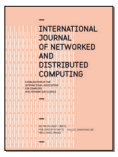
International Journal of Networked and Distributed Computing
Vol. **8**(**3**); June (2020), *pp.* 162–170
DOI: https://doi.org/10.2991/ijndc.k.200515.004; ISSN 2211-7938; eISSN 2211-7946
https://www.atlantis-press.com/journals/ijndc



Research Article

Research on Intelligent Question and Answering Based on a Pet Knowledge Map

Yuan Liu[1,*], Wen Zhang[2], Qi Yuan[1,3], Jie Zhang[2]

[1]*School of Artificial Intelligence and Computer Science, Jiangnan University, No. 1800 Lihu Road, Wuxi 214122, China*
[2]*School of Design, Jiangnan University, No. 1800 Lihu Road, Wuxi 214122, China*
[3]*Department of Industrial Design, Eindhoven University of Technology, Eindhoven, Netherlands*

ARTICLE INFO

Article History
Received 01 March 2020
Accepted 10 April 2020

Keywords
Pet
knowledge map
intelligent Q&A

ABSTRACT

This paper proposes a framework for constructing pet knowledge maps. The schema concept layer is designed and built top-down, and the data layer is constructed from knowledge extracted from semi-structured and unstructured data. In the aspect of entity extraction of unstructured data, a symptom-named entity recognition method combining a Conditional Random Field (CRF) and a pet symptom dictionary is proposed. The method uses a symptom dictionary to identify text and obtain semantic category information. The CRF combines semantic information to recognize and extract symptom entities. Experimental results show the effectiveness of the method. This paper proposes a framework for an intelligent question answering system based on a pet knowledge map. By constructing a named entity dictionary, the problem is abstracted, and the problem is classified by a naive Bayesian text classifier. Through the results of the text classifier, the intent of the natural language question is determined, and the corresponding word order map is matched. The word order map is converted into an OrientDB SQL-like query statement, which is queried in the graph database in which the knowledge map is stored. The example shows that the constructed pet knowledge map and the intelligent question answering system based on the pet knowledge map works well.

1. INTRODUCTION

With the developing economy and society, improvement in people's living standards, increasing work pressures, diminishing urban interpersonal relationships, and many other reasons, families are increasingly owning pets, and pets are increasingly living with people. Changes in family structure and demographics have also led to pets becoming a part of more families. With the "dog person" and "cat addict" becoming a trend in the last several years, the pet economy has continued to grow. According to an analysis in a white paper on China's pet industry in 2018, the number of pets in China exceeded 168 million in 2018, and the main types of pets are cats and dogs. The Internet is one of the most important sources of knowledge about pets and pet medical knowledge. Most pet owners lack pet knowledge. When they need information about their pets, most pet owners use Internet search engines such as Google and Baidu to acquire knowledge. However, this requires considerable time for the pet owner to determine which content contains the information that he or she wants. In many cases, the user wants to acquire further knowledge and needs to read and filter the knowledge again. This leads to lower efficiency of information retrieval, and users are confused by the large amount of information returned by search engines. Therefore, people have a very urgent need for a question and answer system that can return relevant and accurate answers to pet-related questions expressed in natural language.

At present, intelligent question and answering is a hotspot in the field of artificial intelligence research and a trend in technology development. Questions and answers based on knowledge maps allow users to ask questions in the form of natural language and then directly returns the answers needed by users. Intelligent question answering systems are gradually entering various commercial application areas, which will greatly save on labor costs and play a significant role in improving the intelligence of business, industry and service industries. At present, question and answer chat robots based on knowledge bases use Microsoft Xiaoice and Baidu's DuerOS. Therefore, there is significant value in research and applications for building a knowledge database on pets for implementing intelligent questions and answers.

To help users obtain answers from pet encyclopaedias and answers to pet disease problems, this paper constructs an automatic question answering system based on a pet knowledge map. At present, many Internet companies in China and worldwide have built their own knowledge maps and launched searches, questions and answers based on knowledge maps to improve service quality. Many vertical areas have also begun using question answering systems based on knowledge maps. However, there is no mature automatic question answering system based on pet knowledge maps in the pet vertical field. The main work of this paper includes the following:

(1) Pet knowledge map schema (concept) layer construction. According to user needs, a pet knowledge map schema layer is defined and analysed based on a disease encyclopaedia of a pet network.

*Corresponding author. Email: lyuan1800@sina.com

(2) Information extraction. Entity extraction, entity attribute relationship extraction and semantic relationship extraction from different data sources is performed through data crawling, data filtering, cleaning, and parsing to obtain structured pet knowledge. The named entity is obtained by a symptom-named entity recognition model combined with a Conditional Random Field (CRF) and a symptom dictionary. First, by crawling online knowledge a terminology and semantic category information dictionary related to pet medical symptoms is constructed. The semantic category information of the symptoms is added as a feature to the CRF model to obtain more accurate disease symptoms, named entity recognition.

(3) The obtained schema layer data and the instance layer data are stored by the OrientDB [1] graph database, and the OrientDB graph database uses SQL queries.

(4) Building the named entity dictionary. By constructing a dictionary of named entities on pet breeds, disease names, symptoms, and foods, the questions asked by the user are abstracted for later classification by a naive Bayesian text classifier.

(5) Classification of the problem. A naive Bayes-based text classifier is constructed to train the text.

(6) Matching the corresponding word order map. Through the text classifier results, the label of the category corresponding to the problem is obtained, thereby determining the intent of the natural language question and then mapping the determined intent label to the corresponding question template to match the word order graph in the template.

(7) Generating the answer. The word map is converted into an OrientDB class SQL query statement, the answer is queried in the OrientDB graph database with the stored knowledge map, and then the answer to the question is returned to the user.

2. RELATED WORK

In 2012, Google proposed the concept of a knowledge graph [2]. Based on this, a smart search question answering system was built as a new generation information search engine for optimizing the user's search experience. At present, there are many well-known general knowledge maps, such as foreign YAGO [3], Freebase [4], DBpedia [5], Baidu "intimate, Zhishi.me" [6], and Sogou's "Knowledge Cube". The knowledge map provides strong support for natural language understanding, reasoning, questions and answers. Apple's Siri uses knowledge map-related technology, and IBM's Watson system is a knowledge map-based question and answer system. Ali Xiaomi is Ali's customer service chat robot. It uses a combination of knowledge map technology to provide users with personalized services. Service satisfaction has doubled compared with traditional self-service Q&A [7].

Domestic vertical domain knowledge maps, such as the construction of TCM knowledge maps [8], create patterns of knowledge maps based on domain knowledge, and transform structured information in relational databases into RDF data through information transformation. Modules, using multistrategy learning methods, extract information from semi-structured and unstructured data, and finally align data from different data sources.

Research on the construction of bilingual film and television knowledge map [9] and bilingual film and television ontology has been conducted using a semi-automatic method. In the aspect of knowledge links, two entity similarity calculation methods based on word2vec and TFIDF are adopted and based on entity matching, similarity propagation is proposed along with entity matching part of the algorithm. In general, the pet knowledge map of the vertical field in China is still relatively small. In the pet field in China, there is currently no high-quality pet knowledge map.

Usually, the domain knowledge map focuses on the hierarchy of knowledge, and the schema map (schema layer) needs to be constructed in advance. This paper uses a semi-automated knowledge map construction method adopted by most ontology knowledge bases. A pattern diagram (schema layer) is constructed top-down, that is, the concept layer of the pet knowledge map is constructed by hand first, and a data map (data layer) of the pet knowledge map is constructed bottom-up, using various extraction techniques to acquire entities, attributes, and relationships and high-confidence knowledge is extracted into knowledge maps.

In the process of constructing the knowledge map, it is necessary to identify the symptomatic naming entity of the unstructured text data describing the symptoms. Currently, there are many commonly used machine learning models for solving the problem of named entity recognition, such as the Hidden Markov Model (HMM) [10], Support Vector Machine (SVM) [11], and CRF.

The CRF is a statistical sequence labelling algorithm proposed by Lafferty et al. [12] based on the HMM and maximum entropy model. The CRF can effectively overcome the limitations of HMM assumptions and can solve the label offset problem to some extent. The CRF can be seen as an undirected graph model. A commonly used CRF model is the linear chain CRF. Given the sequence of words in the input sentence as the observation sequence, the corresponding output marker sequence is represented, and the conditional probability distribution defined by the CRF is obtained by training to obtain the state sequence at the maximum value. The conditional probability formula for the output sequence in the linear chain CRF is as follows:

$$p(s \mid o) = \frac{1}{z} \exp\left(\sum_i \sum_k \lambda_k f_k(s_{i-1}, s_i, o, i) \right)$$

$$z = \sum_s \exp\left(\sum_{i=1}^{I} \sum_k \lambda_k f_k(s_{i-1}, s_i, o, i) \right)$$

where is the label sequence, is the observation sequence, and is the normalization factor, so that the state sequence probability sum is 1.

The eigenfunction is the corresponding weight. The L-BFGS algorithm is usually used to estimate the CRF. The CRF model has been widely used in medical named entity recognition in recent years and has achieved good results.

At present, there have been many studies on automatic question answering systems worldwide. Pythia et al. [13] is a question and answer system based on an ontology model. Relying on deep language analysis, it is necessary to construct an ontology semantic dictionary. Using the constructed dictionary for semantic understanding can address more complex natural language problems.

TBSL [14] proposed a template-based question and answer method based on the analysis result of the syntax tree, combined with a dictionary to generate the SPARQL template, and then instantiated the SPARQL template through RDF resource mapping to obtain the answer to the question. TBSL has a large number of potential templates for a problem, which is costly and affects system performance.

Yih et al. [15] proposed a question and answer system based on the Freebase knowledge base. The system proposed a new semantic analysis method that maps natural language problems to logical forms of queries and then retrieves the answers. The core of the knowledge base is the knowledge base and the query. It is represented in the form of the graph; the node represents the entity, and the connection relationship between the two entities is represented by the predicate, which simplifies the semantic analysis into the mapping of the problem to the query graph and improves the retrieval efficiency.

Automatic questions and answers based on a knowledge map in domestic vertical fields, such as the e-commerce field question answering system based on the Chinese knowledge map proposed by Zeyu et al. [16], are based on semantic dependence analysis and a reducing dependence algorithm is used to improve the recognition rate of the problem. The semantic slot is extracted, the problem is classified by SVM, the classification result is combined with the semantic slot, and the SPARQL query is constructed, which can better query the e-commerce product.

Chenghao [17] proposed a design and implementation of an automatic question answering system based on a thyroid knowledge map. Based on the maximum matching algorithm, Chinese word segmentation and named entity recognition were performed on the questions, and the entities were classified into questions according to different categories. The problem was to retrieve different templates, analyse the extracted entities with a dependency syntax, obtain the grammatical relationship between the entities, pass the analysed grammatical relations to the invoked template, execute the query in the knowledge map, and obtain the answer to the question.

Automatic question answering systems based on knowledge maps generally have two major problems: understanding the user questions and constructing a knowledge map. The pet knowledge map was constructed in the early stage of this paper. The usual question and answer process is to semantically understand the user's question and then map the question to a structured query statement, such as SPARQL or SQL, to query the entity and relationship in the constructed knowledge map. In this paper, the semantic understanding of user statements is first identified and linked by the entity and then classified by a naive Bayesian algorithm to obtain the intent of the user question. There are many methods for performing entity links, such as keyword matching and similarity calculation based on the neural network method, word2vec.

3. SYSTEM ARCHITECTURE

(1) Construction of the schema layer: The concept layer of the pet knowledge map is constructed in a top-down manner.

(2) Extraction from semi-structured data: Entities, relationships, and attributes are extracted from semi-structured data sources.

(3) Extraction from unstructured data: Named entity recognition and extraction from unstructured data.

(4) Knowledge storage: The pet knowledge map uses the OrientDB graph database storage engine to store the acquired pet knowledge data.

The intelligent question and answer section based on the pet domain knowledge map contains six steps, as shown in Figure 1.

(1) Construction of a named entity dictionary. Build a domain-specific named entity dictionary.

(2) Entity identification and entity linking. Entity identification and entity linking of natural language questions

(3) Natural language abstraction. The user's natural language questions are abstracted to facilitate classification of the classifier.

(4) Classification of problems. The text is classified by the naive Bayesian classification algorithm, which improves the TF-IDF weight calculation method.

(5) Matching the word sequence diagram. The intent obtained according to the problem classifier classification result is mapped to the corresponding question template, and the word order map in the template is matched.

(6) The answer is generated. The word order map is converted into a SQL-like query statement, which is queried in the pet knowledge map, and the obtained result is the answer required by the user.

3.1. Design and Construction of the Schema Layer

The construction of the schema layer is the construction of the entire pet knowledge map framework. The schema defines the

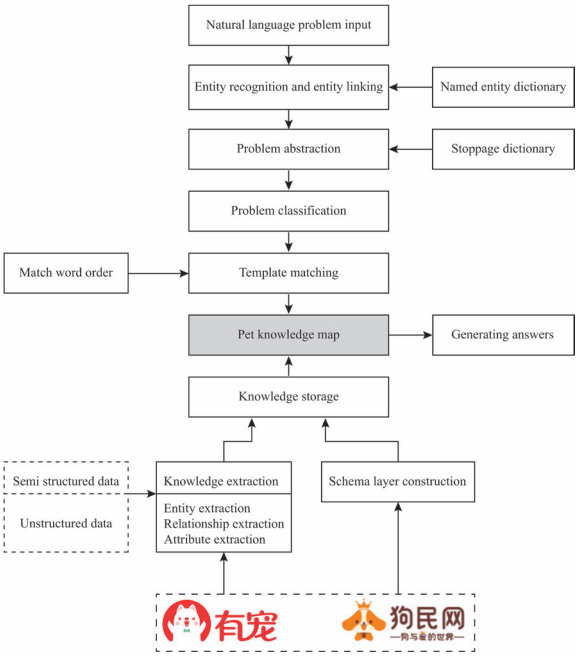


Figure 1 | Flowchart of the intelligent question answering system of the knowledge map of the object domain.

relationship between classes to define the semantic relationship between concepts in the knowledge map.

This paper constructs the knowledge map of the pet field (pet dog- and cat-based), designs and constructs the schema layer of the pet domain knowledge map, and defines the basic four categories: (1) pet breeds, (2) pet diseases, (3) disease symptoms, and (4) pet food.

Following the definition of attributes,

(1) The characteristics of pet breeds include Chinese name, alias, body type, hair length, English name, IQ, origin, weight, life, price, shoulder height, hair colour and function.

(2) The definition of the attributes of pet diseases includes family, overview, cause of the disease, diagnostic criteria, treatment methods and prevention methods.

(3) Attribute definition of pet food: edible.

The above are the attributes of pet species, pet diseases and pet foods are analysed. The symptoms of the disease are quite specific. There are only symptom names, and there is no definition of attribute relationships.

According to the definition of the four categories, three semantic relationships are created:

(1) e_HasDisease (with disease): pet breeds – pet disease, there is a relationship between pet breeds and pet diseases.

(2) e_HasSymptom (symptoms): pet disease – disease symptoms, pet disease and disease symptoms.

(3) e_EatFood (eat food): pet breeds – pet food, there is a relationship between pet breeds and pet food.

The above is the creation of the concept and semantic relationship of the pet knowledge map. The schema of the pet knowledge map is shown in Figure 2.

3.2. Data Sources

The knowledge data in our proposed pet KG were extracted from two Chinese pet websites, namely, "LingDang Pets" and "YouChong"

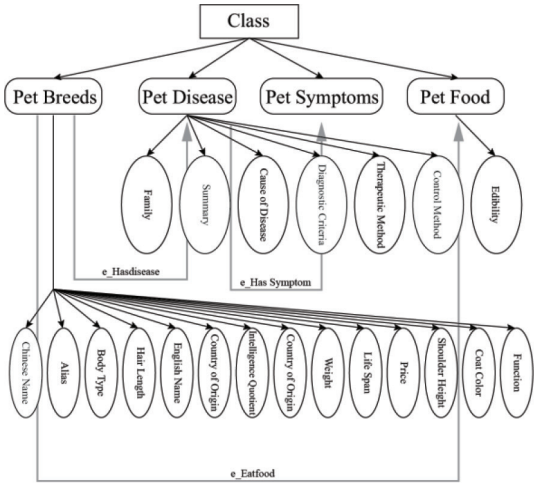


Figure 2 | Pet knowledge map schema layer.

were used to crawl related knowledge materials. Ninety-two kinds of food entities and their properties were extracted from the "LingDang Pets" website. The other 1367 knowledge entities were extracted from the "YouChong" website on encyclopaedic knowledge about pet breeds and pet diseases.

3.3. Semi-structured Data Extraction

First, we use semi-structured data from two source websites to extract the entities of pet breeds, pet diseases and pet foods and the semantic relations among them. In our implementation, a Python library, Beautiful Soup, was used as a parser to parse data from HTML pages. According to the web page layout, the label-based traversal method was used to directly navigate to the key nodes of the DOM tree, which can avoid a large number of nodes traversing operators. Pet breeds and their attributes, pet diseases and their attributes, pet food and the entities of food attributes could all be extracted. Additionally, the semantic relationships among them could also be mined. For example, aspirin poisoning disease for pets is shown in Figure 3.

As shown in Figure 3, the parser extracts a knowledge example, aspirin poisoning disease in pets, where five attributes are also extracted, including the corresponding genus, summary, cause of the disease, diagnostic criteria, and treatment methods of aspirin poisoning. According to our definition of the pet disease attribute, the above information also comprises five "attribute-value" relationships and produces semantics of e_Hasdisease. However, because the symptoms in the attribute text are unstructured, a further symptom-named entity recognition method combined with CRF and a symptom dictionary will be designed to extract detailed symptom entity e_Hassymptom (symptom).

3.4. Unstructured Data Extraction

For disease symptoms, unstructured text analysis methods must be used to identify the entity. In existing machine learning algorithms, CRF can not only use a variety of context features, including words and parts of speech but also external dictionaries. It can achieve good results for the named entity recognition task. In this study, the combined method of CRF and the symptom dictionary was used for symptomatic named entity identification, and the corresponding processing framework is shown in Figure 4.

3.4.1. Data annotations

By searching existing literature and online resources, it was found that there was no publicly available dataset for the identification of symptom-named entities in the pet medical field. Therefore, the experimental corpus was preconstructed, in which a total of 285 texts were used. We chose 100 texts as the training set and 30 texts as the test set, and then the trained model was used to extract symptom entities from 285 unstructured texts.

After the entity analysis from the corpus, the corpus underwent format transform based on the BIESO standard. B-SIGNS, I-SIGNS, E-SIGNS, S, and O marks were used and respectively denoted the head of the symptom, the middle of the symptom, the tail of the

Aspirin Poisoning	
Basic Data	Family: Poisoning Symptoms: Loss of appetite, vomiting
Summary	Inhibition of prostaglandin synthesis, high-dose aspirin can prevent oxidative phosphorylation process, but may also cause hyperglycaemia. At the early stage of the disease, the cat will exhibit shortness of breath, and at the latter stage, it will exhibit inhibited respiration. There will be metabolic acidosis, decreased platelet aggregation and stunted bone development.
Cause of Disease	Accidental ingestion of aspirin (acetylsalicylic acid) or improper dosage of the drug. Puppies are prone to this disease due to the lack of metabolic enzymes, especially those for the synthesis of glucuronides. If the dosage ingested by the dog is more than 60 mg/kg, it can cause poisoning.
Main Symptoms	In the early stage of poisoning, shortness of breath occurs, while in the latter stage, respiration is inhibited; symptoms such as increased body temperature, decreased appetite, vomiting, ulcerative enteritis, and metabolic acidosis will occur; in severe cases, symptoms such as coma, impaired renal function, and bleeding will occur; long-term medicating will cause non-regenerative anaemia in sick dogs; convulsions occasionally occur.
Diagnostic Criteria	Diagnosis and understanding of medical history are very beneficial to the diagnosis of the disease; metabolic acidosis, uric acid, anion gap increase; the content of salicylic acid in serum or urine has certain diagnostic significance. Take 1 ml urine, add 3 drops of 10% ferric chloride after acidification, red indicates that salicylic acid is positive; it could be associated with other diseases causing gastritis and serious metabolic acidosis, such as glycol poisoning, other non-class sterol antibacterial anti-inflammatory drugs, such as ibuprofen poisoning.
Therapeutic Method	Aspirin should be used as early as possible to induce vomiting, wash the stomach, take cathartic drugs and active carbon to further absorb the toxicant; alkalize urine for 36-48 hours to promote discharge of the toxicant: sodium bicarbonate, 50 mg/kg, oral, 2-3 times a day; sodium bicarbonate can also link the collective metabolic acidosis. Supportive therapy: fluid supplement, electrolyte supplement, acid-base balance maintenance; gastrointestinal protective agents and group antagonists (metformin, methylamifoside); alkaline peritoneal dialysis fluid analysis for patients with serious disease.

Figure 3 | Aspirin poisoning disease.

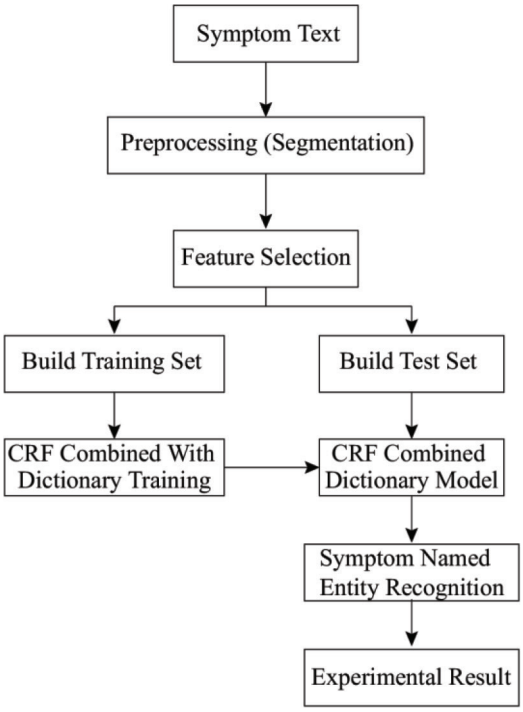


Figure 4 | Key technical framework for symptomatic named entity identification.

Table 1 | Examples of BIESO marked entities

Sentence	BIESO SIGN
The nasal mucosa of the sick dog presents flushing and swelling	Dog/O nasal cavity/B-SIGN mucosa/ I-SIGNS presentation/I-SIGNS flushing/ E-SIGNS/O swelling/S-SIGNS

Table 2 | Category information

Category	Description	Example	Sign
Symptom term	Abnormal performance or discomfort caused by a disease caused by a pet	Vomiting, shortness of breath	BS
Other	Other words in the text	Sick dog, long-term medication	BO

symptom, a single symptom word, and a non-symptomatic word. An example is presented in Table 1.

3.4.2. Named entity recognition method combining CRF and symptom dictionary

To extract symptom entities from unstructured text, a combined method of CRF and a symptom dictionary was introduced. The symptom dictionary was mainly constructed by the information resources from public webs, and two types of descriptions texts were considered, including “symptom label” BS and “non-symptom label” BO. The detailed category information is shown in Table 2.

3.4.3. Feature selection

Feature construction is the key to good recognition performance of symptom entities. In this study, three kinds of features were considered, including "word" linguistic symbol features, part of speech features and symptom dictionary features, as shown in Table 3.

(1) "Word" language symbol feature. The word language symbol feature refers to the word. A word is a linguistic symbol that can be used as a feature to reflect character information. Unlike English, there is no obvious space separator between Chinese words, so the text needs to be segmented before symptom identification. The word segmentation results are then introduced as a word feature.

(2) Part of speech feature. In the entity recognition task of pet disease symptoms, the symptom entity in the text usually appears behind the verb, so the part of speech is characterized as including mainly verbs, nouns and adverbs.

(3) "Dict" dictionary feature. The text contains a large number of professional symptom nouns, so it is necessary to introduce the dictionary features. The dictionary feature is the recognition result of the symptom dictionary to the current word, which includes "BS" and "BO".

3.4.4. Entity extraction experiment

In our study, a total of 285 unstructured texts were used, of which 130 were pre-labelled, 100 texts were used as training data and the other 30 texts were used as test data. To obtain the optimal model parameters, 10-fold cross-validation experiments were performed. The corresponding experimental performance is listed in Table 4. Three typical performance indices, precision, recall and *F*-measure, were used. In the experiments, our hardware platform was A Dell Alienware Aurora R7, CPU 3.7 GHz Intel Core i7, RAM 32 GB, hard disk 2T+512 GB SSD. In addition, two methods, including a single CRF and the combination of CRF and Dict, were carried out to compare the algorithm performance.

$$P = \frac{\text{Number of entities correctly identified}}{\text{Number of entities identified}} \times 100\% \quad (1)$$

$$R = \frac{\text{Number of entities correctly identified}}{\text{Number of entities in standard results}} \times 100\% \quad (2)$$

$$F = \frac{2PR}{P+R} \quad (3)$$

According to the results in Table 4, we can find that the results obtained by CRF + Dict have an obvious improvement over the results obtained by a single CRF. The accuracy, recall rate and *F*-value increased by 5.65%. 6.45% and 6.06%, respectively. From our detailed analysis, the reason was that some symptoms, such as "more drinking to more urine", were rare in the training set, but they could be identified using the symptom dictionary. Based on the trained model, 624 symptom entities on pet diseases were extracted from 285 unstructured texts.

Table 3 | Symptom features

Serial number	Feature	Description
1	Word	Character information of the current word
2	Pos	The part of the word
3	Dict	Semantic categories of current words in symptom terms

Table 4 | Comparison of experimental results

Methods	Precision	Recall	*F*-measure
CRF	0.8413	0.8172	0.8291
CRF + Dict	0.8978	0.8817	0.8897

Table 5 | SQL-like query statement for judging repetitive items

```
"LET $symptom = select from v_symptom WHERE name = '%s';"\
"if($symptom.size()<1){" \
"CREATE VERTEX v_symptom SET name = '%s';" \
"}"%(symptom, symptom)
```

3.4.5. Knowledge storage

To effectively store the knowledge data, the graph database OrientDB was used, which is an open source NoSQL [18] database management system implemented in Java, which is a multimodal library that supports graphs, documents, key-value pairs, object models, and relationships, and provides connectivity between graph data management and logging. The most commonly used query languages are Gremlin [19] and SQL [20], which are used to manipulate property maps and support SQL query data. In detail, the query language of OrientDB introduces some extension of standard SQL statements, so it uses SQL-like statements.

In this study, all extracted entity data in the pet domain were integrated and stored using the OrientDB native database, and the storage language uses the SQL-like language. In the implementation, some schema modes were created, including pet breed (v_Breed), pet disease (v_Disease), food (v_Food), disease symptoms (v_Symptom), disease (e_HasDisease), eat food (e_EatFood) and symptom (e_HasSymptom). Then, all the node information of the corresponding tags and the relationships among these nodes were created. To prevent duplicate node information and duplicate relationships when data information is imported, we used a SQL-like query statement to determine whether there are repetitive items. The corresponding query statement is shown in Table 5.

The above statement first searches the symptom entity in the graph database and then uses the "if" statement to determine whether the symptom entity already exists. If the symptom size() is <1, it means that the symptom entity does not appear in the graph database, then a new entity should be created to represent a new symptom. Moreover, Table 6 lists some statistical information on our created graph database using 285 unstructured texts. Figure 5 gives a visual example of the disease "canine distemper", in which the blue node denotes the entity "canine distemper", the orange nodes indicate nine symptoms of canine distemper, and each edge with "e_HasSymptom" indicates that there is a symptom.

Table 6 | The data statistics of the integrated knowledge base

Statistical item	Numerical values
Number of entity types	4
Relationship type	3
Number of attributes	20
Number of symptom nodes	624
Number of pet breed nodes	357
Number of disease nodes	285
Number of food nodes	92
Total number of nodes	1358
Total number of relationships	79,527

Figure 5 | An example of a pet knowledge map.

3.5. Named Entity Dictionary Construction

The intelligent question and answer system based on the pet knowledge map mainly answers the pet's attribute problems, including alias, price, and IQ. Pet disease's attribute problems include family, symptoms, and prevention, and whether pet food can be eaten. This article is based on pet knowledge. The entities stored in the map construct a dictionary of named entities about pet breeds, disease names, and pet foods and customize the part of the word in the dictionary. As shown in Table 7.

3.6. Entity Identification and Entity Link

At present, there are many open source named entity recognition tools. The mainstream algorithm uses CRF for named entity recognition. However, traditional entity recognition tools cannot effectively identify proprietary domain entities because of their limitations. Only local names, person names, and organization names can be used. Therefore, this article uses a method for building a stopword dictionary. The user enters the natural language, and word segmentation is performed first by the Jieba [21] Chinese word segmentation tool. Then, we establish a stopword dictionary for entity recognition. Our stopword dictionary includes pet breeds and pet diseases

Table 7 | Customized part of speech

Named entity type	Custom part of speech
Pet breed	nm
Pet disease	nd
Pet food	nf

in addition to punctuation marks. Attribute words, such as aliases, prices, symptoms, and what are commonly used, how many, and what the purpose of doing this is to reduce the calculation of the similarity calculated by the identified entity, for example, the question is, "What is the price of a golden retriever?" After our word segmentation, we obtain "What is the price of a golden retriever?" After filtering the stopwords, there is only "golden retriever" left in the question. We need to make physical links. When the user asks, "What is the market price of golden retriever?", after the segmentation is filtered, the entities we need to link to the entity are "golden" and "market" because we mainly link entities to pet breeds, pet diseases and pet foods. Therefore, we filter out the "market" in the similarity calculation and use the entity linked by "golden retriever" to query the answer to the question in the knowledge map.

The entity link links the entities in the text to the entities in the knowledge base. In the text, the entities identified in the user question are linked to the entities in the named entity dictionary. The core of the entity link is to calculate the semantics of the named entity and the candidate entity. Similarly, the candidate entity with the highest semantic similarity is selected as the target entity to be linked [22].

3.7. Problem Abstraction

The abstraction of the problem is to represent the entity that was previously linked by the entity with its corresponding part of speech, mainly for the preprocessing work of the latter problem classification. The pet breed, pet disease name and pet food involved in the user's question are replaced by their part of speech. Consider the following examples:

User source question: What are the symptoms of golden retrievers?

Abstract question: What are the symptoms of nm getting nd?

In the above example, the pet's proper nouns, such as golden retriever, which are involved in the natural language question of the user, will be converted into the golden retriever's part of speech nm after the entity similarity calculation, and the shift is transformed into the distemper word nf instead. The advantage of this is that it can reduce the selection workload of the naive Bayesian classifier feature. Additionally, because there is no special dataset in the pet field, the workload of building the dataset can be reduced, and the required training set can be reduced in size. The specific conversion is shown in Table 8.

3.8. Text Classification Based on Multiple Naive Bayes

This article requires multiple classifications of pet text datasets. At present, there are many machine learning and deep learning

Table 8 | Rule conversion table

Conversion rule	User problem	Abstract problem
Pet breed name — nm	Golden retriever price	Price of nm
Pet disease name — nd	What are the symptoms of golden retrievers?	What is the symptom of nm?
Pet food — nf	Can golden retriever eat grapes?	Can nm eat nf?

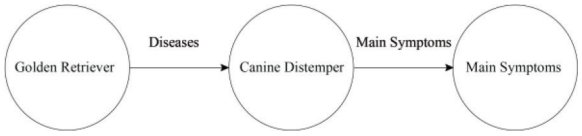


Figure 6 | Example of a word map.

algorithms that can perform multi-classification of texts. Multiple naive Bayes have stable classification efficiency and good performance for small-scale data and multi-classification.

Because there are very few corpora in the pet field, the size of the corpus built in this paper is also very small, so this paper adopts a naive Bayesian text classifier based on polynomials. Based on the knowledge of pet knowledge maps, a total of 24 categories are constructed according to the pet breed, pet disease and pet food attributes. The user's natural language question will match one of the 24 categories after multidirectional naive Bayes classification as the classification results.

3.9. Matching Word Sequence Diagram

Through the classification result of the text quantifier based on multiple naive Bayes, the labels of the categories corresponding to the natural language problem of the user, such as weight, price, and main symptoms, are obtained, which are labels corresponding to the user problem and correspond to natural language questions. Then, the determined intention tag maps the corresponding question template, matching the word order graph in the template. The natural language question basically describes the relationship between the subject and the object, while the graph model can describe the relationship between the node and the node through the edge. The word map is a directed graph, the subject points to the object, and the predicate is used as an edge. In directed graphs, subjects and objects are entities, and predicates are relationships between entities, including attribute relationships. For example, what are the symptoms of a golden retriever with canine distemper? The conversion into a word sequence diagram is shown in Figure 6. This paper constructs a template for a total of 24 types of problems in three major categories. The problem template example is shown in Table 9.

3.10. Generate an Answer

The pet knowledge map is stored in the graph database OrientDB. In this paper, the word map is converted into OrientDB's SQL query statement, the answer is stored in the graph database OrientDB storing the knowledge map, and the answer to the question is returned to the user. The automatic question answering system based on the pet knowledge map supports the origin, price, IQ, disease overview, symptoms, prevention and other issues of pets and can answer three

Table 9 | Example of a problem template

Question type	Problem template
Price	Nm price
The main symptoms	Nm has disease nd main symptoms
Edible	Nm edible nd edible

Figure 7 | Price of a golden retriever.

major questions in total. As shown in Figure 7, the answer is the question of the pet breed attribute, such as the price of a golden retriever.

4. SUMMARY

In this paper, a method for constructing a knowledge map based on data extraction in the pet field is studied, and the whole construction process is described in detail. The knowledge map constructed in this paper is demonstrated by examples, aiming to build a relatively high-quality knowledge base for pet fields.

First, the schema concept layer is constructed in a top-down manner, and the entire pet knowledge map framework is constructed to define the semantic relationship between concepts in the knowledge map. Then, through the extraction of entities, relationships and attributes from semi-structured data, named entity recognition and extraction from unstructured data, in the unstructured knowledge extraction, the named entity recognition of CRF combined symptom dictionary is proposed. The method identifies and obtains the symptom entity. Experiments show that the CRF model combined with the symptom dictionary is better than the CRF model without the dictionary. After obtaining the pet knowledge, the OrientDB native map database is used to store the knowledge, and the built-in visualization of the OrientDB shows an example of the constructed pet knowledge map.

This paper proposes a framework for an automatic question answering system based on a pet knowledge map. The construction process of the automatic question answering system is described in detail, and the built-in intelligent knowledge answering system based on a pet knowledge map is demonstrated by examples.

First, by constructing a named entity dictionary in the pet field, entity identification and entity linking of the user's natural language questions are performed to abstract the problem and facilitate the classification of the latter problem. Then, a naive Bayes-based text classifier is constructed to train the dataset. Through the classification of the text classifier, the label corresponding to the problem is obtained, the intention of the natural language question is determined, and then the corresponding word order map in the template is matched. The word order map is converted into an OrientDB class SQL query statement, which is queried in the graph database

where the knowledge map is stored. The final example shows the built-in knowledge-based automatic question answering system.

Building a question and answer system based on knowledge maps is a complex and long-lasting task. First, the pet knowledge map needs to be improved. For example, pet knowledge is not abundant. It is also necessary to seek more pet knowledge sources to expand the knowledge base and integrate knowledge, including entity alignment and pattern alignment, to study the establishment of the knowledge map update mechanism. Second, the intelligent question answering system based on the pet knowledge map in this paper needs to be improved. For example, the types of pet questions can be increased, the template of the problem can be manually created, the template can be automatically generated, and the reasoning can be answered to make the system smarter.

This paper designs and implements a smart question answering system based on a pet knowledge map, which fills the lack of a question and answer system based on a knowledge map in the Chinese pet field. Additionally, the method to construct an intelligent question answering system based on a knowledge map proposed in this paper has certain reference significance for intelligent question answering systems based on the knowledge map in the vertical field.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

REFERENCES

[1] C. Tesoriero, Getting Started with OrientDB, Packt Publishing Ltd, Birmingham, UK, 2013.
[2] J. Pujara, H. Miao, L. Getoor, W. Cohen, Knowledge graph identification, Proceedings of the 12th International Semantic Web Conference, Sydney, NSW, Australia, 2013, pp. 542–557.
[3] J. Hoffart, F.M. Suchanek, K. Berberich, G. Weikum, YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia, Artif. Intell. 194 (2013), 28–61.
[4] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ACM, Vancouver, BC, Canada, 2008, pp. 1247–1249.
[5] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: a nucleus for a web of open data, International Semantic Web Conference, Busan, Korea, 2007, pp. 722–735.
[6] X. Niu, X. Sun, H. Wang, S. Rong, G. Qi, Y. Yu, Zhishi. me - weaving Chinese linking open data, International Semantic Web Conference, Bonn, Germany, 2011, pp. 205–220.
[7] Z. Shuai, W. Shuai, Y. Yong, et al., Research progress of automatic question answering for knowledge automation, 2017.
[8] R. Tong, S. Chenlin, W. Haofeng, et al., Construction and application of TCM knowledge map, J. Med. Inform. 37 (2016), 8–13.
[9] W. Weiwei, W. Zhigang, P. Liangming, et al., Research on the construction of bilingual film and television knowledge map, J. Peking Univ. (Nat. Sci. Ed.), 52 (2016), 25–34.
[10] S.R. Eddy, Hidden Markov models, Curr. Opin. Struct. Biol. 6 (1996), 361–365.
[11] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Appl. 13 (1998), 18–28.
[12] J.D. Lafferty, A. McCallum, F.C.N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, Eighteenth International Conference on Machine Learning (ICML), Morgan Kaufmann Publishers Inc., San Francisco, CA, 2001, pp. 282–289.
[13] C. Unger, P. Cimiano, Pythia: compositional meaning construction for ontology-based question answering on the semantic web, International Conference on Application of Natural Language to Information Systems, Alicante, Spain, 2011, pp. 153–160.
[14] C. Unger, L. Bühmann, J. Lehmann, A.C.N. Ngomo, D. Gerber, P. Cimiano, Template-based question answering over RDF data, Proceedings of the 21st International Conference on World Wide Web, ACM, Lyon, France, 2012, 639–648.
[15] W-t . Yih, M.W. Chang, X. He, J. Gao, Semantic parsing via staged query graph generation: question answering with knowledge base, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Beijing, China, 2015, pp. 1321–1331.
[16] D. Zeyu, Y. Yan, H. Liang, Question answering system in e-commerce field based on Chinese knowledge map, Comput. Appl. Softw. (2017), 153–159.
[17] M. Chenghao, Design and implementation of an automatic question answering system based on thyroid knowledge map, Intell. Comput. Appl. 8 (2018), 108–113.
[18] R. Cattell, Scalable SQL and NoSQL data stores, ACM SIGMOD Record 39 (2011), 12–27.
[19] F. Holzschuher, R. Peinl, Performance of graph query languages: comparison of cypher, gremlin and native access in Neo4j, Proceedings of the Joint EDBT/ICDT 2013 Workshops, ACM, Genoa, Italy, 2013, pp. 195–204.
[20] H. Hacigümüş, B. Iyer, C. Li, S. Mehrotra, Executing SQL over encrypted data in the database-service-provider model, Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, ACM, Madison, Wisconsin, USA, 2002, pp. 216–227.
[21] J. Sun, 'Jieba' Chinese word segmentation tool, 2012, Available from: https://github.com/fxsjy/jieba (accessed August 25, 2018).
[22] Z. Di, Research and implementation of entity recognition and link, J. Beijing Univ. Posts Telecommun. (2017).