Design of Intelligent Tourism Question Answering System Based on Semantic Web

Hua HU¹,a,* and Shi-zheng ZHU²,b

¹College of Applied technology, Southwest University, Chongqing 401147, China
²ZTE Chongqing software CO., LTD, Chongqing 401121, China

¹huhua@swu.edu.cn, bsz_zhu611@163.com

*Corresponding author

Keywords: Smart Tourism, Semantic Network, Intelligent Question Answering System

Abstract. In order to promote the tourism service more automated and intelligent, this paper designs an intelligent tourism question answering system through the key technology of the semantic web, which makes it easy for users to quickly and accurately acquire the query results and information from massive amounts of data on the network. Greatly improving the semantics of sentence, the algorithm posed in the system provides a new method for question matching as a fusion similarity algorithm based on literal similarity and semantic similarity.

Introduction

With the rapid development of the tourism and the continuous infiltration of information technology, the demand of deep integration between industry and technology is more and more intense, so the notion of smart tourism has been gaining currency recently, particularly as a strategic tool for tourism development. Generally speaking, smart tourism aims to develop information and communication infrastructure and capabilities in order to improve management/governance, facilitate service/product innovation, enhance the tourist experience, and, ultimately, improve the competitiveness of tourism firms and destinations[1]. One of the biggest problems of information processing technology encountered in the development of smart tourism is how to make the users find the accurate information which they need in the sea of massive and complicated tourist information.

As the next generation network, semantic web can effectively solve the problem of intelligent information search. Using semantic technology can make the massive information existing in the network have semantics, and the information can be understood by computer in the process of processing information, so that the information resources on the Internet can achieve the all-round interconnection in the semantic layer, and eventually realize the communication between the human and the computer. From the point of view of input and output, the input of question answering system is a natural language question, not just key words; the output is a sentence or a paragraph, rather than containing a series of links related content.

Question answering system is a research field that can directly reflect computer intelligence, and how to overcome communication obstacle between people and information systems, as well as improving the recall ratio and precision ratio of computer is always the key of intelligent question answering system. From the current research progress, the researchers paying attention to the intelligent question answering system based on semantic web mainly concentrate on two aspects: question matching and answer extraction, while neglect the sentence similarity which plays a crucial role on the accuracy of question matching and answer extraction. There are some shortcomings in the study of the sentence similarity: the similarity of the question is calculated either from the literal similarity of the question, ignoring the semantic similarity of the question, or from the semantic similarity of the question, ignoring the semantic dictionary noise introduced false positive rate.
With consideration of the problems above, this paper designs an intelligent tourism question answering system based on semantic network. Aiming at question matching module in intelligent tourism question and answer system, this paper proposes a fusion similarity algorithm based on literal similarity and semantic similarity. The algorithm takes into account the literal similarity and semantic similarity of the question, which greatly improves the accuracy of question matching.

**Technologies of Semantic Web**

**Sentence segmentation**

Segmentation is a generic concept including three aspects: word segmentation, POS tagging and entity recognition, which is the natural language processing of the basic module. There are many open source toolkit can be used directly, such as Paoding word segmentation, ICTCLAS\(^2\) and so on, and many open source natural language processing platform also provides segmentation features, such as LTP of Harbin Institute of Technology, NLPIR of Beijing Institute of Technology. In addition to word segmentation, word segmentation and naming entity recognition are also needed. Therefore, ICTCLAS word segmentation tool of Chinese Academy of Sciences is the preferred method, which is Multi-level lexical analysis system based on hidden Markov model.

**Keywords extraction**

Keywords extraction is of great significance to question classification, similarity calculation and answer retrieval. The keywords extraction process is: First, we obtain tagging segmentation sequence and named entity sequence by word segmentation processing. Then we implement the stop word filtering and POS filtering for the segmentation sequence, and obtain part of feature words; Finally, we obtain the keywords of question by merging named entity sequence and feature words generated by previous step.

**The research method of sentence similarity**

At present, the methods of researching similarity of question are keyword-based method, semantic-dictionary method, statistic-based method and so on. Most of these methods are based on the analysis of linguistic symbol, unable to be understood from the semantic of sentence.

In this paper, the fusion similarity calculation algorithm is used to combine word matching with word semantic comprehension and improve the accuracy of query similarity calculation greatly.

**Design of Intelligent Tourism Question Answering System**

**The overall framework of system**

The Intelligent Tourism question answering system(QAS) which is put forward in this paper includes five parts: data layer, pretreatment layer, index layer, retrieval layer and presentation layer. The data layer is responsible for answering library acquisition, the maintenance of corpus and the acquisition of potential questions. The potential questions are mainly obtained by frequently asked questions the user voluntarily submit to. The pretreatment layer is responsible for the extraction of knowledge, the pretreatment of question and the construction of vector. In the pre-processing part of the question, this paper adopts regular expression to achieve noise removal, and takes clustering scheme to remove duplicate questions. On the basis of the ICTCLAS word segmentation in Chinese Academy of Sciences (CAS), the vector representation of query questions is obtained by means of word tagging, keyword extraction and keyword semantic extension.

The function of index layer is to construct the forward and inverted index. In this paper, we use the statistic method to filter the front row table and MapReduce framework to construct the inverted table.
The retrieval layer is responsible for querying candidate question set and matching further question according to the user's questions, and question matching is also the focus of this paper. Index retrieval will first use binary search to obtain candidate set of questions, and then take the literal and semantic similarity integrated method to match the most similar questions.

The presentation layer achieve the Q & A show and user interaction by using JAVASCRIPT, HTML and other technologies, and give the final retrieval results or not found the system tips.

**System platform construction**

Based on the server operating system platform independent considerations, this system uses java language and high portability MySQL 5.5 database. Using Spring MVC + MyBatis to build a three tier system architecture, the three tier architecture mainly contains persistence layer, business layer and presentation layer. The database server is used to store all kinds of data and corpus in the system, while the client of the system only needs a browser. According to these techniques, the separation of presentation layer and logic layer is realized, which makes the system have a good expansibility.

**Question Similarity Calculation Algorithm**

The process of question matching is the core processes of intelligent tourism question answering system proposed by this paper, and question similarity computation technology is the key to decide whether question matching is correct or not. This paper proposes a question similarity algorithm suitable for tourism question answering system by using synonyms, synthetically considering the literal similarity and semantic similarity. The question similarity calculation algorithm used in this system firstly preprocesses the questions input by the user and the questions in candidate set, then successively calculate the literal and semantic similarity of user's input question and candidate question, followed by fusion formula to obtain the final similarity; Finally, using the judgment rule similarity to decide whether the user’s input question is similar to the candidate question. This paper uses the vector space model[3] to calculate the literal similarity of question, and target question and each question in question answering system can be used a n dimensional vector \( T = \langle T_1, T_2, ..., T_n \rangle \) to represent. The similarity between the two sentences can be expressed by the cosine of the included angle between the two vectors \( T \) and \( T' \). The calculation formula is as follows:

\[
\text{Similarity}(T, T') = \frac{\sum_{i=1}^{n} T_i \times T'_i}{\sqrt{\sum_{i=1}^{n} T_i^2 \sum_{i=1}^{n} T'_i^2}}
\]  

(1)

The Word forest collected by Harbin Institute of Technology was used to calculate the semantic similarity of questions in this paper. Word forest contains more than 70000 words, using five layers tree structure to organize entries together, with an 8-bit digital hybrid coding to represent each atomic word group, the former seven are the vocabulary of large, medium and small classes, the eighth symbol(#, &, =) represents the relation of each word in atomic word group is semanteme correlative, independent, or the same.

To calculate the semantic similarity of questions, firstly, we need to calculate the semantic similarity of words based on Word forest, and the calculation algorithm of word semantic similarity is as follows:

If the word sense of two words \( w_1 \) and \( w_2 \) is not on the same tree (the first uppercase English letter of word sense encoding is different), then define \( \text{sim}(w_1, w_2) = a \);

Otherwise, if the word sense of two words in the second layer (the second lowercase English letter of word sense encoding is different), then define

\[
\text{sim}(w_1, w_2) = b \times \cos(n \times \Pi / 180) \times (n - k + 1) / n;
\]  

(2)
If the word sense of two words in the third layer (the two third decimal numbers of word sense encoding is different), then define \( \text{sim}(w_1, w_2) = c \times \cos(n \times \Pi / 180) \times (n - k + 1) / n \);

If the word sense of two words in the fourth layer (the fourth uppercase English letter of word sense encoding is different), then define \( \text{sim}(w_1, w_2) = d \times \cos(n \times \Pi / 180) \times (n - k + 1) / n \);

If the word sense of two words in the fifth layer (the two fifth decimal numbers of word sense encoding is different), then define \( \text{sim}(w_1, w_2) = e \times \cos(n \times \Pi / 180) \times (n - k + 1) / n \);

Otherwise, the word sense encoding of two words is same. If the tail code is “=”, then \( \text{sim}(w_1, w_2) = 1 \), otherwise, \( \text{sim}(w_1, w_2) = f \).

Where \( n \) in the similarity calculation formula of layer branch represents the number of the nodes in the current branch layer, \( k \) indicates the distance between the two branches. If two words has multiple word sense, then need to calculate the similarity of multiple word sense, and then take a maximum as the final similarity of two words.

With the semantic similarity between words and words, it is easy to calculate the semantic similarity between sentences. The sentence semantic similarity computation algorithm is shown in figure 1.

![Figure 1. The sentence semantic similarity computation algorithm.](image)

The sentence semantic similarity computation algorithm is described as follows:

Set the two sentences \( A \) and \( B \), and the words \( A_1, A_2, \ldots, A_m \) and \( B_1, B_2, \ldots, B_n \) that are included. The similarity between \( A_i (1 \leq i \leq m) \) and \( B_j (1 \leq j \leq n) \) is represented by \( S(A_i, B_j) \), so that we can get a \( m \times n \) matrix:
Using this matrix, we can use the following formula to obtain the semantic similarity $S(A, B)$ between the two sentences $A$ and $B$:

$$S(A, B) = \frac{\sum_{i=1}^{m} a_i + \sum_{j=1}^{n} b_j}{2}$$  \hspace{1cm} (4)$$

$$a_i = \max(S(A_1, B_1), S(A_1, B_2), \ldots, S(A_1, B_n))$$

$$b_j = \max(S(B_j, A_1), S(B_j, A_2), \ldots, S(B_j, A_n))$$

With the sentence literal similarity and semantic similarity, the fusion similarity of questions can be calculated by the following formulas:

$$S_f = \frac{w_1S_1 + w_2S_2}{w_1 + w_2}$$  \hspace{1cm} (5)$$

Where $S_f$ is the fusion similarity of question, $S_1$ and $S_2$ are the literal similarity and semantic similarity of question, $w_1$ and $w_2$ are the weight of the question’s literal similarity and semantic similarity. Using the above algorithm to calculate the similarity between question, both guarantee the accuracy of question matching, but also to ensure that semantically similar questions will appear in the returned results.

**Experiment**

This paper selects 2000 standard question and answer relating to Chongqing tourism attractions, tourism traffic and tourism catering as the experimental data, and use the accuracy rate of answering questions and false positive rate(FPR), false rejection rate(FRR) as the criterion of intelligence tourism question answering system and question fusion similarity algorithm. Among them, the accuracy rate of answering questions can be calculated using the following formula:

$$\text{Precision} = \frac{|SQ|}{|TQ|}$$

Where $|SQ|$ is the number of users are satisfied with the answer , $|TQ|$ is the total number of questions. The experiment results are shown in Table 1 and Table 2:

**Table 1. Accuracy Comparison.**

<table>
<thead>
<tr>
<th>QAS</th>
<th>QAS in this paper</th>
<th>QAS in Refs[4]</th>
<th>QAS in Refs[5]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy Rate</td>
<td>86%</td>
<td>63%</td>
<td>57%</td>
</tr>
</tbody>
</table>

**Table 2. FPR and FRR Comparison.**

<table>
<thead>
<tr>
<th>Similarity Algorithm</th>
<th>Similarity Algorithm in this paper</th>
<th>Similarity Algorithm in Refs[6]</th>
<th>Similarity Algorithm in Refs[7]</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPR</td>
<td>0.8%</td>
<td>6.8%</td>
<td>5.1%</td>
</tr>
<tr>
<td>FRR</td>
<td>1.3%</td>
<td>8.4%</td>
<td>7.5%</td>
</tr>
</tbody>
</table>

According to the experimental results from table 1 and table 2, we can see that the intelligent tourism question answering system based on semantic web has a high accuracy rate of answering questions, and the fusion similarity algorithm also has a low false positive rate and false rejection rate.
Conclusion

In this paper, we designed and developed an intelligent tourism question answering system based on semantic web, and proposed a fusion similarity algorithm based on the literal similarity and semantic similarity of questions. Experimental results show that the intelligent tourism question answering system can accurately answer the user's questions, and the question similarity algorithm can significantly decrease the false positive rate and false rejection rate.

Acknowledgement

This study was supported by Fundamental Research Funds for the Central Universities (to Hua Hu) No.XDJK2016C184

References


