Front. Electr. Electron. Eng. China 2007, 2(1): 23–28
DOI 10.1007/s11460-007-0004-0

RESEARCH ARTICLE

LIN Hongfei, YANG Zhihao, ZHAO Jing

Question-answering system based on concepts and statistics

Abstract Question-answering systems provide short answers with the use of available information. The implementation mechanism for a question answering system is presented in this paper and is based on concepts and statistics. The system determines the question and focuses on the answer types, making different conceptual expansions for different questions. It applies the latent semantic indexing (LSI) method to retrieve relevant passages. It uses matching algorithms to find a match between questions and sentences stored in a database. It also extracts answers from a frequently asked questions (FAQ) database by finding matching or similar sentences. The answering ability of the system has been improved with the use of LSI and FAQ. The question-answering system introduced in Chinese universities is a developed and proven system capable of precise results.

Keywords question-answering system, concept expansion, latent semantic analysis, similarity of sentence, passage match

1 Introduction

In 1950, Turing, the outstanding mathematician, presented the concept of "machine intelligence", and came out with an approach to determine if computers possess intelligence, known as the "Turing test". The idea behind the approach was to apply natural languages to gauge the intelligence of a computer, and is the earliest prototype of a question-answering system.

Along with the development of the Internet, question-answering systems have also come of age. The question-answering system based on natural languages was first developed in 1993—this was a model capable of answering geographical questions in the Masachuset Institute of Technology (MIT) information laboratory. This was followed by the START system [1] which organized related information into ternary operators of subject-relationship-object to answer questions. Next, Julian Kupiec [2] designed the MURAX system capable of answering general questions based on encyclopedias. It integrated linguistics knowledge with a statistical approach, to search for answers from encyclopedias by means of a Boolean model and a syntactic parser. Thereafter, the FAQ Finder [3] searched for answers from a FAQ database with a search engine in a vector model space.

The next great advancement in this field was the introduction of a question-answering system in the Text Retrieval Conference (TREC) in 1998. The evaluation model of TREC proved effective and easy to operate, evoking much interest among the academic circles. Today, more and more research groups are taking part in the testing of the TREC model, and the TREC question-answering (QA) Track in open mode promotes the development of QA systems across the globe [4].

QA systems also attracted great attention in China [5]. Some QA systems were developed with knowledge representation and reasoning, but these were small in scale as they were limited by the range of the knowledge base and the rules to be acquired. Other QA systems tried to adopt the statistical approach, and were based on the "similar" relationship between questions and text segments. Thereafter, several question-answering systems started employing the syntactic analysis and the semantic analysis to determine the question types and the answer types [6,7].

In this paper, we present the implementation mechanism of a QA system which is based on concepts and statistics. The main idea of the QA system is to first determine the question focuses and the answer classes, and then make a different conceptual expansion to a different question focus. It uses the latent semantic indexing (LSI) to retrieve passages related to the text segment. It also employs 'match algorithms' to find a match between the questions and the sentences, and extracts answers from the FAQ database using the sentence similarity match function. Its answering capability is enhanced by the use of methods such as LSI and FAQ. The QA system introduced in Chinese universities is no doubt a proven product capable of precise results.

Translated from *Journal of Dalian University of Technology*, 2006, 46(2): 280–285 [译自: 大连理工大学学报]

LIN Hongfei (✉), YANG Zhihao, ZHAO Jing
Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024, China
E-mail: hflin@dlut.edu.cn

2 Question analysis

2.1 Question analysis and answer types

The primary and most important task for the QA system is to analyze the focuses of questions—focuses are information relevant to the questions, such as field-type, entity, entity-relation and entity-attribute. These objects stated and described in text, nouns, and noun phrases, are considered as action agents or objects, and are generally focuses of interest to both authors and readers as described by the theory of "attention focus set", while the attention tends to skim over the rest of the language components.

The first task in answering a question is to perform a comprehensive analysis of the question so as to determine the question type. With the help of the Chinese language-segmentation and part-of-speech tagging, the statistical analysis of words appearing in the question is carried out. The extraction rules for question types and answer types are as follows (see Table 1).

2.2 Conceptual expansion at level

After the question type and the key words are determined, the next task is the expansion of the key words [8]. The expansion range is to be strictly limited to avoid importing "noise" or making the meaning too general or too abstract. So the 'expansion at level' has been adopted to control the level of concept expansion.

The mechanism of 'expansion at level' is built on the conceptual dictionary. It is a semantic tree with a hierarchical structure, and the hierarchy of concept is linked to the degree of abstraction.

First level expansion: synonymous expansion to ensure the primary focus is expanded in a strict range and avoids misleading words.

$S^{(1)} = \{s|\ ConceptCode(s) = ConceptCode(t)\}$

Second level expansion: approximate expansion to expand the secondary focus to the same concepts or to the upper concepts.

$S^{(2)} = \{s|\ (ConceptCode(s) = ConceptCode(t)\ )$ OR $(s = ConceptFatherCode(t))$

Third level expansion: additional expansion to expand the additional focus to the same, upper and lower concepts.

$S^{(3)} = \{s|\ ConceptCode(s) = ConceptCode(t)\ )$ OR $(s = ConceptFatherCode(t))$ OR $(ConceptFatherCode(s) = t)\}$

After the question analysis and the conceptual expansion, the question is represented as follows

$AnswerType = \{ENT, ORG, TIM, LOC, NUM, DEF, MUL, ETC\}$

$Question = \{FieldType, AnswerType, FeatureVector\}$

$FeatureVector = \{NP,\ VP,\ TS\} = \{\{NP_1\},\ldots,\ \{NP_p\};\ \{VP_1\},\ldots,\ \{VP_q\};\ \{TS\}\}$

$$NP_i = \{NP_0^i, NP_1^i, ..., NP_{L_i}^i\},\quad i = 1,\ 2, ..., p$$

$$VP_i = \{VP_0^i, VP_1^i, ..., VP_{K_i}^i\},\quad i = 1,\ 2, ..., q$$

$$TS = \{TS_1, TS_2, ..., TS_r\}$$

where NP_i is the expansion set of primary focus, $i = 1, 2,\ldots, p$; NP_0^i is the primary focus; NP_j^i $(j > 0)$is the expansion keyword; VP_i is the expansion set of secondary focus, $i = 1, 2,\ldots, q$; VP_0^i is the secondary focus; $VP_j^i (j > 0)$ is the expansion keyword; TS is the additional focus set.

Each focus in the question is endowed with a weight to indicate its importance and priority in the matching of

Table 1 Question type and answer type

Type(code)	Interrogative words	Question describe and answer type	Examples
Person name (ENT)	Who, whom	Focus + N or NP Focus + …, + N Focus + be + N N + be + Focus Answer-> nr (person name)	Who is the president of Peking University?
Entity (ORG)	Which	NP + V + Focus, Focus + In + NP Answer-> nt or nz (organization name)	Which university has the department of computer?
Date time (TIM)	When, what time, what date, what day	Answer-> t (time)	When did Shandong University established?
Location (LOC)	Where	Answer-> s or ns (location, region ,orientation)	Where was Southeastern University?
Number (NUM)	How much	Answer-> m (numeral)	How many students does Jilin University have?
Definition (DEF)	What	Focus-> be/named/defined/called + … Focus-> … + be/named/defined/called Answer-> n	What is chemistry?
Multi-value (MUL)	Which	1.Focus-> "nation/province/city/district" Answer type -> ns; 2.Focus-> "organization/school/company" Answer type -> nt; 3.Focus-> "presiden/teacher/supervisor/leader" Answer type -> nr	Which university's president is Cheng Gengdong?
Other(ETC)	How,what	Answer-> no limitation	How about the reputation of Dalian University of Technology?

the question and the text. The weight's calculation is as follows

$$
\begin{aligned}
&g(t)=1,\ t\in\{TS\}\\
&g(t)=1,\ t\in NP_i,\ i=1,\ 2,...,p\\
&g(t)=0.75,\ t\in\{VP_i\}\text{AND}(ConceptCode(t)\\
&\quad=ConceptCode(VP_0^i)),\ i=1,\ 2,...,q\\
&g(t)=0.5,\ t\in\{VP_i\}\text{AND}(ConceptCode(t)\\
&\quad=ConceptFatherCode(VP_0^i)),\ i=1,\ 2,...,q
\end{aligned}
$$

The question is denoted by the feature vector $\boldsymbol{Q}=(<t_1, w_1>, <t_2, w_2>,\ldots, <t_n, w_n>)$, where t_i is the keyword, w_i is its weight, $i=1, 2,\ldots, n$.

The expression indicates the importance of the primary focuses and the secondary focuses by means of their weights in the vectors.

3 Passage match based on latent semantic analysis

The focuses of the question and the feature vector are obtained after performing the question analysis. The next task is to search for the related text—the accuracy and the speed of the QA system depends on this feature.

The passages are broken into units which can be retrieved independently. However, the text contains many anaphora relationships and other associations. The simple partitions result in destroying these relationships or even losing them altogether. Each passage is assigned text attributes as additional information, thus making it into a self-governed and meaningful text unit. Reducing the dimensions of the units to be retrieved quickens the answer extraction process.

The physical structure consists of a series of separable text units, denoted by the expressions (title, passage, sentence, word). The passages are fundamental units to describe the text theme, hence they are utilized to express the text theme and field type with attached information from the text.

$$
\begin{aligned}
Text&=\{Title, Passage, Sentence, Keyword\}\\
&\quad+TextTheme+FieldType\\
Passage&=\{Sentence, Keyword\}\\
&\quad+TextTheme+FieldType\\
Sentence&=\{Keyword\}+TextTheme\\
&\quad+FieldType+AnswerType
\end{aligned}
$$

The texts are transferred into passages by the method described, and then the matching function is applied to extract the target passages.

LSI replaces the original word space with K dimensions orthogonal subspace [9], namely, the latent semantic space, and provides higher accuracy in the similarity computation when matching word for word and passage for passage with richer semantic information.

In the passage matching process, each passage is regarded as a text, thus the term-document matrix $\boldsymbol{A}_0=(a_{ij})$ is built as follows

$$
a_{ij}=\frac{tf_{ij}\log\ (1+\frac{N}{n_j})}{\sqrt{\sum_{k=1}^{n}(tf_{ik})^2[\log\ (1+\frac{N}{n_k})]^2}}
$$
$$
i=1,\ 2,...,\ m;\ j=1,\ 2,...,\ n
$$

where tf_{ij} denotes the frequency of keyword t_i in passage P_j, n_j denotes passage frequency of keyword t_i in texts, and N denotes the number of passages in texts.

It is the transmutation of the well-known formula $tf*idf$, and has been normalized, considering the different lengths of passages.

LSI is derived from the matrix decomposition SVD. There exists a matrix decomposition to every $t\times d$ matrix with rank r. $\boldsymbol{A}_0=\boldsymbol{T}_0\boldsymbol{S}_0\boldsymbol{D}_0^{\mathrm{T}}$, where $\boldsymbol{A}_0$ is the original matrix; all column vectors of $\boldsymbol{T}_0$ and $\boldsymbol{D}_0$ are orthogonal vectors, that is, $\boldsymbol{T}_0^{\mathrm{T}}\boldsymbol{T}_0=I$, $\boldsymbol{D}_0^{\mathrm{T}}\boldsymbol{D}_0=\boldsymbol{I}$; $\boldsymbol{S}_0=\mathrm{diag}(\lambda_1, \lambda_2,\ldots, \lambda_r)$, $\lambda_i(i=1, 2,\ldots, r)$ is the latent root of the matrix. $\boldsymbol{A}=\boldsymbol{TSD}^{\mathrm{T}}$, the new matrix $\boldsymbol{A}$ has been generated from the original matrix according to the selected dimension K to approximate the original matrix.

The feature vectors of the example texts are represented by $\boldsymbol{P}_1, \boldsymbol{P}_2,\ldots, \boldsymbol{P}_n$. The feature vector of the question is Q. $\boldsymbol{P}_1, \boldsymbol{P}_2,\ldots, \boldsymbol{P}_n$, and Q are converted into the feature vector of conceptual space with transformation Ω, namely, $(P_1^{(K)}, P_2^{(K)},..., P_n^{(K)})=\Omega(P_1, P_2,..., P_n)$ and $Q^{(K)}=\Omega(Q)$. The two vectors are similar as they are both based on the cosine formula. The output list of the passages is sorted by similarity from high to low, and the passages are extracted as target passages with the threshold as η and the similarity $\theta\geqslant\eta$.

4 Answer extractions

4.1 Match algorithm and threshold selection

The requirement here is to extract accurate answers from the target passages.

The primary task after the partitioning of the text passages is to describe the degree of matching between the text passages and the question. Factors to be considered are keyword-density, limitation of question-type and the constrained condition of focuses. In addition, sentences need to pick up the field-type, text-theme, passage-topic and other related attributes from the text and passage, which help determine their ability to answer the question.

Field-type. It is an important attribute for passage extraction and is a necessary condition for answer search. It is used to filter a large portion of unrelated passages using the Boolean model.

Answer-type. the expectative answer-type is acquired after analysis of the question. Answer-type is classified into 8

types: ENT, ORG, TIM, LOC, NUM, DEF, MUL and ETC. The answer must have the specified answer-type information. Uncertainty-type is used to extract related passages without imitating answer-type.

Factors to be considered are the frequencies of various focuses in passages and the weights of the focuses when the similarities between passages and questions are calculated. As mentioned above, the question sentence is also taken as a passage. Rarely are there repetitive words in one sentence, so each word mostly appears once in a sentence. Therefore, the density of focus and weight of focus are important factors to be considered in the similarity computation. The similarity formulae are as following.

Similarity of primary focus

$$\mathrm{Sim}_1(Q,\boldsymbol{S})=\exp\left(\frac{1}{|Q|}\sum_{t\in Q_1\cap \boldsymbol{S}} g(t)-1\right)$$

Similarity of secondary focus

$$\mathrm{Sim}_2(Q,\boldsymbol{S})=\exp\left(\frac{1}{|Q|}\sum_{t\in Q_2\cap \boldsymbol{S}} g(t)-1\right)$$

Similarity of additional focus

$$\mathrm{Sim}_3(Q,\boldsymbol{S})=\exp\left(\frac{1}{|Q|}\sum_{t\in Q_3\cap \boldsymbol{S}} g(t)-1\right)$$

Similarity between question and passage

$\mathrm{Sim}(Q,\boldsymbol{S})=\alpha\mathrm{Sim}_1(Q,\boldsymbol{S})+\beta\mathrm{Sim}_2(Q,\boldsymbol{S})+\gamma\mathrm{Sim}_3(Q,\boldsymbol{S})$ where, $|Q|$ denotes the length of question, that is, the number of words in a question, $\alpha=0.6/(\mathrm{e}-1)$, $\beta=0.3/(\mathrm{e}-1)$, $\gamma=0.1/(\mathrm{e}-1)$.

Related passages are then extracted from the selected threshold based on the similarities between questions and passages. If required, the matching degree of focus is to be controlled with the constrain-condition of the primary focus and the secondary focus; for example, constrain-condition can be defined as

$\mathrm{Sim}(Q,\boldsymbol{S})>\lambda$ AND $\mathrm{Sim}_1(Q,\boldsymbol{S})*\mathrm{Sim}_2(Q,\boldsymbol{S})>0$ where λ is the threshold of extraction.

4.2 Answer extraction based on question similarity

Questions are often repetitive in nature, that is to say, new questions asked by an individual are frequently very similar to earlier questions, although not identical. Consequently, it is worthwhile to seek the help of the FAQ database to answer new questions. The FAQ database maintains records of frequent answers to questions.

The most important task for the FAQ database is to compare new questions with old questions, and then let the QA system select the answer linked to the same question or to a similar question form the FAQ database. Therefore, the task can be reduced to similarity matching of two sentences [10].

The sentence similarity search consists of three kinds of similarity searches—the syntactic similarity search, the semantic similarity search, and the pragmatic similarity search. It has to be pointed out that the computation of the pragmatic similarity search is comparatively complicated and often provides unsatisfactory results. Consequently, the sentence similarity search depends largely on the syntactic similarity search and the semantic similarity search. The computational approach for the similarity search has to rely on morphologic matching, word-order matching, and conceptual expansion without the support of the Chinese parsers.

1) Morphologic similarity is defined as

$$\mathrm{Sim}_{\mathrm{form}}(\boldsymbol{S}_1,\boldsymbol{S}_2)=\lambda\frac{|\boldsymbol{S}_1\cap\boldsymbol{S}_2|}{|\boldsymbol{S}_1\cup\boldsymbol{S}_2|}+(1-\lambda)\frac{|\boldsymbol{S}_1^*\cap\boldsymbol{S}_2^*|}{|\boldsymbol{S}_1^*\cup\boldsymbol{S}_2^*|}$$

where $\boldsymbol{S}_1$ is the key-word of sentence one, $\boldsymbol{S}_2$ is the key-word of sentence two, $\boldsymbol{S}_1^*$ is the synonymy of $\boldsymbol{S}_1$, $\boldsymbol{S}_2^*$ is the synonymy of $\boldsymbol{S}_2$, and $|\boldsymbol{S}_1\cup\boldsymbol{S}_2|$ is the number of union sets of $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$.

2) Word-order similarity is defined as

$$\mathrm{Sim}_{\mathrm{order}}(\boldsymbol{S}_1,\boldsymbol{S}_2)=1-\frac{\mathrm{Re\,vorder}(\boldsymbol{S}_1,\boldsymbol{S}_2)}{|\boldsymbol{S}_1\cup\boldsymbol{S}_2|}$$

where Re vorder($\boldsymbol{S}_1$, $\boldsymbol{S}_2$) is the number of reverse orders. The computational rule followed by word-order similarity is to take the word order occurred in sentence one as a standard and compares the word order occurred in sentence two with that of sentence one. Re vorder($\boldsymbol{S}_1$, $\boldsymbol{S}_2$) is the sum of times there was inconsistency in the word order. Notably the revorder includes comparison of synonymies, and the order of synonymy of one word in the sentence is same as its order.

3) General similarity is defined as

$$\mathrm{Sim}_{\mathrm{vec}}(\boldsymbol{S}_1,\boldsymbol{S}_2)=\frac{\boldsymbol{S}_1^{\mathrm{T}}\boldsymbol{S}_2}{\|\boldsymbol{S}_1\|\|\boldsymbol{S}_2\|}$$

where $\boldsymbol{S}_1$ is the feature vector of sentence one, $\boldsymbol{S}_2$ is the feature vector of sentence two, $\|\boldsymbol{S}_1\|$ is the length of the feature vector $\boldsymbol{S}_1$, and $\|\boldsymbol{S}_2\|$ is the length of the feature vector $\boldsymbol{S}_2$. General similarity indicates the similarity of sentences in general. It is based on the co-occurrence of key words and is independent of the word order.

4) Integrated similarity is defined as

$$\mathrm{Sim}(\boldsymbol{S}_1,\boldsymbol{S}_2)=\alpha\mathrm{Sim}_{\mathrm{form}}(\boldsymbol{S}_1,\boldsymbol{S}_2)+\beta\mathrm{Sim}_{\mathrm{order}}(\boldsymbol{S}_1,\boldsymbol{S}_2)+\gamma\mathrm{Sim}_{\mathrm{vec}}(\boldsymbol{S}_1,\boldsymbol{S}_2)$$

Based on the results of the integrated similarity search, an appropriate threshold is selected, and the questions are extracted in an orderly manner from the FAQ database. The answers from the FAQ database corresponding to the selected questions are then recommended as the answers to the new questions.

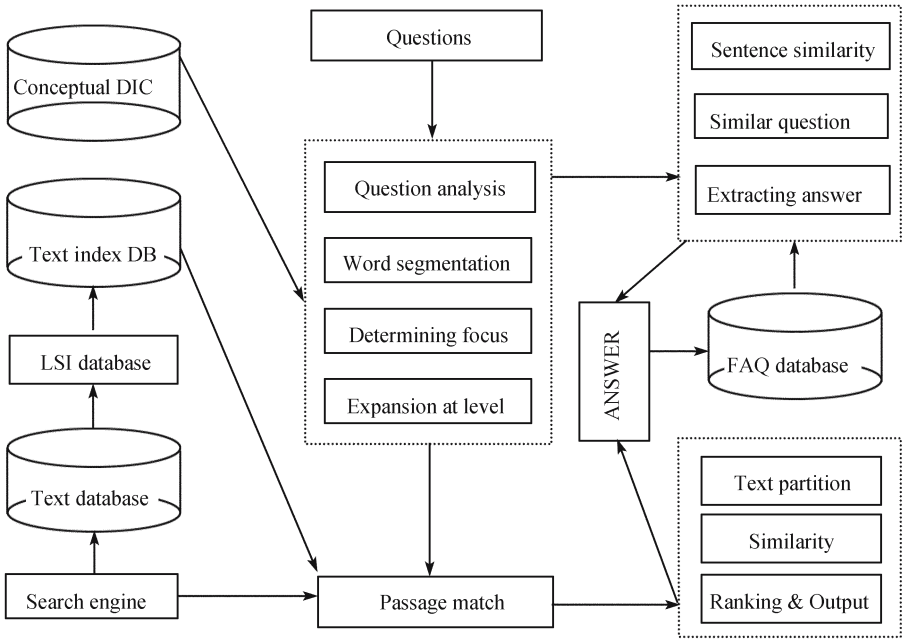

Fig. 1 Flowchart of QA system on universities

5 System flowchart and implementation mechanism

The experimental corpus is a collection of search engines, and consists of general information on about 400 Chinese universities. The relevant text database is built with the use of Chinese word-segmentation and by removing the stop lists, and is indexed by the LSI. In this paper, we select the dimension of the latent semantic space as 100, and then we complete the LSI database. The FAQ database is used to save 119 of the most frequent question-answer records, and is divided into 7 question categories and 23 types of question sentences. All FAQ records are verified and organized by university name and question category. A flowchart of the QA system on universities is shown in Fig. 1.

In consultation with the university entrance examination board, examiners have come up with 360 questions, and some 297 questions are taken as a training set.

The QA system provides 5 different answers, and it endows each answer with a score computed by the formula of TREC QA Track. One point is awarded if the first answer is correct, and half, if the second answer is correct. The rest can be deduced by analogy. A zero is awarded if all given answers are incorrect. Lower the number of correct answers, higher is the score. The accuracy of the QA system can be measured by the sum of the scores; hence the measure is named the mean reciprocal answer (MRAR).

The formula for MRAR is represented by

$$\mathrm{MRAR} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\mathrm{RANK}_i}$$

where $\mathrm{RANK}_i \in \{1, 2, 3, 4, 5\}$ is the number of correct answers.

In addition, the feedback is correct if the 5 answers provided by the QA system include the correct answer, and CF is equal to one if the feedback is correct (see Table 2). As a result, the accuracy of the QA system is measured by the ratio of the correct feedback in simple percentage mode to the validity of the feedback, and the accuracy percentage is calculated by the following formula

$$\mathrm{AP} = \frac{\sum_{i=1}^{n}\mathrm{CF}_i}{n}, i = 1, 2, \ldots$$

where n denotes the question number.

Table 2 Experimental results

No.	Type	Num. questions	I	II	III	IV	V	MRAR	AP/%
1	Name	29	11	6	3	1	1	0.532 7	75.86
2	Entity	76	20	15	11	5	3	0.434 4	71.05
3	Number	61	9	11	11	5	8	0.344 5	72.13
4	Date	41	6	5	8	4	5	0.321 1	68.29
5	Definition	32	6	2	7	3	2	0.327 6	62.50
6	Multi-value	37	4	5	7	1	3	0.261 7	54.05
7	Other	21	4	2	5	1	1	0.338 9	61.90
	Total	297	60	46	52	20	23	0.370 1	67.68

6 Conclusions

The QA system based on concept and statistics is not dependant on domain knowledge and the Chinese parser. The system extracts target passages by passage retrieval, which is based on statistics, and relies on the similarities of questions

and the FAQ records to find answers to new questions. It assigns attributes such as domain-features, question-types and the text-themes to passages, so as to reduce the negative influences of the anaphora resolution. The adaptability and accuracy though, can be improved with the expansion of the related corpus, and increase in the FAQ records.

The disadvantage of the system is that it is short on grammar support and semantic analysis. Although the statistical approach is easier to implement, it fails to well refine questions—for example, problems of privative words and complicated anaphora resolution, still persist. The important task for the future is to fuse the shallow parsing technology and the passage retrieval method based on statistics to improve the efficiency and accuracy of the QA system.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 60373095).

References

1. Katz B. From sentence processing to information access on the world wide web. In; Proceedings on Natural Language Processing for the World Wide Web, AAAI Spring Symposium. California: AAAI Press, 1997, 77–94
2. Kupiec J. MURAX: a robust linguistic approach for question answering using an online encyclopedia. In: Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1993: 181–190
3. Robin B, Kristian H, Vladimir K, et al. Question answering from frequently-asked question files: experiences with the FAQ Finder system. Technical Report TR-97-05. Chicago: University of Chicago, Department of Computer Science, 1997
4. Wu Lide, Huang Xuanjing, et al. FDU at TREC-9: CLIR, Filtering and QA tasks. In: Proceedings of the 9th Text Retrieval Conference. Maryland: NIST Press, 2001
5. Zheng Shifu, Liu Ting, Qin Bing. Overview of question-answering. Journal of Chinese Information Processing, 2002, 16(6): 46–53 (in Chinese)
6. Wang Shuxi, Liu Qun, Bai Shuo, et al. The research on QA system based on dynamic KB. In: Proceedings of the 7th National Joint Conference on Linguistics. Beijing: Tsinghua Press, 2003, 587–592
7. Zhang Gang, Liu Ting, Zheng Shifu, et al. Design and implementation of open domain Chinese question-answering system. In: Proceedings of the 20 Anniversary of Association of Chinese Information Process. Beijing: Tsinghua Press, 2001, 231–235 (in Chinese)
8. Lin Hongfei, Zhan Xuegang, Yao Tianshun. Text structure analysis based on concept. Computer Research and Development, 2000, 37(3): 324–328 (in Chinese)
9. Berry M W, Dumais S, O'brien G W. Using linear algebra for intelligent information retrieval. SIAM Review, 1995, 37(4): 573–595
10. Lu Xueqiang, Ren Feiliang, Huang Zhidan, et al. Sentence similarity model and the most similar sentence search algorithm. Journal of Northeastern University (Natural Science), 2003, 24(6): 531–534 (in Chinese)