Finding Relevant Documents using Top Ranking Sentences: An Evaluation of Two Alternative Schemes

Ryen W. White
Department of Computing Science
University of Glasgow
Glasgow. G12 8QQ
whiter@dcs.gla.ac.uk

Ian Ruthven
Department of Computer and Information Sciences
University of Strathclyde
Glasgow. G1 1XH
ir@cis.strath.ac.uk

Joemon M. Jose
Department of Computing Science
University of Glasgow
Glasgow. G12 8QQ
jj@dcs.gla.ac.uk

ABSTRACT

In this paper we present an evaluation of techniques that are designed to encourage web searchers to interact more with the results of a web search. Two specific techniques are examined: the presentation of sentences that highly match the searcher's query and the use of implicit evidence. Implicit evidence is evidence captured from the searcher's interaction with the retrieval results and is used to automatically update the display. Our evaluation concentrates on the effectiveness and subject perception of these techniques. The results show, with statistical significance, that the techniques are effective and efficient for information seeking.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: - search process, relevance feedback.

General Terms

Experimentation, Human Factors

Keywords

WWW, user studies, sentence extraction

1. INTRODUCTION

Web search engines are an essential tool for finding useful resources on the Internet. However, even though these tools are intended to *facilitate* access to the web, user studies such as [5] demonstrate that searchers display very limited interaction with search engine interfaces. For example, searchers typically input very short queries, around 1 or 2 search terms and typically do not use advanced search facilities to generate more complex queries.

One main conclusion from such user studies is that users are reluctant to view result lists beyond the first page of results, preferring instead to opt for query reformulation and resubmission. For this reason many web page designers place much importance on the creation of web documents that consistently appear in first page of popular web search engines. However, the reluctance of searchers to look beyond the first page of results means that they may miss out on useful documents and will have to generate new search terms for future searches. This type of search behaviour is one in which searchers run many searches, but with limited interaction with the retrieved set of documents. This is not only computationally expensive – many searches have to be run – but it is also cognitively expensive for the searcher to provide additional search terms. Previous research has indicated that reduced querying and increased interaction with the results of web searches can increase search effectiveness, e.g. [9].

In this paper we propose techniques for encouraging searchers to more fully examine the results of a web search, thus reducing the computational and cognitive loads. This research is based on two related research issues. Firstly we propose techniques that *recommend* retrieved pages to searchers. In particular we use sentences from retrieved web pages that highly match the searcher's query as a means of showing the searcher what kind of information has been retrieved. This system is intended to present the searcher with information on the *whole* retrieval result rather than just the top ten pages.

A second system uses a form of relevance feedback, [8], to automatically update the information that the interface displays to the searcher. In this case the system uses *implicit* information – information captured from the searcher's interaction with the interface – to estimate what information may be of use to the searcher. This system is intended to make the search interface less *passive*. That is, the system is intended to make a stronger connection between the searcher's interaction and the information presented.

Both systems are used to encourage searchers to interact more fully with the retrieval results of a web search. Our experiments compare these systems with each other and with an advanced web summarization system. The choice of this type of system as our baseline system is motivated by evidence, e.g. [10], that summarization techniques can also increase the searchers' interaction with search results.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'02, August 11-15, 2002, Tampere, Finland.
Copyright 2002 ACM 1-58113-561-0/02/0008…$5.00.

The remainder of the paper is structured as follows. In section 2 we discuss the motivation behind our investigation and related work. In sections 3-5 we describe the systems we developed and in section 6 we present our experiments. In section 7 we highlight the main results from the experiments and we conclude with a discussion in section 8.

2. MOTIVATION

Most web search interfaces present the user with little information with which to decide whether or not to view a retrieved page. Typically the only information shown is the page title, URL and short (1-2 line) text fragments from the retrieved pages. These text fragments normally contain at least one instance of the query terms and are intended to give the user some notion of the context in which the query terms are used in the web page.

This minimal approach to describing the retrieved set of pages suffers from two flaws: the page descriptions generally do not contain sufficient information to help the user decide whether or not to view the page, [10] and the descriptions only relate to the titles displayed on the current page; there is no indication to the searcher of what other information has been retrieved.

The former problem – insufficient descriptions of the retrieved pages – is typically handled by summarization techniques that give fuller descriptions of the retrieved pages [10, 2]. In this paper we examine the second problem, that of helping the searchers perform more complete analyses of the retrieved pages.

Our intention behind these systems and the experiments we describe is to move away from the web pages themselves and to present the searchers with *indications* of the type of information to be found in the retrieved pages. In both our experimental systems, sections 4 and 5, this takes the form of presenting the searchers with those sentences, from the top thirty ranked pages, that highly match the searcher's query.

By presenting whole sentences to the searcher, we are presenting the query terms in the local context in which they are used within retrieved pages. The use of sentences allow the searcher to see representative samples of a page's text before s/he accesses the full-text of the retrieved page and hence to make better decisions as to what pages to view.

The use of sentences as a context device has been investigated elsewhere, e.g. Magennis and Van Rijsbergen, [6], used sentences to show users the context of expansion terms in relevance feedback. In Magennis and Van Rijsbergen's experiments the sentences came from documents already viewed by the user. In our experiments the sentences come from web pages that have *not* already been viewed by searchers. Sentences have also been the basis of many successful summarization systems. In these systems, sentences from a document which highly matches the searcher's query are used to compose a summary of the page. Our baseline system, section 3, against which we compare our experimental systems is a summarization system built on these principles.

In our experiments we use three search systems. Each system consists of an interface, with underlying functionality, which connects to existing web search engines. In all our experiments the underlying search engine is AltaVista[1]. We chose this search engine because it allows connections from external interfaces[2] and is one of the most-used publicly available search engines.

The first system – our baseline system – is a system which offers the searcher the opportunity to view short summaries of retrieved documents, section 3. The second system, section 4, extends the interface to display top-ranking sentences. The third system, section 5, automatically updates the list of top-ranking sentences in light of the searcher's interaction with the search results.

3. SYSTEM 1 – SUMMARIZATION

Our first system, System 1, is a summarization system. Once the underlying web search engine has performed a retrieval, System 1 downloads and summarises the top thirty ranked web pages.

The summaries are created through a sentence extraction model, presented in [10], in which the web pages are broken up into their component sentences and scored according to factors such as their position (initial introductory sentences are preferred), the words they contain (words that are emphasised by the web page author, e.g. emboldened terms are treated as important), and the proportion of query terms they contain. The latter component – scoring by query terms – biases the summaries towards the query. A number of the highest-scoring sentences are then chosen as the summary.

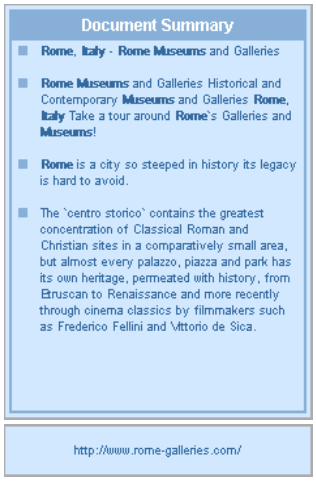

Figure 1: Web page summary

The system displays the titles of the retrieved pages in groups of ten titles. If the searcher passes the mouse over a retrieved title, a summary window appears next to the title. The summary window displays the document title, the top four highest-ranking sentences and the URL of the page, Figure 1.

[1] *http://www.altavista.com*

[2] Unlike some other search engines at system-build/experiment time, e.g. Google.

This form of web page summarization has previously been shown to increase searchers' satisfaction and search task success over conventional search engine interfaces, [10]. Consequently we use this system as it forms a stricter baseline system rather than a more standard web search interface.

4. SYSTEM 2 – TOP RANKING SENTENCES

The second system, System 2 is an extension of System 1. In a similar way, the System 2 interface displays the titles of web pages in groups of ten, and presents summaries on request. However, the System 2 interface also displays a list of *top-ranking* sentences. In this system the same sentence extraction method that is applied to *individual* documents in System 1 is applied to the top thirty documents as a *set*. That is, we split the top thirty documents into component sentences and rank the sentences according to the words they contain and the query words they contain. This allows a ranking of the sentences from the top thirty documents – the top ranking sentences. More than one sentence per document can appear in the list of top-ranking sentences. The sentences themselves are displayed to the searcher as shown in Figure 2.

Figure 2: Top-ranking sentences

If the searcher moves to a subsequent page of results, the system will use the next thirty pages to create the top-ranking sentences, e.g. if the searcher moved to the second page of results, then the sentences from pages in rank positions 11-40 would be used to create the top-ranking sentences.

The intention behind the top-ranking sentences is to encourage searchers to target useful information rather than simply assess the ranked list of titles themselves. That is we are attempting to get the user to make assessments on the content of pages rather than the more usual representations of titles and text fragments.

Figure 3: Document title window

If the searcher passes the mouse over a sentence the system will indicate which page contains this sentence. If the page is in the ten results currently being displayed, the system will simply highlight the title of the page. If the page is outwith this group of ten results, the title of the page containing the sentence will be shown in a small window below the main result list, Figure 3. Passing the mouse over this title will show a summary of the document.

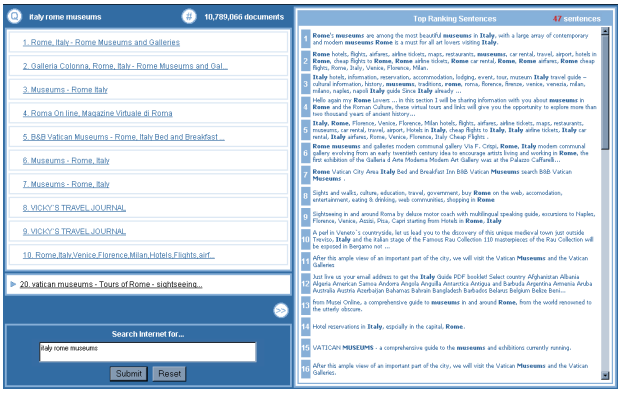

Figure 4: Top ranking sentences interface (Systems 2/3)

The system, thus, encourages searchers to view pages that appear later in the ranking, i.e. pages that have a lower retrieval score than the pages that are currently being listed. The aforementioned components are combined to form the interface of the system, as shown in Figure 4.

5. SYSTEM 3 – IMPLICIT FEEDBACK

The previous system, System 2, presents the searchers with indications of potentially relevant material – the top-ranking sentences. The query terms are shown in the *context* of the sentence in which they appear and the sentences are selected by their similarity to the query.

This notion of context in System 2 is static – the composition of a sentence does not change over time and the query is assumed to be constant within an individual search iteration. Our third system, System 3, uses the context of the search itself to improve the ranking of the sentences displayed to the searcher. The context in this case is the *implicit* information given by a searcher whilst interacting with the search system interface. As will be discussed below, implicit information is the evidence the searcher gives by viewing a page summary. This information is regarded as being implicit as the searcher does not give the information with the explicit purpose of changing the search results.

The context information is used to update the system's representation of the searcher's query, which in turn is used to re-rank the list of top-ranking sentences. Specifically, our notion of context is based on the assumption that searchers will spend a longer time reading interesting (potentially relevant) material and less time reading irrelevant material.

Several studies, e.g. [7], have found a correlation between the positive relevance of a document and viewing time. These studies, focusing on static corpora and full-text documents, have found that if a document is subjected to a lot of 'read wear' [4] it is likely to be relevant. In this study, we assume that these findings will hold for document summaries. That is we assume that summaries which searchers view for longer than expected are those that contain information similar to that desired by users. We discuss how we decide upon a searcher's

expected summary reading time in section 6.2. We use summaries as we can detect which summaries a searcher has assessed and for how long (unlike titles) and searchers tend to view more summaries than web pages leading to more evidence. Any summary that the system believes contains relevant information is used for query modification.

The previous system, System 2, ranks sentences based on the searcher's query. System 3 ranks sentences based on the searcher's query *and* terms taken from the content of the *assumed* relevant summaries. Each time the system believes the searcher has identified a useful summary the content of all useful summaries is used to generate a list of possible query expansion terms. The function we used for this purpose is Robertson's *wpq* formula [8], equation 1.

The top 6 terms, from the *wpq* ranking of terms, are added to the original query, and the new query is used to re-rank the list of top-ranking sentences. This system, then, dynamically updates the list of top-ranking sentences each time the system assumes that the searcher has found a relevant summary.

$$wpq_t = \log\left(\frac{(r_t + 0.5)(N - n_t - R + 0.5)}{(n_t - r_t + 0.5)(R - r_t + 0.5)}\right) \times \left(\frac{r_t}{R} - \frac{n_t - r_t}{N - R}\right)$$

Equation 1: *wpq* formula, where r_t is the number of relevant summaries which contain term t, n_t is the number of summaries which contain term t, R is the number of relevant summaries, N is the number of summaries.

The automatic updating of the top-ranking sentences is intended to reflect the searcher's interaction with the system. That is, we are relating the searcher's search *actions* with the information displayed by the interface. Previous research, [9], has shown that this form of implicit relevance feedback can act as an effective substitute for explicit relevance feedback in which the searcher must *explicitly* indicate, e.g. by checkboxes, what pages or summaries are relevant. Explicit feedback gives more accurate information on what a user finds relevant but searchers are often reluctant to make these assessments, e.g. [3]. Implicit feedback, on the other hand, exploits the searchers' existing interaction with the system.

We use these systems to investigate two main research questions. Firstly, we investigate whether the use of top-ranking sentences encourages searchers to interact more fully with the retrieval results, and whether this leads to more effective searching. This is effectively a comparison between System 1 and Systems 2. Secondly, we investigate the effect implicit relevance feedback on the results obtained from the first research question. That is, does implicit relevance feedback improve searchers' perceptions of the system and does it lead to more effective interaction with the system. This is a comparison between System 2 (no feedback) and System 3 (implicit feedback). In both cases we shall consider measures of both search task success and searchers' perceptions of the systems. In the following section we present the experiments we carried out using these systems.

6. EXPERIMENTS

In this section we discuss the experimental methodology we used in our experiments, the search tasks used, and the experimental subjects who participated.

6.1 Experimental methodology

In our experiments 24 experimental subjects, section 6.4, each completed 3 search tasks, section 6.3, one search task was completed on each of the three systems. The presentation of tasks to subjects was held constant: each subject performed the search tasks in the same order, however the order of presentation of systems was rotated across subjects. Each subject was given 10 minutes to complete each task, although the subjects could terminate the search early if they felt they had completed the task.

The subjects were welcomed and given a short tutorial on the features that were incorporated into the three systems being tested. We also collected background data on aspects such as the subjects' experience and training in online searching. After this, subjects were introduced to tasks and systems according to the experimental design.

Once they had completed a search, the subject was asked to complete questionnaires regarding various aspects of the search. We used Semantic differentials, Likert scales and open-ended questions to collect this data. After the third experiment a final questionnaire was completed that asked searchers to rank the three systems based on their personal preference. In addition, we conducted semi-structured interviews after each search and after the experiment as a whole. Background logging was also used to record user interaction with the systems.

6.2 Timing tasks

Information was also collected on the time taken by subjects to view summaries for use by System 3. This system uses implicit evidence to update the list of top-ranking sentences. As described in section 5 this system assumes that subjects will spend longer viewing summaries that are interesting, or potentially relevant, than summaries that do not appear useful. To allow us to decide when a summary should be counted as useful we need a measure of how long a searcher will take to assess a summary's content. In particular we need to measure how long on average an individual searcher will take to decide whether a summary is relevant and how long the searcher will take to decide the summary is not relevant. To do this consistently we used a timing task before each search task.

We gave each subject a search description and presented them with a prepared list of thirty document titles[3]. We then asked them to view the summary for each document and mark, by clicking the document title, if the summary appears to be relevant to the search description. The time from the appearance of the summary until the subject clicked the document title was recorded for each relevant summary. For the non-relevant summaries, the time taken from when the

[3] These search descriptions and retrieved titles are unrelated to the actual search tasks we use in the main experiment.

summary appeared until the summary disappeared was recorded, i.e. the time from the searcher requesting a summary until moving to the next title. The times measured for the summaries was normalized by the summary length as we expect searchers to take a longer time to read a longer summary. Finally the times are averaged to give a measure of how long an *individual* searcher takes on average to read a *character* in a relevant and non-relevant summary.

The value of time taken for an individual subject to read a character in a relevant summary was used in System 3 to decide when to judge a summary as useful or not. Although the timing information was only used in System 3, this process was repeated, using different search descriptions and titles, before each experimental system was presented. This was to avoid undue attention being paid to System 3 by the experimental subjects.

6.3 Experimental tasks

In our experiments each subject was asked to complete three search tasks. These tasks were chosen to investigate the effectiveness of the three systems for different types of search task: *fact* search, *decision* search and *background* search. The fact search asked subjects to find a single item of information (a named person's *current* email address), the background search asked subjects to find as much information as possible on a given topic and the decision search, Figure 5, forced subjects to make a qualitative decision on the information they retrieved (which Rome museum is the *best* for impressionist painting).

> ***Context***
> *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. Your time in the city is limited to only two (2) hours and as such you would like to look for possible places to visit prior to your departure.*
>
> ***Task***
> *Bearing in mind this context, your task is to find information about the city's best impressionist art museum.*

Figure 5: Simulated work task situation (decision search)

Each search task was placed within a simulated work task situation, [1], (example Figure 5). This technique asserts that subjects should be given search scenarios that reflect real-life search situations and should allow the searcher to make personal assessments on what constitutes relevant material.

6.4 Experimental subjects

We recruited 24 subjects for our experiments. Our recruitment was specifically aimed at targeting two groups of users: experienced and inexperienced searchers.

The experienced searchers were those who used computers and searched the web on a regular, often daily basis. Inexperienced searchers were those who both searched the web and used computers and the Internet infrequently. Per week, inexperienced searchers spent on average 5.1 hours online and experts an average of 29.8 hours online. Overall our subjects had an average age of 24.73 with a range of 33 years (youngest 16 years: oldest 49 years).

The classification between experienced and inexperienced searchers was made on the basis of the subjects' responses to questions about the level of their computing, Internet and web searching experience and their own opinion of their skill level.

7. RESULTS

In this section we present the results from our evaluation of the systems. In presenting these results we are primarily looking for results that relate to our two research questions: the utility of presenting top-ranking sentences (System 1 vs. System 2) and the effectiveness of implicit evidence (System 2 vs. System 3). Our results will concentrate on these comparisons, although fuller comparison between the systems is possible. Tests for statistical significance will be given where appropriate using a Mann-Whitney test, $p \leq 0.05$, unless otherwise stated.

We examine the results in a number of ways. Firstly we analyse the timing task results, section 7.1, then center our discussion around measures that assess the interaction (such as task completion and queries submitted), section 7.2, measures that assess the new interface features, section 7.3 and user preference, section 7.4.

7.1 Timing tasks

The difference between times taken to read a character for relevant and non-relevant summaries allows us to check the hypothesis that searchers spend longer reading relevant summaries. In Table 1, we show the average of these results. This shows that *on average* our experimental subjects spent more time, per character, in assessing a relevant than non-relevant summary. These averaged results are statistically significant using a Mann-Whitney test at $p \leq 0.05$ ($p = 0.01733$). This shows that the basic hypothesis is valid: the time taken to assess a summary can be used to indicate a searcher's interest in the contents of the summary. However, the values reported in Table 1 are averaged results – the results calculated for individual subjects may show a greater or smaller range in times. Although these results have been specifically calculated for our web searching experiments using summaries, this approach is generalisable to other search scenarios and systems.

Table 1: Time in seconds to read a character in a summary

	Relevant summary	Non-relevant summary
Time	0.02620	0.01937

7.2 Interactive measures

In this section we consider the interactive aspects of the search. These aspects relate specifically to differences in how the searchers interact with the different systems and the degree to which this affected their search effectiveness. We start by considering task completion.

Task completion: After each search task the subjects were asked to rate (on a Likert Scale) to what degree they felt they had completed the search task. The responses were on a scale

of 1-5, with a value of 1 reflecting greater task completion. The averaged values are shown in Table 2. As can be seen, System 2 was judged by both experienced and inexperienced subjects as leading to a greater sense of task completion than System 1.

Table 2: Subject responses regarding task completion

	System 1	System 2	System 3
All subjects	3.54	1.63	2.38
Inexperienced	3.58	1.67	2.17
Experienced	3.50	1.59	2.58

The difference between the results given for System 1 and System 2 are statistically significant for all subject groups (all, experienced and inexperienced). The presentation of top-ranking sentences, therefore, does lead to a greater *perception* of task completion for the subjects. We next compare whether the top-ranking sentences leads to *faster* task completion.

In Table 3 we use data collected from background logging and present the average time taken to complete a task, as measured by when the searcher decided they had completed the task. Uncompleted tasks were given a search time of 600 seconds (the maximum 10 minutes allocated per search task) but are *not* used in the computation of the results reported in the Table 3. The table also shows (in brackets) the number of tasks completed on each system.

The difference between search completion time for System 1 and System 2 is statistically significant for all subject groups. The use of top-ranking sentences therefore also leads to faster task completion.

Table 3: Task completion times in seconds

	System 1	System 2	System 3
All subjects	549.40 (15)	496.67 (19)	508.71 (16)
Inexperienced	578.32 (6)	512.43 (9)	520.96 (7)
Experienced	522.74 (9)	486.21 (10)	503.15 (9)

System 3 also gives higher task completion figures compared to System 1. However, the use of implicit evidence to re-rank the top-ranking sentences was less conclusive. From Table 2, System 2 was shown to give a greater sense of task completion than System 3 which used implicit evidence. This result was statistically significant for the inexperienced subjects but not for the experienced searcher, however the results are consistent across the subject groups. The difference between task completion times was not statistically significant but System 2, with no re-ranking, lead to faster task completion, Table 3.

The use of implicit evidence, then, appears to degrade performance. One reason for this, as will be discussed in section 7.3, is that System 3 *removes* sentences from the list of top-ranking sentences. This was an unpopular and distracting feature.

Query iterations: One of the reasons that the subjects completed tasks more quickly with Systems 2 and 3 was that they submitted fewer queries to these systems. In Table 4 we present the average number of queries submitted to each system. This table shows that subjects submitted around 30% more queries to System 1 than either System 2 or System 3. Furthermore, all subject groups submitted significantly more queries to System 1 than System 2 demonstrating that the top-ranking sentences do lead to fewer query iterations.

On all systems, the inexperienced subjects submitted significantly more queries to *any* system than the experienced subjects. This finding points to the fact that less experienced searchers spend more time developing good queries and consequently require more support at the interface.

Table 4: Average queries submitted per subject

	System 1	System 2	System 3
All subjects	8.00	5.21	4.58
Inexperienced	9.33	5.67	6.25
Experienced	7.50	4.75	2.92

Experienced users, on the other hand, tend to submit fewer queries to any system. The difference between the number of queries submitted to System 3 was significantly less than the number submitted to System 2. This means that although experienced subjects tended to take slightly longer to complete their tasks on System 3 and had a lower perception of task completeness, they were interacting differently with the results of the search when using System 3. Experienced users noticed the removal of sentences and re-ranking more than inexperienced users. They felt that the removal hindered rather than helped their interaction, especially on occasions when the system's implicit impression of what was relevant did not match their own. In light of this finding, further research on the use of implicit evidence may be necessary.

Page views: One of the main aims of our work is to encourage subjects to view results that occur after the first page of 10 results. In Table 5 we present the average number of retrieved pages which appear in rank position 11 onwards in the result lists that were viewed by the searcher. This measures the degree to which subjects are accessing pages that appear after the first 10 results.

Table 5: Pages viewed after initial 10 results

	System 1	System 2	System 3
All subjects	4.63	8.21	8.13
Inexperienced	3.08	7.42	8.08
Experienced	6.17	9.00	8.17

From Table 5, it can be seen that subjects view more of these documents when using the systems with top-ranking sentences. The difference between the pages viewed when using System 2 is statistically greater than when using System 1. A similar result is obtained when comparing System 3 with System 1, which is also statistically significant. This indicates that the top-ranking sentences are serving their intended function in

recommending potentially relevant documents encouraging greater interaction by the searcher.

Summary: The top-ranking function of Systems 2 and 3 appear to work well: they help searchers complete tasks more quickly and with a greater sense of task completion. They do this by encouraging searchers to interact more with the results of the search: running less queries but viewing more of the retrieved pages. The implicit feedback in System 3, on the other hand, does not appear to work well. However as we have shown above, may provide useful support for the inexperienced searchers.

7.3 Interface features

In this section we discuss the two novel features of our system: the top-ranking sentences, and the sentence re-ranking through implicit evidence.

Overall the subjects were in favour of the top-ranking sentences component of Systems 2 and 3. In Table 6, we present the averaged results obtained from asking the subjects how effective they found the top-ranking sentences and the degree to which re-ranking helped them find relevant documents. The responses were on a scale of 1-5, with a value of 1 reflecting greater satisfaction. The results are low indicating approval of the top-ranking sentences as an interactive technique.

Comparing the two systems, the results are not significant, i.e. the subjects did not prefer one system over the other. However for both aspects the inexperienced subjects gave better scores to System 3 than the experienced subjects. This suggests that the updating was of more use to the less experienced searchers.

Table 6: Subject responses regarding top-ranking sentences

		Inexperienced	Experienced
Effective	**System 2**	2.00	2.17
	System 3	1.92	2.33
Helped to find relevant documents	**System 2**	1.75	1.67
	System 3	1.25	1.75

This finding is reinforced if we examine the degree to which the subjects found the re-ranking process useful itself. In Table 7, we present the averaged results obtained from asking the subjects the same questions regarding the re-ranking of the top-ranking sentences. On both aspects, the inexperienced subjects gave higher ratings (closer to 1) than the experienced subjects. Both differences are statistically significant. The inexperienced subjects therefore may have obtained more use from the re-ranking than the experienced subjects as it provides more support for query creation. However we would need to investigate this more fully before we would make such a strong assertion.

Table 7: Subject responses regarding re-ranking

	Inexperienced	Experienced
Effective	2.25	2.67
Helped to find relevant documents	2.33	2.67

We also compared the subjects' response to the question (using Semantic differentials) of how *useful* was each of the three interface components: summary, top-ranking sentences and re-ranking of sentences. These results are shown in Table 8. This analysis is intended to compare the subjects' perceptions of the usefulness of the various components of the interface. The results show a positive response (less than 3) by all subject groups to the system components used. There were no statistically significant differences between any of the components (i.e. summary vs. top-ranking) or any user groups (inexperienced vs. experienced).

Table 8: Summarised subject responses

	All subjects	Inexperienced	Experienced
Summary	2.17	2.17	2.17
Top-ranking	1.95	2.08	1.8
Re-ranking	2.08	2.08	2.08

After each search and after the entire experiment we asked the subjects for their view on the systems used. Subjects found the top-ranking sentences helpful and easy to use. Though using the sentences, subjects felt that they were made more aware of potentially relevant documents (especially those that lay outside the first result page) than with the baseline, System 1. Sentence re-ranking did not receive as positive a reaction however.

The subjects had two reservations regarding the re-ranking of sentences in System 3. Firstly, as the re-ranking occurred at the same time as a summary appeared the subjects did not always notice the effect of the re-ranking. The presentation of the updating therefore needs improving in future systems.

Secondly, the top-ranking sentences in System 3 only contained sentences from web pages for which the searcher had not already viewed a summary. If the searcher viewed the summary for a page, then all sentences from a page were removed from the list of top-ranking sentences. We decided on this to increase the degree to which the list of top-ranking sentences would update. However, many searchers stated that they would prefer less updating and no removal of sentences.

Summary: The top-ranking sentences were popular in both systems but rather more popular with inexperienced searchers. The dynamic re-ranking of top-ranking sentences was given better ratings by inexperienced searchers.

7.4 User preference

In this section we will analyse the results obtained when, after the third task, users were asked to rank the three systems in

order of preference. Table 9 shows the average ranking given to each of the three systems *by all users* based on two criteria; which systems users felt helped them to find relevant documents, and which they liked best.

All 24 subjects placed one of the systems that used top ranking sentences (System 2 or System 3) in first place. 19 participants (8 inexperienced and 11 experienced) placed the baseline system in third place, the last ranking position. We applied a Kruskal-Wallis Test, at $p \leq 0.05$, to the results obtained and found that the Systems 2 and 3 obtained rankings significantly better (closer to 1) than System 1. This result holds for both measures and across all user groups (all, inexperienced and experienced). Table 9 also shows that subjects gave System 2, on average, a higher ranking than System 3. This difference is again significant using a Kruskal-Wallis Test at $p \leq 0.05$.

Table 9: Average ranking (range 1-3, lower = better)

	System 1	System 2	System 3
Helped to find relevant documents	2.79	1.5	1.71
Preference	2.75	1.54	1.71
All measures	2.77	1.52	1.71

Summary: All subjects felt systems using top-ranking sentences were more helpful in finding relevant documents, and were preferred to the baseline system. Of those systems with top-ranking sentences, subjects felt the system *without* sentence re-ranking helped locate relevant documents better, and was preferred to, the system *with* sentence re-ranking.

8. CONCLUSIONS

In this paper we present an investigation into techniques for encouraging web searchers to more fully assess a retrieved set of web pages. We presented two techniques: the use of top-ranking sentences and the use of implicit evidence.

We conducted an experiment, based on a sound experimental methodology, in which we used both experienced and inexperienced users. Our experiment used simulated work tasks to generate information needs, and questionnaires using Likert scales, Semantic differentials and open-ended questions to collect data. The results of the experiment have shown, with statistical significance, that presenting searchers with indications of retrieved pages' content – top-ranking sentences – not only leads to greater task completion but is also a popular addition to the interface.

The most encouraging feature of our experiments is that they show that it is possible to get searchers to interact with more than a few search results, thus reducing cognitive and computational loads. In addition, the approach moves away from simply presenting titles to presenting alternative access methods for assessing and targeting potentially relevant information.

We also experimented with implicit evidence to gather more information about a searcher's information need and use this information to suggest unseen documents. This is preferred to our baseline system, however, was not as effective as top-ranking sentences without this additional feature. This may in part be due to some of the design decisions we made in implementing this technique. However, it did appear to be successful in helping inexperienced searchers who may submit poorer initial queries and require greater interface support.

We also investigated the assumption that searchers spend more time reading relevant document summaries than with non-relevant summaries. Our results show that users spend significantly longer viewing relevant document summaries than viewing non-relevant ones.

9. ACKNOWLEDGEMENTS

The work reported in this paper is funded by the UK Engineering and Physical Sciences Research Council grant number GR/R74642/01.

10. REFERENCES

[1] P. Borlund. Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation. **56**. 1. pp 71-90. 2000.

[2] A.L. Berger and V.O. Mittal. *OCELOT: A System for Summarizing Web Pages*. Proceedings of 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp 144-151. Athens 2000.

[3] H. Fowkes and M. Beaulieu. *Interactive searching behaviour: Okapi experiment for TREC-8*. Proceedings of 22nd BCS-IRSG European Colloquium on IR Research. Electronic Workshops in Computing. Cambridge. 2000.

[4] W. C. Hill, J. D. Hollan, D. Wroblewski, and Tim McCandless. *Edit Wear and Read Wear*. Proceedings of ACM SIGCHI'92. pp 3-9. 1992.

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

[6] M. Magennis and C. J. van Rijsbergen. *The potential and actual effectiveness of interactive query expansion*. Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval pp 324-332. Philadelphia. 1997.

[7] D. Oard and J. Kim. *Implicit feedback for recommender systems*. Proceedings of the AAAI Workshop on Recommender System. Madison, Wisconsin. 1998.

[8] S. Robertson. *On term selection for query expansion*. Journal of Documentation. **46**. pp 359-364. 1990.

[9] R. W. White, I. Ruthven and J. M. Jose. *The use of implicit evidence for relevance feedback in web retrieval*. Proceedings of 24th BCS-IRSG European Colloquium on IR Research. Lecture notes in Computer Science 2291. pp 93-109. Glasgow. 2002.

[10] R. W. White, J. M. Jose and I. Ruthven. *A task-oriented study on the influencing effects of query-biased summarisation in web searching*. Information Processing and Management. 2002. *in press.*