Knowledge-Based Systems 66 (2014) 136–145

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities

Guangyou Zhou [a,b,*], Jun Zhao [b], Tingting He [a], Wensheng Wu [c]

[a] College of Computer Science, Central China Normal University, Wuhan, China
[b] National Laboratory of Pattern Recognition, Institute of Automation CAS, Beijing, China
[c] 9201 University City Blvd, Charlotte, NC 28223, USA

ARTICLE INFO

Article history:
Received 13 July 2013
Received in revised form 14 April 2014
Accepted 21 April 2014
Available online 9 May 2014

Keywords:
Community Question Answering
Expert Finding
Topic-Sensitive Model
Yahoo! Answers
User-Generated Content

ABSTRACT

In this article, we study the problem of finding experts in community question answering (CQA). Most of the existing approaches attempt to find experts in CQA via link analysis. One primary challenge of expert finding lies in that how to improve authority score ranking based on the user information. However, these existing link analysis techniques largely fail to consider the interests, expertise, and reputation of users (question askers and answerers). To address this limitation, we propose a topic-sensitive probabilistic model, by extending the PageRank algorithm, more effectively find in the community by incorporating link and user analysis into a unified framework. We have conducted extensive experiments using a real world data set from Yahoo! Answers of English language. Results show that our method significantly outperforms the existing link analysis techniques and advances the state-of-the-art performance on expert finding in CQA.

1. Introduction

Community question answering (CQA) is becoming an increasingly popular online service that has proven to be a very effective way of sharing user-generated contents. In this service, users may pose questions and may also answer questions posed by other users. Well-known CQA services include Yahoo! Answers,[1] Live QnA,[2] and StackOverflow.[3]

Like the open Web, such a community-driven, unmoderated question answering service also suffers from the quality issue (in particular, the quality of answers may vary a lot) and is also very vulnerable to abuse and spam [1]. Although CQA provides many mechanisms to solicit community feedback, e.g., by allowing users to vote on the answers, it may take long time to collect such feedback and there may not be any feedback available for less popular topics. In fact, from a large sample of Yahoo! Answers posts, we observe that there were fewer than 20% of questions that have received any votes from the users. Therefore, it is desirable to find another means of finding experts on a subject, other than relying solely on the community feedback. The found experts on a subject can then be directed to provide quality answers to the questions within their expertise. This would greatly improve the overall quality of answers to the questions posed on the CQA service site [2–4].

The goal of expert finding in CQA is to find community users who can provide a large number of high quality, complete, and reliable answers [5]. The problem has spurred great interests in both in NLP and IR communities [2,3,6,7]. One primary challenge of expert finding lies in that how to improve authority score ranking based on the user information. When performing authority score ranking, existing approaches have mainly focused on finding experts by applying link analysis techniques such as PageRank [8] and HITS [9] algorithms. A key limitation of these approaches is that they have largely ignored the topical interests, expertise, and reputation of users, both askers and answerers. As a result, the experts recommended by these approaches often may not be the good candidates for answering given questions.

To address this limitation, in this article, we extend our previous work [10] and propose a topic-sensitive probabilistic model for finding experts in CQA. Given a set of users in the community, we first automatically distill the topics that users are interested in by analyzing the content of the questions they asked or answered. Based on the distilled topics, we may then construct topic-sensitive question–answer relationships between askers and answerers.

* Corresponding author at: College of Computer Science, Central China Normal University, Wuhan, China. Tel.: +86 027 67868318.

E-mail address: gyzhou@nlpr.ia.ac.cn (G. Zhou).

[1] http://answers.yahoo.com/.

[2] http://qna.live.com/.

[3] http://stackoverflow.com/.

http://dx.doi.org/10.1016/j.knosys.2014.04.032
0950-7051/© 2014 Elsevier B.V. All rights reserved.

This in turn enables us to compute the expert saliency score by taking into account both the link structure and the topical similarity between askers and answerers. Building on this, we propose a probabilistic model to rank the candidate experts by taking into consideration both user expertise and reputation.

To the best of our knowledge, this is the first extensive and empirical study on expert finding in CQA by taking into account both the link structure and the topical similarity between askers and answerers. The topical similarity information has been proved to be very effective for ranking web pages in web search [11,12]. Our goal here is to capture the topical similarity between askers and answerers rather than to calculate the topical similarity of web contents. So far, little work has been done to utilize topical similarity among users for expert finding in CQA. Thus our goal is to fill in this gap. Specifically, we make the following contributions:

- We propose a user-topic model and automatically distill the topics that users are interested in by analyzing the content of the questions and answers they posed (Section 3.1).
- We propose a topic-sensitive expert finding method that takes into account both the link structure and the topical similarity between askers and answerers (Section 3.4).
- We propose a probabilistic model to rank the candidate experts by considering the user expertise and user reputation that can more accurately assess the expertise of users (Section 3.5).
- We have conducted experiments using a CQA data set obtained from Yahoo! Answers. The results show that our proposed method significantly outperforms existing techniques (Section 4).

The rest of this article is organized as follows. Section 2 describes the related work on expert finding in CQA. Section 3 presents our proposed methods. Experimental results are presented in Section 4. Finally, we conclude the article and discuss future research directions in Section 5.

2. Related work

The problem of expert finding in CQA has been widely studied and many algorithms have been developed to solve the problem. Existing methods are largely based on link analysis techniques, in particular, PageRank [8] and HITS [9]. For example, Zhang et al. [7] proposed a PageRank-like algorithm called *ExpertiseRanking* to rank experts in an expertise network by considering how many users involved in asking and answering questions. Jurczyk and Agichtein [3] built a graph based on the question–answer relationships between askers and answerers, and adopted the HITS algorithm [9] to rank the users. Agichtein et al. [1] extracted several graph features such as the degree distribution of users and their PageRank, hubs and authority scores to measure each user's relative importance. However, none of these works utilizes the reputation and authority values obtained from link analysis, which has largely affected the performance of expert finding as shown in our experiments.

There are also approaches that explored user interaction history such as the number of questions and answers user have posed, the number of the best answers, votes users have received and so on [2,5,13,6,14,4,15]. Bouguessa et al. [2] proposed to identify authoritative users based on the number of best answers given by the users. Pal and Konstan [5] identified authoritative users based on preferential selection as question selection bias and showed that authoritative users were more selective than other users. Bian et al. [13] proposed a mutual reinforcement approach for jointly modeling user expertise and answer quality. Kao et al. [6] proposed a hybrid method for expert finding in CQA. Zhu et al. [14] proposed a method for expert finding by leveraging relevant categories. Liu et al. [4] incorporated the best answer selection to infer pairwise comparison of users and then applied two-player competition models to estimate the relative expertise of users. Liu et al. [15] proposed a hybrid approach to effectively find experts in CQA by considering user's subjective relevance, user reputation and the authority of a category. However, none of these existing approaches finds the experts by considering the topical similarity between askers and answerers. Although each category from CQA sites like Yahoo! Answers can be regarded as a topic, we find that there are several sub-topics in the same category, which represent the different aspects of the same category. Therefore, modeling the expert finding at the topical level is more appropriate than at the category level.

There are also studies that used the topical-centric link analysis method for expert finding in other domain (e.g., twitter [16]) or keyphrase extraction [17,18]. For example, Haveliwala [11] and Nie et al. [12] proposed topical link analysis for web search. but they focused on calculating the topical similarity of web contents while we focused on capturing the topical similarity among users. Beyond the expert finding, researchers also focus on other aspects of community question answering, such as question retrieval [19,20], question classification [21], question routing [22], and so on.

Besides CQA, expert finding problem has also been studied in other contexts. Farahat et al. [23] proposed a model that combined social and textual authority and defined authority rank based on the HITS algorithm for finding authoritative hyperlinks on the World Wide Web. Fisher [24] used different features to find users with large out-degree and small in-degree in Usenet newsgroups, assuming that those that reply to many, but are rarely replied to are those who have largely provided the answers to the questions of the community. There are also several efforts that attempted to find authoritative bloggers on blogging sites. Java et al. [25] analyzed the spread of influence on the Blogosphere in order to select an influential set of bloggers which maximize the spread of information on the blogosphere. Java [26] proposed methods to find "blog feeds that matter" using their folder names and subscriber counts. Pal and Counts [27] proposed a number of new user metrics and employed a clustering approach approach that was computationally tractable to run the scenarios that demand near real-time response. In addition, some studies find experts in the TREC enterprise collection by analyzing the text contents of user information [28–31].

In this article, we focus on the widely studied graph-based link analysis techniques for expert finding in CQA. Compared to the previous work, our proposed method is more effective because it finds the experts by taking into account both the link structure and the topical similarity, expertise, and reputation of users.

3. Topic-sensitive expert finding

In this Section, we first propose a topic-sensitive method to find the top N users for each topic as the candidate experts by considering the link structure and the topical similarity among users. Then we develop a general probabilistic model to rank these candidate experts by considering their expertise and reputation.

3.1. Topic distillation

Topic distillation aims to automatically identify the topics that users (askers and answerers) are interested in based on the user profiles.[4] Because our data set is large, it is only feasible to use fully

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