Supporting Searching on Small Screen Devices using Summarisation

Simon Sweeney and Fabio Crestani

Dept. Computer and Information Sciences
University of Strathclyde
Glasgow, Scotland, UK
{simon, fabioc}@cis.strath.ac.uk

Abstract In recent years, small screen devices have seen widespread increase in their acceptance and use. Combining mobility with their increased technological advances many such devices can now be considered mobile information terminals. However, user interactions with small screen devices remain a challenge due to the inherent limited display capabilities. These challenges are particularly evident for tasks, such as information seeking. In this paper we assess the effectiveness of using hierarchical-query biased summaries as a means of supporting the results of an information search conducted on a small screen device, a PDA. We present the results of an experiment focused on measuring users' perception of relevance of displayed documents, in the form of automatically generated summaries of increasing length, in response to a simulated submitted query. The aim is to study experimentally how users' perception of relevance varies depending on the length of summary, in relation to the characteristics of the PDA interface on which the content is presented. Experimental results suggest that hierarchical query-biased summaries are useful and assist users in making relevance judgments.

1 Introduction

The recent trend towards pervasive computing, information technology becoming omnipresent and entering all aspects of modern living [3], means that we are moving away from the traditional interaction paradigm between human and technology being that of the desktop computer. This shift towards ubiquitous computing is perhaps most evident in the increased sophistication and extended utility of mobile devices, such as mobile phones, PDAs, mobile communicators (telephone/PDA) and Pocket PCs. Advances in these mobile device technologies coupled with their much-improved functionality means that current mobile devices can be considered as multi-purpose information tools capable of complex tasks. In terms of services that are available for mobile devices, there are currently thousands of applications for handheld devices for the different handheld operating systems (PalmOS, Windows CE). In fact, many of these devices are now capable of supporting tasks that are normally only associated with the desktop PC, such as creating word-processed documents, spreadsheets, presentation slides. Similarly, for WAP mobile phones there exists a wide variety of network-based services.

However, there remains a significant challenge in the presentation of information on mobile devices given the inherent constraints of a low-resolution small display area and, in the case of mobile phones, limitations on interaction [12]. And whilst the amount of information available on the web is ever increasing the same degree of proliferation of content has not yet been matched for mobile devices [18]. A possible reason for this lag in growth of content for mobile devices could be that there are accepted approaches for supporting information access on the desktop PC, whereas the same approaches may not be appropriate for mobile devices.

Significant improvements have been made in the interface design of applications for mobile device platforms, particularly for supporting web access. WAP is designed specifically for small handheld wireless devices and the limitations of small display screens and the interaction constraints. However, aside from WAP few Web browsers currently available for handheld devices (PDAs) render content to take into account the very different display capabilities of handheld devices.

In this paper we shall highlight work that has been carried out to improve browsing and searching on small screen devices, as a starting point we shall discuss how searching is currently supported on the desktop PC. We shall then briefly describe some automatic summarisation approaches that have been applied in the context of information retrieval (IR). We shall then present the results of our latest experiment that measured user performance in conducting relevance judgments using summaries presented on a PDA device.

The paper is structured as follows. Section 2 describes existing work on supporting searching for both desktop and small screen devices, also including a brief review of some previous work relating to the use of query-biased summarisation. Section 3 outlines an experiment we carried out investigating the effectiveness of presenting hierarchical query biased summaries on a PDA device. Section 4 presents the details of the experimental set-up used. Subsequently, Section 5 presents the experimental results and analysis. Finally, Section 6 reports the conclusions from our findings and present future extensions of this work.

2 Background

The subject of Information Retrieval (IR) is well established and is an underlying feature of the Internet. The function of an IR system is to locate and retrieve relevant data that is subsequently presented to the user. Traditionally, automatic IR systems are accessed using a desktop PC where the results of a search are presented on a large screen display and where user interaction is supported using a mouse or a keyboard. Users range from experienced experts, with possibly formal training in conducting information searches, to novice users who are at the very least computer literate.

The increased capability of mobile computing devices and the development of infrastructures for supporting intensive wireless communications means that mobile devices can now be considered as information terminals. However, as discussed in [13], information access in a mobile environment is considerably different to conventional IR. Firstly, mobile devices by design are multi-purpose and as a consequence may compromise certain useful features to maximise mobility and diversity. These devices tend to have low-resolution small display areas and limitations on interaction, particularly in the case of mobile phones. Using these devices can at times be challenging despite the continued improvements to device displays and means of interaction (stylus, T9 predictive text). Any difficulties experienced may be magnified when the functionality of these devices is extended to support new tasks, such as searching for information on the web. Secondly, the profile of a typical mobile device user differs from that usually associated with an IR system user in the sense that computer proficiency may not be assumed. This is apparent when considering the variety in user profiles of the current mobile phone user population. Finally, the retrieval task differs from that assumed under normal IR circumstances due to the nature of conducting searches in a mobile environment. There is a greater risk that user performance may be influenced by outside factors with the increased potential for distractions of noise and interruptions [8]. Further, a user maybe engaged in other activities at the time of searching. There is also the need to consider the type of information being sought, as there may be significant temporal dependencies. For example, consider the scenario of finding information about possible tourist sites available for visiting. In such a case it would be useful that any suggested tourist sites are first checked to see if they are open given the current time of day, and the expected travelling time to the site. All of these factors then influence the way mobile users will conduct searches and view search results.

2.1 Searching on the desktop

Most search systems present the results of a user query as a list of documents, spanning possibly a number of pages that may or may not be ranked. Users are required to assess each document individually on the basis of relevance to their submitted query. This can be a lengthy process given the often long list of retrieved documents. Approaches have been introduced aimed at reducing the overheads involved in working through the list of retrieved documents, assisting the user in completing their information discovery task.

Ranking the list of retrieved document according to relevance aids the user in this process by presenting those documents the systems considers as best matching the users query higher in the list [1]. Techniques may focus on attempting to improve the quality of the results increasing the number of relevant documents in the retrieved document result set. Relevance feedback is an example of such a technique, where by the system refines a set of results or perform a further search on the basis of user correction [7].

Research in information visualisation focuses on exploring alternative schemes to traditional ranked lists as a means of presenting search results. Many of these schemes make use of colourful highlighting and graphical features to capture aspects of the information access process, with content that is dynamic and can be manipulated by the user [1]. For example, the use of concept 'landscapes' to represent document clusters displayed graphically, in 2D as a 'jigsaw' with the clusters forming the individual pieces, or in 3D as a 'map' with contours describing document similarity and where peaks indicate concentrations of similar documents [5] [25].

Another variation to a plain list of document titles is to include additional information relating to the retrieved document. This additional information then function as a document surrogate providing the user with metadata, such as date of publishing, source, and length of the document, to give more indication about the content of a document [1]. The inclusion of document surrogates in search results lists has become a standard feature among web search engines, possibly the most widely used of the search systems.

Some systems extend document surrogates to include a short automatically generated extract, which may take the form of the first few lines of the document text. And, in recent times there has been an interest in enhancing document surrogates to better represent the content of the source documents. By applying techniques developed in the field of automatic summarisation the properties of the document surrogate can be improved. The outcome of which is document surrogates that are more representative of the document source and can be tuned to be either informative, contain such information from the document text that instantly fulfils the user's information need, or indicative, provide an indication of whether the particular document is relevant [2].

2.2 Searching on the small screen

Small screen devices provide many of the searching functionality found on the desktop PC, ranging from on-device information discovery to searching wider network accessed information resources, such as digital libraries or the WWW. However, whilst similar functionality is provided for such devices in practical terms using such services results in very different user experience [12] [11]. In general terms interfaces for searching on small screen devices have remained largely unchanged, querying is expressed by entry of plain text into a text field and search results are presented as a scrollable list of matches.

Some recent studies have found that supporting information discovery (browsing and searching) on small screen devices, such as PDAs, using interfaces designed for the display area of desktop PCs has a negative influence on task performance [12] [11].

Problems with search interfaces for small screen devices tend to revolve around the scrolling or page requirements when viewing content. Often to make content available for displaying on small screen devices it is not uncommon that long lists of search results are divided into separate pages that contain a reduced number of results. Breaking the content up into smaller manageable chunks is necessary for both transmission requirements and as a means of aiding presentation. However, such techniques have an associated cost that is page-to-page navigation is expensive in terms of user interactions and time [12], both of which may have financial implications (users are likely to be paying for wireless connections or the amount of data they transfer) and may have an impact on the way users use such services. Worst effects are observed if users are required to scroll horizontally [12]. In such cases, it is easy for users to become disorientated and lost within content designed for viewing on much larger screens.

Solutions then to aiding the user in making sense of search results on the small screen can be briefly outlined as follows. As mentioned, presenting only a limited number of results in each result page and limiting the amount of information displayed for each result (Google for the PDA) means that users will not have to view long lists of results. Combining relevance ranking with high precision performance would provide a trade-off to the splitting content over a number of pages and the associated navigation costs. Ideally, the most relevant results would appear in the first couple of pages and would fulfil the users information need reducing the need to go beyond the second page of results [9]. Alternatively, using schemes such as, WebTwig [10] or PowerBrowser [4] are designed specifically to take account of the limited display area of small screen devices and adapt content presentation accordingly. The basis of these schemes is to provide a more direct, systematic approach to viewing content that requires much less scrolling [11]. It is interesting to observe that both these schemes for accessing web on handheld devices have recently incorporated features that use forms of summarisation.

2.3 Applying summarisation to IR results presentation

Automatic summarisation has been used extensively in the content of IR. As a means of supplementing search results thus aiding the user to make the relevance assessments, and for making the IR process more efficient (using a summarised version of documents to build indexes or for storage, in place of the document full text).

Traditionally, automatic document summarisation has been based on sentence extraction approaches [2] [6] [14]. Advances in sentence extraction have seen the introduction of query-biased methods. Query-biased summarisation methods generate summaries in the context of an information need expressed as a query by a user. Such methods aim to identify and present to the user individual parts of the text that are more focused towards this particular information need rather than a generic, non-query-sensitive summary. Summaries of this type can then serve as an indicative function, providing a preview format to support relevance assessments on the full text of documents [16].

Highlighting recent research into the application of summarisation to aid information retrieval tasks, in particular the use of query-biased methods. Tombros and Sanderson investigated and illustrated the application of query-biased methods for text IR [22]. A later study by Tombros and Crestani looked at evaluating the effectiveness of presenting summaries by different means and the effect this has on users' perception of relevance [21]. Results from their study showed that users' ability to make relevance assessments of documents is highly affected by the way they are presented.

Extending the forms of presentation to include small screen devices, Sweeney, Tombros and Crestani looked at the use of query-biased hierarchical based summaries of newspaper articles presented to users on WAP mobile phones [20]. Defining hierarchical summaries as summaries of variable length, increasing from title only, 7%, 15%, to 30% of the original document length, the study investigated how users' perception of relevance varied depending on the length of the summary, and in relation to the specific characteristic of a typical WAP mobile phone interface. This study suggested that hierarchical query-biased summaries are useful when dealing with small screens and assist users in making correct relevance judgments. The results also highlighted, for WAP mobile phones, a preference for concise summaries that are relatively brief, 7% of the document length (up to a maximum of 3 sentences).

3 Presenting hierarchical summaries on a PDA device

We now report a resent study that continues the theme of investigating users' ability to carry out relevance judgements on textual information presented on non-traditional IR platforms. We use the same experimental procedure to the study previously mentioned [20], again using the hierarchical query-biased summaries, however, in this study we shall focus on the effects due to displaying content on a PDA interface. Again we are interested in assessing the variation of user performance in evaluating the relevance of full documents, given hierarchical query-biased summaries, and also determining whether there is an optimal size of summary for this type of interface. Similar to the previous experiments we assume the utility notion of relevance [23] as the basis for evaluating the summaries. Further details describing the context for the users' perception of relevance used in this study can be found in [21].

The summarisation system used to produce the summaries for the experiment was the same as that described in [20]. The system uses a number of sentence extraction methods [15] that utilise information both from the documents of the collection and from the queries used. A detailed description of the system can be found in [21] [22] here we shall only briefly describe the output of the summary generation process.

For the purposes of the experiment summary length was treated as a design variable in our system, corresponding to the level of information a user would be presented with in relation to the original document. Each level is intended to provide more information to the user. We consider this design as producing "hierarchical summaries", the root of which corresponds to the minimum level of information. Proceeding down the hierarchy, more and more information is made available to the user, up to a maximum, which corresponds to the full-text of the document.

Four different summary lengths were used in our experiments. It is established that titles convey useful clues about the contents of a document [17], and based on this fact we used titles as the first level of information (shortest summary) a user would be presented with. The other three summary length values were calculated as a percentage of the number of sentences in the original document. Therefore, for each document a number of sentences equal to the 7%, 15% and 30% of its length (up to a maximum of 3, 6, and 12 sentences respectively) were used.

For the experiment we used a HandSpring Visor running the AvantGo[1] web browser. Prior to the start of each user experiment, experimental content was transferred to the device such that users were only permitted to view content offline thus reducing effects of any outside factors that could influence the results, and ensuring consistency with previous experiments.

4 Experimental Settings

4.1 The Test Collection

The documents used were the same as those in the previous experiments, and are a subset of the 1990-92 Wall Street Journal (WSJ) collection of TREC [24]. The TREC-WSJ collection was used in the study both as a data source and as a standard against which the users' relevance assessments were compared, enabling precision and recall figures to be calculated. For this last purpose the relevance assessments that are part of the TREC collection and that were made by TREC "judges" were used (refer to later discussion on 'Experimental Measures'). We used 50 randomly selected TREC queries and for each of the queries, the 50 top-ranked documents as an input to the summarisation system. The test collection then consisted of a total of 2,220 news articles. To provide an indication of the proportion of relevant documents within those used for the experiment, there was a total of 414 relevant documents in the collection with an average of 8.3 relevant documents per query.

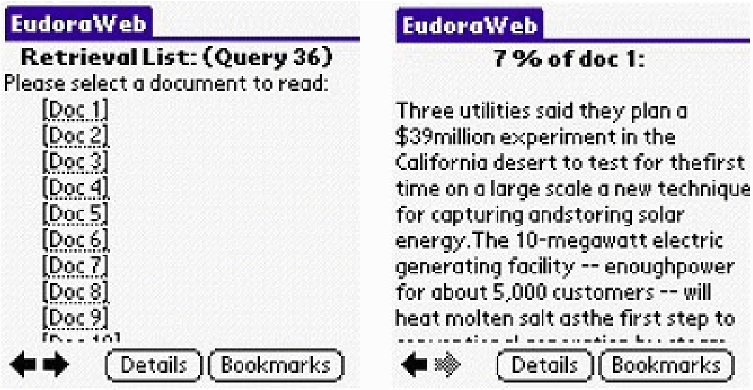
4.2 Experimental Procedure

To enable comparisons the same experimental tasks were used: users were presented with a retrieved document list in response to a query (simulated query), and had to identify as many relevant documents as possible for that particular query within 5 minutes. The information presented for each document was automatically generated, query-biased summaries.

The experimentation was carried out with user group of 10 volunteers with above average experience of using computers and mobile devices (mobile phones, PDAs). Each user was initially briefed about the experimental process, and instructions were handed to the user by the experimenter. Any questions concerning the process were answered by the experimenter at this stage. Users were otherwise uninformed of the purpose of the experiments. Each user was assigned a set of five queries randomly chosen among the 50 used. For each query, the

[1] AvantGo. http://www.avantgo.com.

user was given the title and the description of each query (i.e., the "title" and "description" fields of the respective TREC topic[2]) providing the necessary background to their 'information need' to allow them to make relevance judgements. Once the user indicated to the experimenter that they were ready to proceed the experiment was started. At that point, timing for that specific query started and the user was presented with a ranked document list, composed of the 50 highest ranked documents, and would be allowed to interact with the PDA. Users could select any document from the list and read its contents (see Figure 1). The document title, and the three levels of summary were used to represent document content. Initially, a user would read the title and then make a decision as to whether to mark the document as relevant/non-relevant or to proceed to the next level of summary by selecting "Next". A user can navigate back to the retrieved document list at any point by selecting "Doc List". At any point the subject could stop the system and instruct it to move on to the next document, or instruct it to show again the previous summary of the current document. Documents judged relevant/non-relevant were marked so by the user on an answer sheet that was prepared for each query. In addition, the user marked the level of summary used to make their decision.



Retrieval List: (Query 36)

Please select a document to read:

[Doc 1]
[Doc 2]
[Doc 3]
[Doc 4]
[Doc 5]
[Doc 6]
[Doc 7]
[Doc 8]
[Doc 9]

7 % of doc 1:

Three utilities said they plan a $39million experiment in the California desert to test for thefirst time on a large scale a new technique for capturing andstoring solar energy.The 10-megawatt electric generating facility -- enoughpower for about 5,000 customers -- will heat molten salt asthe first step to

Figure 1. Examples of screen shots.

Once the assigned task was completed (i.e. all the documents were marked or the time elapsed), the user was given the next query and the process was repeated. At the end of the experiment the user was given a questionnaire. The purpose of the questionnaire was to gather additional information on the user's interaction with the system: the utility of the document descriptions, the clarity of reading the description through the PDA interface, the level of difficulty of using the interface, and the level of difficulty of the queries.

[2] Examples of TREC topics are available at http://trec.nist.gov/data/testq_eng.html

There are some limitations to the methodology we used in our experiment. A first limitation pertains to the use of the TREC relevance assessments as the "ground truth" against which user judgments are compared in order to obtain precision and recall values. A second relates to assessing the form-factor of viewing textual content on a PDA device. Current web browsers for handheld devices do not take into account the different display capabilities, and the onus is on the content provider to produce suitable content. The HTML files viewed by user in the experiment were set to "word-wrap" to be consistent with previous experiments, and therefore only partially assessed the effect of page scrolling (horizontally) due to PDA web browser limitations. Finally, a further criticism of our experimental procedure may be the decision to ask the user to identify as many relevant documents as possible within the allotted time. It could be argued that by adopting this approach users maybe encouraged to decide upon the relevance of each document on the minimum amount of information. The result potentially leading to a bias in the decision threshold favouring a "relevant" response. Possibly a better approach would have been to explicitly mention to users that in addition to identifying relevant documents, they must also consider that their performance scores would be penalised if they make mistakes. However, it is fair to assume on the basis of our experimental results (see section 'Results') that the majority of the users (9 out of 10 users) correctly understood the experimental task using the full range of available summaries to make their decisions, with only one user possibly misunderstanding the task and basing their decisions on mainly document "titles".

4.3 Experimental Measures

Experimental measures we used to assess the effectiveness of user relevance judgements were the accuracy and speed of judgements. The speed of user judgements is the time that a user took to assess the relevance of a single document and, to quantify accuracy; precision, recall and decision-correctness were used. In our experiment we focus on the variation of these measures in relation to the different experimental conditions. This is in contrast to the absolute values normally used in IR research.

We define *precision* then as the number of documents marked correctly as relevant (in other words, found to be relevant in agreement with the TREC judges' assessments) out of the total number of documents marked, and *recall* as the number of documents marked correctly as relevant out of the total number of relevant documents seen. A further measure we used to quantify the accuracy of a user's judgment is *decision-correctness*, that is the user ability to identify correctly both the relevant document and the non-relevant (irrelevant) documents. We define decision-correctness as the sum of the number of documents marked correctly as relevant, plus the number of documents correctly marked as non-relevant out of the total number of documents marked for that query.

5 Results

We now report the results of the experimentation outlined in the previous section. A full analysis of all the data produced during the experimentation is outside the scope of this paper. Instead, we present those results that we believe to be most interesting.

Table 1 reports the average precision, average recall, and average time for each user regardless of the summary level used to make the relevance decision. The values report a variation in the precision and recall among the users. It is apparent that some users have high levels of precision (users 4 and 10), while others (in particular user 7) have very low levels. It may be reasoned that the low level of precision cannot be fully explained by a hasty decision, since the fastest two users both show higher levels of precision. It is interesting to notice that the user with the highest level of recall is also the one with the highest level of precision and among the fastest users. The slowest user (user 8) is among the users with the lowest levels of precision.

Table 1. Average precision, recall and time for the overall PDA experiment.

	User										Avg.
	1	2	3	4	5	6	7	8	9	10	
Avg. Precision (%)	66.67	50.00	50.00	83.33	67.50	54.71	32.54	50.00	69.05	73.02	51.20
Avg. Recall (%)	55.83	29.37	75.00	75.00	70.83	83.33	61.11	28.57	91.43	46.28	61.59
Avg. Time (secs)	38.24	21.20	39.29	36.79	31.65	42.62	27.05	47.71	22.89	32.85	34.23

Comparing these results with those of a similar study carried out on a WAP mobile phone interface[3] shown in table 2. We can observe that the overall performance in terms of effectiveness is better for the WAP experiment users. This is maybe the opposite of what we would expect, given that a larger display area allows more content to be read and one could argue therefore reduces some of the cognitive overhead of having to remember what was mentioned in earlier sentences that are out of view. The results for the PDA experiment are skewed by the precision of lowest performing user (user 7). One further interesting observation is that the two highest performing users were consistent in both experiments (users 4 and 10). A possible reason for this is the content of these queries may have been easier for the users to digest, the topic being either the subject of current affairs or common knowledge.

Table 3 reports the average decision-correctness for each user for both the PDA and WAP experiments. These values maybe considered as reflecting the users ability to make correct decisions, identifying both relevant and irrelevant documents correctly. These results show that overall the differences in making

[3] The results reporting in [20] contained errors and the values are incorrect. Instead please refer to [19] for the corrected results.

Table 2. Average precision, recall and time for the overall WAP experiment.

	User										Avg.
	1	2	3	4	5	6	7	8	9	10	
Avg. Precision (%)	46.43	44.05	25.00	75.00	35.00	54.50	41.00	66.67	48.95	71.43	55.84
Avg. Recall (%)	85.71	66.67	16.67	51.67	66.67	65.00	83.33	87.41	67.86	64.25	68.75
Avg. Time (secs)	25.03	42.14	21.43	31.57	35.57	27.70	36.76	22.64	23.82	23.71	29.02

correct decisions for the experiments is in fact smaller, but that performance of WAP users remains higher. A further interesting observation is that the lowest performing users in terms of precision for the PDA experiment (user 7) is actually among the users making the highest correct-decisions (user 7 correctly identified a number of irrelevant documents).

Table 3. Average decision correctness (DC) for the PDA and WAP experiments.

	User										Avg.
	1	2	3	4	5	6	7	8	9	10	
Avg. DC. PDA (%)	76.59	80.45	91.67	79.40	81.17	50.56	74.76	43.71	71.24	45.95	71.97
Avg. DC. WAP (%)	78.01	56.43	90.33	82.67	70.93	67.51	79.54	81.77	76.91	59.54	75.96

Analysing in more detail how users employed the different summary levels to make their decision, table 4 reports on the number of documents that were assessed by each user at different summary levels. In contrast to the users in the WAP experiment, the consistency among our PDA users on employing a particular length of summary is not as apparent, with the notable exception of the 7% summaries. Again, a similar pattern emerges that users tend to base their relevance decisions mainly on the shorter length of summaries (7% of the length of the document). There is however a slight increase in the use of the longer summaries and this has an impact on the total number of documents seen by users that participate in the PDA experiment.

Table 4. Number of documents at the different levels of summary that users utilised to make decisions.

	User										Total PDA		Total WAP	
	1	2	3	4	5	6	7	8	9	10				
Title	9	44	13	6	17	10	29	0	18	25	171	34%	233	41%
7%	19	24	20	14	28	20	24	20	20	20	209	42%	271	48%
15%	11	5	6	12	6	9	5	12	24	0	90	18%	50	9%
30%	4	2	3	12	0	1	5	0	6	0	33	7%	16	3%
Total	43	75	42	44	51	40	63	32	68	45	503	100%	570	100%

Table 5 provides a better insight into the results reported in table 4 where the average precision for different users at different summary levels is reported. Comparing the overall values there is a slight decline in users ability to correctly identify relevant documents for the PDA experiment. This pattern is also evident in decision correctness despite higher values in terms of performance[4]. Within the values shown in table 5, the occurrence of '0.0' precision refers to a decision that was marked as either non-relevant or a series of incorrect decision[5] and 'IND' denotes that a decision was not made using that particular level of summary. The user with the highest precision (user 4) was also amongst those users that showed the highest levels of decision correctness and the user with the lowest decision-correctness (user 8) was also among the users with the lowest precision.

Table 5. Avg. precision (%) for the different levels of summary.

	User										Total	Total
	1	2	3	4	5	6	7	8	9	10	PDA	WAP
Title	50.00	50.00	0.00	0.00	0.00	50.00	16.67	IND	48.75	43.75	53.19	54.10
7%	33.33	50.00	50.00	100.00	58.33	52.14	12.50	50.00	37.50	62.50	51.25	60.71
15%	50.00	100.00	0.00	50.00	50.00	0.00	100.00	33.33	50.00	IND	50.00	41.18
30%	0.00	0.00	50.00	50.00	IND	0.00	0.00	IND	100.00	IND	50.00	28.57
Total	66.67	50.00	50.00	83.33	67.50	54.71	32.54	50.00	69.05	73.02	51.20	55.84

Figure 2 reports another direct comparison, between the average precision for both experiments of the first and last queries presented to users. This comparison highlights the effect of fatigue in the relevance decision process. Whilst fatigue was not an experimental variable being measured it seems a likely reason for the observed drop in performance. This effect is important when comparing the results of the experiments (the first query PDA with the first query WAP). It can be noted that both sets of users perform better for precision (and recall) in the first query as opposed to the last. The effect of fatigue can also be seen in the average time taken to make the relevance decision (not shown in figure 2), users tend to be taking more time for their decision in the first queries, the time notably decreases in the last queries and this reflects on the accuracy of the relevance decision. Effects from fatigue are more apparent in the PDA experiment compare to the WAP experiment.

Another interesting comparison is reported in figure 3. This graph reports the average precision, average recall, and average time for short and long queries.

Long queries were defined as those above a median length value, and short queries defined as those below this value (less than or equal to 6 lines are considered as short). Although not highly pronounced, observations show a difference

[4] Due to constraints on paper length the full results for decision correctness are not report.

[5] There was only one occurrence of consistently incorrect decisions that resulted in a decision-correctness of '0.0' (user 2 at 30%).

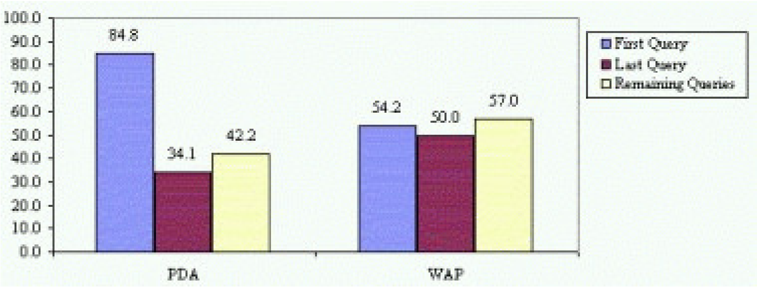


Figure 2. Average precision for the first, last and remaining queries for the PDA and WAP experiments.

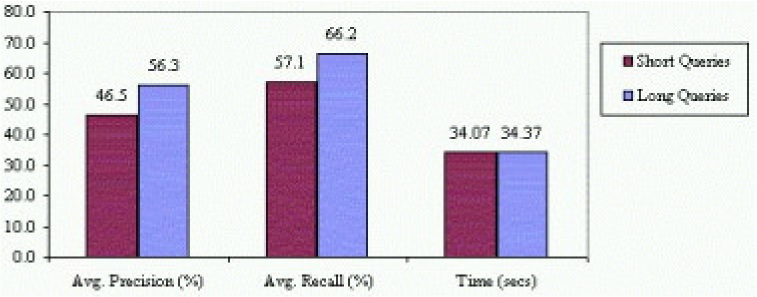


Figure 3. Average Precision, Recall and Time for Long and Short Queries for the PDA Experiment.

in average precision recall for short and long queries. The results of this study agree with the findings of a previous experiment [21], that long queries contain more information for the relevance decision to be taken and therefore enable a user to produce higher levels of recall and precision. Another interesting finding, confirmed by the results, is that longer queries do not require longer times for the relevance decision, this can be attributed the techniques employed by the user to make the relevance assessments. This effect could be due to users employing a ”Keyword Spotting” strategy to making their relevance decisions [22].

From tables 1-5 and figures 2-3, we can conclude that the presentation of documents like news articles on the small screen continues to be both feasible and relatively effective. In fact, considering only the 15% document summary length[6], the results show an improvement in levels of effectiveness compared to those found when users assess the relevance of documents on a WAP mobile

[6] To compare the results with those for other modalities only users' performance at the level of 15% summary length can be used. However, the findings of our experiment show that comparisons based on 15% summary length do not fully represent the overall performance of our PDA users since they were most effective with 7% summary lengths and actually less effective overall than users of WAP also using 7% summary lengths.

phone interface [20]. In particular, the average levels of precision are similar to those found when documents summaries are presented to a user on a computer screen [22] but have slightly lower levels of recall. The average levels of recall are similar to those found when documents summaries are presented in spoken form [21].

6 Conclusions and Future Work

In recent years there has been a large increase in the use of small screen devices. The technological advances of such devices mean that many can now be considered as information terminals, capable of supporting information access tasks normally only associated with the desktop PC. However, despite the advances in device technologies there remains difficulties in supporting user interaction on such devices. This is largely due to approaches designed for the large screen of desktop being applied directly migrated to the small screen. Conducted information retrieval tasks on small devices then proves to be difficult.

The work reported in this paper is a continuation of work assessing the effectiveness of using hierarchical query-biased summaries in the context of IR on non-traditional IR platforms. We propose the use of summaries as a means to improve interfaces for search results presentation on small screen devices. This experiment is aimed at measuring users' perception of relevance of hierarchical query-biased summaries, representing the full text of documents, viewed on a PDA device interface. The difference in users' perception of relevance relating to the judgment conditions and forms of response is compared.

Our results agree with the notion that users' perception of relevance is highly influenced by factors relating to the form of information presentation [21]. The results highlighted, for PDAs, a preference for concise summaries that are relatively brief, 7% of the document length (up to a maximum of 3 sentences) compared with other summary lengths used in the experiment. Questionnaires completed by the users suggest that hierarchical query-biased summaries are useful and assist users in making relevance judgments. The results are consistent with the findings of our previous study that found for small screen displays (WAP mobile phone interface) users showed both a preference and better performance with the shorter summary lengths (7% of the document length) [20]. Further support for presenting concise relatively brief summaries on small screen devices comes from the findings of a recent WAP usability study[7].

Limiting factors of our study include: the use of a small user sample that had similar experience of using current technologies thus representing only a small proportion of the user community, and using the TREC collection to simulate an information discovery task.

As future work, using the results we have presented as a basis for supporting the use of summarisation as a better means of representing the results of a IR search, we intend to investigate the generation and use of adaptive content-aware

[7] Carried out my Nielsen Norman Group. WAP Usability Report, December 2000. Available at http://www.nngroup.com/reports/wap

summarisation techniques that present content to a user on the basis of their means of access, the device being used. We envisage at that such a framework may provide better support for information access that is platform independent.

Acknowledgements

This work is supported by the EU Commission under IST Project MIND (IST-2000-26061). More information about MIND can be found at http://www.mind-project.net/.

References

1. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.
2. R. Brandow, K. Mitze, and L. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685, 1995.
3. J. Burkhardt, H. Henn, S. Hepper, K. Rintdorff, and T. Schack. *Pervasive Computing Technology and Architecture of Mobile Internet Applications*. Addison-Wesley, London, 2002.
4. O. Buyukkokten, H. Garcia-Molina, A. Paepcke, and T. Winograd. Power Browser: efficient web browsing for PDAs. In *Proceedings CHI2000*, pages 430–437, Amsterdam, 2000.
5. H. Chen, A. Houston, R. Sewell, and B. Schatz. Internet Browsing and Searching: User evaluations of category map and concept space techniques. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval*, pages 272–274. Addison-Wesley, 1999.
6. H. Edmundson. New methods in automatic abstracting. *Communications of the ACM*, 16(2):264–285, 1969.
7. D. Harman. Relevance feedback and other query modification techniques. In W. B. Frakes and R. Baeza-Yates, editors, *Information retrieval: data structures & algorithms*, pages 241–263, 1992.
8. A. Jameson, R. Schfer, T. Weis, A. Berthold, and T. Weyrath. Making Systems Sensitive to the User's Time and Working Memory Constraints. In *Proceedings of the 4th International Conference on Intelligent User Interfaces*, pages 79–86, Los Angeles, California, 1998. ACM Press: New York.
9. B. 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):207–227, 2000.
10. M. Jones, G. Buchanan, and N. Mohd-Nasir. Evaluation of WebTwig - a site outliner for handheld Web access. In *Proceedings International Symposium on Handheld and Ubiquitous Computing*, volume 1707, pages 343–345, Karlsrhue, 1999. Lecture Notes in Computer Science, Springer, Berlin.
11. M. Jones, G. Buchanan, and H. Thimbleby. Sorting Out Searching on Small Screen Devices. In *Proceedings of the 4th International Symposium on Mobile HCI*, pages 81–94. Springer, 2002.
12. M. Jones, G. Marsden, N. Mohd-Nasir, and K. Boone. Improving Web Interaction on Small Displays. In *Proceedings of 8th WWW Conference*, Toronto, Canada, May 1999.

13. G. Loudon, H. Sacher, and L. Kew. Design Issues for Mobile Information Retrieval. In *Proceedings of Workshop on Mobile Personal Information Retrieval*, pages 3–14, Tampere, Finland, August 2002. ACM SIGIR 2002.
14. H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal*, pages 159–165, April 1958.
15. C. Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, 26(1):171–186, 1990.
16. J. Rush, R. Salvador, and A. Zamora. Automatic abstracting and indexing. ii. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science*, 22(4):260–274, 1971.
17. T. Saracevic. Comparative effects of titles, abstracts and full texts on relevance judgements. In *Proceedings of the American Society for Information Science*, pages 293–299, 1969.
18. E. Schofield and G. Kubin. On Interfaces for Mobile Information Retrieval. In *Proceedings of the 4th International Symposium on Human-Computer Interaction with Mobile Devices*, Pisa, Italy, September 2002.
19. S. Sweeney. Hierarchical query-biased summaries for WAP mobile phones. Master's thesis, University of Strathclyde, Glasgow, UK, 2001.
20. S. Sweeney, F. Crestani, and A. Tombros. Mobile Delivery of News using Hierarchically Query-Biased Summaries. In *Proceedings of ACM SAC 2002*, pages 634–639, Madrid, Spain, March 2002.
21. A. Tombros and F. Crestan. Users's perception of relevance of spoken documents. *Journal of the American Society for Information Science*, 51(9):929–939, 2000.
22. A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *Proceedings of ACM SIGIR*, pages 2–10, Melbourne, Australia, August 1998.
23. C. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition edition, 1979.
24. E. Voorhees. Overview of TREC 2001. In *Proceedings of the 11th TREC Conference*, Gaithersburg, MD, USA, November 2002.
25. J. Wise, J. Thomas, K. Pennock, D. Lantrip, M. Pottier, and A. Schur. Visualizing the non-visual: Spatial analysis and interaction with information from test documents. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval*, pages 272–274. Addison-Wesley, 1999.