interest over the T topics, e.g., each element $\theta^{'}_{kj}$ denotes the probability that user k is interested in topic j ($p(j|k) = \theta^{'}_{kj}$).

2.2 PageRank for Expert Finding

Based on the topics distilled in subsection 3.1, a directed graph $G = (V, E)$ is formed with the topic-specific question-answer relationships among users. V is a set of nodes representing users (askers and answerers). A directed edge $e \in E$ where $e = (u_i, u_j)$, $u_i \in V$ and $u_j \in V$, indicates that user u_j answers the questions of user u_i. Each edge $e_{ij} \in E$ is associated with an affinity weight $f(i \to j)$ between u_i and u_j. The weight is computed as follows:

$$f(i \to j) = |Q(i) \cap A(j)| \quad (4)$$

where $Q(i)$ is the set of questions asked by u_i, $A(j)$ is the set of questions answered by u_j. Two users are connected if their affinity weight is larger than 0 and we let $f(i \to i) = 0$ to avoid self transition.[4]

The transition probability from u_i to u_j is then defined by normalizing the corresponding affinity weight as follows:

$$p(i \to j) = \begin{cases} \frac{f(i \to j)}{\sum_{k=1}^{|V|} f(i \to k)} & \text{if } \sum f \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where $p(i \to j)$ is usually not equal to $p(j \to i)$, and $\sum_{j'} p(i \to j') = 1$. We use the row-normalized matrix $\widetilde{M} = [\widetilde{M}_{ij}]_{V \times V}$ to describe G with each entry corresponding to the transition probability.

$$\widetilde{M}_{ij} = p(i \to j) \quad (6)$$

In order to make the graph fulfill the property of being aperiodic and $\widetilde{M}$ be a stochastic matrix, the rows with all zero elements are replaced by a smoothing vector with all elements set to $1/|V|$.

Based on the matrix $\widetilde{M}$, the saliency score $R(u_i)$ for u_i can be deduced from those of all other users linked with it and it can be formulated in a recursive manner as in the PageRank algorithm.

$$R(u_i) = \lambda \sum_{j:u_j \to u_i} R(u_j) \cdot \widetilde{M}_{ji} + (1-\lambda)\frac{1}{|V|} \quad (7)$$

where $\lambda \in [0, 1]$ is a damping factor. The damping factor indicates that each vertex has a probability of $(1 - \lambda)$ to perform random jump to another vertex within this graph. The saliency score are obtained by running equation (7) iteratively until convergence.

2.3 Topical PageRank for Expert Finding

In equation (7), the second term is set to be the same value $1/|V|$ for all vertices within the graph, which indicates that there are equal probabilities of random jump to all vertices. However, Haveliwala [6] and Nie et al. [15] proposed a topical PageRank-like algorithm (TPR) and argued that the second term in equation (7) should be set to be non-uniformed. The assumption is that if we assign larger probabilities to some vertices, the final saliency score will prefer these vertices.

The idea of TPR is to run PageRank for each topic separately. Each topic-specific PageRank prefers those users with high relevance to the corresponding topic. Formally, for a specific topic z, we will assign a topic-specific preference value $p_z(u)$ to each user u as its random jump probability $\sum_{u \in V} p_z(u) = 1$. The users who are interested in topic z will be assigned larger probabilities when performing the PageRank. Given a topic z, the TPR-like saliency score are defined as follows:

$$R_z(u_i) = \lambda \sum_{j:u_j \to u_i} R_z(u_j) \cdot \widetilde{M}_{ji} + (1-\lambda) p_z(u_i) \quad (8)$$

The setting of preference value $p_z(u_i)$ in equation (8) will have great influence to TPR. In this paper, we set $p_z(u_i) = p(z|u_i) = \theta'_{iz}$. A large $R_z(u_i)$ indicates a user u_i is a good candidate expert in topic z.

2.4 Topic-Sensitive Expert Finding

The TPR ignores the topical similarity among users when setting the affinity weight. The affinity weight is set by counting the number of questions answered by the two users for a given user's asked questions. For example in Figure 2, the nodes in the figure correspond to users (askers and answerers) and the directed edges represent the question-answer relationships, where $|Q(a) \cap A(b)| = 500$, $|Q(a) \cap A(c)| = 300$. In this case, the transition probability from u_a to u_b is 1.67 times of that of u_a to u_c. As a result, u_b may get the higher ranks than u_c by this topical similarity-free affinity weight although u_b is not interested in topic "coffee making". In other words, topical similarity-free propagation may cause the scores to be off-topic.

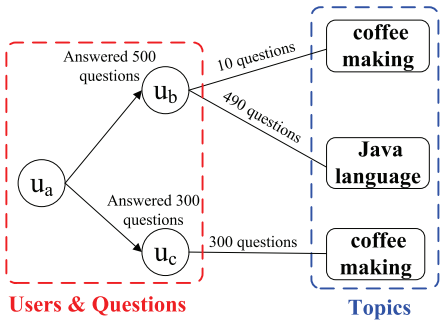

Figure 2: A motivated example in Yahoo! Answers.

In this paper, we propose a topic-sensitive random surfer model (TSPR) by considering the topical similarity among users when setting the affinity weight. The topic-sensitive random surfer model on graph G computes the expert as follows: the random surfer visits each user with certain probability by following the appropriate edge in G. The topic-sensitive propagation method differentiates itself from PageRank and PTR in that the random surfer performs a topic-sensitive random walk (e.g., the transition probability from one user to another is topic-sensitive). By doing so, we can essentially construct a topic-sensitive question-answer relationships between askers and answerers. Given a topic z, the transition probability from the asker u_i to the answerer u_j is defined as:

$$p_z(i \to j) = \begin{cases} \frac{f(i \to j) \times sim_z(i \to j)}{\sum_{k=1}^{|V|} f(i \to k) \times sim_z(i \to k)} & \text{if } \sum f \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where $\sum_{j'} p_z(i \to j') = 1$. $sim_z(i \to j)$ is the similarity from u_i to u_j in topic z. In this paper, we propose to use normalized Kullback Leibler (KL) divergence [10], which is an asymmetric measure. The KL-divergence from u_i to u_j in topic z is computed by $p_{KL}(u_i||u_j) = p(z|u_i)\log\frac{p(z|u_i)}{p(z|u_j)}$, where $p(z|u_i) = \theta'_{iz}$. Then we calculate $sim_z(i \to j)$ as follows:

$$sim_z(i \to j) = \frac{1}{2}\Big\{p_{KL}(u_i||u_j) + p_{KL}(u_j||u_i)\Big\} \quad (10)$$

The larger $sim_z(i \to j) \in [0, 1]$, the more similarity from u_i and u_j in topic z.

Then the new row-normalized matrix $\widetilde{M}^*$ is defined as follows:

$$\widetilde{M}^*_{ij} = p_z(i \to j) \quad (11)$$

Given a topic z, the final TSPR-like saliency score is computed based on the following iterative form:

$$R^*_z(u_i) = \lambda \sum_{j:u_j \to u_i} R^*_z(u_j) \cdot \widetilde{M}^*_{ji} + (1-\lambda) p_z(u_i) \quad (12)$$

For implementation, the initial scores of all users are set to 1 and the iteration algorithm in equation (12) is used to compute the new scores of the users. Usually the convergence of the iteration

[4] In CQA, the users cannot answer their own questions.

is achieved when the difference between the scores computed at two successive iterations for any users falls below a given threshold (0.00001 in this paper).

After ranking the users by using the TSPR or other methods, we select top N users for each topic as topical candidate experts. In this paper, we empirically set N to 1, 5, and 10 shown in the experimental section.

3. EXPERIMENTS

3.1 Data Set

Yahoo! Answers web service supplies an API to allow web users to crawl the existing question answer archives and the corresponding user information from the website. We crawl the data set from Yahoo! Answers, the data set consists of 237,083 resolved questions, and 593,107 answers posted by 286,053 users. In this paper, for all resolved questions, the information of each question includes:

(1) Texts of question and the associated answers, with stop words being excluded[5] and the words being stemmed.[6]

(2) User' IDs of all questions and answers.

(3) Users' rating information (e.g., thumbs up, thumbs down, the best answers and so on.).

Since there is no available benchmark for expert finding for a given topic in CQA, we manually inspect the expert finding results. For each candidate expert u for topic z, we ask two annotators to check whether u is a real expert for the given topic. In this process, the annotators are given the top topic words and user profile. Each identified expert is voted by two annotators with label **Yes** (the user is a real expert for the given topic) or **No** (the user is not a real expert for the given topic). If a conflict happens, a third person will make judgement for the final result. The Cohen's Kappa coefficients of the T topics range from 0.51 to 0.77, showing fair to good agreement.

3.2 Evaluation Metrics

To evaluate the performance of expert finding, we use the three widely studied metrics in information retrieval.

Mean Average Precision (MAP): This metric is the mean of the average precision scores for each topic.

Mean Reciprocal Rank (MRR): This metric is the multiplicative inverse of the rank of the first retrieved expert for each topic.

Average Precision@n (Avg. P@n): This metric denotes the average ratio of the relevant experts in top n identified experts for each topic.

3.3 Parameter Setting

We have several parameters: i.e., Dirichlet hyper-parameters α, β, topic number T, damping factor parameter λ used in PageRank; In this paper, we set Dirichlet priors $\alpha = 50/T$, and $\beta = 0.05$ as Griffiths and Steyvers [5]. We run LDA with 200 iterations of Gibbs sampling. After trying a few different numbers of topics, we empirically set $T = 15$. We choose these parameter settings because they give coherent and meaningful topics for our data set.

For parameter λ, we conduct an experiment on a small development set to determine the best value among 0.1, 0.2, $\cdots$, 0.9 in terms of MAP. This set is also extracted from Yahoo! Answers, and it is not included in the evaluation set (described in subsection 4.1). We find that $\lambda = 0.2$ is the optimal parameter for PR, TPR and TSPR.

[5]http://truereader.com/manuals/onix/stopwords1.html

[6]http://tartarus.org/martin/PorterStemmer/

#	Methods	MAP	MRR	Avg. P@10
1	PR	0.435	0.726	0.331
2	HITS	0.397	0.704	0.266
3	InD	0.369	0.655	0.242
4	ER	0.481	0.773	0.385
5	CB	0.513	0.787	0.392
6	TPR	0.506	0.781	0.388
7	**TSPR**	**0.543**	**0.821**	**0.430**

Table 1: Comparison of expert finding for different methods.

3.4 Experimental Results

3.4.1 Comparison with different methods

To demonstrate the effectiveness of our proposed TSPR method, comparisons against some previous work are also included:

- **PageRank (PR):** This method finds the experts with only link structure taken into account [16].
- **HITS:** Jurczyk and Agichtein [7] proposed to find experts in CQA and estimated the ranking scores by using HITS algorithm.
- **InDegree (InD):** This method identifies the experts based on the number of best answers described in Bouguessa et al. [3]
- **ExpertiseRank (ER):** Zhang et al. (2007) proposed a PageRank-like algorithm called ExpertiseRanking to rank experts in an expertise network considering how many users involved in asking and answering questions.
- **Competition-Based (CB):** Liu et al. [14] proposed to explore the pairwise comparisons inferred from best answer selections to find experts in CQA.
- **TPR:** In subsection 3.2.2, we discuss this method, which is similar to the methods in Haveliwala [6] and Nie et al. [15]. This method focuses on calculating the topical similarity of web contents while we focus on capturing the topical similarity among users.

Table 1 presents the comparison of expert finding for different methods. From this table, we can find that our proposed method significantly outperforms all previous works (row 1, row 2, row 3, row 4, row 5, and row 6 vs. row 7).[7] The results show the effectiveness of the propose method by considering the topical similarity among users, user expertise score and reputation score. Besides, we also note that incorporating the topical preference value into PageRank, the performance can be further improved (row 1 vs. row 6).

3.4.2 Answer quality of the identified experts

In order to further evaluate the effectiveness of our approach, we look at the identified experts and manually evaluated their answers. We expect experts to provide high quality answers for their interested topics, thus answer quality is an indirect evaluation metric. In this paper, we use the quality metric described in Agichtein et al. [1] as the "gold standard" for evaluation. This metric is the confidence score of a binary classifier trained on high and low quality instances. The value of quality score is always between 0 and 1. To avoid manually labeling, we adopt the community and askers' choices used in Li and King [12] to automatically construct a large number of "high quality" and "low quality" instances.

Figure 3 shows the average answer quality scores provided by the identified top 10 experts in three specific topics. *As we can see*

[7]We perform a significant t-test. The comparisons between our method and previous works are significant at $p < 0.05$.

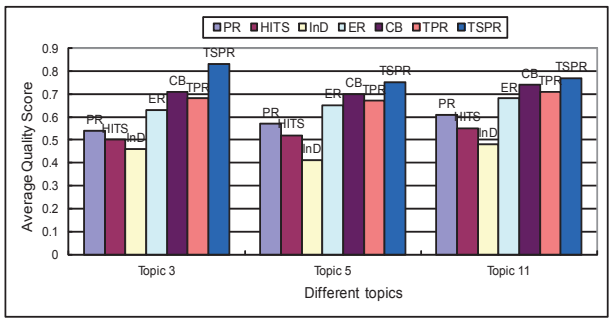

Figure 3: Average answer quality scores provided by the identified experts in three topics.

from this figure, the average answer quality scores of experts using our proposed methods for each topic are generally between 0.74 and 0.83, which is a relatively high quality score. These results represent another source of confirmation concerning the suitability of our approach for finding the experts that contribute significantly to generate high quality answers in CQA. Moreover, such results also indicate that askers are very selective in choosing experts. We can thus recommend the open questions to these experts and enhance the overall quality of content in CQA.

4. CONCLUSION AND FUTURE WORK

In this paper, we propose a topic-sensitive probabilistic model for expert finding in CQA. Compared to the traditional link analysis techniques, our proposed method is more effective because it finds the experts by taking into account both the link structure and the topical similarity between askers and answerers. We conduct experiments on real world data set from Yahoo! Answers. Experimental results show that our proposed method significantly outperforms the traditional link analysis techniques and achieves the state-of-the-art performance.

There are some ways in which this research could be continued. First, we will investigate the proposed method to the full system of CQA (26 categories) or other kinds of data set (e.g., forums and FAQ sites). Second, users' relative saliency scores change over time, so it is necessary to take into account the temporal dimension of questions and answers.

5. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 61070106), the National Basic Research Program of China (No. 2012CB316300), Tsinghua National Laboratory for Information Science and Technology (TNList), Cross-discipline Foundation and the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (No. 5026035403). We thank the anonymous reviewers for their insightful comments.

6. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. 2008. Finding high-quality content in social media. In *WSDM*.

[2] D. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993-1022.

[3] M. Bouguessa, B. Dumoulin, and S. Wang. 2008. Identifying authoritative actors in question-answering forums-the case of Yahoo! Answers. In *KDD*, pages 866-874.

[4] A. P. Dempster, N. M. Laird, D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1-38.

[5] T. Griffiths and M. Steyvers. 2004. Finding scientific topics. *The National Academy of Sciences*, 101:5228-5235.

[6] T. H. Haveliwala. 2002. Topic-sensitive pagerank. In *WWW*.

[7] P. Jurczyk and E. Agichtein. 2007. Discovering authorities in question answer communities by using link analysis. In *CIKM*, pages 919-922.

[8] W. Kao, D. Liu, and S. Wang. 2010. Expert finding in question-answering websites: a novel hybrid approach. In *SAC*, pages 867-871.

[9] J. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604-632.

[10] S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22 (1): 79-86.

[11] J. Lafferty and C. Zhai. 2003. Probabilistic relevance models based on document and query generation. *Language Modeling and Information Retrieval, Kluwer International Series on Information Retrieval*.

[12] B. Li and I. King. 2010. Routing questions to appropriate answerers in community question answering services. In *CIKM*, pages 1585-1588.

[13] Z. Liu, W. Huang, Y. Zheng, and M. Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *EMNLP*, pages 366-376.

[14] J. Liu, Y. -I. Song, and C. -Y. Lin. 2011. Competition-based user expertise score estimation. In *SIGIR*, pages 425-434.

[15] L. Nie, B. D. Davison, and X. Qi. 2006. Topic link analysis for web search. In *SIGIR*.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The pagerank citation ranking: bringing order to the web. *Stanford Digtital Library Technologies Project*.

[17] A. Pal and S. Counts. 2011. Identifying topical authorities in microblogs. In *WSDM*.

[18] A. Pal and J. Konstan. 2010. Expert identification in community question answering: exploring question selection bias. In *CIKM*, pages 1505-1508.

[19] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In *KDD*, pages 569-577.

[20] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. 2004. The author-topic model for authors and documents. In *UAI*, pages 487-494.

[21] D. Schall and F. Skopik. 2011. An Analysis of the Structure and Dynamics of Large-Scale Q/A Communities. In *ADBIS*, pages 285-301.

[22] M. Steyvers, P. Smyth, and T. Griffiths. 2002. Probabilistic author-topic models for informaiton discovery. In *KDD*.

[23] J. Weng, E. -P. Lim, J. Jiang, and Q. He. 2010. TwitterRank: finding topic-sensitive influential twitterers. In *WSDM*.

[24] J. Zhang, M. Ackerman, and L. Adamic. 2007. Expertise networks in online commmunities: structure and algorithm. In *WWW*.

[25] G. Zhou, L. Cai, J. Zhao, and K. Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *ACL*, pages 653–662.