# Web document summarisation: a task-oriented evaluation

Ryen White
University of Glasgow
Scotland
whiter@dcs.gla.ac.uk

Ian Ruthven
University of Glasgow
Scotland
igr@dcs.gla.ac.uk

Joemon M. Jose
University of Glasgow
Scotland
jj@dcs.gla.ac.uk

## Abstract

*In this paper we present a query-biased summarisation interface for web searching. The summarisation system has been specifically developed to act as a component in existing web search interfaces. The summaries allow the user to more effectively assess the content of web pages. We also present an experimental investigation of this approach. Our experimental results shows the system appears to be more useful and effective in helping users gauge document relevance than the traditional ranked titles/abstracts approach.*

**Keywords :** Summarisation, evaluation, WWW.

## 1 Introduction

The Internet has rapidly become an invaluable digital resource. However one of the main features that make the Internet useful - the ease with which information providers can add, update and remove documents - can make it difficult to find relevant information. Current web search engines are useful information access tools but the dynamic nature and size of the Internet result in searches that are incomplete (each web search engine only indexes part of the available information), outdated (pages can change between indexes) or searches that are difficult to manage (the search returns large numbers of documents).

Users of web search engines potentially run large numbers of searches but they will typically have little or no training in how to use these systems effectively. Research such as [5], also indicate that users of web search engines are not inclined to use advanced search facilities. If we are to support this group of users we must incorporate, into the interface, functionalities that help users search more effectively.

This paper contributes to this overall research aim in two ways: firstly we present a summarisation system specifically designed for web search engines, secondly we present a task-oriented evaluation of retrieval techniques for the Internet. This evaluation looks at the effect of subject searching experience, and the user's task on the use and effectiveness of a summarisation-enhanced interface.

Our initial study is a new approach to web search evaluation and involves one-to-one sessions during which users work through a series of simulated information needs [1] on a number of systems.

The paper first describes our motivation for this research, section 2, and describes our summarisation system, section 3. It then looks at IR evaluation and web evaluation, section 4, and our evaluation methodology, section 5. Section 6 presents initial results, we conclude in section 7.

## 2 Motivation

Prior to starting this work on web summarisation we carried out a small pilot study to gauge user opinion about the result pages of two major commercial Internet search engines; AltaVista and Google. This study was intended to elicit difficulties users faced when searching the web. Users were selected to be representative of the web population and incorporated practiced searchers, infrequent searchers and searchers who were relatively new to web searching.

This was an informal study but the results indicated that users require more information about the content of pages. Most users felt that the abstracts presented by the two systems did not provide a sufficient clue about page content, meaning they were forced to visit each page to assess its relevance.

This not only requires effort on the part of the user but also increases the time a user has to spend searching. An effective and efficient method of indicating the content of web pages to users is to present the user with a short summary of the document.

In previous research, [9], we have demonstrated that summarisation techniques can help users of traditional IR systems to filter potentially relevant documents from a list of retrieved documents. Further, summaries that are tailored to the user's query - *query-biased summaries* - can prove more effective than other representations of a document, [10].

For the experiments reported in this paper we developed a retrieval interface, named WebDocSum, which uses query-biased summarisation techniques to enhance the result pages of two search engines. An attempt is also made to incorporate web page media, such as tables and images, into the summary if a document contains insufficient text.

In the remainder of this paper we describe this interface and the experiments we carried out to test its effectiveness in web searching. We conducted the experiments using the evaluative framework reported in [6].

## 3 WebDocSum Retrieval Interface

The summarisation system we developed, WebDocSum, is intended to serve as an adjunct to major commercial search engines. When the user submits a query, the system queries the underlying search engine, parses the results page, dispatches a thread to each page in the result list and creates query-biased summaries of each of these pages. The entire process, from query being submitted to results being displayed takes around 7 seconds. Summaries are created in the background as the results page is being displayed.

### 3.1 Summarisation

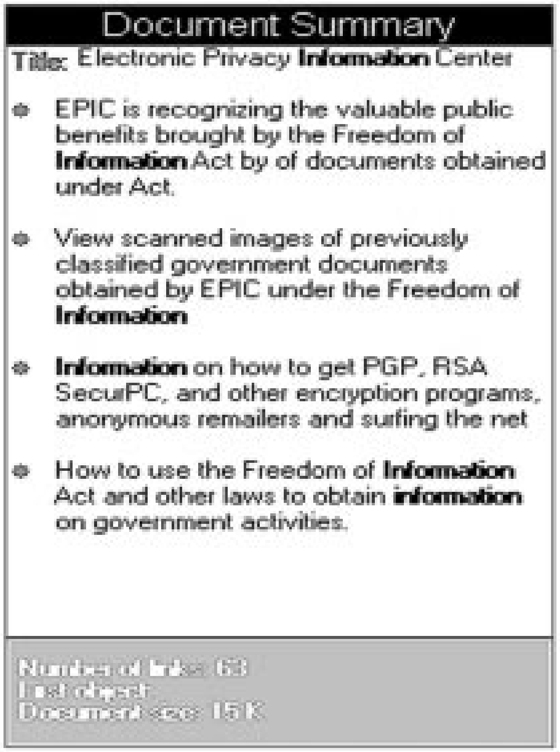
The summaries are created through a sentence extraction model: each web page is split into its component sentences, the sentences are scored according to useful they will be in a summary and a number of the highly-scored sentences are chosen to compose the summary.

Sentences are scored through their position (initial introductory sentences are preferred), the words they contain (words that are emphasised by the user, e.g emboldened words, or words in the document title are treated as important), and the proportion of query terms they contain. This latter component - scoring by query terms - tailors the summaries towards the query.

There are three main parts - summary window, Figure 1, results list and query input. Only the title of the document is shown in the results list. When the user moves the mouse over a document title, the summary window will change to show a summary for that page. If a title is *clicked*, the page will open in a new window. A query form retains the current query for quick and easy reformulation.

### 3.2 Summary Window

Developed using a Java Applet, the summary window will display a summary of a document when the mouse pointer passes over its link in the results list. In its standard form the window displays the page title, each sentence bullet-pointed and all query terms in bold. A panel at the bottom of the window displays the following extra information about the document being summarised:



Document Summary

Title: Electronic Privacy **Information** Center

- EPIC is recognizing the valuable public benefits brought by the Freedom of **Information** Act by of documents obtained under Act.
- View scanned images of previously classified government documents obtained by EPIC under the Freedom of **Information**
- **Information** on how to get PGP, RSA SecurPC, and other encryption programs, anonymous remailers and surfing the net
- How to use the Freedom of **Information** Act and other laws to obtain **information** on government activities.

Number of links: 63
First object:
Document size: 15 K

**Figure 1. The summary window.**

- **Number of Links** - number of links on the page, may help users identify important sites and hubs;
- **First Object** - first non-text object on page, e.g. the first image, used in situations where an alternative summary is needed (see below);
- **Document Size** - the size of the document being summarised.

As well as being able to display textual output, the summary window can also give feedback should a web error occur. Such an error would occur if a web page was unavailable or was taking too long to retrieve. In such circumstances the summary window will show the abstract offered by the underlying search engine and an error message detailing the reason for the web error.

Finally, if the summary generated by the system is not of sufficient length (i.e. more than 25 characters long) the name of the first applet, picture, table or form is displayed in the window. Used in conjunction the abstract from the underlying search engine and the extra information in the panel, this can give a reasonable indication of page content.

## 4 Issues in Evaluation

Evaluative studies are concerned with the assessment of the quality of a system's performance with respect to the needs of its users with a particular context or situation. The direction of such studies is commonly determined, and thus implicitly validated, by the adoption of some kind of structured methodology or 'evaluative framework' [6].

The traditional framework for evaluative studies of information retrieval systems derives from the Cranfield projects in the early 1960s and still survives in the large scale experiments undertaken the annual TREC conference (trec.nist.gov). However, this framework assumes low level of interactivity between users and the systems, uses retrieval effectiveness as the primary measures discarding other aspects such as user satisfaction, efficiency and the entire evaluative process depends on the pre-defined set of queries and relevance assessments. Recently, the drawbacks of this approach have been the topic of intense discussion in the research community [4].

The main exception to the large-scale traditional evaluation paradigm has been the TREC interactive track [7]. However this has not, as yet, examined web searching.

Because of these problems, this traditional framework can not be applied to the evaluation of the web search systems.

### 4.1 Web Evaluation

A number of evaluation studies of web search tools have been carried out, e.g. [2, 3, 5]. These studies either only consider one type of search task,[2], use expert searchers rather than representative end-users, [3], or are based on statistical analyses of web logs rather than interactive aspects of searching, [5]. Hence the methods used in these studies are not appropriate for our study of the effectiveness of a new interaction technique.

In the next section we shall describe our evaluation methodology.

## 5 Task-Oriented Evaluation

It is important to measure systems in actual information seeking situations, and real-world systems can only be meaningfully evaluated in real-world settings. However, we often want to maintain experimental control over the tasks for which a user is searching, to allow system comparison between subjects. An approach known as 'simulated information needs' [1] allows the use of realistic information seeking tasks to be used in a laboratory environment. The careful construction of an information-seeking scenario can serve as a *simulation* of a real information need. This is the approach that we are using during the course of this task-oriented study. Specifically, we will employ an experimental methodology similar to that reported in [6].

We use simulated information needs to investigate the use and effectiveness of our summarisation techniques for different types of searching tasks. We also investigate how the experience of the searchers influence the results. This is particularly important as the user group of the Internet is large and diverse.

### 5.1 Experimental Design

In our evaluative study, we are making use of a within-subjects (repeated-measures) experimental design. The independent variable is system type and each participant will use four systems in total. Separate sets of values of a variety of dependent variables indicative of acceptability or user satisfaction were to be determined through the administration of questionnaires to each subject. Our specific experimental hypothesis was that the system with the query-based summaries prove to be more effective in satisfying the user.

### 5.2 Users

Users are at the center of the evaluation framework. We used 24 users in total, 8 from each of the following three categories; novices (infrequent web searchers), occasional users (moderate frequency web searchers) and experts (high frequency web searchers). Subjective tests and evaluation assess the systems from the perspective of the user. This is done via questionnaires using Likert scales and semantic differentials [8], explained in section 6.

### 5.3 Systems

Four systems were used in our experiments: two commercial web search engines (Google and AltaVista) and a version of each search engine that used WebDocSum.

To eliminate possible bias caused by previous searching experience, and to isolate the effect of the summarisation interface, we gave the user no indication of the specific search engines being used. Wrappers were developed for both search engines that preserved all content, but masked the identity of the search engine. Google and AltaVista were referred to only as System A and System B. The versions of Google and AltaVista that used WebDocSum were labelled as System C and System D.

Both Google and AltaVista show users short descriptions of the content of retrieved documents. These were preserved in Systems A and B. This allowed us to compare the presentation of the original descriptions (A and B) with longer, query-biased summaries of the retrieved documents (C and D).

### 5.4 Search Tasks

Through the use of simulated information needs we are able to place the user mentally in an actual information seeking situation. We used 4 tasks in total and great care was taken to ensure that the tasks were as realistic as possible. The tasks were chosen to reflect different types of information need and are the basis of simulated information need. Each need was framed within a simulated task - the user was given a scenario that indicated what material was required and why the information was needed.

The following list outlines the type and topic of search, the full simulated work task is omitted for brevity.

- **Search for a fact** - finding a named person's current e-mail address;
- **Search for a number of items** - finding five hotels in Paris, France that offer an online booking service;
- **Decision search** - finding information about the 'best' impressionist art museum in Rome, Italy;
- **Background search** - finding information about dust allergies in the workplace.

Each user performed one task on one system; the order of system presentation and allocation of task to system was randomised.

## 6 Results & Analysis

In this section we discuss the data collection and results. We look at both the effectiveness of the searches (time and task success) and the users' perceptions of the systems.

*Semantic Differentials:* Each respondent was asked to describe various aspects of their experience of using each system, by scoring each system on the same set of 11 5-point semantic differentials. 3 of these focused on the task that had been set; 4 of these had been on the search process that the respondent had just been carried out; 4 focused on the summaries/descriptions presented by the system.

Table 1 shows an example semantic differential for the statement: *The task we asked you to perform was*.

| | very | reasonably | neither-nor | reasonably | very | |
|---|---|---|---|---|---|---|
| clear | 1 | 2 | 3 | 4 | 5 | unclear |

**Table 1. Example semantic differential**

We compared the set of 24 scores on each differential for System A (Google) with the corresponding set of 24 scores on each differential from System C (Google with summaries) and System B (AltaVista) with System D (AltaVista with summaries).

Given the ordinal scale of the data, the Mann-Whitney test statistic was used to test the one-tailed experimental hypothesis.

We first compared the differentials regarding the task and found no significant difference between the differentials concerning the task. This indicates that the task distribution was comparable across systems.

We then compared the users' perceptions of the summaries produced by the WebDocSum interface compared to the descriptions produced by the search engines, section 4.3. The users rated the query-biased WebDocSum summaries as more relevant, important, useful and complete (all four assessment categories) compared to the search engine descriptions on both Google and AltaVista. All differences were statistically significant across users and within the three user groups (novice, infrequent, expert). This indicates a user preference for the query-biased summaries as a document representation.

Finally we compared the differentials relating to the overall search. The search process on Google with WebDocSum extension gave significant differences on 3 differentials out of 4 (with the users rating searches with WebDocSum as more relaxing, interesting and restful). Only for the differential easy/not easy was a non-significant difference found although the ratings were in favour of WebDocSum. In comparing the search process on AltaVista with WebDocSum extension all differentials were in favour of WebDocSum and statistically significant. These differences held across users and within the user groups

*Likert Scale:* Each user was invited to indicate, by making a selection from a 5-point Likert scale [8], Table 2, the degree to which they agreed or disagreed with each of five statements about various aspects of their interaction with the system. These statements were phrased in such a way that responses would indicate the extent to which:

- The tasks were familiar and they had an exact idea of the information that they wanted. These are used to measure whether the simulated work task situation placed them in actual user context.
- The use of summaries was helpful to assess the relevance of the page.
- The abstract summaries showed the query in the context.
- Questions about the outcome of their search.

Table 3 shows an example of a Likert scale for the statement: *I had an exact idea of the information I wanted*.

In all counts, users scored similarly with respect to the questions dealing with the familiarity with the task and the

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| I agree completely | | | | I disagree completely |

**Table 2. Example Likert scale**

exact idea of the information need. This indicates that during the evaluation the task contexts were similar across four tasks and four systems.

Regarding the questions about the use of summaries, the systems with query biased summaries scored significantly better on both systems with query-biased summaries. That is, the users rated the summaries as more useful for indicating relevance and reported higher task satisfaction with the summaries than with the descriptions, indicating that the query-biased summarisation was beneficial.

Results from the experiments show that the summarisation component, WebDocSum, significantly reduces the time for a user to complete a task (average search time, Table 3.

| | |
|---|---|
| Google | 8 mins 53 secs |
| AltaVista | 9 mins 21 secs |
| Google + WebDocSum | 6 mins 31 secs |
| AltaVista + WebDocSum | 6 mins 47 secs |

**Table 3. Average time to complete a task**

WebDocSum also increases the number of users who completed a search task (average number of tasks completed on a non-summarising system 2.75 , compared to 4.75 on a summarising system).

Both of these differences are significant using a Mann-Whitney Test.

## 7 Conclusions

In this paper we have briefly described a query-biased summarisation system for web search engines and an evaluation of its effectiveness. The evaluation methodology gives a formal framework for investigating the search process. The results indicate that summarisation techniques, such as the one we propose, are not only more popular than existing document descriptions produced by web search engines but can also lead to more effective user searching. These results hold for users of different search experience and for different types of task.

## 8 Acknowledgements

We would like to thank all the subjects who participiated in this experiment. We would also like to gratefully acknowledge the contribution of Tassos Tombros and the anonymous referees.

## References

[1] P. Borlund and P. Ingwersen. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3):225–250, 1997.

[2] B. J. Dempsey, R. C. Vreeland, R. G. Summer Jr, and K. Yang. Design and empirical evaluation of search software for legal professionals on the www. *Information Processing & Management*, 36(2):253–273, March 2000.

[3] M. Gordon and P. Pathak. Finding information on the world wide web: the retrieval effectiveness of search engines. *Information Processing & Management*, 35:141–180, 1999.

[4] W. R. Hersh. Relevance and retrieval evaluation: Perspectives from medicine. *Journal of The American Society for Information Science*, 45(3):201–206, 1994.

[5] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: A study and analysis of users on the web. *Information Processing & Management*, 36(2):207–227, March 2000.

[6] J. M. Jose, J. Furner, and D. J. Harper. Spatial querying for image retrieval: a user-oriented evaluation. In *Proceedings of the 21st Annual International SIGIR Conference*, pages 232–240. ACM Press, August 1998.

[7] P. Over. Trec-6 interactive track report. In *Proceedings of the Sixth Text Retrieval Conference*, pages 73–82. NIST Special Publication, Novmber 1998.

[8] J. Preece, editor. *Human-Computer Interaction*. Addison-Wesley, 1994.

[9] I. Ruthven, A. Tombros, and J. M. Jose. A study on the use of summaries and summary-based query expansion for a question-answering task. In *ECIR '01*, pages 41–53. electronic workshops in computing, April 2001.

[10] A. Tombros and M. Sanderson. The advantanges of query-biased summaries in information retrieval. In *Proceedings of the 21st Annual International SIGIR Conference*, pages 2–10. ACM Press, August 1998.