Send Orders for Reprints to reprints@benthamscience.ae

The Open Cybernetics & Systemics Journal, 2014, *8,* 645-650

Open Access

Question Recommendation Mechanism under Q&A Community based on LDA Model

Caiyin Wang[1*], Lin Cui[1,2], Baosheng Yang[1] and Xiaoyin Wu[1]

[1]*Intelligent Information Processing Laboratory, Suzhou University, Suzhou, Anhui, 234000, China;* [2]*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu, 210016, China*

Abstract: Aiming at the questions not answered timely under Q&A community, a kind of questions recommendation method based on LDA (Latent Dirichlet Allocation) topic model is proposed, which fully utilizes personalized information of users under Q&A community. The interests distributions of users are expressed through using LDA model and according to the interests distributions of users, questions recommendation lists are calculated out at last. The proposed method can recommend the unsolved problems to users who are interested in these questions, which makes these questions be solved out as soon as possible, and promotes information dissemination and knowledge sharing under Q&A community. Experimental results show that the proposed questions recommendation method based on LDA not only discoveries the unsolved questions quickly, but also recommends the most suitable answers to users compared with PLSA, KL-divergence and Cosine similarity.

Keywords: Latent dirichlet allocation, Q & A community, question recommendation, questions recommendation lists.

1. INTRODUCTION

Nowadays Q&A communities as one of Web 2.0 typical applications have become one of today's most popular social network applications, which provide a platform searching for information and knowledge sharing for Internet users. For example, the popular Q&A communities Yahoo! Answers (https://answers.yahoo.com/), Baidu Zhidao (http://zhidao.baidu.com/) and Sina iask (http://iask.sina.com.cn/) publish tens of thousands of problems every day. Compared to obtaining information through the traditional search engines, users can question problems and answer problems in a rapid and accurate way through Q&A community, rather than spend much time on finding information in a large number of relevant documents returned by the traditional search engines. However, Q&A community provides people with easy access to information services, but there still exist a variety of problems, for example, users have to spend a long time on waiting for answers and sometimes the quality of the obtained answer is very poor. In addition, some users provide a lot of irrelevant answers or even rubbish answers in order to obtain scores provided by the community, which greatly reduces the efficiency that users obtain the required information.

In order to improve the performance of Q&A communities, this paper presents a question recommendation mechanism based on LDA, which aims to help answer users quick access to information and recommend the unsolved interesting problems to the users, so that problems can be answered as soon as possible, thus enhancing knowledge sharing behavior under Q&A communities.

2. RELATED WORK

Question answering system is an advanced form of information retrieval system, which can answer the proposed questions proposed by users through using accurate, concise nature questions raised by users in natural language [1]. Question answering community is also known as "interactive quiz sharing platform" and is different from Q&A system, questions under Q&A community not only are raised by users, but also are answered by users. Q&A community is designed to exploit the power of users from the Internet, gathering intelligence of the public to help answer questions, thereby building up a question for those who need the information to other users or search problem has been resolved network application platform.

With the explosive growth of data traffic under Q&A community, there emerge a lot of theories and algorithms about Q&A community. Agichtein *et al.* used classification framework to integrate all kinds of questions and answers in a text message community to study the question and answer community text quality [2]. Jeon *et al.* adopted translation model to learn semantic similarity between words in order to find a similar problem [3]. Jurczyk *et al.* studied user link structure under the question and answer community, and predicted the degree of authority using HITS algorithm [4]. Bian *et al.* proposed a semi-supervised mutually reinforcing framework computing content quality and user reputation under Q&A community [5]. Adamic *et al.* did a comprehensive analysis on knowledge sharing behavior in Yahoo! Answers, tested the best answer community questions and answers, verified the best answer credibility selected by the

1874-110X/14

askers [6]. In addition, the problem recommendation under Q&A community has always been a very representative content, Bunescu *et al.* departure from repeat discrimination of questions, based on user input inquiries constitute to recommend with repeat relations [7]. Sun utilized limited user voting information and effectively avoided the impact of noise information, improving the accuracy of recommended problem [8]. Li and Manandhar emphasized on analyzing the real intent and demand implicit in user input, questions were recommended based on user needs reflecting by query [9].

3. INTRODUCTION TO LDA MODEL

LDA (Latent Dirichlet Allocation) was proposed by Blei in 2003, which developed from LSA (Latent Semantic Analysis) and PLSA (Probabilistic Latent Semantic Analysis) [10]. LDA is a hidden variable topic model and is trained through unsupervised learning method, regardless of the number of training samples, which is more suitable for handling large-scale text corpus. LDA model is a three-layer Bayesian probability model that contains words, topics and document respectively. As a production model, LDA model has been successfully applied to the field of text classification, information retrieval, and many other text-related fields. LDA topic model is a probabilistic graphical model, which is shown in Fig. (**1**).

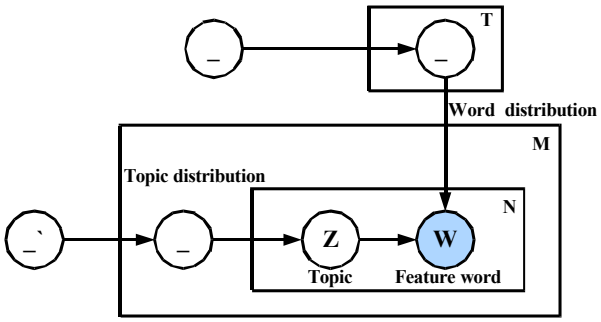

Fig. (1). Representation of LDA model.

Fig. (**1**) shows LDA topic model consists of a collection of documents layer with parameters (α, β), where α reflects the relative strength between the implicit topics in the document collection, β denotes all implicit probability distribution containing topic itself, the process of generating the document are as follows [11].

(a) For each topic *t*, a word multinomial distribution(*t*) on the subject is given by Dirichlet (β) distribution.

(b) For each document *d*, a topic multinomial distribution θ_{d} of the document is obtained by Dirichlet (α) distribution.

(c) For each word w_i in the document, a topic t is extracted from a topic polynomial distribution θ_{d} and a word w_i is also extracted from the word polynomial distributed (t) on the topic (t).

LDA topic model shows that the probability of characteristic words *w* in the questions is calculated as shown in the formula (1):

$$p_{LDA}(w \mid d, \hat{\theta}, \hat{\phi}) = \sum_{z=1}^{T} p(w \mid z, \hat{\phi}) p(z \mid \hat{\theta}, d) \tag{1}$$

where, *z* is the subject corresponding to feature words *w*, *T* is the number of topics and $\hat{\theta}$, $\hat{\phi}$ are priori estimate of parameters θ and ϕ respectively. *V* represents the total number of feature words in the corpus. Aiming at a feature word $w_i \in V = \{w_1, \cdots, w_v\}$ in the question *d*, given a topic $z_i = t$, then, the posterior probability $\hat{\theta}$, $\hat{\phi}$ are calculated as shown in the formula (2) and formula (3) as follows:

$$\hat{\phi}_{w_i}^{z_i} = \frac{n_{w_i,t}^{V,T} + \beta}{\sum_{i=1}^{V} n_{w_i,t}^{V,T} + \beta} \tag{2}$$

$$\hat{\theta}^{d} = \frac{n_{d,t}^{D,T} + \alpha}{\sum_{j=1}^{T} (n_{d,j}^{D,T} + \alpha)} \tag{3}$$

Aiming at the questions under the Q&A community, suppose $\Theta = \{\theta_1, \theta_2, \cdots \theta_k\}$ is the topic model, *k* is the total number of topic model, each topic model is a vocabulary polynomial distribution, then each question *q* is considered to be a sample produced by mixing topic model, as the formula (4) shows:

$$p_q(w) = \sum_{j=1}^{k} \left[\pi_{q,j} p(w \mid \theta_j) \right] \tag{4}$$

where, w denotes terms in problem *q*, $\pi_{q,j}$ is the probability produced by the j-*th* topic models $\sum_{j=1}^{k} \pi_{q,j} = 1$.

4. INTRODUCTION TO QUESTIONS RECOMMENDATION MECHANISMS

The proposed questions recommendation method based on LDA model can recommend the unsolved problems to users who are interested in these unsolved problems, so that the questions can be answered as soon as possible. The proposed question recommendation method is defined as follows:

Given problem sets $Q = \{q_1, q_2, \cdots q_n\}$ and user sets $U = \{u_1, u_2, \cdots, u_m\}$, for each user $u \in U$, the problem $q_u \in Q$ is to recommend to the user and satisfies the formular (5):

$$q_u = \arg\max_{q \in Q} Score_{q,u} \tag{5}$$

where, $Score_{q,u}$ denotes the extent that user *u* is interested in *q*. and the proposed recommendation algorithm based on LDA is as follows:

Firstly, <user, words> co-occurrence information in the problems collection is given out, then LDA model parame-

ters *Pr(w|z), Pr(z|u)* and *Pr(u)* are estimated by using the expectation maximization method, which makes logarithmic likelihood of the whole problems set reach local maximum and obtains the user interest model. Second, a set of problems to be solved *Q* and the user sets *U* are given out, according to the parameters of LDA model, the degree of user interesting in the problems Score is calculated and lastly the problem with the biggest Score value is recommended to the user *U*.

The core issue of the problem recommendation mechanism based on LDA to be solved is as follows:

4.1. Modeling User Interests

When modeling user *u*, using latent variables $z \in \{z_1, z_2, \cdots z_k\}$ in LDA model represents potential topics *z* that user u answers question set. Implicit process that user u answers questions q is firstly to select a topic *z*, and then select questions q based on topics. From the perspective of probability, the joint probability of the user *u*, question *q* and topics *z* is expressed as:

$$\Pr(u,q) = \sum_z \Pr(q \mid z) \Pr(z \mid u) \Pr(u) \tag{6}$$

Among them, $u \in \{u_1, u_2, \cdots u_m\}$ denotes users in the problem collection, $q \in \{q_1, q_2, \cdots q_m\}$ represents the questions in the problem set. However, in the real Q&A community, the number of each user answering questions is usually small, which makes observed value of most <user, problems> concurrence information is zero. To solve the sparsity problem, this study intends to use <user, words> co-occurrence information to model topics, among which, co-occurrence data are input words when the user answers some question. The joint probability of the user and the words is expressed as:

$$\Pr(u,w) = \sum_z \Pr(w \mid z) \Pr(z \mid u) \Pr(u) \tag{7}$$

where, $w \in \{w_1, w_2, \cdots w_m\}$ denotes the input words when user *u* answers question in question set.

LDA model generation process is as follows: From *m* users of the collection, according to the probability $\Pr(u)$, select a user *u*; from the selected user *u*, according to the probability $\Pr(z \mid u)$, select a topic z; then from the selected topic z, according to polynomial distributed $\Pr(w \mid z)$, choose words *w*. As can be seen, LDA model allows each user to have multiple interesting topics and based on these topics, commonly choose words.

4.2. Calculating Similarity

The key issue of question recommendation is calculating the extent $Score_{q,u}$ that user *u* interests in question *q*. Regarding $Score_{q,u}$ as posterior probability $\Pr(q \mid u)$, which represents the possibility that user *u* chooses question *q*, the latter will be seen as the posterior probability, which represents the posterior probability of the likelihood of user choice:

$$\Pr(q \mid u) = \frac{\Pr(u,q)}{\Pr(u)} \tag{8}$$

For the same user *u*, the joint probability $\Pr(u,q)$ may be calculated instead of $\Pr(q \mid u)$, probability values $\Pr(u,q)$ can be calculated out through words probabilities product and the normalized calculation of question length:

$$Score_{q,u} \propto \Pr(u,q) = \left(\prod_t \Pr(u, w_i) \right)^{\frac{1}{|q|}} \tag{9}$$

Among them, w_i is the words in problem q; $|q|$ is the total number of words in problem *q*. So, aiming at user *u*, a question list based on $Score_{q,u}$ sorting is obtained, then you can select one of pre-n questions to recommend to the user.

5. EXPERIMENTAL RESULTS ANALYSIS

5.1. Obtaining Experimental Data Set

In order to measure the performance questions retrieval method proposed in this paper, questions and answering information were collected from one of China's biggest Q&A community - Sina iask by using Web spider, under which, questions set included all the crawled questions, and each question included a number of answers and user information answering to the problems. We removed users whose answering times were less than 3 to generate user set. User and word co-occurrence information $c(u,w)$ was used to denote users set. "Computer "," Health ","Sports" and "Travel" four categories of questions total of 346,245 pieces were lastly collected to build a collection of questions Set_Sina. In order to construct test set, from the questions collection Set_Sina, 500 questions without replacement were randomly selected to build test set Set_Test. The constructed experimental data set is shown in Table **1**.

5.2. Baselines and Performance Indexes

In order to verify the performance of the proposed question recommendation methods, three baselines Cosine Similarity, KL-divergence and PLSA(probability latent semantic analysis method) were selected as comparative methods.

In the aspect of answer recognition, adopting measuring standards proposed by text retrieval conference (TREC) enterprise to evaluate the effect of algorithms [12]. Three evaluation indexes Precision@1(P@1),Mean Reciprocal Rank(MRR) and Precision@3 (P@3) were adopted and their calculation methods are as follows:

Mean Reciprocal Rank (abbreviated as MRR) is the mean of the reciprocals that Q&A experts found that the results the first correct find expert ranking in all of the categories of the user interactive question answering system. Calculating methods of MRR is as shown in the following formula:

Table 1. Statistical information of experimental data set.

Question Sets		Test Sets	
Category Directory	**Number of the Included Questions**	**Category Directory**	**Number of the Included Questions**
COMPUTER	87,931	COMPUTER_TEST	100
HEALTH	96,432	HEALTH_TEST	100
SPORT	90,139	SPORT_TEST	100
TRAVEL	71,743	TRAVEL_TEST	100
Total	**346,245**	**Total**	**400**

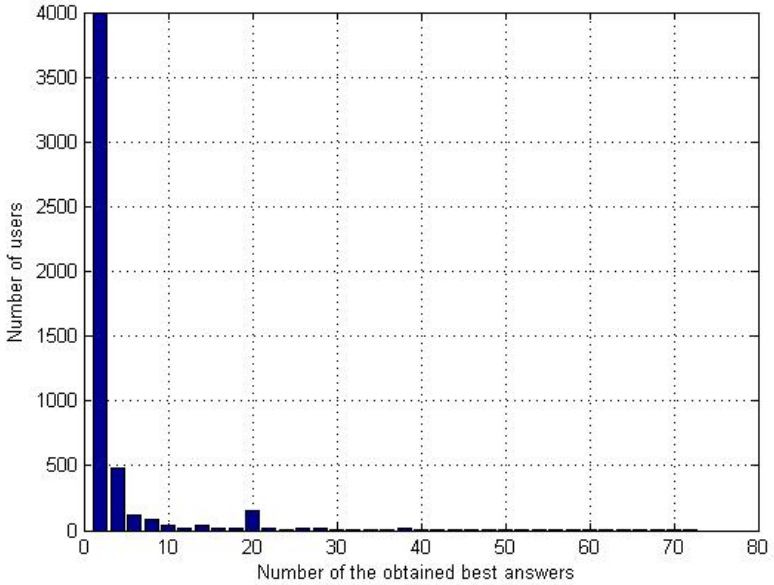

Fig. (2). Statistical charts of the best answer that users obtained under the category of "computer".

$$MRR = \frac{1}{|C|}\sum_{c_k \in C}\frac{1}{Rank(U_{c_k}{}^{R})} \tag{10}$$

where $Rank(U_{c_k}{}^{R})$ is the ranking of Q&A experts users that firstly was found the best answer. Precision@N or P@N denotes precision of the first N data, which shows the right ration on the first n found Q&A experts. Its calculation method is:

$$P@N = \frac{1}{|C|}\sum_{c_k \in C}\frac{|E_{C_k}{}^{R}|}{N} \tag{11}$$

where $E_{C_k}{}^{R}$ denotes the set of the related Q&A experts belong to the category c_k.

5.3. Experimental Results and Comparison Analysis

Firstly taking data from the category "Computer" as an example, Statistics the number of the best answer users get as shown in Fig. (**2**), which shows most users only obtain 1-5 best answers. It can be concluded that the number of experts users in every category only is 0.6% -0. 7% of the total number of users. Therefore, when determining the proportion of optional Q&A expert users and the proportion of expert user in the user sorting results, we choose the parameter value of 1%. That is the top 1% of users are considered to be an expert user, the other are a non-expert users and unrelated users.

Adopting evaluating indexes P@1, MRR and P@3 respectively, the four comparative experimental results are as follows in Fig. (**3**), Table **2**, which were obtained by using the proposed LDA recommendation method, PLSA, KL-divergence and Cosine similarity respectively. It can be observed that the comparative results demonstrate the effectiveness and efficiency of the proposed LDA method in questions recommendation at P@1, MRR and P@3.

As can be seen from the experimental results in Fig. (**3**) and Table **2**, the proposed question recommendation method based on LDA are significantly better than the other three baseline methods, The main reason is that LDA model is able to learn and construct the semantic association between words in the questions and answers from a lot of the corpus of Q&A pairs, and through semantic links between words to achieve short text semantic space mapping. Although the training corpus came from different data sources, but these Q&A pairs from Q&A system are still able to provide suffi-

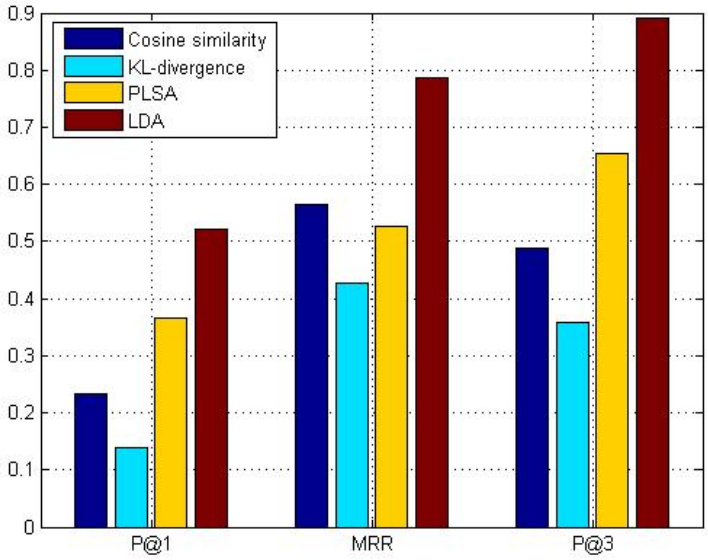

Fig. (3). Comparisons between recommendation algorithms.

Table 2. Comparison data results of the proposed LDA recommendation method and the baseline methods.

Question Recommendation Method	*P* @ 1	MRR	*P* @ 3
Cosine similarity	0.234	0.564	0.489
KL-divergence	0.139	0.426	0.357
PLSA	0.367	0.527	0.653
The proposed LDA	0.521	0.786	0.891

cient semantic knowledge for a model, so that it can be much better than the results of the baseline method.

CONCLUSION AND FUTURE WORK

This paper proposed questions recommendation method based on LDA, aiming at unsolved problems Q and user set U, calculated the degree of Score that users are interested in the problem, and the problem q with the greatest value Score is recommended to the user U. In order to evaluate the performance of the recommended questions based on LDA method, evaluating indexes MRR and $P@N$ were adopted to measure the performance of the proposed question recommendation method based on LDA, and compare with the cosine similarity algorithm, KL-divergence and PLSA. Experimental results show the proposed question recommendation method based on LDA is obviously superior to other three baseline methods. This proposed method does not rely on any structure or social network information relevant to the Q&A community, so it has better adaptability and generalization ability. However, our proposed method does not contain a time-varying relations when modeling topics, in future studies, time should be regarded as one of factors when modeling user interests.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the key project of Anhui Province Colleges and Universities Natural Science Foundation of China (No. KJ2014A250, No. KJ2012A263), the teaching research project at Suzhou University of China (No. szxyjyxm201413), the open project of Intelligent Information Processing Laboratory at Suzhou University of China (No. 2013YKF14, No. 2014YKF40, No. 2014YKF43).

REFERENCES

[1] A. Frank, H. Krieger, F. Xu, U. Hans, C. Berthold, B. Jörg, and U. Schäfer, "Question answering from structured knowledge sources," *Journal of Applied Logic*, vol. 5, no. 1, pp. 20-48, 2007.

[2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, G. Mishne "Finding high quality content in social media," *WSDM'08 Proceedings of the international Conference on Web Search and Web Data Mining ACM*, New York, pp. 183-194, 2008.

[3] J. Jeon, W. B. Croft, J. H. Lee, S. Park "A framework to predict the quality of answers with no textual features," *SIGIR'06 Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval ACM*, Washington, 2006: pp. 228-235.

[4] P. Jurczyk, and E. Agichtein, "Discovering authorities in question answer communities by using link analysis," *Proceedings of the 16th ACM conference on Conference on Information and Knowledge Management ACM*, Lisbon, Portugal, pp. 919-922, 2007.

[5] J. Bian, Y. Liu, E. Agichtein, H. Zha "Finding the right facts in the crowd: Factoid question answering over social media," *International World Wide Web Conference Committee. Proceedings of the 17th International Conference on World Wide Web ACM*, Beijing, pp. 467-476, 2008.

[6] L. Adamic, J. Zhang, E. Bakshy, and M. Ackerman, "Knowledge sharing and yahoo answers:everyone knows something," *Proceedings of the 17th International Conference on World Wide Web ACM*, New York, pp. 665-674, 2008.

[7] R. Bunescu, and Y. Huang, "A utility-driven approach to question ranking in social QA," *Proceedings of the 23rd International Conference on Computational Linguistics*, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 125-133, 2010.

[8] K. Sun, Y. Cao, X. Song, X. Wang, C.-Y. Lin "Learning to recommend questions based on user ratings," *Proceedings of the 18th ACM Conference on Information and Knowledge Management ACM*, New York, NY, USA, pp. 751-758, 2009.

[9] S. Li, and S. Manandhar, "Improving question recommendation by exploiting information need," *Proceedings of the 49th Annual Meeting of the Association for Comutational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 1425-1434, 2011.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[11] D. Ma, L. Rao, and T. Wang, "An empirical study of SLDA for information retrieval," *Proceedings of the 7th Asia Conference on Information Retrieval Technology,* 2011, pp. 84-92.

[12] G. Dror, Y. Koren, Y. Maarek and I. Szpektor, *I Want to Answer, Who Has a Question? Yahoo! Answers Recommender System*, In: *SIGKDD'11*, SanDiego, California, USA, pp. 1109-1117.

Received: September 22, 2014 Revised: November 30, 2014 Accepted: December 02, 2014