Human-Computer Interaction -- INTERACT'03
M. Rauterberg et al. (Eds.)
Published by IOS Press, (c) IFIP, 2003, pp. 220-227

A Granular Approach to Web Search Result Presentation

Ryen W. White[1], Joemon M. Jose[1] & Ian Ruthven[2]

[1] Department of Computing Science, University of Glasgow, Glasgow. Scotland.

[2] Department of Computer and Information Sciences, University of Strathclyde, Glasgow. Scotland.

{ryen, jj}@dcs.gla.ac.uk, ir@cis.strath.ac.uk

Abstract: In this paper we propose and evaluate interfaces for presenting the results of web searches. Sentences, taken from the top retrieved documents, are used as fine-grained representations of document content and, when combined in a ranked list, to provide a *query-specific* overview of the set of retrieved documents. Current search engine interfaces assume users examine such results document-by-document. In contrast our approach groups, ranks and presents the *contents* of the top ranked document *set*. We evaluate our hypotheses that the use of such an approach can lead to more effective web searching and to increased user satisfaction. Our evaluation, with real users and different types of information seeking scenario, showed, with statistical significance, that these hypotheses hold.

Keywords: Web search, Visualisation, User studies

1 Introduction

Web search systems operate using a standard retrieval model, where a user, with a need for information, searches for documents that will help supply this information. Typically, users are expected to describe the information they require via a set of query words submitted to the search system. This query is compared to each document in the collection, and a set of potentially relevant documents is returned.

Devising and submitting a query can be a cognitively expensive and demanding process (Goecks and Shavlik, 2000). However, a searcher may face even more difficulty when interpreting and assessing the relevance of the returned documents. Users of web search engines are typically unwilling to examine large sets of individual documents. Instead users' initial judgments on what documents to view are based on document *surrogates* such as titles, abstracts and URLs. Surrogates can be manually created, e.g. titles or keywords, or automatically created, e.g. summaries.

Information retrieval (IR) systems were originally devised for the retrieval of documents from homogeneous corpora, such as newspaper collections or library index cards. Document surrogates were usually created by experts, such as librarians or professional cataloguers. The growth in size, dynamism and heterogeneity of collections being searched led to the development of automated representation techniques. This led to a reduction in the quality of the surrogates created. However searchers must rely on the indicative worth of surrogates when using them for deciding which documents to download and view.

Presenting lists of document surrogates has remained a popular method of presenting search results. Lists allow documents to be *ranked* in order of their estimated utility to the user. However, lists encourage users to read, interpret and assess documents and their surrogates *individually*.

In this paper we investigate surrogates for web searching. We suggest techniques that encourage a deeper examination of documents at the results interface and blur inter-document boundaries. We shift the focus of interaction from the document surrogate to the document's content. In particular we compare traditional methods of producing surrogates (such as text fragments and titles) against a new method of presenting search results; sentences taken from the retrieved documents, ranked on how closely they match the user's query. This set of sentences can be used to form a *query-specific* overview of the returned document set. We evaluate this approach using real users, realistic search tasks and three interfaces to the popular Google[i] web search engine.

2 Motivation

Web search engines are intended to help people find information that is useful or relevant to completing a task. Finding this information may require running several queries, reading many documents, and making

[i] http://www.google.com

judgments on which documents are worth reading. It is important therefore to design interfaces that maximize the amount of useful information that users can obtain within a search.

Searchers use textual queries to communicate their need with the search system. The query is only an approximate description of the information need (Taylor, 1968), and may fall short of the description necessary to infer relevant documents. Documents in the collection are ranked algorithmically based on this query and returned in a list to the user. These may not be *entirely* relevant, and it is the relevant parts that contribute most to satisfying the user's information need. By ranking *documents* we assume that all of a document conforms to relevance/matching criteria. This assumption is often incorrect as documents can have irrelevant parts. Research into summarisation (Berger and Mittal, 2000) and visualisation (Hearst, 1995) have tackled this problem, but still return document lists to users. Other representations of web search results have been tested. These either present the user with an unfamiliar, graphical interface that imposes an increased cognitive burden (Ahlberg and Shneiderman, 1994) or consider documents as the finest level of granularity for result presentation (Chen and Dumais, 2000).

In result lists users assess document relevance externally, based on what they can infer from their surrogates. On the Internet, authors assign document titles and the extent to which these titles are indicative of content can vary. The short abstracts presented by search systems provide a list of short text fragments containing the query words (see Figure 1). To provide users with representations that are truly indicative, we must delve deeper into the documents, extracting their content at a more fine level of granularity (i.e. with sentences), but with increased information on the context of a document's content.

In our approach, we present whole sentences to users, taken from the top thirty documents in the retrieved document set. These sentences place the query terms in the local context in which they occur in the source document. This technique, known as *sentence extraction*, has been the basis for many successful document summarisation systems (Berger and Mittal, 2000).

These sentences provide a high level of granularity, removing the restriction of document boundaries and shifting the focus from the document to the information it contains. This means that users are not forced to access information through documents but through the actual content of documents in an approach we call *content-driven information seeking*. Through ranking this information with respect to the query, the user is given a *query-specific overview of the content* of the returned set. A document list is biased towards the user's information need at the document level. Documents that are a close match to the user's query appear near the top of the list. In our approach we bias at the sentence level. Sentences that are a close match to the user's query are shown near the top of a ranked list of sentences.

In the next section we describe the interfaces created to test our hypotheses that presenting a list of ranked sentences extracted from the retrieved document set can lead to increased user perceptions and to more effective searching.

3 Interfaces

Three interfaces were used in our experiments. System 1 is a traditional web search system, and is used as a baseline. System 2 uses sentences extracted from top ranked documents at retrieval time to create more detailed summaries, but presents its results in the same way as System 1. System 3 uses the same extraction methods as System 2, but presents all sentences from the top thirty documents in a ranked list. In this section we describe each of the interfaces.

3.1 Web Search Baseline (System 1)

This interface uses the Google commercial search engine to search the Internet. In response to a query submitted by the user, the system returns a ranked list of document titles, abstracts and URLs, ten per page. After perusing the list, s/he can reformulate the query or proceed to the next ten results. Showing ten results per page is the Google default.

To eliminate possible bias caused by previous searching experience no indication was given of the commercial search engine users were using. They submitted queries and were presented with an interface that masked the search engine's identity, yet preserved all content (see Figure 1).

Figure 1. Masked web search baseline

3.2 Sentences as Surrogates (System 2)

System 2 is similar to System 1. It returns a ranked list of document titles, abstracts and URLs, ten per page. However, the size and nature of the abstract differ.

In this system, the top ten documents are downloaded and all sentences from each document are extracted. Each sentence is assigned a score, using an algorithm similar to that in (White, Jose and Ruthven, 2003). This uses factors such as position of the sentence in document and the presence of any emphasised words. In addition sentences receive additional scores depending on the proportion of query terms contained within the sentence. This *query-biasing* component biases the scoring mechanism to

sentences that use words contained within the user's query. Instead of the Google abstract, top scoring sentences from that document are combined to form a real-time summary of that document, up to a maximum of 20% of the document length (Figure 2).

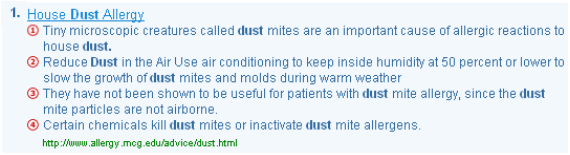

Figure 2: Document summary (composed of sentences)

The sentences are shown under the document title in descending order of score. If the user clicks 'next', documents in the range 11-20 will be downloaded and summarised in the same way.

3.3 Sentences as a List (System 3)

System 3 uses the same underlying mechanics for retrieving and extracting sentences from documents as System 2. However, the number of documents downloaded, the post-extraction operations and presentation of search results differ. The system downloads a set of *thirty* documents at retrieval time and extracts sentences from these. It then pools the top scoring sentences from each document (up to 20% of the document length), and ranks all sentences in a 'global' list: a list of *top-ranking sentences*. This list is a query-biased overview of the returned document set. The sentences are shown individually, with query terms highlighted (shown by ① in Figure 3). There are typically around 40-50 sentences per list.

In System 3 the users are not shown a list of retrieved document titles and URLs: only the list of top-ranking sentences is shown. Initially there is no direct association between the sentence and its source document, i.e. there is no indication to the user of which document supplied each sentence. To view the association, the user must move the mouse pointer over a sentence. When this occurs, the sentence is highlighted and a window pops up next to it (shown by ② in Figure 3). Displaying this window next to the sentence, instead of in a fixed position on the screen, makes the sentence-document relationship more lucid.

In the window the user is shown the document title, URL and the rank position and content of any other sentences from that document that occur in the list of top-ranking sentences. If no other sentences appear an appropriate message is shown.

To visit a document the user must click the highlighted sentence, or any sentences in the pop-up window. It is the sentences (content) that drive the interaction. When the user has clicked a sentence and visited that document, all sentences from that document are marked to reflect this.

②

①

Figure 3: Sentence List interface

The way the results are presented is the main difference between System 2 and System 3. The same depth and detail of information is shown, but in different ways. In System 2 the information-seeking interaction is driven by document surrogates and the ranking of *documents*. In System 3 the interaction is driven by content and the ranking of *components* of the document content; the top-ranking sentences.

We use these systems to investigate three main research questions. Firstly, we examine whether using a small set of query-biased top-ranking sentences is a viable alternative for a standard search engine abstract. This is effectively a comparison of System 1 and System 2. Secondly, we investigate whether presenting a list of ranked, content bearing sentences increases searcher awareness of the returned document set's content and increased granularity preferred by end-users as an alternative to traditional forms of result presentation. This is a comparison of Systems 1 and 2, against System 3. Finally, we investigate whether showing a list of sentences in this way leads to improved perceptions of task success, actual task success and whether these feelings traverse all tasks and agree with quantifiable measurements such as task completion time. This is again a comparison of Systems 1 and 2, versus System 3. In the next section we outline the experiments undertaken.

4 Experiments

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

4.1 Methodology

A total of 18 subjects took part in the experiment, each completed 3 tasks, one on each of the 3 search systems. Tasks and systems were allocated according to a Greco-Latin square design. To reduce learning effects we rotated both the order in which tasks *and* systems were presented across participants. It was important to rotate both of these, as one of our hypotheses investigated the usefulness of our approach for

different types of search task. The negative effect of task-bias was minimised by this rotation. 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 time restriction ensured consistency between subjects.

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.

When they completed a search, the subjects were 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 search a final questionnaire was completed that asked searchers to rank the three systems based on their personal preference. Subjects were also asked to rank the tasks based on their level of difficulty. This allowed us to discern whether the task impacted on subject preference and/or search success. Background logging was also used to record user interaction with the systems.

4.2 Subjects

We recruited 18 subjects for our experiments. Our recruitment was specifically aimed at targeting two groups of users: experienced and inexperienced searchers. We recruited 9 users for each group.

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.

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 4.2 hours online and experts an average of 32.6 hours online. Overall our subjects had an average age of 24 with a range of 32 years (youngest 17: oldest 49).

4.3 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 (dust allergies) and the decision search forced subjects to make a qualitative decision on the information they retrieved (find Rome's *best* museum for impressionist art). Each search task was placed within a simulated work task situation, (Borlund, 2000). This technique asserts that subjects should be given search scenarios that reflect real-life search situations and allow the searcher to make personal assessments on what constitutes relevant material.

The choice of tasks was particularly important for these experiments, as we planned to investigate the effectiveness of the system for different types of task. For this reason we chose tasks we had used in previous experiments (White, Jose and Ruthven, 2003; White, Ruthven and Jose, 2002) where the impact of task bias was not significant.

5 Experimental Results

In this section we present results from our system evaluation. In particular, we focus on results pertinent to each of our three research questions: the usefulness of top-ranking sentences as document abstracts, the value of such sentences for improving result set awareness and general user perceptions, and the impact of our approaches on perceived/actual task success. Tests for statistical significance will be given where appropriate with $p \leq .05$, unless otherwise stated. S_1, S_2 and S_3 denote System 1, System 2 and System 3 respectively. We analyse the results individually by research question. *M* is used to denote the mean.

5.1 Top-Ranking Sentences as Abstracts

In this section we compare System 1 and System 2, analysing the results obtained from questionnaires and system logging. These systems both display their results via document lists, but their document abstracts differ in size, currency (recency) and quality. The term 'abstract' will be used in this section to refer to the document snippets in System 1 and the sentences presented underneath the document titles in System 2.

5.1.1 Participant Feedback

Participants provided feedback on the search process, the document abstracts and the interface features.

Search process: After each search task subjects were asked to rate (using 5-point semantic differentials between 1-5, with 1 reflecting greater agreement) how *relaxing*, *interesting*, *restful* and *easy* their search process had been. Search systems have a 'duty of care' for those who use them, and those that impose an increased affective burden will hinder, rather than help, searchers. We ran MANOVA to test for an overall difference between experienced and inexperienced subjects. We found that there was no significant difference between the groups ($F_{4,33} = 1.847$, $p = .326$). Repeated measures two-way ANOVAs were run on each of the differentials. To reduce the number of Type I errors i.e. rejecting null hypotheses that were true, we set the *alpha level* (α) to .0125 i.e. .05 divided by 4, the number of tests performed. The results showed that System 2 made the search process

significantly easier (M_{S1} = 2.50, M_{S2} = 1.78, p = .0042), across all user groups. All other differences were not significant.

Abstracts: Using 5-point semantic differentials, subjects were asked to rate how *helpful*, *beneficial*, *appropriate*, *relevant*, *useful* and *effective* the abstracts presented by each of the systems were. We ran MANOVA to test the results, and found that they were significant between systems ($F_{6,27}$ = 2.649, p = .0391). The close proximity of p to the upper bound of significance (e.g. p ≤ .05) motivated us to use repeated measures two-way ANOVAs for each differential. The results showed that the abstracts presented by System 2 were significantly more useful, relevant and effective than those shown by System 1. The mean differentials are shown in Table 1, where a rating closer to 1 reflects stronger agreement. The differences are significant between *systems*, but not *user groups*.

	Inexperienced		Experienced	
	S_1	S_2	S_1	S_2
useful	3.00	2.12	3.24	2.22
relevant	3.24	2.37	3.55	2.56
effective	3.21	2.87	3.66	3.02

Table 1: Subject perceptions of document abstracts

Usefulness of interface features: Subjects were asked to rate, after each search task, how useful each of the features on the interface had been. The features used in this comparison are document titles, abstracts, URLs and the 'next' button. In Table 2 we present the average results obtained. Responses were on a scale of 1-5, *with 1 reflecting greater usefulness*. The ratings for the 'next' button and URLs are consistently 'low', indicating that subjects did not find them useful.

	Inexperienced		Experienced	
	S_1	S_2	S_1	S_2
Document title	3.13	3.02	2.96	2.97
Abstract	3.28	2.32	3.54	1.95
URL	4.31	4.43	4.65	4.33
'next' button	4.33	4.78	4.54	4.76

Table 2: Responses on interface feature usefulness

There are no significant inter-system/inter-subject differences for document title, URL or 'next' button with MANOVA ($F_{3,30}$ = .94, p = .438). This is to be expected, as there is no difference in the way these features are presented in each system. The difference in the 'usefulness' of the abstracts is significant with repeated measures two-way ANOVA ($F_{1,32}$ = 13.15, p = .001, α = .0125). Subjects found the sentence-composed abstracts in System 2 significantly more useful for their tasks than standard web search engine abstracts in System 1. The difference was not significant between user groups.

5.1.2 System Logging

We present the average number of query iterations and average task completion times (in seconds) across user groups and with each system (Table 3).

	Inexperienced		Experienced	
	S_1	S_2	S_1	S_2
Time	543.22	529.04	453.72	422.59
Queries	7.65	6.44	6.77	4.53

Table 3: Subject perceptions of document abstracts

We ran MANOVA to see if there was an overall difference between systems and between subjects. We found this to be the case ($F_{2,31}$ = 7.47, p = .002). System 2 reduces task completion time and the number of query iterations.

5.2 Top-Ranking Sentences Increase Awareness and Subject Perceptions

In this section we analyse the impact of presenting subjects with a query-biased list of top-ranking sentences extracted from the top ranked documents in a result set. This is effectively a comparison of System 1/System 2 and System 3. To this end, we examine both questionnaire responses and system logs. We report results from each method.

5.2.1 Participant Feedback

Participants used questionnaires to provide feedback on the systems used. Here, we analyse their feelings about the abstracts/sentences presented to them, the usefulness of interface features, their awareness of retrieved set content and the clarity of the sentence/abstract-document relationship.

Search process: As before, subjects were asked to rate the search process they had just performed using the semantic differentials *relaxing*, *interesting*, *restful* and *easy*. We applied a MANOVA to test for differences between systems and groups of subjects. We found that there was no significant overall difference between systems or subjects ($F_{8,90}$ = 1.65, p = .243). We applied ANOVA to each differential and found that System 3 made the search process significantly easier than System 1 and 2. (M_{S1}= 2.34, M_{S2} = 2.15, M_{S3} = 1.85, p = .0121, α = .0125). This spanned all user groups.

Abstracts/sentences: Again using 5-point semantic differentials, subjects were asked to rate how *helpful*, *beneficial*, *appropriate*, *relevant*, *useful* and *effective* the abstracts or sentences were. Table 4 shows the statistically significant differentials, where a rating closer to 1 reflects stronger agreement.

MANOVA showed that there was a significant overall difference between S_1, S_2 and S_3 ($F_{8,90}$ = 5.32, p =

.0032) but not between groups of searcher. Repeated measures two-way ANOVAs showed that the differences were significant apart from between those differentials shown in bold. This means that subjects did not find the sentences shown in System 3 to be significantly more *relevant* than those shown in System 2 ($F_{4,45}$ = 1.88, p = .127, α = .0125). The way in which sentences are generated is identical in these two systems, only the way they are presented differs (i.e. underneath document titles in System 2 and in a 'global' list in System 3).

	Inexperienced			Experienced		
	S_1	S_2	S_3	S_1	S_2	S_3
helpful	3.30	2.98	2.17	3.35	3.33	2.11
useful	3.00	2.37	1.91	3.34	2.22	2.02
relevant	3.24	**2.12**	**1.93**	3.65	**2.56**	**2.18**
effective	3.21	2.87	2.32	3.76	3.02	2.39

Table 4: Perceptions of document abstracts/sentences

Usefulness of interface features: Subjects were asked to rate (on Likert scales), after each search task, how useful each of the features on the interface had been. The features used in this comparison are document titles, URLs, the 'next' button, the sentence list and the sentence pop-up (Figure 4). In Table 5 we present the average results obtained when we asked the subjects to rate how useful each of the features were[ii]. The responses were on a scale of 1-5, *with a value of 1 reflecting greater usefulness*. The dashes in the table indicate which features subjects were not asked to rate on certain systems. For example, System 1 does not use top-ranking sentences.

MANOVA was used to test the significance of overall difference between experienced and inexperienced subjects across document title, URL and the 'next button'. No significant difference was found ($F_{3,50}$ = .99, p = .222). Following this we did separate ANOVAs for each dependent variable and found that document titles were significantly *less* useful in System 3 than in System 1 and System 2 ($F_{4,45}$ = 5.43, p = .005, α = .0167). Paired *T*-tests established that there were no significant inter-group differences for the top-ranking sentences (T_{16} = .98 p = .176) and the document pop-up (T_{16} = 1.64 p = .062). The implications of these findings are that web subjects do not appear to perceive benefit from, nor make use of, document URLs and the 'next' button. Presenting a ranked list of content bearing sentences had a marked impact on how useful subjects regarded the document titles. The sentences shifted the focus of interaction from surrogate to content, reducing the need for these titles. This was appreciated by both inexperienced and experienced subjects.

[ii] System 3 does not use document abstracts and these are not included in this part of the analysis.

	Inexperienced			Experienced		
	S_1	S_2	S_3	S_1	S_2	S_3
Document title	3.13	3.02	4.51	2.96	2.97	3.65
URL	4.31	4.43	4.85	4.65	4.33	4.53
'next' button	4.33	4.78	4.79	4.54	4.76	4.64
Sentences	–	–	1.86	–	–	1.74
Pop-up	–	–	2.36	–	–	1.81

Table 5: Responses on interface feature usefulness

Awareness of returned document set: One of the main aims of our approach was to increase subject awareness of the returned document set contents. After each search task we asked subjects, based on the information provided at the interface, how aware they were of the contents of the top-ranking documents in the retrieved document set. Table 6 shows the average Likert scale responses for this awareness and the clarity of the abstract/sentence-document relationship.

	Inexperienced			Experienced		
	S_1	S_2	S_3	S_1	S_2	S_3
Awareness	3.15	2.64	2.02	3.24	2.11	1.54
Clarity	1.46	1.51	3.24	1.11	1.37	2.14

Table 6: Responses on interface feature usefulness

Repeated measures two-way ANOVAs for each of the dependent variables showed that System 3 significantly increased awareness of result set content ($F_{4, 45}$ = 14.86, p = .001, α = .025). However, the system made it less clear which document a sentence came from ($F_{4,45}$ = 10.51, p = .0018, α = .025). These differences held for both experienced and inexperienced subjects. This was to be expected as unlike Systems 1 and 2, the interface in System 3 showed no direct association between the abstract/sentence unless the searcher actively sought this association (i.e. by passing over a top-ranking sentence with the mouse).

5.2.2 System Logging

As subjects performed tasks, we logged their interactive behaviour with the systems. In this section we report on what was gleaned from these logs. In particular we examine their task completion time (i.e. the time they took to complete a task), the number of queries they submit per task, and the number of pages viewed outwith the top 10 results.

Task times: Table 7 shows the average task completion time, across all tasks, and the average number of query iterations on each of the three

systems. Repeated measures two-way ANOVAs for each of the dependent variables showed that System 3 significantly reduced task completion time (in seconds) ($F_{4,45}$ = 9.31, p = .004, α = .025). These differences held for both experienced and inexperienced subjects.

	Inexperienced			Experienced		
	S_1	S_2	S_3	S_1	S_2	S_3
Time	543.2	529.0	502.8	453.7	422.6	386.8
Queries	7.65	6.44	2.43	6.77	4.53	2.50

Table 7: Average task time and average query iterations

Page views: Another aim of this work was to reduce the importance of a document's result list location, focusing on document content over document ranking. To this end, we recorded the number of page views outwith the first result page (i.e. pages at position 11 and onwards in the result ranking). Although System 3 presented sentences from thirty documents on one page, subjects could easily access a similar number of document abstracts on the other systems by clicking the 'next' button. However, subjects appeared reluctant to do this, opting to reformulate and resubmit their query (Table 8).

	Inexperienced			Experienced		
	S_1	S_2	S_3	S_1	S_2	S_3
Page views	2.21	2.11	5.43	3.12	3.44	6.29

Table 8: Page views outwith first 10 documents

We applied repeated measures ANOVA to test the significance of our result. The analysis shows with significance ($F_{4,45}$ = 9.72, p = .001) that using sentences in this way encourages subjects to peruse more documents outwith the first 10 results. This holds for both user groups.

System ranking: Subjects were asked to rank the systems on a scale of 1-3 based on personal preference, where 1 was their preferred option. The results showed, with statistical significance with ANOVA ($F_{2,51}$ = 10.82, p = .001) that subjects preferred System 3 over Systems 1 and 2 and System 2 was preferred to System 1.

5.3 Perceptions of Task Success

We investigate the performance of our approach with different types of task. To do so, we use the same results obtained in our second research question, analyse them for each task and enhance them with data gathered for this purpose from the questionnaires. According to our experimental setup each task is attempted 7 times on each system.

After each search task, subjects completed a set of semantic differentials on the task they had just attempted. The differentials asked them to comment on the how *clear*, *complex* and *familiar* the task had been. We ran MANOVA and found no significant differences between tasks for each subject group ($F_{6,92}$ = 1.54, p = .186).

Table 9 shows the average task completion time, per task and the average number of tasks successfully completed. The time is taken from the subject starting the search, until they complete the task. The difference in task time between Task 1 (fact) and Tasks 2 (decision) and 3 (background) is significant, however the difference in the number of tasks actually completed (shown above each bar and out of 10) is not significant, for any task comparison, on any system.

	Inexperienced			Experienced		
	T_{fact}	$T_{dec.}$	$T_{back.}$	T_{fact}	$T_{dec.}$	$T_{back.}$
Time	443.2	549.0	512.8	323.9	494.3	398.3
Tasks	5	5	7	6	5	7

Table 9: Average completion time and number (per task)

System usefulness: After each search, subjects were asked to rate on a 5-point Likert scale, *where 1 indicates most agreement*, how useful the system was for the task they had just attempted. Table 10 shows these results. We applied a two-way ANOVA and found that only System 3 yielded any significant differences ($F_{4,45}$ = 10.14, p = .004).

	S_1	S_2	S_3
Fact	2.81	2.53	3.87
Decision	3.24	2.86	2.54
Background	3.68	3.31	1.39

Table 10: System usefulness for each type of task

Our subjects found Systems 1 and 2 perform best for the fact search, whereas on System 3 did so for the background search. Providing an overview of document set content may therefore be useful in gathering background information and decision making, not for pinpointing facts.

Task ranking: Subjects were asked to rank the tasks from 1-3 based on their difficulty, the results from this ranking were not significant with ANOVA ($F_{2,51}$ = 2.02, p = .401).

6 Discussion

Using sentences extracted from documents at retrieval time as a substitute for traditional web search engine abstracts made the search process easier, led to significant reductions in task completion time, and meant subjects submitted fewer queries. Subjects also found the abstracts composed of sentences significantly more useful, relevant and effective. 16 out the 18 participants (9 experienced and 7 inexperienced) ranked System 2 above System 1. Those who did not,

found the answer on System 1 after two iterations, giving them little time to compare the two systems. Our first hypothesis, that top-ranking sentences could viably replace web search abstracts was supported.

We then assessed whether using sentences, extracted from top ranked documents and presented in a query-biased ranked list would perform better than traditional forms of presentation. System 3 made the search more interesting than the traditional systems and easier than the baseline system. This result was mimicked in measures such as task completion time, task completeness and average number of queries submitted. Subjects also found the sentences in System 3 significantly more helpful, useful and effective than both baselines. This is surprising as the sentences were created in the same way as System 2, and often the same sentences appeared in both systems. Therefore, this difference could only be attributable to the result *presentation*.

The sentences and associated interface features were liked by subjects. The document titles were of less use in System 3 as user attention was focused on document *content*. To make sound judgments on the effectiveness of a submitted query, subjects should be able to assess the actual content of the document set, not just document surrogates. System 3 made subjects more content-aware.

Presenting a list of sentences in this way encouraged subjects to view documents outwith the first page of results. On the traditional systems subjects would rather reformulate and resubmit their queries than deeply peruse the documents returned to them. By doing so they discard potentially relevant documents without giving them due consideration. The document list returned is only an algorithmic match to the user's query, something that typically only contains 1 or 2 query terms (Jansen et al., 2000) Unless the information need is very specific (i.e. someone's name, such as in the fact search) the system may struggle to provide a ranking that is a match for the user's information need. This problem is amplified if the system only ranks whole documents as small highly relevant sections may reside in documents with a low overall ranking.

16 of the 18 subjects (9 experienced and 7 inexperienced) preferred System 3 to System 2, and the same 16 preferred System 3 over System 1. It is worth noting that the two participants that did not rank System 3 highly used System 3 for the fact search *and did not complete the task.* Our second hypothesis, that top-ranking sentences improve result set awareness and general user perceptions, was supported.

According to our experimental results, our tasks were of a similar level of difficulty. System 3 appears most useful for tasks that involve gathering and assessing information, but not as useful for pinpointing facts. The final hypothesis, that the approach has a positive impact on aspects of task success and there was no associated task bias, was supported.

7 Conclusions

In this paper we present an investigation of an approach for presenting web search results. An approach that shifts the focus of perusal and interaction away from the document surrogates, such as document titles, abstracts and URLs, to the actual content of the document. The results of our experiments have shown, with statistical significance, that ranking the content of the retrieved document set rather than the documents themselves leads to increased searcher efficiency, effectiveness and personal preference.

Acknowledgements

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

References

Ahlberg, C. and Shneiderman, B. 'Visual information seeking: Tight coupling of dynamic query filters with starfield displays.' *Proceedings of ACM SIGCHI*, 313-317. 1994.

Berger, A.L. and Mittal, V.O. 'OCELOT: A System for Summarizing Web Pages'. *Proceedings of ACM SIGIR*, 144-151. 2000.

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

Chen, H. and Dumais, S. 'Bringing order to the web: automatically categorizing search results'. *Proceedings of ACM SIGCHI*, 145-152. 2000.

Goecks, J. and Shavlik, J. 'Learning users' interests by unobtrusively monitoring their normal behavior'. *Proceedings of the International Conference on Intelligent User Interfaces*, 129-132. 2000.

Hearst, M. 'TileBars: Visualization of Term Distribution Information in Full Text Information Access'. *Proceedings of ACM SIGCHI*, 59-66. 1995.

Jansen, B.J., Spink, A. and Saracevic, T. 'Real life, real users and real needs: A study and analysis of users on the web'. *Information Processing and Management* **36**(2), 207-227 2000.

Taylor, R.S. 'Question-negotiation and information seeking in libraries'. *College and Research Libraries* **29**, 178-194. 1968.

White, R.W., Jose, J.M. and Ruthven, I. 'A task-oriented study on the influencing effects of query-biased summarisation in web searching'. *Information Processing and Management*. 2003. in press.

White, R.W., Ruthven, I. And Jose, J.M. 'Finding Relevant Documents Using Top-ranking Sentences: An Evaluation of Two Alternative Schemes'. *Proceedings of ACM SIGIR*, 57-64. 2002.