Question Similarity Calculation for FAQ Answering

Wanpeng Song[1,2,3] Min Feng[2] Naijie Gu[1,3] Liu Wenyin[2,3]
[1]Department of Computer Science and Technology, University of Science & Technology of China, Hefei, China
[2]Department of Computer Science, City University of Hong Kong, Hong Kong, China
[3]Joint Research Lab of Excellence, CityU-USTC Advanced Research Institute, Suzhou, China
wadeswp@mail.ustc.edu.cn
{emmwcity, csliuwy}@cityu.edu.hk

Abstract

Frequently Asked Question (FAQ) answering is a very useful module in automatic question answering (QA) systems where calculation of question similarity is a key problem. In this paper, we propose a new method for measuring the similarity between users' questions and the questions in a FAQ database. Both statistic measure and semantic information are employed. Statistic similarity is calculated based on dynamically formed vectors while semantic similarity is calculated by utilizing word similarity based on WordNet. Overall similarity is a combination of statistic similarity and semantic similarity. Preliminary results show that our method achieves a good performance.

1. Introduction

Question answering (QA) has become a popular Web-based service. Unlike search engines which return a few relevant documents, QA systems give one or several exact answers for each user question, which is more preferable. For these QA systems, FAQ answering is a useful aid to improve the efficiency of the whole system.

Unlike other automatic QA systems which focus on generating new answers, FAQ answering module retrieves answers from existing question-answer pairs in the database which were previously posted. When a new question is asked by a user, it is sent to the FAQ answering module to be compared with frequently asked questions (FAQs). If a similar question is found, the corresponding correct answer is returned immediately to the user as the final answer. Within the whole procedure, question similarity calculation is the key step which mainly determines the answer quality.

There have been many research works on the FAQ answering systems. FAQ Finder [1] heuristically combines statistical similarity and semantic similarity between users' questions and FAQs. Auto-FAQ [2] applies shallow language understanding into automatic FAQ answering, where the matching of a user question to FAQs is based on keyword comparison enhanced by limited language processing techniques. Sneiders [3] proposed a template-based FAQ retrieval system. FALLQ [4] used case-based knowledge for FAQ answering. Berger et al. [5] proposed a statistical lexicon correlation method. User click log has also been used to find similar queries in [6].

The rest of the paper is organized as follows. In Section 2, we give a brief introduction to simple NLP. The proposed method for question similarity calculation is presented in detail in Section 3. Section 4 shows the experiment result. Finally, we draw a conclusion in Section 5.

2. Simple NLP

Before we calculate the similarity between questions, we first conduct simple Natural Language Processing (NLP) which mainly includes three parts: POS tagging, stemming and stop words pruning.

POS tagging provides the POS information which is necessary for searching in the semantic dictionary WordNet [7]. It is also useful to help us find similar word when we compute the similarity between words. In our method, we use the rule-based tagger [8] for POS tagging.

Stemming is used to reduce inflected words to their root forms. After stemming, two words sharing the same root can be recognized as one word even though they are in different morphemes when we calculate the semantic similarity.

Stop words refer to the words which are useless in the process. Stop words pruning is useful when we calculate the question similarity. All stop words such as "the, a, an", are meaningless and should not be considered during the process of similarity calculation. We construct a stop word list which is slightly different from the normal one to do this pruning. Some words such as "is, are", which are in normal stop word list, is not contained in our stop word list because we think they imply some structure information.

3. Question similarity calculation

The proposed method in this paper for question similarity calculation is a linear combination of statistic similarity and semantic similarity. For the statistic similarity, we consider the word co-occurrence; for the semantic similarity, we utilize the word semantic similarity. At first, we calculate the two parts respectively, and then we do a linear combination. Unlike other methods, our method calculates statistic similarity based on dynamically formed vectors which avoids the sparse vector space problem. Semantic similarity is calculated by using a bipartite mapping. To calculate the semantic similarity, we not only map users' questions to the FAQs but also map the FAQs to users' questions. In this way, the semantic relationship is better described.

3.1. Statistic similarity

Most traditional document similarity measures represent document using high dimensional space, each word in the collection of documents as one dimension. This method works well in large document retrieval applications because the documents share a lot of words. However, it is not suitable for the similarity calculation between questions. Since there are only a few words in one sentence, if we use the high dimension representation method the vector space will be very sparse which will lead to a poor performance. Hence, we represent a question by using a low dimensional vector which is formed based on the word set of the two compared questions instead of the word set of all questions. In this way, we can avoid the sparse vector space problem.

Given two questions, Q_1 and Q_2, we define the word set as follows:

$$QS = Q_1 \cup Q_2$$

The word set QS contains all the distinct words in Q_1 and Q_2. Each question can be represented by a vector v. The vector's dimensionality is equal to the word number of QS. Each component in the vector represents the corresponding word in QS. The value of each component is determined as follows:

1. If the word do not appear in this question, the value of the component is set to 0;
2. If the word appears in this question, the value of the component is equal to its frequency in the question.

For example, consider the following two questions:

Q_1: What is the color of rose?

Q_2: What color is the Yangtze River?

After stemming and pruning the stop words, the word set is formed as follows:

QS = {what, is, color, of, rose, Yangtze, river}.

Notice that, "is" and "of" are not in our stop word list because they contain some structure information.

Therefore the dimension of the vector is 7. The two vectors are as follows:

$$v_1 = (1,1,1,1,1,0,0)$$

$$v_2 = (1,1,1,0,0,1,1)$$

Given two vectors, we can measure the statistic similarity of Q_1 and Q_2 by calculating their cosine product:

$$Sim_{statistic} = \frac{v1 \cdot v2}{\|v1\| \cdot \|v2\|}, \qquad (1)$$

where $Sim_{statistic}$ is the statistic similarity of Q_1 and Q_2.

3.2. Semantic similarity

Semantic similarity computing between words is often based on semantic knowledge base. Here we use WordNet [7] to calculate the word semantic similarity. WordNet [7] is structured as a hierarchical net where words are organized into synonym sets (synsets) with semantic pointers to other synsets. In WordNet [7], we can see that if two words are near to each other, they are more semantically similar, while they are less semantically similar if they are far from each other. Hence, we use the minimum length of path [9] to measure the similarity of two words. The length of a path ranges from 0 to infinite while the range of similarity is [0, 1]. When the path length is zero (i.e. two words are in the same synset), we set the similarity as 1; when the path length increases to infinite, the similarity should monotonically decrease to 0. Given two words, w_1 and w_2, we use the following formula to calculate the semantic similarity of two words:

$$sim(w_1, w_2) = \frac{1}{dis(w_1, w_2)+1}, \qquad (2)$$

where $sim(w_1, w_2)$ is the semantic similarity of w_1 and w_2, $dis(w_1, w_2)$ is the path length between them. In this way, the similarity decreases as the distance increases.

Based on the word semantic similarity, we can define the semantic similarity between two questions by summing up the semantic similarities between the words they have. We only consider content words and prune all function words when calculating semantic similarity. This is for two reasons. Firstly, function words have little meaning, which is useless in semantic similarity measuring. Secondly, searching in semantic dictionary is very time-consuming.

We calculate the semantic similarity of two questions Q_1 and Q_2 by using a bipartite mapping. Firstly, Q_1 is mapped to Q_2. Then the same process is done in the inverse direction. In mathematics, the semantic similarity score of the two questions is defined by the following formula:

$$Sim_{semantic} = \frac{1}{2}\left(\frac{\sum_{a_i \in Q_1} \max ssim(a_i, Q_2)}{|Q_1|} + \frac{\sum_{b_j \in Q_2} \max ssim(b_j, Q_1)}{|Q_2|}\right), \qquad (3)$$

where $Sim_{semantic}$ is the semantic similarity between question Q_1 and question Q_2, $|Q_1|$ and $|Q_2|$ are the numbers of content words in Q_1 and Q_2 respectively, and $\max ssim(a_i, Q_2)$ and $\max ssim(b_j, Q_1)$ are defined as follows:

$$\max ssim(a_i, Q_2) = \max(sim(a_i, b_1), sim(a_i, b_2), \ldots\ldots sim(a_i, b_{|Q_2|})), \qquad (4)$$

$$\max ssim(b_j, Q_1) = \max(sim(b_j, a_1), sim(b_j, a_2), \ldots\ldots sim(b_j, a_{|Q_1|})), \qquad (5)$$

where a_i and b_j are the content words in Q_1 and Q_2 respectively.

3.3. Overall similarity

With the statistic similarity and semantic similarity calculated, we can calculate the overall similarity between two questions by a linear combination:

$$Sim_{overall} = (1-\delta)Sim_{statistic} + \delta Sim_{semantic}, \qquad (6)$$

where $Sim_{overall}$ is the overall similarity constituted by statistic similarity and semantic similarity, and δ is a constant between 0 and 1 which decides the contribution of semantic similarity.

4. Experiments

To conduct the experiment, we get 500 question-answer pairs as FAQs from our QA system - BuyAns [12]. Most of these questions are factoid questions, like “What is the color of rose?” while the others are not, like “How to exercise to keep ourselves fit?” The length of these questions is from 5 words to 18 words and most of them are about 10. Based on these questions, we manually construct 58 questions by the following four approaches: (1) the same wording with different meaning, (2) different wording with the same meaning, (3) similar structure with different words, and (4) sub-words substituting. We use these 58 questions as users’ questions in our experiment.

The completeness [10] of the FAQ set is the factor which could influence the system performance. If there is no question which is similar to the user’s question in the FAQ set, we cannot obtain a correct answer. To eliminate the influence of the completeness problem, we only consider the questions which have similar questions in FAQs and those with no similar questions in the FAQs are ignored. Within the constructed 58 questions, there are 42 questions which have similar questions in the FAQs.

In the experiment, we calculate the similarities between a user’s question and each question in the FAQ database. We find the most similar question and return the corresponding answer as the final answer. The value of the parameter δ in Eq. (6) is also set to 0.7 empirically.

For performance evaluation, we use success at *n* (S@*n*) [11] as the performance metrics, which means the proportion of queries for which a correct answer is within the top *n*. Since the aim of FAQ answering is to find the most similar question, we use S@1 in our experiment. For example, S@1=50% means that the correct answer is at rank 1 for 50% of the queries,.

We evaluate the performance of our method with different value of parameter δ and compare our combined similarity measure with the statistic similarity measure and semantic similarity measure. Table 1 shows the experimental results using our combined method with different value of δ and Table

2 shows the experimental results using different measure.

Table 1. S@1 using different value of parameter

	δ =0.5	δ =0.6	δ =0.7	δ =0.8	δ =0.9
S@1	59.5%	61.9%	64.3%	52.4%	57.1%

Table 2. Experimental results using different method

	S@1
Statistic	50.0%
Semantic	57.1%
Combined	64.3%

The experimental results show that our method can achieve a good performance by combining the statistic similarity and the semantic similarity.

5. Conclusions and further work

In this paper, we propose a new method for question similarity calculation, which combines statistic similarity and semantic similarity. The statistic similarity is calculated based on the dynamically formed vectors. The semantic similarity is calculated based on WordNet [8] by utilizing the semantic similarity between words in the two compared questions. The experimental results show that our method achieves a good performance.

In our further work, we will consider the question type information and use it to help us to measure the similarity between questions.

6. Acknowledgement

The work described in this paper was fully supported by a grant from City University of Hong Kong (Project No. 7002137), the National Grand Fundamental Research 973 Program of China under Grant No.2003CB317002, and the Ph.D. Programs Foundation of Ministry of Education of China under Grant No.20060358011.

7. References

[1] R. Burke, K. Hammond, V. Kulyukin, S. Lytinen, N. Tomuro, and S. Schoenberg, “Question Answering from Frequently Asked Question Files”, *AI Magazine*, 18(2):57--66, 1997.

[2] S. D. Whitehead, “Auto-FAQ: an Experiment in Cyberspace Leveraging”, *Computer Networks and ISDN Systems*, 28(1-2):137-146, 1995.

[3] E. Sneiders, “Automated Question Answering using Question Templates that Cover the Conceptual Model of the Database”, *Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems-Revised Papers*, pages 235－239, 2002.

[4] M. Lenz, A. Hbner, and M. Kunze, “Question Answering with Textual CBR”, *Proceedings of the International Conference on Flexible Query Answering Systems* (Denmark), pages 236–247, 1998.

[5] A. Berger, R. Caruana, D.Cohn, D. Freitag, and V. Mittal, “Bridging the Lexical Chasm: Statistical Approaches to Answer-Finding”, *Proceedings of SIGIR*, pages 192--199, 2000.

[6] H. Kim, J.Y. Seo, “High-Performance FAQ Retrieval using an Automatic Clustering Method of Query Logs”, *Information Processing and Management*, 42(3):650-661, 2006.

[7] C. Fellbaum, WordNet: An Electronic Lexical Database, the MIT Press, 1998.

[8] E. Brill, “A simple rule-based part of speech tagger”, *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.

[9] R. Rada, H. Mili, E. Bichnell, and M. Blettner, “Development and Application of Metric on Semantic Nets”, *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):17-30, Jan, 1989.

[10] E. Sneiders, “Automated FAQ Answering on WWW Using Shallow Language Understanding”, thesis in partial fulfillment of the requirements for the degree of Licentiate of Technology, Dept. of Computer and Systems Sciences, Stockholm University / Royal Institute of Technology, Sweden, 1999.

[11] D. Hawking, N. Craswell, Very Large Scale Retrieval and Web Search, the MIT Press, 2005.

[12] Buyans. http://www.buyans.com/