A task-oriented study on the influencing effects of query-biased summarisation in web searching ⋆

Ryen W. White [a] Joemon M. Jose [a] Ian Ruthven [b]

[a] *Department of Computing Science, University of Glasgow, Glasgow G12 8QQ. Scotland*

[b] *Department of Computer and Information Sciences, University of Strathclyde, Glasgow G1 1XH. Scotland*

Abstract

The aim of the work described in this paper is to evaluate the influencing effects of query-biased summaries in web searching. For this purpose, a summarisation system has been developed, and a summary tailored to the user's query is generated automatically for each document retrieved. The system aims to provide both a better means of assessing document relevance than titles or abstracts typical of many web search result lists. Through visiting each result page at retrieval-time, the system provides the user with an idea of the current page content and thus deals with the dynamic nature of the web.

To examine the effectiveness of this approach, a task-oriented, comparative evaluation between four different web retrieval systems was performed; two that use query-biased summarisation, and two that use the standard ranked titles/abstracts approach. The results from the evaluation indicate that query-biased summarisation techniques appear to be more useful and effective in helping users gauge document relevance than the traditional ranked titles/abstracts approach. The same methodology was used to compare the effectiveness of two of the web's major search engines; AltaVista and Google.

⋆ This work builds on White, R., Ruthven, I. and Jose, J. *Web Document Summarisation: A Task-Oriented Evaluation*, a paper presented at the 1st International Workshop on Digital Libraries, Munich, Germany, 3-7 September 2001 and White, R., Jose, J.M. and Ruthven, I. *Query-Biased Web Page Summarisation: A Task-Oriented Evaluation*, a poster paper presented at The 24th Annual International SIGIR Conference on Research and Development in Information Retrieval, New Orleans, USA, 9-13 September 2001. Both papers reported only preliminary results.

Email addresses: whiter@dcs.gla.ac.uk (Ryen W. White), jj@dcs.gla.ac.uk (Joemon M. Jose), Ian.Ruthven@cis.strath.ac.uk (Ian Ruthven).

1 Introduction

The role that the Internet plays in today's society is an integral one. However, one of the main features that makes the Internet useful - the ease with which information providers can add, update and remove documents - can make it difficult to find relevant information. The dynamic nature and size of the Internet can result in searches that are incomplete (each web search engine only indexes part of the available information), outdated (pages and page content can change after index construction) or - due to the large number of documents returned - difficult to manage.

Users of web search systems potentially run large numbers of searches but will typically have little or no training in how to best utilise web search engines. Users also tend to refrain from using the advanced search facilities that many web search systems now offer (Jansen et al., 2000). If we are to support these users we must incorporate, into the interface, functionalities that help them search more effectively.

In this paper we will firstly present a summarisation system specifically designed for web search engines. Then we present a formal, task-oriented, comparative evaluation for the Internet. We compare two *types* of system; one that uses a summarisation-enhanced interface, and one without such an enhancement. Coupled with this comparison, the evaluation also examines how subject searching experience and tasks allocated influence the use and effectiveness of both these types of system. We use Google [1] and AltaVista [2], two of the web's most popular web search engines, as the baseline systems in this comparison. Using the same methodology, we have also included a comparison of these two search systems.

The motivation behind this study is presented in Section 2, along with an outline of the methodology and results of a pilot study undertaken to gauge user opinion of web search systems. Also in this section, we explore the purpose of the study and introduce the research hypothesis. WebDocSum, the query-biased web page summariser developed to test this hypothesis is introduced in Section 3. In Section 4, the experimental design adopted for the evaluation of the developed summarisation system will be presented, and the wrappers used to disguise the two web search engines used in this evaluation will be introduced. In Section 5, we present and analyse the experimental results, and we discuss these results in Section 6. Finally, Section 7 presents the conclusions that can be drawn from the work reported in this paper and outlines some scope for future research.

[1] http://www.google.com
[2] http://www.altavista.com

2 Motivation

The motivation for the research reported in this paper is twofold. It emanates both from a need to plug a research gap and carry out a task-oriented evaluation on the web, and a need to provide users with a better means by which they can effectively assess document relevance without referring to the full text of a web document.

If we are to provide users with a more effective method of Internet searching, it was felt necessary to elicit the difficulties they faced when using current web search systems. Prior to starting our work on summarisation we carried out a small pilot study to gauge user opinion on the result pages of two major commercial search engines; AltaVista and Google. Users were selected to be representative of the web population and incorporated high, medium and low frequency web searchers. Experiments were informal and carried out in a variety of settings. Users carried out four short searching tasks; two using the standard and two using the advanced features of each web search engine.

In total, 6 participants took part in this informal study. It was not our intention to derive statistically significant results from these experiments, more to informally gauge current user opinion on web search engines. Participants felt that:

- The refinement of queries and the management of the often large document sets proved problematic;
- Document abstracts were ambiguous and too short;
- 'Advanced' searching, in general, tended to return a lower number of matching documents, but led to no 'real' difference (i.e. in one instance 4 million documents were found with the 'advanced features' and 6 million with 'standard features');
- The number of documents found was the first 'feedback' users received. Only a few results actually appeared 'before the fold' (i.e. on the page before any scrolling was necessary), forcing users to scroll down the results list. A large number of retrieved documents tended to dissuade users from viewing the results and persuade them to reformulate their query. Users reported feelings of 'failure' if the search engine returned a large number of documents.

Although this study was informal, the results indicated that initial query interfaces were confusing, result lists were difficult to interpret and more information was required on the content of pages. In particular, most users felt that the document abstracts presented by each system did not provide a sufficient clue about page content, meaning they were forced to visit each page to gauge its relevance. Such an activity only increases cognitive load on the user, and a better solution is to present them with a short summary of the document.

Previous research into this area (Ruthven, Tombros and Jose, 2001; Tombros and Sanderson, 1998) shows that a summary biased to the user's information need (or to be more exact, their query) can be beneficial.

Both Google and AltaVista present short document abstracts for each document in their result lists, based on page content at the time of indexing. Google uses query-biased techniques and presents the query in the context it occurs in the document. To provide this context, Google uses leading and trailing non-query terms to create short snippets of text centered on the query. These snippets, separated by ellipses (i.e. '...') are combined to construct the document abstract. AltaVista, in contrast, does not use query-biased techniques in the construction of its document abstracts, instead opting for a *tf.idf* approach to assign each sentence a score based on the cumulative scores of their constituent words. Both approaches provide short, at times irrelevant, abstracts that may be outdated and hence not truly reflect page content *at query-time*.

For the experiments reported in this paper [3] we developed a retrieval interface, *WebDocSum*, that uses real-time, query-biased summarisation techniques to enhance the result pages of two web search engines. An attempt is also made to incoporate web page media, such as tables and images, into the summary if a document contains insufficient text. If a web page is unavailable, and thus cannot be summarised, an appropriate error message is displayed and the abstract from the underlying search engine is shown instead.

2.1 Purpose of this Study

This study has a twin purpose. We will test the hypothesis that the presence of query-biased web page summaries, for each document in a ranked list, is expected to improve user search effectiveness. We also propose a generic evaluation methodology for web searches, which allows us to evaluate the effectiveness of two internet search engines; AltaVista and Google.

The system, WebDocSum, is only a means that enables us to test this hypothesis. Even though the system can stand on its own, the primary focus is the evaluation and the experimental results obtained. In the next section we will outline the WebDocSum interface.

[3] These experiments were performed during February 2001, when both Google and AltaVista allowed connections from external interfaces.

3 WebDocSum Retrieval Interface

WebDocSum, the summarisation system we developed, 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 list, dispatches a thread to each page in this list and creates query-biased summaries on each of these pages. The entire process, from submission of the query to presentation of results takes around 7 seconds. Summaries are created in the background as the results page is being displayed.

3.1 Summarisation

Numerous researchers have addressed the automation of document summarisation, although the most common method has been through the selection of sentences from the original document (Luhn, 1958; Edmundson, 1969; Paice, 1981; Brandow et al., 1995; Salton et al., 1997). This approach is commonly known as *sentence extraction*, and is capable of producing acceptable summaries that are domain independent.

The summaries are created via a sentence extraction model in which each web page is split into its component sentences, the sentences are scored according to how useful they will be in a summary and in reflecting page content, and typically the top 4 highest-scoring sentences are chosen to compose the summary.

The rationale behind sentence extraction is to find a subset of the source document that is indicative of its contents, typically by scoring words and then sentences according to specific rules. The rules are mainly concerned with the identification of clues for the importance of each sentence in the source document. WebDocSum employs four main types of sentence scoring - title, location, relation to query and text formatting (i.e. **bolded** terms) - all of which are based on the sentence extraction. Each of these will now be explored in turn.

3.1.1 Title

This approach is based on the idea that the author of a document 'reveals' the main concepts in the title of his work. This method also assumes that when an author divides his work into sections, he does so in a standard manner, selecting appropriate headings for each of these divisions. Sentences containing terms that appear in the title and headings are given more weight than those without. Edmundson (1969) experimented with this method, and assigned a

greater importance to terms that appear in the title than in the section headings. The final sentence score for each sentence could then be found through the sum of the weights of each title word in the sentence.

3.1.2 Location

This approach is derived from the belief that sentences located under certain headings in a document convey significant content and are therefore relevant, and that important sentences tend to occur near the start, or near to the end, of a document and its paragraphs (Edmundson, 1969; Brandow et al., 1995).

3.1.3 Query-biased

The motivation for this work comes from the belief that if users could see the sentences in which their query terms appeared they would be able to make a better assessment of document relevance. The work of Tombros and Sanderson (1998) involved calculating a 'query score' for each sentence in the document, based upon its relevance to the query. This comes from the view that the larger the number of query terms in a sentence, the more relevant the sentence is likely to be. The formula used to calculate this score is shown below, where n is the number of query terms in the sentence, q is the total number of query terms and k is an arbitrary constant (set to 2 in this case), included to give the query score k times more weight than the other sentence extraction methods:

$$score = k \cdot \left(\frac{n^2}{q} \right)$$

The 'query score' is added to the score computed by the other extraction methods used, and a final sentence score is calculated. The summary for a document is generated through selecting the top scoring sentences, until the desired summary length was reached. This is defined to be 15% of the document length, or a maximum of four sentences and concurs with previous work in this field (Edmundson, 1964; Brandow et al., 1995). It was felt that such a length of summary would be sufficiently indicative of web document content.

3.1.4 Text Formatting

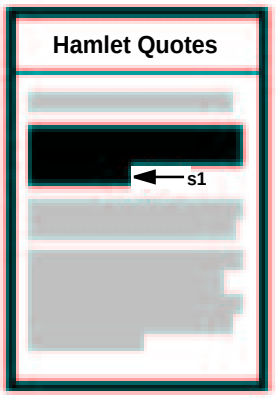
The notion behind this concept is derived from the idea that a web page author may make seek to emphasise important terms (or keywords) in some way. When using the HyperText Markup Language (HTML) that most web documents are written in, the author has access to a vast array of formatting tools, such as **bolded**, *italicised* and <u>underlined</u> text. Intelliscope (Lernout and

Hauspie, 2001), a commercial text summarisation system, uses these means and other heuristics to create document summaries.

When terms of this nature occur in a sentence, the sentence score is incremented by an appropriate amount ($\frac{1}{10}$ in our case) for each term. These values are only intuitive and based on trial-and-error. No empirical testing has been done to test their validity. If a term is formatted in two or more ways, say bold and italic, then the score for that sentence is incremented for each piece of formatting separately.

The methods above are applied to a sentence in the following sequence (title → location → query-biased → text-formatting). All methods are given an opportunity to weight sentences, in reality a large proportion of a sentence's score is derived from its relation to the query. Example 1 (below) shows how a sentence *s1* is scored in our system in relation to a given query.

Example 1: *Sentence scoring*



Web Document

document title : "Hamlet Quotes"

query : "slings arrows Horatio"

sentence(s1) : "Whether 't is nobler in the mind to suffer the slings and arrows of outrageous fortune, or to take arms against a **sea** of troubles, and by opposing end them?"

Scoring process: sentence scoring methods are used sequentially →

Method	**title**	**location**	**query-biased**	**text-formatting**
Applicable terms	-	-	slings, arrows	outrageous, sea
Method score	0	1	$2 \cdot \left(\frac{2^2}{3}\right) = \frac{8}{3}$	$2 \cdot \left(\frac{1}{10}\right) = \frac{1}{5}$
Cumulative score	0	1	$\frac{11}{3}$	$\frac{58}{15} \approx 3.8667$

The sentence scoring methods are applied sequentially (from title to text-formatting) to each sentence in the document. In the example above, the sentence (s1) does not contain any terms from the document title - "Hamlet Quotes" - resulting in a title score of 0. The sentence resides near the start of the document, so receives a location score of 1. The sentence contains two of the three query terms (one occurence of each) and receives a 'query score' of $\frac{8}{3}$ (≈ 2.6667). Two terms in the sentence have extra formatting ('outrageous'

: underlined and 'sea' : bolded), for which they increase the sentence score by 0.2. The scores from all methods are added together to compute a final sentence score, in this case $\frac{58}{15}$ (≈ 3.8667).

Much of the research to date has focused on the summarisation of homogeneous, restricted domain document collections. WebDocSum applies the above methods to web documents in an attempt to provide users with a means by which they can effectively assess relevance.

3.2 Summary Window

The WebDocSum interface, Figure 1, has three main parts - the summary window, the 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. The query input form retains the current query for quick and easy reformulation.

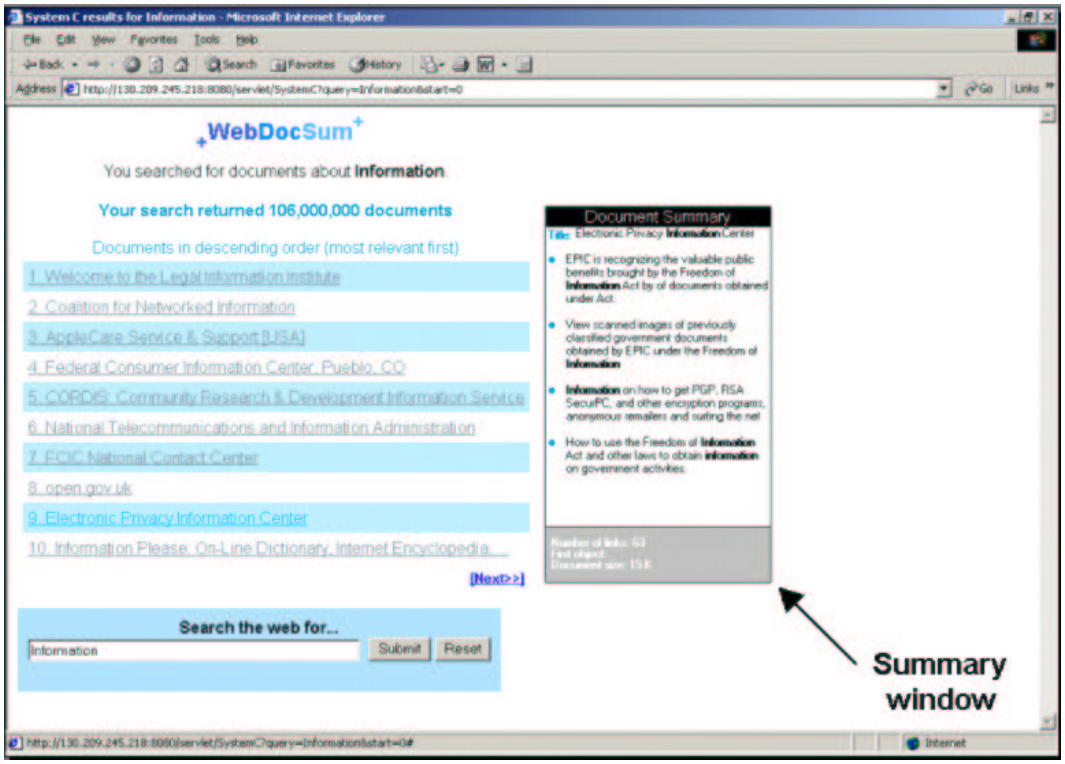


Fig. 1. WebDocSum Interface

The **summary window**, Figure 1 (marked), is the most important part of the interface. 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 result list. In its standard form, the window displays the page title, each

sentence bullet-pointed and query terms in bold. A panel at the bottom of the summary window displays the following extra information about the document being summarised:

- **Number of Links** - number of outlinks on a page, may be useful to users in identifying important sites and hubs;
- **First Object** - the media type of the first non-text object on page, e.g. 'IMAGE', used in situations where an alternative summary is needed, and;
- **Document Size** - the size of the document being summarised.

We decided against including the query-biased summaries underneath the document titles, as is standard in web search result lists. We thought that having a four sentence summary under each document title would lead to a cluttered interface.

3.3 Error Reporting

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 the time taken for a page to load was outwith acceptable bounds (set to 4 seconds in this case), if the page could not be found, if the page was framed or if there was insufficient text to create a summary (see next subsection). 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.

It was hoped that this extra information would help users to make an informed decision about whether to visit a page or not. For example, if it was indicated that a page could not be found, the user will very probably not want to visit it. If it has simply taken too long to create the summary this may point to a slow server and depending on the user, they may be willing or unwilling to attempt to visit the document and wait a little longer for it to be shown.

3.4 Media Incorporation

If, as is the case with a number of pages on the web, there is very little text on a page, the sentence extraction techniques outlined earlier will not be particularly effective. 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.

The rationale behind this approach is to present the user with a sufficient

amount of information about a web page to allow them to make an informed decision about that page's relevance. The incorporation of web page media (type and name, not actual media) into the summaries, is an attempt to do just this. Used in conjunction with the abstract from the underlying search engine and the extra information in the panel, this can give a reasonable indication of page content. Through an informal examination of a large number (around 100) of low-text, high-media web documents, we found that data on filenames and media types may be of some use in relevance assessment. However, this was an informal analysis and a more detailed examination of this concept may be beneficial.

We will now outline the experimental methodology used during this study.

4 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. However, we often want to maintain experimental control over the tasks a searcher is using, to allow system comparison between subjects. An approach known as 'simulated work task situation' (Borlund, 2000) involves the creation of an effective information-seeking scenario *simulating* a real need. This approach will be used during the course of this task-oriented study and our experimental methodology is similar to that used by Jose et al. (1998).

4.1 Issues in evaluation

Relatively few large-scale studies have been conducted to monitor how real users search the web. Most tend only to assess historical data gathered by search engines or logging systems (Jansen et al., 2000; Crovella and Bestavros, 1996; Ding and Marchionini, 1996). The Jansen, Spink and Saracevic study analysed the transaction logs of around 51,000 queries posed by 18,113 users of the web search system Excite [4] . This was a fact and figure analysis, focusing only on the log analysis of user queries. On the web, most system evaluations tend to be large-scale, anonymous and lack real experimental control (Amitay and Paris, 2000; Perlman, 2000). The study proposed in this paper is a new approach to web-based evaluation and involves one-to-one sessions during which users work through a series of simulated work tasks on a number of systems.

[4] http://www.excite.com

Interactive evaluative studies of web search tools are rare, and only a few have been carried out in recent years (Dempsey et al., 2000; Gordon and Pathak, 1999). However, they either consider only one type of search task (i.e. the search for law-related information on the world wide web (Dempsey et al., 2000)), or use expert searchers rather than end-users representative of the general web populace (Gordon and Pathak, 1999). The Dempsey et al. study was restricted to the web's legal information resources and focused solely on legal professionals, those likely to be most interested in the retrieval of such information. The Gordon and Pathak study used searchers who were 'highly qualified to search the web'. Participants (who did no searching themselves) provided skilled searchers with a detailed description of their information need, who then searched the web on their behalf and returned a set of documents to the partcipants for relevance assessment. The methods detailed in these studies were too specific and therefore not appropriate for our study.

4.2 Overview

In contrast to the pilot study detailed earlier, this task-oriented evaluation was formal and served to test the research hypothesis that the presence of query-biased web page summaries, for each document in a ranked list, is expected to improve the search effectiveness of system users.

The evaluation involved the comparison of four systems:

1. Google	2. Google with WebDocSum
3. AltaVista	4. AltaVista with WebDocSum

Each subject uses each of the four systems in a pre-determined, randomised order (see Section 4.4.1 for more details).

Users are at the center of our evaluative framework and the systems are evaluated from the users' perspective. Emphasis is placed on users' perception of search success, rather than the actual outcome of the searches. By doing this we can compare how effective *users felt* the systems were. These results are supplemented by more traditional measures such as actual search success (based on whether users completed tasks in the allotted 10 minutes) and task completion times.

We make use of a within-subjects (repeated-measures) experimental design (Miller, 1984). The independent variable is *system type* and each participant uses four systems in total. Separate sets of values of a variety of different variables indicative of acceptability or user satisfaction are determined through the administration of questionnaires to each subject.

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 results can be influenced by the experience of searchers. This is particulary important considering the size and diversity of the Internet user group.

In what follows in this section we will outline the experimental procedure, the actors involved (both human and automated), the search tasks used and how user opinion and interaction were captured. This section will end with a detailed account of the comparative analysis and a summary of section contents.

4.3 Experimental procedure

The experiment lasted around one-and-a-half hours, with 10 minutes being allocated to each task and a total of 50 minutes for all questionnaires, training and possible interviews. We met one user at a time, each on a separate occasion. For each subject, the procedure was as follows:

- an introductory orientation session;
- a background questionnaire;
- the following 4 steps were carried out for each system:
  - a short training session on the system with which the subject has to interact;
  - hand-out instructions for the task;
  - a 10 minute search session in which the subject interacted with the system in pursuit of the task;
  - a post-search questionnaire;
- a final questionnaire, and;
- informal discussion (optional) [5].

4.4 Actors

There are two types of 'actor' with roles in this experiment; *participants* and *systems*. We will outline each in the subsections that follow.

[5] The informal discussion was initiated at the subject's request. An opportunity to take part in such a discussion was given to all participants.

4.4.1 Participants

Our recruitment was specifically aimed at targetting three groups of users, meant to provide a sample representative of the web populace. There are 24 users in total, 8 from each of the following three categories:

- **Novices** - infrequent web searchers (i.e. participants from a variety of non-computing disciplines, not all are academics, who rarely searched the web)
- **Occasional users** - moderate frequency web searchers (i.e. technically oriented university students with a reasonable level of web searching experience)
- **Experts** - high frequency web searchers (i.e. academics and research staff)

This categorisation was used to investigate the effects of our system on users with different experience levels.

A Greco-Latin square design (Tague-Sutcliffe, 1992) was used to control the learning effects from one session to the next. The order in which the systems were used to undertake the tasks was changed and rotated completely every 4 users. Table 1 shows the experimental design employed (the total number of users attempting a particular task on a particular system is shown in brackets). Subjective questions in questionnaires were used to assess the systems from the perspective of the user. Likert scales and Semantic differentials, explained in Section 5, were used to this end. Participants were recruited through a variety of means and were expected to have both general computer knowledge and some previous web search experience (minimal in the case of Novice users). There was a 50/50 split between the number of males and females participating.

Table 1
Greco-Latin Square design used during experiments

User	Task 1	Task 2	Task 3	Task 4
1	Google (6)	Google[WDS] (6)	AltaVista (6)	AltaVista[WDS] (6)
2	AltaVista[WDS] (6)	Google (6)	Google[WDS] (6)	AltaVista (6)
3	AltaVista (6)	AltaVista[WDS] (6)	Google (6)	Google[WDS] (6)
4	Google[WDS] (6)	AltaVista (6)	AltaVista[WDS] (6)	Google (6)
5	Google (6)	Google[WDS] (6)	AltaVista (6)	AltaVista[WDS] (6)
...	...	...	...	...

4.4.2 Systems

Four systems were used in total: two commercial web search engines (Google and AltaVista) and versions of each search engine that used the WebDocSum

interface.

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 commercial search engine they were using. Wrappers were developed for both search engines that preserved all content, but masked their identity. Google and AltaVista were referred to only as System A and System B.

The wrappers present the document abstract below the title and do not make use of a summary window. It was felt that the influence of this decision on the experimental results would be minimal. Searchers were asked to focus on the summary *content* not *presentation*, shifting the focus away from the appearance of extra interface features (i.e. the summary window).

Two systems using WebDocSum are also included. These systems aim to enhance Google and AltaVista and therefore enabled a summarisation versus non-summarisation comparison. These two systems were referred to as Systems C and D. System C is System A (Google) with WebDocSum and System D is System B (AltaVista) with WebDocSum.

At the end of the evaluation, users are asked to rank the systems based on their personal preference. It was felt that this task would be complicated if the interfaces were all identical. For this reason, a different colour was used for the interface of each system. The colours were made as distinct as possible, to avoid any confusion.

The systems were only referred to by their given names (A through D) for the course of the evaluation, every effort was made to mask the identity of the underlying search engine. All hyperlinks, page titles and HTML source referred only to the systems' evaluation names.

4.5 Search Tasks

Simulated work tasks are intended to replicate an actual information seeking session. Four tasks were used 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 the 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 topics and the tasks themselves are briefly outlined below:

- **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*

The complete tasks can be found in the Appendix.

4.6 User opinion and interaction

Three methods were used during the experiments to collect data on user opinion and their interaction with the search systems. These methods were:

- **Questionnaires**
  - Pre-search - prior to searching.
  - Post-search - after each task.
  - Final - after all tasks.
- **Think-aloud** - user informs and explains to the experimenter his or her actions.
- **Automatic Logging** - built into system (see next section).

4.6.1 Logging

Automatic logging was used throughout the course of the experiment and recorded the following:

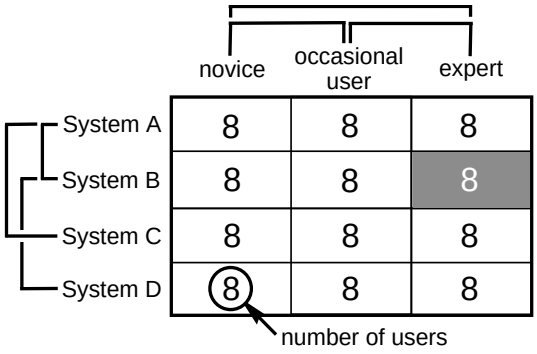
- The total number of documents returned (unique and duplicate).
- The number of summaries requested, and the number presented.
- The number of results pages viewed.
- Time for each session.

The results obtained from this background logging were used to compare the search systems. Section 5.2 describes in more detail the use to which these data were put.

4.7 Comparative Analysis

As mentioned earlier, the independent variable in these experiments is system type. Therefore, the focus of the analysis we undertake involves comparing user perceptions and search effectiveness between different combinations of these

systems. Figure 2 outlines the comparatory analysis, with system comparisons being shown by the solid lines on the left of the figure and translated on the right. By using this figure we attempt to clarify which comparisons are made and which are not.



The lines joining the system names can be translated into:

A - B → Google - AltaVista
A - C → Google - Google(WDS)
B - D → AltaVista - AltaVista(WDS)

Fig. 2. Comparative analysis

It may be apparent from the figure that not all possible inter-system comparisons are considered (i.e. we do not use {B - C}, {A - D} and {C - D}). Such comparisons are outwith the scope of this paper and we only seek to compare those systems that are relevant to testing our research hypothesis.

In Figure 2 we consider a *cell* to be a single user-group/system combination. For example the 'expert/System B' cell is shown in grey.

We concentrate on one user group at a time and *do not* compare each cell for that group with every other cell in the column (i.e. A vs B vs C vs D). Only the comparisons outlined on the right of the figure are considered and at any time we only compare at most two systems for a single group of users (i.e. A vs B *or* A vs C *or* B vs D). Mann-Whitney Tests are used to test for significance in the results obtained. This strategy is repeated for each of the three user groups.

We use the non-parametric Mann-Whitney Test for two reasons:

(1) uncertainty over whether the data is normally distributed means that parametric tests such as an ANOVA are not applicable, and;
(2) the small sample size in the comparisons again invalidates parametric testing.

Mann-Whitney Tests are used to calculate the significance of any inter-group differences (shown by the solid lines at the *top* of Figure 2). This is a pair-wise comparison of different user groups on the *same system* and as such, restricted to comparisons between two cells *in the same row* of Figure 2. All inter-group comparisons - for any one system - are considered (i.e. Novices vs Occasional Users vs Experts). These comparisons do not cover more than one system at a time and are therefore not intrinsic to testing the research hypothesis. They

are only included to test for any significant differences that could be attributed to user experience level.

4.8 Summary

The experimental design adopted to investigate the influencing effects of summaries in web searching has been outlined in this section. The experimental methodology was outlined, and the reasoning behind the design decisions made was given. The main experiment was described in detail and the main 'actors' were identified and their respective roles outlined. The means by which we captured user opinon and interaction were described, and our methodology for system (and user-group) comparison was outlined. In the next section, the results obtained from the described experimental procedure will be presented.

5 Experimental Results

The experiment described in the previous section was executed to test the research hypothesis stated, and through these means provide an assessment of the extent to which query-biased, web page summarisation helps users gauge document relevance. In this section we discuss the data collection and results. We look at both the effectiveness of the searches (including time and task success) and the users' perceptions of the systems.

5.1 Questionnaires

Questionnaires were the main way in which user opinion and feeling were captured. As is normal practice with such means (Lewis, 1995) there were a mixture of qualitative and quantitative questions that served to provide a rounded view of user opinion. In these questionnaires, two main methods were used to elicit opinion from the participants: *Semantic Differentials* and *Likert Scales*.

5.1.1 Semantic Differentials

Each respondent was asked to describe the various aspects of their searching 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 focused on the search process that the respondent had just carried out;

4 focused on the the summaries/descriptions presented by the system (Table 2).

Table 2
Semantic differentials

The task that we asked you to perform was...						
	very	reasonably	neither-nor	reasonably	very	
clear	**1**	**2**	**3**	**4**	**5**	**unclear**
complex	**1**	**2**	**3**	**4**	**5**	**simple**
familiar	**1**	**2**	**3**	**4**	**5**	**unfamiliar**
The search process you have just performed was...						
	very	reasonably	neither-nor	reasonably	very	
stressful	1	2	3	4	5	relaxing
interesting	1	2	3	4	5	boring
tiring	1	2	3	4	5	restful
easy	1	2	3	4	5	difficult
The summaries/abstracts presented by the system were...						
	very	reasonably	neither-nor	reasonably	very	
irrelevant	1	2	3	4	5	relevant
unimportant	1	2	3	4	5	important
not useful	1	2	3	4	5	useful
incomplete	1	2	3	4	5	complete

The result was a set of 1056 scores on a scale of 1 to 5: 24 respondents scoring each of the 4 systems on each of the 11 differentials. On the questionnaires, differentials were displayed as shown in Table 2. The arrangement of positive descriptors (e.g. 'interesting', 'relevant') and negative descriptors (e.g. 'boring', 'irrelevant') was randomised so that a positive assessment was sometimes represented by a low score (i.e. approaching 1) and sometimes a high score (i.e. approaching 5). This forced users to apply a little more thought to the completion of each differential. These scores were reversed at the analysis stage so that all positive descriptors were respresented by a low score.

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).

Considering the ordinal scale of the data, the Mann-Whitney test statistic

with a significance of $p < 0.05$ was used to test the one-tailed experimental hypothesis that the presence of query-biased web page summaries, for each document in a ranked list, is expected to improve the search effectiveness of system users.

Firstly we compared the task differentials (shown in bold at the top of Table 2) for *all* users on *all* systems, and found no significant difference between these. The task distribution was therefore comparable across all systems and no task-system bias was evident. We tested for this bias using the significance values from the task differential comparison, and if $p < 0.05$ (with a Mann-Whitney Test) we assumed bias existed. Placing one anomalous result aside ($p = 0.0303$ when comparing Experts on System A and System C) the semantic differentials pointed to no significant task bias. Indeed, for every other user group/system combination, $0.239 < p < 0.645$, well outside the $p < 0.05$ threshold.

On the effectiveness of summaries, we found that all user groups feel they benefit from the introduction of query-biased summarisation ($p = 0.0012$ for System C and $p = 0.0018$ for System D). All user groups, with the exception of the Experts using System C ($p = 0.4654$) and Infrequent Users on System D ($p = 0.1239$), appeared to benefit from the use of summarisation.

All user groups preferred the abstracts/summaries presented by the WebDocSum systems ($p = 0.0000$ for both System C and System D). When each user group was analysed in turn, all preferred the summaries/abstracts offered by the query-biased summarisers. Indeed, the largest significance value for the Mann-Whitney Test was 0.0253, well inside the 0.05 boundary imposed by the 95.95% confidence assumption. This shows that users preferred the query-biased, web page summaries over the standard search engine abstracts.

We used the same methodology to compare the abstracts offered by the two ranked titles/abstracts systems, AltaVista and Google. The difference between the two was minimal, indeed across all the semantic differentials AltaVista's average differential was 3.08 and Google's was 2.87 ($p = 0.3617$ at $p < 0.05$ with a Mann-Whitney Test). The only significant difference was how 'useful' the abstracts were. In the users' opinion, the document abstracts presented by Google (average = 3.17) were significantly more useful that those offered by AltaVista (average = 3.50). This was significant at $p = 0.011$, $p < 0.05$, again using a Mann-Whitney Test. This applied to all user groups (novices: $p = 0.034$, infrequent users: $p = 0.012$, experts: $p = 0.027$).

Finally, we compared the differentials of the overall search. The search process on Google with the 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/dificult' was

an no statistically significant difference found although the ratings did appear to favour 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 all users and within user groups.

5.1.2 Likert Scales

Each user was invited to indicate, by making a selection from a 5-point Likert scale (Preece, 1994), as shown in Table 3, 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 they wanted. These are used to measure whether the simulated work task situation placed them in actual context. We would expect the same level of response for all systems. An error here would indicate a potential error in the formulation of the tasks.
- The use of the summaries was helpful to assess the relevance of the web page.
- The abstracts/summaries showed the query in the context in which it occured in the web document being summarised.

Table 3
Example Likert scales for: *I encounter a task similar to this one frequently*

1	2	3	4	5
I agree completely				I disagree completely

Many Likert scales were used, Table 4. In a similar way to the semantic differentials, the postive descriptors (e.g. 'I agree completely') and the negative descriptors (e.g. 'I disagree completely') were randomised. Again, this was to ensure that subjects applied thought to the responses they gave.

Table 4
Likert scales used in the questionnaires

- *I encounter a task similar to this one frequently.*
- *I had an exact idea of the type of information I wanted.*
- *The abstracts/summaries helped me to assess the pages for relevance.*
- *The abstracts/summaries showed my query terms in context.*
- *I am happy with the results I found.*
- *I believe that I have succeeded in my performance of the task.*

In all counts, users scored similarly in questions dealing with the familiarity with the task and the exact idea of the information need. This indicates that during the evaluation, the task contexts were good across four tasks and four systems.

Users perceived the WebDocSum systems as being beneficial. They rated the summaries produced by the query-biased systems as more useful for indicating relevance ($p = 0.013$) and reported higher task satisfaction with these summaries than with the traditional descriptions ($p = 0.009$).

The two ranked title/abstracts systems, AltaVista and Google, were also compared using the same method. Google scored significantly better in the analysis of users' satisfaction with the results they obtained ($p = 0.012$) and how successful they felt they were at the task ($p = 0.017$). In the other scales individually there was no significant difference, but Google did fare better across them all ($p = 0.037$). All tests were carried out using a Mann-Whitney Test at the 99.95% confidence level.

5.1.3 Qualitative Data

In addition to the data collected through the aformentioned means, indicative qualitative data were also collected. The users were asked to express their opinions on the systems they had just used. These comments were collected both during the experiments using 'think-aloud', and in voluntary interviews conducted afterwards. These comments generally focused on three main aspects of the systems; interface design, results presentation and document abstracts/summaries. We will present the, at times contrasting, views of our users about each of these areas.

- **Interface features** - In general, users liked both styles of interface. The "familiarity" and "simplicity" of the ranked titles systems was cited as a major strong point by most users. The "conciseness", "simplicity" and "ease of use" of the WebDocSum interfaces were mentioned by users also.

- **Results features** - The results lists that were presented by systems of both kinds were subject to some criticism. Users disliked the need to scroll down the list in the ranked titles systems, and the need to move the mouse over a summary in the WebDocSum systems. One user commented that one of the main advantages of the WebDocSum systems, the presentation of all results on one page, was "tarnished by the need to move your mouse over a title to view a summary". This user felt that we were just replacing one problem (i.e. only viewing a few titles and abstracts at one time) with another, where it was only possible to view one summary at a time. The WebDocSum system also faced much criticism for not making the document URL clearly

visible.

- **Abstracts/summaries** - The abstracts presented by the ranked titles systems (A and B) were described as being "brief and irrelevant", with System B facing criticism from one user for failing to "show query terms in context". The WebDocSum systems performed well in this regard, with one user commenting that "summaries are detailed and in-depth". However, one user did remark that the "query terms still not shown in context but better than the A and B systems". This was despite the fact that the query terms were made bold in the sentences in which they resided.

Comments outwith these categories were also provided by some users. Four users commented on the ranked titles/abstracts approach. These comments varied, but included: "I hate these types of systems", "why dont they give me the results I want?", "quick, easy and reasonably effective" and "simple but not enough [content] in the abstracts". When invited to do the same about the WebDocSum approach, 2 users did so. The views voiced included a suggestion to "incorporate small summaries into the results list and larger ones in the summary window" and "I would definitely use this system if it was commercially available on the web".

5.1.4 Final Questionnaire

In the post-search questionnaire, subjects were asked to rank the four systems in order of preference. Table 5 shows the average ranking given to each of the four systems based on two criteria; which systems they felt helped relevance assessment, and which they liked best.

Table 5
Average ranking given to each system (range 1 - 5, lower = better)

	System A	System B	System C	System D
Relevance	3.38	3.29	1.92	1.46
Preference	3.38	3.29	1.88	1.46
All	3.38	3.29	1.90	1.46

23 out of the 24 respondents placed *a system* that used WebDocSum in either first or second place. 19 placed *both* WebDocSum systems in the top two ranking positions (i.e. one system in first and one in second place).

When scored on a scale of 1 (high rank) to 4 (low rank), the ranked title/abstract systems (A and B) obtained a mean score of 3.335 whereas the WebDocSum systems (C and D) obtained a mean score of 1.68. We applied the Kruskal-Wallis Test to the data and found that the WebDocSum systems

are significantly different from other systems (for System C, Z = -3.07 and for System D, Z = -3.92) at $p < 0.05$. The difference in the rankings obtained by AltaVista and Google was not significant.

The respondents who preferred the WebDocSum systems liked the way in which the results were presented, the simplicity of the interface and the increased length of summary. The single respondent who did not place any WebDocSum systems in the top two blamed the "unfamiliarity of the interface" as a major contributing factor.

5.2 Automatic Logging

Throughout the experiment, the systems' response to user activity was logged continuously in the background.

Table 6
Results from Automatic Logging

System	Web documents	Unique web documents	Queries
System A	1322	803	160
System B	1655	1236	187
System C	1202	843	86
System D	954	740	76

The average number of web addresses given by the search engine across the non-summarised systems is 1488.5 (62.02 or just over 6 result pages per user); across the summarised systems is 1123 (46.79 or just around 4 and a half result pages per user). 68.5% of the pages retrieved by the non-summarised systems were unique, whereas 73.4% of the pages retrieved by summarised-systems were unique. A web document is 'unique' if it is only suggested by the system once per search task. We can use 'uniqueness' to measure a system's reponsiveness to query evolution (i.e. a high level of uniqueness implies a high level of responsiveness to a changing query). Query reformulation implies some degree of user dissatisfaction with the current result list. A system that, in response to this changing query, presents a high proportion of documents that were offered before reformulation can be thought as being less responsive to query evolution and less helpful to users. The total number of search iterations (queries submitted) on the non-summarised systems was 173.5 (mean of 7.22 per user) and on summarised systems this figure was 81 (mean of 3.375 per user). We used a Mann-Whitney Test at $p < 0.05$, and were able to show that the difference in the total number of search iterations (hence queries) was significant ($p = 0.002$). In contrast, the system comparisons using 'web documents' and 'unique web documents' were not significant ($p = 0.0565$ and

p = 0.2334 respectively).

Out of the 1322 web addresses that Google (System A) presented, 803 (or 60.7%) of were unique. In contrast, 74.7% of those presented by AltaVista (System B) were unique. This was significant, p = 0.0385, using a Mann-Whitney Test at $p < 0.05$. This suggests that AltaVista is more responsive to a changing query than Google. There was no significant difference between the number of queries submitted to each system (p = 0.117).

5.3 Search tasks

There were 24 subjects in total, and each user carried out 1 task on each of the 4 systems. The tasks were allocated as shown Table 1 in section 4.4.1.

The session times for each task will be compared. The session time gives an idea of how long it took to complete a task. If a user failed to complete a task in the 10 minute time allocated, a time of 10 minutes would be recorded for that user/task combination. Unless they had completed a task, on no occasion did a user stop before the 10 minute period has passed. Table 7 shows the average session time for all 4 systems.

Table 7
Average search session time

System	System A	System B	System C	System D
Average Time	8mins 53secs	9mins 21secs	6mins 31secs	6mins 47secs
Normalised Time	20mins 43 secs	21mins 48secs	6mins 31secs	6mins 47secs

The results above show an apparent link between the length of time taken to complete a task and the type of system used. To verify the significance of these results, a Mann-Whitney Test was carried out at $p < 0.05$. This test showed that the results were significant (p = 0.034) and the hypothesis that WebDocSum reduces search times was proven. The response times of the WebDocSum (7 seconds) and traditional ranked titles (3 seconds) systems differed by around 4 seconds. We take this difference into account and show the *normalised* task time in the second row of Table 7. The difference between the two system types is even more marked.

We compute this normalised time by calculating how much faster the traditional systems (A and B) are than the WebDocSum systems (C and D) and multiplying the result (i.e. $\frac{7}{3} \approx 2.333$) by the traditional system times. Such calculation gives us the *expected time* had system response speeds been the same. We do not alter the times for systems C and D as these are the bases for this comparison.

It is also worth looking at the number of people who completed each task using the two different types of system. Due to the limitation in the number of tasks, it is worth grouping the results from System A and B together, and doing the same for C and D. Table 8 shows the results obtained.

Table 8
Number of users who fully completed tasks

Task	Non-summarisation	Summarisation
Task 1	3	5
Task 2	4	4
Task 3	2	4
Task 4	4	6

There appears to be a slight correlation between the number of tasks fully completed and the type of system used. The summarisation systems have an average of 4.75 users completing their tasks, whereas the non-summarised systems have 2.75. This difference is significant ($p = 0.043$ using the Mann-Whitney Test) despite the small nature of the sample involved.

There was no significant difference between task time ($p = 0.5080$) or number of tasks completed ($p = 0.651$) when comparing the ranked titles/abstracts systems.

6 Discussion

In the previous section we outlined the experimental results obtained. It is possible to categorise the results in three ways; the validity of the experimental tasks, the role of summarisation and system comparisons. We will now endeavour to analyse these results in more depth, category by category.

6.1 Experimental Tasks

To ensure that our tasks (shown in Appendix) were not biased toward any particular system, we used semantic differentials to elicit user opinion on the clarity, difficulty and familiarity of the tasks. We tested the statistical significance of the differences in results and found, with one exception (Expert users in the Google/Google with WebDocSum comparison) there was no experimental task bias. This anomolous result defies explanation. For all other groups (including Expert users in the AltaVista/AltaVista with WebDocSum

comparison) such bias is not evident and all tasks appeared to be of a similar level of difficulty.

6.2 Role of Summarisation

6.2.1 Summaries

The experimental questionnaires showed that users rated the WebDocSum systems' *summaries* as more 'relevant', 'important', 'useful' and 'complete' (all four of the assessment categories) than those presented by the traditional ranked titles/abstracts systems. All differences were statistically significant across users and within each of the three user groups (novice, infrequent, expert). This indicates a user preference for the query-biased summaries as a document representation.

Every user and every user group indicated a significant preference (taken from *all* 'search process' and 'summary relevance' Semantic Differentials) for the query-biased summarisation systems ($0.0001 < p < 0.0362$). There was one exception who cited the 'unfamiliarity' of the WebDocSum interface as the reason for their dislike. It may also be useful to note that this participant was the oldest subject in the evaluation (62 years of age).

Users assessed the systems using Likert Scales based on (among others) the beneficial effect of the summaries, their success/failure at the task and their happiness with the results they obtained. All user groups appeared to prefer the summarised systems ($0.0000 < p < 0.0123$) with the exception of Expert users on Google with WebDocSum ($p = 0.0931$).

Through using query-biased summarisation, we hoped that as well as producing larger, real-time summaries of web documents we would also be able to improve the way in which query terms were shown in context. In our experiments, users pointed towards a significant improvement in our systems compared with AltaVista ($p = 0.001$), but no such significance in the comparison with Google ($p = 0.0741$). Google incorporates a certain degree of query-biasing in the abstracts it shows. Users were asked to rate (on a Likert Scale of 1 to 5, where 1 is the highest) the extent to which each system placed query terms in context. Although there was a trend toward the WebDocSum systems; Google: $\bar{x}$ [6] $= 2.96$, WebDocSum: $\bar{x} = 2.21$, the difference was not significant at $p < 0.05$. The problem was perhaps not that our systems' query-biasing did not rival, or even better, that of Google, but more that *the differences were not apparent to all users*. Expert users gave a large difference between Google and WebDocSum (Google : $\bar{x} = 3.25$, WebDocSum

[6] $\bar{x}$ is used to denote the mean or average

: $\bar{x} = 2$ where p = 0.0545), whereas Novice users did not (Google : $\bar{x} = 2.88$, WebDocSum : $\bar{x} = 2.5$ where p = 0.4354). The results for infrequent users lay inbetween these two user groups (Google : $\bar{x} = 2.75$, WebDocSum : $\bar{x} = 2.13$ where p = 0.2988). As the experience level of the user groups increased, there was an improvement in their ability to distinguish between the summaries offered by Google and those offered by WebDocSum. When we compared AltaVista (which does not query-bias its summaries) and WebDocSum, all users preferred how the WebDocSum system showed the query terms in context.

6.2.2 Interface

Users seemed to prefer the simpler, more concise interface of the WebDocSum systems over their ranked-list counterparts. Users felt that the summaries were of a sufficient length and depth to allow a good assessment of document relevance to be undertaken without referring to the full text.

Participants disliked the need to scroll up and down the traditional web search result lists to view all entries. We sought to counteract this problem by elimating the need to scroll. However, one user felt that the restriction imposed by WebDocSum that only one summary could be viewed at a time was simply replacing one problem with another. It could be argued that WebDocSum systems therefore encouraged a more methodolgical perusal of the results list, but this was not tested as part of this study.

6.2.3 Time to Search

We found that users took longer to complete tasks on the ranked titles/abstracts systems than on the WebDocSum systems ($\bar{x}$ for traditional web search systems: 9mins 14secs, $\bar{x}$ for WebDocSum systems: 6mins 39secs). This difference was yet more marked when we normalised for the differences in system response time. We also found that users completed more tasks on the WebDocSum systems ($\bar{x} = 4.75$ users per task) compared with the ranked titles/abstracts systems ($\bar{x} = 2.75$ users per task), and submitted signicantly fewer queries. The task completion difference is significant at p = 0.043, and the query difference at p = 0.002. The summarisation systems appear to provide the user with a means by which their search effectiveness - in terms of tasks completed and time taken - can be improved.

6.3.1 WebDocSum and Traditional Web Search Systems

Participants were asked to rank the systems based on both personal preference and relevance assessment. Our results pointed to a preference for WebDocSum systems over the traditional web search systems in both of these ranking. There also appeared to be a link between relevance and personal preference. In this study, users (22 out of 24) assigned the same ranking for relevance assessment as they did for personal preference. However, it is also possible for users to have ranked the relevance of summaries placing an undue weight on personal preference for a particular system, or type of system.

6.3.2 Google and AltaVista

The same experimental methodology that was used to compare summarised and non-summarised systems was also used to compare two traditional web search engines, Google and AltaVista.

There was a significant difference in the number of 'unique' web pages (a document that is only suggested once per task) returned over all the tasks and all the users. 60.7% of the pages returned by Google were unique, opposed to 74.7% in the case of AltaVista ($p = 0.038$). From this, it may be argued that AltaVista is more responsive to a user's changing query, suggesting a greater amount of *different* documents as the query evolves.

The 'usefulness' of a summary was defined as the extent to which it was of use in determining the relevance of a document. Users appeared to find the abstracts presented by Google more 'useful' (novices: $p = 0.034$, infrequent users: $p = 0.012$, experts: $p = 0.027$) than those offered by AltaVista. In the other differentials; 'relevance', 'importance' and 'completeness', there were no significant differences. Users felt more happy with Google's result list and with the notion of task completeness on Google ($p = 0.012$ and $p = 0.017$ respectively). Overall, across all Likert Scales and all user groups, Google was preferred to AltaVista. However, it should be noted that this was a comparison of the search engine's document abstracts and their link to search effectiveness, and *not* a comparison of the whole interface's role.

There was no significant difference in the personal preference and relevance assessment rankings that users assigned to the Google and AltaVista systems. In this study, the trend points to a preference for Google however, with users placing more weight on the quality of a search engine's document abstracts than any other factors.

The experimental methodology developed to test the influencing effects of query-biased summarisation in web searching, was also used to compare the effectiveness of two major commercial search engines. The methodology's generic nature would make it equally applicable for the comparison of any web search systems.

The results and observations establish the usefulness of query-biased, web page summarisation systems and in parts point to differences between the quality of summary and effectiveness of AltaVista and Google.

7 Conclusions and Future Work

7.1 Conclusions

The main aim of this paper was to investigate the effectiveness of automatically generated summaries of web documents that were tailored to the needs of the user expressed in a textual form via the query. The research was motivated by the lack of such a system for the web and the lack of a formative, methodological task-oriented web study. The work reported in this paper has outlined a query-biased web page summarisation system, its related evaluation and the experimental results obtained. Coupled with this, we have also proposed a generic experimental methodology for evaluating web searches.

The conclusions drawn from the experimental results were that automatically generated web page summaries allowed the user to gauge document relevance more effectively than those presented by the traditional ranked title/abstract approaches. On WebDocSum systems users completed more tasks, submitted fewer queries, viewed less result pages and completed tasks up to 27% faster than traditional ranked titles/abstracts systems. Novices and Infrequent Users appeared to gain the most benefit from the use of the summarisation systems, although Experts' search effectiveness did improve as well

In addition, the experimental methodology was useful in comparing the effectiveness of traditional web search systems. We found that AltaVista showed more 'unique' web documents than Google, suggesting that AltaVista may be more responsive to query evolution. However, Google provided a higher level of task satisfaction and its document abstracts were significantly more 'useful' (as an aid to relevance assessment) than AltaVista's. There was no significant difference in search effectiveness between Google and AltaVista.

In this paper we have described a query-biased summarisation system for web search engines and an evaluation of its effectiveness. 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.

An experimental methodology developed to test our research hypothesis and outlined in this paper, has been proposed as a generic evaluatory medium for testing the effectiveness of web searches.

We have used this methodology to compare the summaries of AltaVista and Google, and shown that in parts Google performs better than the AltaVista. Our results show that users appear more happy with their results, completed the tasks more frequently and find the abstracts of more use when using Google, opposed to AltaVista.

7.2 Future Work

The results presented in this paper have proved the effectiveness of the query-biased, web page summariser. It was apparent that the presence of automatically generated summaries, providing users with a more in-depth, detailed summary than is currently provided by web search systems, had a positive effect on relevance assessment. However, this research is by no means definitive, and there are a few areas on which future research could focus. These center on ways to improve the experimental methodology and modify the system.

7.2.1 Experimental Design

- The system has been evaluated in an experimental environment, where great care has been taken to control the situational variables that can impart upon a user's search experience. Unfortunately, web searching is not like this and users often have to use search systems in operational, practical environments such as the workplace. In such arenas, the cognitive load placed on users is increased significantly and the extent to which they can concentrate on the task in hand decreases. It may be worthwhile to carry out an ethnographical study to determine whether the results obtained follow a similar pattern to its laboratory equivalent.

7.2.2 System

- The sentence scoring algorithm used to score the sentences in the summarisation system uses a number of numerical constants. These constants have only been created via intuition and no research has been carried out to test their validity. Through testing, we were able to tailor the constants to produce what was thought to be the most effective summary. However, the validity of these assumptions has not been fully tested and empirical research into their effectiveness may be necessary in the future.
- Different ways of displaying the result list and summary window combination could be tested to ensure that the best use of the available screen space is made.
- Finally, different approaches to focus user's attention towards the context in which the query terms appear within the full text of the web page (i.e. allowing the user to view the summarised document with summary sentences highlighted) would be an interesting area for future research.

Acknowledgements

We would like to thank the fellow members of the Information Retrieval Group at the University of Glasgow for their thoughts and assistance. The comments and enthusiasm of our experimental subjects were greatly appreciated; our thanks goes out to them. We would also like to thank the anonymous reviewers, whose comments were very helpful.

Appendix : Search Tasks

Search for a fact

Assume that you are a research student and have just finished reading a very interesting article from a popular journal in your area of research. It has been five years since the article was first published, but you note that the author is Jan-Jaap IJdens from the Robert Gordon University, Aberdeen. You have a keen interest in what the article discusses and would like to send an electronic mail to the author. However, you contact the university and find that Dr IJdens has moved, leaving no forwarding e-mail address. Your task is to find his current e-mail address.

Search for a number of items

Next weekend, a close friend of yours is hoping to go on a short-break to Paris, France. He has recently moved house and does not have a phone line installed. As a result he has asked you to look for hotels in the city on his behalf. Both of you are not too confident about your French speaking skills and would like to find hotels that offer an online registration service. Your friend expects to get Internet access again soon and he would like both the registration page's web address and the page ID from five such hotels in the city, so that he can pursue the booking himself.

Decision search

You are about to depart on a short-tour along the west coast of Italy. The agenda includes a visit to the country's capital, Rome, during which you hope to find time to pursue your interest in impressionist paintings. As your time in the city is limited to only two(2) hours you would like to save time and find information about the city's best impressionist art museum prior to your departure.

Background search

You work in an old building and one of your colleagues has developed a severe dust allergy which you believe is caused by his working environment. He is writing a letter to complain about the lack of cleanliness in your working environment and has asked you to help find information about dust allergies.

References

Amitay E. and Paris C. (2000). Automatically summarising web sites - is there a way around it?. In *Proceedings of ACM 9th International Conference on Information and Knowledge Management*. November 6-11, 2000. (pp. 173-179). Washington DC, USA.

Brandow, R., Mitze, K. and Rau, L.F. (1995) Automatic condensation of electronic publications be sentence selection. In *Information Processing and Management*, 31 (5). (pp. 675-685).

Borlund, P. (2000) Evaluation of information retrieval systems, PhD Thesis, Department of Information Studies, Abö Akademi University, Finland.

Borlund, P. (2000) Experimental Components for the evaluation of interactive information retrieval systems. In *Journal of Documentation*, 56 (1). (pp. 71-90).

Crovella, M. and Bestavros, A. (1996) Self-similarity in world wide web traffic: evidence and possible causes. In *Proceedings of SIGMETRICS'96*. May 23-26, 1996. Philadelphia, PA, USA.

Ding, W. and Marchionini, G. (1996) A comparative study of web search service performance. *ASIS 1996 Annual Conference Proceedings*. October 19-24, 1996. (pp. 136-142). Baltimore, MD, USA.

Dempsey, B.J., Vreeland, R.C., Summer Jr., R.G. and Yang, K. (2000) Design and empirical evaluation of search software for legal professionals on the WWW. In *Information Processing and Management*, 36 (2). (pp. 253-273).

Edmundson, H.P. (1964) Problems in automatic abstracting. In *Communications of the ACM*, 7 (4). (pp. 259-285).

Edmundson, H.P. (1969) New methods in automatic abstracting. In *Journal of the ACM*, 16 (2). (pp. 264-285).

Ellis, D. (1996) *Progress and Problems in Information Retrieval*. London: Library Association Publishing

Gordon, M. and Pathak, P. (1999) Finding information on the world wide web: the retrieval effectiveness of search engines. In *Information Processing and Management*, 35 (2). (pp. 141-180).

Jansen, B.J., Spink, A. and Saracevic, T. (2000) Real life, real users, and real needs: a study and analysis of users on the web. In *Information Processing and Management*, 36 (2). (pp. 207-227).

Jose, J.M., Furner, J. and Harper, D.J. (1998) Spatial querying for image retrieval: a user-oriented evaluation. In *Proceedings of the 21st Annual International SIGIR Conference on Research and Development in Information Retrieval*. August 24-28, 1998. (pp. 232-240). Melbourne, Australia.

Kupiec, J., Pedersen, J. and Chen, F. (1995) A trainable document summarizer. In *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. July 8-13, 1995. (pp. 68-73). Seattle, WA, USA.

Lernaut & Hauspie Inc. (online) *Intelliscope Retrieval Toolkit* : http://www.lhsl.com/tech/icm/retrieval/toolkit/default.asp

Lewis, J.R. Computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. In *International Journal of Human-Computer Interaction*, 7 (1). (pp. 57-78).

Luhn, H.P. (1958) The automatic creation of literature abstracts. In *IBM Journal of Processing and Development*, 2 (2). (pp. 159-165).

Miike, S., Itoh, E., Ono, K. and Sumita, K. (1994) A full-text retrieval system with a dynamic abstract generation function. In *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. July 3-6, 1994. (pp. 152-161). Dublin, Ireland.

Miller, S. (1984) *Experimental design and statistics (2nd edition)*. New York: Routledge.

Over, P. (1998) TREC-6 interactive track report. In *Proceedings of the 6th Text Retrieval Conference*. NIST Special Publication. (pp. 73-82).

Paice, C.D. (1981) The automatic generation of literature abstracts: an ap-

proach based on the identification of self-indicating phrases. In Oddy, R.N., Robertson, S.E., van Rijsbergen, C.J. and Williams, P.W. (eds.) *Information retrieval research*. London: Butterworths. (pp. 172-191).

Perlman, G. (online) *Web-based User Interface Evaluation with Questionnaires* : http://www.acm.org/ perlman/question.html

Preece, J. (ed.) (1994) *Human-Computer Interaction*. Addison-Wesley.

Robertson, S.E. and Hancock-Beaulieu, M.M. (1992) On the evaluation of IR systems. In *Information Processing and Management*, 28 (4). (pp. 457-466).

Ruthven, I., Tombros, A. and Jose, J.M. (2001) A study on the use of summaries and summary-based query expansion for a question-answering task. In *Proceedings of the 23rd BCS European Annual Colloquium of Information Retrieval Research*. April 4-6, 2001. (pp. 41-53). Darmstadt, Germany.

Salton, G, Singhal, A., Mitra, M. and Buckley, C. (1997) Automatic text structuring and summarisation. In *Information Processing and Management* 33 (2). (pp. 193-207).

Sparck Jones K. (ed.) (1981) *Information Retrieval Experiments*. London: Butterworths.

Swanson, D.R. (1986) Subjective versus objective relevance in bibliographic retrieval systems. In *The Library quarterly*, (56). (pp. 389-398).

Tague-Sutcliffe, J. (1992) The pragmatics of information retrieval experimentation, revisited. In *Information Processing and Management*, 28 (4). (pp. 467-490).

Tombros, A. and Sanderson, M. (1998) The advantages of query-biased summaries in information retrieval. In *Proceedings of the 21st Annual International SIGIR Conference on Research and Development in Information Retrieval*. August 24-28, 1998. (pp. 2-10). Melbourne, Australia.

White R., Jose. J.M. and Ruthven, I. (2001) Query-biased web page summarisation : a task-oriented evaluation. In *Proceeings of the 24th Annual International SIGIR Conference on Research and Development in Information Retrieval*. September 9-13, 2001. (pp. 412-413). New Orleans, USA.

White R., Ruthven, I. and Jose, J. (2001) Web document summarisation : A task-oriented evaluation. In *Proceedings of the the 1st International Workshop on Digital Libraries*. September 3-7, 2001. Munich, Germany.