User authority ranking models for community question answering[1]

Yanghui Rao[a], Haoran Xie[b,*], Xuebo Liu[a], Qing Li[c], Fu Lee Wang[d] and Tak-Lam Wong[b]

[a]*School of Data and Computer Science, Sun Yat-sen University, China*
[b]*Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong*
[c]*Department of Computer Science, City University of Hong Kong, Hong Kong*
[d]*Caritas Institute of Higher Education, Hong Kong*

Abstract. The proliferation of knowledge-sharing communities has generated large amounts of data. Prominent examples of how user-generated content can be harnessed include IBM's Watson question answering sytem and Apple's Siri, the question answering application in iPhones. Facing such massive data, user authority ranking is important to the development of question answering and other e-commerce services. In this study, we propose three probabilistic models to rank the user authority of each question. Compared to the existing approaches focused on the user relationship primarily, our method is more effective because we consider the link structure and topical similarities between users and questions simultaneously. We use a real-world dataset from *Zhihu*, a popular community question answering website in China to conduct experiments. Experimental results show that our model outperforms other baseline methods in ranking the user authority.

Keywords: User authority ranking, topic modeling, community question answering

1. Introduction

Question Answering (QA) systems involve different techniques for question analysis, answer extraction, answer presentation, and so forth. The successes of IBM's Watson and Apple's Siri have highlighted QA researches and commercialization opportunities [2]. For example, Community Question Answering (CQA) websites including *Quora* (www.quora.com), *Stackoverflow* (stackoverflow.com), *Yahoo! Answers* (answers.yahoo.com) and *Zhihu* (zhihu.com), are extremely popular in recent years. On CQA websites, users can raise their questions, read answers of these questions, and answer the questions which are posted by others. It is interesting that many CQA websites have strong social characteristics, which are different from the traditional ones such as *Baidu Knows* (zhidao.baidu.com). Take *Zhihu* as an example, the content read by user u in his/her home page depends on the people that u has followed (i.e., "friends" of u). Therefore, a user is able to see the questions from all his/her friends, their answers and their "liked" answers. By allowing the interaction between users, CQA websites have accumulated millions of questions and their answers over time [3].

Although the CQA website such as *Yahoo! Answers* has attracted many users to contribute knowledge, it brings several new challenges to provide high quality services [4]: (1) **Ineffective matching**: It is very challenging to recommend questions to answerers with both sufficient knowledge and strong motivation, which results in suboptimal answers and delay of responses. (2) **Unavailable**

---

*Corresponding author. Haoran Xie, Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong. Tel.: +852 29488545; Fax: +852 29487726; E-mail: hrxie2@gmail.com.

answer: As it is difficult to match questions with suitable answerers, many questions will be always without answer. (3) **Redundant questions**: If a user does not find a satisfied answer for one question, he/she may post redundant similar questions for keeping attract potential answerers. Thus, user authority ranking is important for CQA websites to recommend the best-matching user to answer these questions.

Traditional methods of ranking the user authority are based on the relationship between users primarily [5, 6], which are insufficient to get a personalized ranking list for each question. For instance, a movie star may have many followers and achieve a high ranking if we rank the user authority based on the relationship between users only. However, he/she may not be appropriate to answer the questions about computers. We here propose a new framework based on the link structure and topical similarities between users and questions, which can get a personalized user authority ranking for each question. In our framework, all answers posted by a user are considered as a document, the description and all answers to a question are also treated as a document; we then apply latent Dirichlet allocation (LDA) [7] to extract topics from both users and questions. After generating these topics, we can measure the topical similarities between questions and answers, and rank users by the relationship between users and the topical similarity. The main contributions of this work are summarized as follows. First, unlike the user authority ranking using the topic-sensitive transition matrix [1], we propose two methods of ranking the user's authority by a topic-independent matrix. Second, we further compare the above models in terms of both performance and efficiency.

The rest of the paper is organized as follows. In Section 2, we introduce the related work on user authority ranking, similar question recommendation, topic-sensitive expert recommendation and question answering system. In Section 3, the proposed framework is described in detail. In Section 4, we illustrate the employed dataset, evaluation metrics and experimental results. We conclude our paper and make a future plan in Section 5.

2. Related work

In this section, we firstly focus on studies on user authority ranking by relationships, similar question recommendation, and topic-sensitive expert recommendation, which will shed light on topic-sensitive user authority ranking for CQA websites. To obtain a holistic view of our research question in QA systems domain, we also give a general review of the existing techniques and models for QA systems.

2.1. User authority ranking by relationships

The existing ranking algorithms computed the user authority on CQA websites by the relationship between users primarily. Bouguessa et al. [8] proposed a method to find the expert based on the best answer that the user has posted. Jurczyk and Agichtein [9] applied HITS algorithm to rank the user authority by the followed-following relationship. Zhang et al. [10] 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 field he/she is not skilled in. Thus, some other algorithms based on topics were proposed. Guo et al. [11] proposed an algorithm to explore the similarity between users and askers by the tags of users. Liu et al. [12] proposed a probabilistic language model to predict the best answerer of questions.

2.2. Similar question recommendation

For similar question recommendation, Jeon et al. [13] proposed a statistical approach that explored the semantic features to measure the question similarity. Li et al. [14] designed a strategy via machine translation rather than the simple cosine similarity approaches. Hao et al. [15] developed a pattern-based algorithm by seed patterns and a semi-supervised approach. Wu et al. [16, 17] proposed an approach that exploited both the user interest and feedback, in which the historical data was utilized to derive the user interest. These studies showed the benefits of using topic model over the traditional methods. Our method is similar to some of these works in that we explore the semantics of questions using topic models. However, our method also considers the link structure of users, which was omitted in most similar question recommendation methods.

2.3. Topic-sensitive expert recommendation

Recently, some algorithms of expert recommendation have been proposed to consider both the

relationship between users and the similarity of topics. TwitterRank [18] is a typical algorithm developed to estimate the influence of Twitter users based on the followed-following relationship and topical similarities. Zhou et al. [19] proposed a TSPR algorithm for expert recommendation on CQA websites. The algorithm first employed LDA to extract the topics of users, and then recommended experts by the number of answers a user posted and the similarity between the user and the asker [20]. Zhao et al. [21] designed a method to model experts and topics jointly by incorporating each user's contribution dynamically. Chen et al. [22] developed a rating system of user reputation based on user comments. Pal et al. [23] proposed an algorithm based on Gaussian mixture models to identify topical authorities in microblogs. Liu et al. [24] proposed a new topic model for expert recommendation on CQA websites.

Different from these algorithms, our method considers both the link structure of users and the topical similarity between users and questions. To the best of our knowledge, it is the first time to consider the topical similarity between users and questions for user authority ranking on CQA websites.

2.4. Question answering system

QA systems have been extensively studied in recent years. In 2007, IBM Research undertook a challenge to build a computer system named Watson that could compete at the human champion level in real time on the American TV quiz show, Jeopardy [25]. Watson has been performed at human expert levels in terms of precision, confidence, and speed at the Jeopardy quiz show, the results of which strongly suggested that DeepQA is an effective and extensible architecture in the field of QA [25]. To alleviate the problem that the typical triplet does not represent a faithful representation of the semantic structure of the natural language question, Unger et al. [26] presented a novel approach that relies on a parse of the question to produce a SPARQL template that directly mirrors the internal structure of the question.

More recently, with the rapid growth of open, distributed and structured semantic data on the Internet, how to address the question answering task over RDF data has been a popular research topic [27]. Marx et al. [28] presented a modular and extensible open-source question answering framework to tackle the problems such as imposing minimal hurdles to their users, retrieving the desired data, and developing and evaluating question answering systems. Furthermore, Zou

et al. [29] proposed a systematic framework to answer natural language questions over RDF repository from a graph data-driven perspective, which adopted a semantic query graph to model the query intention in the natural language question in a structural way. The significant contributions of this research were (i) the RDF question answering has been reduced to the general subgraph matching problem; and (ii) the ambiguity of natural language questions has been resolved at the time when matches of query are found [29].

Another popular research direction is to generalize the conventional text-based question answering to multimedia objects (named as visual/image question answering). Specifically, given an image and a natural language question about the image, the task is to provide an accurate natural language answer [30]. Ren et al. [31] proposed to use neural networks and visual semantic embeddings, without intermediate stages such as object detection and image segmentation, to predict answers to simple questions about images. Furthermore, Malinowski et al. [32] attempted to address this problem by combining the Convolutional Neural Networks (CNN) for visual recognition and Recurrent Neural Networks (RNN) for sequence modeling.

3. Proposed models

The proposed Topical User Authority Ranking (TUAR) models are to rank the user authority for each question, by exploiting 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 [5]. In addition, as each question may involves various topics and users are familiar with different topics, we take the similarities among users and topics into consideration.

3.1. Topic extraction of users

We apply LDA [7] to perform the topic extraction of users from the combined text of their answers. 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 the Gibbs sampling algorithm [33]. The detailed process of user topic extraction in our TUAR models is as follows:

- Firstly, we consider all answers which are posted by each user as a text, thus 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 questions, 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. The definition of UZ is given below.

Definition 1. UZ: the matrix of $U \times Z$. U is the number of users. Z is the number of topics. UZ_{ij} denotes 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 extant studies on social networks, the followed-following relationship between users [18] and the number of answers [19, 24] can be employed to generate the user authority transition matrix via the iterative approach [34]. Different from the traditional social networks, the approval mechanism is introduced into the CQA website such as *Zhihu*. Thus, we can generate the user authority transition matrix based on the following principles. First, the more "likes" a user receives, the higher authority he/she will achieve. Second, 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. Third, the "likes" given by a user who

seldom makes approval achieves a larger weighting value than those who often deliver "likes". Therefore, we consider the user's authority ranking as a *Markov Chain* [34], and the transition matrix is shown as follows:

$$T_{i,j} = \frac{V_{j \to i}}{\sum_k V_{j \to k}} \times sim_{i,j}, \quad (1)$$

where T represents the user authority transition matrix, in which $T_{i,j}$ represents the influence of user i to user j. $V_{j \to i}$ represents the number of "likes" that user j gives to user i, and $\sum_k V_{j \to k}$ is the total number of "likes" that user j gives to all users. $sim_{i,j}$ represents the similarity between user i and j, which can be estimated as follows:

$$sim_{i,j} = 0.5 \times \frac{UZ'_i \cdot UZ'_j}{||UZ'_i|| \times ||UZ'_j||} + 0.5, \quad (2)$$

where UZ' is the row-normalized form of UZ, i.e., the L_1-norm of each row is 1. UZ_i and UZ_j represent the degree of interest of user i and user j in topics, respectively. The above equation is the *cosine similarity* between UZ_i and UZ_j after normalization. Due to that each element in the transition matrix of *Markov Chain* stands for the transition probability ranging from 0 to 1 while the value of cosine similarity varies from –1 to 1, we normalize the similarity from 0 to 1 by using two adjusting factors 0.5.

Another way to measure the similarity between user i and user j is to employ the normalized Euclidean distance between UZ_i and UZ_j, as follows:

$$sim_{i,j} = \frac{1}{1 + \sqrt{\sum_{k=1}^{Z} \left(UZ'_{ik} - UZ'_{jk} \right)^2}}, \quad (3)$$

where Z is the number of topics, UZ'_{ik} reperesents user i's interest in topic k, which is the k-th element of UZ_i. $\sqrt{\sum_{k=1}^{Z} \left(UZ'_{ik} - UZ'_{jk} \right)^2}$ is the Euclidean distance ED between UZ_i and UZ_j, and we use $\frac{1}{1+ED}$ to normalize the value.

We also use a topic-sensitive transition matrix to rank the user's authority [1], as follows:

$$T_Z(i, j) = \frac{V_{j \to i}}{\sum_k V_{j \to k}} \times sim_z(i, j), \quad (4)$$

where 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. $sim_z(i, j)$ represents the similarity between user i and j in topic z.

The similarity function in the topic-sensitive transition matrix is:

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

where 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. The above equation is the *relative entropy* between UZ_i and UZ_j after normalization. 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 $\ln \left(\frac{UZ'_{jz}}{UZ'_{iz}} \right)$ tend to approximate 0 while the value of *sim* tends to approximate 1. Otherwise, the value of *sim* will be small. 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, TUAR models take the approval relationship between users and the topical similarity into account to compute the authority ranking of users in topic z:

$$UR_z = \lambda T \times UR_z + (1 - \lambda) \times UZ''_z, \quad (5)$$

where 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 is the user authority transition matrix as 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 $QZ = \varphi_q$. The definition of QZ is as follows:

Definition 2. 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), the user authority ranking of each question can be estimated by the following *Bayes*'s rule:

$$QR = QZ \times UR, \quad (6)$$

where 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. The detail of our TUAR models is described as follows:

Algorithm 1. The proposed TUAR models

Input: The user's information (answers, "likes" received and the content of all answers he/she posted), and the question's information (its description, the content of all its answers).
Output: The user authority ranking of every question QR.
Parameters: topic number Z, damping factor parameter λ, iteration times C

1. Use LDA to train the documents of all users and questions and get UZ
2. **for** *each user i* **do**
3.   **for** *each user j* **do**
4.   Estimate $T_{i,j}$ according to Equations 1 and 2/3
5.   Estimate $T_z(i, j)$ according to Equation 4
6.   **end for**
7. **end for**
8. **for** *each topic z* **do**
9.   **for** $i = 1, \ldots, C$ **do**
10.   Update UR_z according to Equation 5
11.   **end for**
12. **end for**
13. Use the trained LDA to infer the documents of questions and get QZ
14. Estimate QR according to Equation 6

---

4. Experiments

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

4.1. Dataset

A real-world dataset from *Zhihu* was employed for the experiment. *Zhihu* is one of the most popular

question answering communities in China. Different from *StackOverflow* and *Yahoo! Answers*, we can get all users who "like" an answer from *Zhihu*, thus the user authority transition matrix can be generated by users' "like" relationships. According to Equation 1, the user's authority ranking is modeled as a *Markov Chain* based on the numbers of "likes". For instance, the more "likes" user j gives to user i, the higher the influence of user i to user j.

We have collected 576 questions which were raised by 9,043 *Zhihu* users. The total amount of answers is 209,309, which include all answers of the 576 questions, and all answers which were posted by the 9,043 users. The latter was collected to extract the topics of users, as shown in Section 3.1. 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.

We used the *Jieba Chinese Text Segmentation* (http://github.com/fxsjy/jieba) to perform Chinese word segmentation, and filtered stop words to reduce the vocabulary size.

Table 1
Top five words of different topics discovered by LDA

Topic 1	英雄 (Hero)	Dota (Dota)	技能 (Skill)	比赛 (Contest)	游戏 (Game)
Topic 2	时间 (Time)	学习 (Study)	事情 (Matter)	很多 (Many)	工作 (Work)
Topic 3	情况 (Situation)	很多 (Many)	来说 (In terms of)	时间 (Time)	两个 (Two)
Topic 4	知乎 (Zhihu)	http (http)	用户 (User)	内容 (Content)	网站 (Website)
Topic 5	用户 (User)	产品 (Product)	互联网 (Internet)	微信 (WeChat)	需求 (Requirement)
Topic 6	人类 (Human)	科学 (Science)	研究 (Research)	数学 (Math)	理论 (Theory)
Topic 7	中国 (China)	政治 (Politics)	社会 (Society)	台湾 (Taiwan)	国家 (Country)
Topic 8	工作 (Work)	北京 (Beijing)	很多 (Many)	公司 (Company)	几个 (Few)
Topic 9	历史 (History)	中国 (China)	皇帝 (Emperor)	时代 (Era)	文化 (Culture)
Topic 10	经济 (Economy)	发展 (Develop)	中国 (China)	政府 (Government)	市场 (Market)
Topic 11	喜欢 (Like)	生活 (Life)	世界 (World)	人生 (Life)	爱情 (Love)
Topic 12	新闻 (News)	时间 (Time)	媒体 (Media)	视频 (Video)	节目 (Program)
Topic 13	设计 (Design)	设计师 (Designer)	照片 (Photo)	效果 (Effect)	图片 (Picture)
Topic 14	汽车 (Car)	司机 (Diver)	车辆 (Vehicle)	开车 (Car Driving)	驾驶 (Drive)
Topic 15	中国 (China)	美国 (USA)	日本 (Japan)	国家 (Country)	国内 (Domestic)
Topic 16	公司 (Company)	企业 (Enterprise)	投资 (Invest)	业务 (Business)	保险 (Insurance)
Topic 17	建筑 (Architecture)	设计 (Design)	房子 (House)	城市 (Urban)	保护 (Protect)
Topic 18	电影 (Movie)	作品 (Works)	动画 (Animation)	故事 (Story)	导演 (Director)
Topic 19	比赛 (Contest)	球员 (Player)	球队 (Team)	防守 (Defense)	足球 (Football)
Topic 20	用户 (User)	软件 (Software)	功能 (Function)	系统 (System)	手机 (Cellphone)

Model	*MRR*	*nDCG*	Complexity
TUAR1	0.8158	0.8712	O(ZC) + O(U^2)
TUAR2	0.8194	0.8714	O(ZC) + O(U^2)
TUAR3	0.8411	0.8789	O(ZC) + O(U^2Z)

4.2. Parameters

We have several parameters in our models, i.e., the Dirichlet hyper-parameters α and β, the number of topics Z, the damping factor parameter λ used in PageRank. In this paper, we set the Dirichlet priors $\alpha = 50/Z$, and $\beta = 0.05$ as in [35]. We run LDA with 1000 iterations of Gibbs sampling. After trying several different numbers of topics, we empirically set $Z = 50$. We choose these parameter settings because they give coherent and meaningful topics for our dataset. Table 1 shows top five words of 20 topics generated by LDA.

When computing the user authority ranking of each topic, the damping factor parameter λ is set to be 0.85. The reason of setting λ as 0.85 is that we optimize the parameter in the training set (80% of the entire dataset). That is, we further divide the training data into a tuning set (64%) and a validation set (16%), and obtain the optimized setting of λ through 5-fold cross-validation.

4.3. Evaluation metrics

To measure the performance of different methods, two evaluation metrics commonly used in information retrieval and ranking were chosen:

– *Mean Reciprocal Rank (MRR)*: This index is the multiplicative inverse of the rank of the first retrieved expert for each topic;
– *nDCG@K*: This index measures the quality of a recommendation system based on the graded relevance of the recommended entities. It varies from 0.0 to 1.0, with 1.0 representing the ideal ranking of all entities, as follows: $nDCG@K = \frac{1}{Q} \sum_{q \in Q} \frac{\sum_{j=1}^{K} \frac{1}{\log_2(j+1)} score(M_{q,j})}{IdealScore(K,q)}$, where Q is the set of questions. $M_{q,j}$ is the j-th expert generated by method M for question q. $score(M_{q,j}) = 2^{v(M_{q,j})} - 1$, and $v(M_{q,j})$ 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.

4.4. Comparison with baselines

As stated in Section 3.2, we have proposed two models of user authority ranking using the topic-independent transition matrix. For these two methods, the one of estimating the similarity between user i and user j by the normalized *cosine similarity* is denoted as **TUAR1**, and the one of using the normalized *Euclidean distance* to measure the similarity is denoted by **TUAR2**. The method of user authority ranking using the topic-sensitive transition matrix is represented as **TUAR3**. To compare the performance of different models comprehensively, we also implement the following baselines:

(i) In-degree by number of followers (**IDF**): 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;
(ii) In-degree by number of "likes" (**IDV**): 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;
(iii) PageRank by number of followers (**PRF**): the algorithm generates the user ranking by applying PageRank with the number of followers;
(iv) PageRank by number of "likes" (**PRV**): the algorithm generates the user ranking by applying PageRank with the number of "likes";
(v) Topic-Sensitive PageRank (**TSPR**): 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 [19].

The *MRR* and *nDCG@K* of different methods are presented in Figs. 1 and 2, respectively, from which we can observe that the proposed TUAR models outperformed other methods for both metrics. Compared to the baseline methods of IDF, IDV, PRF, PRV and TSPR, the performance of TUAR1 improved 7.8%, 9.4%, 3.4%, 2.4%, 28.0% in terms of *MRR* and 1.9%, 2.8%, 1.0%, 0.8%, 13.1% in terms of *nDCG*. With respect to the same baselines, the performance of TUAR2 improved 8.3%, 9.9%, 3.9%, 2.9%, 28.6% in terms of *MRR* and 2.0%, 2.8%, 1.0%, 0.8%, 13.1% in terms of *nDCG*. Compared to the
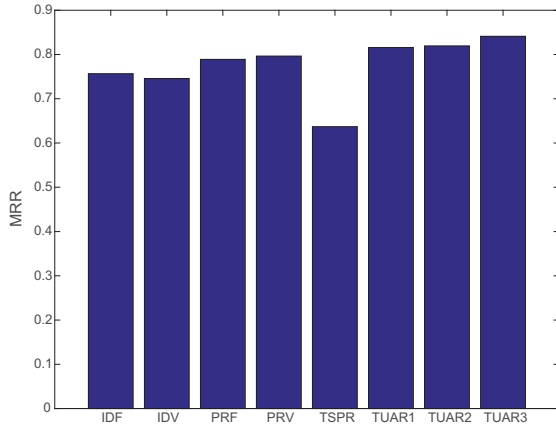
Fig. 1. *MRR* of different methods.
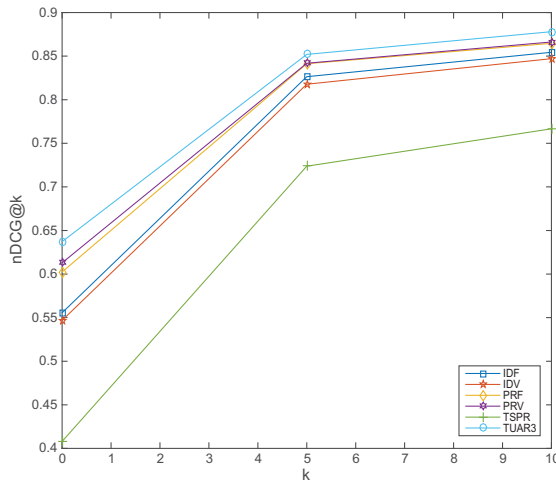
Fig. 2. *nDCG@K* of models with different *K* values.

above baseline methods, the performance of TUAR3 improved 11.1%, 12.8%, 6.6%, 5.6%, 32.0% in terms of *MRR* and 2.8%, 3.7%, 1.9%, 1.7%, 14.1% in terms of *nDCG*. The results indicated that it is effective to consider the topical similarity between questions and users when ranking the user authority. Due to that TUAR1, TUAR2 and TUAR3 performed nearly the same in terms of *nDCG*, we plot the values of TUAR3 in Fig. 2 for clarity.

4.5. Comparison with topic-sensitive models

The performance and efficiency of our TUAR models are shown in Table 2, in which *U* is the number of users, *Z* is the number of topics, and *C* is the iteration times. For example, the computational cost of TUAR1 is consisted of two parts. The first one is to

extract topics and the second one is to generate the user authority transition matrix.

Although the performance of TUAR1 and TUAR2 are worse than that of TUAR3 in terms of *MRR*, we can find that the topic-independent ones (TUAR1 and TUAR2) are much faster than TUAR3. The reason is that the topic-independent models only need to compute once for each pair of users when estimating the transition matrix, but the topic-sensitive model need to estimate the transition matrix for each topic. With respect to *nDCG*, the performance of these models are nearly the same with each other.

5. Conclusion

Many promising question answering system application areas, including education, health, and defense have been identified [2]. This paper proposed an effective framework to estimate the user authority ranking on CQA social networks, which can integrated with various multimedia primitive features [36–38] for social media organization [39, 40], sentiment analysis [41–44], emotion detection [45–47] and personalization [48–50] in the big data era. We evaluated the framework by the real-world dataset from *Zhihu*, and the experimental results verified the effectiveness of our models when compared to other existing methods.

In the future, we plan to design a new probabilistic model and try other recent learning models [51–53] for short text topic extraction when there are few answers to a question. The sparsity of content in short documents brings new challenges to topic modeling and user authority ranking. On one hand, the frequency of words, which is quite important to model lengthy text, plays limited discriminative role in short documents. On the other hand, inferring topics and ranking the user authority from large-scale short documents becomes a critical task for many areas. It follows that user authority ranking in short text deserves further research.

Acknowledgments

References

[1] X. Liu, S. Ye, X. Li, Y. Luo and Y. Rao, ZhihuRank: A topic-sensitive expert finding algorithm in community question answering websites, in *Proceedings of the 14th International Conference on Web-based Learning (ICWL)*, 2015, pp. 165–173.

[2] H. Chen, R.H.L. Chiang and V.C. Storey, Business intelligence and analytics: From big data to big impact, *MIS Quarterly* **36**(4) (2012), 1–24.

[3] Y. Liu, S. Li, Y. Cao, C.-Y. Lin, D. Han and Y. Yu, Understanding and summarizing answers in community-based question answering services, in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling)*, 2008, pp. 497–504.

[4] E. Agichtein, C. Castillo, D. Donato, A. Gionis and G. Mishne. Finding high-quality content in social media, in *Proceedings of the 1st ACM International Conference on Web Search and Data Mining (WSDM)*, 2008, pp. 183–194.

[5] S. Brin and L. Page, The anatomy of a large-scale hypertextual web search engine, *Computer Network and ISDN Systems* **30**(1-7) (1998), 107–117.

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

[7] D.M. Blei, A.Y. Ng and M.I. Jordan, Latent dirichlet allocation, *Journal of Machine Learning Research* **3** (2003), 993–1022.

[8] M. Bouguessa, B. Dumoulin and S. Wang, Identifying authoritative actors in question-answering forums: The case of yahoo! answers, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2008, pp. 866–874.

[9] P. Jurczyk and E. Agichtein, Discovering authorities in question answer communities by using link analysis, in *Proceedings of the ACM 16th Conference on Information and Knowledge Management (CIKM)*, 2007, pp. 919–922.

[10] J. Zhang, M.S. Ackerman and L. Adamic, Expertise networks in online communities: Structure and algorithms, in *Proceedings of the 16th International World Wide Web Conference (WWW)*, 2007, pp. 221–230.

[11] J. Guo, S. Xu, S. Bao and Y. Yu, Tapping on the potential of QA community by recommending answer providers, in *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, 2008, pp. 921–930.

[12] X. Liu, W.B. Croft and M. Koll, Finding experts in community-based question answering services, in *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management (CIKM)*, 2005, pp. 315–316.

[13] J. Jeon, W.B. Croft and J.H. Lee, Finding similar questions in large question and answer archives, in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 2005, pp. 84–90.

[14] S. Li and S. Manandhar, Improving question recommendation by exploiting information need, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2011, pp. 1425–1434.

[15] T. Hao and E. Agichtein, Finding similar questions in collaborative question answering archives: Toward bootstrapping-based equivalent pattern learning, *Information Retrieval* **15**(3-4) (2012), 332–353.

[16] H. Wu, Y. Wang and X. Cheng, Incremental probabilistic latent semantic analysis for automatic question recommendation, in *Proceedings of the 2nd ACM Conference on Recommender Systems (RecSys)*, 2008, pp. 99–106.

[17] M. Qu, G. Qiu, X. He, C. Zhang, H. Wu, J. Bu and C. Chen, Probabilistic question recommendation for question answering communities, in *Proceedings of the 18th International World Wide Web Conference (WWW)*, 2009, pp. 1229–1230.

[18] J. Weng, E.-P. Lim, J. Jiang and Q. He, Twitterrank: Finding topic-sensitive influential twitters, in *Proceedings of the 3rd International Conference on Web Search and Data Mining (WSDM)*, 2010, pp. 261–270.

[19] G. Zhou, S. Lai, K. Liu and J. Zhao, Topic-sensitive probabilistic model for expert finding in question answer communities, in *Proceedings of the 21st ACM Conference on Information and Knowledge Management (CIKM)*, 2012, pp. 1662–1666.

[20] T.H. Haveliwala, Topic-sensitive pagerank, in *Proceedings of the 11th International World Wide Web Conference (WWW)*, 2002, pp. 517–526.

[21] T. Zhao, N. Bian, C. Li and M. Li, Topic-level expert modeling in community question answering, in *SIAM International Conference on Data Mining (SDM)*, 2013, pp. 776–784.

[22] B.-C. Chen, J. Guo, B. Tseng and J. Yang, User reputation in a comment rating environment, in *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2011, pp. 159–167.

[23] A. Pal and S. Counts, Identifying topical authorities in microblogs, in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*, 2011, pp. 45–54.

[24] 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 *Proceedings of the 22nd International Conference on Information and Knowledge Management (CIKM)*, 2013, pp. 99–108.

[25] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur and N. Schlaefer, Building Watson: An overview of the DeepQA project, *AI magazine* **31**(3) (2010), 59–79.

[26] C. Unger, L. Bühmann, J. Lehmann, N. Ngomo, A.C. Gerber and P. Cimiano, Template-based question answering over RDF data, in *Proceedings of the 21st International World Wide Web Conference (WWW)*, 2012, pp. 639–648.

[27] V. Lopez, C. Unger, P. Cimiano and E. Motta, Evaluating question answering over linked data, *Web Semantics: Science, Services and Agents on the World Wide Web* **21** (2013), 3–13.

[28] E. Marx, R. Usbeck, A.C.N. Ngomo, K. Höffner, J. Lehmann and S. Auer, Towards an open question answering

architecture, in *Proceedings of the 10th International Conference on Semantic Systems (SEM)*, 2014, pp. 57–60.

[29] L. Zou, R. Huang, H. Wang, J.X. Yu, W. He and D. Zhao, Natural language question answering over RDF: A graph data driven approach, in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 2014, pp. 313–324.

[30] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick and D. Parikh, VQA: Visual question answering, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2425–2433.

[31] M. Ren, R. Kiros and R. Zemel, Exploring models and data for image question answering, in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 2935–2943.

[32] M. Malinowski, M. Rohrbach and M. Fritz, Ask your neurons: A neural-based approach to answering questions about images, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1–9.

[33] 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) (1984), 721–741.

[34] J.R. Norris, Markov chains. Cambridge University Press, 1998.

[35] T. Griffiths and M. Steyvers, Finding scientific topics, *Proceedings of the National Academy of Sciences of the United States of America* **101** (2004), 5228–5235.

[36] Z. Pan, Y. Zhang and S. Kwong, Efficient motion and disparity estimation optimization for low complexity multiview video coding, *IEEE Transactions on Broadcasting* **61**(2) (2015), 166–176.

[37] Y. Zheng, B. Jeon, D. Xu, Q.M. Wu and H. Zhang, Image segmentation by generalized hierarchical fuzzy C-means algorithm, *Journal of Intelligent and Fuzzy Systems* **28**(2) (2015), 961–973.

[38] J. Li, X. Li, B. Yang and X. Sun, Segmentation-based image copy-move forgery detection scheme, *IEEE Transactions on Information Forensics and Security* **10**(3) (2015), 507–518.

[39] H. Xie, Q. Li, X. Mao, X. Li, Y. Cai and Q. Zheng, Mining latent user community for tag-based and content-based search in social media, *Computer Journal* **57**(9) (2014), 1415–1430.

[40] H. Xie, Q. Li and Y. Cai, Community-Aware resource profiling for personalized search in folksonomy, *Journal*

of *Computer Science and Technology* **27**(3) (2012), 599–610.

[41] X. Li, H. Xie, L. Chen, J. Wang and X. Deng, News impact on stock price return via sentiment analysis, *Knowledge-Based Systems* **69** (2014), 14–23.

[42] X. Li, H. Xie, Y. Song, S. Zhu, Q. Li and F.L. Wang, Does summarization help stock prediction? a news impact analysis, *IEEE Intelligent Systems* **30**(3) (2015), 26–34.

[43] Y. Rao, J. Lei, W. Liu, Q. Li and M. Chen, Building emotional dictionary for sentiment analysis of online news, *World Wide Web* **17**(4) (2014), 723–742.

[44] J. Lei, Y. Rao, Q. Li, X. Quan and W. Liu, Towards building a social emotion detection system for online news, *Future Generation Computer Systems* **37** (2014), 438–448.

[45] Y. Rao, Contextual sentiment topic model for adaptive social emotion classification, *IEEE Intelligent Systems* **31**(1) (2016), 41–47.

[46] Y. Rao, Q. Li, W. Liu, Q. Wu and X. Quan, Affective topic model for social emotion detection, *Neural Networks* **58** (2014), 29–37.

[47] Y. Rao, Q. Li, X. Mao and W. Liu, Sentiment topic models for social emotion mining, *Information Sciences* **266** (2014), 90–100.

[48] H. Xie, D. Zou, R. Lau, F.L. Wang and T.L. Wong, Generating incidental word learning tasks via topic-based and load-based profiles, *IEEE MultiMedia* **23**(1) (2016), 60–70.

[49] H. Xie, X. Li, T. Wang, R.Y.K. Lau, T.L. Wong, L. Chen, F.L. Wang and Q. Li, Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy, *Information Processing and Management* **52**(1) (2016), 61–72.

[50] H. Xie, Q. Li, X. Mao, X. Li, Y. Cai and Y. Rao, Communty-aware user profile enrichment in folksonomy, *Neural Networks* **58** (2014), 111–121.

[51] X. Wen, L. Shao, Y. Xue and W. Fang, A rapid learning algorithm for vehicle classification, *Information Sciences* **295**(1) (2015), 395–406.

[52] B. Gu, V.S. Sheng, Z. Wang, D. Ho, S. Osman and S. Li, Incremental learning for ν-support vector regression, *Neural Networks* **67** (2015), 140–150.

[53] B. Gu and V.S. Sheng, A robust regularization path algorithm for ν-support vector classification, *IEEE Transactions on Neural Networks and Learning Systems* (2016). DOI: 10.1109/TNNLS.2016.2527796