Working Paper Series
ISSN 1170-487X

Using Keyphrases as Search Result Surrogates on Small Screen Devices

Steve Jones, Matt Jones and Shaleen Deo

Working Paper: 07/03
September 2003

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

Using Keyphrases as Search Result Surrogates on Small Screen Devices

STEVE JONES[1], MATT JONES, SHALEEN DEO[2]

Department of Computer Science

University of Waikato

Private Bag 3105

Hamilton

New Zealand

{stevej, m.jones}@cs.waikato.ac.nz

Tel: +64 7 838 4021

Fax: +64 7 858 4095

This paper investigates user interpretation of search result displays on small screen devices. Such devices present interesting design challenges given their limited display capabilities, particularly in relation to screen size. Our aim is to provide users with succinct yet useful representations of search results that allow rapid and accurate decisions to be made about the utility of result documents, yet minimize user actions (such as scrolling), the use of device resources, and the volume of data to be downloaded. Our hypothesis is that keyphrases that are automatically extracted from documents can support this aim. We report on a user study that compared how accurately users categorized result documents on small screens when the document surrogates consisted of either keyphrases only, or document titles. We found no significant performance differences between the two conditions. In addition to these encouraging results, keyphrases have the benefit that they can be extracted and presented when no other document metadata can be identified.

Keywords: *Small screen devices, Searching, Usability evaluation, Keyphrase extraction.*

---

[1] Author for correspondence.

[2] Now with DataCom Ltd, Hamilton, New Zealand.

1. Introduction

David Beckham, Michael Schumacher and Rubens Barrichello seem to have no difficulty in mastering their small screen devices—as portrayed in a recent major mobile telephony marketing campaign. For most users, however, the mobile interactive experience is less successful. Small screens, limited input-output capabilities and expensive connection charges lead to frustrating, confusing and costly interactions.

Table 1 compares the display characteristics of a number of mobile devices: a PDA, a Pocket PC, a mobile communicator (telephone/PDA) and a mobile telephone. The characteristics of a laptop computer and desktop screen are also shown.

Devices A, B and C are, at the time of writing, some of the most sophisticated consumer handheld devices available. Obviously, a strong design consideration with such devices is portability and consequently display size is substantially restricted in comparison to laptop computers or desktop monitors. In fact these three devices offer only 6-7% of the display area of device E and 3-4% of that of device F. Device D is even more restricted, with 2% and 1% of the display areas of device E and F respectively.

In spite of these restrictions such devices commonly support standard desktop productivity tools such as word processors, spreadsheets and presentation designers. They also support Web access via wireless network capabilities using WAP or in some instances (e.g., devices B and C) fully-fledged HTML Web browsing software. The distinction between the tasks supported by these devices and those by laptop or desktop computers is blurred, particularly when versions of standard operating systems and attendant user interfaces are tailored to handheld devices (e.g., Microsoft Pocket PC on device B). In reality, however, common user activities, such as searching the Web, result in quite different user experiences on small screen devices.

Figure 1 compares Google search results displays on devices B and F (1920x1440 resolution). In the image on the left we see that only two result document details (document *surrogates*) can be fully displayed using the Microsoft Internet Explorer 'shrink to fit' option. Document related navigational links, such as 'Cached' and 'Similar pages', have also been replaced by small icons.

Additionally, the high-level navigational tools leading to other Google indexes, search preferences and so on are not immediately available. Not only is the full set of navigational links available on a desktop display (right-hand image of Figure 1), but five times as many document surrogates can be viewed on a single screen without the need for scrolling. As with many other search engines Google selects and displays a brief document extract for each result item, yet this information is substantially truncated in the small screen display, restricting the information provided to the user about each item.

1.1 User issues

Limited display area impacts on the effort required by users in their interaction with software on handheld devices and can reduce their ability to complete search-type tasks. In the search result example (Figure 1) on the small screen users must scroll through the Web page to consider half of the results available on the desktop screen with no scrolling. In addition, a further result page navigation is required to consider all of the results shown on the desktop screen. Interestingly, a finding that users rarely view more than the first page of results (Jansen et al., 2000), appears to hold for small screen devices (Jones et al., 2002). In the handheld case, this means that users may only consider half of the possibilities shown on the desktop display.

On small screen devices, there can be a very high interaction cost associated with exploring Web pages (Jones et al., 2002; Jones et al., 1999b). As Figure 2 illustrates, users can quickly become lost and disorientated as they attempt to make sense of content designed for a large display through the peephole of a mobile phone or handheld computer (Figure 2).

Users of wireless handheld devices often incur further costs both in monetary terms and response time. Wireless data transfer rates (via mechanisms such as Bluetooth or mobile telephony) are generally slower than those available on networked desktop computers. Response times to data requests are longer, and unproductive user 'wait time' increases. Wireless data transfer charges are normally also higher than non-wireless charges. For example, a leading service provider in New Zealand charges $75 (local currency) for 50 megabytes on a

mobile data plan, and $49 (local currency) per *500* megabytes for residential fixed-line customers.

A design aim is, therefore, to reduce user effort and costs. In the context of a task where a user is trying to satisfy an information need, this implies support for user identification of relevant documents as efficiently as possible. The information presented to users about documents should support immediate and accurate relevance judgements to minimize consideration of irrelevant documents and the consequent effort, time and financial costs.

Although we have exemplified the issues with a Web-searching task, the problems also arise when users search for information on their own devices. Some problems are ameliorated in this context, as the monetary cost of data transfer is not an issue, and search response time should be substantially reduced. However, the over-riding requirement to identify items of interest accurately and efficiently still applies. Some additional cues to aid users in this task will be present, such as filenames, storage locations and creation dates, yet as users' file stores increase in size and complexity, these cues become less helpful.

1.2 Our approach

Our aim is to support users in their Internet and personal data search tasks when using a small-screen device. The primary requirement is to enable users to make rapid and accurate assessments regarding the utility of result documents within the constraints of limited display space.

Conventional surrogate displays, of which Google is representative, contain at least document titles. Unfortunately the provision of title surrogates and their utility is dependent upon their creation by the document or Web-page authors. In many cases this information is missing, uninformative or cannot be extracted from the source documents. Extensions include topic categories, URL, and short text segments that contain query terms. However, these items offer little information about the topics covered by the document. Document extracts are query term related and give no indication as to whether the document addresses the query topic in a substantial way.

There is a need, then, for concise surrogates that can be provided when other metadata is unavailable or inappropriate, with minimal effort, and which properly represent the range of topics focussed on by a given document. Our approach is to

replace conventional surrogates with sets of keyphrases that have been automatically extracted from document text. We believe that they hold promise in supporting fast and accurate user judgments about result documents.

The remainder of this paper considers the use of keyphrases in information seeking tasks and report on our study of the efficacy of keyphrases as document surrogates.

1.3 Structure of the paper

In Section 2 we discuss related work, considering a number of textual and graphical representations of search results and their application in the small screen context. In Section 3 we report on how keyphrases have previously been used to support information seeking activities. We also describe the Kea system, which we used to identify document keyphrases for our study, and establish that the quality of Kea phrases is suitable for our purpose. Section 4 describes the methodology and procedures for our user study, with Sections 5 and 6 respectively reporting the results and discussing their implications. We then summarise our conclusions and propose further avenues for exploration that lead from this work.

2. Making Sense of Search Results

A well known human-computer-interaction phenomenon is the "QWERTY-isation" of keyboards. The QWERTY arrangement of keys was designed to avoid "crashes" of mechanical levers, with descending ones colliding with those just rising. The layout aimed to ensure that commonly co-located letters in text were positioned far enough away from each other to avoid such problems. With the arrival of digital word processors, there was no longer a technological need for QWERTY. However, despite research into more human-centred, easier-to-learn, and faster-to-use keyboard arrangements the majority of keyboards produced today still use QWERTY. Individual users and developers have become conditioned to the technology and it will be very difficult to shift to alternative designs.

A similar phenomenon seems to be occurring with search services – we might call it the "Googlisation" of search. When processing and memory power were much lower and more expensive than today, dynamic, flexible, graphically-rich search schemes were not viable. A "paper-based" approach, though, was possible, with

matches shown as paginated lists, each result represented by a series of surrogates such as the document title, location and some summary information.

Such simple schemes might, however, not best serve users needs. Hearst (1998), for instance, suggests services should allow users to gain an overview of the result set and to manipulate it to further understand the contents and to make sensible selections. Both of these user-centred qualities are poorly supported in conventional schemes.

There has been much research into richer, dynamic search schemes that are now viable on standard consumer computers (and soon will be on small screen devices). Again, however, it appears a technological inertia is setting in, with developers and users being reluctant to consider the alternatives to ranked, mainly textual result lists.

While commercial developers appear indisposed to explore alternative presentations for conventional, large-screen devices, they might well have to consider alternative and additional support for small screen users (Jones et al., 2002).

2.1 Richer presentations

Information visualisation is a well-established research area (Card et al., 1999). Much work has been put into the use of highly graphically sophisticated approaches to help the user make sense of large sets of information. Such graphical schemes have been applied to the fields of information retrieval and exploration in an attempt to overcome search problems on conventional displays. For instance, the Information Visualiser (Card et al., 1991) allows users to manipulate an animated 3-D categorical view of search results.

While many such schemes are radically different from the conventional ranked retrieval lists, other researchers have considered complementing textual presentations with additional information to help users discriminate better between result choices. For example, the TileBars system (Hearst, 1995) combines text and compact graphical surrogates. After the system retrieves documents, a graphical bar is shown next to each title in the results list showing how well the result item corresponds to each query term a user entered.

As Figure 3 illustrates, each retrieved document is represented as a rectangular bar, as shown on the left of the display, is subdivided into rows corresponding to

the query terms (in this case, two terms). The bar is also divided vertically to represent automatically parsed passages within the document. The distribution and frequency a term is used within a document is shown through the shading in of these "tilebars". The darkness of any tilebar gives an indication of the number of times that term occurs at that point in the document. The pattern thus created is meant to convey which of the query terms are most significant to the document, where they are occur and to quickly present those terms which are mentioned in passing.

In the small screen context, however, the appropriateness of many of the novel information-visualisation schemes is questionable. Even if the display technology can deliver the high resolution required, the available screen space is not necessarily adequate for meaningful presentations and manipulation. Adaptations of certain approaches may, though, be possible. For instance, the Starfield scheme (Ahlberg & Shneiderman, 1994) has been prototyped for the PalmPilot (Dunlop & Davidson, 2000).

Approaches that are not graphical, but provide a greater degree of freedom for user manipulation than the conventional schemes, have also been proposed for large screen devices. One such approach is the Scatter/Gather scheme explored in (Cutting et al., 1992). Here, similar documents are automatically clustered together and key term summaries can be displayed for each cluster. The aim is to enable users to gain an understanding of the topics available by scanning the cluster descriptions. Although small user studies indicate it may improve user effectiveness (Pirolli et al., 1996), there are problems including computational costs and difficulty in providing meaningful description of clusters (Chen & Dumais, 2000).

An alternative to clustering is categorisation. Instead of attempting to group results into generated groups, documents are assigned to an existing categorical structure. Many search engines (e.g. Yahoo!) already provide categorical browse access to their content. In Yahoo!, for instance, users can select from top-level categories such as Entertainment, Government and Health, and browse further sub-categories to help them gain an understanding of the sorts of material available. Category detail can also be shown for each search result—in Yahoo!, the hierarchical category information is given in addition to the document title, URL and text summary surrogates.

Chen and Dumais (2000) present a search system that automatically categorises Web search results into an existing hierarchical category structure. In an attempt to make good use of the (large) screen space, the system initially presents the all-important first result page in a way that gives an overview of how the results are distributed across the categorical structure. Only a top-level category view is shown, and just the top twenty results are given. The user can expand and contract categories to see more of the hierarchical structure and how the results relate to it, and additional matches can also be displayed on demand. A user-study showed that users not only liked the new approach but that they were 50% faster at finding information than in the ranked-list scheme.

2.2 Schemes for small screens

Approaches that group results (such as categorization and clustering) seem potentially valuable for the small screen context as a significant amount of information about query results can be displayed in a small space without the need for graphical sophistication.

The WebTwig browser for small screens (Jones et al., 1999a) gives users an hierarchical outline view of a Web site. They can expand and contract portions of the tree display to view more or less detail about the information structure and content. User evaluations of the system suggest benefits of the approach and the PowerBrowser investigations amplify the findings (Buyukkokton et al., 2000b). To extend the approach to Web searching, we developed LibTwig (Jones et al., 2002) that presents search results relative to the hierarchical Web site structure (see Figure 4). The rationale is that the outline view not only limits the amount of scrolling required to make sense of the search results, but provides context information which should help users to make decisions about which alternatives to pursue. Initial pilot study user evaluations are encouraging (Buchanan et al., 2002).

The PowerBrowser uses a similar approach to WebTwig for browsing, but a very different type of overview scheme for cross-site searching (Buyukkokton et al., 2000a). With each new keyword entered, the number of Web pages matching the search is updated and shown to the user. Individual page details are only shown when the user feels the number of pages in the retrieval set is small enough to deal with on the small screen.

3. Keyphrases and information seeking

Document metadata (information about documents) has long been used to support information seeking. Conventional libraries devote considerable resources to the accurate cataloguing of their holdings, associating metadata—such as title, author, date of publication, subject descriptors, classification labels, keywords and so on—with items to ease organization and management of holdings, and support end user access to them. By comparison, descriptive information about documents indexed by search engines, held in digital libraries or provided by other electronic sources is sparse, even though the volume of documents accessible by a single service is often many times that of physical libraries. For example Google claims to index more than 3 billion documents (Web pages).

These services commonly return query result sets containing hundreds or thousands of documents, making it infeasible for users to examine each complete document to determine whether or not it might be useful. Instead, metadata-based document surrogates such as titles, bibliographic information, extracts containing query terms and summaries help users to identify documents of interest.

Some types of document contain a list of keywords specified by the author. These keywords and keyphrases—we use the latter term to subsume the former—are a particularly useful type of summary information. They condense documents into a few words and phrases, offering a brief and precise description of their content.

Recent research work has investigated the potential of keyphrases to support information seeking in a number of ways, including the classification or clustering of documents (Jones & Mahoui, 2000; Zamir & Etzioni, 1999), search and browsing interfaces (Gutwin et al., 1999; Jones & Paynter, 1999), retrieval engines (Arampatzis et al., 1998; Jones & Staveley, 1999), and thesaurus construction (Kosovac et al., 2000; Paynter et al., 2000).

A number of other systems exploit phrases to enhance user interaction. The Journal of Artificial Intelligence Research (http://extractor.iit.nrc.ca/jair/keyphrases/) can be accessed through an interface that is based around keyphrases produced by Extractor (Turney, 2000). Larkey (1999) describes a system for searching a database of patent information. Within

the system phrases are used to suggest query expansions to users based on the search terms that have been specified. Similarly, Pedersen et al (1991) use phrases to support query reformulation in their Snippet Search system. Krulwich and Burkey (1997) exploit heuristically extracted phrases to inform InfoFinder, an 'intelligent' agent that learns user interests during access to on-line documents.

The Stanford Digital Library group has carried out some investigations into the use of extracted keyphrases for browsing Web pages on the small screen (Buyukkokton et al., 2001). In this work, the use of such meta-data was reported to reduce the time to complete information seeking tasks by a factor of 3 to 4 relative to when the user had to contend with the full Web page.

3.1 Kea

We believe that document keyphrase sets can form useful document surrogates to aid users in determining the relevance of query result items. In particular we hypothesise that they have a promising application on small screen devices, given that they concisely reflect document content. In some cases documents will have been provided with keyphrases by their authors or professional cataloguers, although this is relatively rare. Manual creation of keyphrases for large collections of documents is clearly infeasible due to cost and time constraints, leading to a need for automatic keyphrase identification. A number of automated systems have been developed for this purpose such as Extractor (Turney, 2000) and Kea (Witten et al., 1999).

For the purposes of our study we chose to use the Kea system.

3.1.1 Kea keyphrase extraction process

Kea (www.nzdl.org/Kea) identifies keyphrases in a two-step process. The first step is to build a model reflecting characteristics of desirable keyphrases. This is achieved by providing a set of training documents that have keyphrases associated with them (either by authors or some other authoritative source). Kea extracts all potential phrases from a training document and calculates three attributes for each phrase: whether it is in the predefined list of phrases for the document, how far into the document it first occurs, and how specific it is to the document. These

candidate phrases are combined into one dataset from which a Naïve Bayes classifier is built.

The model can then be applied to the remaining (non-training) documents. Candidate phrases are extracted, their distance and specificity attributes are computed, and then used by the classifier to calculate the probability that a candidate is suitable as a keyphrase. Candidates are then output in ranked probability order.

When suitable training data is not available an existing general-purpose keyphrase model can be applied. One such model is *aliweb*, which was derived from a Web-page corpus created by Turney (2000). Topics within the corpus are varied and include micro-breweries, law libraries, text processing and university departments. This is the model that we have used to extract keyphrases for use in our system and evaluation.

3.1.2 Quality of Kea keyphrases

One factor affecting the efficacy of keyphrase search result surrogates is the quality of the keyphrases themselves. Previous research into the quality of Kea keyphrases has adopted a number of approaches, including precision and recall comparisons against author keyphrases, subjective evaluation of individual keyphrases and subjective evaluation of keyphrase sets. These results have been positive. Witten et al (1999) report that about 11% of Kea keyphrases matched those specified by authors. Although this figure seems rather low, the result is ameliorated by two factors. First, Kea is unable to identify author phrases that do not appear in a document's text (which is often the case). Second, this figure does not account for the fact that many more Kea phrases were generated for each document than had been specified by authors.

Perhaps more useful are the subjective evaluations of Kea. Jones and Paynter (2001) found that the *aliweb* model extracted keyphrases that were judged to be as good as those provided by authors, with 80% deemed representative of the document from which they were extracted. Also, Kea ranked phrases appropriately, assigning appropriate relative scores, so that users could be confident that the first N phrases would be the best N phrases. However, this study investigated the quality of individual phrases. A further study reported by Jones and Paynter (2002) establishes the quality of Kea phrases when presented as

sets, as is the aim in our study. Overall, *aliweb* phrase sets were viewed positively by assessors, and no significant difference was established between the quality of *aliweb* phrase sets and author-specified phrase sets.

4. User study

We designed and administered a user study in order to measure the utility of keyphrases as result document surrogates in a representative user task. The study also investigated the comparative performance of surrogates based on document titles in the same task. We were interested in the impact on accuracy and task completion times of using keyphrases versus titles.

4.1 The task context

The evaluation of the two types of surrogates took place in the context of a document categorization task. A subject's task was to categorize a set of query result documents—*based solely on their surrogates*—according to a pre-defined category hierarchy.

A categorization task has a number of desirable characteristics for our purposes:

- subjects are required to reason about the overall content of a document to determine its most suitable categorization;
- inter-document comparisons are necessary, in order to determine which documents should be similarly or differently categorized;
- predefined document categories allow subject accuracy to be measured by the similarity between subject responses and predefined categories;

4.2 Experimental conditions

The experiment employed a between-groups design involving sixteen subjects split randomly into two groups of eight participants. There were two experimental conditions:

- Condition 1: The document surrogates contained only document titles
- Condition 2: The document surrogates contained only document keyphrases

The first group carried out tasks under condition 1, and the second group carried out tasks under condition 2.

4.3 Subjects

All sixteen subjects were either undergraduate or postgraduate Computer Science students at an English speaking university. Ten were male and six female, and four students did not have English as their first language (although they had met all English language requirements for university study). All of the subjects had substantial prior experience of Internet searching, using Web-based search engines every working day. All but one subject had prior experience of using a small screen device such as a mobile telephone, and six had previously used a PDA such as a Palm Pilot.

4.4 Materials

A collection of 45 Web page documents was sourced via the Yahoo! search engine (www.yahoo.com). Web sites indexed by Yahoo! are categorized by editors according to a predefined category hierarchy, which contains 14 major categories at the top level, such as 'Arts & Humanities', 'Business & Economy' and so on. Each document in the experiment collection was described by a path through the hierarchy such as Entertainment->Movies and Film->Theory and Criticism, where 'Entertainment' is the most general level and 'Theory and Criticism' is the most specific. All documents belonged to one of three general interest top level categories—entertainment, government and health—with a third (15) belonging to each. For each category we selected documents so that we had a set that was drawn from the range of subcategories. In examining potential documents we checked they were of a similar length (one thousand words) and had substantial content rather than simply being "homepages" or pages of links etc. For each set of candidate documents, we selected the first ones that met our criteria in order to avoid an experimenter bias towards good or poor titles.

The titles were extracted from each document by hand, and keyphrases were automatically extracted from each document using the Kea system. Five keyphrases were extracted from each document.

Two pseudo-result lists of the same 15 documents were prepared for each top-level category. One list contained keyphrase-only surrogates, and the other

contained title-only surrogates. Each list contained the 15 documents assigned to the category by Yahoo!, plus an additional 3 documents that were unrelated to the category. The unrelated documents were included to assess the effect of the two conditions on subjects' ability to make coarse (essentially binary) judgements about the relevance of a document to a category.

Surrogate lists were presented to subjects in Microsoft Internet Explorer on a Compaq iPAQ H3870 running Pocket PC 2002. The display size of this device is 240x320 pixels, with a physical size of 2.26 inches wide by 3.02 inches tall. Each surrogate was labelled alphabetically to make reference to specific items easier for subjects. Sample presentations for the same set of result documents are show in Figure 5 (left-hand-side, titles) and (right-hand-side, keyphrases).

Representations of the three category hierarchies were prepared so that subjects could easily place surrogate labels at the desired location to indicate their categorization decisions. One of the hierarchy representations (for the entertainment category) is shown in Figure 6.

At the coarsest level, subjects made a binary decision about whether a particular surrogate could be categorized as, for instance, entertainment-related or not. If not, the surrogate label would be placed in the 'Unclassifiable' box. If a surrogate was determined to be entertainment-related, subjects would place its label at the most descriptive place in the entertainment hierarchy. Each surrogate could appear in the hierarchy only once.

4.5 Procedure

On arrival each subject was provided with, and instructed to consider, written instructions, a document describing their rights as an experimental participant and a consent form. Once consent was obtained their task was described verbally, replicating the written instructions.

Each subject undertook three tasks in one condition only. The nature of each task was identical. The subjects were instructed to consider a set of document surrogates and categorize each one as specifically as possible according to the provided category hierarchy. The three tasks differed only in the category under consideration—entertainment, government and health—and category order was randomized for each subject. Once completed, the materials for each task were

removed, and the materials for the next were provided. Unlimited time was available for the subjects to complete the tasks.

4.6 Data captured

The time taken for completion of each task by each subject was recorded. Timing commenced when a subject began a task and finished when all surrogates had been categorized to the subject's satisfaction. Each categorization hierarchy sheet (exemplified in Figure 6) was retained for analysis.

5. Results

In this section we present a detailed set of performance data under the two conditions.

5.1 Categorization accuracy

The effect of the two conditions on subjects' ability to accurately categorize the surrogates was analyzed. Accuracy was measured by how closely subject responses match the predefined Yahoo! categorization for a given document. Each categorization made by a subject was awarded a score of 0, 1, 2 or 3. A score of 0 was awarded if a document was placed in a category tree when its true category was 'unclassifiable' or vice versa. A score of 3 was awarded when a document was placed at the correct leaf node of the category tree or correctly identified as unclassifiable. Scores of 1 and 2 reflect the distance along the branch of the correct leaf node that the subject placed a document, 2 being awarded for selecting the parent of the correct leaf, and 1 for selecting the correct grandparent.

5.1.1 Overall accuracy

There were a total of 864 categorizations (16 subjects x 18 surrogates x 3 tasks), with half occurring under each of the two conditions. Of these, a total of 474 (55%) were completely correct—they exactly matched the predefined categorizations—with 234 (27%) observed under the keyphrase condition and 240 (28%) under the title condition.

Table 2 shows the accuracy data for subjects for the three categories combined. The percentage of categorisations that were completely correct is almost identical under both conditions (55.6% and 54.2%). The percentage of partially correct

categorizations is slightly higher (6%) for the keyphrase condition, for which there are also slightly fewer (4%) incorrect categorizations.

Table 3 shows accuracy with respect to conditions and categories. There was found to be a significant category effect within conditions. In both the titles and keyphrases conditions, entertainment documents were harder to categorize. In the keyphrases case, health documents were also easier to categorize than those from both of the entertainment and government categories.

5.1.2 Binary categorization accuracy

At the coarsest level we considered the subjects' ability to make binary relevance judgements about whether a document belonged to the top-level category under consideration or not. This data is summarised in Table 4. 144 of all categorizations (16 subjects x 3 surrogates x 3 categories) should have been assigned into 'unclassifiable' because 3 out of each set of 18 surrogates were unrelated to the category at hand. 105 (73%) of these assignments were correct—44 (31%) under the keyphrase condition and 61 (42%) under the title condition. 39 (27%) of these assignments were incorrect.

The remaining 720 categorizations (16 subjects x 15 surrogates x 3 categories) should have been made within the category under consideration and not assigned 'unclassifiable'. Of these, 624 (87%) were correctly placed somewhere within a category hierarchy (46% under the keyphrase condition, and 41% under the title condition), and 96 (13%) were not.

Therefore, 135 (16%) of *all* categorizations were incorrect by the coarsest measure, comprising 39 false-positives and 96 false-negatives. 96 categorizations (11%) were erroneous assignments to 'unclassifiable' (4% under the keyphrase condition and 7% under the title condition). For the remaining (erroneous) categorizations 3% occurred under the keyphrase condition and 2% under the title condition.

5.1.3 Overall accuracy

Each categorization was awarded a score of 0, 1, 2 or 3 to reflect its accuracy. The mean categorization score by condition and category is shown in Table 5. Higher

scores reflect more accurate categorizations, with a maximum of 3. For all categories combined there is little difference between the accuracy under the two conditions (2.08 for keyphrases, 2.03 for titles).

Under both conditions the category effect is the same—the entertainment category resulted in the least accurate categorizations and the health category resulted in the most accurate.

Using the Wilcoxon rank-sum test (accounting for ties), there is no significant difference between the accuracy under the two conditions for all categories combined ($p=0.787$), entertainment ($p=0.631$), government ($p=0.508$) or health ($p=0.087$).

5.1.4 Accuracy within category hierarchies

We considered more closely the accuracy with which surrogates were placed within category hierarchies. Therefore we examined the 624 categorizations that were correctly placed somewhere within a hierarchy (as opposed to incorrectly placed under 'unclassifiable'). Inaccuracies occured in these categorizations when the subjects selected an incorrect leaf node, or categorized a document at a less specific location in the hierarchy.

Table 6 shows this mean accuracy for both conditions across all categories and for each category individually. The means for the two conditions were almost identical—2.39 for the keyphrase condition and 2.45 for the titles condition. When considered by category we see that accuracy was again virtually identical for entertainment (2.22 keyphrases, 2.27 titles). Titles supported better accuracy than keyphrases in the government category, but not the health category, although with small differences. In the case of the government category, however, the difference was significant ($p=0.015$, Wilcoxon rank-sum). There was no significant difference in accuracy for all categories combined ($p=0.35$), entertainment ($p=0.736$) or health ($p=0.377$).

We calculated similar accuracy measures for the responses that placed a surrogate correctly within a category hierarchy but at the incorrect location. These measures indicate whether the erroneous placements were minor or extreme, and are shown in Table 7.

Overall, mis-categorizations were slightly less severe under the keyphrase condition (mean=1.49) than the titles condition (mean=1.46), although the difference was not significant (p=0.688, Wilcoxon rank-sum). The differences for each category were also not significant.

5.2 Task completion time

The overall mean times for individual task completion were calculated for both conditions and are shown in Table 8. The mean times for the two conditions are very similar, at just under 6 minutes for task completion, with subjects in the keyphrase condition taking 13 seconds longer on average.

As the large standard deviations suggest, the performance times varied significantly between the participants within each condition (p=0.040 for titles, p=0.009 for keyphrases). Between the conditions there was no significant difference between the times taken (Wilcoxon rank-sum, p=0.627).

The potential effect of different characteristics of the three category and document set pairings was considered for each condition. Table 9 shows the mean task completion times by category for the two conditions. Mean task completion times range between just over 5 minutes and almost 7 minutes. Mean times were lower under the titles condition for two of the categories: entertainment and health and lower under the keyphrase condition for the government category. However, by ANOVA, no significant category effect was identified at the 5% confidence level. There was also no significant difference (Wilcoxon rank-sum) between completion times under the two conditions for entertainment (p=0.328), government (p=0.424) or health (p=0.775).

6. Discussion

Before drawing some overall conclusions, the results have led us to consider two interesting "speculations" and these are discussed, here.

6.1 Category effects

The results in Table 3 suggest that while titles and keyphrases are useful and adequate surrogates for documents in the government and health categories, they are less supportive for the entertainment category. Looking at the documents in this category we saw that there were more examples than in the other categories of

titles that gave little hint of their nature: e.g., "Matthew and Jake's Adventures" could have been categorized under "Movies and Film", "Humour", and "Genres" and a number of sub-categories within these based on the title. Keyphrases might have fared poorly due to the particular importance of meta-textual information in entertainment type documents: for instance, it might be hard to distinguish a piece of irony from a thriller-style document simply on keywords.

The results in Table 3 and Table 4 suggest that keyphrases result in fewer high-level categorization errors (whether or not a document can be placed in a category hierarchy or not). For errors that did occur at this level, the keyphrase condition provided marginally more false-positives than false-negatives—the fact that a document was unrelated to the category was not evident. However, the difference is small. By contrast, the title condition produced four times the number of false-negative responses than it did false-positives—subjects clearly had some difficulty determining that documents were related to the category from their titles.

The difference in accuracy between conditions marginally favours titles (from Table 5) for both the entertainment and government categories, but more markedly favours keyphrases for the health category. This may be because titles of such documents can require domain knowledge while the keyphrases can present more context to the lay reader. For example, one document in our study had the title "Sarcoma" while the keyphrase list read, "tumor, Sarcoma, tissue, bone, cancer"

Both of these observations lead to a hypothesis that providing complementary surrogates (e.g. titles, keywords and categorizations) may allow users to make better overall use of search results.

6.2 Quicker judgements with keyphrases?

Table 10 provides details of the length of the surrogates presented to subjects on the handheld device (which on average, using a proportional font, supported 39 characters per line and 14.5 lines per screen). The data shows that keyphrase surrogates required just more than one and a half times the display space of title surrogates.

Table 11 shows surrogate length details broken down into each of the three categories considered by subjects. The small variations between categories are explained by differences of length of individual keyphrases (which where all one, two or three words in length) even though five keyphrases were extracted from each document.

Whereas the title surrogates required one and a half screens of text to be displayed, the keyphrases required an additional screen on average. We would expect this to have a noticeable impact on task completion times because of the increased navigation overhead imposed by additional scrolling. However, we note that overall, mean completion times were similar under the two conditions.

One possible explanation is that although there was additional scrolling time, this was offset by users making faster (and yet as accurate) judgements with keyphrases than they did with titles. If this is the case, then one implication is that result lists that do not require scrolling on the small screen will be assessed more quickly using keyphrases than titles. The amount of scrolling in the keyphrase case could be reduced by decreasing the number of keyphrases presented for each document; however, the speed-of-use gain that may follow would need to be traded-off against the potential drop in accuracy of judgements about a document's contents.

7. Conclusions and Future Work

Overall, we found no evidence to support the initial hypotheses. That is, we have to reject the notion that keyphrases used on their own provide for better categorization accuracy or support faster categorization than titles. The results though, do suggest that keyphrase are as good as titles; this is encouraging. While the provision of titles is dependent on authors, keyphrases can be automatically extracted. Our results indicate that where no title, or a poor title (including those that require domain knowledge), is provided by the author, keyphrase surrogates might well aid a user to make sense of the document.

Keyphrases as surrogates are also attractive as they can be used not only in the presentation of search results but also to assist the skim reading of target documents. Others, notably in the XLibris project (Schilit et al., 1999), have implemented keyphrase highlighting. In our ongoing work, we have implemented

a handheld computer prototype that allows the user to view documents with keyphrases, sentences and even paragraphs highlighted (see Figure 7). We are currently evaluating the usability impact of this scheme.

While we have explored the efficacy of keyphrases in the context of Web-based information, our findings are also of interest to those developing access mechanisms for local, personal information spaces. Conventional desktop and laptop computers already have vast long-term local storage capabilities (for example, a laptop computer used to write this paper has a hard disk capacities of 18Gb). Mobile small screen devices are also beginning to provide storage for large quantities of local information: one of the authors, for instance, has an HP IPaq with 1Gb local storage and a 30Gb Apple iPod.

It is more likely that the documents on these devices will have "poor" titles. For example, Figure 8 shows the contents of a folder from one of our workspaces. There are condensed (e.g. "uktalk1"), general (e.g. "call_for_papers"), coded (e.g. "2003_COMP245AE") and vague (e.g. "staying03") titles. Overt time, and in the context of thousands or tens of thousands of documents, the contents of these files might be hard to discern from the titles alone. Additional surrogates such as keyphrases, both while browsing available files and when using local search engines might be very helpful.

Although mobile small screen devices can be useful, everyday information appliances, their impoverished interfaces present major challenges to content providers. Without careful interaction design, their promise may be lost. Improving usability of these devices will lead to large-scale benefits. With a billion mobile phones and hundreds of millions handheld computers already in use, just a 10 second reduction in wasted user time, means over 4500 person years saved, each day.

Automatically extracted keyphrases seem to hold potential to improve the use of such devices and are certainly worthy of further investigation. They can provide small screen users with concise, succinct descriptions of documents, independently of a human intermediary. They have the further advantages of being useful in the full presentation of a document (for, say, skim reading) and in contexts where poor or no other surrogates are available.

Acknowledgements

We gratefully acknowledge Lyn Hunt's wise insights and practical statistical assistance. Helpful comments were provided by Paul Cairns. The New Zealand Digital Library is funded by the New Zealand New Economies Research Fund.

References

Ahlberg, C., & Shneiderman, B. (1994). Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays, Proceedings of CHI'94: Human Factors in Computing Systems (pp. 313-317, 479-480): ACM Press.

Arampatzis, A. T., Tsoris, T., Koster, C. H. A., & Van der Weide, T. P. (1998). Phrase-based information retrieval. Information Processing & Management., 34(6), 693-707.

Buchanan, G., Jones, M., & Marsden, G. (2002). Exploring Small Screen Digital Library Access with the Greestone Digital Library. In M. Agosti & C. Thanos (Eds.), Proceedings 6th European Conference on Research and Advanced Technology for Digital Libraries (Vol. 2458, pp. 583-596): Springer-Verlag.

Buyukkokton, O., Garcia-Molina, H., & Paepcke, A. (2000a). Focused Web Searching with PDAs. Computer Networks (Proceedings of the 9th International World Wide Web Conference), 33(1--6), 213-230.

Buyukkokton, O., Garcia-Molina, H., & Paepcke, A. (2001). Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices., Proceedings of the 10th International World Wide Web Conference (WWW10). Hong Kong.

Buyukkokton, O., Garcia-Molina, H., Paepcke, A., & Winograd, T. (2000b). Power Browser: Efficient Web Browsing for PDAs, Proceedings of CHI'00: Human Factors in Computing Systems (pp. 430-437): ACM Press.

Card, S., Mackinlay, J. D., & Shneiderman, B. (1999). Readings in Information Visualization - Using Vision to Think: Morgan Kaufmann.

Card, S. K., Robertson, G. G., & Mackinlay, J. D. (1991). The Information Visualizer, an information workspace, Proceedings of CHI'91: Human Factors in Computing Systems (pp. 181-186). New Orleans: ACM Press.

Chen, H., & Dumais, S. T. (2000). Bringing Order to the Web: Automatically Categorizing Search Results, Proceedings of CHI'00: Human Factors in Computing Systems (pp. 145-152): ACM Press.

Cutting, D. R., Karger, D., Pedersen, J. O., & Tukey, J. W. (1992). Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, Proceedings of SIGIR'92: the 15th International Conference on Research and Development in Information Retrieval (pp. 318-329): ACM Press.

Dunlop, M., & Davidson, N. (2000). Visual Information Seeking on PDAtop Devices, Proceedings of BCS Human Computer Interaction 2000 (Vol. 2, pp. 19-20).

Gutwin, C., Paynter, G. W., Witten, I. H., Nevill-Manning, C., & Frank, E. (1999). Improving Browsing in Digital Libraries with Keyphrase Indexes. Journal of Decision Support Systems, 27(1-2), 81-104.

Hearst, M. (1995). TileBars: Visualization of Term Distribution Information in Full Text Information Access, Proceedings of CHI'95: Human Factors in Computing Systems (pp. 59-66). Denver, Colorado, USA: ACM Press.

Hearst, M. (1998). User Interfaces and Visualization. In Baeza-Yates & Biberio-Neto (Eds.), Modern Information Retrieval: Addison-Wesley Longman Publishing Company.

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 & Management, 36(2), 207-227.

Jones, M., Buchanan, G., & Mohd-Nasir, N. (1999a). Evaluation of WebTwig — a Site Outliner for Handheld Web Access. Proceedings of the International Symposium on Handheld and Ubiquitous Computing. Lecture Notes in Computer Science., 1707, 343-345.

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 the 8th World Wide Web Conference.

Jones, S., & Mahoui, M. (2000). Hierarchical Document Clustering Using Automatically Extracted Keyphrases, Proceedings of the Third International Asian Conference on Digital Libraries (pp. 113-120). Seoul, Korea.

Jones, S., & Paynter, G. (1999). Topic-based Browsing Within a Digital Library Using Keyphrases. In E. Fox & N. Rowe (Eds.), Proceedings of Digital Libraries'99: The Fourth ACM Conference on Digital Libraries (pp. 114-121). Berkeley, CA: ACM Press.

Jones, S., & Paynter, G. W. (2001). Human Evaluation of Kea, an Automatic Keyphrasing System, Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 148-156). Roanoke, VA, USA: ACM Press.

Jones, S., & Paynter, G. W. (2002). Automatic Extraction of Document Keyphrases for Use in Digital Libraries: Evaluation and Applications. Journal of the American Society for Information Science and Technology, 53(8), 653-677.

Jones, S., & Staveley, M. (1999). Phrasier: a System for Interactive Document Retrieval Using Keyphrases. In M. Hearst & F. Gey & R. Tong (Eds.), Proceedings of SIGIR'99: the 22nd International Conference on Research and Development in Information Retrieval (pp. 160-167). Berkeley, CA: ACM Press.

Kosovac, B., Vanier, D. J., & Froese, T. M. (2000). Use of Keyphrase Extraction Software for Creation of an AEC/FM Thesaurus. Electronic Journal of Information Technology in Construction, 5, 25-36.

Krulwich, B., & Burkey, C. (1997). The Infofinder Agent - Learning User Interests Through Heuristic Phrase Extraction. IEEE Intelligent Systems & Their Applications, 12(5), 22-27.

Larkey, L. S. (1999). A Patent Search and Classification System, Proceedings of Digital Libraries'99: The Fourth ACM Conference on Digital Libraries (pp. 179-187). Berkeley, CA: ACM Press.

Paynter, G. W., Witten, I. H., & Cunningham, S. J. (2000). Evaluating Extracted Phrases and Extending Thesauri, Proceedings of the Third International Conference on Asian Digital Libraries (pp. 131-138). Seoul, Korea.

Pedersen, J., Cutting, D., & Tukey, J. (1991). Snippet Search: a Single Phrase Approach to Text Access, Proceedings of the 1991 Joint Statistical Meetings: American Statistical Association.

Pirolli, P., Schank, P., Hearst, M., & Diehl, C. (1996). Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection, Proceedings of CHI'96: Human Factors in Computing Systems (pp. 213-220): ACM Press.

Schilit, B. N., Price, M. N., Golovchinsky, G., Tanaka, K., & Marshall, C. C. (1999). The Reading Appliance Revolution. IEEE Computer, 32(1), 65-.

Turney, P. D. (2000). Learning Algorithms for Keyphrase Extraction. Information Retrieval, 2(4), 303-336.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). KEA: Practical Automatic Keyphrase Extraction. In E. A. Fox & N. Rowe (Eds.), Proceedings of Digital Libraries '99: The Fourth ACM Conference on Digital Libraries (pp. 254-255). Berkeley, CA: ACM Press.

Zamir, O., & Etzioni, O. (1999). Grouper: A Dynamic Clustering Interface to Web Search Results. Computer Networks and ISDN Systems, 31(11-16), 1361-1374.

			Display characteristics		
			Resolution (pixels)	Viewable dimensions (inches)	Colours
Device	A	Palm Tungsten T PDA	320x320	2.3 width 2.3 height	16-bit (65,536 colours)
	B	Compaq iPAQ 5400 series Pocket PC	240x320	2.26 width 3.02 height	16-bit (65,536 colours)
	C	Nokia Communicator 9290 Mobile telephone/PDA	240x320	4.3 width 1.4 height	12-bit (4096 colours)
	D	Nokia 6800 Mobile telephone	130x130	(approx) 1.2 width 1.2 height	12-bit (4096 colours)
	E	Apple Titanium Powerbook 15"	1280x854 (max)	(approx) 12 width 8 height 15.2 diagonal	Millions of colours
	F	Philips 202P monitor	2048x1536 (max)	16 width 12 height	Millions of colours

Table 1: Comparison of display characteristics of mobile and conventional devices

	exactly correct	partially correct	completely incorrect
keyphrases (n=432)	234 (54.2%)	131 (30.3%)	67 (15.5%)
titles (n=432)	240 (55.6%)	107 (24.8%)	85 (19.7%)

Table 2: categorizations observed for each level of accuracy for all participants and categories

			exact	partial	incorrect
Entertainment	titles	n=144	61 (42.4%)	49 (34%)	34 (23.6%)
	keyphrases		56 (38.9%)	55 (38.2%)	33 (22.9%)
Government	titles		85 (59%)	31 (21.5%)	28 (19.4%)
	keyphrases		72 (50%)	51 (35.4%)	21 (14.6%)
Health	titles		94 (65.3%)	27 (18.8%)	23 (16%)
	keyphrases		106 (73.6%)	25 (17.4%)	13 (9%)

Table 3: categorization accuracy by category

			Surrogate type	
			Keyphrases	Titles
'unclassifiable' documents	correctly identified as 'unclassifiable'	n=144	44 (31%)	61 (42%)
	incorrectly identified as classifiable		23 (16%)	16 (11%)
classifiable documents	incorrectly identified as 'unclassifiable'	n=720	32 (4%)	64 (9%)
	correctly identified as classifiable		328 (46%)	296 (41%)

Table 4: accuracy of participant judgements about whether surrogates were classifiable or 'unclassifiable'

		Mean categorization score (s.d.)			
		All categories combined	Entertainment	Government	Health
Surrogate type	Keyphrases	2.08 (1.15)	1.76 (1.19)	2.02 (1.12)	2.44 (1.01)
	Titles	2.03 (1.21)	1.82 (1.21)	2.06 (1.23)	2.20 (1.18)

Table 5: mean categorization accuracies for all responses.

		Mean categorization score (s.d.)			
		All categories combined	Entertainment	Government	Health
Surrogate type	Keyphrases	2.39 (0.81)	2.22 (0.81)	2.32 (0.82)	2.63 (0.73)
	Titles	2.45 (0.80)	2.27 (0.78)	2.52 (0.79)	2.54 (0.80)

Table 6: mean categorization accuracies for responses correctly placed at some position within the classification tree.

		Mean categorization score (s.d.) excluding completely correct categorizations			
		All categories combined	Entertainment	Government	Health
Surrogate type	Keyphrases	1.49 (0.50)	1.55 (0.50)	1.51 (0.50)	1.32 (0.47)
	Titles	1.46 (0.50)	1.61 (0.49)	1.37 (0.48)	1.30 (0.46)

Table 7: mean categorization accuracies for responses correctly placed within the classification tree, but at the incorrect location.

Condition (surrogate type)	Mean times (mins)	Std. Deviation (mins)
Titles	5:44	1:49
Keyphrases	5:57	1:46

Table 8: Mean times for individual task completion.

Condition (surrogate type)	Entertainment (mean, std dev) (mins)	Government (mean, std dev) (mins)	Health (mean, std dev) (mins)
Titles	5:35, 1:53	6:29, 1:59	5:08, 1:13
Keyphrases	6:46, 2:28	5:45, 1:06	5:22, 1:01

Table 9: Mean task completion time by category for the two conditions.

		Mean surrogate length		
		Characters per surrogate	Lines per surrogate	Pages per category
Surrogate type	Keyphrases	57	2.0	2.48
	Titles only	33	1.2	1.54

Table 10: Mean surrogate lengths in the two conditions.

			Surrogate length		
			Mean characters per surrogate	Mean lines per surrogate	Pages required
Surrogate type	Keyphrases	Entertainment	59	2.1	2.62
		Government	57	1.9	2.34
		Health	55	2.0	2.48
	Titles	Entertainment	30	1.2	1.45
		Government	36	1.3	1.59
		Health	33	1.3	1.59

Table 11: Mean surrogate lengths for the categories in each condition.

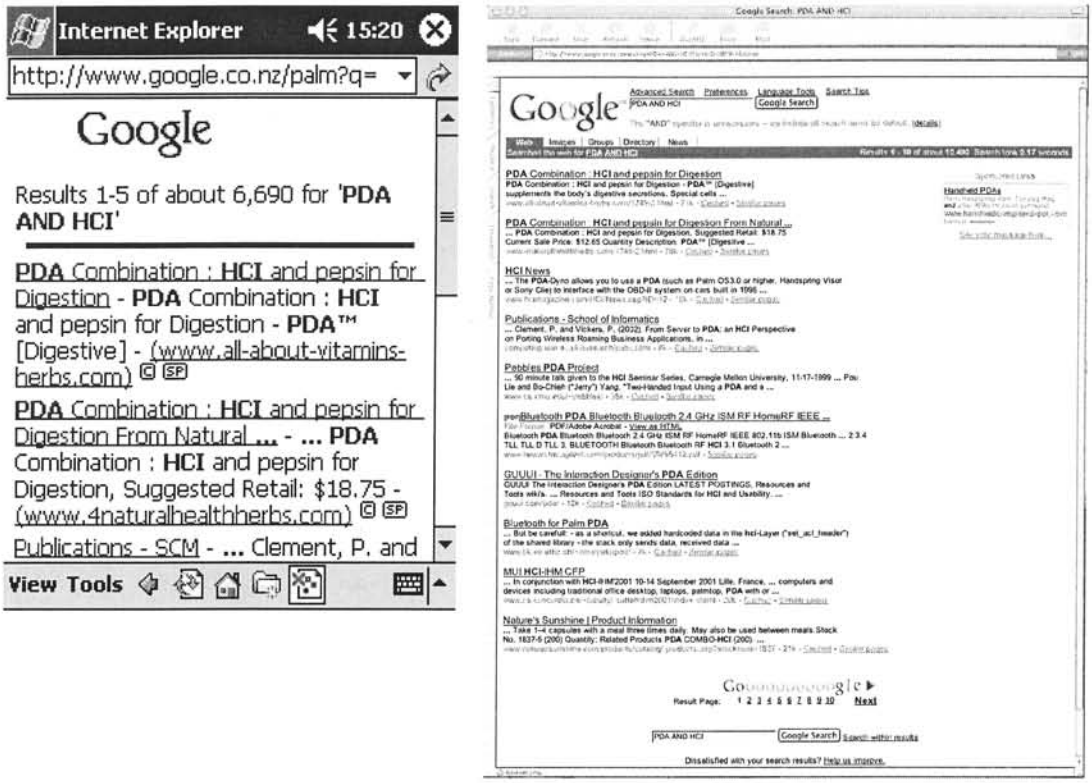


Figure 1: Google viewed on a handheld computer (left); and a conventional, large screen device (right)

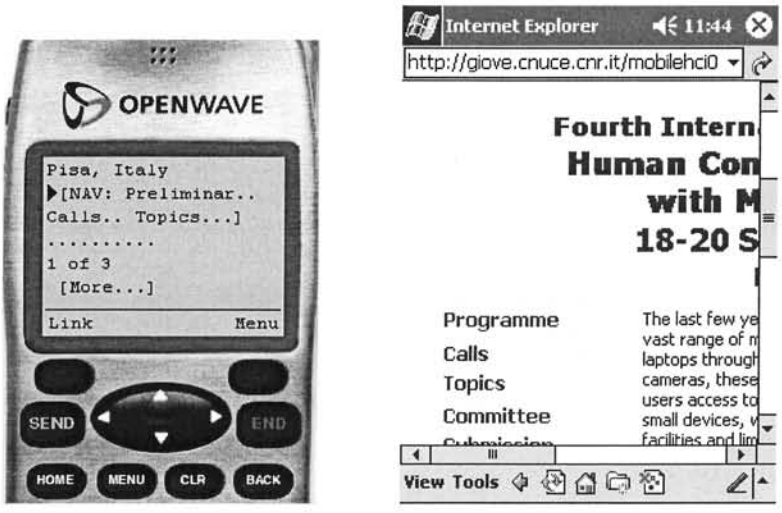


Figure 2: Search target documents displayed on a WAP phone simulation (left) and a handheld computer (right).

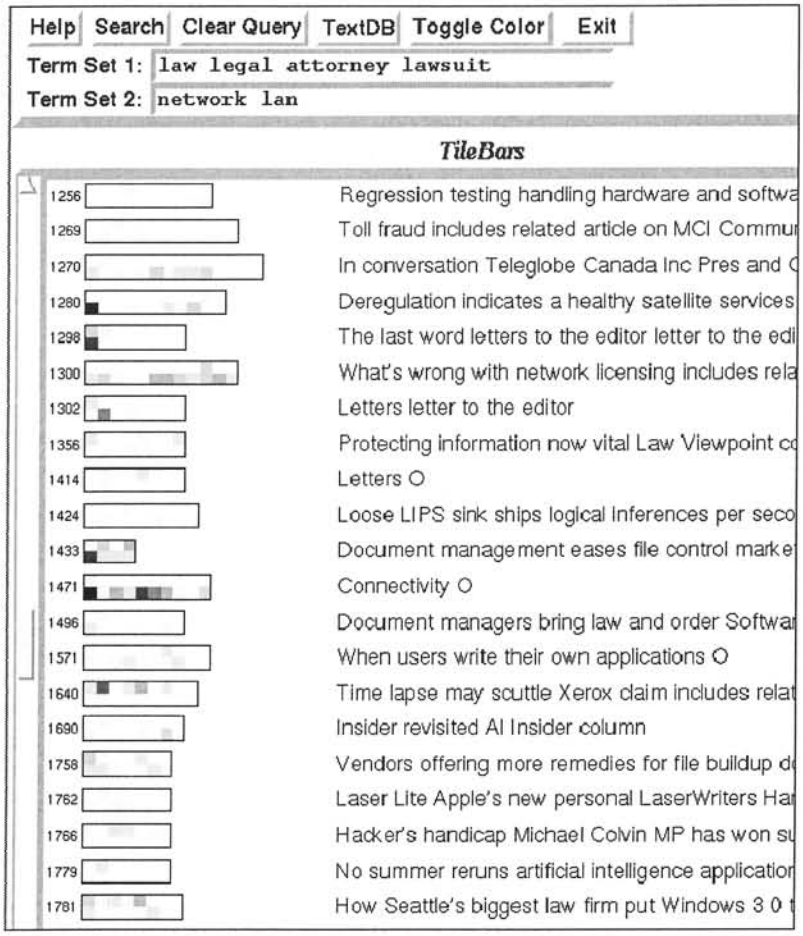


Figure 3: the TileBars interface (from Hearst 1995).

Figure 4: LibTwig view of search results. User can view context of each document match – in this example the "Farming snails…" document is seen to be part of the "Better Farming…" category which itself is part of the "Agriculture and Food Processing" section.

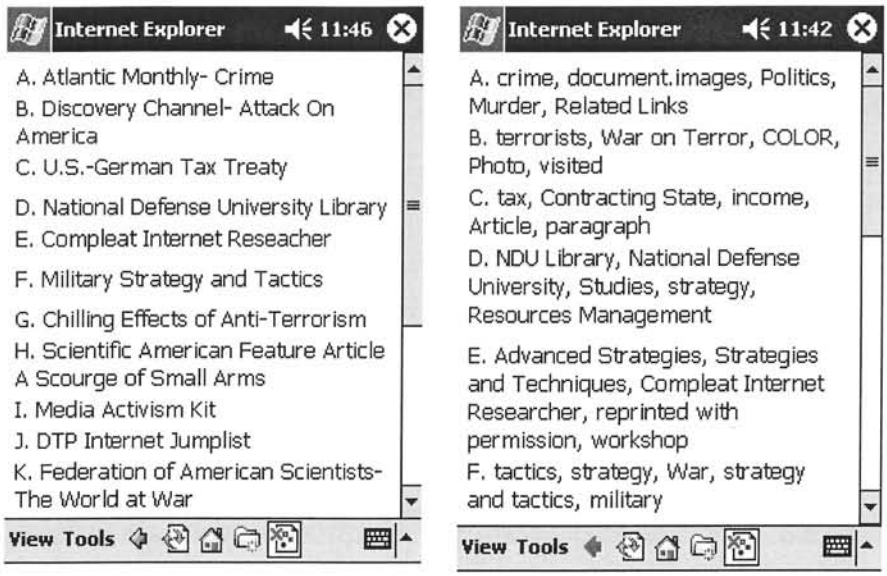


Figure 5: Example presentations of titles (left-hand-side) and keyphrases (right-hand-side) on small screen of a Pocket PC

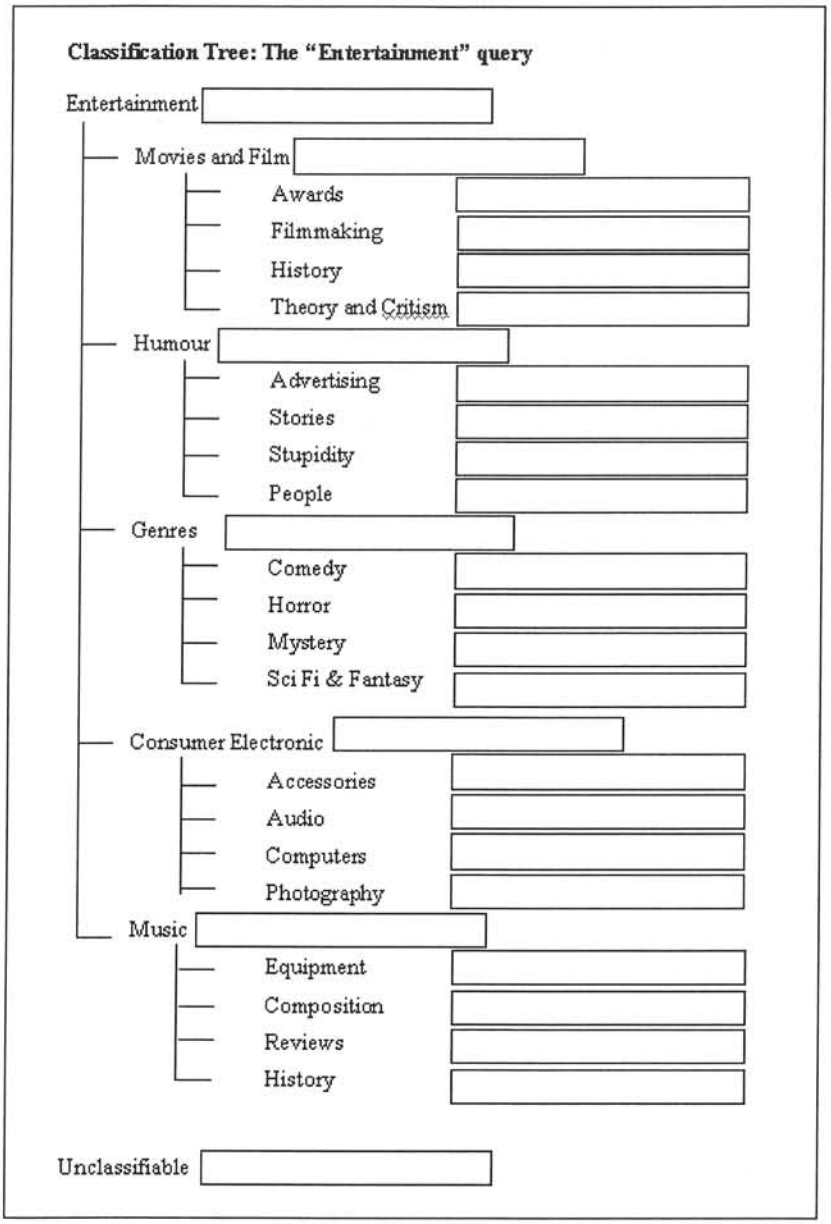
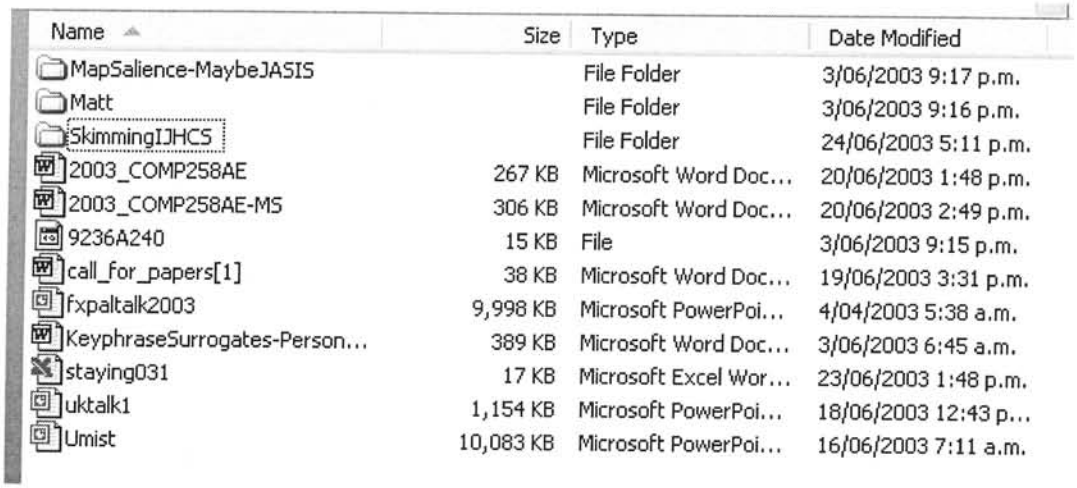


Figure 6: Hierarchy representation for the Entertainment category. Similar representations were created for the government and health categories.

Figure 7: Our prototype skimming interface, showing keyphrase emphasis (left) and sentence emphasis (right)



Name	Size	Type	Date Modified
MapSalience-MaybeJASIS		File Folder	3/06/2003 9:17 p.m.
Matt		File Folder	3/06/2003 9:16 p.m.
SkimmingIJHCS		File Folder	24/06/2003 5:11 p.m.
2003_COMP258AE	267 KB	Microsoft Word Doc...	20/06/2003 1:48 p.m.
2003_COMP258AE-MS	306 KB	Microsoft Word Doc...	20/06/2003 2:49 p.m.
9236A240	15 KB	File	3/06/2003 9:15 p.m.
call_for_papers[1]	38 KB	Microsoft Word Doc...	19/06/2003 3:31 p.m.
fxpaltalk2003	9,998 KB	Microsoft PowerPoi...	4/04/2003 5:38 a.m.
KeyphraseSurrogates-Person...	389 KB	Microsoft Word Doc...	3/06/2003 6:45 a.m.
staying031	17 KB	Microsoft Excel Wor...	23/06/2003 1:48 p.m.
uktalk1	1,154 KB	Microsoft PowerPoi...	18/06/2003 12:43 p...
Umist	10,083 KB	Microsoft PowerPoi...	16/06/2003 7:11 a.m.

Figure 8: Contents of a personal folder taken from a laptop computer.