ORIGINAL RESEARCH

The improvement of question process method in Q&A system

Yonghe Lu*, Shuo Wang

School of Information Management, Sun Yat-sen University, Guangzhou, China

Received: July 19, 2015 **Accepted:** September 15, 2015 **Online Published:** October 10, 2015
DOI: 10.5430/air.v5n1p36 **URL:** http://dx.doi.org/10.5430/air.v5n1p36

ABSTRACT

Life service information plays an important role in people's life, such as weather conditions, so the study of how to get life service information has important significance. This paper put forward a question processing method called "integrated semantic algorithm" in Q&A System of life service information. The new algorithm was based on the semantic web, word order similarity algorithm and the syntactic similarity algorithm. When matching the question templates, especially for some question templates which are characteristic of certain fields, the new algorithm can identify the type of questions, narrow the matching range of the question templates, and improve the matching accuracy. In the experiment, we chose "weather field" as the experimental subject. In the first experiment, we built the question syntactic templates and semantic web of weather, and collected 55 questions of weather title as test set. Then we used the word similarity algorithm, the syntactic similarity algorithm and integrated semantic similarity algorithm to match question templates with the test question set. The experimental results show that the integrated semantic algorithm is better than the other two algorithms in matching accuracy. In the second experiment, we randomly selected some questions from different fields, then we used the three similarity algorithms in the first experiment to do the field distinguishing experiment. The experiment shows that only the integrated semantic algorithm can recognize questions of different fields.

Key Words: Life service information, Q&A System, Semantic web, Integrated semantic similarity algorithm

1. INTRODUCTION

Today, the Internet has become the main source of information. We can get life service information about the finance, lifestyle, transportation and so on from the Internet. Because this kind of information has timeliness, so it only has value in a period of time. In terms of different kinds of life service information, this valuable time is also different. For example, weather information is valuable in one day, but financial index such as stock is only valuable in few minute or second.

In the face of a huge amount of Internet information, users can't retrieve and filter the information easily. As a result, the concept of Question Answer System (QAS) is put forward. In QAS, the input is a question asked in natural language and the output is the answer in natural language too, but not a lot of related documents.[1] This paper introduces a simple semantic web into traditional method to improve the method of question processing in the life service QAS. The improved method realizes a better reply in the life service QAS.

2. RELATED WORK

2.1 Q&A System

At present, the study about Q&A system is divided into two parts——the limited-domain and open-domain. For the

*Correspondence: Yonghe Lu; Email: luyonghe@mail.sysu.edu.cn; Address: School of Information Management, Sun Yat-sen University, No.132, East Waihuan Road, Higher Education Mega Center, Panyu District, Guangzhou City, Guangdong Province, China.

research of the opening-domain QAS, zheng-tao yu, xiaozhong made the computer understand the natural language by using question sentence corpus and the "Hownet" (a Chinese Wordnet).[2] Based on the traditional TF-IDF similarity algorithm, the semantic factor is introduced to improve the accuracy of the similarity calculation. Ye Zheng, Lin Hong-fei present a computation approach of question similarity based on split vector space model and semantic concept according to the common question characteristic research.[3] Oleksandr Kolomiyets and Marie-Francine Moens suggest a general question answering architecture that steadily increases the complexity of the representation level of questions and information objects.[4] Paloma Moreda *et al.* present two proposals based on semantic information, semantic roles and WordNet for the answer extraction module of a general open-domain question answering (QA) system.[5] Alessandro Moschitti *et al.* described question and answer pairs by means of powerful generalization methods,[6] and exploited the application of structural kernels to syntactic/semantic structures. For the research of the limited-domain Q&A System, Cui Huan, DongFeng CAI put forward a method of answer extraction based on sentence similarity calculation, and they designed a Chinese QAS using Internet search engine, this system do well in question about name, number and time.[7] Xianling Mao divided the Q&A system into three categories:[8] the structural data based question and answer, the free-text based question and answer, the question-answer pairs based question and answer. The current studies of QAS are almost based on a collection of documents and existing question-answer collection. These two collections have some delay, and there is few research on life service information.

Research about Life service information system mainly concentrate in the field of finance, transportation and production process. Sun Xue used XML analyze the data on the Internet, which can identify the life service information.[9]

2.2 Question process

Generally, users ask questions in natural language when using Q&A system, so it is critical to understand user's questions.

Jia Junzhi and Mao HaifeI described the syntax relations of the questions by depending on dependency relationship and matching the question with the template in the storehouse, and it defines the valent pattern of the question, realizes semantic annotation with frame elements.[10] Kwanho Kim, Beom-suk Chung *et al.* proposed a novel kernel, called language independent semantic (LIS) kernel,[11] which is able to effectively compute the similarity between short-text documents without using grammatical tags and lexical databases. In fact, question template matching is also one kind of short-text classification. Jin Yaohong put forward a texts similarity algorithm,[12] which can deal with the domain of the text and the semantic role of the object, computes the synonymy, polysemy and the combination among concepts. Jia Junzhi and Tai Yangfang realized semantic analysis on question sentence based on semantic frame.[13] It can identify the target words of different type of question sentences on the basis of Chinese syntactic dependency relation, finally identify the question focus and question type of sentences. Zhao Zhen *et al.* described a method for semantic similarity computation based on the Multi-feature of a Sentence (MFS) which integrates the weights of words, word semantics and structure with together.[14]

In the question and answering system based on question-answer pairs, to do the question matching, firstly, we should artificially collect and classify questions as question templates, in order to match and classify questions to be tested. Secondly, we should do the similarity calculation and question template matching, details are as follows:

(1) Word similarity calculation
Dekang Lin considerd that the similarity of any two things depends on their commonality and differences.[15] The formula is as 1.

$$Sim(A,B) = \log P[common(A,B)] / \log P[description(A,B)] \quad (1)$$

Where the concept of synonyms was introduced into the general calculation, HowNet was used to search keywords for sememe.[16] Yang Sichun described an improved method of sentence similarity including the induction of synonyms in sentence similarity definition.[17]

(2) Syntactic similarity calculation
Hang Cui introduced syntactic dependency tree into Q&A system, used the fuzzy matching relation of the statistical model to realize a better understanding of the sentence.[18] Wang Pin *et al.* used syntax analyzer to do the sentence analysis, and then calculated the similarity based on semantic dependency structure.[19] To express the meaning of a sentence, we should consider not only the keywords and the arrangement of words, but also the semantic. Because the two methods mentioned above have their own advantages and disadvantages, so the sentence similarity $S(S_1, S_2)$ between sentence S_1 and sentence S_2 can be computed as the formula(2):[19]

$$S(S_1,S_2) = \lambda_1 \times S_{word}(S_1,S_2) + \lambda_2 \times SIM(S_1,S_2) \quad (2)$$

Where $S_{word}(S_1, S_2)$ is the Word similarity between

S_1 and S_2, $SIM(S_1, S_2)$ is the syntactic similarity between S_1 and S_2, λ_1 and λ_2 are the weighting coefficient and $0 < \lambda_1 \leq \lambda_2 < 1, \lambda_1 + \lambda_2 = 1$.

3. Improvement of Question Process Method Based on Semantic Web

Life service information is a special kind of information. In life service QAS, the most important step is to deal with questions, because answer processing and obtaining depend on the information obtained from the question processing flow. The purpose of question processing is to match the existing question templates with the test questions. The core technology of matching is the similarity calculation of questions. The method of similarity calculation mentioned above is a general method. In the specific fields, semantic information relying solely on the thesaurus is still too weak, and the method can't match well in different fields. Therefore, this paper introduces the method of the Field Semantic Web into the traditional method. The improved algorithm is called Integrated Semantic Algorithm, which can improve the efficiency of question template matching. It also improves the ability of identifying fields and extract the necessary criteria for answer obtaining.

Part of the improvement is shown in Figure 1:

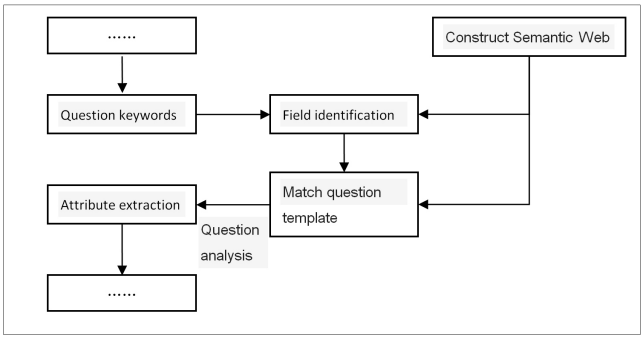

Figure 1. Improvement of Question Process Method

3.1 Construction of field semantic web and field identification

Firstly, in the pre-construction stage of the QAS, we should construct Semantic Web in various fields. The first layer of child nodes in the Semantic Web include time, location, attributes, entities, *etc*. We should collect questions on the online Q&A platforms about these life service fields. Then we manually classify different points that the question asked. The types of the classifications about these points make up the first layer of child nodes. And the different points of the question make up the second layer of child nodes below their classifications, the child nodes may be an adjective or noun. Secondly, after extracting keywords and before matching question templates, we can match the keywords with the Semantic Web to identify the type of problem and narrow the matching range of the question templates. Thereby, the amount of matching calculation can be reduced and efficiency can be improved. Identification rules are as follows:

Rule 1: If the keywords of the sentence appear in root node of Filed Semantic Web, such as "weather", "train" and so on, regard it as the primary field of similarity calculation.

Rule 2: If the keywords of the sentence aren't the same as the root node, we should search the other nodes except the time and place nodes, sum up the number of keywords that appear in the nodes, and find out the first three fields that have the maximum matching number, then calculate the similarity in these three fields. Here, in order to both simplify the calculation and ensure the accuracy, we choose the first three fields.

3.2 Integrated semantic algorithm

When matching the question templates, question similarity computation is the most important step. Integrated semantic algorithm has some improvements compared with the traditional methods. The core improvement is shown in Formula (3), where the concept of Field Semantic Web is introduced to the question similarity computation. As we all know, different fields have some different question templates. Because the integrated semantic algorithm consider more about the templates' characteristics of different fields. It can improve the matching accuracy when matching the question templates, especially for some question templates which are characteristic of certain fields.

$$S(S_1, S_2) = \lambda_1 \times S_{word}(S_1, S_2) + \lambda_2 \times SIM(S_1, S_2) + \lambda_3 \times S_{sn}(S_1, S_2) \tag{3}$$

In formula (3), λ_1, λ_2 and λ_3 are the weighting coefficient and $0 < \lambda_1 \leq \lambda_2 \leq \lambda_3 < 1; \lambda_1 + \lambda_2 + \lambda_3 = 1, S_{sn}(S_1, S_2)$ is the parameter of Field Semantic Web and the computation rule of the $S_{sn}(S_1, S_2)$ is shown as follow:

(1) If the root words of Semantic Web appear, such as "weather", "train" and so on, it indicates the possibility that the sentence has high correlation with the field. But in order to distinguish the case of child node, the value of $S_{sn}(S_1, S_2)$ should be more than 1, but if the value is too high it will cause excessive weight so that the role of child nodes is weakened. Here, when computing the similarity between question and question template in its fields, through simple test, we take 5 as

value of $S_{sn}(S_1, S_2)$;

(2) Based on Rule 1, if the keyword in the template and keyword in the target question appear in the same child node of semantic network, adds the number of the keywords to $S_{sn}(S_1, S_2)$.

4. EXPERIMENTS

The experiments aim to verify the matching rate of the questions on the field of life service information by using the above three algorithms. After accurately determining the type of question, we can extract effective answers for different types of questions and then return the answers to users in natural language.

4.1 Pre-process

In the experiments, the paper chose the "weather" field as an instance. Weather information is one of the most common life service information in daily life. First, we built question templates of weather field and summarize various question types of weather field artificially. Then some questions were collected as the representation of each type of questions, and the expected answers should be summarized. Question templates in weather field are shown in Table 1:

Table 1. Question Template in Weather Field

Type	Template sentence	Expected answer
Weather	1-How's the weather today? 1-I want to know the weather situation today.	It is cloudy today.
Weather attribute	2-Is today a fine day?	Yes, it's fine today.
	2-Is it cloudy today?	No, it's fine today.
Temperature	3-What's the temperature today?	The temperature is 17℃-25℃.
Air quality	4-What about the air quality today? 4-Is't air pollution today?	The air quality rated an "excellent".

Secondly, we constructed Field Semantic Web in weather field. Figure 2 shows part of the Semantic Web.

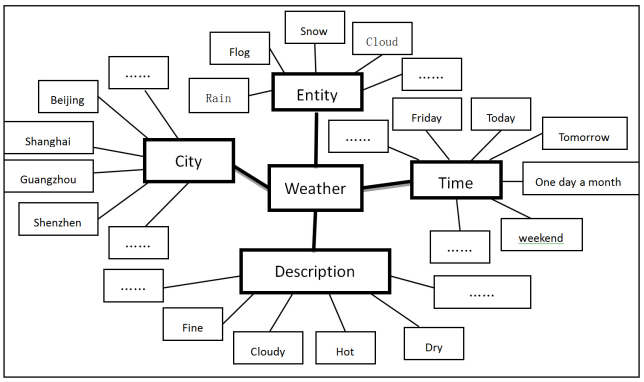

Figure 2. Semantic Web in weather field

4.2 Experiment data set

After finishing the pre-preprocessing, we collected questions in the weather field as experimental sets. Currently, there is no public test set about Chinese question sentences in the world. Testing corpus is usually built artificially. Selection of similar sentence mainly depends on the subjective judgment because there is no uniform standard. Consequently, it is not easy to get high-quality corpus. The experimental sets used in this paper were collected from two large online platforms, namely "Baidu Know Platform" and "Soso Know Platform". Table 2 shows part of the experimental set of questions.

Table 2. Part of the experimental set of questions

Number	Question	Type of expected answer
1	Is it a little cloudy today?	2
2	How is the weather today in Tibet?	1
3	What was the weather like today?	2
4	How is the weather today?	1
5	What is the temperature in Fuzhou today?	3
6	The weather is going to rain today?	2
……	……	……

4.3 Experimental steps

The experiment is divided into two parts, the first part is the comparison of word similarity algorithm, syntactic similarity algorithm and the integrated semantic similarity algorithm. We compared these three similarity algorithms to detect the accuracy in question template matching. The second part, we chose questions in different fields, and used three methods to identify the field of the questions.

The word similarity algorithm calculates the number of the same or similar words between two sentences. The experi-

ment uses X-similarity java program to compute. And syntactic similarity algorithm is based on the syntactic interdependent relationship. It's better to consider the syntactical meaning of the sentence, rather than the meaning of single word. Syntactic analysis uses Stanford-parser. And integrated semantic algorithm is a combination of word order similarity, syntactic similarity and the semantic web. The paper used the calculation method of synonyms in X-similarity program, the syntax analysis method in the Stanford parser and the semantic similarity algorithm written by ourselves.

The experimental steps are as follows:

(1) Write the java programs. We used ICTCLAS segmentation code to segment words of a sentence, then removed the stop word, and wrote the program code of three similarity algorithms;

(2) Question template matching. Compute the similarity between the question of the experiment and the question template by using three algorithms. According to the Wang Pin's experiment, The best value of λ_1, λ_2 in formula (2) should be 1/2.[19] The same as the formula (2), the value of $\lambda_1, \lambda_2, \lambda_3$ in formula (3) should be 1/3. Then, we regarded the template which had the highest similarity as the matching template. If the type of the matching template was the same as the excepted answer, we regarded it as "successful matching", otherwise "failed matching". We got the experiment data of three similarity algorithms by traversing all sentences in the experiment set.

(3) Field identifying. We randomly selected two questions in weather field and some other questions in other fields, and computed the similarity with three algorithms, then compared the highest similarity value and matching results, detected the ability of discrimination in different field.

4.4 Experiment result and analysis

4.4.1 *Question template matching experiment*

In the first part of the experiment, the experiment computed the similarity of the 55 different questions with three algorithms. The results of matching accuracy are shown in Table 3.

Table 3. Matching Accuracy of Three Algorithms

	Integrated semantic algorithm	Syntactic algorithm	Word algorithm
Matching accuracy	80.00%	70.91%	65.45%
Matching number	44	39	36

Table 4. Part of the Similarity Results and Matching Results

Number	Results of integrated semantic algorithm	Highest similarity	Results of syntactic algorithm	Highest similarity	Results of word algorithm	Highest similarity
1	1	3.22381	0	0.360714	1	0.660633
2	1	2.55291	1	0.329365	1	0.730154
3	1	3.238095	0	0.392857	0	0.769231
4	1	2.680556	1	0.520833	1	0.8
5	1	3.984127	1	0.47619	0	0.754325
6	1	3.296296	1	0.444444	0	0.729167
…	…	…	…	…	…	…

Table 3 shows the matching accuracy rate of three algorithms. The matching accuracy increases in this order - "word algorithm", "syntax algorithm" and "integrated semantic algorithm". By using the integrated semantic algorithm, successful matching rate reaches 80%. Part of the similarity results and matching results are shown in Table 4.

In Table 4, we regard the highest similarity matching template as the correct template of the question, and the "1" in "result column" represents successful matching. "0" is on behalf of the failed matching. The definition of successful matching and failed one are explained in experimental step 2.

4.4.2 *Field identifying experiment*

In the second part of the experiment, the results of similarity values and matching in different fields are shown in Table 5.

Table 5 shows that the similarity calculation results of questions 4 and 8, whose field belongs to weather, are obviously more than 1 by using integrated semantic algorithm, and the results of the other questions are less than 1. So integrated semantic algorithm can distinguish sentences in various fields and make the appropriate response, and it will not misidentify the sentences in other fields. The other two algorithms can't effectively distinguish whether the questions are in weather fields or not.

In summary, the integrated semantic algorithm proposed in this paper is effective and can be used to question process in automatic Q&A system.

Table 5. Results of Similarity Values and Matching

Number	Question	Word algorithm	Syntactic algorithm	Integrated semantic algorithm	Matching results
1	Has the train from Beijing to Guangzhou today?	0.862116	0.625	0.708333	none
2	Is there any restaurant nearby?	0.634964	0.071429	0.047619	none
3	Can we drink the boiled water in the second day?	0.647432	0.027778	0.018519	none
4	Is it going to rain tomorrow?	0.842538	0.235714	1.490476	Weather Type 3
5	How to let the smell of perfume released into the surrounding air?	0.560401	0.113636	0.075758	none
6	What is the meaning of "Winning a thousand miles"	0.685763	0.071429	0.047619	none
7	How much throw-away lunchbox are used per day in the world?	0.680413	0.13125	0.0875	none
8	What's the weather like today?	1	0.75	2.833333	Weather Type 1
……	……	……	……	……	……

5. Conclusion

The paper puts forward a kind of template matching algorithm in question processing based on the semantic web. And it is combined with the word order similarity algorithm and the syntactic similarity algorithm. In the experiment, we chose "weather" as the experimental field. The experiment used the word similarity algorithm, the syntactic similarity algorithm and integrated semantic similarity algorithm to match question template with the question set. The experimental results show that the integrated semantic algorithm is better than the other two algorithms in matching accuracy and only the integrated semantic algorithm can recognize questions of different fields. However, the paper still has the problems, for example, there aren't sufficient experimental question templates and experimental set of data. Further work can be carried out in the parameter optimization of integrated semantic algorithm by using some of the existing algorithms, such as ant colony algorithm and genetic algorithm. Finally the answer accuracy of the automatic Q&A system can be increased.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC Grant No: 71373291). This work also was supported by the National High Technology Research and Development Program of China (863 Program) under Grant 2012AA101701.

References

[1] Guo Q, Fan X. Question Answer System Based on Natural Language Understanding. Computer Engineering. 2004; 30(13): 11-12. http://dx.chinadoi.cn/10.3969/j.issn.1000-3428.2004.13.005

[2] Yu Z, Fan X, Song L. Restricted Domain Question Answer System Based on Question Sentence Corpus. Computer Engineering and Applications. 2003; (36): 28-31. http://dx.chinadoi.cn/10.3321/j.issn:1002-8331.2003.36.009

[3] Ye Z, Lin HF, Yang Z. Chinese FAQ system based on sentence similarity. Computer Engineering and Applications. 2007; 43(9): 161-3. http://dx.chinadoi.cn/10.3321/j.issn:1002-8331.2007.09.047

[4] Kolomiyets O, Moens M. A survey on question answering technology from an information retrieval perspective. Information Sciences. 2011; 181(24): 5412-34. http://dx.doi.org/10.1016/j.ins.2011.07.047

[5] Moreda P, Llorens H, Saquete E, *et al*. Combining semantic information in question answering systems. Information Processing and Management. 2011; 47(6): 870-85. http://dx.doi.org/10.1016/j.ipm.2010.03.008

[6] Moschitti A, Quarteroni S. Linguistic kernels for answer re-ranking in question answering systems. Information Processing and Management. 2011; 47(6): 825-42. http://dx.doi.org/10.1016/j.ipm.2010.06.002

[7] Cui H, Cai D, Miao X. Research on Web-based Chinese Question Answering System and Answer Extraction. Journal of Chinese Information Processing. 2004; (18): 24-31. http://dx.chinadoi.cn/10.3969/j.issn.1003-0077.2004.03.004

[8] Mao X, Li X. A Survey on Question and Answering Systems.

Journal of Frontiers of Computer Science and Technology. 2012; (6): 193-206. http://dx.chinadoi.cn/10.3778/j.issn.1673-9418.2012.03.001

[9] Sun X. Web Data Mining for Life service Information Based on XML. Taiyuan University of Technology. 2006; (5).

[10] Jia J, Mao H. Study on the Question Processing in Chinese FrameNet Question Answering System. Library and Information Service. 2008a; 52(10): 55-7.

[11] Kim K, Chung B, Choi Y, *et al*. Language independent semantic kernels for short-text classification. Expert Systems with Applications. 2014; 41(2): 735-43. http://dx.doi.org/10.1016/j.eswa.2013.07.097

[12] Jin Y. Text Similarity Computing Based on Context Framework. Computer engineering & application. 2004; (16): 36-8. http://dx.chinadoi.cn/10.3321/j.issn:1002-8331.2004.16.013

[13] Jia J, Tai Y. The Design and Implementation of Question Analysis for Q&A System Based on Chinese FrameNet. New Technology of Library and Information Service. 2008b; 165(6): 11-5. http://dx.chinadoi.cn/10.3969/j.issn.1003-3513.2008.06.003

[14] Zhao Z, Wu N, Song P. Sentence Semantic Similarity Calculation Based on Multi-feature Fusion. Computer Engineering. 2012; 38(1): 171-3. http://dx.chinadoi.cn/10.3969/j.issn.1000-3428.2012.01.054

[15] Lin D. An Information Theoretic Definition of Similarity Semantic distance in WordNet. Proceedings of the Fifteenth International Conference on Machine Learning. 1998.

[16] Lin L, Xue F, Ren Z. Modified word similarity computation approach based on HowNet. Journal of Computer Applications. 2009; 29(1): 217-20.

[17] Yang S. An Improved Model for Sentence Similarity Computing. Journal of University of Electronic Science and Technology of China. 2006; 35(6): 956-9. http://dx.chinadoi.cn/10.3969/j.issn.1001-0548.2006.06.027

[18] Cui H, Sun R, Li K, *et al*. Question Answering Passage Retrieval Using Dependency Relations. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; 2005. pp. 400-6. http://dx.doi.org/10.1145/1076034.1076103

[19] Wang P, Huang G. Sentence Similarity Computation in Information Retrieval. Computer Engineering. 2011; 37(12): 38-40.