An Algorithm Research on Potential Problem Reactor in Medical Online Question Answering Community

Hu Zhongkai
Zhejiang University, China

Abstract:
Objective: Based on the online question-and-answer community of Internet medicine, a weighted user-personal model is proposed in view of the pain point of the users’ queries in the medical question-and-answer community which can not be effectively answered in time. Predicts whether the recommended user can respond timely and effectively to the problem to improve the accuracy of the proposed results.

Methods: Using the user's historical activity data in the community, a model of user's personal domain was proposed based on query-like language model to describe the degree of user's concern in different field.

Results: In the experiment based on Baidu medical question-answering system data, compared with the existing algorithms, the potential problem respondents in this paper have better recommendation effect.

keywords: online medical data; Potential problems; a reply person; discovery algorithm

1 Preface

With the development of information technology and the Internet, a lot of data is produced every day. People are exposed to massive amounts of information. However, while the Internet brings information to the whole society, it also brings information overload, too much information is far beyond people's ability to deal with it. At the same time, redundant information insubmerges information that people really need to drown in the information oceans, resulting in inefficiencies. The online question-and-answer community has developed, given the growing demand for information and the difficulty of using traditional search engines to locate information. As a representative traditional information retrieval system based on health search engine, medical online question-answering community has overcome some shortcomings of existing document retrieval system, and can answer users' questions in natural language with accurate and succinct natural language. Its biggest feature is the integration of community interaction with Internet users, allowing problems with natural language expression as query input, and the return result is a precise answer to the problem provided by community users, and users do not need to retrieve the information needed from the traditional search results list.

However, due to the professional requirements of medical knowledge and the lack of general Internet users in medical knowledge, in order to be able to solve ordinary Internet users in medical questions, The online question-and-answer community has become one of the most important applications in the online question-and-answer community. But because the number of users in these communities and the number of questions raised is too large, users' questions are difficult to get feedback in a short time. The doctor or patient who wants to answer questions can't find a solution to a large number of problems that match his or her expertise or experience. And since many of the answers provided by many users are limited to the user's personal experience in the problem domain, not all answers are correct in the answers provided by the user, giving the questioner a great deal of trouble identifying the correct answers.

In view of the above-mentioned problems, this paper
proposes a potential problem-answer algorithm based on the consideration of user domain and user's expertise. The method uses query-like language model to model the attention degree of users in order to find the respondents who are interested in the problem. By using the maximum entropy model to estimate the user's response quality, a user-specific model based on the quality assessment of the answer was established to distinguish the interlocutors from the medical question-and-answer community. This method takes into account many different factors such as attention, expertise and so on, which can be used to model the user and distinguish the user's personalization. The results show that the algorithm has good effect on finding suitable potential problems.

2 Literature review

2.1 Medical Online Question Answering Community

Medical online question-and-answer community, also known as medical interactive knowledge-and-answer sharing platform. The questions in the medical question-and-answer community are addressed by the user, as well as by the user. The medical question-and-answer community website provides users with the ability to input questions and retrieve problems, while providing users with questions of interest and offering their own answers. As a typical application of medical social network service in knowledge sharing, medical question-and-answer community fully embodies the vigorous development of grass-roots culture. At present, the common medical question-and-answer community has "WebMD", "Baidu asked a doctor", "search-seeking medicine network" and so on.

Early medical question answering system mainly consists of automatic question answering system. The goal of automatic question answering system is to make semantic analysis of knowledge set through natural language processing technology, and build knowledge representation and form knowledge base. Retrieve the possible knowledge points corresponding to the retrieval problem in the knowledge base. Organize these knowledge points, generate the representation forms that you can understand, and form the answer back to the user. However, this knowledge-based system has its limitations. First of all, because the medical knowledge system is very large, it is difficult to establish a complete medical knowledge base. Second, natural language is a very complex expression form. On the base of large knowledge base, the efficiency and accuracy of analysis and retrieval are a problem. The present technology has yielded satisfactory results only in the knowledge systems of specific areas. By defining the problem in a particular field, you can use the method of the domain expert to assist, define ontology and other metadata, and establish a structured knowledge representation. In the commonality medical question answering system, the effect of this method is often not satisfactory, and there is no effective solution at present. Since automatic question answering system mainly relies on natural language processing, knowledge representation and reasoning, machine learning and other techniques[1], at this stage, a satisfactory open-domain automatic question-answering system can not be generated because of the limitations of technology.

In recent years, with the development of the Internet and the development of Web 2.0 and social network services, a question-and-answer system, such as "WebMD" and "Baidu asked a doctor", is produced, which is the concept of "online medical question-and-answer community" in this paper. The online question-and-answer community uses a form of mutual assistance to answer questions posed by other users in the community. Different people have different backgrounds and expertise, and the online question-and-answer community provides a platform for mutual assistance to allow professional users to help other users answer questions and form the complementarity of knowledge. This approach bypasses both the complexities of natural language processing and the complexity of the knowledge base, shifting the focus to the matching of questions and answers, and the questioners can use less standard language descriptions, since the respondent is an expert in the field. Before natural language processing and knowledge-base technology break through existing bottlenecks, the pattern remains the mainstay of the question-and-answer system.

2.2 The Status Quo of Online Question Answering Community in Medicine
2.2.1 Automatic Categorization of Problems

In the medical question-and-answer community, because of the wide range of questions raised by the user, the automatic classification of questions becomes an important research direction if they can effectively divide these questions to improve the efficiency of answering questions in question-and-answer communities. Zhang et al.\[^2,3\]\(\text{to machine learning to classify the problem. The authors find that SVM is superior to the nearest neighbor model, decision tree and naive Bayes model. In this paper, a tree kernel function for sentence structure is proposed and nonlinear transformation is introduced to further improve the effect of SVM.}

Solorio et al.\[^4\]\(\text{proposes a sort of unrelated classification algorithm. Previous work was mostly based on specific languages and used language-related processing. The author uses the characteristic of sentence grammar and combines the Internet to study SVM, which can be easily applied to different languages. The algorithm has good results in English, Spanish and Italian languages.}

People\[^5\]\(\text{used semi-supervised learning to categorize the problem. In the feature selection, high-frequency keywords are used as characteristics and weights are adjusted based on semantic similarity. For the training process, a small number of tag data training models are first used, then adjusted parameters are adjusted based on unmarked data, and data that need further markup are given. Experiments show that this method has good performance in coarse-grained and fine-grained.}

2.2.2 Similarity Problem Retrieval

In the online question-and-answer community, an artificial answer raises another question: the repeatability of the problem. Many times, there are questions that the questioner wants to ask in the system, and there are good answers that can be found by searching. But because questioners are usually not familiar with the field, the current search engine technology does not understand semantics, usually requiring some skills or knowing some jargon to accurately describe existing problems. The effect of traditional retrieval technology is that the calculation method of similarity degree of document does not take into account the semantic information, and the question raised by the questioner is not satisfied with the similarity of word frequency vector. Therefore, the similarity detection algorithm is needed to solve the problem.

Jon et al.\[^6\]\(\text{presents a problem-search algorithm which considers semantic similarity. The author uses the language translation model to calculate the similarity of the problem while considering the answer information of the problem at the same time. This method can detect the problem that word frequency is not similar and semantically similar.}

Wang et al.\[^7\]\(\text{uses a syntax tree-based retrieval algorithm. This method is a heuristic method by analyzing the syntax tree of the document instead of using the data training method. Experiments show that the algorithm is 8.3% higher than the machine learning method based on vocabulary and tree kernel, and the semantic feature is more than 50%.}

2.3 Question Recommendation

The online question-and-answer community produces a lot of new questions every day, and it's hard to get answers in time. In other words, the magnitude of the problem has led to a similar overload of information. The classification of the problem helps to alleviate this problem, but this sort of treatment is extensive in the connotation of the problem, and it is not feasible to spend a lot of manpower on the precise and structured description of the problem. In addition, the knowledge composition of the responder is complex and can't be solved completely by simple subscription classification. Therefore, the introduction of recommendation system is very necessary, recommender system based on user's personal information, problem content and answer history, choose the most likely answer to this problem for each question, and do not need to do hard matching according to the explicit classification of the problem. According to the classification and similarity detection system, the question-and-answer recommendation system can greatly improve the efficiency of the problem solving.

Guo et al.\[^8,9,10]\(\text{uses probabilistic latent semantic analysis model to build the recommendation model. Training the potential vectors of the user through the user's answer history; For the new problem, give a list of the most}
matching users of the problem and push the issue to these users. One example of the analogy to the medical question-and-answer community is that Li et al.\cite{11} has studied the problem backers recommendation in the programmer community. First, analyze the source code in the problem, construct the "conceptual network" according to the semantic information such as inheritance relation and call relation in code. Secondly, through the forum's users to answer the activity, build a "subscriber network". When a user replies to a post, the post will be given by the publisher, which deals with the field of "expert score". Finally, through the integration of the problem of the "conceptual network" and the user's "subscriber network", according to the problem belongs to the conceptual domain and the user good area, the relevant issues posts are recommended to the corresponding experts. People\cite{12} build user models based on the response relationship between user's personal data and posts and the content of posts. These cases are also worthy of reference in the study of the medical question-and-answer community.

3 Potential Problem Reactor Discovery Algorithm Construction

In the medical question-and-answer community, users have to spend more and more time waiting for other users to answer questions due to the fact that there are so many new questions each day. Difficult to find answers to questions that match their expertise or experience to answer questions; And not all answers are correct in the answers provided by the user, and many of the answers are limited to the user's personal experience in the area of the problem. So in this section we implemented a potential problem-answer discovery algorithm in question-and-answer communities. When the new problem arises, it is recommended to the appropriate user to answer so that the problem can be resolved in time, through the problem recommendation to reduce the time for the user to wait for the question to be answered. Raise user satisfaction.

3.1 Definition of "potential respondent of the problem"

In order to express and display the discussion of this article more clearly, this paper gives a formal definition and explanation of the problems in this paper.

Definition 2.1: Potential Respondents of Questions

The potential answer to the question can be defined as: a given user set \( U = \{u_1, u_2, \ldots, u_n\} \) For the new problem \( q_r \), the system identifies a specific user \( u_{i1}, u_{i2}, \ldots, u_{ik} \). Turk answers the new question \( q_r \), and potential respondents are satisfied:

\[
U_t = \text{TOPN}\{\text{UR}(u_i, q_r)\}.
\]

\( \text{UR}(u_i, q_r) \) is the appropriate level for \( u_i \) to answer \( q_r \), and \( \text{TOPN}\{\text{UR}(u_i, q_r)\} \) indicates that the appropriate level of answer \( q_r \) is ranked in the former N user; The greater the \( \text{UR}(u_i, q_r) \), the greater the suitability of the user's response to the problem \( q_r \).

3.2 Research Framework and Process Design

For potential respondents, the following two aspects should be considered: (1) The potential respondents should have greater awareness of the subject's theme and therefore need to build a user-domain attention model to evaluate the user's interest in the problem; (2) In the field of attention, potential respondents should have the ability or expertise to answer questions, so they also need to build a user-expertise model to measure whether the user's expertise matches the problem domain.

When a user asks a new question, in order to find a user who is interested in the problem and is suitable for answering the question, we first set up two user models for all potential problem responders: user domain attention model and user expertise model, Evaluate the potential problem respondents' ability to focus and solve the problem in the area of the problem, respectively. Then, by synthesizing the scoring situation of the users in the two aspects, according to the comprehensive score to sort the users, select the higher-scoring users as the appropriate question reply, and push the issue to them.

3.3 Modeling user domain based on query-like language model

In the medical question-and-answer community, because of the wide range of medical knowledge involved, the average user rarely answers the questions of multiple medical directions, so individual users show special attention to certain areas. Moreover, users' focus on these particular areas remains stable for a long time, meaning that users are usually only interested in issues specific to certain topics.
At the same time, the vocabulary included in a topic has its own uniqueness, and the vocabulary contained in different subjects is different. Through the analysis of "Baidu knows", we find that the probability of "Gynecology", "Pregnancy", "miscarriage" and other words in the "medical health" category of "Hospital Correlation" is very high, while the problems in the class of "treat" include the words such as "menstruation", "Suggestions", "operation" and so on.

If a word that makes up a problem can be generated by a user already answering a question set, it can be inferred that the user may have a higher profile of the problem. Therefore, we can use query-like language model to measure the user's interest in the problem by calculating the probability of generating the problem from the user's answering question. A user's response to a problem usually indicates that the user has a degree of interest in the subject's topic, so in order to find the right answer to the new problem, the user domain attention model can be first established, and then the potential problem answer is found. According to the matching degree of attention and problem in the user domain, the appropriate responders are identified.

Excavate the user's domain attention, can start from the user's answer history. We can assume that the user's domain knowledge and interests are relatively stable, so you can get the user's domain attention from the user's response history. Specifically, determine if the user is fit to answer the question by calculating the similarity between the new problem and the user's response. If the problem is similar to the problem that the user has answered, give the user a higher priority to answer the question.

Figure 1 illustrates this method of query-like model users' interest in the problem, among which, for each user i, we first use the user's answered question set Q_r(u_i) to train its corresponding language model, θ_{Q_r(u_i)}, Then calculate the probability P(q_r | θ_{Q_r(u_i)}) of the problem q_r generated by the language model, ranking all users in accordance with that value, ranking by the previous user having a higher priority.

First we train the language model of each user, θ_{Q_r(u_i)}, and then calculate the problem q_r generation probability of the user language model, or query likelihood ratio, to calculate the user's attention in the field of the problem. We use n-gram as a language model. Previous research has shown that higher-order n-gram increases the complexity of the model, but does not yield significant improvements. We use unigram as a user's language model. The unary model does not take into account the effect of preposition on the target word, which is equivalent to the hypothesis of the polynomial distribution of the word distribution in the document. The term of the upcoming document is regarded as the result of a number of random trials, and this document is represented by this multinomial distribution parameter. The one-dimensional model P(q_r | θ_{Q_r(u_i)}) is a multinomial distribution:

\[
P(q_r | θ_{Q_r(u_i)}) = \prod_{w=1}^{m} P(w | θ_{Q_r(u_i)}) = \prod_{w=1}^{m} P(w | θ_{Q_r(u_i)})^c(w, q_r) q_r = w_1w_2...w_m, \quad c(w, q_r) represents the number of times a word w appears in the problem q_r.
\]

For the training of one-dimensional model, we know the probability that each word appears, i.e. \(P(w | θ_{Q_r(u_i)})\), according to maximum likelihood estimation, the parameter estimation of language model can be obtained:

\[
P_m(w | θ_{Q_r(u_i)}) = \frac{c(w, q_r)}{|Q_r(u_i)|}
\]

In which, \(c(w, Q_r(u_i))\) denotes the number of words w in the problem set \(Q_r(u_i)\), \(|Q_r(u_i)|\) sets the total number of words for the problem set. But the maximum likelihood estimate has the problem of fitting, that is, for a word that does not exist in a training center, the probability that uses maximum likelihood is \(P(w | θ_{Q_r(u_i)}) = 0\), which does not conform to general logic common sense. For this reason, we use the smoothing technique to eliminate the over-fitting problem caused by sparse data. In this paper, using Jelinek-
Mercer smoothing method\cite{14}, on the basis of maximum likelihood estimation, we add all users as a whole language model, and do linear interpolation, and get the language model:

\[ P_w(\theta|Q_{(u)}) = (1 - \lambda)P_m(\theta|Q_{(u)}) + \lambda P(w|\text{corpus}) \]

Among them, \( \lambda \) is the smoothing parameter, \( \lambda \in [0,1] \), corpus for all the user's problem sets. When entering \( \lambda=0 \), the model is equivalent to the maximum likelihood estimation of single user. When entering \( \lambda=1 \), the model becomes a single full-user language model, and there is no difference between users. \( P(w|\text{corpus}) \) is a language model for the entire dataset. In the end, the user's attention to the domain of the problem is as follows:

\[ P(Q_t|\theta|Q_{(u)}) = \prod_{w\in Q_t} P_w(\theta|Q_{(u)})^{c(w, Q_t)} \]

\[ = \prod_{w\in Q_t} [(1 - \lambda)\frac{c(w, Q_t)}{|Q_t|} + \lambda \frac{c(w, \text{corpus})}{|\text{corpus}|}]^{c(w, Q_t)} \]

3.4 Data Experiment and Analysis

The experimental data are automatically extracted from the "Baidu knows" system by the program. We capture the complete question-and-answer data (questions and answers) from 1 May 2017 to the complete question and answer of the May 1, 2018, from 1 May 2017 to the "medical health" of the "Baidu knows", and the total number of the answers to the total number of 502359 is the 228099. The number of questioners is the 159283 and the respondent number is the 180970. After statistical analysis, we found that the average number of questions asked users was 1.432, and the average number of respondents was 2.775, and the average answer was 2.202.

First of all, the user answers historical information to use Lucene: the system carries on the Chinese participle, then based on the relevant information to the user domain attention to carry on the modelling and the specialty analysis. In this paper, the following methods are tested and compared:

Base model (BM, Base Model), we use the probability potential semantic analysis to build user domain attention model based on probabilistic latent semantic analysis. A query-like language model (IoFM and Interview of Field Model) is constructed based on query-like language model. A discovery algorithm (IoFM + Exp) that takes into account the user's expertise. To build user domain attention model based on query-like language model, we use formula (2.10) to calculate the user's score and find recommendations accordingly. This article begins with \( \lambda=0 \) and iterates step by 0.1. The IM method experiment results are shown in Figure 2.

It can be seen that when \( \lambda > 0 \), that is to add data smoothing, which can improve the non-smooth model, especially in the range \( \lambda \) from 0 to 0.1, the recommended effect is the greatest. However, as the proportion of global data increases, the recommendation effect decreases as the proportion of global data is greater than 0.1. When \( \lambda=1 \), the model degenerates into a global language model, loses the personalization element, all the user scores are the same, the recommendation effect also becomes very poor. This phenomenon is in line with our understanding of the model, and a certain global smoothing will help to improve the fitting problem, but over-smoothing causes the model to lose the area division, unable to generate an effective user sort, and loses the personalization recommendation element.

4 Conclusion

In view of the phenomenon that the questioners mentioned in the medical question-and-answer community don't get a quick and effective answer, this paper puts forward a kind of potential problem-answer algorithm based on the factors such as the attention degree of users. The method uses query-like language model to model the attention degree of user domain to describe the user's degree of interest in different fields, so as to find the respondents who are interested in the problem. The maximum entropy model is used to estimate the user's response quality to describe the user's expertise in answering the medical questions,
and to establish a user-specific model based on the quality assessment of the answer to distinguish the interlocutor from the medical question-and-answer community. This method takes into account many different factors such as attention, expertise and so on, which can be used to model the user and distinguish the user's personalization. In the experiment based on Baidu medical question-answering system data, compared with the existing algorithms, this paper presents a better recommendation effect for the potential problem respondents in this paper.

References


