Design And Implementation of Intelligent Question-answer System Based on Campus Network Service

Cao Yuanyuan[*1], Huang Jiaqi[*2], Xu Zhiyong[*3], Zou Yuanjun[*4], Su Wei[*5], Pan Xuefeng[*6], and Hou Xiaorui[*7]

School of Medical Information, Changchun University of Chinese Medicine, Jilin, Changchun 130117
[*1] E-mail: 957764134@qq.com
[*5] Corresponding author, E-mail: suwei@ccucm.edu.cn

Abstract: In this paper, the intelligent question-answer system based on the campus network service knowledge graph was proposed. The intelligent question-answer system was convenient to the customer of the campus network, and it can promote the efficiency of campus information services. The entity keyword in the question of the customer was extracted by part-of-speech tagging, then the question template that was mostly closed to the question semantic in the system was calculated by the hybrid algorithm based on the term frequency-inverse document frequency (TF-IDF) algorithm and naive Bayesian classifier. Finally, according to the question type of the question template and the entity keywords in the question, the Cypher statement was constructed to complete the retrieval of the answer from the knowledge graph, and the answer would be returned to the customer. The intelligent question-answer system based on the campus network service knowledge graph can help the operation and maintenance staff of the campus network to reduce the workload. The question-answer accuracy test of the system was conducted in this paper, and the results illuminated that the accuracy of answering questions reached to 86%. The compatibility testing of this system was performed, and the results indicated the system compatible with all mainstream operation systems and browsers. At last, the concurrent request processing capability of the system was measured by Jmeter, the results indicated that for 20 users, the shortest response time is 153ms, the longest response time is 1445ms, the average response time is 395ms, the median response time is 254ms, and the error rate is 0, the throughput is 4.1/sec.

Keywords: Knowledge Graph, Intelligent Question-answer System, Campus Network, Fault Question-answer

1. INTRODUCTION

The concept of knowledge graph was formally proposed by Google in May 2012. With the continuous development of intelligent information services and applications, the knowledge graph plays an important role in applications such as intelligent search, intelligent question answering, big data risk control, and recommendation systems. The intelligent question and answer system is a very important research field in natural language processing, which specifically refers to the computer can automatically answer the questions which queried from the user by analysis of the human language of user[1] . At present, the methods used in the question-answer system based on the knowledge graph mainly employ the technologies as following: information extraction, semantic analysis, and modeling based on space vector [2] . Currently, knowledge graph technology has become an indispensable information organization method for intelligent question-answer service systems [3] . The question-answer systems used in knowledge graphs of various industries fields, such as medical, finance, operation and maintenance, have become common relatively[4] [5] [6] [7] .

With the rapid development of information technology, a large number of information systems and facilities used in teaching, scientific research, office and home life have been built in many universities based on their campus networks[8]. At present, domestic universities usually provide network and information services through face-to-face verbal communication or remote telephone communication. This method not only takes up a lot of time for the customer of teachers and students, but also increases the workload of staff, and the efficiency of operation and maintenance will be very low.

For solving the above problems, the intelligent question-answer system based on the campus network service knowledge graph was proposed in this paper, it can provide intelligent network and information services to the customers online, and improve the efficiency of campus network service, reduce the workload of campus networks operation and maintenance staff.

[Fund Project] CERNET Innovation Project: "IPv6 Smart Campus Network Customer Service Intelligent Voice Question Answering System Development" (NGII20190616); Jilin Traditional Chinese Medicine Science and Technology Project: "Research on Key Technologies of Remote Intelligent Chinese Medicine Interrogation System Based on Internet Plus" (2020001)

2. SYSTEM DESIGN

This system is mainly composed of three parts: the construction of knowledge graph and intelligent question-answer system, and the construction of question answering interactive platform. The overall flow chart is shown in Figure 1. Among them, the realization of the campus network service question-answer system, that is the intelligent question-answer system in the above context, is divided into three steps: the first step is the classification of the question, it is obtained by the model of naïve Bayesian classifier which is acquired by model training. In this model, the characteristic matrix of term frequency-inverse document frequency (TF-IDF) is used for input data, and the templates of question classification according to the knowledge graph of campus networks are used for the model training set, the TF-IDF is employed for the text features extraction, word segmentation of training data set, and vector transformation after word segmentation. The second step is entity recognition. The part-of-speech tagging and matching of entity closely are used for entity recognition. The entity keywords in the questions put forward by users are obtained. Finally, according to the question type of the question template and the entity keywords in the question, the Cypher statements which is transformed from natural language and which can run in the neo4j graph database is constructed to complete the retrieval of the answer from the knowledge graph and return it to the user.

2.1. Construction of The Campus Network Service Knowledge Graph

First of all, the data source of the campus network knowledge graph is acquired from two documents provided by Changchun university of Chinese medicine. It is a guide to solving campus network problems for the teachers and students' customers on the campus. Which content includes some unstructured text and pictures. By analyzing the documents, the six types of entities are extracted to represent the question-answer system, which are the campus network access service, the campus network failure service, the usage of campus network, the campus ID card service facilities, and the campus ID card service business. The 29 entities related to the question and answer of campus network services and 5 types of relationships between entities are extracted. According to the type of problem involved in each entity, the entity's attributes and attribute values are set. For the image content that appears in the attribute value, it is displayed as an absolute path, and add <img_start> and <img_end> tags before and after the path URL to facilitate the subsequent reading of the image content.

The entities, relationships, and attributes involved in the campus network knowledge graph are imported to the neo4j graph database by the py2neo module in python programming language, then the nodes, relationships, and attributes are created in neo4j. The knowledge graph is shown in Figure 2. The entities with the same color in the figure represent for the same entity concept category, and the edges with the same color represent for the same relationship type.

2.2. Implementation of Intelligent Question-answer Module

The intelligent question-answer module is the core part of the system, which mainly includes three parts, namely question classification, entity recognition, and answer query.

2.2.1 Question Classification

The question classification module mainly includes the corresponding training set designed for different templates of query questions[9]. TF-IDF was used for text features extraction, word segmentation of training data set and vector transformation after word segmentation. The model was trained for obtaining the naive Bayesian classifier model. The 51 training sets for question classification are constructed in the system, and they are stored in form of TXT documents.

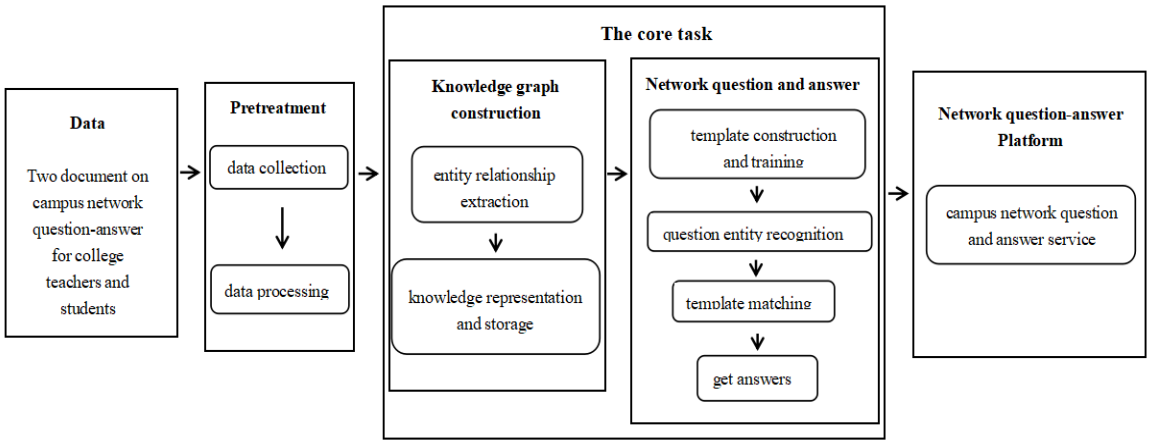

Figure.1 The Overall Flow Chart of The Question-answer System

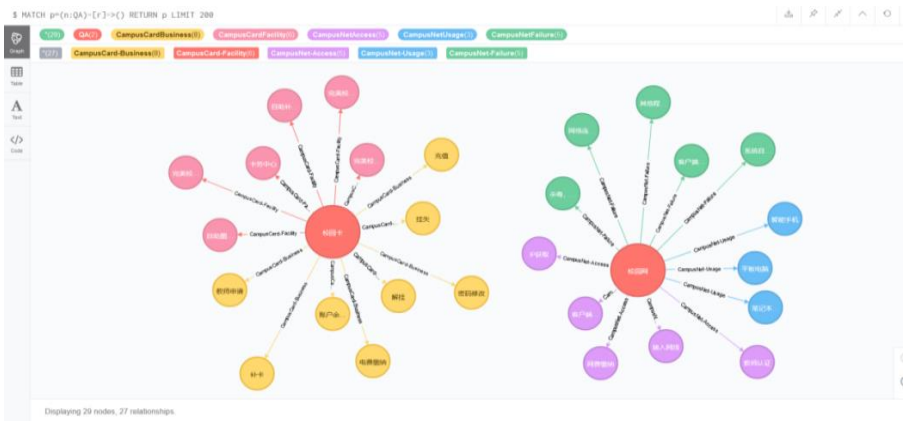

Figure.2 The Campus Network Information Knowledge Graph

i. Text features extraction algorithm based on TF-IDF

TF-IDF algorithm is employed for features extraction and vector transformation of training data. There is an extraction function TfidVectorizer in the Sklearn, a module of python, it is used to transform the training data set into TF-IDF feature matrix, which was used as the input data of naive Bayesian classifier.

TF-IDF is a kind of statistical approach used for information retrieval and text mining commonly, and its value is used to express the importance of each word in a document to enhance the effect of question classification[10] . The term frequency (TF) are used for standing for the frequency of a word (keywords) appearing in the document, it may be computed as follows**:**

$$tf_{ij}=\frac{n_{ij}}{\sum_k n_{kj}} \tag{1}$$

Where, n_{ij} is the frequency of the word appearing in the document d_j , and the denominator is the sum of the number of occurrences of all words in the document d_j.

The inverse document frequency (IDF) stands for the IDF of a particular word, which can be computed as follows:

$$idf_i=\log\frac{|D|}{1+|\{j: t_j\in d_j\}|} \tag{2}$$

Where, $|D|$ is the total number of documents in the corpus, and $|\{j: t_j\in d_j\}|$ stands for the number of documents that contain words t_i (that is, the number of documents that $n_{ij}\neq 0$).

The high-frequency words of a particular document and the low-frequency documents of the word in the corpus can produce TF-IDF with high weight[11] . Therefore, common words tend to be filtered out and important words are retained in TF-IDF.

$$TF\text{-}IDF=TF\times IDF \tag{3}$$

ii. Naive Bayesian classification model

The naïve Bayesian classification model is a simple method to construct classifiers[12] . The Multinomial NB classifier is used for question classification in this question-answer system. In the naive Bayesian classification model, problems were divided into two categories: eigenvectors and decision vectors. The eigenvectors of problems were assumed to independent of decision vectors, and they are non-correlation with each other[13] .

Assume that the eigenvector of the problem is X, $X_i=\{X_1,X_2,\cdots\cdots,X_n\}$ is one of the characteristic attributes. And $X_1, X_2,\cdots\cdots$, Xn are independent of each other. Then P(X|Y) can be decomposed into the product of multiple vectors, namely:

$$P(X|Y)=\prod_{i=1}^{n}P(X_i|Y) \tag{4}$$

According to the Bayesian theorem, formula (4) will be solved by the naïve Bayesian classifier:

$$P(Y|X)=\frac{P(Y)\prod_{i=1}^{n}P(X_i|Y)}{P(x)} \tag{5}$$

Where P(X) is a constant; A priori probability P(Y) can be estimated by the proportion of each class of samples in the training set. Given Y = y, for estimating the classification of test sample X, the posterior probability of Y obtained from naive Bayesian classification is:

$$P(Y=y|X)=\frac{P(Y=y)\prod_{i=1}^{n}P(X_i|Y=y)}{P(X)} \tag{6}$$

At last, the results will be obtained by finding Y, which makes formula $P(Y=y)\prod_{i=1}^{n}P(X_i|Y=y)$. maximization.

2.2.2. Entity Recognition

In this question-answer system, part-of-speech tagging hard extraction and similarity calculation were used for entity recognition. There are the function of word segmentation and part-of-speech tagging in the “Jieba”

module[14] . First, a custom dictionary of campus network service entity names is constructed. A total of 17 kinds of part of speech entity data are constructed, which are saved as TXT documents and added to Jieba to realize the hard extraction of input statement entities.

As for similarity calculation, the entity similarity matching method is used, and it is divided into two parts for calculating. Firstly, the overlap ratio and edit distance between the candidate words are obtained by word segmentation using Jieba and the words in the constructed entity library are calculated, and then, the average of two is used for the score of candidate words. Finally, the word with the highest score was suggested as the approximate result of entity recognition[15] . All entities involved in the custom entity dictionary were stored in the constructed entity library.

2.2.3. Question Query

The question query can be performed when the question classification and entity identification are completed. The natural language input by the user is converted into Cypher query statements that can be executed in the neo4j graph database. The entities are located at the campus network knowledge graph by executing specific Cypher statements, the answers of question attributes of the corresponding entity are acquired.

3. IMPLEMENTATION & TEST OF THE SYSTEM

3.1. Implementation of Question-answer System

The system is implemented using python programming language. Firstly, the system was connected to the neo4j graph database by the three-part module, py2neo. The natural language input by the user is converted into Cypher statements by the method mentioned in the previous section, which was executed to query the graph database and get the answer to the question.

Take the natural language question " How to access the campus network?" as an example to illustrate the implementation process of the system. In the query module, two input parameters, list after part-of-speech tagging of questions and the query template of questions, are imported. The name of the entity is obtained by the word position that is located according to the part of speech, and the number and the abstract templates are stored in the template for matching the corresponding query statements.

3.2. Implementation & Test of System Front-end Interactive Platform

The interactive platform is implemented based on the lightweight web framework Flask. GET mode is selected as the interactive model of the interface.The Ajax interface is called to transmit the information of the user's question to the web back-end in HTML and acquire and the answer of the question.For the display of the picture in the answer, the split() function was used to split the answer character string. The image content is displayed by the image tag, and the processed results are spliced in a regularization expression mode.

Take the natural language question "What is the picture connected to the network cable?" as an example to conduct a question-answer test on the system front-end interactive platform. The screenshot of the test interface is displayed in Figure 3, and the corresponding picture of the question is shown on the screen correctly.

3.3. System Performance Test Analysis

In addition to testing the system functions, the performance testing of the question-answer system is equally important.

3.3.1. The Question-answer Accuracy Test of The System

A total of 100 samples test set data were simulated for question-answer tests from the three major categories of "campus network access and usage", "campus network failure", and "campus ID card service business". The accuracy test results of the question-answer system are shown in Table 1.

From the test results, the system accuracy is about 86%, which indicates that there is an excellent ability to answer-

Figure.3 The Screenshot of Picture Display Test in System Front-end Interactive Platform

Table.1 The Accuracy Test Results of the Question-answer System

The classification of test samples	The number of test samples	The number of samples with correct answers	The correct rate
The campus network access and usage	35	31	88.57%
The campus network fault diagnosis	35	28	80.00%
The campus card service and business	30	27	90.00%
The total	100	86	86%

Table.2 The test aggregation report with Jmeter

Lable	#Sample	Average	Median	90%Line	95%Line	Min	Maximum	Error%	Throughput
HTTP Request	20	395	254	823	972	153	1445	0.00%	4.1/sec
TOTAL	20	395	254	823	972	153	1445	0.00%	4.1/sec

questions system. The content with a relatively high error rate is distributed in the fault diagnosis part of the campus network. If users cannot accurately describe the problem, the question-answer system can easily provide incorrect answers or fail to answer.

3.3.2. The Compatibility Test of The System

The compatibility test is performed to the question-answer system for its compatibility with various browsers, such as Chrome, IE, Firefox, and 360 security browsers, and the results indicate that there is good compatibility to the various browsers, it can be accessed smoothly in various browsers.

3.3.3. The Concurrent Performance Test of The System

The 20 simulate user processes are used for the concurrent requests performance test of system access functions within 5 seconds by the Jmeter automated testing tool. The hardware environment of the performance test machine is Intel Core i5-8265U single processor, 8GB memory, and the operating system is Windows 10. A monitor listener is added for observing the aggregate report, and the results are shown in Table 2.

From the test results, for 20 users, the shortest response time is 153 milliseconds, the longest response time is 1445 milliseconds, the average response time is 395 milliseconds, the median response time is 254 milliseconds, and the error rate is 0, the throughput is 4.1/sec, that is, the number of requested processes completed per second is 4.1. The results of the test indicate that there is a good concurrent performance of this system.

4. CONCLUSION

This paper proposes a design and implementation method of an intelligent question-answer system based on the campus network service knowledge graph. Combining the TF-IDF text feature extraction algorithm and naive Bayesian classification algorithm for question classification, and by identifying the question classification template, and then this question template is used for responding to the Cypher query sentence in the corresponding campus network service knowledge graph, querying the graph, and obtaining the answer for the user's question. For the system Q&A accuracy test, the results showed that the system has an excellent ability to answer questions, with an accuracy of about 86%. It also shows good compatibility with various web browsers. The concurrent request test of the system is conducted, results of which indicate that there is a good concurrent performance. The shortest response time of the process is 153 milliseconds, the longest response time is 1445 milliseconds, the average response time is 395 milliseconds, and the median response time is 823 milliseconds.

Compared with other question and answer methods, this system uses template matching to achieve question and answer, which has better accuracy and stability of question answering. However, there are still some limitations in this system. Firstly, the system only involves the campus network field of Changchun university of Chinese medicine and provides services to the campus teachers and students. At the same time, based on the template matching method, when a new problem template is added to the expanded knowledge

base, the system also needs to be retrained, and there will be a problem of poor scalability of the system.

REFERENCES

[1] WANG Zhiyue, YU Qing, WANG Nan. Survey of intelligent question-answering research based on the knowledge graph. Computer Engineering and Applications, 2020, 56（23）：1-11.

[2] Guo Tianyi, Peng Min, Yimulan. research progress of automatic question-answering in the field of natural language processing[J]. Journal of Wuhan University (Science Edition), 2019, 65(05): 417 - 426.

[3] Zhang Yunzhong, Zhu Rui.The Construction of Knowledge Graph in the LIS Academic Field of Knowledge Question-Answering System: A Multi-Source Data Integration Perspective：a multi-source data integration perspective [J].Information Science, 2021,39(05):115-123.

[4] Wang Jingwei, Xiao Li, Yan Junfeng. Research on the construction of knowledge graph of "Treatise on Febrile Diseases" based on Neo4j[J].Computer and Digital Engineering,[J].Computer System Applications,2010,30(04),2021,30(04):93-98.

[5] Wan Qian, Ouyang Feng, Zhao Ming. Knowledge Mapping of Operation Big Data Analysis for Broadcasting Network[J]. Broadcasting and Television Technology, 2018,45(12):79-86.

[6] Du Zeyu,Yang Yan,He Liang. E-commerce question answering system based on Chinese knowledge graph[J].Computer Applications and Software,2017,34(5):153-159.

[7] Du Zeyu,Yang Yan,He Liang. E-commerce question answering system based on Chinese knowledge graph[J].Computer Applications and Software,2017,34(5):153-159.

[8] Hou Ying, Chen Wensheng, Wang Danning, Cheng Chen, Niu Shichuan, Ji Yao. Research on Intelligent Question Answering Technology in Network Operation and Maintenance Services[J].Software Engineering, 2020,23(09):9-12.

[9] Sun Yi, Li Zhi. Intelligent question answering system for college entrance examination academic planning using Bayesian classification[J].Computer System Applications,2021,30(04):93-98.

[10] Cao Mingyu, Li Qingqing, Yang Zhihao. Primary liver cancer knowledge question and answer system based on knowledgeAD graph[J].Chinese Information Journal,2019,33(06):88-93.

[11] Zhao Shenghui, Li Jiyue, Xu Bi, Sun Boyan. TFIDF-based community question answering system question similarity improvement algorithm[J].Journal of Beijing Institute of Technology,2017,37(09):982-985.

[12] Han Dongfang, Tuerdi Toheti, Eskar Aimdula. Survey of question classification methods in question answering systems[J]Computer Engineering and Applications,2021,57(06):10- twenty-one.

[13] Lin Lingwu, Zhang Chuqi, Zhang Wenliang, Ye Yingya, Chen Ke. Research on automatic question and answer systems in smart wearable devices[J].Journal of Guangdong Institute of Petrochemical Technology,2019,29(01):56-60.

[14] Tian Li, Hong Fubin. Research on the application of natural language processing in the field of power intelligent question answering[J].Technology and Innovation,2021(08):5-8.

[15] Wang Xinlei, Li Shuaichi, Yang Zhihao, Lin Hongfei, Wang Jian. Chinese knowledge graph question answering system based on pre-trained language model[J].Journal of Shanxi University (Natural Science Edition), 2020,43(04):955-96