Effective search results summary size and device screen size: is there a relationship?

Simon Sweeney and Fabio Crestani*

Department of Computer and Information Sciences

University of Strathclyde

Glasgow G1 1XH, Scotland, UK

24th June 2005

Abstract

In recent years, small screen devices have seen widespread increase in their acceptance and use. Combining mobility with increased technological advances many such devices can now be considered mobile information terminals. However, user interactions with small display devices remain a challenge due to the inherent input restrictions and limited display capabilities. These challenges are particularly evident for tasks, such as information seeking. For the presentation of retrieval results we consider that a personalised and context dependent approach could offer benefits, particularly for retrieving information in a non-traditional environment. As a starting point, in this paper we report an investigation into the effects of summary length as a function of screen size, where query-biased summaries are used to present retrieval results. Following a brief description of our proposed system, we report a user study aimed at exploring whether there is an optimal summary size for three types of device (smartphone, PDA and laptop), given their different screen sizes.

Keywords: Information Retrieval, results presentation, small screen devices, summarisation.

*Author to whom correspondence should be sent. Tel: +44-141-548 4303. Fax: +44-141-548 4523. Email: f.crestani@cis.strath.ac.uk.

1 Introduction

Technology is having an ever increasing impact on our lives, this is evident in the increasing advances in information technology to support access to information on demand, anywhere and anytime. The emergence of these technologies and services, often referred to as pervasive computing, or ubiquitous computing (Weiser, 1991, 1993), can be seen as supplementing traditional paradigms of human and computer interaction, the desktop PC.

The acceptance of 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 fact, many of these devices can now support tasks that were normally only associated with the desktop PC, such as creating word-processed documents, spreadsheets, presentation slides. Similarly, for mobile phones (both for 2nd and 3rd generation) there exists a wide variety of network-based services (Crestani, M. Dunlop, & Mizzaro, 2004).

Significant improvements in display technologies and awareness of issues for interface design of applications for mobile devices offer the potential for an improved experience for users. Technologies such as Wireless Application Protocol (WAP), designed specifically for small handheld wireless devices, and research efforts in information visualisation also contribute to more desirable interaction. Despite the mentioned advances, however, the reality is that a challenge remains to effectively present content on mobile devices. This challenge is compounded by information overload. Challenges in presentation are due largely to the inherent constraints of a small display area and, in the case of mobile phones, limitations on interaction. Consequently, the standard approaches that exist for supporting information access on traditional platforms are not appropriate for mobile devices (Jones, Marsden, Mohd-Nasir, & Boone, 1999b).

In this paper we focus on the presentation of search results and the effects due to screen size. The paper is structured as follows. Section 2 presents the motivations for investigating the effects of screen size on retrieval results presentation. Section 3

reports related work in retrieval results presentation in terms of approaches for normal and small screen devices, and summarisation techniques used. Section 4 presents our proposal for the personalisation of retrieval results, outlining a system topology and rasing some open issues related to its design. Section 5 reports an investigation into the relationship between the summarisation of search results and the effects of screen size, presenting the results of user studies using a range of devices (smartphone, PDA and laptop). This section presents the experimental procedure, the results and the analysis of the study carried out. Finally, Section 6 presents some conclusions and directions of future work.

2 Motivations

The inherent characteristics of devices play an important role in supporting information access. This is particularly the case for mobile devices with their limited display area, constrained interaction (stylus, T9 predictive text) and their other device related factors (bandwidth, limited processing capabilities, battery life, etc.). It is also interesting to note that despite future advances both in mobile device technologies and the communication infrastructures, challenges will remain for these devices due to physical form of the device, the need to be palm sized and portable, and the resolution thresholds for what is legible by the human eye.

Previous studies (Jones, Buchanan, & Thimbleby, 2002) have suggested that improved user performance can be achieved through optimisations for user capabilities, a reflection of the device being used. With this in mind we consider that for mobile devices (phones, PDAs, laptops) the presentation of results should be personalised and context dependent. Central to this approach is the assumption that mobile devices are by definition personal devices. By personalised and context-aware delivery we refer to content that reflects the specific interest of a user (both current information requests and long standing interests), whilst also taking account of situational and environmental considerations, such as the user's current location, local time, and the device being used. We intend to use contextual information as a filtering mechanism to refine the results following the retrieval process.

There are several parameters to personalisation, particularly in the case when the user is mobile. Considering as one parameter the device and it's associated charac-

teristics then content delivery should be optimised to suit the interactions supported by the device. One means of adapting search results presentation is to employ summarisation techniques, the aim being to summarise the results with minimal loss of user perception of relevance. Indeed, some forms of summarisation can improve user perception of relevance, using for example query-biased techniques (Tombros & Sanderson, 1998). However, existing approaches to adapting content for devices do not support the needs of the individual user (e.g. (Buyukkokten, Garcia-Molina, & Paepcke, 2001)), content-device optimisations are generic and would be the same for all users.

As a starting point to developing an approach that is device and user sensitive we have investigated the effects of screen size on retrieval results presentation. This paper presents our findings on exploring the following research question: is there an optimal a-priori (i.e. not related to personal preferences) summary size, given the device screen size? In other words, is search results presentation affected by screen size, and is there an optimal summary size, for the document retrieved in response to a search, that is a function of the screen size? These findings would be useful for addressing the "cold start" problem, since personalisation would then enable a user to change summary size to suit their preference. The work reported in this paper focuses on the delivery of news, which are a particular type of document.

The hypothesis we use for the user study relate to the presentation of search results and effects due to screen size, and can be outlined as the following. In order to keep the user perception of relevance at a constant level, should summary length be related to screen size? Also, considering the effectiveness of a summary related to how it enables a user to perceive the relevance or not of a document, is the intuition that a long summary is more effective for large displays and less effective for small displays true? Conversely, is it true that a short summary is more effective for small displays and less effective for a large displays? We expand on the details of our research questions in Section 4.2.

3 Background

Traditionally, information retrieval (IR) systems are accessed using a desktop PC where the results of a search are presented on a large screen display and there

exists a rich environment for interaction. Considering the topology of user in this traditional setting they might range from experienced experts, possibly with formal training in conducting information searches, to novice users who are at the very least computer literate. In most circumstances, when engaged in an information access (IA) task it is reasonable to assume that users will devote their complete attention to the task. Comparing IA in a mobile environment to the conventional setting appears to show substantial differences (Loudon, Sacher, & Kew, 2002), one being the cognitive effort devoted to the task of accessing information. IA on a mobile device seems to require considerably more effort than on a desktop PC. Differences in device have an impact since mobile devices by design are multi-purpose and as a consequence may compromise certain useful functionality to maximise mobility and diversity (small display, limited interaction, etc). Any difficulties experienced in using these devices may be magnified when the utility is extended to support new tasks, such as searching for information in digital collections, or on the web.

In addition, the profile of a typical mobile device user is different from that usually associated with an IR system user in the sense that computer proficiency, and indeed acceptance of technology as a whole, may not be assumed. This is apparent when considering the variety in user profiles within the mobile phone user population.

The retrieval task itself differs from that assumed under normal IR circumstances due to the nature of conducting searches in a non-static, transient environment, where there is a greater risk that user performance may be influenced by outside factors with the increased potential for distractions of noise and interruptions (Jameson, Schfer, Weis, Berthold, & Weyrath, 1998). These distractions may even be user driven, as the user may be engaged in other activities at the time of searching and cannot commit their full attention to the task of accessing information.

Finally, for mobile IA there is an increased prominence to consider the type of information being sought as there may be significant temporal and/or locational 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 the mentioned factors then influence the way mobile users will conduct searches and view search results.

3.1 Searching on the desktop

Most search systems present the results of a user query as a serial list of documents, often represented by document title, 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 as many search engines provide large result sets that may span many pages. In reality, studies have shown that users are only prepared to look at the first 10-20 results (Kirsch, 1998). To reduce the overhead involved in working through the list of retrieved documents, approaches have been developed to assist users in completing their information discovery task.

Applying *ranking* to the list of retrieved documents, is one such approach. In IR query-relevance ranking normally takes precedence, presenting those documents the system considers as best matching the users query higher in the list (Baeza-Yates & Ribeiro-Neto, 1999). Other measures that can be used for ranking results include time, or document size. Complimentary techniques may focus on attempting to improve the quality of the results ranking, attempting to increasing the position and number of relevant documents in the retrieved document result set. *Relevance feedback* is an example of such a technique, whereby the system refines a set of results or performs a further search on the basis of user explicit or implicit correction (Harman, 1992).

Information Visualisation techniques explore alternatives to presenting search results in contrast to traditional ranked lists. Many of these schemes make use of colourful highlighting and graphical features to capture aspects of the information access process, presenting content that is dynamic and can be interactively manipulated by the user (Baeza-Yates & Ribeiro-Neto, 1999). For example, the use of concept "landscapes" to represent document clusters (categories of similar documents) displayed graphically, in 2D as a "jigsaw" (Kohonen Feature Maps (Chen, Houston, Sewell, & Schatz, 1999)) with the clusters forming the individual pieces,

or in 3D as a "map" (ThemeScapes (Wise, Thomas, Pennock, Lantrip, Pottier, & Schur, 1999)) with contours describing document similarity and where peaks indicate concentrations of similar documents.

Another variation to a plain list of document titles is to accompany the document title with supplementary information describing the retrieved document. This additional information functions as a *document surrogate* providing document metadata, such as date of publishing, source, and length of the document, to give more indication about the content of a document (Baeza-Yates & Ribeiro-Neto, 1999).

More relevant to our study, 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. Further by applying techniques borrowed from automatic text summarisation the extracts for the document surrogate can be improved to be more representative of the source document, being informative, instantly fulfilling the user's information need, or indicative, providing an indication of whether the particular document is relevant (Brandow, Mitze, & Rau, 1995).

3.2 Searching on the small screen devices

Small screen devices provide much 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. Whilst similar functionality is provided, in practical terms using such services results in a very different user experience (Jones et al., 1999b). 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 retrieved matches.

Recent human-computer interaction (HCI) studies have found that supporting information discovery tasks, both browsing and searching, on PDAs using interfaces designed for the display area of desktop PCs has a negative influence on task performance (Jones et al., 1999b; Jones et al., 2002).

These studies highlight problems with search interfaces for small screen devices as being associated with the within-page vertical scrolling or paging requirements for viewing content. Vertical scrolling describes the action of viewing the content

outside the screen display area shown in a progressive manner, serially. By contrast, paging permits access to the next full screen's worth of content without any further action by the user. To make content available for displaying on small screen devices, it is not uncommon for long lists of search results to be divided into separate pages that contain a reduced number of results. Breaking the content up into smaller, more manageable chunks may be necessary for transmission requirements,[1] as well as a means of aiding presentation. Nevertheless, there is an associated cost with this approach; page-to-page navigation increases the amount of user interaction and reading time (Jones et al., 1999b). Both of these factors may have financial implications (users are likely to be paying for wireless connections or the amount of data they transfer) and as a consequence may impact on the way users use such services. Continuing with navigational issues, the worst effects are observed if users are required to scroll horizontally. In such cases, it is easy for users to become disorientated and lost within content designed for viewing on much larger screens (Jones et al., 1999b).

Solutions to assist the user in making sense of search results on the small screen can be briefly outlined as the following. Limiting number of results in each results page and limiting the amount of information displayed for each result (document surrogate) has the benefit of reducing the long lists of results instead of splitting them over multiple pages.

The Google interface for the PDA displays only the top 5 results per page and for each result. A further difference is the use of symbols to represent features expressed more completely in the full version. However, there is an increased requirement for page-to-page navigation which is a negative effect. Combining relevance ranking with algorithms that favour high precision performance may provide a trade-off to splitting content over multiple of pages and the associated navigation costs. Ideally, the most relevant results would then 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 (Jansen, Spink, & Saracevic, 2000). Also, providing quality document surrogates in the retrieved document list may enable the user to be more selective in choosing documents to view, again potentially reducing the need to visit all

[1]WAP protocols require that a deck (unit of deliverable content) to be no greater than 1000 bytes.

retrieved documents.

Notable schemes for accessing web content on mobile devices are WebTwig (Jones, Buchanan, & Mohd-Nasir, 1999a) and PowerBrowser (Buyukkokten, Garcia-Molina, Paepcke, & Winograd, 2000) both were 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. It is interesting to observe that both these schemes have more recently incorporated features that use forms of summarisation (Buyukkokten et al., 2001; Jones, Jones, & Deo, 2004). This is the approach we intend to follow.

For a discussion of content adaption techniques for small screen viewing refer to (MacKay & Watters, 2003) where an number of approaches are discussed under the headings of direct manipulation, data modification, data suppression and data overview.

3.3 Applying summarisation to retrieval results presentation

Automatic summarisation has been used extensively in the context of IR. Both as a means of supplementing search results, and therefore aiding the user to make relevance assessments, and for making the IR process more efficient using. For example, a summarised version of documents can be used to build indexes or for storage, in place of the document full text.

Traditionally, automatic document summarisation has been based on sentence extraction approaches (Brandow et al., 1995; Edmundson, 1969; Luhn, 1958). Advances in sentence extraction have seen the introduction of query-biased summarisation 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 a 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 (Rush, Salvador, & Zamora, 1971).

Highlighting recent research in the application of summarisation to aid infor-

mation 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 (Tombros & Sanderson, 1998). The results from their evaluation indicate that the use of query biased summaries significantly improves both the accuracy and speed of user relevance judgements compared with the typical output of an IR system, that is a static predefined summary composed of the title and first few sentences of retrieved documents. A later study by Tombros and Crestani evaluated the effectiveness of presenting summaries by different means and the effect this has on users' perception of relevance (Tombros & Crestani, 2000). Results from this study showed that users' ability to make relevance assessments of documents was 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 (Sweeney, Crestani, & Tombros, 2002). 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).

Other related research in search results summarisation combines more recent trends in multi-document text summarisation includes approaches based on linguistic analysis (Radev & Fan, 2000), and with focus on small screen delivery (Radev, Fan, & Zhang, 2001; Boguraev, Bellamy, & Swart, 2001). More general work on text summarisation for small screen devices has seen specific approaches targeted at email processing/viewing (Corston-Oliver, 2001), financial news delivery (Yang & Wang., 2003) and web page viewing (Buyukkokten et al., 2001).

4 Personalisation of results presentation

The case for device optimised personalised content delivery has been presented, with the objective that content should reflect the constraints, or indeed take full advantage of the functionality, of the end user platform and take account of users interests. Automatic summarisation offers a solution for presenting device-friendly content. There are many ways to produce summaries and query-biased techniques; (Tombros & Sanderson, 1998) present one example of a current user-centred approach.

A limitation of this approach, however, is that the summary generated only takes account of the users current request, and not of a long standing interest, such as a user profile. An effective way of summarising for personalisation would be to learn from the users how to make the best summary for their particular information need and specific devices being used. Additionally, contextual information could be used to further refine the content of the summary.

4.1 An architecture for personalised and context-aware summarisation

By experimenting with the system, proposed above, we aim to investigate the potential merits of a personalised and context aware approach. The underlying architecture for our system is based on the traditional client-server model where the server communicates with all types of client devices. Using a central server promotes independence and consistency among the possibilities of client devices since the software responsible for content adaption is not tied to particular device technology. There is also the added advantage of reducing processing overheads for thin clients (mobile phone, PDAs).

An illustration of the overall system processes is shown in Figure 1, where the different types of device all communicate through the central server. The results returned to the user in respect of their initial request for information are different for different devices, for different tasks, and for different users.

The architecture of the server is component based, dividing the overall process into a number of separate stages. Conceptually the server is split into two parts: the (a) push and pull technologies, and the (b) adaption engine. We intended to

use existing IR and information filtering (IF) technologies to provide the retrieval and filtering capability for the system and will focus on the adaption engine as the main part of our work.

4.2 Open issues

To realise a solution that is personalised to the user and adapted to the device being used to access information there are a number of open issues that require consideration. Comparing devices on a discriminating factor such as screen display area and regarding this dimension in isolation, that is assuming the user (or personalisation) is a constant, then there are a number of issues for the delivery of content. *Should we consider the screen size when presenting the results of a search?* Existing research would suggest that screen size does have an impact (refer to Section 3.2) and therefore should be taken into account for content delivery. If the answer is yes and screen size does have to be considered, so that users should see different summaries for different devices then, *is there a relationship between screen size and the amount of information the user wants to see?* In other words, is the screen size a variable to be consider in the task of generating summaries of retrieval results? We must also establish are users expectations. Do they expect a certain amount of information given a particular screen size? Finally, *is it effective to summarise retrieval results considering the device screen size?* Assuming users do indicate a preference for content length on the basis of screen size, is this actually effective? Does it provide improvements in performance, which in this study, refers to users' ability to carry out correct relevance judgements.

The experiments described in the following section aims to investigate these open questions.

5 Investigation into the relationship between retrieval results summarisation and screen size

The following sections report on a series of experiments that investigate the relationship between retrieval results summarisation and screen size. The general theme of the investigations is users' ability to carry out relevance judgements on textual

information presented on non-traditional IR platforms. We are interested in assessing the variation of user performance in evaluating the relevance of full documents, given query-biased summaries of different lengths, and also determining whether there is an optimal size of summary for a given type of device interface. The range of mobile devices used in the experiments comprised a smartphone, PDA (personal digital assistant) and a laptop. For the study we assume the utility notion of relevance (van Rijsbergen, 1979) as the basis for evaluating the summaries. Further details describing the context for the users' perception of relevance used can be found in (Tombros & Crestani, 2000).

5.1 Research questions for the study

In this study we focus on the effects of summary length as a function of device screen size, and leave investigating aspects of personalisation and use of context in our proposal for another time.

Relating to the open issues which we have identified in Section 4.2, the study aims to explore the following: Is there an optimal a-priori (i.e. not related to personal preferences) summary size, given the device screen size? In other words, is search results presentation affected by screen size, and is there an optimal summary size, for the document retrieved in response to a search, that is a function of the screen size?

To test this research question we devised the following hypotheses which are based on our initial intuitions. Given the advantages of a larger screen that,

1. users would visit a greater number of documents using the larger screen compared to using the smaller screens;

2. users would make the majority of decisions using the longer levels of summary (15% and 30%) on the larger screen and make fewer decisions using the shorter levels of summary (title and 7%), while on the smaller screens users would make more decisions using the shorter summaries and make fewer decisions using the longer levels of summary;

3. for both large and small screens, the time taken to make decisions would be similar given that larger screen users may tend to read more and scroll less,

while for the smaller screen read less and scroll more;

4. users would achieve higher performance (a higher level of decision correctness, precision and recall) using the larger screen compared to a lower level of performance using the smaller screens;

5. users would achieve higher performance with longer summaries on the larger screen, and similarly would achieve higher performance with shorter summaries on the smaller screen. In other words, that longer summaries would be more effective for the larger screen, while for the smaller screen shorter summaries would be more effective.

5.2 Presenting summaries on different devices

We will present here the experimental environment in which the study was carried out.

5.2.1 Devices used in the experiment

The devices for the experiment were chosen to represent a range of screen sizes from the spectrum of mobile devices available. Comparing the different devices based on the characteristics of their display then we can group devices as micro displays (mobile phones, smartphones), small displays (PDAs, Pocket PCs and other handheld devices), and normal displays (tablet PCs to Desktop PCs), as reported in Figure 2. Also apparent is that along with increasing screen size there is also increase in screen display quality.

In our experiments we used an SPV E200 smartphone, a HandSpring Visor PDA, and an Acer TravelMate 529TXV laptop. Figure 3 provides an illustration of two of the devices, the third being a 14 inch laptop. The summaries were authored as web pages and displayed through web browsers on the devices[2]. Also, we consider the benefits of using a desktop sized screen in addition to the laptop redundant for our study despite recognising it as a point on the scale of device displays (see Figure 2).

Table 1 provides a comparison of the device displays used.

[2]We used Microsoft's Pocket IE for Windows Mobile 2003 for Smartphone, AvantGo's web browser for the Palm OS v4.x (http://www.avantgo.com), and IE 6.0 for Windows.

5.2.2 Query-biased Summarisation

The summarisation system employed in the experimentation described in this paper is based on work by Tombros and Sanderson. The system uses a number of sentence extraction methods (Paice, 1990) 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 (Tombros & Sanderson, 1998); here we shall only briefly describe the summary generation process.

For the studies reported in this paper each individual document of the collection was passed through the summarisation system, and as a result a score for each sentence of each document was computed. This score represents the sentence's importance for inclusion in the document's summary. Scores are assigned to sentences by examining the structural organisation of each document, and by utilising within-document term frequency information. Information from the structural organisation of the documents was utilised in three ways. Terms occurring in the title section of a document were assigned a positive weight (title score) in order to reflect the fact that headlines of news articles tend to reveal the major subject of the article. In addition, a positive ordinal weight was assigned to the first two sentences of each article, capturing the informativeness of the leading text of news articles. Finally, a heading score was assigned to each one of the sentences comprising a within-article section heading, reflecting the fact that such headings provide evidence about the article's division into semantic units. By using the number of occurrences of a term in a document (term frequency - TF), we can establish a list of "significant" terms for that document (i.e., terms whose TF value is greater than a specific threshold). The summarisation system then locates clusters of significant terms within a sentence, and computes a significance factor for each sentence (Luhn, 1958).

In addition to the scores assigned to sentences, information from the queries that were used in the experiments and supposed to be issued by the users was also employed in order to compute the overall score for each sentence. A query score was thus computed, intended to represent the distribution of query words in a sentence. The rationale for this choice was that, by allowing users to see the context in which the query terms occurred, they could better judge the relevance of a document to

the query. The actual measure of significance of a sentence in relation to a query is derived using a query length normalisation process.

The final score for each sentence is calculated by summing the partial scores discussed above. The summary for each document is then generated by selecting the desired number of top-scoring sentences, and outputting them in the order in which they appear in the original document. 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.

Four different summary lengths were used in our experiments. It is established that titles convey useful clues about the contents of a document (Saracevic, 1969), 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. Previous summarisation research (e.g. (Brandow et al., 1995)) has suggested that summaries of roughly 20% of the original document's length can be successful in providing relevance clues to users. These sentences were the top-scoring sentences of the summarisation system, output in the order that they appear in the original document as was previously explained. Sample summaries used in the experiment are shown in Figure 4.

As mentioned, for the experiment web browsers were used to view the summaries. 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 among the experiments.

5.3 Experimental Settings

The following sections report on the test collection, the experimental procedure and the evaluation measures used.

5.3.1 The Test Collection

The documents for the experiment were a subset of the 1990-92 Wall Street Journal (WSJ) collection of TREC (Voorhees, 2002). 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. We used 50 randomly selected TREC queries (referred as topics in TREC) and for each of the queries, the 50 top-ranked documents as an input to the summarisation system. The test collection 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.

5.3.2 Experimental Procedure

To enable comparisons among devices 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 correctly 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.

Experiments were carried out with user groups of 10 volunteers with above average experience of using computers and mobile devices (mobile phones, PDAs). A total of 30 users participated in the study (10 users per experimental condition). At the outset, 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 user was given the title and the description of each query (i.e., the "title" and "description" fields of the respective TREC topic providing the necessary background to their 'information need' to allow them to make relevance judgements. When 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 device. Users could select any document from the list and read its contents (see Figure 5). The document title, and the three levels of summary (7%, 15% and 30% of the original document length) were used to represent document content.

The arrangement for the experiment is as follows (see Figure 6). Users are presented with a ranked document list (element A of Figure 6) and are able to select a document from the list and read its contents. The content for the documents are divided into Title (B), 7% (C), 15% (D) and 30% (D) of the document and this is the order that the content is presented. A user can read the title of a summary and then make a decision as to whether to mark the document as relevant/non-relevant and return to the retrieved document list to select the next document title (A), or proceed to the next level of summary by selecting Next (B). A user can navigate back to the retrieved document list at any point by selecting Doc List (A).

At any point the subject could stop the system and move on to the next document, or re-display 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.

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. A post-experiment questionnaire was used to gather additional information on each user's interaction with the system: the utility of the document descriptions, the clarity of reading the descriptions through the device interface, the level of difficulty of using the interface, and the level of difficulty of the queries.

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 mobile devices. Designing web content for small screen devices can be difficult due to the variation in display capabilities, and the onus is on the content

provider to produce suitable content. The HTML files viewed by users in the experiments were set to "word-wrap" to be consistent with previous experiments, and therefore only partially assessed the effect of page scrolling (horizontally). 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.

5.3.3 Experimental Measures

The main experimental measure used to assess the effectiveness of user relevance judgements was accuracy. 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. This definition corresponds to the standard definition of precision. *Recall* is defined 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.

5.4 Results

In this section we now report on the results of the experimentation outlined in the previous section. We present only those results that we believe to be most interesting. For results of a similar study comparing the performance of PDA and WAP mobile phone emulator refer to (Sweeney et al., 2002).

Figure 7 shows the number of documents used to make decisions as a percentage of the documents viewed for all device types in the study. An interesting observation is the similarity in the distribution of documents for the different levels of summary for both PDA and smartphone. Indeed, comparing on the basis of screen size then according to the results there exists a split in terms of an increased number of smaller summaries (title and 7%) for the smaller screens (PDA and smartphone). The opposite is also evident for the larger screen of the laptop, in that users based their relevance decision in greater number on longer summaries (15% and 30%). However, for all three devices the largest proportion of decisions were made using the 7% summaries. Interestingly, comparing the overall number of documents viewed using the different devices then there does not appear to be a great difference.

In terms of users ability to make correct decisions, that is identify both relevant and non-relevant documents, Figure 8 shows that there is a slight degradation in the overall performance with decreasing screen size. However, this difference is not significant. The Kruskal-Wallis non-parametric equivalent to the ANOVA test was used to test for significance. An interesting observation from the overall results is the similarity in performance of the PDA and smartphone. The deviations in decision performance are more evident looking at the individual levels of summary. Considering the summaries as the previously mention distinct groups, short and long summaries, then the shorter summaries (title and 7%) can be seen as outperforming the longer summaries (15% and 30%) for all devices, with the exception of the 30% summary viewed on the PDA. The variance in correct decisions at the different levels of summary is most evident for the laptop, and less clear for both the smaller screen devices.

Figure 9 provides an additional insight into users ability to make precise decisions, that is identify relevant documents only. From the results we can observe a drop in performance levels from making correct decisions, and making precise deci-

sions. It appears that users perform less well in identifying relevant documents, and therefore it is possibly to speculate from the results that the better performance in decision correctness must be achieved by correctly identifying non-relevant content (see also comments on recall later). Again, for the overall view of average precision, the previously observed similarity in performance remains for the PDA and smartphone, both achieving considerably less precision compared to the level for the laptop. There is also a more pronounced variation in performance at the individual levels of summary. The results show a pattern of decline in performance with decrease in screen size, and increase in summary length. However, the results for the PDA contradict this pattern, remaining at a consistent level for all three devices.

Recall may not be considered as good a measure for mobile IR, since one could argue that mobile users will be less inclined to investigate all the relevant items, but would rather be satisfied with the first few relevant items. The low levels for recall in Figure 10 show clearly that users have difficulty in recognising relevant documents. It is difficult to suggest the cause for this poor performance as it may be attributed to a number of factors: the specialist domain coverage of WSJ articles, the lack of exemplar relevance-decisions given to the user, or the quality of the summaries generated. Another possibility that might explain the low levels of recall might be the fact that the topics for the study were not of direct interest to the user and therefore did not motivate the user to look for documents. According to the users responses to the questionnaire (reported in the Appendix), a summary of which is presented in Table 2, they did score the level of difficulty of the queries as tending to complex (see Question 1 in Figure 6).

Figures 11 and 12 provides a basis for comparing the relative performance of the different levels of summary at the different screen sizes for correct and precise decisions. In all cases using the title and the 7% summary seems to enable the user to provide consistently precise and correct decisions. There is sufficient evidence to suggest that the title and the 7% summary achieve a higher performance for the laptop and smartphone screen sizes, while on PDA it appears that the summary length does not provide similar degrees of variation on decision correctness and precision.

In summary, the results presented here suggest that there is a relationship between screen size and summary length. The relationship does not necessarily influence the performance of users' relevance decisions, but it does influence in the choice of summary from which they make the relevance decision. Thus, while users tend to make relevance decision on long summaries when using large screens and on short summaries when using small screen, this does not have any significant effect on either the decision precision or decision correctness. Kruskal-Wallis p-value for overall decision correctness on the different devices is 0.722, and for overall precision is 0.145, both greater than 0.05. In fact, our study shows that users make precise and correct relevance decision using small summaries, whatever the size of the screen of the device they are using.

5.5 Discussions

In terms of our initial expectations, some of the results provide an unexpected outcome, a few of these cases are now discussed.

At the outset, based on our intuitions, we had some hypotheses, which we describe in Section 5.1. Given the advantages of a larger screen it seemed reasonable to have the mentioned preconceptions. However, the results show that for the majority of cases the initial expectations were proved wrong. In fact out of the listed hypotheses, only 2 and 3 were achieved. Indeed, the results confirm the pattern that users would use a greater number of longer summaries (15% and 30%) on the larger screen and the opposite for the smaller screen. Also, that the time taken to make decisions is similar[3]. More importantly, our initial view for performance stated in point 5 were proven not to be the case: that longer summaries would be more effective for the larger screen, while for the smaller screen shorter summaries more effective. In terms of an optimal summary length for the different devices, our experimental results suggest it is best to show the same level of summary on all devices, and that our experiment users were most effective with the shorter summary lengths (title and 7%), whatever the screen size.

There results have important implications in the design of our personalised and context-aware result presentation for mobile devices, in that we will now use short

[3]Note, as task duration was a controlled variable in the experiments (5 minutes per query) we are not able to report any data relating to the time spent at the different levels of summary.

summaries independently from the screen size of the device being used. It will now be our strategy to allow the user to request longer summaries and adapt the default summary length to user preferences. However, in the absence of any user preference, a short summary will be delivered.

So, to the question posed in the title of this paper on whether there is a relationship between search results summary size and device screen size, our answer, supported by the findings of this study, will have to be no.

6 Conclusions and Future Work

This paper presents an investigation into the effectiveness of using query-biased summarises of varying length to present retrieval results with respect to different screen sizes. The aim was to establish experimentally whether screen size needs to be considered in the presentation of search results, and if the case, if there is an optimal summary size for a particular size of screen. Whilst the results of the experiments indicate that summary size should not be selected in relation to screen size, it does show that screen size is an important factor and needs to be considered in the design of personalised results presentation, reflect by the varied usage and preferences of the users. In terms of an optimal summary size for the different screen displays, on the basis of effectiveness for the task in our experiment, the results hint at shorter summaries being more effective for all screen sizes.

As future work, we intend to fully develop the system for personalised and context-aware search results presentation sketched in Section 4.1. We believe that such a tool may provide better support for a platform independent and user-centered information access.

Acknowledgements

This work is supported by the EU Commission under IST Project PENG (PErsonalised News content programminG Information). More information about PENG can be found at `http://www.peng-project.org/`. We would also like to acknowledge all the experimental participants and the reviewers for their useful feedback.

7 References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley.

Boguraev, B., Bellamy, R., & Swart, C. (2001). Summarisation Miniaturisation: Delivery of News to Hand-Helds. *NAACL 2001 Workshop on Automatic Summarization*. Pittsburgh.

Brandow, R., Mitze, K., & Rau, L. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing and Management, 31*(5), 675–685.

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001). Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. *Proceedings of 10th International WWW Conference (WWW10)* (pp. 652–662). Hong Kong, China.

Buyukkokten, O., Garcia-Molina, H., Paepcke, A., & Winograd, T. (2000). Power Browser: efficient web browsing for PDAs. *Proceedings CHI2000* (pp. 430–437). Amsterdam.

Chen, H., Houston, A., Sewell, R., & Schatz, B. (1999). Internet Browsing and Searching: User evaluations of category map and concept space techniques. In R. Baeza-Yates, & B. Ribeiro-Neto (Eds.), *Modern Information Retrieval* (pp. 272–274). Addison-Wesley.

Corston-Oliver, S. (2001). Text compaction for display very small screens. *Proceedings of Automatic Summarization Workshop*, the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL). Pittsburgh, PA, U.S.A.

Crestani, F., M. Dunlop, M., & Mizzaro, S. (Eds.). (2004). *Mobile and ubiquitous information access: Proceedings of the Mobile HCI 2003 workshop on mobile and ubiquitous information access*. No. 2954 in Lecture Notes in Computer Science. Heidelberg, Germany: Springer-Verlag.

Edmundson, H. (1969). New methods in automatic abstracting. *Communications of the ACM, 16*(2), 264–285.

Harman, D. (1992). Relevance feedback and other query modification techniques. In W. B. Frakes, & R. Baeza-Yates (Eds.), *Information retrieval: data structures & algorithms* (pp. 241–263).

Jameson, A., Schfer, R., Weis, T., Berthold, A., & Weyrath, T. (1998). Making Systems Sensitive to the User's Time and Working Memory Constraints. *Proceedings of the 4th International Conference on Intelligent User Interfaces* (pp. 79–86). Los Angeles, California: ACM Press: New York.

Jansen, B., Spink, A., & Saracevic, T. (2000). 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.

Jones, M., Buchanan, G., & Mohd-Nasir, N. (1999a). Evaluation of WebTwig - a site outliner for handheld Web access. *Proceedings International Symposium on Handheld and Ubiquitous Computing*, Vol. 1707 (pp. 343–345). Lecture Notes in Computer Science, Karlsrhue: Springer, Berlin.

Jones, M., Buchanan, G., & Thimbleby, H. (2002). Sorting Out Searching on Small Screen Devices. *Proceedings of the 4th International Symposium on Mobile HCI* (pp. 81–94). Springer.

Jones, M., Marsden, G., Mohd-Nasir, N., & Boone, K. (1999b). Improving Web Interaction on Small Displays. *Proceedings of 8th WWW Conference*. Toronto, Canada.

Jones, S., Jones, M., & Deo, S. (2004). Using keyphrases as search result surrogates on small screen devices. *Personal Ubiquitous Computing*, *8*(1), 55–68.

Kirsch, S. (1998). Infoseek's experiences searching the internet. *SIGIR Forum*, *32*(2), 3–7.

Loudon, G., Sacher, H., & Kew, L. (2002). Design Issues for Mobile Information Retrieval. *Proceedings of Workshop on Mobile Personal Information Retrieval* (pp. 3–14). ACM SIGIR 2002, Tampere, Finland.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal*, 159–165.

MacKay, B., & Watters, C. (2003). The Impact of Migration of Data to Small Screens on Navigation. *IT & Society*, *1*(3), 90–101.

Paice, C. (1990). Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, *26*(1), 171–186.

Radev, D., & Fan, W. (2000). Automatic summarization of search engine hit lists. *ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval* (pp. 99–109). Hong Kong.

Radev, D., Fan, W., & Zhang, Z. (2001). WebInEssence: A Personalized Web-Based Multi-Document Summarization and Recommendation System. *NAACL 2001 Workshop on Automatic Summarization*. Pittsburgh.

Rush, J., Salvador, R., & Zamora, A. (1971). 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.

Saracevic, T. (1969). Comparative effects of titles, abstracts and full texts on relevance judgements. *Proceedings of the American Society for Information Science* (pp. 293–299).

Sweeney, S., Crestani, F., & Tombros, A. (2002). Mobile Delivery of News using Hierarchically Query-Biased Summaries. *Proceedings of ACM SAC 2002* (pp. 634–639). Madrid, Spain.

Tombros, A., & Crestani, F. (2000). Users's perception of relevance of spoken documents. *Journal of the American Society for Information Science*, *51*(9), 929–939.

Tombros, A., & Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. *Proceedings of ACM SIGIR* (pp. 2–10). Melbourne, Australia.

van Rijsbergen, C. (1979). *Information retrieval*. London: Butterworths, second edition edition.

Voorhees, E. (2002). Overview of TREC 2001. *Proceedings of the 11th TREC Conference*. Gaithersburg, MD, USA.

Weiser, M. (1991). The computer for the 21st century. *Scientific American*, *265*(3), 94–104.

Weiser, M. (1993). Hot topic: Ubiquitous computing. *IEEE Computer*, 71–72.

Wise, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., & Schur, A. (1999). Visualizing the non-visual: Spatial analysis and interaction with information from test documents. In R. Baeza-Yates, & B. Ribeiro-Neto (Eds.), *Modern Information Retrieval* (pp. 272–274). Addison-Wesley.

Yang, C., & Wang., F. (2003). Automatic Summarization for Financial News Delivery on Mobile Devices. *The Twelfth International World Wide Web Conference*. Budapest, Hungary.

Appendix

Post Experiment Questionnaire

The following questionnaire was completed by users after carrying the 5 queries. The questionnaire was completed in paper format. The response to each question was on a Likert-type scale as follows:

Q1. How do you rate the complexity of the queries in general?

1 = very easy

2 = easy

3 = neither complex nor easy

4 = complex

5 = very complex

Q2. How much do you think the summaries (any, including title) helped you in your judgements?

1 = very unhelpful

2 = unhelpful

3 = neither helpful nor unhelpful

4 = helpful

5 = very helpful

Q3. Mark how useful the summaries were for making your judgements (rating each level of summary)? (Q3.1 - Title, Q3.2 - 7%, Q3.3 - 15%, Q3.4 - 30%)

1 = very ineffective

2 = ineffective

3 = neither useful nor ineffective

4 = useful

5 = very useful

Finally, an open ended question for further comments was asked of the participants. This was often a basis for a brief discussion.

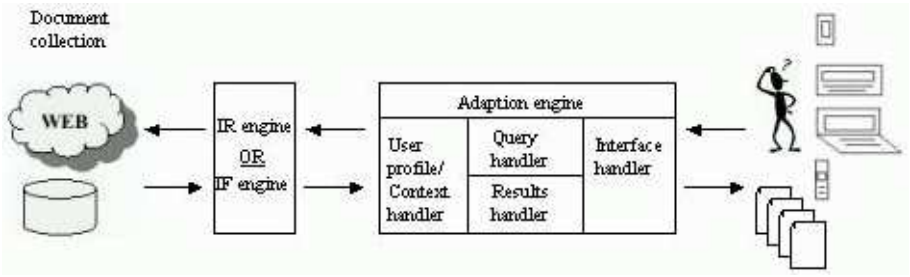


Figure 1: Architecture of a Personalised and Context-aware System.

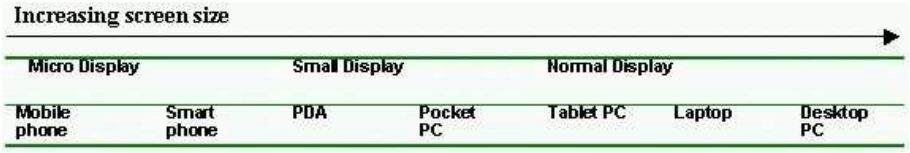
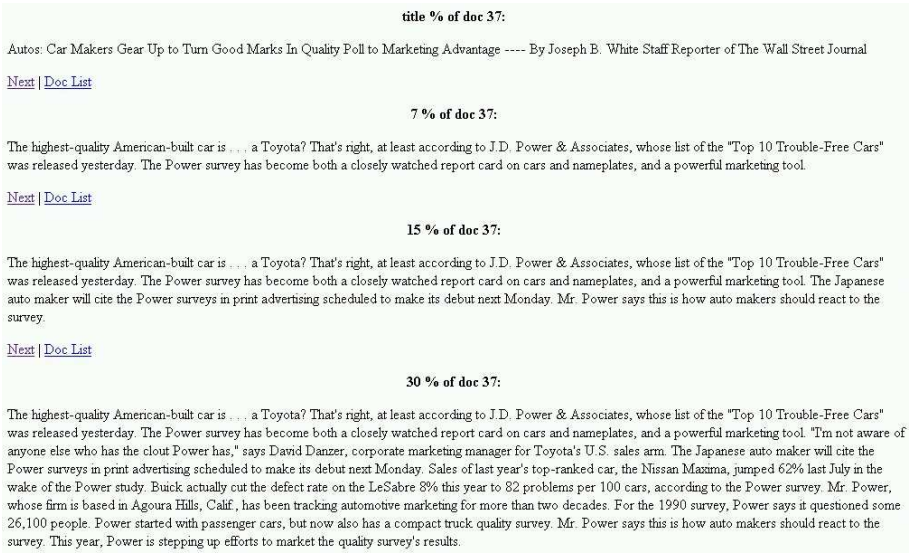


Figure 2: Range of device displays.

Figure 3: Illustration of the mobile devices used for the experiments (left: SPV E200 Smartphone, right: Handspring Visor PDA).

	Smartphone SPV E200	PDA Visor	Laptop Acer TravelMate 520 series
Type	Colour TFT LCD	Black & white	Colour TFT Active Matrix
Dimension (l x b)	1.69" x 1.37" 4.3cm x 3.5cm	3.18" x 2.38" 8.1cm x 6.0cm	11.25" x 8.5" 28.5cm x 21.5cm
Resolution	176 x 220	160 x 160	1024 x 768
Colour depth	16-bit 65k colour	2-bit greyscale	24-bit (16.7 million colours)

Table 1: Device displays for the experiment.



title % of doc 37:

Autos: Car Makers Gear Up to Turn Good Marks In Quality Poll to Marketing Advantage ---- By Joseph B. White Staff Reporter of The Wall Street Journal

Next | Doc List

7 % of doc 37:

The highest-quality American-built car is . . . a Toyota? That's right, at least according to J.D. Power & Associates, whose list of the "Top 10 Trouble-Free Cars" was released yesterday. The Power survey has become both a closely watched report card on cars and nameplates, and a powerful marketing tool.

Next | Doc List

15 % of doc 37:

The highest-quality American-built car is . . . a Toyota? That's right, at least according to J.D. Power & Associates, whose list of the "Top 10 Trouble-Free Cars" was released yesterday. The Power survey has become both a closely watched report card on cars and nameplates, and a powerful marketing tool. The Japanese auto maker will cite the Power surveys in print advertising scheduled to make its debut next Monday. Mr. Power says this is how auto makers should react to the survey.

Next | Doc List

30 % of doc 37:

The highest-quality American-built car is . . . a Toyota? That's right, at least according to J.D. Power & Associates, whose list of the "Top 10 Trouble-Free Cars" was released yesterday. The Power survey has become both a closely watched report card on cars and nameplates, and a powerful marketing tool. "I'm not aware of anyone else who has the clout Power has," says David Danzer, corporate marketing manager for Toyota's U.S. sales arm. The Japanese auto maker will cite the Power surveys in print advertising scheduled to make its debut next Monday. Sales of last year's top-ranked car, the Nissan Maxima, jumped 62% last July in the wake of the Power study. Buick actually cut the defect rate on the LeSabre 8% this year to 82 problems per 100 cars, according to the Power survey. Mr. Power, whose firm is based in Agoura Hills, Calif., has been tracking automotive marketing for more than two decades. For the 1990 survey, Power says it questioned some 26,100 people. Power started with passenger cars, but now also has a compact truck quality survey. Mr. Power says this is how auto makers should react to the survey. This year, Power is stepping up efforts to market the quality survey's results.

Figure 4: Sample summaries used in the experiment for single document (note, the layout here is for presentation purposes and was not the format used in the experiment).

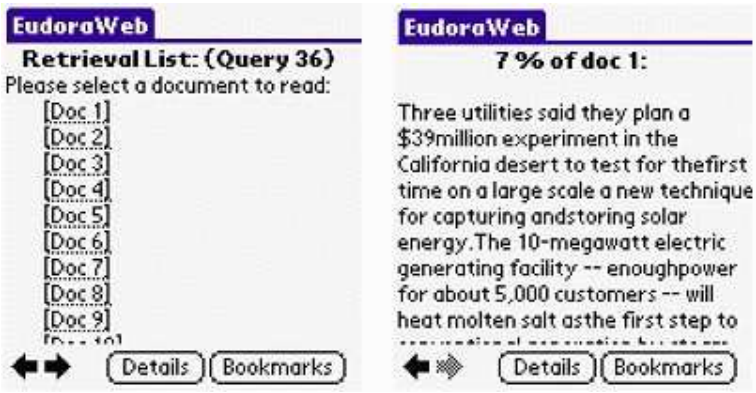


Figure 5: Examples of screen shots (for PDA).

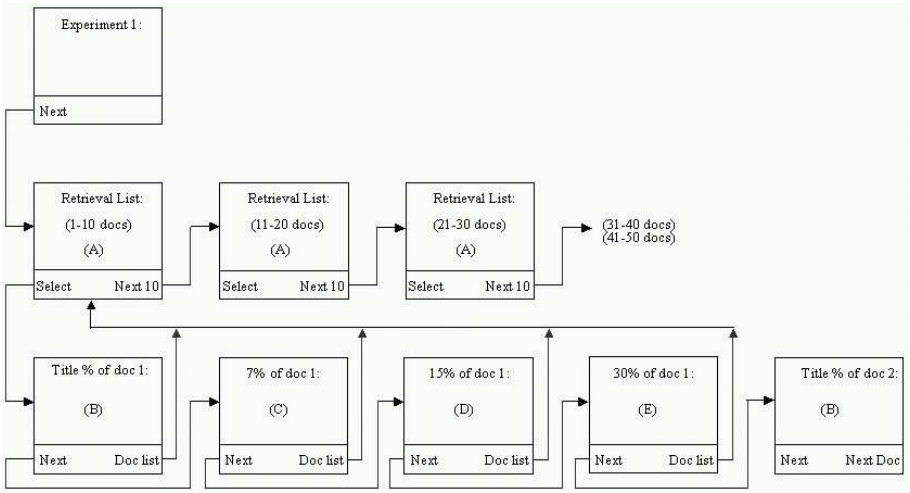


Figure 6: Illustration of the experimental arrangement.

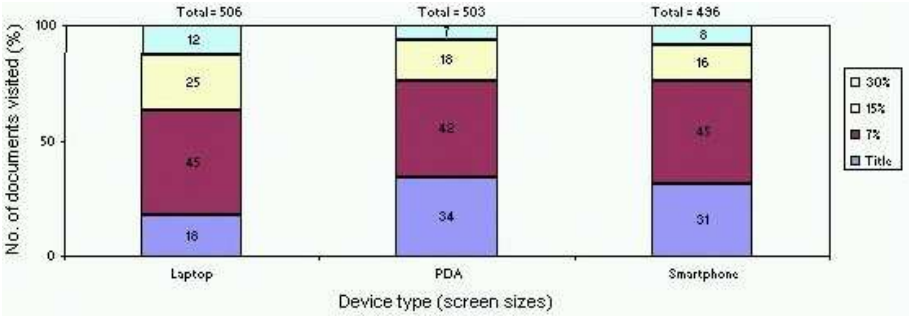


Figure 7: Number of documents at the different levels of summary that users utilised to make decisions.

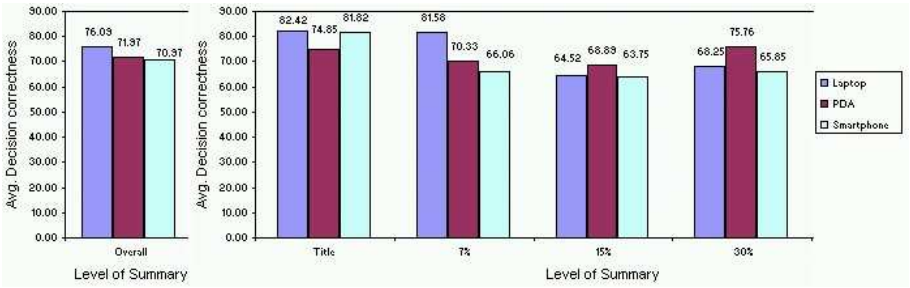


Figure 8: Average decision correctness for the experiments.

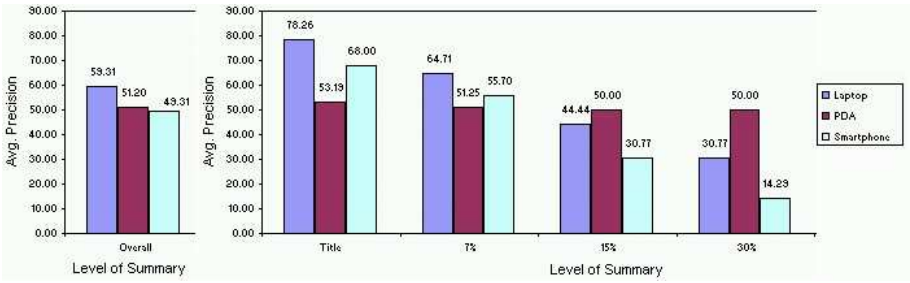


Figure 9: Average precision for the experiments.

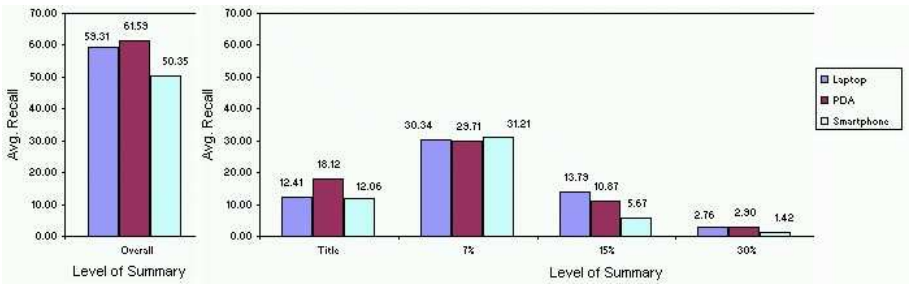


Figure 10: Average recall for the experiments.

	User										
	1	2	3	4	5	6	7	8	9	10	Avg.
Q1	3.0	3.7	3.7	3.8	4.0	4.7	4.0	3.7	3.0	4.0	3.8
Q2	3.0	3.0	2.7	3.2	2.8	3.3	2.7	2.0	3.7	4.7	3.1
Q3.1	2.3	3.3	3.0	3.8	2.7	3.0	3.3	2.7	3.0	4.7	3.2
Q3.2	3.3	3.7	3.7	4.7	3.0	4.3	3.0	3.7	3.7	2.7	3.6
Q3.3	3.0	2.0	3.3	4.0	3.0	3.7	3.3	3.0	3.7	1.7	3.1
Q3.4	3.0	1.0	3.0	2.0	4.0	4.0	3.0	3.0	2.0	1.0	2.6

Table 2: Questionnaire responses averaged over all experiments.

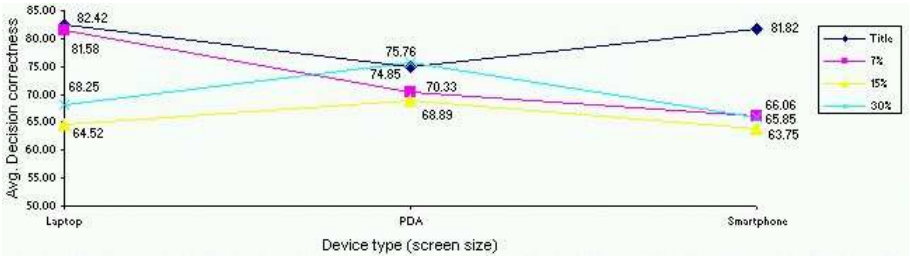


Figure 11: Performance of users to make correct decisions.

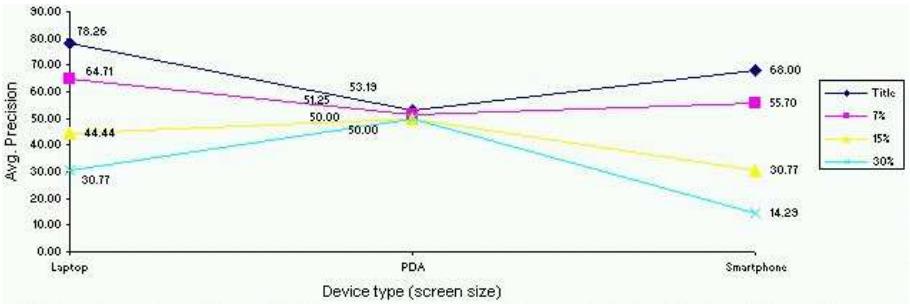


Figure 12: Performance of users to make precise decisions.