Web based Question Answering with Aggregation Strategy

Dell Zhang[1,2]

[1] Department of Computer Science, School of Computing, National University of Singapore,
S15-05-24, 3 Science Drive 2, Singapore 117543
[2] Computer Science Programme, Singapore-MIT Alliance
E4-04-10, 4 Engineering Drive 3, Singapore 117576
dell.z@ieee.org

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 question answering system LAMP, which can instantly find answers from the Web. LAMP is publicly accessible on the Web, and its performance is comparable to the state-of-the-art question answering systems. One novel characteristic of this system is that its answer selection method is based on aggregation, which is shown to be more effective than the common methods based on individual or redundancy.

1 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 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 question answering system LAMP[3], which can instantly find answers from the Web. LAMP is publicly accessible on the Web[4], and its performance is comparable to the state-of-the-art question answering systems. One novel characteristic of this system is that its answer selection method is based on aggregation, which is shown to be more effective than the common methods based on individual or redundancy.

2 System

[3] LAMP stands for Learning and Answering Program.

[4] http://hal.comp.nus.edu.sg/cgi-bin/smadellz/lamp_query.pl

Given a natural language question, the system transforms it into an appropriate query and submits the query to a search engine like Google[5], then extracts all plausible answers from the search results according to the question type identified by the question classification module, finally selects the most plausible answers to return.

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 [10], it has been used as a running sample in [8] and [9] as well.

2.1 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) [6] to classify the questions. While trained by about 6,000 labeled questions, the question classification SVM can achieve above 90% accuracy. Details of this algorithm and the experiments are included in another paper [13].

2.2 Query Formulation

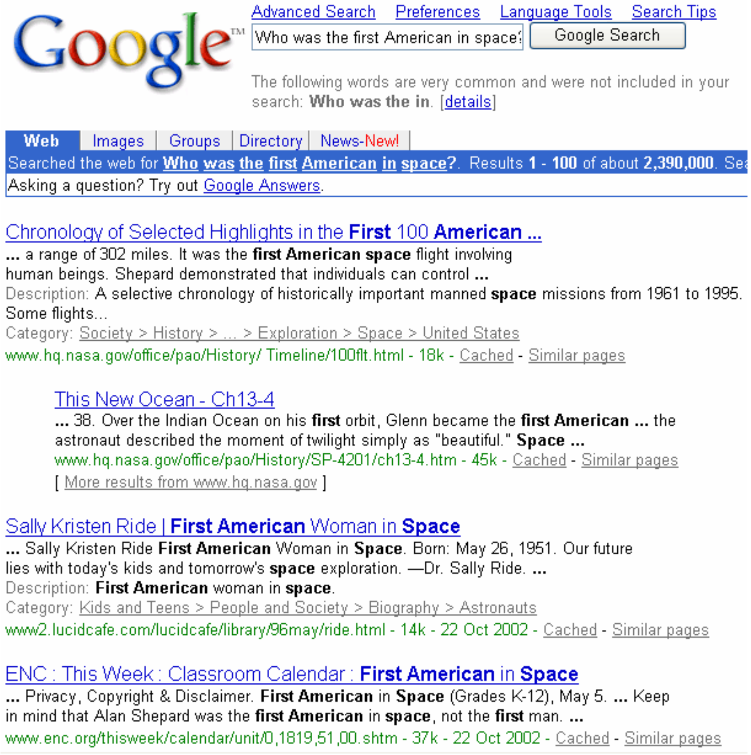
The question should be transformed into a specific query before being submitted to a search engine, since most search engines are not straightly designed for natural language questions. Using search engine specific queries instead of the raw question might significantly improve the effect of question answering [2]. We have used the following two methods for query formulation.

- Interrogative Words Deletion. Interrogative words such as “who”, “what”, “how much”, etc. usually do not occur in the target text. However, many interrogative words are not in the stop-word list of the search engine. Deleting interrogative words makes the query shorter, and often increases recall of the search engine without affecting precision. Currently we have about 70 regular expressions to recognize the interrogative words and automatically remove them from the query.
- Verb Form Conversion. For questions with an auxiliary do-verb and a main verb, the target sentence is likely to contain the verb in the conjugated form rather than separate verbs. For instance, the answer to the question “When did Nixon visit China?” would more likely to occur in the target text as “…… Nixon visited China ……” rather than “…… did Nixon visit China ......”. So we convert the main verb from the original form to the “third person singular” form if the auxiliary do-verb is “does”, or to the “past tense” form if the auxiliary do-verb is “did”. To correctly locate the main verb in the question, we parse the question using the MEI parser [5]. To know the specific form of a verb convert the main verb, we employ PC-KIMMO [1].

[5] http://www.google.com/

2.3 Search Engine

The system submits the question to the prestigious search engine Google and grabs its top 100 search results. A search result usually contains the URL, the title, and some string-segments of the related web document. We call these titles and the string-segments in the search results "snippets". 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 a "snippet-tolerant" property is important for an online question answering system to be practical.



Who was the first American in space?

The following words are very common and were not included in your search: **Who was the in**. [details]

Searched the web for **Who was the first American in space?**. Results **1 - 100** of about **2,390,000**. Sea

Asking a question? Try out Google Answers.

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 **...**
Description: A selective chronology of historically important manned **space** missions from 1961 to 1995. Some flights...
Category: Society > History > ... > Exploration > Space > United States
www.hq.nasa.gov/office/pao/History/ Timeline/100flt.html - 18k - Cached - Similar pages

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 ...**
www.hq.nasa.gov/office/pao/History/SP-4201/ch13-4.htm - 45k - Cached - Similar pages
[More results from www.hq.nasa.gov]

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. **...**
Description: **First American** woman in **space**.
Category: Kids and Teens > People and Society > Biography > Astronauts
www2.lucidcafe.com/lucidcafe/library/96may/ride.html - 14k - 22 Oct 2002 - Cached - Similar pages

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. **...**
www.enc.org/thisweek/calendar/unit/0,1819,51,00.shtm - 37k - 22 Oct 2002 - Cached - Similar pages

Fig. 1. Google search results for the sample question.

2.4 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 [4] as well as some heuristics rules.

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. 2. The snippets in the above sample search results and the extracted plausible answers (displayed in bold italic font).

2.5 Answer Selection

A snippet *S* containing a plausible answer A describes one occurring context of *A*. It can be represented as a bag-of-words feature vector $\mathbf{s} = (s_1, s_2, ..., s_n)$, where n is the number of all words and s_i is the occurring frequency of the *i*-th word in *S*. The question *Q* is also represented as a vector $\mathbf{q} = (q_1, q_2, ..., q_n)$ in the same way.

The traditional method for answer selection is based on individuals [8,9]. A snippet *S* in the search result is assessed individually by the similarity between $\mathbf{s}$ and $\mathbf{q}$. Then the plausible answers contained in the top few snippets are selected.

A recently emerged method for answer selection is based on redundancy [7]. A plausible answer *A* is accessed by how many times it occurred in the search result, i.e., how many snippets contain it. The underlying idea is that the correct answer to a question usually occurs more often than the incorrect ones on the search results. This assumption has been empirically proved to be true given the tremendous amount of information available on the Web [7].

Here we propose a novel answer selection method based on aggregation, which distinguishes our question answering system with others. For each plausible answer *A*, the system will aggregate all snippets containing *A* into a cluster C_A. Moreover, the snippet clusters of different answers referring to the same entity should be merged into one. For example, it is obvious that “Sally Kristen Ride” and “Sally Ride” are two variants of the same person name, so their snippet clusters should be merged.

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. 3. The snippet clusters constructed on the above sample snippets.

The snippet cluster C_A of a plausible answer A summarizes A’s occurring context. It can also be represented as a vector $\mathbf{a} = (a_1, a_2, ..., a_n)$, where n is the number of all words and a_i is the occurring frequency of the i-th word in C_A. In other words, $\mathbf{a} = \sum_{A\in S_k} \mathbf{s}_k$. We think $\mathbf{a}$, the feature vector of the snippet cluster C_A, may characterize its corresponding plausible answer A very well and can be used for answer selection.

The standard Vector Space Model in Information Retrieval area uses the cosine value of the angle between the query and document vectors to measure their similarity [3]. However, we think only the similarity information is not enough for our application. Since the correct factual statements have more chances to be replicated on the Web [7], the size (norm) of the answer’s feature vector which partially reflects its redundancy should not be neglected.

We propose to use the following score function to rank the plausible answers, $score(Q, A) = \|\mathbf{a}\| \cos\theta = (\mathbf{q}\cdot\mathbf{a}) \big/ \|\mathbf{q}\| = \left(\sum_i q_i a_i\right) \Big/ \sqrt{\sum_i q_i^2}$, where $\mathbf{q}$ is the feature vector of Q, $\mathbf{a}$ is the feature vector of C_A, and θ is the angle between them. This score function incorporates both similarity and redundancy information for answer selection. Actually the value of $score(Q, A)$ is the length of the “projection” of $\mathbf{a}$ on $\mathbf{q}$.

All the plausible answers are then ranked by their scores, and several most plausible answers will be returned to the user.

3 Evaluation

3.1 TREC

The Text Retrieval Conference, TREC[6], has launched a QA track to support the competitive research on question answering, from 1999 (TREC8). The focus of TREC-QA is to build fully automatic open-domain question answering systems, which can answer factual questions based on very large document collections. Today, TREC-QA [10,11,12] is the major large-scale evaluation environment for open-domain question answering systems.

Several experiments were conducted using the test questions and answer patterns dataset from TREC-QA. The questions with typo mistakes, the questions asking for definitions 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 patterns were discarded. In the following experiments, all the Web search results were retrieved from Google in Nov. 2002.

3.2 Web as Answer Source

It turns out that the answers to most of the TREC-QA questions can be found in the Web search results, as shown in Table 1, 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 Web data 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 1. How many answers to TREC questions can be found in the snippets.

dataset	q#	w#	percentage
TREC8	196	144	73.5%
TREC9	438	348	79.5%
TREC10	312	260	83.3%
TREC11	444	380	85.6%
total	1390	1132	81.4%

3.3 System Performance

In TREC8, TREC9 and TREC10 QA tracks [10,11,12], 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 $MRR = \sum_{i=1}^{n} (1/r_i)$, 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.

[6] http://trec.nist.gov/

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 $CWS = \left(\sum_{i=1}^{n} p_i \right) \Big/ n$, 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 2, 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. 4. The CWS score will place our system roughly at the 3rd position in TREC11 QA track.

Table 2. The MRR and CWS of the LAMP system on TREC11 questions.

type	q#	MRR	CWS
PERSON	74	0.72	0.83
ORGANIZATION	15	0.56	0.49
LOCATION	101	0.56	0.65
DATE	95	0.65	0.81
QUANTITY	60	0.22	0.26
PROPERNOUN	53	0.39	0.59
OTHER	46	0	0
total	444	0.48	0.62

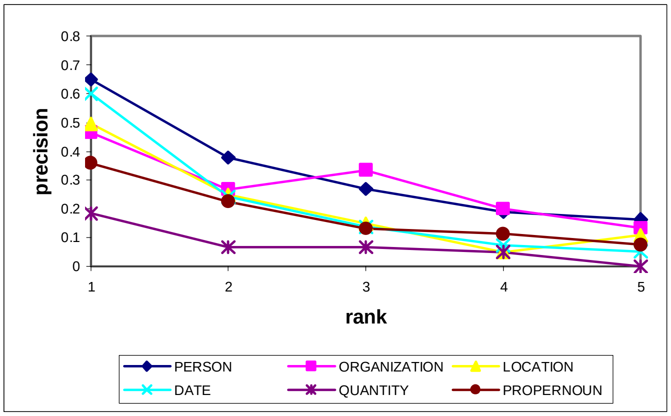


Fig. 4. The relationship between the precision of answers and their ranks.

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 patterns (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.

3.4 Contributions of Components

Furthermore, we analyze the contributions of different components to the overall system performance. Table 3 reports the performance change when different components of the system are removed or changed. Here “baseline” means the actual complete LAMP system.

The “-transforming” system is the system without the “query formulation” component that transforms the given question into an appropriate query, i.e., it straightly submits the raw question to the search engine.

The “-merging” system is the system that does not merge the snippet clusters.

The “-stopwords” system is the system that does not consider the common words such as “at”, “of”, “for”, etc. when forming the bag-of-words feature vectors.

The “cos_score” system is the system that uses the cosine value of the angle between the question vector and the answer vector as the score function for answer selection.

Table 3. The contributions of components.

	MRR	MRR drop	CWS	CWS drop
baseline	0.48		0.62	
-transforming	0.46	4%	0.61	2%
-merging	0.46	4%	0.59	5%
-stopwords	0.47	2%	0.60	3%
cos_score	0.31	35%	0.28	55%

3.5 Comparison with Other Strategies

Table 4 compares the performance of systems with different answer selection methods mentioned in Section 2.6. Clearly the answer selection method based on aggregation which is proposed in this paper outperforms the common methods based on individual or redundancy.

Table 4. The comparison with other strategies.

	MRR	MRR drop	CWS	CWS drop
aggregation	0.48		0.62	
individual	0.33	31%	0.28	55%
redundancy	0.45	6%	0.55	11%

4 Related Work

The main features of LAMP and other state-of-the-art question answering systems are listed in Table 5.

Table 5. The main 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 & snippets	Y	Y
LCC-Web	Web documents	URLs & snippets	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

Here TREC systems are the question answering systems dedicated to TREC-QA tasks, including Qanda, Falcon, Webclopedia, AskMSR, Insight, etc. [10,11,12]. These systems have to find answers from a large local news text corpus. They usually run slowly in offline mode, because they have about one week time to submit their results for several hundred test questions. Their answer selection methods are mostly based on individual.

The researchers from Microsoft have tried to use the snippets in Web search results to reinforce their TREC-QA system [7]. Their answer selection method is based on redundancy.

The following systems are all online, and publicly accessible on the Web: START[7], QuASM[8], SiteQ/E[9], IONAUT[10], AskJeeves[11], LCCWeb[12], and AnswerBus[13]. However, they are still not ready to return the exact answers for the questions.

The question answering systems closest to LAMP are Mulder [8] and NSIR [9], which were published on the recent World Wide Web conferences. 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. In 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 or

[7] http://www.ai.mit.edu/projects/infolab/

[8] http://ciir.cs.umass.edu/~reu2/

[9] http://ressell.postech.ac.kr/~pinesnow/siteqeng/

[10] http://www.ionaut.com:8400/

[11] http://www.ask.com/

[12] http://www.languagecomputer.com/demos/

[13] http://misshoover.si.umich.edu/~zzheng/qa-new/

CWS score, so it can not be compared directly. The MRR score of NSIR, over 200 TREC8 test questions, is 0.15 [9].

5 Conclusion

The main contributions of this paper are as follows. (1) We show that high performance question answering based on Web search results is feasible. (2) We describe a question answering system LAMP, which can instantly find answers from the Web. It is fast because it only takes advantage of the snippets in the search results returned by a search engine. (3) We propose a simple but effective answer selection method based on aggregation, which is shown to be more effective than the common methods based on individual or redundancy.

References

1. E. L. Antworth. *PC-KIMMO: A Two-level Processor for Morphological Analysis*. Dallas, TX: Summer Institute of Linguistics, 1990.
2. E. Agichtein, S. Lawrence and L. Gravano. Learning Search Engine Specific Query Transformations for Question Answering. In *Proceedings of the l0th World Wide Web Conference (WWW)*, pp.169-178, 2001
3. R. Baeza-Yates and B. Ribeiero-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
4. 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.
5. E. Charniak. *A Maximum-Entropy-Inspired Parser*. Technical Report CS-99-12, Brown University, Computer Science Department, August 1999.
6. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.
7. S. Dumais, M. Banko, E. Brill, J. Lin and A. Ng. Web Question Answering: Is More Always Better?. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2002, pp. 291-298.
8. C. Kwok, O. Etzioni, and D. S. Weld. Scaling Question Answering to the Web. In *Proceedings of the 10th World Wide Web Conference (WWW)*, Hong Kong, 2001.
9. 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 (WWW)*, Hawaii, 2002.
10. E. Voorhees. The TREC-8 Question Answering Track Report. In *Proceedings of the 8th Text Retrieval Conference (TREC)*, pp. 77-82, NIST, Gaithersburg, MD, 1999.
11. E. Voorhees. Overview of the TREC-9 Question Answering Track. In *Proceedings of the 9th Text Retrieval Conference (TREC)*, pp. 71-80, NIST, Gaithersburg, MD, 2000.
12. E. Voorhees. Overview of the TREC 2001 Question Answering Track. In *Proceedings of the 10th Text Retrieval Conference (TREC)*, pp. 157-165, NIST, Gaithersburg, MD, 2001.
13. D. Zhang and W. S. Lee. Question Classification using Support Vector Machines. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Toronto, Canada, 2003, pp. 26- 32.