Query-Biased Web Page Summarisation:
A Task-Oriented Evaluation

[Extended Abstract]

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

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

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

ABSTRACT

We present a system that offers a new way of assessing web document relevance and new approach to the web-based evaluation of such a system. Provisionally named WebDoc-Sum, the system is a query-biased web page summariser that aims to provide an alternative to the short, irrelevant abstracts typical of many web search result lists. Based on an initial evaluation the system appears to be more useful in helping users gauge document relevance than the traditional ranked titles/abstracts approach.

1. INTRODUCTION

For many users, retrieving documents from the Internet can prove a difficult task. Users can find it hard to satisfy their information need, construct queries and assess the relevance of documents based on just a few words selected by the search engine. The system we have developed will hopefully address the latter of these three concerns; relevance assessment.

Building on the work of Tombros and Sanderson [2], our system uses query-biased summarisation techniques to enhance the result pages of the Google and AltaVista. An attempt is also made to incorporate web page media, such as tables and images, into the summary if a document contains insufficient text. We study the retrieval effectiveness of these two popular systems.

Most web evaluation studies tend only to assess historical data gathered by search engines or logging systems. In contrast, our study is a approach to web-based evaluation that involves one-to-one sessions during which users work through a series of simulated work tasks [1] on a number of systems.

2. SYSTEM DESCRIPTION

Our system, WebDocSum, is intended to serve as an adjunct to a major commerical search engine. 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.

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.

Summaries appear in a summary window which will be discussed next.

2.1 Summary Window

Developed using a Java Applet, the summary window will display a summary of a document when the mouse pointer passes over the page title in the results list. If a title is clicked, the page will open in a new window. In its standard form the summary 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:

- **Number of Links** - number of links on the page, may help users identify important sites and hubs;

- **First Object** - the first non-text object, e.g. an image, on the page, used in situations where an alternative summary is needed. For example if a page contains too little text to summarise (less than 25 characters), the summary window will display the name of the first non-text object. 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.

- **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.

3. EXPERIMENTAL METHODOLOGY

It is important to measure systems in actual information seeking situations, and real-world systems can only be meaningfully evaluated in real-world settings. An approach known as 'simulated work task situation' [1] involves the creation of an effective information-seeking scenario simulating a real need. This is the approach that we are using during the course of this task-oriented study.

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.

3.1 Users

Users are at the center of the evaluation framework. We recruited 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.

3.2 Systems

Two of the systems used in our evaluation are commercial web search engines. To eliminate possible bias caused by previous searching experiences it was thought best to give the user no indication of the search engines being used. Wrappers were developed for both search engines that preserved all content, but masked the identity of the search engine. The systems were referred to only as System A and System B, and these two will be compared in our evaluation.

Two systems using WebDocSum are also included. These systems aim to enhance the two search engines used for Systems A (Google) and B (AltaVista) and will therefore enable a summarisation versus no-summarisation comparison. These two systems will be referred to as Systems C (Google + WebDocSum) and D (AltaVista + WebDocSum).

3.3 Search Tasks

Through the use of simulated work tasks we are able to place the user mentally in an actual information seeking situation. They serve to make the user undertake the objective of the search. There are 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 work task.

- **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.

Table 1: The average semantic differential value assigned to the summaries of each system (lower = better, rang 1 - 5)

System	Relevant	Important	Useful	Complete
System A	3.04	2.88	3.17	3.75
System C	1.79	2.04	1.96	2.21
System B	3.54	3.17	3.63	3.63
System D	1.67	1.92	2.13	1.91

Table 2: The average Likert scale rating based on how well systems place query terms in context (lower = better, range 1 - 5)

System	Average Rating
System A	2.75
System C	1.75
System B	3.63
System D	2

4. RESULTS

24 users participated in our evaluation . After using each system, users answered questionnaires to assess the summaries/abstracts presented. Users felt the systems with query-biased summaries (i.e. C and D) produced more relevant, important, useful and complete summaries than A and B, the traditional ranked titles systems. This can be seen in Table 1.

Participants were asked to rank the four systems in terms of which system they preferred. For personal preference, 23 out of 24 users ranked a WebDocSum system in first place and 19 out of 24 placed both WebDocSum systems WebDocSum system as their top two. There also appears to be a loose correlation between document relevance and personal preference.

As part of the post-system questionnaire, users used Likert scales to rate how well the summaries placed the query terms in context. Table 2 shows that the query-biased summarised systems (C and D) appeared to perform better than the non-summarised systems (A and B).

The results to date also indicate the the WebDocSum interfaces help users complete more tasks and complete their tasks more quickly.

5. ACKNOWLEDGEMENTS

We would like to thank the experimental subjects, the anonymous reviewers and Tassos Tombros for their valuable contribution to this work.

6. 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] A. Tombros and M. Sanderson. The advantanges of query-biased summaries in information retrieval. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–10. ACM, 1998.