Modelling and Implementation of a Knowledge Question-answering System for Product Quality Problem Based on Knowledge Graph

Hanxu Liu¹, Fangxu Dong¹, *, Meiqing Wang¹ and Qiu Lin²

¹School of Mechanical Engineering and Automation, Beihang University, Beijing 100191, China
²Beijing Institute of Control and Electronic Technology, Beijing 100038, China

*Corresponding author email: dongfang@buaa.edu.cn

Abstract. Aiming at the problem of difficulty in understanding the semantics of the problem in the traditional quality problem management system, the knowledge retrieval technology of product quality problem based on the knowledge graph is carried out. The process model for knowledge retrieval of quality problem based on semantic templates is constructed. A domain corpus is built, which consisting of thousands of quality problem handling records. The TF-IDF (Term Frequency-inverse Document Frequency) algorithm was used to extracted the vocabulary from the quality problem analysis reports. A natural language question semantic classification process model based on Naive Bayes classifier is established to improve the accuracy of semantic template matching. On the basis of theoretical study, a quality problem knowledge question-answering system-QQ-KQAS based on knowledge graph is developed, and the effectiveness of the proposed method is verified through examples.

Keywords: Quality problem, knowledge graph, natural language retrieval, semantic templates, naive bayes classifier.

1. Introduction

With the continuous deepening of the informatization construction of aerospace enterprises, the aerospace product quality problem knowledge graph, as an important part of the aerospace quality big data management platform, has become the basis for the development of product quality management in the direction of intelligence [1-2]. Question-answering based on the knowledge graph is an important application of the quality problem knowledge graph which provides a natural and intuitive way for the retrieval of quality problem knowledge, and provides an effective way to improve the reuse rate and effectiveness of the quality problem knowledge.

In recent years, question-answering technology based on knowledge graph has received extensive attention from scholars in China and abroad. In addition to its large-scale application in open fields such as the Internet [3], it has also carried out preliminary study and application in vertical fields such as military, medical and manufacturing industries [4-6]. In summary, the question-answering technology based on knowledge graph mainly includes two categories: the method based on semantic analysis and the method based on representation learning. The method based on semantic analysis is to parse the natural language question into the grammatical logic form according to the grammar of the natural language question, and then convert it into the query statement corresponding to the knowledge graph and obtain the answer in the knowledge base [7]. The method based on representation learning regards question-answering as a semantic matching process. The numerical
vector of low dimensional space is obtained through the representation learning of knowledge base and problems, and the answer with the largest semantic similarity with question is obtained by numerical calculation [8]. Representation-based learning methods need large-scale corpus for model training. Mourad et al. [9] have developed a fully automatic QA system in the biomedical domain, which can deal with four types of biomedical questions. Shekarpour et al. [10] have proposed an approach that is able to automatically generate explanations for the QA pipeline, which paves the way for further developing the explainable QA systems. Yang et al. [11] have proposed a Two-stage Multi-teacher Knowledge Distillation method for web QA system, and have tackled the challenge that appears in previous model compression methods, which is the information usually loss during the model compression procedure. Phan and Do [12] have developed a Vietnamese QA system through experimenting on three models, and have achieved both high accuracy and time improvement with the combination of deep learning and knowledge graph. Sawant et al. [13] have presented a system that can deal with the full spectrum of query styles between keyword queries using knowledge graph and web corpus.

Compared with the QA system of knowledge graph in open fields such as the Internet, the quality problem knowledge QA system has the characteristics of limited problem forms and limited data resources, and it needs to make deep use of quality problem for knowledge reasoning. As a result, it is difficult to obtain effective language models through representation-based learning. Therefore, based on the constructed quality problem knowledge graph. The remainder of the paper is organized as follows. First, related semantic templates and the process model of knowledge retrieval based on the templates are proposed in Section 2. Then, methods and advantages on constructing the quality problem domain corpus are presented in Section 3. Question classification method for knowledge retrieval of quality problem based on Naive Bayes classifier are introduced in Section 4. The implementation of the QA system is performed in Section 5 where experimental question and the answer for it with the knowledge graph are presented. Conclusion and future work are finally presented in Section 6.

2. Knowledge Retrieval Method of Quality Problem Based on Semantic Template

2.1. Semantic Template Oriented to Knowledge Retrieval of Quality Problem

The semantic template for knowledge retrieval of quality problem is a formal semantic description model created to describe formatted question text. Semantic templates are generally composed of elements such as “question object”, “question type”, “basic concept”, “event” and “constraint condition”. These elements can be combined to form various types of semantic templates. For example, in the template “which [product] has the most [scrap]?”, “question object” is “product”, “question type” is “interrogative sentence”, “basic concept” is “scrap”, and “constraint condition' is “most”. Among them, “scrap” and “product” in square brackets are parts that users can input on their own. Through this variable parameter standard question statement, a standardized formatting template can be formed. Users can improve the efficiency of retrieval by selecting templates according to their requirements. At the same time, since the semantic relationship of templates is clear when they are defined, the use of semantic templates is helpful for the decomposition of questions and the study of semantic association between components, and the accuracy of knowledge retrieval can be improved. However, with the increasing number of semantic templates, the accurate matching of semantic templates becomes more important.

2.2. The Process Model of Knowledge Retrieval of Quality Problem Based on Semantic Template

The process of knowledge retrieval of quality problem based on semantic template is a process of natural language word segmentation, keyword extraction, keyword-based semantic template matching and entity-relation subgraph query. The process model is shown in Figure 1.

(1) Input a section of knowledge retrieval problem text based on semantic template, and use word segmentation tool to segment the problem text. In the process of word segmentation, domain corpus is needed to improve the accuracy of word segmentation;
(2) Analyse and match the keywords obtained after word segmentation with the entities, concepts, and attributes in the quality problem knowledge graph, and obtain knowledge elements that match the retrieval problem—key entities, concepts, or attributes;

(3) Calculate the correlation between the question text and the semantic template using Naive Bayes classifier, then match and select the semantic template according to the level of correlation;

(4) Integrate the key entities and concepts of the retrieval problem with the matching semantic template to construct a retrieval subgraph, and conduct knowledge retrieval oriented to the quality problem knowledge graph;

(5) Return the search results and visualize them.

In the process of knowledge retrieval of quality problem based on semantic template, the construction of domain corpus is the basis for improving the accuracy of natural language problem recognition, and the classification algorithm based on Naive Bayes classifier is the key to achieving accurate template matching.

Figure 1. Knowledge retrieval process model of quality problem based on semantic template.

3. Construction of Quality Problem Domain Corpus

The key of quality problem knowledge retrieval based on natural language is the understanding of question semantics, and the key of semantic understanding is the accuracy of the word segmentation of natural language question. In this paper, HanLP (Han Language Process), a Chinese language processing package, is used as the word segmentation tool. Since its corpus is constructed by the corpus training of People’s Daily in recent ten years, it has a good word segmentation effect in the general domain. But in the domain of aerospace quality problem, the effect of word segmentation is not satisfactory. For example, the domain-specific phrase “quality problem close loop” cannot be identified by the original word segmentation tool. Therefore, in order to make the word segmentation in natural language processing more accurate, it is necessary to construct a domain corpus for knowledge of aerospace quality problems.
3.1. Strategies for the Construction of Quality Problem Domain Corpus

In the process of quality problem solving, quality manager input a large number of quality problem case data into the database. Each quality problem case information includes dictionary data such as product, model, problem occurrence stage and problem cause type et al, and text data such as description of the quality problem phenomenon, corrective measures and cause analysis of the problem et al. For the dictionary data, they can be directly imported into the domain corpus; for text data such as description of the quality problem, the general description form is: a certain failure phenomenon occurs in the function and performance of a product. Most of the core vocabulary are mainly subject-verb phrases and gerunds, which have the characteristics of refined wording, strong independence, language standardization, and a certain degree of repetition. Therefore, this paper uses natural language processing technology to extract keywords from the description text in the quality problem information and import them into the domain corpus to lay the foundation for more accurate word segmentation in natural language processing. Figure 2 shows the model of the construction process of the quality problem domain corpus.

![Figure 2. Construction of quality problem domain corpus.](image)

3.2. Domain Keyword Extraction for the Descriptive Text

The quality problem case information contains descriptive text information such as description of the problem phenomenon, cause analysis of the problem, corrective measures, and the report of quality problem close loop, which contains a large amount of domain knowledge. In order to improve the accuracy of natural language problem retrieval, it is necessary to extract keywords that appear frequently in the descriptive text and have a certain degree of representativeness, and use them as a component of the quality problem domain corpus.

In the process of extracting keywords from descriptive texts, since professional domain vocabulary is generally mixed with general words and they usually appear at the same time, the frequency of general vocabulary will be higher than that of professional vocabulary. If keywords are only extracted by counting word frequencies, a few real key domain words may be missed. This paper uses Term Frequency-Inverse Document Frequency (TF-IDF) algorithm to extract key domain vocabulary from descriptive texts [14]. TF-IDF is a commonly used weighting technology for information retrieval and information exploration. The process model of the construction of a keyword dictionary based on the descriptive text is shown in Figure 3. Firstly, the noise data that do not meet the specification are eliminated, and then the descriptive text is segmented and analysed by HanLP to filter out the identified stop words. Then the useless auxiliary words, conjunctions and other words are filtered out based on the part-of-speech filtering. The next step is to use the TF-IDF algorithm for keyword recognition with the assistance of the synonym dictionary to calculate the scores of each word. The keywords are ranked according to the calculated scores. Then the domain experts will identify the keywords based on actual experience. The keywords that have passed the review are added to the domain corpus.
4. Semantic Template Matching Based on Naive Bayes Classifier

4.1. Principle of Naive Bayes Classification

Naive Bayes classification is derived from Bayes theorem. It is a widely used classification model based on probability theory. Bayes theorem is used to describe the numerical relationship between the probability $P(B|A)$, the probability of event B under the premise of event A, and the probability $P(A|B)$, the probability of event A under the premise of event B. The formula is:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{1}$$

Where $P(A)$ represents the probability value of event A, $P(B)$ Represents the probability of occurrence of event B.

The Naive Bayes classification algorithm assumes that the feature conditions are independent of each other, and then obtains the classification algorithm of different class probabilities on the basis of the Bayes principle. The category with the highest probability among the classification categories corresponding to the predicted value is the result of the classification. The process is as follows:

1) Suppose $x = \{a_1, a_2, ..., a_m\}$ is an item to be classified, where $a$ is an independent characteristic property of $x$;

2) $C = \{y_1, y_2, ..., y_n\}$ is the set of all categories;

3) Calculate each category’s conditional probability $P(y_1|x)$, $P(y_2|x)$, ..., $P(y_n|x)$;

4) Assuming that $P(y_k|x)$ is the maximum value of each conditional probability, then $x \in y_k$.

The calculation method of the conditional probability value in step (3) is as follows:

1) Prepare a set of items to be classified as a training sample set, and the classification of the items to be classified is known;

2) Calculate the conditional probability estimation value $P(a_j|y_i)$ corresponding to each category of each feature attribute under each category (where $j=1, 2, ..., m$; $i=1, 2, ..., n$);

3) It can be found through the Bayesian formula $P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)}$ that the probability of occurrence of $x$ is the same. For the item to be classified, the value of the denominator $P(x)$ in the Bayes formula is the same. Therefore, when calculating the posterior probability and comparing the magnitudes, it is enough to find the maximum value of the numerator. The value of the numerator can be expressed as follows because the characteristic attributes are independent of each other:

$$P(x|y_i)P(y_i) = P(a_1|y_i)P(a_2|y_i) ... P(a_m|y_i)P(y_i) = P(y_i) \prod_{j=1}^{m} P(a_j|y_i)$$
4.2. Question Classification Method for Knowledge Retrieval of Quality Problem
The key problem of natural language retrieval in the field of quality problem lies in the computer's understanding of the input natural language questions. The Naive Bayes classifier is used as the basis of the text classification algorithm in this paper. Firstly, a problem classification set is constructed according to business requirements, and a semantic template set corresponding to the classification is manually constructed as a training set for Bayesian classification. Because the quality problem domain is obviously problem-oriented and closely integrated with business requirements, the feature words of the training set are determined manually. After the feature word set number is vectorised, the training set is used as the input of the Naive Bayes classifier to train and form a classifier. Then the natural language question input by the system can be classified and predicted to realize the computer's understanding of the natural language question. The quality problem classification process model based on Naive Bayes classifier is shown in Figure 4.

There are many ways to express natural questions, and many ways to ask questions for the same question category. Although the questions can be asked in many ways, the intentions of users' inquiries are similar. By classifying natural language questions with consistent query intentions, and manually defining category labels, accurate classification of questions can be achieved. The Naive Bayes classifier realizes the classification of the problem by identifying the characteristic words in the natural language question.

**Table 1.** Concept entity -- character code.

| Conceptual entity                | Character code |
|----------------------------------|----------------|
| Quality problem                 | qm             |
| Model                           | mq             |
| Product                         | pt             |
| Work stage                      | we             |
| Problem cause type              | pn             |
| Department                      | dt             |
| Model development stage         | de             |
| Standard terms                  | sd             |

After defining the problem classification template, the feature words in the problem training set are manually extracted. For the identified key concept entities in the quality problem domain, the feature word codes should be used to represent the key entity features for the computer to understand, as
shown in Table 1. The identified key concept entities include quality problem, model, product, work stage, problem cause type, department, model development stage, and standard terms. The corresponding feature word codes are “qm”, “ml”, “pt”, “we”, “pn”, “dt”, “de”, and “sd”.

5. Implementation and Case Study
On the basis of the product quality problem knowledge graph established by the Quality Engineering Laboratory of Beihang University, the "Quick Quality-Knowledge Question-answering System (QQ-KQAS) " was developed. The system is based on B/S architecture, using front-end and back-end separation technology. The back-end uses Springboot integration framework, while the front-end uses Vue.js framework. Neo4j is used as the graph database. And Echarts plug-in is used to realize the visualization of knowledge graph. Besides, the system uses HanLP word segmentation tool to realize the integration of natural language word segmentation, synonyms and domain corpus. The system realizes the definition of semantic templates and the function of quality problem knowledge retrieval based on semantic template on the basis of the constructed quality problem knowledge graph. Figure 5 shows the query results and visualization of the results based on the natural language problem "what nonconformity quality problems have occurred in the pipeline centrifugal pump" as an example. There are three nonconformity problems in this product, they are “damage of pump body of pipeline centrifugal pump”, “flow is lower than the designed flow” and “mechanical seal failure of centrifugal pump”.

6. Conclusions and Future Work
This paper studies on the knowledge retrieval technology of model quality problem based on the knowledge graph, and then establishes the process model of quality question knowledge retrieval based on semantic templates. A method for constructing quality problem domain corpuss for dictionary data and descriptive text data is proposed in order to segment natural language questions more accurately. The Naive Bayes classification model is introduced to realize the semantic classification based on natural language questions to solve the problem of accurate matching of semantic templates, which improves the efficiency of knowledge retrieval of quality problem based on
The developed "QQ-KQAS for quality problem based on knowledge graph" has been initially applied in an aerospace company and achieved good application results. Manually constructing semantic templates is time-consuming and laborious. Therefore, in the next step of the thesis, we will focus on the automated construction of semantic templates and the precise matching of semantic templates based on deep learning to further improve the level of intelligence of the question-answering system.

References
[1] Yang B., Mingyan S., Tong Z. Research on the application of aerospace equipment quality knowledge graph [J]. Information Technology and Standardization, 2021(5):27-30.
[2] Yanan C., Pengwu H., Wenbo H., et al. Research on quality management method of aerospace equipment based on knowledge graph[J]. Quality and Reliability, 2019(6):56-60.
[3] Hoy Matthew B. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants [J]. Medical reference services quarterly, 2018, 37(1): 81-88.
[4] Ben D. A study on automatic generation of template library in Intelligent question and answering system in automobiles domain [D]. Shanghai Jiao Tong University, 2017.
[5] Daiyi L., Jie S., Yunxing L., Zongmin M. Military weapon QA system based on knowledge graph[J]. Command Information System and Technology, 2020, v.11; No.65(05):62-69.
[6] Chenhao M. Design and implementation of automatic question answering system based on thyroid knowledge map [J]. Intelligent Computer and Application, 2018, 008(003): P.102-107.
[7] Berant J., Chou A., Frostig R., et al. Semantic parsing on freebase from question-answer pairs[C]. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, 2013:1533-1544.
[8] Cui W., Xiao Y., Wang H., et al. KQBA: learning question answering over QA corpora and knowledge bases [J]. Proceedings of the VLDB Endowment, 2017, 10(5):565-576.
[9] Sarrouti M, Saïd O. SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions[J]. Artificial intelligence in medicine, 2020, 102(Jan.):101767.1-101767.16.
[10] Shekarpour S, Nadgeri A, Singh K. QA2Explanation: Generating and Evaluating Explanations for Question Answering Systems over Knowledge Graph[J]. 2020.
[11] Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, Daxin Jiang. Model Compression with Two-stage Multi-teacher Knowledge Distillation for Web Question Answering System[P]. Web Search and Data Mining,2020.
[12] Truong H. V Phan, Phuc Do. BERT+vnKG: Using Deep Learning and Knowledge Graph to Improve Vietnamese Question Answering System[J]. International Journal of Advanced Computer Science and Applications (IJACSA),2020,11(7):479-487.
[13] Uma Sawant, Saurabh Garg, Soumen Chakrabarti, Ganesh Ramakrishnan. Neural architecture for question answering using a knowledge graph and web corpus[J]. Information Retrieval Journal,2019,22(3-4):
[14] Stephen Robertson. Understanding Inverse Document Frequency: On theoretical argument for IDF [J]. Journal of Documentation, 2004, 60(5): 503-520.