2017 International Conference on Mathematics, Modelling and Simulation Technologies and Applications (MMSTA 2017)
ISBN: 978-1-60595-530-8

# Intelligent Question and Answering System Based on SVM and Cosine Similarity

Wan-li SONG*, Wei-wei CHEN and Ming-zhu ZHANG

Key Laboratory of Trusted Cloud Computing and Big Data Analysis, School of Information Engineering, Nanjing Xiao Zhuang University, Nanjing, China

*Corresponding author

**Keywords:** Intelligent system, Question classification, Question similarity, Cosine similarity.

**Abstract.** It is a tough task for teachers to answer all questions from students effectively and timely. In this paper, we design and implements an intelligent question answering system using Natural Language Processing, template classification, support vector machine. This system also calculates the similarity between the question and answer pairs by cosine similarity algorithm, and returns the most similar answer. If the user is not satisfied with the answer, the system will write the question into the public section to fall back on other users. The answer will be evaluated and added to the QA base if it is passed with the corresponding question. So that the questions and answers in the QA base continue to expand. We use the QA base of a network forum as the basic library to carry out the experiments. The implementation and experimental results indicate that the proposed approach is achievable.

## Introduction

Intelligent answering system, also known as QA system, with the continuous development of Internet technology and natural language processing technology, the research of intelligent answering system is also continuously promoted [1]. It allows the user to ask questions in natural language, and the system extracts the text information from a certain information source through reasoning analysis, and feeds back valid answers to users. Intelligent question answering system is divided into two categories: open area and closed area. Closed area limits users' questions to a certain area. Open area does not set the scope of the question. The questioner can ask any question of his interest and can Get a satisfactory answer from the system. Currently popular Q & A robot mostly based on the open field, such as Microsoft ice based on the Internet expected and user click log, Baidu voice search assistant based on Baidu search log and so on. These developmental robots are unable to give accurate answers to specific areas such as government, finance, insurance and education. Q & A system of bank, financial question and answer system, e-commerce machine customer service and other question answering system is for a specific area. Such as the Beijing Institute of Technology natural language processing laboratory developed by the Bank of intelligent question answering system BAQS [2], Harbin Institute of Technology graduate face-to-face financial question and answer system [3], they can always answer the user's problems, great savings Human resources.

Students in the learning process will encounter various problems, need to be promptly answered. This paper studies and designs a set of question answering system which is aimed at the basic knowledge of computer. The system can meet the user's demand for quiz in this field. Intelligent question answering system consists of three parts: problem analysis, information retrieval and answer extraction [4]. This paper is based on the Q & A area of the FAQ (Frequently Asked Questions) library. Mainly on the course of keyword extraction, Chinese question classification, Chinese question type classification, Chinese question similarity calculation research and implementation. The value of this system is to increase students' interest in learning and the efficiency of learning, and to promote the teaching to the direction of intelligence.

## System Architecture

The system process begins with the user's question, and first uses the word segmentation tool to deal with the questions, including the removing of the stop words and user-defined word extraction. Then it judges the type of the question according to the rule-based question classification method, and then classify the question to different courses by the trained SVM model. Then, it calculates the cosine similarity between the user question and the question in the FAQ library, and take out the highest similarity question answer feedback to the user. Since the FAQ library can not contain all the question answers asked by users, the similarity in this case will be very low, the feedback from the system will not satisfy all the users. In this case, the user can issue the question to the public question and answer section to seek help from other users, and the user can select a satisfied answer from all the answers submitted by the other users. Finally, the system will review the question and answer, and if the audit is successful, add the question answer pair to the FAQ. In the process of using, the FAQ library will be constantly improved, then the ability of the system to feed back the correct answer will become more and more [5-7]. The question and answer system architecture is shown in Fig. 1.

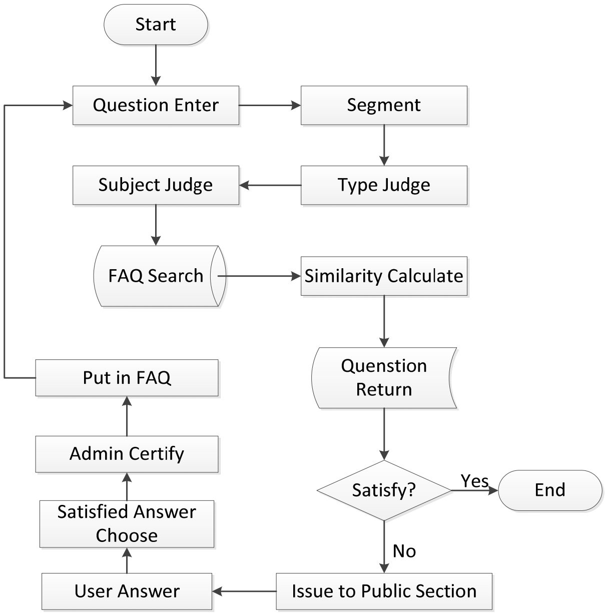

Figure 1. QA system architecture.

## Approach and Implementation

### Extracting Course Key Words Using TF-IDF Algorithm

This paper divides the course into four categories: "data structure", "database introduction", "computer network", "other courses". Each course has its key word [8-9], and the key word is to judge the course of the question. For example, “TCP”, “routers”, “switches”, “network protocols”, and so on are clearly the key words of the computer network. This paper uses the TF-IDF [10] algorithm to collect the key words of each course, and then combine the artificial collection to complement the key words. If a word rarely appears in other articles, but it appears many times in this article, we think that the word is the key word we are looking for. The importance adjustment coefficient is the inverse document frequency (IDF), which is inversely proportional to the common degree of a word. Knowing word frequency (TF) and inverse document frequency (IDF), by multiplying them, you can get a word's TF-IDF [11]. The greater the TF-IDF value, the more important the word is for this article, that is, the key word.. We can formulate the word frequency and inverse document frequency as Eq. 1 and Eq. 2.

$$\text{Term Frequency} = \frac{\text{the number of times that term occurs in document}}{\text{number of words in document}} \quad (1)$$

$$\text{Inverse Document Frequency} = \log\left(\frac{\text{total number of documents in the corpus}}{\text{number of documents where the term appears} + 1}\right)) \quad (2)$$

### Judging the Course Category Using LibSVM

LibSVM[12] is a simple, easy to use, fast and efficient SVM pattern recognition and regression software package developed by Professor Lin Chih-Jen of National Taiwan University and so on. We have multiple courses to categorization, which belong to multiple classification problems. SVM is a two classifier. When encountering multiple classes, [13] generally adopts the following two strategies. One is the one to many method: in training, we classify one sample into one class, and the rest of the samples belong to another class, so that K class samples construct K SVM. In classification, the unknown sample is classified as the class with the maximum value of the classification function. The other is one to one method: the practice is to design a SVM between any two classes of samples, so the sample of the K class needs to design K (k-1) /2 SVM. When an unknown sample is classified, the last category that gets the most votes is the category of the unknown sample. The multiple classification in LibSVM is implemented according to second methods.

We use the key words of the three courses collected by TF-IDF as the attributes of the classification. We specify that the "data structure" of the course is class 1, the course "database introduction" is class 2, and the course "computer network" is labeled as class 3. For example, there are 100 questions in the database of computer network courses. We will select 80 items to be used as training data, and the other 20 to test training models. We divide each question into participles, and compare these words and attribute word sets. If the participle of question appears in the attribute word set, the attribute is 1 and the remaining attribute is 0. In this way, the data structure, database introduction and computer network question of training set are converted to this data format, and transmitted to LibSVM for training and obtaining models.

### Question Classification Based on Pattern Matching

The classification of questions is very important to the question answering system. It can reduce the range of data search and determine the accuracy of the answer extraction at a certain level. By collecting the problem set and analyzing the proportion of each question, this paper puts forward a set of Chinese question classification rules suitable for the field of computer basic question. The system divides the questions into cause, comparison class, description, example and the other class. All of the categories is shown in Table 1.

Table 1. Question classification.

| Type of the question | Feature words | Examples |
|---|---|---|
| Cause | why, for what reason, what for | Why can JAVA cross platform? |
| Comparison | Difference, distinction, divide, similar, relation | What is the difference between an abstract class and an interface? |
| Description | What about, how, describe, introduce | What is the three-way handshake? |
| List | What, which, list, in | What are the features of the object-oriented? |
| Others | none | Hello, JDBC |

Firstly, the question sentence is words-segmented, such as: why can JAVA cross platform? It can be segment into the words "JAVA", "why", "can", "cross", "platform". Comparing the words with the feature words of the cause type questions. It is clear why it belongs to the feature word of the cause type questions. If they cannot match well, it is compared with the feature words of the comparison type, the description type and the list type. If there is no match, we will check whether the feature words in this field are contained in the question sentence. For example, users enter the "data structure", it is clear that there is no question of the appearance of the word. But "data structure" is a feature word in this field, so we consider the user to ask "what is the data structure?" as default. It is about to be classified as a descriptive question. If no interrogative words have no feature words in the field, they are classified as other types of questions. The flow chart of question classification is shown in Fig.2.

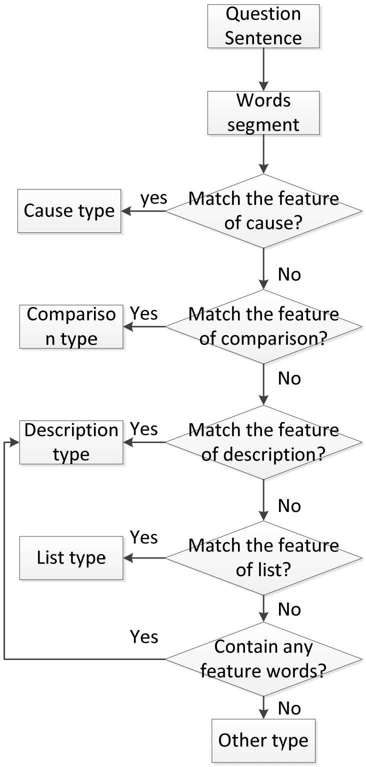
Question Sentence
Words segment
Match the feature of cause?
yes
Cause type
No
Match the feature of comparison?
Yes
Comparison type
No
Match the feature of description?
Yes
Description type
No
Match the feature of list?
Yes
List type
No
Contain any feature words?
Yes
No
Other type

Figure 2. Question classification flow.

## Calculating of Questions Similarity Using Cosine

The computation of question similarity is a key step in this system. Users input the questions, then the answers to the most similar questions are returned to the user [14] from the same type of questions. Firstly, the system converts the question sentences to vectors through word segmenting. Then it calculates the similarity of two sentences using cosine theorem. For example, here are two vectors of two questions sentences processed by our system.

Question A: [1, 1, 1, 1, 1, 0, 0, 0]

Question B: [0, 1, 0, 0, 1, 1, 1, 1]

The above two vectors can be believed that they are all starting from the origin, and the coordinates of the origin are (0, 0, 0, 0, 0, 0, 0, 0). The starting point is the same, but they are not in the same direction. Their angles can explain their similarity. If the angle is smaller, the more similar the direction they refer to, and on the contrary, the more dissimilar. Assuming that A and B are two n-dimensional vectors, A is [A1, A2,..., An], B is [B1, B2,..., Bn], then the cosine formula of A and B.We can formulate the cosine formula as Eq. 3.

$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{3}$$

## Performance Evaluation

### Experimental Settings

The intelligent answer system studied in this paper is based on the FAQ library, and the test data is the answer to all the questions in the FAQ base. The first is to test the course classification model, and there are 1000 common problems in the FAQ base. Among them, "data structure" question 300, "database introduction" question 300, "computer network" question 300, other courses question 100. Each department was taken 80% as the training data, and the remaining 20% was used as the test data of the model. The accuracy rate of the system feedback answers is counted, and the accuracy is influenced by the type of question sentence and the calculation of similarity. 30% of the questions are taken out randomly as input, and the statistical system feedback the accuracy of the answer. In order to obtain better test results, the extracted questions should be evenly distributed in different question categories.

### Experimental Results

The results based on the FAQ gained from the Internet are shown in Table 2. The accuracy rate of category judgment reached 86.4%. The accuracy of data structure type judgement is 86.7%, and the accuracy of database generalization is 85.4%, and the accuracy of computer network is 87.3%. The correct answer rate is 87%, as shown in Table 3.

Table 2. Test results of the question classification.

| Categories | Total number of questions | Accuracy |
|---|---|---|
| Data structure | 60 | 86.7% |
| Database | 60 | 85.4% |
| Computer network | 60 | 87.3% |
| Average accuracy | | 86.5% |

## Summary

In this paper, we showed how to design and implement of an intelligent question answering system. We take the technologies of words segmentation tool to process the question, and filtering stop words; TF-IDF technology to extract the keywords as question curriculum classification basis of courses; question classification method based on pattern matching; LibSVM training question classification model to determine the curriculum of the question; cosine similarity algorithm to calculate similarity between questions. Experiments show that the intelligent answering system proposed in this paper can satisfy users' need for questions and answers, improve students' interest in learning and learning efficiency, and embodies the intelligent of teaching. The next step is to study the semantic understanding of Chinese in order to improve the correct rate of feedback in the system.

## Acknowledgement

This work was partially supported by the following research grants: (1) No. BM2013123 from the Jiangsu Engineering Research Center for Networking of Elementary Education Resources, (2) No. 2016NXY46 from the Research Foundation of Nanjing Xiaozhuang University and (3) the Excellent Graduation Projects (theses) Cultivation Project of Nanjing Xiaozhuang University.

## References

[1] Y. Liu, Research and Application of Question Answering System based on Course Knowledge. Master thesis, Dalian Maritime University(2010).

[2] X.Zh. Fan, H.Q. Li, L.F. Li, et al. Research and implementation of BAQS in banking field Chinese automatic question answering system [J]. Journal of Beijing Institute of Technology, 2004 (6) 528-532.

[3] Y.M. Li, Design and implementation of the automatic question answering system for the bank client service. Master thesis, Harbin Institute of Technology(2012).

[4] Sh. F. Zheng, T. Liu, B. Qin, Sh. Li, Overview of Question-Answering. Journal of Chinese Information Processing, 2002(6) 46-52.

[5] W. Zhang, Research on question-answering system mixed with FAQ, ontology and reasoning technology. Ph.D. thesis, Taiyuan University of Technology(2011).

[6] J.Y. Duan, J. Li, M. Zhang, L. Ma, Research on automatic question-and-answering systems in restricted domains. Journal of North China University of Technology, 2010(1) 23-27.

[7] S.C. Cheng, Design and implementation of a Chinese question answering system based on semantic understanding. Journal of North China University of Technology, 2013(5) 76-83.

[8] Y.Q. Niu, J.J. Chen, L.G. Duan, W. Zhang, Study on classification features of Chinese interrogatives. Computer Applications and Software, 2012(3) 108-111.

[9] F. Jiang, G.H. Li, X. Yue, Semantic-based keyword extraction method for document. Application Research of Computers, 2015(1) 142-145.

[10] Beel J., Gipp B., Langer S., et al. Research-paper recommender systems: a literature survey. International Journal on Digital Libraries, 2015, 17(4) 1-34.

[11] Chang C.C., Lin C.J. LIBSVM: A library for support vector machines. ACM, 2011.

[12] J.E. Zhang, A Chinese keywords extraction approach based on TFIDF and word correlation. Information Science, 2012, 30(10) 1542-1544+1555.

[13] X.Y. Zhuang, Research and implementation of a SVM-based Chinese test categorization system. Master Thesis, Jilin University(2007).

[14] Y.M. Zhou, H. Tao, J.J. Chen, Z.Y. Zhang, Study on sentence similarity approach of automatic ask and answer system. Computer Technology and Development, 2012, 22(5) 75-78.