Question Recommendation Based on User Model in CQA

Junfeng Wang, Lei Su(✉), Jun Chen, and Di Jiang

School of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650093, China
w327918069@163.com, s28341@hotmail.com

Abstract. At present, people no longer meet the way of communication between users and the Internet. And more and more people choose the interaction between users and users to get information. The community question answering system is one of the new information sharing model. In the community question answering system, users are not only the questioner but also the answer and the question is the link between the users. With the increasing number of users and the increasing number of questions and answers, it makes many questions which just were raised disappear in the category pages of the home page. Leading to the efficiency of the questions be answered greatly reduce. Aim at the recommended user's interest, ability and time. In this paper we construct a dynamic user interest model and user expertise model. Experimental results show that the recommendation mechanism improves the efficiency of the recommendation to a certain extent.

Keywords: Community question answering system · Question recommendation · User' dynamic interest · User' expertise

1 Introduction

With the rapidly development of the internet, community question answering system (CQA) [1] has become an important way of people to get information and to share knowledge. Baidu knows is one of the largest community question and answer system, which has accumulated tens of millions of question and answer right now. And these questions and answers are provided by the user who use the community question and answer system. To help users find the questions they are interested in and answer them, more and more scholars are actively involved in and come up with some good models and methods. At present, the content of the research is divided into two categories.

The first category is to establish statistical language model and the theme model. In terms of statistical language model, Liu et al. [2] has built a language model through the user's information file. Put the content of the question into this model to calculate the extent of the user's interest in the question, so as to complete the recommendation of question. Zhang et al. [3] has combined with language model to form a mixed model to realize the recommendation of question on the basis of probability latent semantic. The result shows that the addition of semantic information of potential hybrid model is superior to the traditional language model. In establishing the subject model, Wu [4]

J. Wan et al. (Eds.): CloudComp 2016, SPNCE 2016, LNICST 197, pp. 91–101, 2018.
https://doi.org/10.1007/978-3-319-69605-8_9

has put forward an incremental question recommendation mechanism on the basis of probabilistic latent semantic analysis. This incremental embodied in the two aspects of new users and new problems. And the community Q&A system based on incremental to update the topic model. Guo et al. [5] has proposed UQA topic model. The model extracts semantic information from user's information to recommend the questions. Yu [6] has put forward a kind of personalized recommendation model based on social network. The model is based on the social relationship of trust among users. And get information to better reflect the personalized target users from other users with a high degree of relevance to the target user. To obtain more information in line with the needs of its personalized recommendation to reduce the blindness of the recommendation, improve the accuracy of the recommendation. Whether it is the use of statistical language models or the use of topic model, the recommended users to a large extent prefer to be interested in this question.

The second category is the research based on link analysis. Link analysis method is derived from the search engine and it through the link analysis algorithm to get the value of the authority of the web page. According to the idea of link analysis, researchers believe that in the community Q&A system the relationship between the user and the user is mainly a question and answer relationship. Through this relationship can also find expert users. According to this view, Agichtein and Jurczyk [7, 8] first established a Q&A community in the user's directed graph. This graph reflects the question and answers relation between the user and the user, and then uses the link analysis method to calculate the user's expertise value which is similar to the authority of the web page. If a user's expertise is higher, this user is an expert user. Finally, the problem is recommended to the experts who have been found. At present, the two link analysis methods used in the community question answering system are the PageRank algorithm proposed by Google's founder, Page [9] and the HITS algorithm proposed by Kleinberg [10]. However, using link analysis methods in community Q&A system to find expert users also has shortcoming. The reason is that the initial value of the different users is the same. However each user will give a different quality of answers, so each user can't be treated as equally.

According to the task characteristics of the recommendation system in community Q&A system, this paper studies the construction of user interest model and the construction of user expertise model in community Q&A system and puts forward a kind of recommendation mechanism which contains two methods based on the user model. This paper uses the data from the "Baidu know" to do recommend experiments. And the result shows that using this method to recommend new question has a significant improvement in the accuracy (P@N-Percent).

In the second section, this paper introduces the user's dynamic interest model based on the time weight and the user expertise model based on the PageRank algorithm. In the third section, this paper puts forward a synthesis algorithm for the question recommendation in line with recommendation question mechanism based on the user model constructed in section second. In the fourth section, this paper designs an experiment bases on the question recommendation mechanism proposed in this paper and verify the superiority of this algorithm. The fifth section summarizes the research work in this paper.

2 Question Recommendation Based on User Model

2.1 User Dynamic Interest Model

The degree of users' interest can be determined according to the number of the questions which the users answered. If the user answers more questions in a particular category, indicates that the user is more interested in the category. The problem is that the interest degree of the old user may be far greater than the new user's, whichcan't reflect the interest degree of the users in recently. The first method is to construct a user dynamic interest model based on time weight. When a new problem is raised, the degree of the user's interest in the question category can be calculated by calculating the time weight of each question the answered by the historic records of user's answer. The bigger sum of the time weights is more able to explain users are more interested in this question category in recently.

2.1.1 Time Weight

In the community Q&A system, according to study the information of the users' answered history this paper summarizes the following two characteristics: the user's interest in certain category is dynamic; the user's interest is divided into persistent and transient type.

The above problem shows that the questions the user answered in recently have a more important role for recommending the questions the user may be interested in the future. The early questions the user has answered impact on the user's interest is relatively small. This is because over time, the user's interest in changing, and the user's interest in a short period of time is relatively stable and unchanging. Therefore, the questions the users may be interested are similar with the questions they have answered in recently. In this paper, the concept of time weight is introduced when the user interest model is established, and so then the user's dynamic interest model is constructed.

Time weight refers to that each answered question has a time weight (*WT*) in the history record of the user answer questions (see the formula (2.1)).

$$WT(u,q) = (1-a) + a\frac{D_{uq}}{L_u} \tag{2.1}$$

$a \in (0,1)$ is the weight growth coefficient. If a is bigger, the time weight change is bigger. In this paper a is 0.7. *Duq* is the time interval of answering the question Q and of answering certain questions earliest. Lu is the time interval of raising questions and answering certain question.

In formula (2.1), Lu represents the time horizons of user to answer the questions that is. The time span also is the time interval of answering the question earliest and answering the question recently. The advantage of the original formula is that the time weight can be calculated off-line, saving the system time overhead. The disadvantage is that there is no consideration the liveness of user interest. When the users in all the time to answer the questions, it is easy to get the degree of users' interest of some type; however if the user in recently for too long a period of time not to answer the question, it is impossible to determine the degree of users' interest of some type. The Lu proposed in this paper taking into account the liveness of the user interest. If the user does not participate answer the question in a long time, the liveness of the user's interest would become low. Then the time weight is relatively small. And from the side it also can reflect the user's online situation in the recent period time, so as to determine whether the user can answer the question in time.

2.1.2 User Dynamic Interest

After determining the category of the question, find out the user groups that have answered the category of the question according to the category of the problem. Then combines the number of questions that each user answered with the time weight to calculate user's dynamic interest in this category online (see the formula (2.2)).

$$I_{Class(u_i)} = \log_2\left(\sum_{q \in Q(Class(u_i))} WT(u_i, q) + 1\right) \tag{2.2}$$

In this formula, Q(Class(ui)) is a set of certain categories questions in the history of user's answered record. User dynamic interest model not only considers the influence of early questions answered to the user's interest but also pay more attention that the impact to the user's interest of the user's answered which distance from closer to the stage of asking questions.

2.2 User Expertise Model

User's expertise length refers to the professional level of the user to answer questions in certain category.

There is a disadvantage that no considers the different quality of the answer given by the different user in using the traditional link analysis method to find the user's expertise of answering questions. In this paper, the research on the construction method of user expertise model basses on weight link. According to the answer adoption mechanism of community Q&A system calculate the value of the user's expertise. So that it can be more obvious to distinguish the user's expertise length.

2.2.1 Traditional PageRank Algorithm in the User Expertise Found

The discovery of web page authority is based on the link relationship between web pages. And there exists the Q&A relations that equivalents the relationship of link in and link out among the users in the community Q&A system. So the PageRank algorithm can be applied to the user expertise found by the relations. According to the Q&A relationship between the user A with user B and user C, a link is established between the user A to the user B and the user A to the user C. In order to be more intuitive to see in the picture omit the relationship between the user and the problem or the answer and get the Fig. 1. This graph reflects the Q&A relationship between users and this graph is used as the relations of links of the user in this paper. Can be seen from the figure, in addition to the user B to answer the A user's question, the user C also gives the answer. According to the idea of PageRank algorithm assuming that the user A a total of 1 point, then the user B and user C will get 0.5 points from the user A.

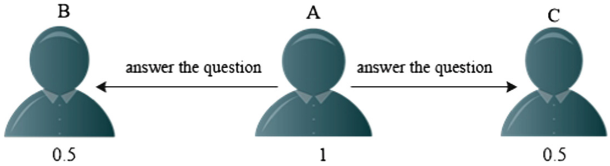

Fig. 1. User link diagram

In this graph, the length of a user's expertise length is determined by the number of the user's link in and the chain out of the user's specific length.

2.2.2 Calculate the Initial Expertise Length

In the papers related to the traditional PageRank algorithm published by the relevant researchers [11–13], analyzed the shortcomings of traditional PageRank algorithm in finding user expertise. In this paper, the user's link diagram is improved into the weighted user link diagram and the quality of the user's answers is considered indirectly. And aiming at the problem that the traditional PageRank algorithm does not consider the type of the questions in user expertise found, this paper improves the PageRank algorithm. So that gets the specific length of a user in a certain category (see the formula (2.3)).

$$E_{Class(u_i)} = (1-d) + d\sum_{j=1}^{n}\left(\frac{W_{ij}}{\sum W_j}E_{Class(u_i)}\right) \tag{2.3}$$

In this formula, n is the number of chain target users u_i in certain class; W_{ij} is the weight of the edge that the u_j to the user u_i; $\sum W_j$ is the sum of weights of all edges that the u_j to this type.

By the formula (2.3), it is known that computing user expertise is a constant iteration and recursive process. The recursive ends until all the chain out users have not answered the questions of the category. And the length of the recursive to the boundary of the user is $(1-d)$, $(1-d)$ can be understood as 1* $(1-d)$ that is the initial length of all users is 1 and the proportion of the length of the original length is $(1-D)$. However, analysis of the user's personal center in community Q&A system can find the expertise information from the user's personal information. It is not accurate to regard the initial length of the user as the same value in formula (2.3), so can use the expertise information to represent the user's initial expertise length. The personal home page that Baidu knows has the domain the user to be good at; the domain is the category which the mark user is good at. Because Baidu knows set three tier categories, the field of expertise can be one of the layers of the category. In this paper the user's expertise length of the corresponding category for the third tier categories, according to the field of expertise to set the user in a class of the initial length.

2.2.3 User Specific Length

The proposed user initial expertise length is integrated into the weighted PageRank algorithm to calculate the length of the user expertise to the class (see the formula (2.4)).

$$E_{Class(u_i)} = (1-d)EI_{Class(u_i)} + d\sum_{j=1}^{n}\left(\frac{W_{ij}}{\sum W_j}E_{Class(u_i)}\right) \tag{2.4}$$

When calculating the length of the user's expertise, according to the users answer historical records of the categories involved to calculate the length of the line under these categories of users.

3 Problem Recommendation Algorithm Based on User Model

According to the user's dynamic interest model and user expertise model proposed above, a question recommendation algorithm is proposed to match the user with the category of the proposed question (see the formula (3.1)).

$$QR(Class(q), u_i) = I_{Class(u_i)} \cdot E_{Class(u_i)} \tag{3.1}$$

$I_{Class(u_i)}$ is the level of user's interest in categories of questions, $E_{Class(u_i)}$ is the user's expertise length in the categories of questions. The product of the two represents the degree of matching between a user and the categories of questions and the matching values expressed in.

This paper references the category system of the Baidu Known and this system has three layer categories. First of all classifies the proposed questions so that the problem of the category specific to the third tier categories. Due to the characteristics of the third tier category is more detailed, it can be considered that the same questions for a user with the same degree of matching. In this paper, the recommend thought of questions and user matching is transformed into the thought of category of questions and user matching.

According to the above description can determine the course of the recommend questions. After determining the category of the problem, calculate the user's dynamic interestingness to this category. At the same time, this category is matched with the corresponding categories of user's expertise length that have been calculated offline so that get the user's expertise length for this category. According to the formula (3.1) calculate $QR(Class(u_i))$ and sort the $QR(Class(u_i))$. Eventually form a list of recommended users. The flow chart of the question recommendation is shown in Fig. 2:

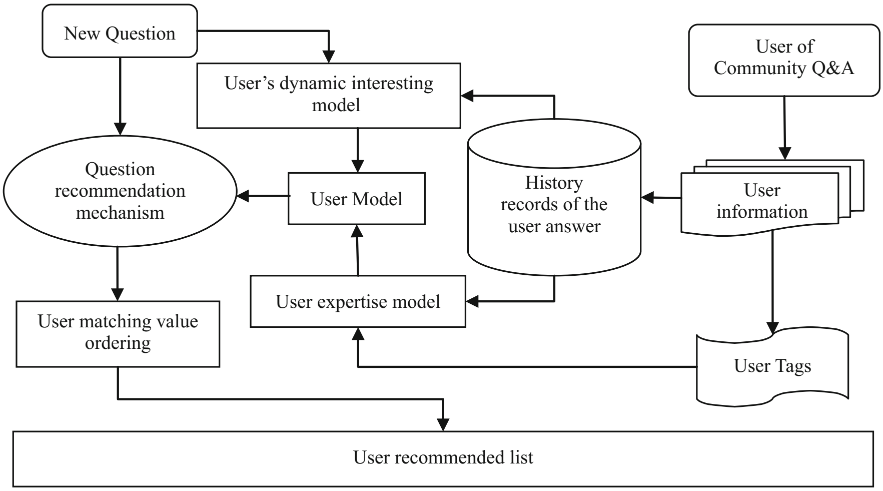
New Question
User's dynamic interesting model
User of Community Q&A
Question recommendation mechanism
User Model
History records of the user answer
User information
User matching value ordering
User expertise model
User Tags
User recommended list

Fig. 2. Flow chart of problem recommendation

4 Experiment and Analysis

4.1 Experimental Data

The study of recommending question is based on Baidu known so the experimental data comes from Baidu know. And the experimental data is divided into three parts. The first part is the category information, including all of the three tier categories that Baidu known. The second part is the information of the question, including the user's name, the title of the question, and the time of the question; the third part is the answer information, including the answer description, user name, user label, feedback type, answer time.

This experiment crawls 5 categories of questions as the experimental data, in order to evaluate the effect removes the questions which answer is not adopted. The statistics of specific crawling data are shown in Table 1.

Table 1. The experimental data statistics of Baidu Known

Question categories	Question quantity	Number of users	Answer quantity
Basketball	2680	3623	9978
Computer	2749	3865	11231
Exercise	2451	3015	8781
Tourism	2552	3279	8567
Mobile phone	2893	4032	12608

4.2 Evaluation Method

To check the recommendation effect of the problem recommendation algorithm, not only depends on whether the user is able to answer questions raised but also knows how high the probability of the user's answer will be adopted. According to this idea, this paper uses the accuracy value P@N-Percent to test the recommendation effect of the question recommendation algorithm. The formula for calculating the exact value of P@N-Percent is shown in the formula 4.1.

$$P@N - Percent = \frac{\sum_{i=1}^{n} isHit(u_i, q_i, n - percent)}{n} \tag{4.1}$$

$isHit(u_i,\ q_i,\ n - percent)$ indicates whether the user u_i that is adopted by the question q_i is matched to the user recommended list of the top percent N, the value of $isHit(u_i,\ q_i,\ n - percent)$ is 1 or 0; n is the number of test set questions.

4.3 Experimental Design and Results Analysis

In this paper, the questions of the 5 categories of the experimental data are sorted according to the order of the time of questioning. Take the top 80% of the question and answer data as a training corpus, the remaining 20% question and answer corpus as a test corpus. Adopt the question recommendation method of the Table 2 to test and compare.

Table 2. Question recommended methods

Methods	Description
Category based PageRank	It can find the expertise of users but not consider the user's recent interest and the quality of the user's answer
User dynamic interest model combines with PageRank (DIM-PageRank)	It can find the expertise of users at the same time to consider the user's recent interest, but still do not consider the quality of the user's answer
User dynamic interest model combines with user expertise model (DIM-EM)	It can find the expertise of users at the same time to consider the user's recent interest, and in the expertise model to add weight and initial expertise length

According to these three methods, the precision values of P@1-Percent, P@5-Percent and P@10-Percent 3 are used to test recommendation effect of the question recommendation algorithm. The results of the experiment are shown in Tables 3, 4 and 5.

According to the results of these 3 tables, it can be seen that there is a problem with black-bordered font data only in the P@1-Percent and P@5-Percent. In P@1-Percent the recommended effect of the DIM-PageRank is worse than the traditional PageRank recommended in the two categories of computers and tourism. And the recommended effect of traditional PageRank in the travel category is better than the recommended effect of DIM-EM. In the P@5-Percent the recommended effect of DIM-EM is worse

Table 3. The test statistics for each method in p@1-percent

Question category	Number of questions	PageRank	DIM-PageRank	DIM-EM
Basketball	536	0.312	0.323	0.344
Computer	550	0.290	**0.281**	0.307
Exercise	490	0.245	0.256	0.259
Tourism	510	0.307	**0.298**	**0.302**
Mobile phone	579	0.268	0.273	0.298

Table 4. The test statistics for each method in p@5-percent

Question category	Number of questions	PageRank	DIM-PageRank	DIM-EM
Basketball	536	0.357	0.391	0.459
Computer	550	0.312	0.374	0.442
Exercise	490	0.309	0.355	0.423
Tourism	510	0.337	0.409	0.488
Mobile phone	579	0.329	0.386	**0.381**

Table 5. The test statistics for each method in p@10-percent

Question category	Number of questions	PageRank	DIM-PageRank	DIM-EM
Basketball	536	0.489	0.554	0.612
Computer	550	0.442	0.481	0.597
Exercise	490	0.391	0.463	0.552
Tourism	510	0.404	0.529	0.600
Mobile phone	536	0.489	0.554	0.612

than the DIM-PageRank in the mobile phone category. And other cases, the recommended effect of DIM-EM is better than the recommended effect of DIM-PageRank; the recommended effect of DIM-PageRank is better than that of traditional PageRank.

In order to observe the whole situation, calculate the average P@N-Percent of PageRank, DIM-PageRank and DIM-EM in 5 categories. Statistical results are shown in Table 6 and the growth trend chart is shown in Fig. 3.

Table 6. The average P@n-Percent test statistics for each method

Methods	P@1-Percent	P@5-Percent	P@10-Percent
PageRank	0.284	0.329	0.423
DIM-PageRank	0.286	0.383	0.495
DIM-EM	0.302	0.439	0.571

According to the average P@N-Percent of the three methods in the 5 categories concluded that: the recommended effect of DIM-PageRank is better than the PageRank,

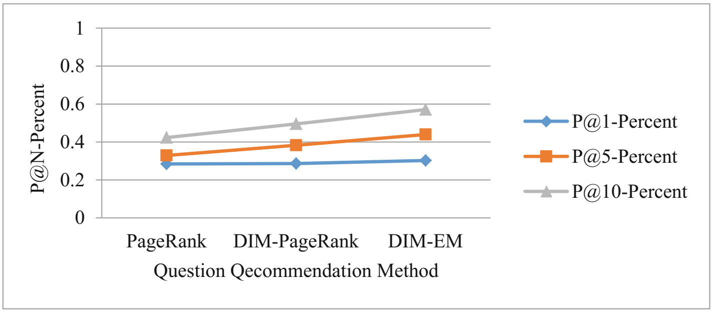

Fig. 3. The average P@N-Percent growth trend of each method

and the recommended effect of DIM-EM is better than the DIM-PageRank. Therefore, the question recommendation method based on the user model proposed in this paper has a better recommended effect.

5 Conclusions

Community Q&A system has become a new and important way for users to obtain t information and share knowledge. The work of raising question and answering is the core of community Q&A system, and the classification of questions and the problem can be answered in a timely become a key part of the work of question-answering. The question recommendation mechanism is the important link between the questioner and the answerer, which can greatly promote the development of the community Q&A system. In view of the problems in the community question and answer system, the study in this paper focused on the construction of user dynamic interest model and the construction of user expertise model. A question recommendation method based on user model is proposed in this paper, and the experimental result shows that the proposed method is effective.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No. 61365010).

References

1. Zhang, Z.F., Li, Q.D.: Review of community question answering system. Comput. Sci. **37** (11), 19–23 (2011)
2. Liu, X.Y., Bruce Croft, W., et al.: Finding experts in community-based question-answering services. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 315–316 (2005)
3. Zhang, J., Tang, J., Li, J.: Expert finding in a social network. In: Kotagiri, R., Krishna, P.R., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, vol. 4443, pp. 1066–1069. Springer, Heidelberg (2007). doi:10.1007/978-3-540-71703-4_106

4. Wu, H., Wang, Y., Cheng, X.: Incremental probabilistic latent semantic analysis for automatic question recommendation. In: Proceedings of the 2008 ACM Conference on Recommender Systems, pp. 99–106 (2008)
5. Guo, J., Xu, S., Bao, S., Yu, Y.: Tapping on the potential of Q&A community by recommending answer providers. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 921–930 (2008)
6. Yu, S.H.: Research on key technologies of personalized social network based on recommendation system. Nat. Def. Sci. Technol. Univ. 24–40 (2011)
7. Jurczyk, P., Agichtein, E.: Hits on question answer portals: exploration of link analysis for author ranking. In: Proceedings of 30th Annual International ACM SIGIR Conference, pp. 845–846 (2007)
8. Urczyk, P., Agichtein, E.: Discovering authorities in question answer communities by using link analysis. In: Proceedings of ACM 17th Conference on Information and Knowledge Management, pp. 919–922 (2007)
9. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Stanford Digital Library Working Paper SIDL-WP-1999-0120
10. Kleinberg, J.: Authoritative sources in a hyper linked environment (1998)
11. Duan, W.C., Hu, P.: An improved PageRank algorithm based on topic feature and time factor. Comput. Eng. Des. **31**(4), 866–868 (2010)
12. Yang, J.S., Ling, P.L.: Improvement of PageRank algorithm for search engine. Comput. Proj. **35**(22), 35–37 (2009)
13. Deng, D.J., Zhou, C.L.: Improved PageRank algorithm based on content correlation and time analysis. Comput. Digit. Eng. **39**(1), 25–27 (2011)