Evaluation of a Query-biased Document Summarisation Approach for the Question Ansnwering Task

Mingfang Wu[1] **Ross Wilkinson**[1] **Cécile Paris**[2]
CSIRO ICT Center,
[1]Gate 7, 71 Normanby Road, Private Bag 10, South Clayton 3169
[2]Locked Bag 17, North Ryde, NSW 1670, Australia
{Mingfang.Wu, Ross.Wilkinson, Cecile.Paris}@csiro.au

Abstract

This paper presents an approach on a query-biased document summarisation and an evaluation of the use of such an approach for question answering tasks. We observed a significant difference in the user's performance by presenting a list of documents customized to the task type, compared with a generic document summarization approach. This indicates that paying attention to the task and the searcher interaction may provide substantial improvement in task performance.

1 Introduction

People are searching information to meet their information needs from their tasks at hand. However, most search engines interact with user in a "one size fits all" fashion and ignore the user's preferences, search context or the task context. The burden is then placed on the user to scan, navigate, and read the retrieved documents to identify what s/he wants. We believe that paying attention to the nature of the information task and the needs of the searcher may provide benefits beyond those available through more accurate matching. As Saracevic [9] pointed out, the key to the future of information systems and searching processes lies not only in increased sophistication of technology, but also in increased understanding of human involvement with information.

In the study presented in this paper, we examine searchers' ability to carry out a question answering task [13]. Unlike the task of the non-interactive TREC question answer track [10], where the question answering focuses on fact-based, short answer questions such as "Who is the first prime minister of Australia". We looked at the type of question answering task that is more complex than the task of finding a single fact. The answer to this type of questions would not generally be available in a single document, but would require facts to be extracted from several documents. For example, an Australian cattle farmer would like an information access system that could tell s/he "which countries are the top ten importers of Australian beef?". An ideal answer should consist of a list of country names together with corresponding beef import data. This answer could be synthesized from scattered information collected from various sources, such as a news article about Japanese meat imports and an analysis report on Australian beef in the European market.

The successful completion of such a task requires an answer to be obtained, citing the relevant source documents. If we assume that we do not have an advanced language engine that can understand such questions and then synthesize answers to them, a searcher will be involved in the process, beyond simply initiating a query and reading a list of answers. Some of the elements that might lead to successful answering might include:

- support for query formulation (and re-formulation)
- effective matching and ordering of candidate documents
- delivery of a useful form of the list of the candidate documents
- support for extraction of answers from documents
- synthesis of the answer

There has been quite a bit of study on how to support query formulation, and the bulk of IR research has been devoted to better matching. Research into question answering technology (for automatic approaches) or text editing (for manual approaches) is needed for the last two activities. In this work, we have concentrated on the task of delivering a useful form of the list of the candidate documents. The research question we investigated is: given a same list of retrieved documents, will

the variation in document summary/surrogate improve searcher's performance on question answering task?

Under the evaluation framework of the TREC (Text REtreieval Conference) interactive track [7], we conducted two experiments that compared two types of candidate lists in two experimental systems. One system (the control system) uses the document title and the first N words of a document as the document's summary, while the other system (the testing system) uses the document title and the best three "answer chunks" extracted from the documents as the document's summary. The second confirming experiment repeated the first experiment, but with different search engine, test collection and subjects. The purpose of the second experiment is to confirm the strong results from the first experiment and to test whether the methodology could be generalized to web data.

The rest of the paper is organized as follows: Section 2 discusses our motivation and approach. Section 3 describes the experimental setup, including experiment design and test collections. Section 4 presents the experiments' results and provides detailed analysis. Section 5 provides the conclusions we have drawn.

2 Motivation and approach

In our previous studies [12], we investigated the use of clustering and classification methods to organize the retrieved documents, we found that while subjects could use the structured delivery format to locate groups of relevant documents, the subjects often either failed to identify a relevant document from the document summary or were unable to locate the answer component present within a relevant document.

We hypothesize that one of the reasons for potential gains from structured delivery not being realized is that in our previous test systems the tools that were provided to differentiate the answer containing documents from non-answer containing documents were inadequate for the task of question answering.

In our previous testing systems, a retrieved document is represented by its title. While a document's title may tell what the document is about, very often an answer component exists within a small chunk of the document, and this small chunk may not be related to the main theme of the document. For example, for the question "Which was the last dynasty of China: Qing or Ming?", the titles of the first two documents presented to a searcher are: "Claim Record Sale For Porcelain Ming Vase" and "Chinese Dish for Calligraphy Brushes Brings Record Price". The themes of the two documents are Ming vases and Chinese dishes respectively, but there are sentences in each document that mention the time span of the Ming Dynasty and of the Qing Dynasty. By reading only the titles, searchers miss a chance to easily and quickly determine the answer, even the answer components are in the top ranked documents.

In this work, we still use document retrieval, but focus on the surrogate or summary of the retrieved documents. Some experiments have evaluated the suitability of taking extracted paragraphs or sentences as a document summary [2], [6], [8]. The produced summary by these methods is purely based on individual document and basically a condensed version of a document - it requires the user less reading time to get to know the gist of the document. There are little studies that have shown whether the use of these summaries is suitable for the interactive question answering task.

In our approach, a document is summarized and represented by its title and the three best answer-indicative sentences (AIS). The three best AIS are dynamically generated after each query search, based on the following criteria:

- An AIS should contain at least one query word.
- The AIS are first ranked according to the number of *unique* query words contained in each AIS. If two AIS have the same number of unique query words, they will be ranked according to their order of appearance in the document.
- Only the top three AIS are selected.

Our hypothesis is that the title and answer-indicative sentences should provide a better indication of whether a document might help answer a given question. This is because documents can easily be completely off the topic of interest to the searcher, but still be relevant because they contain a part of the answer to the question. Therefore, our experiment focused on the comparison and evaluation of two systems using different summaries. The control system First20 uses the title and the first twenty words as the document summary, and the test system AIS3 uses the title and best three answer-indicative sentences as the document summary. Performance will be evaluated in terms of searchers' abilities to locate answer components, searchers' subjective perceptions of the systems, and the efforts required by searchers to determine answers.

3 Experimental setup

Experimental design

The experimental design concerns three major factors: system, question, and searcher, with focus

on the comparison of two experimental systems. Thus, we adopted a factorial, Latin-square experiment design. In this design, each searcher uses each system to search a block of questions; questions are rotated completely within two blocks. For an experiment involving two systems and eight questions, a block of sixteen searchers is needed.

System description

In each experiment, the two experimental systems use the same underlying search engine. The Experiment I used the MG [11] search engine, while the Experiment II used the Padre search engine [4]. In each experiment, the two experimental systems provide natural language querying only. For each query, both systems present a searcher with the summary of the top 100 retrieved documents in five consecutive pages, with each page containing 20 documents. Each system has a main window for showing these summary pages. A document reading window will pop up when a document title is clicked. If a searcher finds an answer component from the document reading window, s/he can click the "Save Answer" button in this window and a save window will pop up for the searcher to record the newly found answer component and modify previously saved answer components.

The difference between the two systems is the form and content of the result presented in the main windows. The main window of the control system (First20) is shown in Figure 1. The main windows of the test systems (AIS3) are shown in Figure 2 and Figure 3.

The AIS3 windows in each experiment are slightly different. In Experiment I (Figure 2), each answer-indicative sentence is linked to the content of the document and the sentence in the document is highlighted and brought to the top of the window. In Experiment II (Figure 3), we remove these links to make the interface closer to the interface of First20 and the three AIS truly the summary. There is a save icon beside each AIS (in Figure 2) or each document title (in Figure3) in AIS3, this icon has the same function as the Save Answer button in the document reading window. If a searcher finds a fact from the following three answer-indicative sentences, s/he can save the fact directly from this (summary) page by clicking the icon.

Document collection

The document collection used by Experiment I contains all newswire articles. Experiment II used a partial collection from the main web track (W10G) [1]. This collection is a snapshot of the WWW; all documents in the collection are web pages. To concentrate on document summaries instead of browsing, we removed all links and images inside a web page - for the purpose of this experiment; each web page was treated as a stand-alone document.

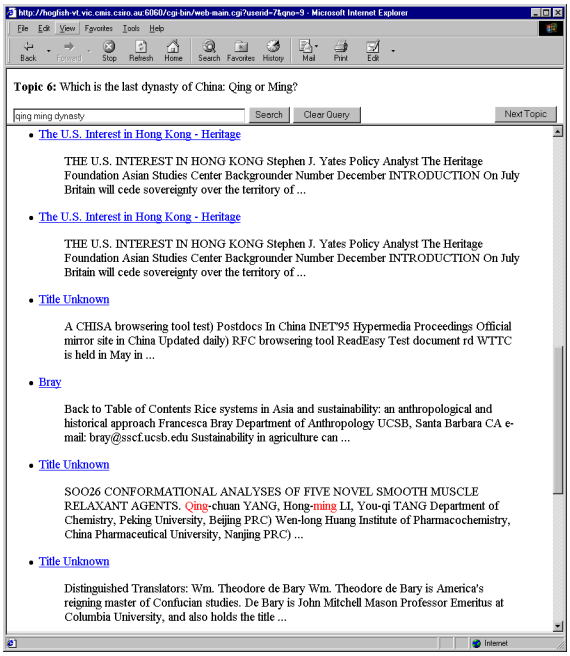

Figure 1. The interface of the First20

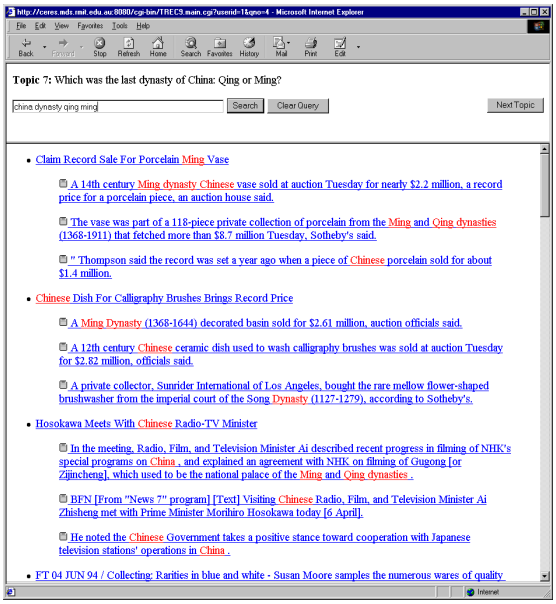

Figure 2. The interface of the AIS3 system in Experiment I

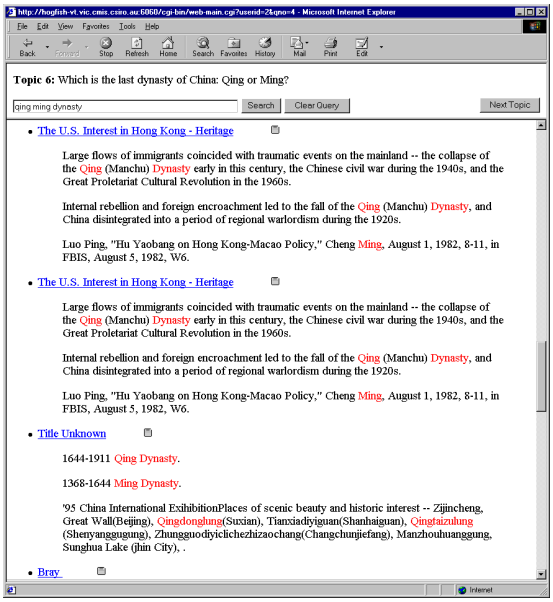

Figure 3. The interface of the AIS3 system in Experiment II

Questions

There are two types of questions in the experiments. The Type 1 questions are of the form <<find any n Xs>>; for example, "Name four films in which Orson Welles appeared.". The Type 2 questions are of the form <<compare two specific Xs>>; for example, "Which was the last dynasty of China: Qing or Ming?". For the Type 1 questions (question 1–4), a complete answer consists of n answer components, plus a list of supporting documents. For the Type 2 questions (question 5–8), two facts are usually needed to make the comparison, plus supporting documents.

Experiment I used a set of eight questions developed by TREC9i participants. To prepare a set of questions for Experiment II, we started with the eight questions from TREC9i. We then removed those questions that could not be fully answered from the document collection used in Experiment II. Additional questions were added either by modifying questions from the main web track, or were developed by an independent volunteer.

Evaluation

A searcher's performance is evaluated in terms of the success rate. For each search question, the saved answers and their supporting documents were collected for judging. There are two levels of judgement: one is whether the searcher finds the required number of answer components (for questions of Type 1) or whether the found facts are enough to infer the answer (for questions of Type 2); another is whether the submitted facts (or answers) are supported by the saved documents. For the success rate, a search session is given a score between 0 and 1: each correctly identified fact supported by a saved document contributes a score of 1/n to the search score, where n is the number of required answer components (or facts) for the question

Experimental procedure

Generally, we followed the procedure recommended by the TREC interactive track [7]. During the experiments, the subjects performed the following tasks:

- Pre-search preparation: consisting of introduction to the experiment, answering a pre-search questionnaire, demonstration of the main functions of each experimental system, and hands-on practice.
- Searching session: each subject attempts four questions on each of the systems, answering a pre-search questionnaire and a post-search questionnaire per question, and a post-system questionnaire per system. Subjects have a maximum of five minutes per question search.
- Answering an exit questionnaire.

Subjects

All searchers were recruited via an internal university newsgroup: all were students from the department of computer science. The average age of searchers was 23, with 4.7 years of online search experience.

Subjects were asked about their familiarity about each question. Overall, subjects claimed low familiarity with all questions (all under 3 on a 5-point Likert scale). In experiment I, the average familiarity of questions from each system is 1.5 (AIS) and 1.58 (First20). In experiment II, the scores are 2.1 (AIS) and 2.0 (First20). No significant correlations are found between familiarity and success

4 EXPERIMENTAL RESULTS

To determine the success of a system at supporting a user performing an information task, it is important to know how well the task is done, how much effort is required, and whether an information system is perceived as helpful. We use independent assessment for performance, system logging for effort, and questionnaires for perception.

4.1 Searcher performance

Experiment I

We aimed to determine whether searchers could answer questions more successfully with the

First20 system or the AIS3 system. Our results show that searchers using AIS3 had a higher success rate than those using First20 for all questions except for Question 5. Overall, by using AIS3, searchers' performance is improved by 38%.

In this experiment, the three variables to consider are the question, the searcher, and the system. Although the Latin-square design should minimize the effect of question and searcher, it is possible that question or searcher effects may still occur. An ANalysis Of Variance (ANOVA) model was used to test the significance of individual factor and the interactions between the factors. Here, the success rate is the dependent variable, and system, question, and searcher are three independent variables. A major advantage of using the ANOVA model is that the effect of each independent variable as well as their interactions are analyzed, whereas for the t-test, we can compare only one independent variable under different treatments. Table 1 shows the result of the three-way ANOVA test on success rates. It tells us that the system effect and question effect are significant, but that the searcher effect and the interaction effects are not.

Table 1. Experiment I: summary of ANOVA model for the success rate

Source	p-value
System	0.041
Question	0.000
Searcher	0.195
System * Question	0.414
Question * Searcher	0.691
System * Searcher	0.050

Experiment II

Experiment II was aimed to confirm the strong result from the experiment I. We planned to repeat the above experiment with a quite different document collection, another set of questions, and different searchers. However, we found that the technique for selecting AIS used in Experiment I could not be applied directly to web documents. Unlike news articles that have coherent text with a well-define discourse structure, web pages are often a chaotic jumble of phrases, links, graphics, and formatting commands. On the other hand, compared with news articles, web documents have more structural information. Although their mark-up is more for presentation effect than to indicate their logical structure, some information between two tags (for example: <LI>…</LI>) can be regarded as a semantically coherent unit and treated as a sentence. Therefore, in addition to the techniques used in Experiment I to segment documents into sentences, we also used some document mark-up as "sentence" indicators.

Table 2 shows the ANOVA test on the experiment II data. The table shows results similar to those in Table 1: only the system and the question have significant effect on the success rate. Overall, AIS3 leads to a performance improvement of 34% over First20.

Based on the searchers' performance in both experiments, our hypothesis that the AIS is a better form of document summary than the first N words for the question answering task is supported.

Table 2. Experiment II: summary of ANOVA model for the success rate

Source	p-value
System	0.020
Question	0.018
Searcher	0.547
System * Question	0.248
Question * Searcher	0.808
System * Searcher	0.525

4.2 Searcher effort

The effort of a searcher in determining answers to a question can be measured by the number of queries sent, the number of summary pages viewed, and the number of documents read.

On average, searchers sent fewer queries, viewed fewer summary pages, and read fewer documents from *AIS3* than from *First20* in both experiments (refer to Table 3).

We note that searchers generally did not use more than one summary page per query, nor did they need to read many documents to carry out the task. Considering the summary page of *AIS3* displays more text than that in *First20*, we may tentatively conclude that searchers read similar amount of text, but *AIS3* provides higher quality information than the *First20* does, since we know searcher performance is better.

Searcher preference

The perception of searchers of the systems is captured by three questions in exit questionnaire. The three questions are

- Q1: Which of the two systems did you find easier to learn to use?
- Q2: Which of the two systems did you find easier to use?

Table 3. Searchers' interactions with two systems

	Experiment I		Experiment II	
	First20 Mean(SD)	*AIS3* Mean(SD)	*First20* Mean(SD)	*AIS3* Mean(SD)
No. of unique queries sent	2.14(0.56)	1.73(0.57)	2.0(1.2)	1.7(1.0)
No. of surrogate pages viewed	2.80(1.64)	1.98(0.97)	2.4(1.4)	2.0(1.3)
No. of documents read	3.42(1.22)	2.66(0.77)	4.2(2.8)	3.2(2.7)

Table 4. Searchers' perceptions of two systems

	Experiment I			Experiment II		
	Q1	Q2	Q3	Q1	Q2	Q3
First20	3	4	5	2	2	2
AIS3	8	11	11	10	12	13
No difference	5	1		4	2	1

- Q3: Which of the two systems did you like the best overall?

The distribution of the searchers' choices is shown in Table 4. Combining the results from the two experiments' questionnaires, for question 1, 15% of subjects selected *First20,* while 56% of subjects selected *AIS3*; for question 2, 19% of subjects selected *First20,* while 71% of subjects selected *AIS3*; for question 3, 22% of subjects preferred *First20*, while 75% preferred *AIS3*.

5 Conclusion

In this paper, we report two user studies on interactive question answering task. By constructing a delivery interface that takes into account the nature of the task, we saw that searchers:

- issued fewer queries
- read fewer documents
- found more answers

We conducted two experiments that would allow us to determine searcher performance, searcher effort and searcher preference. Our results show that searchers' performance when using an AIS3 system is improved over using a First20 system, based on objective assessment; this result is consistent in both experiments. The performance difference between two experimental systems is statistically significant. The data suggests that searchers using AIS3 require less effort, although cognitive load experiments are required to confirm this. Finally, AIS3 is preferred by most searchers. Thus, the experiments support our hypothesis that AIS3 is a better indication of document suitability than First20, for the question answering task.

Different search tasks may require different delivery methods. For example: the clustering of retrieved documents can be used for the task of finding relevant documents [5], and the classification of retrieved documents can be used for the purposing of browsing. However, for the task of question answering, we found that none of these delivery methods performed better than a ranked list [12]. The experiments presented in this paper indicate that a relatively simple document summary can significantly improve the searcher's performance in question answering task.

References

[1] Bailey P., Craswell N. and Hawking D. *Engineering a Multi-purpose Test Collection for Web Retrieval Experiments.* Information Processing and Management. 2001.

[2] Berger A. L. and Mittal V. O. *OCELOT: A System for summarizing web pages.* In Proceedings of the 23rd ACM SIGIR Conference. July 24-28, 2000, Athens, Greece (pp. 144–151).

[3] D'Souza D., Fuller M., *et al. The Melbourne TREC-9 Experiments.* In Proceedings of the Ninth Text Retrieval Conference (TREC-9) (pp. 366-379). November 2000, Gaithersberg, MD, USA.

[4] Hawking D., Craswell N. and Thistlewaite P. *ACSys TREC-8 Experiments.* In Proceeding of Seventh Text Retrieval Conference (TREC-8), November 1999, Gaithersburg, MD, USA.

[5] Hearst M. A. and Pedersen J. O. *Reexamining the cluster hypothesis: Scatter/gather on retrieval results.* In Proceedings of the 19th ACM SIGIR conference, August 1996, Zurich, Switzerland (pp. 76-84).

[6] Kupiec J., Pedersen J. and Chen F. *A Trainable Document Summarizer.* In Proceedings of the ACM SIGIR conference, July 1995, New York, USA (pp. 68–73).

[7] Over P. *TREC-9 Interactive Track – Basics.* In Note papers of TREC-9. November 2000, Gaithersberg, MD, USA (pp 721-728).

[8] Salton G., Singhal A., Mitra M. and Buckley C. *Automatic Text Structure and Summarization.* Information Processing and Management. Vol. 33 (2) 193-207, 1997.

[9] Saracevic T. and Kentor P. *A Study of Information Seeking and Retrieving: I. Background and Methodology.* Journal of the American Society for Information Science. Vol. 39(3), 161-176. 1988.

[10] Voorhees E. M. *The TREC-8 Question Answering Track Report.* In proceedings of the Nipth Text Retrieval Conference (TREC-8). Novemeber 1999, Gaithersberg, MD, USA.

[11] Witten I., Moffat A. and Bell T. *Managing Gigabytes: Compressing and indexing documents and images*. Van Nostrand Reinhold. 1994.

[12] Wu M., Fuller M. and Wilkinson R. *Using clustering and classification approaches in interactive retrieval.* Information Processing & Management, 37 (2001) 459-484.

[13] Wu, M., Fuller, M. and Wilkinson, R. *Searcher performance in question answering.* In Proceedings of 24th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 375-381, 2001.