

Research and Implementation of Discovering and Recommending Experts in Medical Community-Based Q&A

Lan FANG*

Department of Electronic and Information Engineering
Tongji University
Shanghai, China
e-mail: flnam9241@gmail.com

Yu FANG

Department of Electronic and Information Engineering
Tongji University
Shanghai, China
e-mail: fangyu@tongji.edu.cn

Abstract—With the continuous development of the Internet, community-based Q&A (CQA) system has become an effective platform for people to access online information and help. People can submit their own question in CQA and get answers from others. There will be an enormous amount of questions all the time, and new questions which put forward by users are very easy to be buried. Even though there is no new question raising in a short time, the difficult questions will also be buried, because there are no enough users with professional approach to discover and answer the questions in time. To solve such problems, the technology of recommending experts emerged. We propose an optimized LDA recommendation model which combined with medical features in this thesis, by means of the recommendation technology, we can get around 3~4 potential topics of each experts interested in and the corresponding probabilities with the improved topics model. Thus we can obtain a relationship matrix. Then we can get a scores ranking according to the customized scoring approach of experts, and recommend. The experimental results show that, compared to the conventional recommendation algorithm, the results and accuracy of the recommendation algorithm proposed in this thesis has improved.

Keywords—health;experts-recommendation;relationship matrix

I. INTRODUCTION

CQA system is an interactive Q & A platform. With the rapid development of Web2.0 which is a user-centered product pattern, this pattern which pays more attentions to the interaction of users get increasingly grace from the majority of users. At present, the major CQA systems are ma Baidu knows, Yahoo Answer! Soso and so on in China [1]. In recent years, with the increasingly improved CQA system, it has been the important way for people to search and seek the answers, because of the wide range of users, huge information and unique ways of interacting. More and more users join the CQA system to interact and search for information through this platform. Users in CQA act both as contents providers and beneficiaries. Compared to traditional pure search engines, the advantage of CQA is having better performance at recognizing users' questions. It integrates the functions of database, the answers put forward by other users can be as results of the search, and be provided to other users who seek the similar questions. So then it can achieve sharing answers.

One of the disadvantages of CQA is that users have different quality. There will be huge amount of users doing nothing about, many questions won't get replied after being submitted, there are certain users just advertise in CQA. In order to provide the better recommendations service, we need to identify high-quality users. In CQA system, it usually defines the users who behavior actively and provide high-quality answers as high-quality users. Because diverse problems are divided into different into categories based on different knowledge areas, and experts will be places into different categories according to the answers they have provided as well. In general, experts will good at one or some knowledge domains instead of every domain. If we can identify expert users in user groups, and send the new submitted questions to them which makes the questions get professional replied in time, expands the knowledge database. Thus, discovering the experts can be research focus in CQA.

II. RELATED WORK

In recent years, the research of recommendation system gets more and more attentions, a lot of researchers have spent a lot of energy working on the recommendation system. Recommendation algorithm, which as the most important part of recommendation system, determines the performance of the recommendation system. Therefore, the recommendation algorithm gets widespread attentions; the research involves Cognition, Approximate Theory, Information Retrieval, Management Science and Market Modeling and many other aspects [2]. There are three main Q&A expert's recommendation algorithms.

A. The Method of Discovering Q&A Experts Based-on Language Model

In terms of establishing a language model, Liu et al. to solve the problems of recommending experts by using a model of various languages and multilingual user information files, and ultimately draw a conclusion that it's more beneficial to discover the Q&A experts by enriching users' information files [3]. Cao et al. propose a two-stage language model, which consists of co-occurrence model and the relevance model. These two models regard the statistical language model as theoretic basis. Co-occurrence model is used to judge the relevance of users and query term, and relevance model shows the relevance of text documents and

query term. Zhang et al. use PLSA (Probabilistic latent semantic analysis) as basis to establish a mixed model to discover Q&A experts [4]. The introduction of latent semantic information makes the method is superior to the traditional language model, Zhang et al validate their ideas in practice [5].

B. The Method of Discovering Q&A Experts Based-on Link Analysis

Researchers find that, it's a kind of Q&A relationship between each user except to find experts according to users' files. In the same time, it also increases the probability to find the Q&A experts. From this point of view, Agichtein and Jurczyk et al in order to find the experts in CQA, firstly, they establish the Q&A relationship between users by directed-graph, and then use sophisticated link analysis method which belongs to search domain to distinguish the experts [6][7]. The PageRank algorithm [8] which designed by Page.L who is the founder of Google and HITS algorithm [9] designed by Kleinberg.J are the most two widely used link analysis method in interactive Q&A system. The research based on link analysis follow these conventions: Q&A experts are the users who have high-authority, and authority analogy to the authority degree of webs in search engines.

C. The Method of Discovering Q&A Experts Based-on Topic-Based Model

In terms of establishing topic-based models, Wu and others design a kind of problem of increments recommendation strategy based on PLSA [10]. There are two aspects embodying Increments: new users and new problems. The system updates model base on increments so that the model is maintained at a relatively new status. Guo et al propose UQA topic-based model, it's able to extract semantic information from the user information to help the problem recommendation [11].

The method of discovering experts which bases on language model is prone to lack the underlying semantically analysis. For example, in the field of medical Q&A, there is professional vocabulary and terminology. We need preprocess before doing the text analysis. For the method of discovering experts which bases on link analysis, the link relationship is built by Q&A relationship, but it's easy to overlook some hidden links. And for new entrants, to establish contact with experts is less than ideal. For questions and the links which built for the new experts are not ideal. The weak point of the method of discovering experts which bases on based topic model of expert is considering few factor, maybe the questions which recommended by this method and the fields which expert is good have a strong correlation. But it ignores some, for example the efficiency of answering questions, the answers' quality and other factors.

We can not only use forefathers' studies to solve the problems of discovering experts in medical CQA. We should improve forefathers' algorithm. For example: we should add Lexical Analysis of medical field, consider more factors when recommending the experts. In order to research

a kind of more perfect technology of discovering and recommending experts in medical CQA, in this thesis, we make health as insertion points to research the technology of in discovering and recommending experts in CQA.

III. RECOMMENDATION TECHNOLOGY

A. Relation Matrix of Doctor and Potential Medical Topics

Firstly, we classify the medical documents which we collect. According to the research of forefathers show that: we can use perplexity model to measure classification effect about topics. In medical field, classify the topics into 110 can achieve the best results. Classification process as shown below:

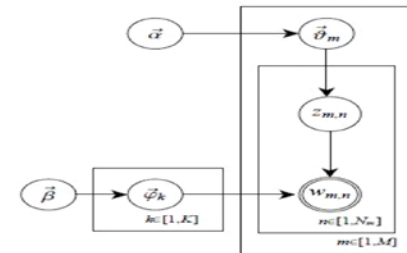


Figure 1. The process of LDA

TABLE I. DECLARATION OF PARAMETERS IN CLASSIFICATION PROCESS

Parameter	Declaration
α	probability distribution of topics' probability, parameter of Dirichlet
Φ_m	the probability of topic distribution of document m
$Z_{m,n}$	the topic make by the document m of the n-th vocabulary
$W_{m,n}$	the vocabulary made by topic
Φ	the probability of topic vocabulary, β make Φ meet the parameters of Dirichlet distribution

After finish classifying 110 medical topics by topic model, we can get a topic model which the value of k equal 110. Then we can record the historical Q&A record of each doctors into individual documents, and compare each document with the topic model we have gotten before to get each similarity degree, we can get 3 the most similar topics of each doctors through the results and the corresponding probability. Then we can get a distribution matrix of doctor and potential medical topics like Table II. One doctor may good at more than one medical topic. (e.g. user 14).

TABLE II. DISTRIBUTION MATRIX OF DOCTOR AND POTENTIAL MEDICAL TOPICS

potential medical topic distribution of Q	a	b	c	d
	User43	User348	User4	User24
	User75	User2	User775	User536
	User14	User432	User843	User43

	User348	User4800	User3289	User14

B. Expert Scoring Algorithm

The experts scoring algorithm proposed in this thesis is divided into two parts: Professional degree matching and expert's historical performance score. The specific formula is:

$$\text{Score} = \sum_{p \in P} (Q_p * U_p + S_{(t,d,r,c,o)}) \quad (1)$$

When a new question is put forward, we can use topic model to get several medical topics which are similar with the new question. Then we can get elementary experts set which is matching the new question through the matrix of medical topics and experts with their corresponding probability. According to the research of literature, we can assume that the quality of answers depends on the performance of doctor's historical answer records. On the basis of the available information in medical CQA, this thesis adopts five different aspects which contain answer-question time, the relevance between question and answer, the difficult level of question, whether the answer is adopted and support or objection count of the answer to model the performance of doctor's historical answer records. These factors all need historical data to measure, and influence the score of the performance of each doctor in certain medical topic.

TABLE III. DECLARATION OF EACH PARAMETER IN EXPERTS SCORING ALGORITHM

Parameter	Declaration
Qp	the distribution probability of topic p in the question
Up	the distribution probability of topic p of user's files-trained model
t	answer-question time
d	the relevance between question and answer
r	the difficult level of question
c	whether the answer is adopted
o	support or objection count of the answer

The specific declaration of each parameter in experts scoring algorithm are: Qp means the matching degree of topic p and new question. Up means the distribution probability of topic p of user's files-trained model. Score means the matching degree of doctor and the new question, which the result is the sum of the distribution of each medical topic and the probability of topic in each doctor's files-trained model. T means doctor's answer-question time. D means the relevance between question and answer, but there is no better method to measure the relevance up to now. So in this thesis, we just use the keywords of the Q&A, the similar degree of the topic to approximate the result. R means the difficult level of question. We also approximate it in this thesis. More answers the question gets much difficult the question is. C means whether the answer is adopted, the score will be added if the answer is adopted.

O means the support or objection count of the answer, we use the difference. In the experts scoring algorithm, we also introduce self-tuning parameter to adjust the proportion of the matching degree of profession and the performance of each expert's historical answer records.

IV. EXPERIMENTS AND ANALYSIS

A. Experimental Setting

We use 122511 CV of doctors from 120ask website in the experiments. The CV of each doctor contains the introduction and the historical answer records. Table IV shows the experimental data we static.

TABLE IV. THE STATISTIC OF THE EXPERIMENTAL DATA

Number of doctors	Number of questions	Number of answers	Status of question
122511	644615	4464792	closed
Average number of answers of the question	Average number of votes of answer	Average time of the question closed	Status of question
6.93	80.13	27.31(mins)	closed

The data which bases on LDA shows: For a data set, LDA divides the concept of health into a medical topic through semantics which has inter connection, it can describe the underlying structure of semantics of health data according to the declaration of topics. Each potential topic can be regarded as one-dimensional characteristics. Dimension of feature space can be gotten by perplexity [12]. Perplexity is a kind of model for calculating:

$$P_e = \exp \left\{ \frac{\sum_{i=1}^m \log p(d)}{\sum_{i=1}^m l_i} \right\} \quad (2)$$

In the formula, d means the number of the words. Much lower the perplexity is, much better the topic model is. We divide doctor's individual into two parts: CV and historical Q&A, 80% of the data are used to train the model and 20% are used to evaluate the performance. To achieve building the model and perplexity matrix through Stanford modeling toolset, when the number of the potential topics changes, perplexity can be valued as chart one. What we can know from chart one is: when the number of potential topics is 110, the value of the perplexity is the lowest. Thus, for a given question, it can be expressed as mixed by 110 semantic characteristics.

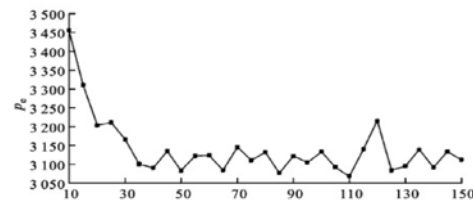


Figure 2. The changing curve of perplexity with the number medical topics

B. Comparison of Recommendation Performance

In the experiment, we can evaluate the recommendation technology by two aspects: 1. Use 20% of Q&A to evaluate the result of the recommendation. 2. Compare the algorithm in this thesis with others.

We also use two standards to evaluate the recommendation results, they're accuracy and MRR. Table V shows that: what the value of the accuracy is when we adjust the self-tuning parameter λ . MRR is a universal evaluating mechanism to evaluate the search algorithms, the formula is:

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^N \frac{1}{r(S_y^{+(i)})} \quad (3)$$

TABLE V. THE VALUE OF ACCURACY WITH CORRESPONDING λ

λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
accuracy	0.31	0.34	0.357	0.42	0.58	0.625	0.49	0.32

TABLE VI. THE VALUE OF MRR WITH CORRESPONDING λ

λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
MRR	0.32	0.44	0.522	0.67	0.734	0.66	0.57	0.38

In the experiment, we use two basic recommendation algorithm to compare with. They're collaborative filtering based on content and association rules algorithm. We can regard the value of MRR as the evaluation parameters. Chart 2 shows the result of comparison. It demonstrates that, to surmise several potential topics of each doctor is better than class doctors as one classification.

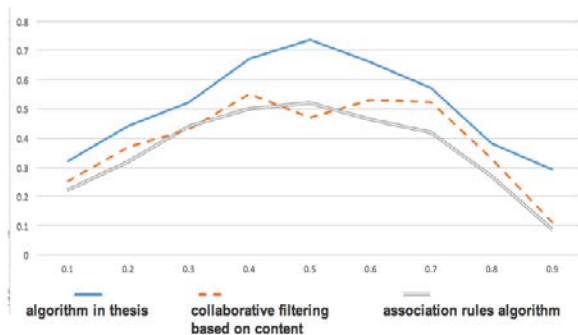


Figure 3. The value of different recommendation algorithms with corresponding λ

V. CONCLUSION

The expert's recommendation method in medical CQA in this thesis contains 2 main steps: To build a model of

doctors and corresponding potential major distribution and establish a relation matrix: in term of the relevance of Q&A, the answer time, the difficult degree of question and other factors to evaluate the performance of the doctor. At last, we recommend the experts according to the scores they get. The result of the experiment shows that, the recommendation technology proposed by this thesis has nice accuracy and can cover some shortages of the application of medical expert's recommendation.

Due to the quality of Chinese corpus sets which we used in the modeling and experiments should be improved; and the number of the classification is not small. It leads that: The degree of similarity of some topic is very high. In the future, we should improve the quality of the corpus set, discover the better classification method of semantics topics. They are the problems which need to continue to be solved.

REFERENCES

- [1] J. Wang, H. T. Yang, and H. F. Lin, "Question Similarity Calculation Based on Feature Fusion in Community Question Answering," *Journal of Jiangxi Normal University: Natural Science*, vol. 37, 2013, pp. 125-129.
- [2] R. Mooney, P. Bennett, and L. Roy, Book Recommending Using Text Categorization with Extracted Information[C].Proceeding of AAAI 1998/ICML 1998 Workshop on Learning for Text Categorization, pp. 49-54.
- [3] G. Somolo, A. Howe, Adaptive Lightweight Text Filtering[C].Proceeding of the 4th International Conference on Advances in Intelligent Data Analysis, 2011, pp. 319-329.
- [4] Y. Zhang, J. Callan, and T. Minka, Nobilty and Redundancy Detection in Adaptive Filtering[C].Proceeding of the 25th Annal International ACM SIGIR Conference on Research and Development in Information Retrieval, 2002, pp. 81-88.
- [5] T. Robertson, Threshold setting and performance optimization in adaptive filtering[J]. *Information Retrieval*, vol. 5, 2002, pp. 239-256.
- [6] Y. Zhang, J. Callan, Maximum likelihood estimation for filtering thresholds[C].//Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001, pp. 294-302.
- [7] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, 2005, pp. 734-749.
- [8] G. Linden, B. Smith, York J, Amazon. com recommendations: Item-to-item collaborative filtering[J]. *Internet Computing, IEEE*, vol. 7, 2003, pp. 76-80.
- [9] F. Ricci, L. Rokach, and B. Shapira, Introduction to recommender systems handbook[M]. Springer US, 2011.
- [10] L. Zhu, A. Z. Yue, "Routing Health-Oriented Questions to Appropriate Doctors," *Journal of xi'an jiaotong university*, vol. 48, 2014, pp. 57-62.
- [11] Iwasaki, Makoto, "Disturbance observer-based practical control of shaking tables with nonlinear specimen," *IFAC Proceedings*, vol. 38, 2005, pp. 251-256.
- [12] W. X. Che, T. Liu and B. Qin, " Solution Algorithm of String Similarity Based on Improved Levenshtein Distance[J]," vol. 14, 2004, pp. 15-19.