World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:4, 2007

Latent Semantic Inference for Agriculture FAQ Retrieval

Dawei Wang, Rujing Wang, Ying Li, and Baozi Wei

***Abstract*—**FAQ system can make user find answer to the problem that puzzles them. But now the research on Chinese FAQ system is still on the theoretical stage. This paper presents an approach to semantic inference for FAQ mining. To enhance the efficiency, a small pool of the candidate question-answering pairs retrieved from the system for the follow-up work according to the concept of the agriculture domain extracted from user input .Input queries or questions are converted into four parts, the question word segment (QWS) 、the verb segment (VS), the concept of agricultural areas segment (CS), the auxiliary segment (AS). A semantic matching method is presented to estimate the similarity between the semantic segments of the query and the questions in the pool of the candidate. A thesaurus constructed from the HowNet, a Chinese knowledge base, is adopted for word similarity measure in the matcher. The questions are classified into eleven intension categories using predefined question stemming keywords. For FAQ mining, given a query, the question part and answer part in an FAQ question-answer pair is matched with the input query, respectively. Finally, the probabilities estimated from these two parts are integrated and used to choose the most likely answer for the input query. These approaches are experimented on an agriculture FAQ system. Experimental results indicate that the proposed approach outperformed the FAQ-Finder system in agriculture FAQ retrieval.

***Keywords*—**FAQ, Semantic Inference, Ontology.

I. Introduction

THE explosive pace at which information appears on the Internet implies that search engines have become fundamental to Internet usage. Navigating the Internet to obtain specific information is occasionally frustrating, labor intensive, and time-consuming. In the recent decade, question answering (QA) systems have been designed to identify the most similar question-answer pairs with respect to user queries by conventional keyword-based methods. However, keyword-based methods only focus on high-recall retrieval of relevant documents. Inadequate information from keyword query reduces the retrieval precision of the keyword-based approach. Natural language queries can easily capture more complete and precise query information than keyword-based queries. More sophisticated semantic representation and matching method for natural language queries are necessary to obtain high precision retrieval [1], [2]. Rather than generating the desired answers in QA systems, FAQ systems retrieve existing QA pairs from the files moderated by the domain experts. FAQ and question formulation systems using a natural language interface have recently attracted significant attention. Ontology has been adopted for semantic Interpretation [4], [5]. Shallow filtering patterns have been adopted to improve search precision [6]. Burke et al. combined the lexicon and semantic features to extract semantic components automatically from a corpus of questions in FAQ-Finder [3]. Soricut and Brill considered a question as a distorted answer in their statistical noisy channel model [7]. Sneiders interpreted query structures by prioritizing keyword matching and question templates based on entity slots and created an FAQ finding system using case-based reasoning [8]. Chung-Hsien Wu adopts syntactic parser to parse the question category segments of the query and question. The similarity between two question category segments is estimated from the similarity between two QS parse trees [9].

Although in the past open-domain systems were a hot topic for QA systems and information retrieval. A domain-specific FAQ retrieval system with high content-to-noise ratio is still a core research topic for practical consideration. This investigation focuses on agriculture domain FAQ retrieval using independent aspects. Semantic-based FAQ system usually calculates the similarity of user input queries with all of the system's QA pairs. With the popularity of the Internet and the rapid increase of the network information the database increases rapid. The system is inefficiencies to match user's questions with the QA pairs of FAQ database one by one.

This study presents a method that extracts the concept of the agriculture domain from user input for get a small pool of the candidate question-answering pairs and can exactly pinpoint a user's question category by structurally analyzing the Noun-Phrase of a sentence to find questions with the same areas of concern. Input queries or questions are converted into three parts, question word segment (QWS) 、keyword segment (KS) and Noun-Phrase segment (NPS). In this investigation, a query is defined as the user's input in the retrieval phase and a question is the question part of the question-answering pairs in the training corpus. Similarly, the keyword segment (KS) is fed into both question matcher and answer matcher for concept matching based on HowNet and the vector space model, respectively. Finally, a Scoring strategy is adopted to estimate the combination factors to yield the best performance.

This work was supported by the china national 863 Program (Number: 2006AA10Z237)

Authors are with Institute of Intelligent Machines, CAS, Hefei, 230031, China and Department of Automation, University of Science and Technology of China, Hefei, 230026, China.

II. Framework of the Agriculture FAQ System

Adomain-specific FAQ retrieval system with high content-to-noise ratio is a core research topic for practical consideration. In this investigation, a query is defined as the user's input in the retrieval phase and a question is the question part of the question-answering pairs in the training corpus. This study focuses on agriculture domain FAQ retrieval using independent aspects. Semantic-based FAQ system usually calculates the similarity of user input queries with all of the system's QA pairs. With the popularity of the Internet and the rapid increase of the network information the database increases rapid. The system is inefficiencies to match user's questions with the QA pairs of FAQ database one by one. This paper presents a method that extracts the concept of the agriculture domain from user input for get a small pool of the candidate question-answering pairs. Queries using natural language are first converted into word sequences by Part of Speech (POS). Four segments are extracted from the query, the question word segment (QWS) 、the verb segment (VS), the concept of agricultural areas segment (CS), the auxiliary segment (AS), the auxiliary segment is a set of nouns and adjectives which is an additional interpretation part of the concept of agricultural areas segment. Then, a probabilistic mixture model based on the independent aspects provides a powerfully transparent for the causal relation and identifies the desired answer from the answer part of the QA pairs. Fig. 1 illustrates the framework of the described FAQ retrieval system.

A thesaurus constructed from the HowNet, a Chinese knowledge base, is adopted for word similarity measure in the matcher. The similarity between the query and a question is obtained by combining the similarity four segments. In the answer matcher, we used the method mentioned by Chung-Hsien Wu [9], the vector space model (VSM) is applied to measure the similarity between the keyword strings of the query and every answer in the pool of the candidate by checking whether the collocation determined by the keyword segment of input corresponds to the term that may be the answer. The question-answering pair candidates with similar scores are eventually ranked and fed to the mixture model scoring strategy, which is employed to combine the scores from the above matching processes. A list of ranked question-answer pair candidates relevant to the input query is derived from the combined scores.

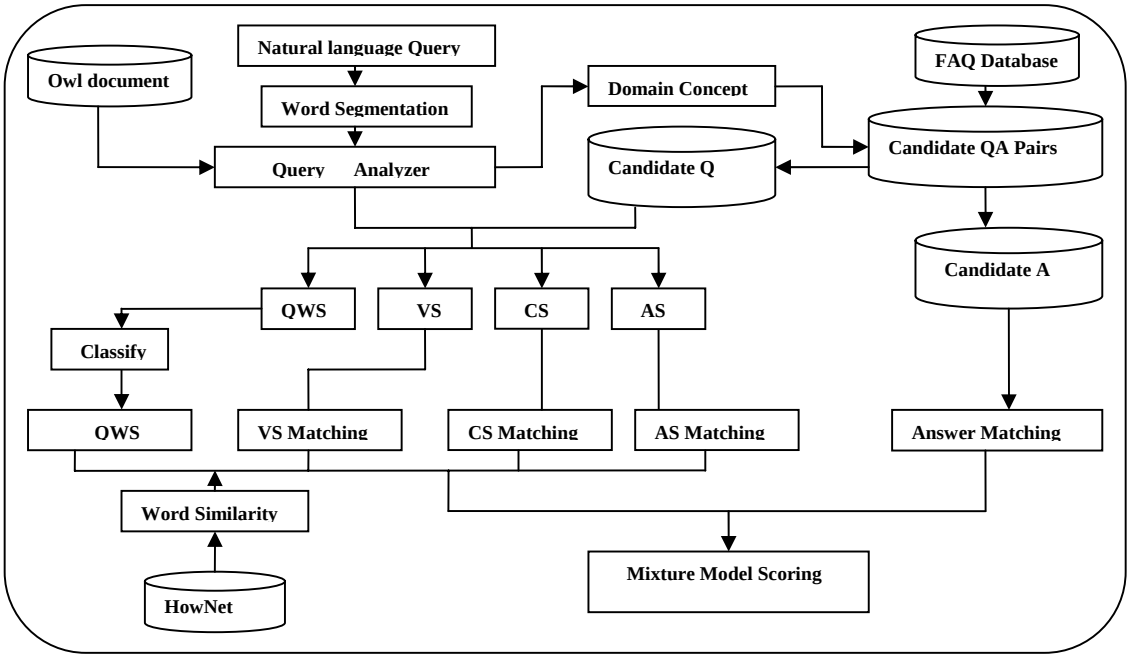

Fig. 1 Framework of the internet agriculture domain FAQ retrieval system

III. Conceptual Models

Ontology is the specification of conceptualizations which provides a shared and common understanding of domain knowledge that can be communicated, integrated, and reused among people and application systems. Ontology comprises primarily two necessary parts, the definition of concepts and the relations among them. Formal ontologies that structure underlying data for the purpose of comprehensive and transportable machine understanding are foundation of the Semantic Web. With the rapid development of Semantic Web, many domain ontologies have been built. For built agriculture domain Conceptual Models, We got the ontology library from Food and Agriculture Organization of the United Nations (http://www.fao.org/aims/tools_onto.jsp). Two classes described below as examples:

```
1.<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_8174">
 <rdfs:label xml:lang="EN">Vegetables</rdfs:label>
 <rdfs:label xml:lang="ZH">蔬菜</rdfs:label>
 <rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_8171"/>
   </owl:Class>
2.<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_8478">
 <rdfs:label xml:lang="EN">Yams</rdfs:label>
 <rdfs:label xml:lang="ZH">山药</rdfs:label>
 <rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_8174"/>
   </owl:Class>
```

By using conceptual model of domain ontology, we can get a conceptual tree, as shown in Fig. 2:

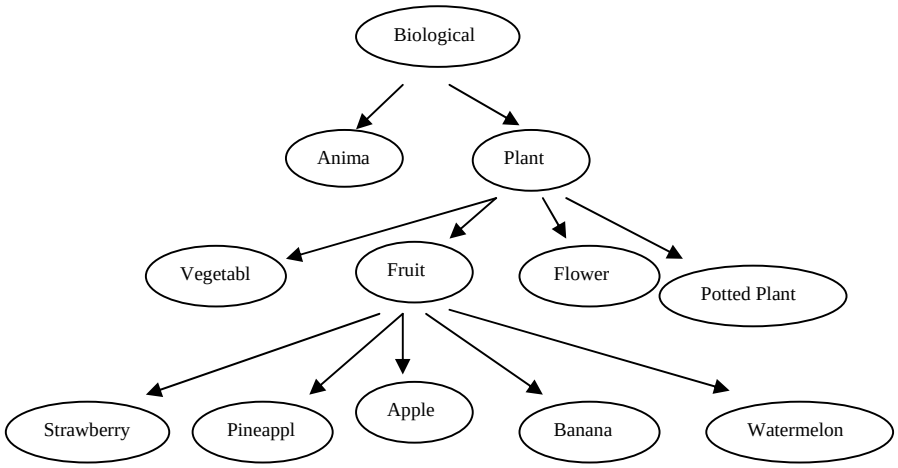

Fig. 2 Concept tree of agriculture

IV. Candidate Pool

In the experiment, we found that in general the centre word and the question words of the question are the core of the question. We discovered that in question time the general users only care about the specific aspects of the problem. For example: when the user input: "how to grow watermelon" or "What will be the prospects of the cow market", the user only wants to know the specific aspects of the "cow" and the "Watermelon". Therefore, we only need to calculate QA pairs of this regard in the FAQ database which was similar with the users' queries. By using agricultural Ontology this paper extract the concept of knowledge in agriculture, index FAQ database. When inquiry, first search the area concept in order to get the set options.

V. Mixture Similarity

The FAQ retrieval system compares the user query with every one of the pool of the candidate QA pairs. In this paper QA pairs can be interpreted based on a set of independent aspects, S= { s_{qws} , s_{vs} , s_{cs} , s_{as} , s_a }.The similarity QA pairs and their associated queries is defined by the mixtures as $Sim(QA,q)=\sum_{s\in S} w_s sim(QA_s,q_s)$, where q denotes the query and QA is a QA pair. The similarity of two word according to The Word Similarity Computing method Proposed by Qun Liu[11].

A. Concept Segment Similarity

The similarity of the CS between of the query and the question is derived from the distance of two concepts,

$Sim(QA_{cs}, q_{cs}) = \frac{\alpha}{d+\alpha}$ Where d is the distance of QA_{cs}, q_{cs} in concept tree of agriculture, α is an adjustable parameters.

B. Question -Words Similarity

In the question matcher, the question stemming words are divided into eleven categories. We expanded the question stemming words of the Hownet for this study. The category of the query and the FAQ questions are used for question words segment matching. The word similarity of the question stemming words between the query and the question is also determined. In the FAQ database, intension is the important factor for an interrogative sentence. Therefore, eleven categories for question stemming words listed in Table I for the questions is derived based on the question stemming keywords. The question stems are used as coarse clues in the identification of the expected answer types. The similarity of the one QA pair with respect to aspect QWS and query q is defined as:

$$\mathrm{Sim}_{qsw}(\mathrm{query}_{qsw}, Q_{qsw}) = \sigma \mathrm{Sim}_{word}(\mathrm{query}_{q\text{-}word}, Q_{q\text{-}word})$$

Where σ =1 if the query's question-words and question's question-words belong to the same category and σ =0 if not.

C. Verb Segment Similarity

A set of verb were extracted from the query and the question. All word of the query's set of verb has a weight, which was determined by the length between this word and the center word of the sentence, the center word in this study is the concept of the agriculture field which extracted from the query. The weight varies inversely with the length. For all word of the query's set of verb, select a verb from the question's set of verb, which most similar to it .If there is not a word similar to it, it matches with null. The Verb Segment Similarity is translating into the weighted average of every verb.

D. Auxiliary Segment Similarity

This similarity is same to the Verb Segment. Final similarity is the weighted average similarity of each word in the auxiliary set.

E. Answer Similarity

In calculate the similarity of Answer part of the query and the answer, the calculation method [9] is adapted to computing. The query and the answer in a QA pair are compared with the VSM-based approach [12]. Two sets of words other than stop words, extracted from the query sentence and the answer in the QA pair, are represented as two descriptions vectors $u = \{a_1, a_2, \ldots\ldots a_n\}$ and $v = \{b_1, b_2, \ldots\ldots b_n\}$, respectively, by the TF-IDF measure.

TABLE I
TAXONOMY OF THE QUESTION STEMMING

Type	Function Word	Intension
1	怎样(what is the state)|怎么样(what is the state)|如何(what is the state)	状态(state)
2	怎(how)|怎么(how)|怎样(how)|怎么样(how)|如何(how)|怎的(how)|怎么办(how to do)|什么方法(the method is)|怎么样办(how to do) |什么办法(the method is)	方法(how)
3	哪些(what for plural)|哪样(what one)|何许(what)|什么(what)|何种(what kind)|哪种(what kind) |何谓(what is)||什么是(what kind)||是什么(what kind)||什么意思(what means)|	什么(what)
4	哪(where)|哪儿(where)|哪里(where)|何在(where)|何处(where)|什么地方|什么地点|哪有(where)|哪个地方(where)|哪些地方(where)	地点(where)
5	为何(for what)|为啥(for what)|为什么(why)|缘何(for what)|干吗(why)|何以(why)|何故(why)|什么原因(what is the reason)	原因(why)
6	多(how)|多么(how)|何等(what degree is)|何其(how)|什么程度(how)	程度(degree)
7	若干(how about)|多少(how about)|几多(how many/much)|几(how about)|几个(what is the number)|几许(how about)	数量(quantity)
8	是…吗(is)|是否(is…… or not)|有无(have…… or not)|有没(have…… or not)|有没有(have…… or not)|是不是(is…… or not)	正反(whether)
9	什么关系(what is the relation)|什么关联(what is the relation)	关系(relation)
10	可否(may)|能否(may)|能不能(can or not)|能够不能够(can or not)|能不能够(can or not)|可以不(can or not)|可以不可以(can or not)|能不(can or not)|可能不(can or not)	能力(capability)
11	多久(how long)|何时(what time)|几时(what time)|什么时候(what time)|什么时间(what time)|哪时(what time)|哪个时候(what time)	时间(when)

VI. EXPERIMENTAL RESULTS

An agriculture domain FAQ retrieval system was constructed in order to evaluate the proposed approach. The QA pairs were collected from the websites and contain a total of 11208 agriculture QA pairs. 100 Artificial queries are given the answer part of the QA pairs. On average, each query corresponds to the answers from 25.40 QA pairs, and there are 5.50 terms per query. The baseline FAQ-Finder system [6] and Keyword FAQ were evaluated for comparison. Finally, the experimental results demonstrate that this study can effectively improve the performance of the agriculture domain FAQ retrieval system compared to the baseline FAQ-Finder system.

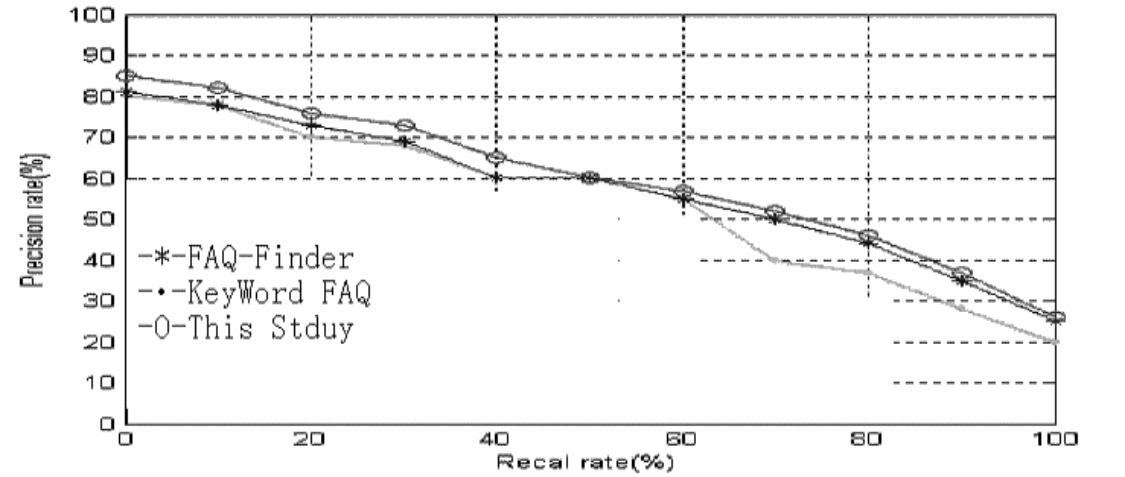

Fig. 4 Precision-recall rate curves for FAQ system

REFERENCES

[1] S. Oyama,T. Kokubo, and T. Ishida, Domain-Specific Web Search with Keyword Spices .EEE Trans. Knowledge and Data Eng.vol.16, no. 1, pp. 17-27, Jan. 2004.

[2] C.O.Kwok , O. Etzioni , and D.S. Weld, Scaling Question Answering to the Web, ACM Trans. Information Systems, vol. 19,no. 3, pp. 242-262, 2001.

[3] R.D. Burke, K.J. Hammond, V.A. Kulyukin, S.L. Lytinen, N.Tomuro, and S. Schoenber, Question Answering from Frequently-Asked Question Files Experiences with the FAQ Finder System, Technical Report TR-97-05, Univ.of Chicago, pp. 1-38,1997.

[4] C.H. Wu, J.F. Yeh, and M.J. Chen, Domain-Specific FAQ Retrieval Using Independent Aspects, ACM Trans. Asian Language Information Processing, vol. 4, no. 1, 2005.

[5] D. Camacho, "Using Hierarchical Knowledge Structure to Implement Dynamic FAQ System, Proc. Fifth Int'l Conf. Practical Aspects of Knowledge Management (PAKM '04), 2004.

[6] V. Jijkoun, J. Mur, and M. de Rijke, Information Extraction for Question Answering: Improving Recall through Syntactic Patterns,Proc. Int'l Conf. Computational Linguistics, 2004.

[7] R. Soricut and E. Brill, Automatic Question Answering: Beyond the Factoid," Proc. Human Language Technology Conf., 2004.

[8] E. Sneiders, Automated Question Answering Using Question Templates that Cover the Conceptual Model of the Database,Natural Language Processing and Information Systems, Proc.Int'l Workshop Applications of Natural Language to Information Systems, pp. 235-239, 2002.

[9] Chung-Hsien Wu, Senior Member, IEEE, Jui-Feng Yeh, and Yu-Sheng Lai, Semantic Segment Extraction and Matching for Internet FAQ Retrieval, IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 7, July 2006

[10] Zhou Qiang, Huang Changning, An Improved Approach for Chinese Parsing Based on Local Preference Information, Journal of Software, Vol.10, No.1, pp1-6, 1999

[11] Qun Liu, Sujian LI .Word Similarity Computing Based on How-net. Computational Linguistics and Chinese Language Processing,China(Taiwan), 2002(7): 59 ～ 76.

[12] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison-Wesley, 1999.