ZhihuRank: A Topic-Sensitive Expert Finding Algorithm in Community Question Answering Websites

Xuebo Liu, Shuang Ye, Xin Li, Yonghao Luo, Yanghui Rao

School of Mobile Information Engineering, Sun Yat-sen University, Zhuhai, P.R.China, 519082

Abstract. Expert finding is important to the development of community question answering websites and e-learning. In this study, we propose a topic-sensitive probabilistic model to estimate the user authority ranking for each question, which is based on the link analysis technique and topical similarities between users and questions. Most of the existing approaches focus on the user relationship only. Compared to the existing approaches, our method is more effective because we consider the link structure and the topical similarity simultaneously. We use the real-world data set from *Zhihu* (a famous CQA website in China) to conduct experiments. Experimental results show that our algorithm outperforms other algorithms in the user authority ranking.

Keywords. Community Question Answering; Expert Finding; PageRank; Latent Topic Modeling

1 Introduction

Community Question Answering(CQA) websites, such as *Quora*[1], *Stackoverflow*[2], *Yahoo!Answer*[3] and *Zhihu*[4], are extremely popular in recent years. In CQA websites, users can raise their own questions, answer the questions posted by others and read the corresponding answers to a question. It is interesting that the CQA websites have strong social characteristics, which are different from the traditional ones, such as *Baidu Knows*[5]. Take *Zhihu* as an example, the content read by a user u in his/her home page depend on the people u have followed (i.e., "friends" of u). Therefore, a user is able to see the questions raised by all his/her friends, their answers and their "liked" answers. *Zhihu* develops rapidly

[1] http://www.quora.com/
[2] http://stackoverflow.com/
[3] https://answers.yahoo.com/
[4] http://zhihu.com/
[5] http://zhidao.baidu.com/

because of the high quality content[6] provided by its users. By the end of March 2015, the number of registered users in *Zhihu* have reached 17 million[7].

Although the CQA website has attracted many users, it brings several challenges to provide high quality services: (1) **Poor expertise matching**: Many questions are difficult to be recommended to the expert with the best-matching interest and ability to answer them, which results in suboptimal answers and delay of satisfaction [1]. (2) **No answer**: Since it is hard for users to find questions that they are really interested in, there are many questions with no answer. (3) **Similar questions**: If a user can not find a satisfied answer for one question, he/she may post a new similar question. Thus, it is necessary to use expert finding algorithms in CQA websites to solve these problems.

Traditional algorithms of expert finding are based on the relationship between users primarily [2,3], which are insufficient to get a unique user authority ranking for each question. For example, the appropriate users to answer different questions may differ. If we rank the user authority based on the relationship between users only, the user who is a movie star will have many followers and achieve a high ranking. However, he/she may not be appropriate to answer the questions about computers. We here propose a new algorithm based on the link analysis technique and topical similarities between users and questions, which can get a unique user authority ranking for each question. In our algorithm, we consider all answers posted by a user as a document, the description and all answers to a question are also treated as a document, then we apply latent Dirichlet allocation (LDA) [4] to extract topics from both users and questions. After generating these topics, we can measure the topical similarity between questions and answers, and rank users by the relationship between users and the topical similarity between users and questions.

The rest of the paper is organized as follows. Section 2 introduces the relative works on user ranking in social networks. In section 3, the *ZhihuRank* algorithm is described in detail. Section 4 illustrates the dataset used and the experimental results. We conclude our paper and make a future plan in section 5.

2 Related work

The existing algorithms computed the user authority in CQA websites by the relationship between users primarily. *Bouguessa et. al.* [5] proposed a method to dig out the expert user, which is based on the best answer the user posted. *Jurczyk and Agichtein* [6] applied HITS algorithm [3] to their works and compute the user ranking by the followed-following relationship. *Zhang et. al.* [7] proposed an algorithm based on the users' specialty. Although the algorithms that analyzing the relationship between users have achieved the desired results, there are some challenging problems that are difficult to be solved by these algorithms, e.g., the expert discovered could not give the satisfied answer to the

[6] `http://www.pcdigest.com/2014/01/zhihupopular/`

[7] `http://appsearch.m.sogou.com/i/m21ywJS-QrY?from=groupmessage&isappinstalled=0`

field he/she is not skilled in. Thus, some other algorithms based on topics were proposed. *Guo et. al.* [8] proposed an algorithm to explore the similarity between users and the askers by the tags of users. *Liu et. al.* [9] used language model and LDA to detect the best answerer.

Recently, there are algorithms that composed the relationship between users and the similarity of topics. *TwitterRank* [10] is a typical algorithm which is able to figure out the influence of Twitter users based on the followed-following relationship and topic similarities. *Zhou et. al.* [11] proposed a TSPR algorithm for expert recommendation in CQA websites. In the TSPR algorithm [12], the topics of users were first extracted by LDA, and then experts were recommended by the number of answers a user posted and the similarity between the user and the asker. *Zhao et. al.* [13] designed a method to generate experts and topics simultaneously by incorporating the users' contribution dynamically. *Chen et. al.* [14] established the user reputation rating system based on user comments. *Pal et. al.* [15] proposed an algorithm based on Gaussian mixture models to identify topical authorities in microblogs. *Liu et. al.* [16] proposed a new topic model for expert recommendation in CQA websites.

Different from these algorithms, our method considers both the link structure of users and the topical similarity between users and questions, which achieved a competitive performance. To the best of our knowledge, it is the first time to consider the topical similarity between users and questions for expert recommendation problems in CQA websites.

3 ZhihuRank Algorithm

ZhihuRank algorithm is able to figure out the user authority ranking in each question, which is based on the relationship between users and the topical similarities between users and questions. Conventionally, the more "likes" a user receives, the higher authority he/she will achieve, which is similar to the ranking approach of websites. It is important to note that the weight of each "likes" given to an answer is different from the others. Assume that both *User B* with higher authority and *User C* with lower authority give "likes" to *User A*, the "likes" from *User B* is often more powerful to improve the authority of *User A* than *User C*. Therefore, we compute of influence of users based on the iteration method of PageRank [2]. In addition, since each question is consisted of different topics and the topics those users are familiar with are varies, we take also the similarity between users and topics into consideration.

3.1 Topic Extraction of Users

We apply LDA [4] to perform the topic extraction of users. LDA is an unsupervised topic extraction model, which is based on the bag of word assumption. It treats each text as a vector whose characteristic of each dimension is the number of a word that appears in the text. Each text can be expressed as the probability distribution of a series of topics and each topic can be expressed as

the probability distribution of a series of words. LDA is a nature model for topic extraction of long text, in which the probability distribution of topics for each text and words for each topic can be estimated by Gibbs sampling algorithm [17]. The detailed process of user topic extraction in *ZhihuRank* is as follows:

Firstly, we consider all answers of a user posted as a text and the one-to-one mapping between a user and its text is established. Meanwhile, all answers to a question are treated as another text. There is also a one-to-one mapping between a question and its text. Secondly, we use LDA to train the text of all users and question, in which, the probability distribution of the topics corresponding to each text θ and probability distribution of the words corresponding to each topic φ can be estimated. Thirdly, we keep θ unchanged and carry out the Gibbs sampling only with the input of user text. Finally, we get a new φ_u and let $UZ = \varphi_u$, which represents the topic distribution of a text.

Definition 1 *UZ: the matrix of $U \times Z$. U is the number of users. Z is the number of topics. UZ_{ij} represents the number of words that assigned to topic z_j appearing in all answers posted by User u_i.*

3.2 User Authority Transition Matrix

In the past researches on social networks, the relationship of following and being followed between users is often used to generate the user authority transition matrix by iterative computation [18]. Different from the traditional social networks, the approval mechanism is introduced into the CQA websites. Conventionally, the more "likes" a user receives, the higher authority he/she will achieve. Meanwhile, the weights of a "likes" given by different users are varied, i.e., the approval by the expert in a certain field is more powerful to improve the authority of the user whom he/she give "likes" to in this field. In addition, the "likes" given by a user who seldom make approval is of higher value in comparison to those who often delivers "likes". Therefore, we consider the users' authority ranking in each topic as a *Markov Chain* [18], and the transition matrix of topic z is shown as follows:

Definition 2

$$T_z(i,j) = \frac{V_{j \to i}}{\sum_{for\ every\ user\ k} V_{j \to i}} \times sim_z(i,j) \tag{1}$$

T_z represents the user authority transition matrix of topic z. $T_z(i,j)$ represents the influence of User i to User j in topic z. $V_{j \to i}$ represents the number of "likes" that User j gives to User i, and the denominator is the summation of the number of "likes" that User j gives to all users. $sim_z(i,j)$ represents the similarity between User i and j in topic z.

Definition 3

$$sim_z(i,j) = 1 - 0.5 \times ((UZ'_{iz} - UZ'_{jz}) \times ln\left(\frac{UZ'_{iz}}{UZ'_{jz}}\right)) \tag{2}$$

UZ' is the row-normalized form of matrix UZ, i.e., the L_1-norm of each row is 1. UZ'_{iz} reflects the degree of interest of User i in topic z. If the degrees of interest of User i and j in topic z are close , both $ln\left(\frac{UZ'_{iz}}{UZ'_{jz}}\right)$ and $UZ'_{iz} - UZ'_{jz}$ tend to approximate 0 while sim tends to approximate 1,otherwise sim will be small. If there is a huge gap between the degrees of interest of User i and j in topic z, the value of $ln\left(\frac{UZ'_{iz}}{UZ'_{jz}}\right)$ will be positive infinity. The larger the value of sim is, the more similar User i and j in topic z will be.

3.3 User Authority Ranking for Each Topic

In Section 3.2, we get the user authority transition matrix iteratively. Next, *ZhihuRank* takes the approval relationship between users and the topical similarity into account to compute the authority ranking of users in topic z:

Definition 4

$$UR_z = \lambda T_z \times UR_z + (1 - \lambda)UZ''_z \tag{3}$$

UR_z represents the user authority ranking of topic z. λ is a weighting parameter between 0 and 1. A lager value of λ indicates that the approval relationship between users has a greater influence on the authority ranking. While a smaller value of λ indicates that the degree of interest of the user to topic z has a greater influence on the authority ranking. T_z is the transition matrix described in Section 3.2. UZ''_z is the column-normalized form of matrix UZ, i.e., the L_1-norm of each column is 1. It represents the degree of interest of each user to topic z.

After convergence, we get the final result of the user authority ranking for each topic.

3.4 Topic Extraction of Questions

We consider all answers to a question as a document. Then we apply LDA trained in Section 3.1, i.e., keep θ unchanged and carry out the Gibbs sampling with the input of question document again. Finally, we get φ_q and let $QZ = \varphi_q$.

Definition 5 *QZ: the matrix of $Q \times Z$. Q is the number of questions. Z is the number of topics. QZ_{ij} represents the number of words that assigned to topic z_j in all the answers of question q_i.*

3.5 User Authority Ranking for Each Question

Since we get matrix QZ (the topic distribution of every question) and UR (the user authority ranking of every topic), multiply the two matrices (i.e., *Bayes*'s rule) then we get the user authority ranking of each question.

Definition 6

$$QR = QZ \times UR \tag{4}$$

QZ represents the topic distribution of each question. UR represents the user authority ranking of each topic. The multiplication result is the user authority ranking of every question.

4 Experimental Analysis

The previous section introduced the details of our algorithm. In this section, the dataset we used and the experimental results are to be shown.

4.1 Dataset

We use the real-world data from *Zhihu* for experiments. *Zhihu* is one of the most popular question answering communities in China. Different from *StackOverflow* and *Yahoo!Answer*, we can get all users who "like" an answer, thus we can get user authority transition matrix based on users' "like" relationship.

We crawled 576 questions, 9043 users and 209309 answers from *Zhihu*, and employed *Jieba Chinese Text Segmentation*[8] to perform the Chinese word segmentation. The detail process of preparing the above dataset is as follows:

- For each question, its description, contents of all its answers and the real ranking of all answers were crawled.
- For each user, the number of friends, followers, answers, "likes" received and the contents of all answers he/she posted were crawled.

4.2 Parameters

When applying LDA in topic extraction, the hyper parameters α and β are set to be 0.1 and 0.01, respectively. We selected 50 topics, performed 1000 times of iterations and finally got these topics.

When computing the user authority ranking of each topic, λ is set to be 0.85 based on cross-validation.

4.3 Evaluation Metrics

To measure the accuracy of different algorithms, two evaluation metrics commonly used in information retrieval were chosen:

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

$$nDCG@K = \frac{1}{Q}\sum_{q\in Q}\frac{\sum_{j=1}^{K}\frac{1}{log_2(j+1)}score(M_{q,j})}{IdealScore(K,q)} \tag{5}$$

  In the above, Q is the set of questions. $M_{q,j}$ is the j-th expert generated by method M for question q. $score(M_{q,j}) = 2^{v(M_{q,j})} - 1$. $v(M_{q,j})$ is the ground truth score for the expert $M_{q,j}$. $IdealScore(K,q)$ is the ideal ranking score of the top K experts for question q.

[8] `https://github.com/fxsjy/jieba`

4.4 Comparison with Baselines

In this part, we will introduce the baselines and the criterion to measure the experimental results. The baselines used are:

- **In-degree by number of followers**: this algorithm measures the authority of users according to the number of followers. The more followers a user has, the higher value of the user authority will be.
- **In-degree by number of "likes"**: the algorithm measures the authority of users according to the number of "likes" received. The more "likes" a user owns, the higher value of the user authority will be.
- **PageRank by number of followers**: the algorithm generates the user ranking by applying PageRank with the number of followers.
- **PageRank by number of "likes"**: the algorithm generates the user ranking by applying PageRank with the number of "likes".
- **Topic PageRank** [11]: the algorithm generates the user ranking of each question according to the following aspects: (1) the user topical similarity between the asker and other users; (2) the number of times that users answered the questions raised by the asker.

For convenience of description, the algorithms are denotes as: *ZhihuRank* (ZR), *In-degree by number of followers* (IDF), *In-degree by number of "likes"* (IDV), *PageRank by number of followers* (PRF), *PageRank by number of "likes"* (PRV) and *Topic PageRank* (TSPR). The following table shows our experimental results:

Alogrithm	**MRR**	**nDCG**
IDF	0.75676	0.85459
IDV	0.74568	0.84730
PRF	0.78899	0.86280
PRV	0.79643	0.86455
TSPR	0.63710	0.77037
ZR	**0.84114**	**0.87893**

Table 1: Performance of expert finding for different methods.

From Table 1, we can observe that the proposed ZR outperformed other methods for both metrics. The results indicate that it is effective to consider the topical similarity between questions and users when computing the user authority ranking.

5 Conclusion

This paper proposed an effective algorithm to estimate the user authority ranking in CQA social networks, which is based on the relationship between

users and the topical similarity between users and questions. We evaluated the algorithm by the dataset from *Zhihu* and experimental results demonstrated the effectiveness of our model when compared to other existing methods.

In the future, we plan to test the algorithm by more dataset, and design a new topic model to cover the shortcomings of LDA in short text topic extraction when there are few answers to a question.

References

1. E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne: Finding high-quality content in social media. In WSDM, (2008).
2. S. Brin and L. Page: The anatomy of a large-scale hypertextual web search engine. Computer Network and ISDN Systems, 30 (1-7) pp. 107-117, (1998).
3. J. M. Kleinberg: Authoritative sources in a hyperlinked environment. Journal of the ACM, 46 (5) pp. 604-632, (1999).
4. D. M. Blei, A. Y. Ng, and M. I. Jordan: Latent dirichlet allocation. Journal of Machine Learning Research, pp. 993-1022, (2003).
5. M. Bouguessa, B. Dumoulin, and S. Wang: Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In KDD, pp. 866-874, (2012).
6. P. Jurczyk and E. Agichtein: Discovering authorities in question answer communities by using link analysis. In CIKM, pp. 919-922, (2007).
7. J. Zhang, M. S. Ackerman, and L. Adamic: Expertise networks in online communities: structure and algorithms. In WWW, pp. 221-230, (2007).
8. J. Guo, S. Xu, S. Bao, and Y. Yu: Tapping on the potential of qa community by recommending answer providers. In CIKM, pp. 921-930, (2008).
9. X. Liu, W. B. Croft, and M. Koll: Finding experts in community-based question answering services. In CIKM, pp. 315-316, (2005).
10. J. Weng, E.-P. Lim, J. Jiang, and Q. He: Twitterrank: finding topic-sensitive influential twitterers. In WSDM, pp. 261-270, (2010).
11. G. Zhou, S. Lai, K. Liu, and J. Zhao: Topic-Sensitive Probabilistic Model for Expert Finding in Question Answer Communities. In CIKM, pp. 1662-1666, (2012).
12. T. H. Haveliwala: Topic-sensitive pagerank. In WWW, pp. 517-526, (2002).
13. T. Zhao, N. Bian, C. Li and M. Li: Topic-level expert modeling in community question answering. In SDM, (2013).
14. B.-C. Chen, J. Guo, B. Tseng, and J. Yang: User reputation in a comment rating environment. In KDD, pp. 159-167, (2011).
15. A. Pal and S. Counts: Identifying topical authorities in microblogs. In WSDM, pp. 45-54, (2011).
16. Y. Liu, M. Qiu, S. Gottipati, F. Zhu, J. Jiang, H. Sun, and Z. Chen: CQARank: Jointly Model Topics and Expertise in Community Question Answering. In CIKM, pp. 99-108, (2013).

17. S. Geman and D. Geman: Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (6) pp. 721-741, (1984).

18. J. R. Norris: Markov chains. Cambridge University Press. (1998).