Text Compression Heuristics: Coping with Small Displays

Daniel Gonçalves, Pedro Gomes, Sérgio Tostão, Joaquim Jorge
Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa Portugal
+351 2184 1772, +351 2131 00363
djvg@gia.ist.utl.pt, pnag@netc.pt, sergio.tostao@clix.pt, jorgej@acm.org

ABSTRACT
Most approaches to read web pages on portable devices require special versions of these pages and do not deal adequately with small screens found on PDAs. We present an approach that copes with display limitations by analyzing the content to display and organizing it into abstract visualization levels the user can zoom in and out of. Levels are defined by heuristics, discovered through task analysis and usability studies. Those studies provided meaningful insights about trade-offs between information filtering (compression) and text comprehension. They showed that significant compression could be achieved without hindering comprehension.

Keywords
Zoomable Interfaces, Web-Clipping, Morphological Text Analysis

INTRODUCTION
Mobile computing devices, such as Personal Digital Assistants (PDAs) are becoming widespread. These devices usually have small screens and reduced storage and processing capacities. While the desire to read World-Wide-Web (WWW) documents on PDAs is increasing, most web documents aren't designed to cope with these limitations. Most solutions to overcome this problem usually require alternate, trimmed-down, versions of documents to be prepared beforehand. Some of the most popular solutions, such as Web-Clipping, developed by Palm, Inc, or AvantGo (http://www.avantgo.com) do so. This is undesirable because it involves an increased effort in creating and maintaining alternate versions of a site, and because only prepared sites can be read. Also, it doesn't deal with the problem of having a small screen. Long documents might become too cumbersome to read in such a fashion.

We propose an approach that enables the user to adapt pre-existing sites and allows their visualization and access in a PDA, without undue changes to their contents. To achieve this and cope with display limitations, the system will allow users to navigate on the text using abstract levels of information, with a zoomable interface [2][5]. Also, a great level of customization is possible. The user can specify which sections of a page he wants to read on the PDA (thus getting rid of publicity, navigation bars, and other content-poor items). The system also has the advantages of existing solutions: fast during clipping phase and simplicity of use. In this paper we concentrate on the user interface and heuristics to make it possible to display longer texts on PDA screens without sacrificing text comprehension. We also present results where we have evaluated several heuristics for text compression as to their impact on text comprehension, and shown that a surprisingly high level of compression can be achieved while keeping acceptable comprehension levels.

ARCHITECTURAL FRAMEWORK
The system is divided in two main components: the retrieval and conceptual analysis of the information inside the web page (*Clipping System*) and the visual manipulation on the PDA (*Visualization System*). The former takes place on the user's PC, and is responsible for the analysis and transformation of the content to be read. The latter exists on the user's PDA, and consists of a zoomable interface that can be used to read that content.

The Clipping system allows the user to specify what parts of a web page are of real importance to him. It starts by parsing the page and building an internal representation of its structure and the relations between its components. The user then specifies a filter for the relevant information. Blocks of text can be included or excluded according to the occurrence of keywords, of their importance in the text (corresponding to the headings level they appear under), the font type and size used or the frame or table cell they appear in. A filter can be defined globally, for all sites transformed by the system, or on a per-site basis. In fact, most sites tend to present their information after a predetermined fashion or style. The user can tune these preferences to better adjust the filters. Thus, after a filter is tuned for a given site, it can be used until the site's layout or structure changes significantly, doing away with a special version of the site as required by other approaches.

COMPRESSING THE TEXT
While the clipping filter reduces the amount of information to be displayed, it remains far larger than what can easily be accommodated on a PDA's display. To help alleviate that problem, several criteria are used to establish levels of detail in the text that, while still allowing it to be understood for better display and navigation. While the user can ultimately zoom into the original text, he will seldom do so

if he understands it on a more abstract level. A questionnaire was undertaken to find in what ways do people usually reduce the size of a text, and to validate several hypothesis about that process. The results showed that three techniques are used to compress the text: *morphological analysis, abbreviations* and *heuristics*.

Morphological Analysis

A parsing application, SMORPH [1] is used to classify every word on the text according to the grammatical category it belongs to. Different classes contain words whose roles in a sentence are more or less crucial to its understanding.

Abbreviations

Several well-known abbreviations can also be applied. These include not only standard dictionary abbreviations, but also others that are of common use nowadays, such as those used on SMS or Internet messages. Examples of abbreviations can be found on the following table:

abbr.	Abbreviation	IMHO	In my humble opinion
fig.	Figure	U	You
masc.	Masculine	AFAIK	As far as I know

Heuristics

Some heuristic criteria are also used. From the inquiries, the following heuristics were chosen:

- **Remove internal vowels:** all internal vowels, except for those that precede or succeed another vowel, are removed (exchangeable→ exchngeable)
- **Remove 'e' from the end of a word** (service→servic)
- **Replace the '-ly' suffix of adverbs with '/'** (friendly→friend/)
- **Remove the 'u' after a 'q'** (quiet→qiet). This heuristic is an example of a more general instance: replacing words with others that sound the same when read out loud.
- **Remove all text within parenthesis.**

While the nature of these heuristics can be language dependant (the removal of the 'u' after a 'q' can be an example of this, since there is no difference in sound in the Portuguese language), the principles they embody can be extended to different languages.

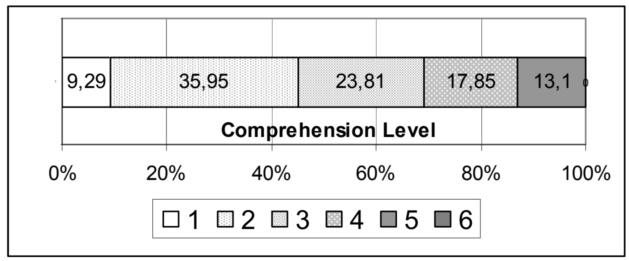
The Zoom Levels

Although several criteria for reducing text size have been presented, an important question remains unanswered: in what way can we combine them to define relevant zoom levels? Several reading comprehension questionnaires were made. Texts were presented using different combinations of criteria. Questions were then asked about the contents of those texts. The levels that proved to allow both a fair compression level and a good comprehension level were:

Level	Contents
1	Names + Verbs + All heuristics
2	Names + Verbs + Adjectives + All heuristics
3	Names + Verbs + Adjectives + Pronouns + Adverbs + All heuristics
4	All morphological classes + All heuristics
5	Names + Verbs + Adjectives + Pronouns + Adverbs + Parenthesis Suppression + Adverb Supp.
6	Original Text

RESULTS

After coding the zoomable interface on the PDA, we performed usability tests to try and evaluate compression versus comprehension trade-offs. By default, the system starts at zoom level 1. We measured reader's comprehension level as they zoomed in trying to understand the text, through questionnaires. The following graphic condenses the results.



As can be seen, nearly 70% of the users only needed to zoom back to level 3 to understand the content. None of the subjects needed the original text to do so. Furthermore, at level 3 the average compression level is 55%. Thus, our approach is able to display twice as much text as uncompressed systems, at a reasonable comprehension level.

CONCLUSIONS

The main concern that led to this work is the need to display large amounts of text on small displays. We have shown that, given the right criteria to summarize text, large levels of compression can be achieved without significant loss to the comprehension level. In fact, most users were able to fully understand the text looking at a summary half of original size, and none needed to read the entire text. The choice of the criteria and their combination is critical. As future work, other criteria should be considered, along with a more thorough analysis of the text, including, perhaps, semantic information.

REFERENCES

1. Ait-Molahtar, S. L'analyse pré syntaxique en une seule etape. PhD Thesis, Université Blaise Pascal, GRIL, 1998
2. Bederson, B., Meyer, J., Good, L., Jazz: An Extensible Zoomable User Interface Graphics Toolkit in Java, In ACM UIST 2000, pp.171-180.
3. Euralex 2000 Tutorial – Homepage, European Association for Lexicography, http://www.ims.uni-stuttgart.de/euralex/conferences/elx2000/tutorial/
4. Gomes, P., Tostão, S., Gonçalves, D. Jorge, J. Web Clipping: Compression Heuristics for Reading Text on a PDA, Proceedings MobileHCI, 2001.
5. Stuart, P. Context and Interaction in Zoomable User Interfaces, Published in the AVI 2000 Conference Proceedings, pp 227-231.