Journal of Southeast University (English Edition) Vol. 23, No. 3, pp. 369 – 372 Sept. 2007 ISSN 1003—7985

Knowledge presentation model for QnA web forums

Yu Shitao Yuan Xiaojie Shi Jianxing

(College of Information Technical Science, Nankai University, Tianjin 300071, China)

Abstract: For an extract description of threads information in question and answer (QnA) web forums, it is proposed to construct a QnA knowledge presentation model in the English language, and then an entire solution for the QnA knowledge system is presented, including data gathering, platform building and applications design. With pre-defined dictionary and grammatical analysis, the model draws semantic information, grammatical information and knowledge confidence into IR methods, in the form of statement sets and term sets with semantic links. Theoretical analysis shows that the statement model can provide an exact presentation for QnA knowledge, breaking through any limits from original QnA patterns and being adaptable to various query demands; the semantic links between terms can assist the statement model, in terms of deducing new from existing knowledge. The model makes use of both information retrieval (IR) and natural language processing (NLP) features, strengthening the knowledge presentation ability. Many knowledge-based applications built upon this model can be improved, providing better performance.

Key words: QnA web forum; knowledge presentation; semantic link; statement model; knowledge confidence

1 Related Work

The question and answer (QnA) web forum is an important type of Internet service, where users can help each other by asking a question, answering a question or voting for a best answer. Live QnA, Yahoo Answers, Baidu Knows and Wondir are examples of this type. Popular QnA web forums usually keep large quantities of QnA threads data with manually labeled "best answer", while only several simple services mentioned above are offered.

In fact, much more valuable applications can be built upon forum data sets, such as question answering (QA)[1], automatic encyclopedia, expert search[2], etc. These issues have become quite popular in recent years, especially since the corresponding tracks appeared at TREC. However, these valuable applications on QnA data sets were mostly built[3–5] with IR methods previously, paying more attention to statistical features of query keywords.

Some recent studies have tried to extract head chunks and related words from sentences, or place a hierarchy classification on questions, aiming at a more exact description on the questions[6–7]. It was also proposed in TREC 2006 to bring semantic relations into the QA system[8], but it mainly focused on well-formed factoid and list questions offered by TREC, and was not quite suitable for QnA web forum data.

In a QnA web forum, the original QnA data sets are in the form of QnA threads, thus the meaning of a word or a sentence depends much on the thread context[9]. In addition, the contents of QnA threads are full of colloquial words and spelling mistakes, thus much noisier than enterprise data sets offered by TREC. So it is difficult to extract or retrieve information directly from QnA threads. To put more QnA knowledge in use, it is necessary to build a new presentation model for QnA web forum data sets, paying more attention to semantic and grammatical information[10]. What's more, some opinions in QnA threads may be disagreed with by other people. So knowledge confidence should also be brought into the model.

In this paper a knowledge presentation platform in the English language is proposed to provide a more exact presentation of knowledge in QnA threads. Many knowledge-based applications can be built on this platform in the future, no longer directly on original QnA threads.

2 System Overview

QnA data sets, QnA knowledge presentation platforms and QnA applications constitute a complete QnA knowledge system. Fig. 1 gives an overview of the system. The original QnA threads data is crawled from QnA web forums and stored in QnA data sets. The QnA knowledge presentation platform is built upon

Received 2007-05-18.

Foundation items: Microsoft Research Asia Internet Services in Academic Research Fund (No. FY07-RES-OPP-116), Tianjin Technological Development Program Project (No. 06YFGZGX05900).

Biographies: Yu Shitao (1981—), male, graduate; Yuan Xiaojie (corresponding author), female, doctor, professor, yuanxj@nankai.edu.cn.

QnA data sets, and provides query services to high-level QnA applications.

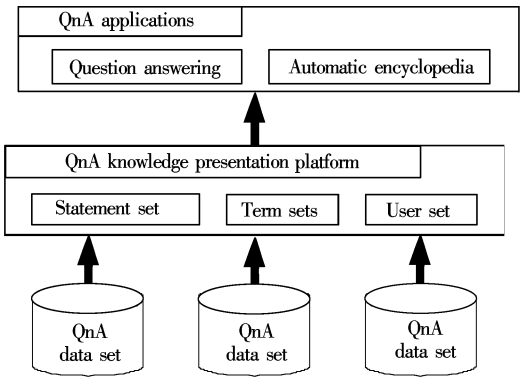

Fig. 1 Architecture of the QnA knowledge system

3 Data Gathering

QnA data gathering can be divided into three steps: data crawling, thread content extraction and data storage. QnA web pages are downloaded by some thread crawlers, and then are parsed into thread contents such as question title, question body, question author, posted time, answer body, answer author, answer voted count, etc. This work is not too difficult to finish since QnA web pages are usually rendered from a well-structured template. Then the parsing results can be stored in relational or XML-enabled databases. After data gathering, all the QnA data sets in Fig. 1 can be built.

4 Platform Description

As is shown in Fig. 1, the QnA knowledge presentation platform is composed of the statement set, term sets and other assist components as the user set. The platform is the most important part of the QnA system. Then various components of the platform are described in this section.

4.1 Terms with semantic links—using semantic information

In the QnA knowledge presentation system, terms are collected from QnA data sets as well as from a predefined dictionary. A term is a basic semantic unit in the form of a word, group of words, or a phrase, but its headword should be a notional word. All terms are stored in their stemmed form, and organized by their parts of speech (POS), which is defined as the POS of their headwords. Thus there will be noun term set, verb term set, adjective term set, adverb term set, etc. A multi-POS term can be found in several term sets, but occurs at most only once in one set. These terms will be used as statement elements in the statement set.

What's more, various types of semantic links can be added to related terms in each term set. For example, undirected antonym links and directed hyponym links can be found in the noun term set; and undirected antonym links and undirected synonym links can be found in the adjective term set. WordNet renders much help in this work. Examples of term sets are shown in Fig. 2.

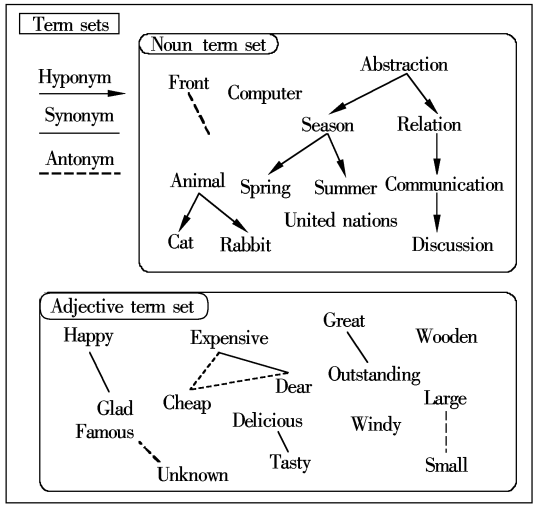

Fig. 2 Term set examples in the QnA knowledge presentation platform

These semantic links can help in extracting new knowledge from existing knowledge. For example, it is known that "rabbits eat vegetables" and the term "vegetable" has a hyponym link to the term "carrot". Then it can be deduced that "rabbits eat carrots". It may not be the truth, while it can be a truth in high probability.

4.2 Statement model—using grammatical information

The statement set, composed of statement models, is the most important part of the QnA knowledge presentation platform. It is the carrier of QnA knowledge. In QnA threads, there are all types of natural language sentences, with active or passive voice, with natural or inverted order. For knowledge presentation, all the sentences should be converted into a unified form. Thus a sentence model called the statement is defined, which all sentences of question/answer (Q/A) pairs can be converted into. A statement looks like a declarative clause with active voice and natural order, with elements of subject, predicate, object, attributive and adverbial. But it is a knowledge model, not a natural language sentence. Possible structures of a statement are shown in Tab. 1.

Tab. 1 Possible structures of a statement

Subject part	Predicate part		
Subject	Factitive verb	Object	
		Direct object	Indirect object
	Factitive verb	Object	
		Direct object	Indirect object
	Link verb	Predicative	

Each statement element is a term, a paragraph, or another statement. If it is a term, the element will be indexed and associated to the same term in term sets. There are several steps from Q/A pairs to statements, including spelling mistake correction, POS tagging, stemming, term generation, stop words removing, sentence pattern conversion, etc. Thanks to various NLP tools, this work can be done more facilely now. For example, the Minipar tool produces a quite detailed grammatical analysis of natural language sentences, and its output can be used as an intermediate result for the conversion to a statement. Fig. 3 shows some conversion examples. The definition of the POS tag in the figure can be found in Brown Corpus.

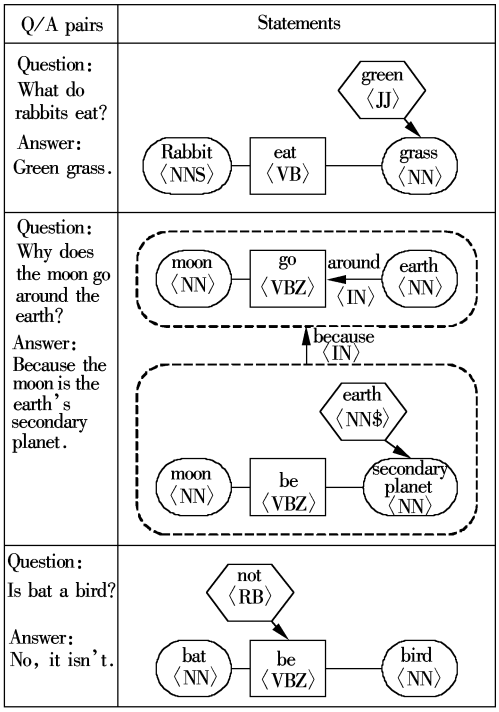

Fig. 3 Conversion examples from Q/A pairs to statements

Compared with statistical IR methods, the statement model can benefit greatly in the answer extraction. For example, if a Q/A pair "What do rabbits eat?/Green grass." is converted into a statement "rabbits eat green grass", and then both "What do rabbits eat?" and "Who eats green grass?" can be answered exactly.

4.3 Knowledge confidence

Along with semantic and grammatical information, knowledge confidence also plays an important role in QnA threads. As is known, not all the contents of QnA threads are valid. Some opinions may be disagreed with by other people, or they may even be wrong. So a confidence coefficient called statement confidence is defined, which is a decimal fraction in the interval of [0, 1], expressing how confident a certain statement is. It is mainly calculated by the voting count of the source answer and is added to each statement as an extended feature.

Then the confidence of semantic links can be defined, also a decimal fraction in the interval of [0, 1], expressing how relational a pair of terms are. With semantic link confidence, statement confidence can be recalculated and forwarded from existing knowledge to the new knowledge. A simple way to re-calculate new statement confidence is to multiply existing statement confidence by semantic link confidence.

The knowledge confidence in the form of statement confidence and semantic link confidence can be used by high-level QnA applications. For example, in QA application, the knowledge confidence is treated as part of a query result, and exercises a great effect on the final rank order.

Furthermore, there is a user set in the QnA knowledge presentation platform recording the information of statement users. The statement user is the user in original QnA threads, who has offered the source answer of a certain statement. So a user confidence can also be defined from the confidence of all statements related to the user. All the information of statement users can help generate experts in expert search applications.

5 QnA Applications

Many knowledge-based applications, such as QA, automatic encyclopedia and expert search, can be built upon the QnA knowledge presentation platform. Here are two typical ones.

5.1 Question answering

A QA application can be built upon the QnA knowledge presentation platform. It allows users to submit a specific question, and returns a ranked list of answers to the question. To implement the query process from a QnA knowledge model, the QA application should include components of question analyzer, statement query processor, statement ranker, and answer generator. Especially, the result answers are ranked not only by how much the question statement and answer statements match, but also by the knowledge confidence of answers. The architecture of QA application is shown in Fig. 4.

5.2 Automatic encyclopedia

An automatic encyclopedia application can be built as well upon the QnA knowledge presentation platform. An automatic encyclopedia is a web site like Wikipedia or Baidupedia. While it is not manually

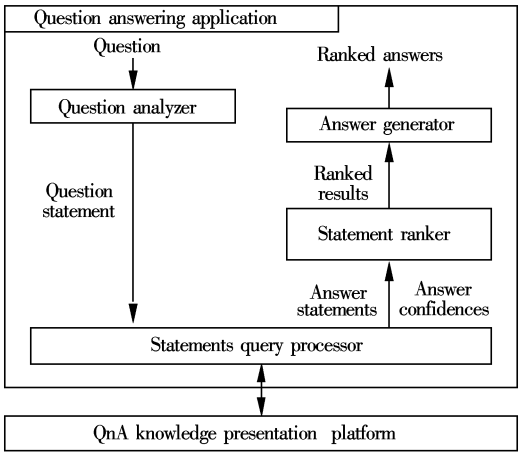

Fig. 4 Architecture of QA application

built, it is automatically generated from QnA knowledge. In brief, each term in a noun term set can be treated as a lemma in an automatic encyclopedia. Then for each lemma, a web page is generated from all the statements related to it.

6 Conclusion

QnA web forum is an important type of Internet service. Many valuable applications can be built upon forum data sets. To put more QnA knowledge in use, a knowledge presentation model is proposed in this paper. It is a new way to present knowledge in QnA threads, paying more attention to grammatical and semantic information, and even knowledge confidence.

As is discussed in this paper, the platform built with the QnA knowledge presentation model has the ability to provide a more exact presentation for knowledge in QnA threads. So many knowledge-based applications can be built upon this platform in the future, for better performance.

References

[1] Harabagiu S, Moldovan D, Clark C, et al. Employing two question answering systems in TREC 2005 [C]//*Proc of the* 14*th Text REtrieval Conference*. Gaithersburg, Maryland, 2005.

[2] Balog K, Azzopardi L, de Rijke M. Formal models for expert finding in enterprise corpora [C]//*Proc of the* 29*th ACM SIGIR Conference*. Seattle, Washington, USA, 2006: 43 – 50.

[3] Jijkoun V, de Rijke M. Retrieving answers from frequently asked questions pages on the web [C]//*Proc of the* 14*th ACM CIKM Conference*. ACM Press, 2005: 76 – 83.

[4] Ramakrishnan G, Chakrabarti S, Paranjpe D, et al. Is question answering an acquired skill? [C]//*Proc of* 13*th International WWW Conference*. ACM Press, 2004: 111 – 120.

[5] Soricut R, Brill E. Automatic question answering using the web: beyond the factoid [J]. *Information Retrieval*, 2006, **9** (2): 191 – 206.

[6] Wu Youzheng, Zhao Jun, Duan Xiangyu, et al. Research on question answering and evaluation: a survey [J]. *Journal of Chinese Information Processing*, 2005, **19**(3): 1 – 13. (in Chinese)

[7] Wen Xu, Zhang Yu, Liu Ting, et al. Syntactic structure parsing based Chinese question classification [J]. *Journal of Chinese Information Processing*, 2006, **20**(2): 33 – 39. (in Chinese)

[8] Ka Kan Lo, Wai Lam. Using semantic relations with world knowledge for question answering [C]//*Proc of the* 15*th Text REtrieval Conference*. Gaithersburg, Maryland, 2006.

[9] Lawrence S. Context in web search [J]. *IEEE Data Engineering Bulletin*, 2000, **23**(3): 25 – 32.

[10] Sun R, Jiang J, Tan Y F, et al. Using syntactic and semantic relation analysis in question answering [C]//*Proc of the* 14*th Text REtrieval Conference*. Gaithersburg, Maryland, 2005.

问答网络论坛的知识表示模型

于士涛　袁晓洁　师建兴

(南开大学信息技术科学学院,天津 300071)

摘要:为精确描述问答网络论坛的主题信息,提出构建面向英语自然语言的问答知识表示模型,进而提出包括数据采集、平台搭建和应用设计在内的问答知识系统的完整解决方案. 模型借助先验词典和自然语言语法分析方法,将语义信息、语法信息和知识置信度引入信息检索技术,并以语句模型集合和带有语义链接的标引项集合的形式表现出来. 理论分析表明:语句模型能够突破原始问答模式限制,精确表达知识陈述,满足各种查询需求;标引项之间的语义链接则可辅助语句模型,在现有知识基础上推导出新知识. 模型同时应用了信息检索和自然语言处理特征,增强了知识表达能力. 诸多知识系统应用可在此模型上得到改善,提供更好的性能.

关键词:问答网络论坛;知识表示;语义链接;语句模型;知识置信度

中图分类号:TP391