A Novel Similarity Calculation Method Based on Chinese Sentence Keyword Weight★

Yangxin Yu
Faculty of Computer Engineering,Huaiyin Institute of Technology, Huai'an 223003 China
Email: hyyyx@hyit.edu.cn

Liuyang Wang
Faculty of Computer Engineering,Huaiyin Institute of Technology, Huai'an 223003 China
Email: wangly@hyit.edu.cn

Abstract—**Question Answering system (*QA*), is a kind of new information retrieval system which can be queried with natural language and return knowledge directly.By improving the traditional Chinese sentence similarity model, this paper proposes a Chinese sentence similarity calculation method based on keywords weight.Its key question is that questions asked by user and questions in the Frequently Asked Question (*FAQ*) carry on similarity calculation, discover the closest question in the *FAQ* and return the question answer stored in advance. It can also automatically update and maintain *FAQ*.The experiment indicates that precision of question match can be improved compared with traditional sentence similarity model.**

Index Terms—**Chinese sentence similarity, Keywords weight, Similarity calculation**

I. Introduction

With the rapid development of Internet, there is more and more information in Web, and the search engine development has been largely convenient for users to query information. The growing network information makes users difficult to quickly find their required information from large amounts of returned information of the search engine[1].The users may have diverse backgrounds and different expectations for a given query, some search engines try to personalize their results in order to better match the overall interests of an individual user[2].Therefore, People put forward higher requirements for network information retrieval, and want to get the information they need more rapid, accurate and detailed by searching. *QA* system is developed in order to meet these desires of people. It has provided people with the means of communication to ask questions in natural language, and directly return the answers users need rather than the relevant pages with a convenient, fast and efficient[3].

Currently the *QA* system has become very popular research domain, including the English *QA* system, Chinese *QA* system in recent years[4].There are many studies on English *QA* systems. Due to the Chinese characteristics, *QA* system in Chinese has some characters, sometimes it is completely different from English and other languages of *QA* system.Natural Language Processing (*NLP*) is the key technology in *QA* system.To improve the *QA* system, *NLP* technology, especially the semantics analysis of questions and answers, need to be improved[5]. But at present, the natural language semantics analysis technique is still in the primary stage. So, most of *QA* systems were not involved the semantic analysis or were only based on shallow studies of the semantic analysis.How to improve the semantic level of understanding in the *QA* systems should be the key to improve the level of *QA* system.

As natural language processing application, natural language *QA* system is still in its infancy. Although many foreign research institutions and companies have achieved some results on natural language *QA* system due to combining natural language processing, knowledge representation and information retrieval on the integration, *QA* system development is highly dependent on advances in these areas. Currently its research is still very limited.

The domestic research on the Chinese *QA* system is less, because the requirements of some areas for the Chinese *QA* research are higher than the natural language *QA*[6]. Compared with natural language, Chinese syntactic analysis and semantic understanding are more difficult, and cause the slow development of the Chinese *QA* system.

The *QA* system of *FAQ* is a good model of *QA* realization. It combines *FAQ* and related answers on the integration. When users propose questions, the system can quickly give the answer according to the corresponding relations of the *FAQ*, and provides a more convenient quick way of eliminating doubt with a strong practical value without the complex process of information retrieval and answer extraction.

★Project supported by the Social Science Foundation of the Colleges and Universities of Jiangsu Province (No.2012SJD870001) and the Science and Technology Research Project of Huai'an (No.HAG2012055).

∗ Corresponding author: hyyyx@hyit.edu.cn (Yangxin Yu).

doi:10.4304/jsw.9.5.1151-1156

Sentence similarity calculation is an important theoretical basis of the automatic *QA* system and key implementation technology. Applied it to the *FAQ*, the research accuracy will be greatly improved[7]. This paper presents a novel similarity calculation method based on keyword weight, and the application of the method is given in the *QA* system.

II. Common Sentence Similarity Calculation

The semantic meaning of a phrase or a sentence mainly consists of two parts: the meaning of components of the phrase or the sentence, and that of the structure. Research on sentence types is very important for the linguistics in the syntax structure level[8]. Sentences of a language are infinite, but the sentence types are finite. Through the study of the finite sentence types to grasp the infinite sentences is the main goal of sentence types research. For *QA* system, the sentcence types of interrogative sentences are more closely relationless with the interrogative semantic meanings. By analyzing the sentence types of interrogative sentences, the questions can be accurately understanded. Generally, the sentence similarity calculation is divided into three levels: syntax similarity, semantic similarity and pragmatic similarity. Pragmatic similarity is rather difficult, and the result is not ideal. This paper only uses the syntax similarity and semantic similarity.

A. Syntax Similarity

Syntax similarity is based on an integrated approach of word formation, sentence length and word sequence of words. It thinks sentence similarity is determined by the similarity of the word formation, and sentence length similarity and word sequence of the three factors. The word formation similarity plays a major role, sentence length similarity and word sequence similarity playe a secondary role[9]. The calculation process is as follows:

Supposed, the length of the sentence *X* is the number of words in the sentence , denoted by *Len(X)*. *SameWC(A,B)* expresses the number of the same word occurrencing in sentences *A* and *B*. If the number of occurrences for word is not the same, we will count the less number of occurrences.The word formation similarity of sentence *A* and *B* is calculated as follows:

$$WordSame(A,B) = 2 \times Same(A,B)/(Len(A)+Len(B)) \tag{1}$$

The sentence similarity marked with sentence length also reflects the similarities of sentence formation in a certain extent, the length of the sentence *X* is the number of words in the sentence *X*, denoted by *Len(X)*.The similarity of sentence *A* and *B* is calculated as follows:

$$LenSim(A,B) = 1 - |Len(A)-Len(B)/(Len(A)+Len(B))| \tag{2}$$

Oncews(A,B) expresses a word occurring one time in sentence *A* and *B* at the same time, *Pfirst(A,B)* expresses the composed vector of the word location serial number of *Oncews(A,B)* occurring in sentence *A*, *Psecond(A,B)* expresses the generated vector by the component in the *Pfirst(A,B)* according to the sorted *sequence* of the corresponding word in sentence *B*, *RevOrd(A,B)* expresses the reverse number of adjacent components of the *Psecond(A,B)*.The similarity of sentence *A* and *B* is calculated as follows:

$$Oncews = \begin{cases} 1 - \dfrac{\mathrm{Re}\,vOrd(A,B)}{|Oncews(A,B)|-1} & if\ |Oncews(A,B)|>1 \\ 1 & if\ |Oncews(A,B)|=1 \\ 0 & if\ |Oncews(A,B)|<1 \end{cases} \tag{3}$$

The similarity of sentence *A* and *B* is the similarity weight sum of word form, sentence length and word sequence. The advantage of this method is a comprehensive consideration of the sentence structure and the number of same words impacting on the similarity. Algorithm is simple and low complexity. However, this method only considers the shape matching based on the words between words without considering the semantic information and distinguishing the different effects for sentence between the different part speech words.Therefore,it often appears the irrational phenomenon that the similarity of the semantic similar sentence is lower.

B. Semantic Similarity

There are many calculation methods of semantic similarity model.This paper uses the sentence similarity calculation method based on the semantic dictionary[10]. Semantic dictionary method mainly uses knowledge net, synonymous with the word forest and other existing more mature semantic resources to calculate the questions similarity by the word similarity between the questions. Synonymous with the word forest is seen as the system knowledge resources in this paper.

Supposed, word similarity of any two sentences *A* and *B* can be calculated in some way. *A* contains the words A_1, A_2 , ..., A_m, *B* contains the word as B_1, B_2,... , B_n, the similarity between the word A_i $(1\le i\le m)$ and B_j $(1\le j \le n)$ is expressed as $S(A_i,B_j)$, and the semantic similarity *Sim(A,B)* between sentences *A* and *B* is calculated as follows:

$$Sim(A,B) = (|\sum_{i=1}^{m} a_i \Big/ m| + |\sum_{j=1}^{n} a_j \Big/ n|)/2 \tag{4}$$

Where, $\begin{matrix} a_i = \max(s(A_1,B_1), s(A_1,B_2), \cdots, s(A_1,B_m)) \\ b_j = \max(s(B_1,A_1), s(B_1,A_2), \cdots, s(B_1,A_n)) \end{matrix}$ .This method takes full account of the depth information of each word in the sentence while calculating the sentence similarity,so that the same deep meaning words with different surface are exhumed. However, the theoretical immaturity of semantic annotation and semantic dictionary are not comprehensive so as to calculate some errors[11]. This method selects the biggest similarity matching words to calculate the sentence similarity, and does not consider the structure of the sentence.Therefore, the accuracy is also difficult to achieve a satisfactory level.

III. Improved Sentence Similarity Calculation Method

A. Keyword Extraction

In the *FAQ* database of *QA* system, not all the words of each question play a role on the matching function after the statistics of large amounts data. Known by the linguistic knowledge, any sentences are constituted by key ingredient (*Subject*, *Predicate* and *Object*, etc.) and modifiers (*Attributive*, *Adverbial* and *Complement*,etc.). The key component of the sentence plays a major role,and the modification of the sentence components playes a secondary role[12]. It only considers the sentence key components in calculating the sentence similarity while calculating sentence similarity.

Under normal circumstances, *subject* and *object* of a sentence are a *noun* or *pronoun* while predicate is *verb* or *adjective*. Therefore, all *nouns*, *pronouns*, *verbs*,*adjectives* or *adverbs* in a sentence may take as keywords, and these keywords are only considered in the calculation of the sentence similarity.

A *noun*, *pronoun*,*verb*,*adjective* or *adverb* is not necessarily the *subject*, *object* or *predicate* composition of the sentence for particular sentence. Keyword sequence with a certain syntactic structure infomation is more important in the sentence structure composed of all words, the similarity calculation based on this basis is more accurate than generally based on the word.

B. Semantic Similarity Keyword Weight

The question is usually composed by a number of words, but the importance of each word is not the same.

In the course of practice, nouns and verbs in the sentence play a very important role. A sentence is basically around the center of *verbs* and *nouns* to expand. We have purposely increased the weight degree of the *verbs* and *nouns* during sentence similarity calculation, and the center of gravity of the sentence falls on the *nouns* and *verbs*. Weight *W* expresses this characteristic,and W_1, W_2,..., W_n respectively expresses the weight of the words x_1, x_2, ..., x_n. Therefore, Equation(1) is improved as follows:

$$WordSame1(A,B) = 2\times(\sum_{t=1}^{k} w_t)/(\sum_{i=1}^{m} w_i x_i + \sum_{j=1}^{n} w_j y_j) \quad (5)$$

Where, *k* expresses the same number of words in the sentene *A* and *B*, *m* and *n* respectively expresses the length of sentences *A* and *B*. Equation *(4)* is improved as follows[13]:

$$Sim1(A,B) = ((|\sum_{i=1}^{m} w_i a_i|/\sum_{i=1}^{m} w_i) + (|\sum_{j=1}^{n} w_j b_j|/\sum_{j=1}^{n} w_j))/2 \quad (6)$$

In this paper, the weight value of sentence similarity calculation is as follows: the *W* of *noun* class and *verb* class is *0.3,* the *W* of *adjectives* class and *adverbs* class is *0.2*.

C. Sentence Similarity Calculation of Keyword Weight

Sentence keyword extraction and different parts of speech (*POS*) word are given different weights by the above method, you can calculate the sentence similarity from the two aspects of word formation and word meaning.

Supposed, the compared two sentences is *S1* and *S2*, their similarity is denoted by *SentenceSim(S1, S2)*:

$$SentenceSim(S1,S2) = \lambda 1 Sim1(S1,S2) + \lambda 2 WordSim1(S1,S2) + \lambda 3 LenSim(S1,S2) + \lambda 4 Oncews(S1,S2) \quad (7)$$

Where, *Sim1(S1,S2)* is the improved word meaning similarity of *S1* and *S2,WordSim1(S1,S2)* is the improved word form similarity of *S1* and *S2*, *LenSim(S1,S2)* is the sentence length similarity of *S1* and *S2*, *Oncews (S1,S2)* is the word sequence similarity of *S1* and *S2*. In this paper, $\lambda 1 + \lambda 2 + \lambda 3 + \lambda 4 = 1$, $\lambda 1 > \lambda 2 > \lambda 3 > \lambda 4 > 0$, $\lambda 1 = 0.5$, $\lambda 2 = 0.3$, $\lambda 3 = 0.15$ and $\lambda 4 = 0.05$.

Compared with the original algorithm, the keyword extraction part of the algorithm involves word segmentation and *POS* tagging as well as the weight problem of the different *POS* (The original algorithm only involves word segmentation). The algorithm has three characteristics:Taking into account the different effects of different *POS* sentence.

The extracted keywords and the given weight can be approximated by some syntactic structure information. As computing the sentence similarity, considering the two levels similarity of the syntax and semantics, merging their strengths and overcoming their shortcomings, computing the similarity with high accuracy.

IV. Sentence Similarity In The FAQ

A. Automated QA System Processes Based on the FAQ

As Fig.1 shows that the general work processes of Chinese automatic *QA* is: Firstly,the *QA* gets the porposed questions by the users, and forms a combination of keywords after the segmentation module is processed. Secondly, analyzing the formed keywords to determine the type and focus of the problem.Thirdly, finding normally question base, and comparing the proposed questions by the users with the same sort of problems in normally question base.The system directly returns the answer to user when the similarity is greater than a threshold[14]. The problem matching can be seen as an application of the two sentence similarity calculation.

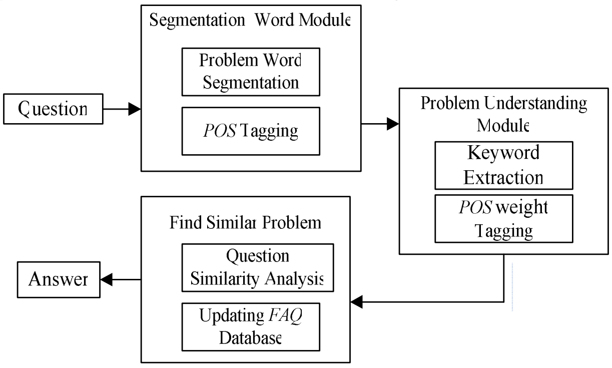

Fig.1. The general work processes of automatic *QA* system

B. Establishment of Candidate Question Set

The purpose of building candidate question set is to narrow the search range, more complex process of subsequent similarity calculations is executed within the relatively small range of candidate question set. In this paper, the words of the question sentence are taken as the basic unit so as to establish the candidate question set of the inverted index.

Supposed, there are *n* words in the entered question sentences by the user (Referred to as the target question sentence): W_1, W_2,…,W_n, and there are *m* question sentences in *FAQ* base while n_i words exists in the *ith* question sentence: *Q1, Q2 ,…,Qn*. The number of overlapping words between the *ith* question sentence and the target question sentence is denoted by Num_i. That is, $Num=\{W_1,W_2, \ldots,W_n\}\cap\{Q_1,Q_2,\ldots,Q_{ni}\}$, and selecting the largest top *50%* Num_i value of the *FAQ* question sentence to form the candidate question set.

C. Updating FAQ Database

There are new problems to be added, so new problems often are extended to *FAQ* database. It must judge whether the new entered question is the same and similar problems of the *FAQ* or not while new problems are extended[15].

Calculating the similarity between the entered target question sentence and each question sentence in the *FAQ* database by using the method of *No.3* section.If the maximum similarity is greater than a certain threshold value *m*, so that the maximum similarity corresponding the question sentence and the user's goal question sentence are taked as the same question, and directly output the corresponding answer of this question sentence to the user, otherwise this problem will be entered the *FAQ* database to be updated.

V. Experiment And Analysis

A. System evaluation criteria

Using the above method to calculate the similarity of each sentence and the sentence form similarity between the sentence, if the maximum similarity is larger than a predetermined threshold value σ, then the sentence with the most similarity is taken as the needed sentence. If maximum likelihood is smaller than σ, then the sentence is not the composition of sentence hierarchy.

Comprehensive test of the automatic *QA* system uses artificial question, determines the degree of accuracy at the system matching and estimates the reliability of the answers. In this paper, the threshold value is *0.7*. For the question matching of the similarity more than *0.7*, the retrieved answers may meet with the raised issues the correctrate is very strong.On the contrary, the correctrate of the raised issues is poor. The test result is calculated as:

$$\mathit{Correctrat\,e} = \mathit{The\ nuber\ of\ tested\ correct\ sentences/}\ \mathit{The\ total\ of\ tested\ sentences} \tag{8}$$

B. Similarity Calculation

As mentioned above, the composition of a sentence is composed by sentence phrases and characteristic word, and grammar of these ingredients is in different levels. Making sentences identified, we must separate the composition of the sentence to sentence as a template conduct statute, after the sentence corresponding vectors before you can be matched with the sentence vectors to determine whether the sentence is the sentence model. The method calculating the similarity between sentences and sentence type models to identify the sentence's type. Namely, through the calculation of the relationship of the characteristic words, part of speech, and the sentence sequence of sentences, the sentence type of a sentence was identified.

According of idea about syntax isomorphic and sentence systems, sentence elements include syntactic, word sequence, grammatical form length, characteristic word. Correspondingly, the similarity between sentences and phrases related with these elements, mainly includes three types: sentence characteristic word similarity, sentence length similarity, sentence sequence similarity.

Sentence characteristic word similarity is calculated by the Weight *W*, shown as Equation (3). Sentence length similarity is calculated by Equation (6).Sentence sequence similarity is is calculated by Equation (7).

The authors use BaiDu search engine to get a total of 6000 search results with a variety of phrases. Manual processing carry out these results, weed out some of the repetitive or advertising sentence, then select 1000 sentences as the testing set, and. manual processing mark out sentence characteristic word, sentence length and sentence sequence.Further, the testing sentences are divided into two parts: the correct syntax sentence testing set and the erroneous syntax sentence testing set. By the forementioned analysis, the testing results are shown as in Table 1,and the overall sentence recognition accuracy was 62.2%.

The authors used BaiDu search engine to get a total of 6000 search results with a variety of phrases. Manual processing carry out these results, weed out some of the repetitive or advertising sentence, then select 1000 sentences as the testing set, and. manual processing mark out sentence characteristic word, sentence length and sentence sequence. Further, the testing sentences are divided into two parts: the semantic sentence testing set and the syntax sentence testing set. By the forementioned analysis, the testing results are shown as in TABLE I, and the overall sentence recognition accuracy was 64.4%.

TABLE I.
Testing Results of Semantic and Syntax Sentence

	Sentences Total	Results Correct Sentence	Correct rate (%)
Testing	1000	644	64.4
Semantic Sentence	432	422	97.7
Syntax Sentence	568	222	39.1

Comparing the similarity of keyword weight with the similarity of the word form, sentence length word sequence and the similarity of the word meaning.The test results are shown in TABLE II if similarity value is greater than or equal to *0.7*.

TABLE II.
Experimental Results Comparison

Method	Tested sentences	Results Ccorrect sentence	Correct rate (%)
The similarity based on word form, sentence length and word sequence	500	320	64
The similarity based on the word meaning	500	400	80
The similarity based on keyword weight	500	430	86

Experimental results show that the accuracy of search result based on keyword weight is the highest by comparing the similarity of keyword weight with the similarity of the word form, sentence length, word *sequence* and the similarity of the word meaning, and shows that the keyword weight similarity model has played an important role in the matching problem between users' questions and answers library so that the system accuracy has been greatly improved.

CONCLUSIONS

Similarity model is a core part of the automatic *QA* system.You can design a query system to meet user requirements as taking full advantage of the characteristics of the Chinese language itself, sentence composed of words and semantic information.

In this paper, the proposed sentence similarity method of key words weight can significantly improve the accuracy of the calculation by keyword extraction and POS assignment, increasing the importance of nouns and verbs in the sentence and taking into account two aspects of the syntax and semantics in sentence similarity. Although the experiment obtains a better accuracy rate, there are a lot of inadequacies. In order to achieve better results, we need to further improve the accuracy of synonyms and synonyms expansion.

ACKNOWLEDGMENT

This work was supported by the Social Science Foundation of the Colleges and Universities of Jiangsu Province (No.2012SJD870001) and the Science and Technology Research Project of Huai'an (No.HAG2012055).The authors would like to express our sincere thanks to the editors and the reviewers for giving very insightful and encouraging comments.

REFERENCES

[1] S.L.CHen, Y.CH.Song, W.L.Li,"Exact Phrases in Information Retrieval for Question Answering", Proceedings of the 2nd workshop on Information Retrieval for Question Answering, Manchester, UK, 2008, pp.9-16.

[2] Chaogai Xue, Lili Dong, Guohua Li, "An Improved Immune Genetic Algorithm for the Optimization of Enterprise Information System based on Time Property", Journal of Software, Vol.6, No.3, pp.436-443, Mar 2011.

[3] Zhongjun Li, "Research on the Website Keywords Seeding System Based on SEO", Journal of Computers, Vol.6, No 1, pp.75-82, Jan 2011.

[4] Tian Xia, Yanmei Chai, "An Improvement to TF-IDF: Term Distribution based Term Weight Algorithm", Journal of Software, Vol.6, No.3, pp. 413-420, Mar 2011.

[5] K.Boris, M.Gregory, B.CH.Gary, et al,"The START Natural Language Question Answering System", http://start. csail.mit.edu, 2006.

[6] Xueping Peng, Zhendong Niu, Sheng Huang,et al. "Personalized Web Search Using Click through Data and Web Page Rating", Journal of Computers, Vol.7, No.10, pp. 2578-2584, Oct 2012.

[7] CH.L Song, M.Y Dai, H.CH Yan, et al, "A Template Alignment Algorithm for Question Classification", ISI2008, 2008, pp. 136-140.

[8] Haijiang He, "A Co-Ranking Algorithm for Learning Listwise Ranking Functions from Unlabeled Data", Journal of Computers, Vol.6, No.11, pp. 2302-2309, Nov 2011.

[9] B.Barla Cambazogl and Cevdet Aykanat, "Performance of Query Processing Implementations in Ranking-based Text Retrieval Systems Using Inverted Indices", Information Processing & Management, 2006, pp. 875–898.

[10] Minsu Jang, Joo-Chan Sohn, Hyun Kyu Cho,"Automated Question Answering using Semantic Web Services", IEEE Asia-Pacific Services Computing Conference, 2007, pp. 344-348.

[11] Liyi Zhang, Mingzhu Zhu, Wei Huang, "A Framework for an Ontology-based E-commerce Product Information Retrieval System", Journal of Computers, Vol.4, No.6, pp.436-443, Jun 2009.

[12] Xinyue Liu, Hongfei Lin, Liguo Zhang, "An Attractive Force Model for Weighting Links in Query Dependant Web Page Ranking", Journal of Computers, Vol.7, No.1, pp.124-129, Jan 2012.

[13] Jun Zhai, Yan Chen, Yi Yu, et al, "Fuzzy Semantic Retrieval for Traffic Information Based on Fuzzy Ontology and RDF on the Semantic Web", Journal of Software, Vol. 4, No.7, pp. 758-765, Sep 2009.

[14] Lin Sun, Ning Zhong, Jiucheng Xu, "Granularity Based User Centric Multi-Strategies and Application in Knowledge Retrieval", Journal of Computers, Vol.8, No.8, pp.2101-2109, Aug 2013.

[15] Wei Li, Wenhua Zeng, "Structure Feature Collection of Chinese Language Sentence Patterns", Journal of China Universities of Posts and Telecommunications，2006, 13, pp.74-77.

Yangxin Yu was born in Taizhou of Jiangsu Province, China in 1970 and respectively received both his B.Sc degree(1995) and the M.Sc degree(2007) from College of Computer Science and Technology,Wuhan University of Technology and School of Computer Science and Technology, Suzhou University.

He is currently an associate professor in the Faculty of Computer Engineering, Huaiyin Institute of Technology, Information society member of Jiangsu Province and anonymous reviewers of Library and Information Service. His research covers a broad range of topics within AI including information management and information systems, information retrieval,intelligent information processing and personalization techniques.

Prof.Yu has authored or co-authored about 40 journal and conference papers, a few of which have got the honor of Huaian Committee natural science outstanding papers and has won the third prize of Science and Technology Progress Award of Huai'an Committee, edited three textbooks in chief.

Liuyang Wang, was born in Huaiyin of Jiangsu Province, China in 1974 and respectively received both his B.Sc degree(1998) and the M.Sc degree(2009) from College of Computer Science and Technology, Nanjing University of Science and Technology.

He is currently an associate professor in the Faculty of Computer Engineering, Huaiyin Institute of Technology, outstanding young teachers. His research covers a broad range of topics within AI including information management and information systems, information security and software Testing.

Prof.Wang has authored or co-authored about 20 journal and conference papers, chaired and completed a number of research projects; a few of which have got the honor of Huaian Committee natural science outstanding papers and has won the third prize of Science and Technology Progress Award of Huai'an Committee, edited four textbooks in chief.