Research on Industry Question Answering System Based on Near Neighbor Vector Search

Ming Gao¹, Heng Wu², Huangsheng Hua³, Wenfei Zhang⁴*, Zhuozhuo Chen⁵, Dehui Wang⁶
¹²³⁴⁵⁶ Guangzhou Power Supply Bureau, Guangdong Power Grid Co., Ltd. Guangzhou, Guangdong, China, 510630
*1018205008@tju.edu.cn

Abstract. In the field of natural language processing, language-based question answering has been widely studied and made great achievements. As an important form of search engine, question answering system can further improve the search quality compared with traditional search engines. Traditional tree-based indexing methods are inefficient in high-dimensional space. This paper focuses on the application of k nearest neighbor search in large-scale image databases, and researches on high-dimensional indexing technology based on vector approximation method. Fast-Hessian detector is used to detect feature points and generate SURF feature description vector; Then, the initial matching point pair is obtained by fast approximate nearest neighbor search algorithm, and then the unidirectional matching results are matched bidirectionally; at the same time, it makes a detailed analysis of the situation that the events in the sentences are verbs of "abstract change", so as to realize the automatic question and answer after the change of abstract relations between entities.

1. Introduction
With the development of the Internet today, especially the rapid development and wide popularization of broadband Internet, the proportion of multimedia data, especially image data, in the Internet is increasing day by day[1]. In view of this situation, many powerful search engines came into being. These search engines receive the user's retrieval request and return the retrieved web pages related to the retrieval request to the user, and the user looks at these web pages one by one to finally get the information they need [2]. With the rise of deep learning and artificial intelligence, and the great progress of modern computer computing ability, it is possible to use deep learning algorithm to train models. The successful application of Q&A system and Q&A robot in commercial field makes more and more people study it, and becomes one of the hottest research tasks in the field of Natural Language Processing (NLP).

Another extremely popular research direction in machine learning is deep learning. It can be said that deep learning is a higher level development of machine learning [3]. This paper is an exploration of Chinese question answering system, which includes three main components: question processing, information retrieval and answer extraction. For the questions submitted by users, we must first master the users' questioning intentions, classify the questions, determine the requirements that the answers need to meet, and extract the keywords and their extensions from the users' questions. Named entity recognition, entity linking and other technologies are mainly used to obtain answers to questions through knowledge base.

2. System construction and platform design
2.1 Overview of system structure
This system is developed on the basis of the original Bit Question Answering Robot 1 in the laboratory. For the limitation of the knowledge base of the original question answering system, the Internet-based method can solve the unknown questions well. The original vector is represented by fewer bits, so as to reduce the I/O time of the system. By searching in the look up table, the subject words are replaced by the corresponding embedded expression, that is, the word vector. The main tasks to be completed in the problem handling part are: Question pretreatment, question types and answers need to meet the requirements, extract the key words of questions, and expand the key words appropriately according to the types of questions.

This system is mainly composed of three modules: question understanding, information retrieval and answer extraction [4-5]. Each module is interdependent and realizes the final effect of the system together. However, each module is independent of each other, and its task division is different. Each module of the system will be introduced in detail below, and the system flow chart is shown in Figure 1.

---

![Figure 1 Flow chart of question answering system based on web search.](image)

For the problems input by users, the system will first hand them over to the problem understanding module for processing, classify the problem types, and extract the key words of the problems. Then, the generated retrieval formula is given to the information retrieval module to crawl the relevant web pages. Finally, the answer extraction module extracts the correct answer from the retrieved related document set.

2.2 Components of question answering system
The question answering system consists of three modules: question processing, information retrieval and answer extraction. How to fully understand the user's questioning intention in the question processing stage, how to search out the paragraphs containing relevant information in the information retrieval module, and how to accurately generate the answers required by the user in the answer extraction module are the main problems that the question answering system needs to solve.
Deep learning model is closely related to information processing and communication modes in biological nervous system, such as neural coding that tries to define the relationship between various stimuli and related neuron responses in brain [6]. When the dictionary dimension is very large, this representation method will take up a lot of space, resulting in the whole dictionary being extremely sparse, which will cause dimension disaster. According to the approximate vector, the upper and lower bounds of the distance between the feature vector and the query vector can be calculated. When accessing the approximate vector sequentially, most of the data can be filtered out by the upper and lower bounds of the distance, and the search process can be completed only by accessing a few original vectors. The determinant and eigenvalue of the matrix can be used to determine whether it is an extreme point. If the determinant value of Fast-Hessian matrix is positive and the two eigenvalues are not positive or negative at the same time, it is considered as an extreme point.

After the information retrieval module returns the relevant document set, the first job of the answer extraction module is to extract the text of the webpage. In a small interval \([i, j]\), the value of \(y\) will suddenly increase at \(i\), and suddenly decrease at \(j\), showing a continuous peak-shaped area containing the maximum value in the distribution map. Therefore, the interval \([i, j]\) is the text of the webpage, so as long as the sudden increase point \(i\) and the sudden drop point \(j\) are found, the text of the webpage can be determined. The following is the specific flow chart of the algorithm, as shown in Figure 2.

![Figure 2 Flow chart of text extraction algorithm based on line block distribution function](image)

In the process of information retrieval, the keywords input by users do not match the keywords in resources, which leads to the low retrieval efficiency. After receiving the analysis results of the user's questions, the information retrieval module searches the document database within the scope to find the answers, and returns the relevant documents and web pages instead of the exact answers, and the returned contents are disordered, which needs to be processed by the next module. Reasonable features are designed to reflect the similarity of the data in the sense of corresponding types, but are insensitive to irrelevant components in the data. Then, the method based on feature similarity matching is used to index, retrieve and sort the data efficiently. In the case of high dimension, its retrieval performance is still better than the sequential search algorithm, while other index structures cannot guarantee that the retrieval performance is still better than the sequential search method when the dimension is high enough [7]. In fact, the basic idea of hashing is to establish a hash function to project high-dimensional feature vectors into a set of short binary codes, so that similar data samples can be mapped into similar hash codes.

Assume that the current \(k\)-th nearest neighbor distance is \(knn_u[k] dist\). If \(l^m > knn_u[k] dist\) \((1 \leq m \leq d)\)
Then \( l_i \geq l_i^m > k \text{nn} _u[k] \text{dist} \) holds.

According to the above theorem, the vector \( P_i \) can be excluded without high-dimensional calculation, thus reducing the computational complexity of the algorithm. The initial \( m \) value is 1, if \( l_i^m > k \text{nn} _u[k] \text{dist} \), the vector \( P_i \) is excluded; Otherwise, the following PDS algorithm is adopted, starting from \( m = 2 \), and calculating \( l_i^m \) for each \( m, m = 2, \ldots, 2^L \).

\( P_{i,1} \) and \( Q_1 \) are assumed to be the first dimension components of the transform domain vectors \( P_i \) and \( Q_i \), respectively. The approximate vector of \( P_i \) is \( a_i \), \( a_{i,1} \) is the first dimension component of \( a_i \), the approximate vector of \( Q_i \) is \( a_q \), and \( a_{q,1} \) is the first dimension component of \( a_q \). When using the query vector \( Q \) to query, query the data page where \( Q_1 \) is located from the \( B^+ \) tree, which is called the initial access page. From the initial page access, the data pages are accessed in ascending and descending directions, as shown in Figure 3.

![Figure 3 Data page access mode](image)

In the transform domain PDS algorithm in vector quantization, starting from the initial matching code word, the order of ascending and descending cross access is adopted. Because the code book is stored in memory, this access mode is the best access mode.

The work to be done in the question preprocessing stage is: for questions submitted by users in natural language, we need to extract useful information for subsequent retrieval, but not all the words in the questions are helpful for information retrieval. The biggest function is to represent an entity in the knowledge base as a vector, which brings great convenience for calculating similarity and other subsequent processing. Generally speaking, only a small number of candidate vectors need to be calculated. If the distance lower bound of a candidate vector is larger than the current \( k \)-th nearest neighbor distance, the whole query process will end, and the feature vector set to be accessed in this stage is denoted by \( N_2 \). Therefore, this module will focus on problem type identification. The main idea of question classification is to judge the type of expected answer according to the characteristics of interrogative words and other words in questions.

### 2.3 Problem keyword extraction

Keyword extraction not only directly affects the retrieval accuracy of related webpages in the information retrieval module, but also involves the similarity calculation of candidate answers in the answer extraction module. For general interrogative words, we often need another word called "question focus" in the question to know exactly what the user is asking. Different types of questions often lead to different forms of answers. However, long codes not only have a low recall rate, but also increase the
storage capacity of Hamming codes. This gives the cyclic neural network a powerful ability to process time series information. The cyclic neural network model generates the corresponding output according to the current model state and the input at this moment, and finally updates the model state. The feature vector itself is only an approximate representation of multimedia content, rather than an accurate representation, so accurate neighbor search is not of precise significance. For multimedia information retrieval, researchers put forward the concept of approximate query, and put forward many index structures supporting approximate query.

According to the approximate vector \( a^i \), we can calculate the upper and lower bounds \( u^i \) and \( l^i \) of the distance between the query vector \( q \) and \( p^i \). Let's assume that \( l^i,j \) and \( u^i,j \) are the upper and lower bounds of the distance between \( q \) and \( p^i \) in the \( j \)-th dimension space, and \( l^i \) and \( u^i \) are [8]:

\[
\begin{align*}
    l^i &= [l_{i1}, l_{i2}, \ldots, l_{id}] W[l_{i1}, l_{i2}, \ldots, l_{id}]^T \\
    u^i &= [u_{i1}, u_{i2}, \ldots, u_{id}] W[u_{i1}, u_{i2}, \ldots, u_{id}]^T
\end{align*}
\]

According to the definition of the upper and lower bounds of distance, \( l^i \leq d_{p^i,q} \leq u^i \) can be obtained.

In order to make the descriptor rotation invariant, the orientation of feature points is obtained first. According to statistics, the possibility that people use identical words to describe the same object is less than 20%, and the shorter the keywords, the more common the mismatch; From the perspective of machine learning, this is a dimension reduction operation. The exploratory selection of feature vector and metric distance, as well as the subjectivity of user judgment, ensure that the near neighbor search is a retrieval mechanism that meets the application requirements. Because in the prediction, it may be necessary to consider the remote context information, but the traditional cyclic neural network has no such ability of long-distance learning. Using short hash code, as long as the similar samples of data samples are well preserved in Hamming space, fast approximate nearest neighbor (ANN) search can be realized with constant time complexity. Calculate the similarity of the candidate answers, sort them according to the level of similarity, select the best matching answer with the user's question according to the question analysis result returned by the question analysis module, and finally output it to the user.

There is also an important function in the Retrieval class, whose function is to calculate the weight of the extracted paragraphs. The input of this function is the information of retrieved paragraphs output by parretrieve(Word wd) function, and the output is the weight value of each retrieved paragraph. When calculating the weight of paragraphs, two formulas [9] are needed:

\[
IDF^i = \log(N / f^i_t) \tag{3}
\]

\[
Wd = \sum_{i=1}^{k} (KW^i_t \times TF^i_t \times IDF^i_t) + D \tag{4}
\]

\( f^i_t \) is the number of times the keyword \( t \) appears in the paragraph; \( N \) is the total number of words in the paragraph;

The other one is:

\( KW^i_t \) is the weight set by the part of speech for the keywords (including their extensions) output by the problem processing module; \( TF^i_t \) is the frequency of keywords appearing in paragraphs; \( IDF^i_t \) is the inverse frequency of keywords appearing in paragraphs for convenience of experiment, the reciprocal of the longest distance of keywords is taken as the density of keywords.

It is the key task of the problem processing module to extract the keywords from the user's questions and send them to the following information retrieval module. The performance of this part will affect the
effect of information retrieval and the extraction of the final answer. Then, according to the "Stop Words List", the worthless stop words in the question is removed, and the meaningless words such as function words in the question are removed according to the part of speech and the common modal particle lexicon, and only the words such as nouns, verbs, gerunds and adjectives are retained as keywords. However, the words in the answer paragraphs are often not the key words when asking questions, but their synonyms or near meanings. In this kind of hashing method, the nonlinear relationship between data samples can not be solved well.

3. Result analysis
Parallel indexing technology is one of the effective methods to improve retrieval performance. Parallel indexing here mainly refers to distributing data to multiple computers according to certain partition rules to establish index structures respectively. The object of answer extraction is the retrieval result submitted by the information retrieval module. Usually, the information retrieval module returns a bunch of web pages, while the question answering system needs to return short and highly accurate answers. To verify the performance of the algorithm, the algorithm is implemented in C++ language under the environment of VS2005. Through matching experiments on several different types of image pairs, it is found that the stereo matching algorithm proposed in this paper has high accuracy. Besides the questions frequently asked by users, it also includes the problems in the use of the system. The establishment of this answer database can help users make better use of the functions of the system.

The system has been tested for 3 times, and each time there are 50 original documents containing "drawing and changing" actions, and the content forms in the documents are as follows (some statements may not contain "drawing and changing" actions, but only simple content patterns are given below). When answering such questions, the experimental results of the problem processing module are shown in Table 1.

| Frequency | Average accuracy of problem type location | Average accuracy of answer type positioning |
|-----------|------------------------------------------|--------------------------------------------|
| 1         | 99%                                      | 90.8%                                      |
| 2         | 100%                                     | 99.3%                                      |
| 3         | 97.6%                                    | 89.1%                                      |

The successful application of Fourier transform in signal processing and image compression implies the potential of frequency domain-based methods in the face of complex image problems. Information retrieval is based on QA pairs and community networks in FAQ library. If no completely accurate answers are found between QA pairs and community networks, then continue to search for relevant questions. Therefore, this paper avoids the two difficulties of question classification and answer construction, and takes the frequently asked question set as the main retrieval object of the system, supplemented by network retrieval; Global image features are extracted using CNN network. Finally, the question features and image features are fused together to generate the answer. The fusion between problem features and image features is various, for example, the two features are multiplied or connected in series. Finally, according to the weight, the retrieved paragraphs are sorted and sent to the answer extraction module in turn.

In view of some shortcomings of the original classification system, this paper puts forward several improvement measures. In the following, according to these improvement types, the improvement performance will be tested separately. First of all, aiming at the first improved method, that is, the improvement of character description question type, it is changed to definition type. The experimental data of specific test set are shown in Table 2.

| Type | Total number of questions | Classification correct number | Accuracy |
|------|--------------------------|-------------------------------|----------|

Table 2 Classification performance of improved character description questions
It can be seen from the above table that the overall accuracy of the original system is 78.1% in the set of 360 test questions constructed in this paper. A large number of problem types can be identified accurately.

The main idea is that when the qualifier of unknown answer is extracted in the question analysis module, and then the answer is extracted from the text, the candidate answer closest to the qualifier is more likely to become the final accurate answer. When users ask questions, they will find out all the paragraphs that meet the conditions from the index table according to the keywords extracted by the system, and return them by intersection. Because the index structure is memory resident, the retrieval speed is fast, and once the paragraph is located, it can be further semantic analyzed and processed.

4. Summary

According to the needs of practical application, this topic designs and implements a real-time online question answering system based on search engine search results in detail. Different from general search systems, question answering systems want to analyze users' needs from a semantic level. In order to understand users' real intentions, they need to use natural language methods to deal with problems. By introducing the deep hash method, the high-dimensional pixel-level image is compressed and encoded into a short hash code of tens of bits, and the Hamming distance is calculated by simple XOR operation, so that the similar image which is close to the nearest neighbor can be found, which greatly accelerates the image retrieval speed. Question answering system allows users to ask questions in the form of natural language, and its ultimate goal is to accurately express sentences in the form of simple natural language or submit correct answers to users in the form of abstracts.

References

[1] Zhang chuhan, Zhang jiaqiao, feng Jianlin. aknn-qalsh: plug-in for high-dimensional spatial approximation nearest neighbor retrieval in PostgreSQL system [J]. journal of sun yat-sen university (natural science edition), 2019, 058(003):79-85.
[2] Huang Xianying, Xie Jin, Long Shuyuan. Problem classification method based on word vector and BTM model [J]. Computer Engineering and Design, 2019, 040(002):384-388.
[3] Liu Gehong, Li Jinze, Li Bianting, et al. Research on Intelligent Question Answering System of Contract Law Based on Text-CNN Joint Classification and Matching [J]. Software Engineering, 2020, v.23; No.252(06):8+12-16.
[4] Qiu Yu, Cheng Li, Alghazzawi D. Research on Non-factual Problems Based on Semantic Retrieval in Specific Field Question Answering System [J]. Journal of Peking University (Natural Science Edition), 2019, 55(01):58-67.
[5] Ouyang linyan. application of VSM in automatic question answering system of tourism [J]. journal of Shanxi coal management cadre institute, 2019, 032(002):97-99.
[6] Jiang Longquan, Zhang Bo, Hu Zhipeng, et al. A semantic kernel function based question classification algorithm in question answering system [J]. Journal of Shanghai Normal University (Natural Science Edition), 2018, 047(001):53-56.
[7] Yang Kaiping, Li Mingqi, Qin Siyi. Lawyer Evaluation Method Based on Network Response [J]. Computer Science, 2018,45 (9): 237-242.
[8] Ju Xingxing, Chen Jiawei, Zhang Junrong, Li Gaoli, Xingxing J U, Jiawei C, et al. Connectivity of approximate solution sets for generalized vector equilibrium problems with parameters [J]. Applied Mathematics and Mechanics, 2018, 39(10):1206-1212.
[9] Guo zhaoyi, qi Li. research and development of intelligent question answering system for enrollment field based on word vector [J]. information and computer (theoretical edition), 2019, 426(08):91-92.