A Web-based Question Answering System

Dell Zhang and Wee Sun Lee

***Abstract*— The Web is apparently an ideal source of answers to a large variety of questions, due to the tremendous amount of information available online. This paper describes a Web-based question answering system LAMP, which is publicly accessible. A particular characteristic of this system is that it only takes advantage of the snippets in the search results returned by a search engine like Google. We think such "snippet-tolerant" property is important for an online question answering system to be practical, because it is time-consuming to download and analyze the original web documents. The performance of LAMP is comparable to the best state-of-the-art question answering systems.**

***Index Terms*— question answering, information retrieval, search engine, web mining.**

I. Introduction

WHAT a current information retrieval system or search engine can do is just "document retrieval", i.e., given some keywords it only returns the relevant documents that contain the keywords. However, what a user really wants is often a precise answer to a question. For instance, given the question "Who was the first American in space?", what a user really wants is the answer "Alan Shepard", but not to read through lots of documents that contain the words "first", "American" and "space" etc.

The Text Retrieval Conference, TREC (http://trec.nist.gov/), has launched a QA track to support the competitive research on question answering, from 1999 (TREC8). The focus of TREC QA track is to build a fully automatic open-domain question answering system, which can answer factual questions based on very large document. Today, the TREC QA track [7,8,9] is the major large-scale evaluation environment for open-domain question answering systems.

The Web is apparently an ideal source of answers to a large variety of questions, due to the tremendous amount of information available online. This paper describes a Web-based question answering system LAMP, which is publicly accessible (http://hal.comp.nus.edu.sg/cgi-bin/smadellz/lamp_query.pl). A particular characteristic of this system is that it only takes advantage of the snippets in the search results returned by a search engine like Google (http://www.google.com/). We think such "snippet-tolerant" property is important for an online question answering system to be practical, because it is time-consuming to download and analyze the original web documents. The performance of LAMP is comparable to the best state-of-the-art question answering systems.

Manuscript received October 31, 2002.

Dell Zhang is with the Department of Computer Science, School of Computing, S15-05-24, 3 Science Drive 2, National University of Singapore, Singapore 117543, and Singapore-MIT Alliance, E4-04-10, 4 Engineering Drive3, Singapore 117576 (e-mail: dell.z@ieee.org).

Wee Sun Lee is with the Department of Computer Science, School of Computing, S15-05-24, 3 Science Drive 2, National University of Singapore, Singapore 117543, and Singapore-MIT Alliance, E4-04-10, 4 Engineering Drive3, Singapore 117576 (e-mail: leews@comp.nus.edu.sg).

II. System

A. Overview

The system architecture is depicted in Fig. 1. Given a user's natural language question, the system will submit the question to a search engine, then extract all plausible answers from the search results according to the question type identified by the question classification module, finally select the most plausible answers to return. A screen snapshot of the system is shown in Fig. 2.

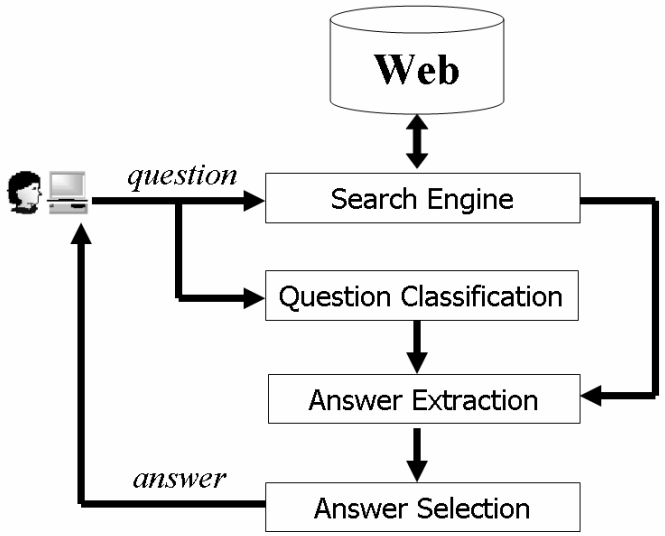

Fig. 1. The system architecture.

To illustrate our approach, we would like to use the question "Who was the first American in space?" as a running sample. This question was the No.21 test question

in the TREC8 QA track, and has been used as a running instance in [5] and [6] as well.

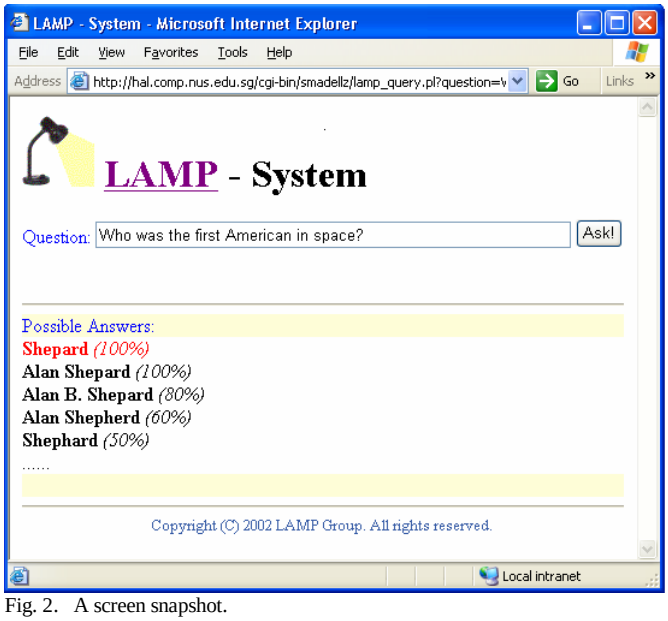

Fig. 2. A screen snapshot.

B. *Search Engine*

The system submits the question to the search engine Google (http://www.google.com/) and grabs its top 100 search results. Part of the search results of Google for the sample question are shown in Fig. 3.

Fig. 3. Some search results of Google for the sample question.

Each search result usually contains the URL, the title, and some string segments of the related web document. We call these title and the string segments in the search results "snippets". The snippets in the above sample search results are shown in Fig. 4.

The system only takes advantage of the snippets in the search results, because it is time-consuming to download and analyze the original web documents. We think such "snippet-tolerant" property is important for an online question answering system to be practical,

> Chronology of Selected Highlights in the First 100 American.
>
> a range of 302 miles.
>
> It was the first American space flight involving human beings.
>
> Shepard demonstrated that individuals can control.
>
> This New Ocean - Ch13-4.
>
> 38.
>
> Over the Indian Ocean on his first orbit, Glenn became the first American.
>
> the astronaut described the moment of twilight simply as "beautiful."
>
> Space.
>
> Sally Kristen Ride | First American Woman in Space.
>
> Sally Kristen Ride First American Woman in Space.
>
> Born: May 26, 1951.
>
> Our future lies with today's kids and tomorrow's space exploration.
>
> Dr. Sally Ride.
>
> ENC : This Week : Classroom Calendar : First American in Space.
>
> Privacy, Copyright & Disclaimer.
>
> First American in Space (Grades K-12), May 5. Keep in mind that Alan Shepard was the first American in space, not the first man.
>
> .....

Fig. 4. The snippets in the above sample search results.

C. *Question Classification*

In order to correctly answer a question, usually one needs to understand what type of information the question asks for, e.g., the sample question "Who was the first American in space?" asks for a person name. Here we only consider the factual questions, i.e., TREC-style questions.

The system utilizes a Support Vector Machine (SVM) [3] to classify the questions. While trained by about 6,000 labeled questions, the question classification SVM can achieve above 90% accuracy.

D. *Answer Extraction*

After the question type has been identified, the system extracts all such type information from the snippets as plausible answers, using a HMM-based named entity recognizer [2] as well as some heuristics rules. For instance, some plausible answers to the sample questions are shown in Fig. 5, they are just all the person names occurred in the snippets.

Chronology of Selected Highlights in the First 100 American.

a range of 302 miles.

It was the first American space flight involving human beings.

Shepard demonstrated that individuals can control.

This New Ocean - Ch13-4.

38.

Over the Indian Ocean on his first orbit, ***Glenn*** became the first American.

the astronaut described the moment of twilight simply as "beautiful."

Space.

Sally Kristen Ride | First American Woman in Space.

Sally Kristen Ride First American Woman in Space.

Born: May 26, 1951.

Our future lies with today's kids and tomorrow's space exploration.

Dr. ***Sally Ride***.

ENC : This Week : Classroom Calendar : First American in Space.

Privacy, Copyright & Disclaimer.

First American in Space (Grades K-12), May 5. Keep in mind that ***Alan Shepard*** was the first American in space, not the first man.

……

Fig. 5. The plausible answers extracted from the above sample snippets.

E. Answer Selection

For each plausible answer, the system constructs a snippet cluster which is composed of all the snippets containing that answer. Moreover, the snippet clusters of different answers referring to the same entity should be merged into one. For example, it is obvious that the two plausible answers "Sally Kristen Ride" and "Sally Ride" are two variants of the same person name, so their snippet clusters will be merged. The snippet clusters constructed from the above sample snippets are shown in Fig. 6.

The snippet cluster of a plausible answer describes its occurring context. We take the snippet cluster as the feature of the corresponding plausible answer, and represent it as a vector $\mathbf{a} = (a_1, a_2, ..., a_n)$, where a_i is the number of all words and n is the occurring frequency of the i-th word. The question is also represented as a vector $\mathbf{q} = (q_1, q_2, ..., q_n)$ in the same way.

The standard Vector Space Model in IR (Information Retrieval) area uses the cosine value of the angle between the query and document vectors, i.e., the inner product between the normalized query and document vectors, to measure their relevance [1]. However, we found that only the angle information is not good enough for this application.

It has been discovered that the correct answer to a question usually occurs more often than the incorrect ones on the search results of that question [4]. Maybe the reason is that the correct factual statements have more chance to be replicated. So the size (norm) of the answer vector should also be considered for answer selection. This has been validated by our experiments.

We propose to use the following score function which has incorporated both the angle and the norm information to rank the plausible answers,

$$score(\mathbf{q},\mathbf{a}) = \|\mathbf{a}\|\cos\theta = \frac{\mathbf{q}\bullet\mathbf{a}}{\|\mathbf{q}\|} = \frac{\sum_i q_i a_i}{\sqrt{\sum_i (q_i)^2}}$$

where $\mathbf{q}$ is the question vector, $\mathbf{a}$ is the answer vector, and θ is the angle between them. Actually the value of $score(\mathbf{q},\mathbf{a})$ is just the length of the "projection" of $\mathbf{a}$ on $\mathbf{q}$, as shown in Fig. 7.

Shepard demonstrated that individuals can control.

First American in Space (Grades K-12), May 5. Keep in mind that ***Alan Shepard*** was the first American in space, not the first man.

……

Over the Indian Ocean on his first orbit, ***Glenn*** became the first American.

……

Sally Kristen Ride | First American Woman in Space.

Sally Kristen Ride First American Woman in Space.

Dr. ***Sally Ride***.

……

……

Fig. 6. The snippet clusters constructed from the above sample snippets.

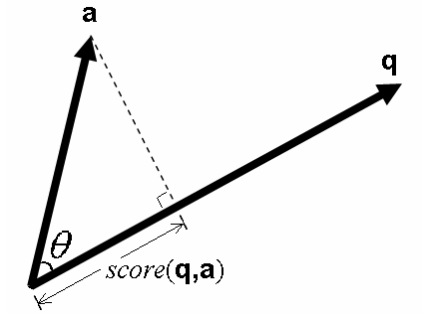

Fig. 7. The geometrical meaning of $score(\mathbf{q},\mathbf{a})$.

III. EVALUATION

Several experiments have been done using the dataset from TREC QA track. The questions with typo mistakes, the definition style questions like "Who is Colin Powell?", the questions which are syntactic rewrites of earlier

questions (TREC9 test questions No.701-893), and the questions with no associated answer regular expressions have been removed from the datasets. Note all the Web search results were retrieved from Google in the period Sep -- Oct 2002.

It turns out that the answers to most of the TREC questions can be found in the snippets, as shown in Table I, where q# means the number of test questions, and w# means the number of questions whose correct answer can be found in the snippets of Google's top 100 search results. The abundance and variation of data on the Web allows the system to find correct answers in high probability, because the factual knowledge is usually replicated across the Web in different expressing manners.

TABLE I
HOW MANY ANSWERS TO TREC QUESTIONS CAN BE FOUND IN THE SNIPPETS

datasets	q#	w#	percentage
TREC8	196	132	67.3%
TREC9	438	345	78.8%
TREC10	312	263	84.3%
TREC11	444	373	84.0%
total	1390	1113	80.1%

In TREC8, TREC9 and TREC10 QA tracks [7,8,9], a question answering system is required to return 5 ranked answers for each test question, the results are evaluated by the MRR metric. MRR stands for "Mean Reciprocal Rank", it is computed as $MRR = \sum_{i=1}^{n} \frac{1}{r_i}$, where n is the number of test questions and r_i is the rank of the first correct answer for the i-th test question.

In TREC11 QA track, a question answering system is required to return only one exact answer for each test question, and all answers returned should be ordered by the system's confidence about their correctness, the results are evaluated by the CWS metric. CWS stands for "Confidence Weighted Score", it is computed as $CWS = \sum_{i=1}^{n} p_i \Big/ n$, where n is the number of test questions and p_i is the precision of the answers at positions from 1 to i in the ordered list.

The performance of this system has been evaluated using the test questions from TREC11. The MRR and CWS scores are shown in Table II, where q# means the number of test questions. The relationship between the precision of answers and their ranks in the returned answer list is shown in Fig. 8.

TABLE II
THE MRR AND CWS OF THE SYSTEM ON TREC11 QUESTIONS

question type	q#	MRR	CWS
PERSON	74	0.64	0.75
ORGANIZATION	15	0.51	0.61
LOCATION	101	0.53	0.63
DATE	95	0.69	0.81
QUANTITY	60	0.24	0.28
PROPERNOUN	53	0.35	0.55
OTHER	46	0	0
total	444	0.47	0.60

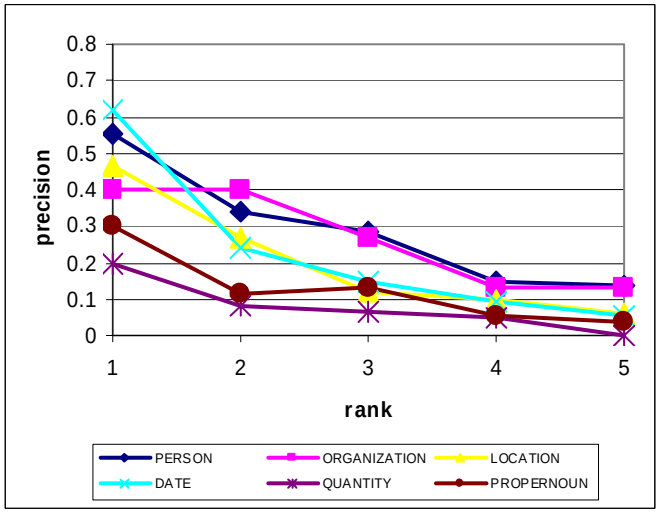

Fig. 8. The relationship between the precision of answers and their ranks.

The MRR score of LAMP is not as high as that of the best question answering system in TREC. This discrepancy is due to many reasons. One important factor is that the answer regular expressions provided by TREC are quite limited, many correct answers such as "Alan B. Shepard, Jr." are judged wrong since they do not occur in the TREC specified document collection. Another interesting issue is time, the correct answers to some questions like "Who is the U.S. president?" will change over time. The Web is also messier than the TREC document collection.

LAMP performs very well on some types of questions such as PERSON, LOCATION, and DATE, this observation suggests us to try the "divide and conquer" strategy in the future.

IV. RELATED WORK

Table III compares the feature of LAMP with other state-of-the-art question answering systems.

Here TREC QA systems are the systems dedicated to the TREC QA track tasks, including Qanda, Falcon, Webclopedia, AskMSR, Insight, etc. [7,8,9]. These QA systems have to find answers from a large local news text corpus. And, these QA systems usually run in offline mode, because they have about one week time to submit their results for several hundred test questions.

TABLE III
THE FEATURES OF LAMP AND OTHER STATE-OF-THE-ART QUESTION ANSWERING SYSTEMS

System	Data Source	Result Format	Open Domain	Web Accessible
TREC8,9,10	local documents	fixed-length string segments	Y	N
TREC11	local documents	exact answers	Y	N
START	miscellaneous	miscellaneous	N	Y
QuASM	online databases	named-entities & passages	N	Y
SiteQ/E	several websites	named-entities & passages	Y	Y
IONAUT	Web documents	named-entities & passages	Y	Y
AskJeeves	Web documents	URLs	Y	Y
AnswerBus	Web documents	sentences	Y	Y
Mulder	Web documents	exact answers	Y	N
NSIR	Web documents	exact answers	Y	N
LAMP	Web search results	exact answers	Y	Y

The following systems are all online, and publicly accessible on the Web. However, they are still not ready to return the exact answers for the questions.
START (http://www.ai.mit.edu/projects/infolab/)
QuASM (http://ciir.cs.umass.edu/~reu2/)
SiteQ/E (http://ressell.postech.ac.kr/~pinesnow/siteqeng/)
IONAUT (http://www.ionaut.com:8400/)
AskJeeves (http://www.ask.com/)
AnswerBus (http://misshoover.si.umich.edu/~zzheng/qa-new/)

The question answering systems closest to LAMP are Mulder [5] and NSIR [6], which were published on the recent World Wide Web conferences (WWW2001 and WWW2002).

Both Mulder and NSIR are claimed to be Web accessible. However, they are actually not available while we write this paper. Both Mulder and NSIR have to download and analyze the original Web documents, which are time-consuming. On contrast, LAMP only uses the snippets from the Web search result. This "snippet-tolerant" property makes LAMP system very efficient.

The performance of Mulder was not measured by MRR score, so it can not be compared directly. The MRR score of NSIR, over the 200 test questions from TREC8 QA track, is 0.15, which is lower than that of LAMP.

V. CONCLUSION

The Web is apparently an ideal source of answers to a large variety of questions, due to the tremendous amount of information available online. This paper describes a Web-based question answering system LAMP, which is publicly accessible. A particular characteristic of this system is that it only takes advantage of the snippets in the search results returned by a search engine like Google. We think such "snippet-tolerant" property is important for an online question answering system to be practical, because it is time-consuming to download and analyze the original web documents. The performance of LAMP is comparable to the best state-of-the-art question answering systems.

REFERENCES

[1] R. Baeza-Yates and B. Ribeiero-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
[2] D. Bikel, R. Schwartz, and R. Weischedel. "An Algorithm that Learns What's in a Name". *Machine learning*, 34(1-3) pp. 211--231, 1999.
[3] C. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.
[4] S. Dumais, M. Banko, E. Brill, J. Lin and A. Ng. "Web Question Answering: Is More Always Better?". In *Proceedings of SIGIR'02*, pp. 291-298, Aug 2002.
[5] C. Kwok, O. Etzioni, and D. S. Weld. "Scaling Question Answering to the Web". In *Proceedings of the 10th World Wide Web Conference (WWW2001)*, Hong Kong, 2001.
[6] D.R. Radev, W. Fan, H. Qi, H. Wu and A. Grewal. "Probabilistic Question Answering from the Web". In *Proceedings of the 11th World Wide Web Conference (WWW2002)*, Hawaii, 2002.
[7] E. Voorhees. "The TREC-8 Question Answering Track Report". In *Proceedings of the 8th Text Retrieval Conference (TREC8)*, pp. 77-82, NIST, Gaithersburg, MD, 1999.
[8] E. Voorhees. "Overview of the TREC-9 Question Answering Track". In *Proceedings of the 9th Text Retrieval Conference (TREC9)*, pp. 71-80, NIST, Gaithersburg, MD, 2000.
[9] E. Voorhees. "Overview of the TREC 2001 Question Answering Track". *In Proceedings of the 10th Text Retrieval Conference (TREC10)*, pp. 157-165, NIST, Gaithersburg, MD, 2001.

Dell Zhang is a research fellow in the National University of Singapore under the Singapore-MIT Alliance (SMA). He has received his BEng and PhD in Computer Science from the Southeast University, Nanjing, China. He is currently focusing on machine learning and information retrieval.
Wee Sun Lee is an Assistant Professor at the Department of Computer Science, National University of Singapore, and a Fellow of the Singapore-MIT Alliance (SMA). He obtained his Bachelor of Engineering degree in Computer Systems Engineering from the University of Queensland in 1992, and his PhD from the Department of Systems Engineering at the Australian National University in 1996. He is interested in computational learning theory, machine learning and information retrieval.