An Answer Passage Retrieval Strategy for Web-based Question Answering

Work in progress

Xin Li[1, 2, 3], Dawei Hu[1, 2, 3], Tianyong Hao[3], Enhong Chen[1, 2], Liu Wenyin[2, 3]
[1]Department of Computer Science and Technology, University of Science & Technology of China, Hefei, China
[2]Joint Research Lab of Excellence, CityU-USTC Advanced Research Institute, Suzhou, China
[3]Department of Computer Science, City University of Hong Kong, Hong Kong, China

xinli@mail.ustc.edu.cn, dwhu@mail.ustc.edu.cn, tianyong@cityu.edu.hk
ehchen@ustc.edu.cn,csliuwy@cityu.edu.hk

ABSTRACT

A passage retrieval strategy for our web-based Question Answering (QA) system is proposed in this paper. We utilize *Google* to retrieve web documents for answer passage finding. We propose a new method to rewrite the query for passage retrieval. We calculate the relevancy between the query and the passage by combining the term frequency and semantic relevancy. The method has been found effective in the experiment on factoid questions of TREC 2003.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval– *retrieval models, search process*.

General Terms

Algorithms, Experimentation

Keywords

Question Answering, Passage Retrieval, Semantic Pattern

1. INTRODUCTION

Most research on question answering is to build an open-domain question answering system, which can return exact answers for questions, instead of a list of documents [6]. We build a web-based QA system to utilize World Wide Web (WWW) as a knowledge source for question answering, because it has tremendous amount of freely available online information. Passage retrieval is added as an intermediate stage between document retrieval and answer extraction in order to reduce the text size to be processed [5]. Non-relevant candidate passages are often retrieved by traditional density-based or language model based methods because they ignore the constraint relations between words in a phrase or neighborhood. We use a semantic pattern model proposed by *Hao et al*. [1] for users to submit questions and then analyze the question to obtain its question target and keywords. This information is used to optimize queries and to calculate the relevancy between the query and the passage, which involves both the *tf-idf* likelihood and semantic relevancy.

2. OVERVIEW OF OUR SYSTEM

Our web-based QA system has four components: question analysis module, document retrieval module, passage retrieval module, and answer extraction module, as shown in Figure 1. The system firstly makes an analysis of a question. One or more queries are then formed and are submitted to the search engine to retrieve relevant documents. The documents are divided into passages. Relevant passages are retrieved by the passage retrieval module and answers are extracted from the passages. In this paper, we will focus on examining the performance of our passage retrieval strategy.

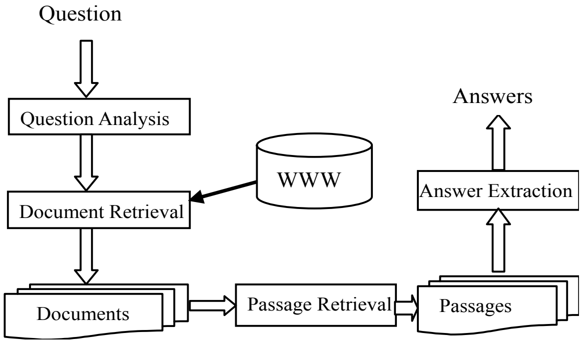

Figure 1. Architecture of our web-based QA system

3. PASSAGE RETRIEVAL

We firstly retrieve a set of web documents using *Google* and then extract the plain text of the web documents. Passages are formed by adjacent sentences and the number of sentences is no more than three, because longer passages contain more extra information which may improve the difficulty of analyzing and increase the possibility to return error answers. We focus on factoid questions and use a Semantic Pattern Matching (SPM) [1] method for question analysis in order to obtain the question target, keywords and the relation between keywords. We then utilize this information to optimize queries for passage retrieval.

INFOSCALE 2007, June 6-8, Suzhou, China
Copyright © 2007 ICST 978-1-59593-757-5
DOI 10.4108/infoscale.2007.217

3.1 Query Rewriting for Passage Retrieval

We apply several heuristics to rewrite the queries for passage retrieval.

a) When a word is a noun, it and its immediate modifier must be taken together as a keyword. For example, given question "*what is the longest river in china?*", we take "*longest river*" together as a keyword. This heuristics is also utilized when we calculate the frequency of keywords of a passage.

b) When an adjective or an adverb follows an interrogative "how", the adjective or adverb together with "how", are transformed to a noun that relevant to the question's semantic category. For example, for question "*how far is it from Earth to Mars?*", its question type is DISTANCE, so we transform "*how far* " to "*distance*" .

c) When a noun follows an interrogative "what", the noun is removed. Take question "*What country is Aswan High Dam located in?*", "*country*" is removed from the query.

d) When a predicate verb is followed by a preposition to form a phrase, the verb and the preposition are remained together. An example is that: for question "*What are pennies made of?*", "*made of*" is taken together as a keyword.

A keyword may have several different morphological forms such as noun plurality, verb preterits; we use a famous stemmer Porter's stemmer [3] to transform them to their stemmed form. For example, "made" is transformed to "make", "books" is transformed to "book" and "killed" to "kill". The same transform is also performed on the words of the passages.

3.2 Passage Scoring

Given a query, the score of a passage consist of two parts: the first is the *tf-idf* likelihood score and the second is the semantic relevancy score between the query and the passage. The *tf-idf* likelihood score, as shown in Eq. (1), is a vector space model proposed in [4], where the term frequency of query and passage is involved. However, in our experiment, we do not take into account the frequency of query terms because it is not necessary for passage retrieval. The semantic relevancy score, as shown in Eq. (2), is to calculate the number of the semantic relevant words between the query and the passage. The score of passage is calculated using Eq. (3).

$$Score_1(p,q) = \sum_{t \in p \cap q} \frac{1+\ln(1+\ln(tf(t,p)))}{(1-s)+s\frac{|p|}{avgpl}} \times \ln\frac{N+1}{df(t)} \qquad (1)$$

$$Score_2(p,q) = \sum_{t \in p} \delta(t) \qquad (2)$$

$$Score(p,q) = Score_1(p,q) + \alpha \times Score_2(p,q) \qquad (3)$$

p: the passage
q: the query
$tf(t,p)$: the frequency of term t occurs in passage p
s: a parameter to balance the length of passages
$|p|$: the length of passage p

N: the number of candidate passages for q

$df(t)$: the number of passages that contain term t

$avgpl$: the average length of passages

$\delta(t)$: if term t has a semantic relevancy word in query q, then the value is 1, otherwise 0.

α : the balance parameter between $score_1$ and $score_2$.

The semantic relevancy is defined as follow: for word w_1 and word w_2, if w_1 is a synonym, hypernyms or hyponyms of w_2 in WordNet [2], then they have a semantic relevancy.

4. EXPERIMENT RESULT

We select factoid questions from TREC-12 [6] for passage task as our test set. Table 1 shows the experiment result: The MRR for *what-type* question is 0.3135. The MRR value for *what-type* question is better than that of how-*type* and *when-type* question, whose MRR values are 0.2827 and 0.2416 respectively. The percentages of questions with no answer passage for the questions of the three types are 52.4%, 21.4%, and 50%, respectively. The overall MRR for all test questions is 29.8% and the percentage of questions with no answer passage is 43.5 %.

Table 1. The number of questions with correct or relevant passages and MRR

Rank	1	2	3	4	5	NIL	MRR
what	28	11	7	9	4	65	0.314
How	5	3	10	16	10	12	0.283
when	2	2	4	2	0	10	0.242

5. CONCLUSIONS AND FUTURE WORK

We proposed a passage retrieval strategy, which includes a query optimization method and a passage scoring function. Experiment results show that our passage retrieval strategy performs well for factoid questions. Anyway, we are still improving our method and hope to implement it in *BuyAns* [7], which is our user-interactive QA system, to automatically answer users' questions based on the passages found from the Web.

6. ACKNOWLEDGMENTS

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 Research Fund of Microsoft Research Asia.

7. REFERENCES

[1] Hao, T.Y., Zeng, Q. T., Liu, W. Y.Semantic Pattern for User-Interactive Question Answering. Proc. of Second International Conference on Semantics, Knowledge, and Grid (SKG'06), 2006.

[2] Miller, G.A. WordNet: A Lexical Database. Communication of the ACM, vol 38: No11, pp 39-41, 1995.

[3] Porter, M.F. An algorithm for Suffix Stripping. Program, 1980, 14(3), 130-137.

[4] Singhal, A. Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 24(4):35-43, 2001.

[5] Tellex, S., Katz, B., Lin, J., Fernandes, A. and Marton, G. Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering. Proc. of SIGIR' 03, 2003.

[6] Voorhees, E.M. Overview of the TREC 2003 Question Answering Track. Proc. of TREC-12, 2003.

[7] http://www.buyans.com/