An intelligent medical guidance system based on multi-words TF-IDF algorithm

Y. S. Lin [1], L Huang [1], Z. M. Wang [2]

[1]Cooperative Innovation Center for Internet Healthcare & School of Information Engineering
Zhengzhou University

[2]Cooperative Innovation Center for Internet Healthcare

Keywords: Internet healthcare; Intelligent medical guidance; TF-IDF algorithm; Cosine similarity algorithm.

Abstract. He traditional human-aided medical guide service costs heavy human resources. The on-line medical guide service is almost human-aided and has uncertain waiting time for patients. Existing medical guidance systems still have insufficiency for improvement in reliability and correctness. This paper proposed an intelligent medical guide system based on multi-word TF-IDF algorithm, which has been implemented on Android mobile platform. This system applies the improved TF-IDF algorithm and the cosine similarity algorithm to calculate the possibility of disease for patients, and gives the user reasonable guidance. The related experiments proved that the improved TF-IDF algorithm in this paper increased the correctness and reliability of medical guidance.

1. Introduction

Human-aided medical guide service is adopted by most of the current hospital, and it takes heavy human resources. Some Internet medical service website and mobile app, use on-line human-aided medical guide service actually, which cannot assure responding to the user timely. Therefore, the use of computer intelligent medical guide system instead of human-aided medical guide service, will effectively save running costs of hospital and medical web sites.

At present, the intelligent medical guide system has two main ways: 1) intelligent medical guide system which based on knowledge [1-2]; 2) intelligent medical guide system which based on similarity calculation [2]. Intelligent medical guide system which based on knowledge base constructs the knowledge base by knowledge and experience of medical experts, and uses fixed diagnosis condition to judge the disease which user may suffer from and to provide medical guide service in the form of expert system. The realization of the knowledge base and diagnosis condition is according to the knowledge and experience of medical experts in different fields and departments, the input and maintenance of the knowledge base mainly rely on the professional, and the upgrade of the knowledge base sometimes needs to change the original knowledge base and diagnosis condition, therefore, the realization and maintenance of the knowledge base is complicated and inefficient, and currently the function of medical guide system which based on the knowledge base is not perfect. For example, the literature [8] deduces department by using symptom only, and do not predict disease. Intelligent medical guide system which based on similarity calculation, develops vector space model by symptoms which user input and symptoms of disease, judges the disease which user may suffer from and provides medical guide service using the similarity calculation. Compared to intelligent medical guide system which based on knowledge base, it is more efficient, easy extension, easy maintenance, see [2]. However, the similarity calculation method ignores the effect to the reliability of comparability of disease and medical guide results, when several symptoms which user input occur in certain diseases at the same time.

To deal with the problems in the intelligent medical guide system, this paper designs and realizes an intelligent medical guide system based on the improved TF-IDF algorithm. The system develops the vector space model by natural language symptoms which patient input and symptoms of disease, calculates weight of the symptoms through the improved TF-IDF algorithm, and then calculate the disease which the user may suffer from by the cosine similarity algorithm, in order to provide users the disease forecast before diagnosis, guide the user for right medical treatment, and improve the efficiency of the patient visiting.

2. Intelligent medical guide system

Preparing the new file with the correct template. This paper designed an intelligent medical guide system based on TF-IDF algorithm, the system framework as shown in figure 1, it consists three modules: 1) UI module, this module include user register, login, input symptoms and display medical guide results. 2) NLP (Nature language symptom) module, the module is the first segment the natural language symptom which is input by user. Word segmentation uses the IK Analyzer which is open-source and based on java. 3) MGC (Medical guide calculation) module, is the core module of this system.

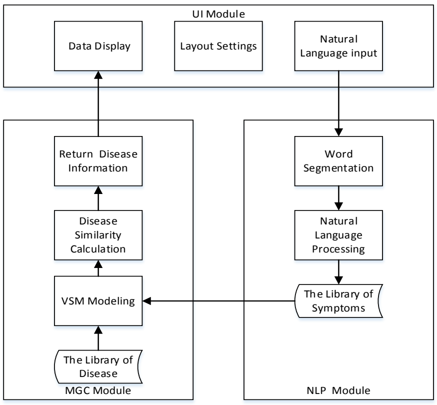

Figure 1. System Framework.

Analysis and improvement for the Algorithm. This paper designed an intelligent medical guide system based on TF-IDF algorithm, the system framework as shown in figure 1, it consists three modules: 1) UI module, this module include user register, login, input symptoms and display medical guide results. 2) NLP (Nature language symptom) module, the module is the first segment the natural language symptom which is input by user. Word segmentation uses the IK Analyzer which is open-source and based on java. 3) MGC (Medical guide calculation) module, is the core module of this system. Respectively VSM m and d for VSM vector modeling. Thus $\vec{m} = (W_{1,m}, W_{2,m}, \ldots, W_{t,m}, W_k)$, $\vec{d} = (W_{1,d}, W_{2,d}, \ldots, W_{t,d}, W_k)$, W_k is the weight of the same symptoms, $W_{i,m}$ is the weight i-th symptoms from m except the same symptoms, $W_{i,d}$ is the weight i-th symptoms from d except the same symptoms, $W_{i,m}$ and $W_{i,d}$ use the TF-IDF algorithm [2] to calculate the weights, W_k using the modified TF-IDF algorithm to calculate weight. Cosine of the Included Angle between two vectors to calculate the similarity between user input and matching diseases. The similarity of Sim (m, d) is given by:

$$Sim\left(m,d\right) = \cos\theta = \frac{\vec{m}\cdot\vec{d}}{\left\|\vec{m}\right\| \times \left\|\vec{d}\right\|} \tag{1}$$

3. INTELLIGENT MEDICAL GUIDE SYSTEM

TF-IDF algorithm is a method to calculating weight, is a kind of effective method always used in information retrieval and data mining. [1] proposed to use the TF-IDF with semantic information to calculate similarity, [3] proposed to use TF-IDF to identify user's hand writing, [4] proposed to use TF-IDF to calculate similarity in the radar intelligence analysis, etc.

Conventional TF-IDF algorithm. Term Frequency and Inverse Document Frequency consists of the weight of the keywords of TF-IDF method. TF is the frequency of a word appears in the document, IDF is the frequency of a word that appears in all the documents, IDF is a measure of the importance of a word.

The formula for the calculation of weight is given as follows:

$$W = TF \times IDF \tag{2}$$

Where in Intelligent Guide System, TF means the frequency of a symptom for a given word appears in the disease, the formula is given by:

$$TF = \frac{n}{m} \tag{3}$$

Where n is the number of this symptom appeared in one of diseases, m is the number of symptoms which be included in this disease.

The word's frequency has the same value due to a symptom just appear once in a disease. The more common symptom, the more difficult to judge the name of the disease. On the contrary, the less common symptom, the much importance it has in the diagnosis of the disease. Thus, in this paper, we use the method of Baidu Index [2], to calculate the weight of a single symptom word. The formula is:

$$TF = \frac{\min(u_i)}{u_i} \tag{4}$$

Where u_i is the symptom of Baidu Index, min (u_i) is the smallest of all symptoms of the disease in Baidu Index.

IDF is the frequency of a symptom appears in all diseases, the calculation formula of IDF is given as:

$$IDF = \log(\frac{A}{A_{in} + 1}) \tag{5}$$

Where A is the total number of the diseases, A_{in} is the number of the diseases which contain same symptom at the same time.

Thus:

$$W = \frac{\min(u_i)}{u_i} \times \log(\frac{A}{A_{in} + 1}) \tag{6}$$

4. Analysis and improvement for the Algorithm

The following two criteria are common used to determine the size of the weights for the symptoms from the perspective of the statistical properties:

1. A smaller Baidu index of Symptoms values, a higher weight value of symptoms;
2. The fewer number of symptoms in all diseases, the higher weight value of the symptoms;

It should be noted that the determination of the disease cannot be judged by one symptom, it can be obtained after combining a variety of symptoms and factors. There will be a varity of clinical symptoms for each disease in disease library, meanwhile, one or more symptoms may be same for a variety of diseases.

Suppose the user enters two symptoms , say n_1 and n_2, then there are three cases of the diseases in the diseases library which contains the symptoms n_1; n_2: only symptom n_1 is included for the disease; only symptom n_2 is included for the disease and both the two symptom n_1 and n_2 are included. The weight value of the disease which contains symptoms n_1 and n_2 cannot be distinguished effectively due to the effect of the first two cases. When the two symptoms n_1 and n_2 are included at the same time for someone disease, say disease D, then the probability of suffering the disease will be increased for the users. However, when applying the original TF-IFD algorithm, the probability of suffering this disease D is lower, and unable to effectively extract the disease which the users may be suffering. Analysis shows that the reason is that the calculation of weights of symptoms n_1 and n_2 are calculated separately, i.e., the number of occurrences for symptoms n_1 and n_2 in all diseases are considered separately, and not taking the case that symptoms n_1 and n_2 occur at the same time into consideration. The disease which the users may be suffering can be clearly distinguished from the medical perspective when two or more symptoms occurs at the same time. However, TF-IDF algorithm does not take this into consideration. Based on the above analysis, when two or more symptoms are output at the same time, the original algorithm cannot effectively analyze the disease which the user may be suffering.

Therefore, when there two or more symptoms occur simultaneously, it cannot separately consider the calculation of the weight for different symptoms, combination weight value of variety of symptoms should be considered. Based on the above analysis, we propose a third criterion:

3) Take the symptoms as a combination symptom when there are two or more symptoms occur simultaneously, and the IDF values is the number of co-occurrence of the symptoms.

Based on the above three criteria, we propose an improved weight calculation method: TF-IDF algorithm based on multi-word. The formula for calculating the weight values of symptoms of the improved TF-IDF algorithm is given by:

$$W = TF \times IDF = \frac{\sum_{i=1}^{k} t_i}{N} \times \log\left(\frac{A}{A_{in}}\right) \tag{7}$$

where K(K >1) is the number of co-occurrence of the symptoms, N is the number of symptoms which be included in disease D, t_i is the number of symptom i occurs in disease D, A is the total number of the diseases, A_{in} is the number of the diseases which contain k symptoms at the same time.

The simulation. We implemented our designated intelligent medical guide system on the Android platform, we use the MySQL as the backend database.

The purpose of the experiment. We implemented our intelligent medical guide system respectively under the traditional TF-IDF algorithm and the modified TF-IDF algorithm. And then the reliability of our system and the reliability of the modified algorithm could be verified through the comparing experiment. During the experiment, a correctly medical guidance means that the disease with the highest possibility in the possible diseases list which is responded by the system to the user is exactly the disease when the patient is diagnosed by a doctor. The correctness means the percentage the system diagnosed the disease correctly over the whole sample set. The reliability implies the value of the highest possibility of the disease when the intelligent medical guide system is used. A high reliability demonstrates the disease which is diagnosed with the highest possibility when using our intelligent medical guide system should with the highest possibility to be the exact disease a patient have, the opposite of the above is also true.

The data of the experiment. The source of the data of our experiment comes from two parts: Some of the data is derived from the medical guidance data which is collected by the diagnosis platform of the people's hospital of Henan province, this data set contains the diseases and the symptoms patients consulted and the diagnosis result the nurse gave to them, we name this data set as H. The other part of the data is collected form consultations form the patients and the corresponding exactly correct answers form the doctors in websites such as Haodaifu, Youwenbida, we name this data set as T.

The data set H and T consist of our whole experimental data. The data set H contains 300 items while the T contains 100 randomly chosen items. In our experiment, we value the answers from the nurses and doctors when they make medical guidance is correct, then we can verify the correctness of the intelligent medical guide system using data set H and the improving ratio of the reliability of the system using data set T.

Table 1. Margin settings for A4 size paper and letter size Paper

Algorithm	Correctness
Traditional TF-IDF	79.7%
Modified TF-IDF	86%

Analysis of the experiment. After test the data set H using two systems we implement under the traditional TF-IDF algorithm and the modified TF-IDF algorithm respectively. We computed, compared and analyzed the result of the medical guidance. The figure 1 presents the result of the correctness under the two algorithms. Our result proves that the modified algorithm can achieve a higher correctness, with 86%, compared with the traditional ones, 79.7%. The average improving percentage of the correctness is 6.3%. Through the analysis, we know that the main reason why a higher correctness cannot be achieved is the insufficiency of the processing technology of the synonymous in natural language process.

Latter, we apply the data set T to test the result of the medical guidance using the two systems we mentioned above, also the comparison and analysis of the reliability of the result is listed below.

We divide the user inputs into two categories, which are the user inputs contain single symptom and inputs which contain multiple symptoms. The experimental result lists as figure 2. We can figure out that when user inputs contain single symptom, the reliability of each algorithm shows no difference when compared with another one, however when the user inputs contain multiple symptoms, the reliability of the modified algorithm shows obvious improvement compared with the traditional ones with the peak point of the percentage of the improvement is 35% and the average percentage of the improvement 11.8%, which means that the modified algorithm shows superb advantage in the reliability of the medical guidance over the traditional TF-IDF algorithm.

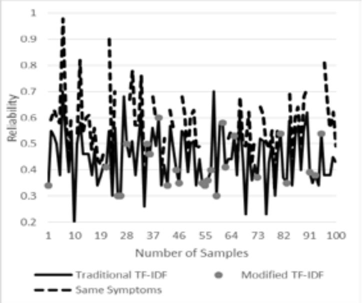

Figure 2. Caption of a typical figure. Photographs will be scanned by the printer.

5. Conclusion

For the traditional TF-IDF algorithm cannot tackle the problem processing multiple symptoms in the intelligent medical guidance system efficiently, this paper proposes an intelligent medical guidance system based on the multi-words TF-IDF algorithm. The experiment shows that the modified TF-IDF algorithm can achieves higher correctness than the traditional TF-IDF algorithm. According to the reliability, on the condition of multi-works inputs, the modified algorithm still shows obvious benefit. At the same time, our system cannot process the symptoms sufficiently in the natural language process, and this could be our focus of the further study.

6. References

[1] Huang C.H et al.2011.A text similarity measurement combining word semantic information with TF-IDF method, Chinese Journal of Computers, 2011, 34(5): 98-106.

[2] Liang, L. 2014. Research on the intelligent medical guide system basing on the VSM weight improved algorithm. M. E. Zhengzhou University.

[3] Quang, A. B et al. 2011. Writer identification using TF-IDF for cursive handwritten word recognition: Kaizhu Huang (ed.), International Conference on Document Analysis and Recognition. Beijing, 844-848 2011.

[4] Su, Y et al.2011.The improvement of VSM model based on semantics. Computer Applications and Software, 28(8):158-161. (VSM)

[5] Song, R. 2013. The research and implementation of method for domain Chinese word segmentation. D. E. Beijing University of Technology.

[6] Wu, J. C. 2010. Research of intelligent guide medical system based on MAS. M. E. Jinan University.

[7] Yu, M et al.2012. Research on intelligence distribution based on TF-IDF classifier, Computer Engineering and Design, 33(5):1822-1826.

[8] Yang, B. 2014. Design and implementation of a comprehensive outpatient medical guide system based on patient's symptoms. M. E. Southwest Jiaotong University.