Text Summarization of Web pages on Handheld Devices

Orkut Buyukkokten **Hector Garcia-Molina** **Andreas Paepcke**

Digital Libraries Lab (InfoLab), Stanford University, Stanford, CA, 94305
{orkut, hector, paepcke}@cs.stanford.edu

Abstract

We present a design for displaying and manipulating HTML pages on small handheld devices such as personal digital assistants (PDAs), or cellular phones. We introduce methods for summarizing parts of Web pages. Each page is broken into text units that can each be hidden, partially displayed, made fully visible, or summarized. A variety of methods are introduced that summarize the text units. We found that the combination of keywords and single-sentence summaries provides significant improvements in access times and number of required pen actions, as compared to other schemes.

Introduction

Wireless access to the World-Wide Web from handheld personal digital assistants (PDAs) is an exciting, promising addition to our use of the Web. Frequently, we know that the information we need is online, but we cannot access it, because we are not near our desk, or do not wish to interrupt the flow of conversation and events around us. PDAs are in principle a perfect medium for filling such information needs right when they arise.

Unfortunately, PDA access to the Web continues to pose difficulties for users. The small screen quickly renders Web pages confusing and cumbersome to peruse. Entering information by pen, while routinely accomplished by PDA users, is nevertheless time consuming and error-prone. The download time for Web material to radio linked devices is still much slower than landline connections. The standard browsing process of downloading entire pages just to find the links to pursue next is thus poor for the context of wireless PDAs.

We have been exploring solutions to these problems in the context of our Power Browser Project [1,2,3,4,5]. The Power Browser provides displays and tools that facilitate Web navigation, searching, browsing, and input entry from a small device. The Power Browser uses proxy technologies to improve performance by doing computation-intensive operations on behalf of the client. The proxy filters irrelevant content (i.e., HTML tags, multimedia) and transforms the Web page into an appropriate format to be displayed on the handheld device.

We were able to show that the facilities outlined above are effective for searching and browsing. This summary addresses methods for text summarization of Web pages for small devices.

1 Page Summarization

The page summarization facility is employed after a user has searched and navigated the Web, and wishes to explore in more detail a particular page. At this point, the user needs to gain an overview of the page, and needs the ability to explore successive portions of the page in more depth. Our proxy server, in collaboration with the PDA, provides two levels of summarization: a macro level, and a micro level.

We apply ‘Macro-level’ summarization, which relies on structural analysis of Web pages. These summaries allow users to expand and contract pages based on their relative structural nesting. An additional, integrated ‘Micro-level’ summarization uses information retrieval techniques to outline portions of the text for the user.

1.1 Macro-Level Summarization

The proxy begins by partitioning the page into 'Semantic Textual Units' (STUs). STUs are page fragments such as paragraphs, lists, or ALT tags that describe images. In a second step, the proxy then uses font and other structural information to identify a hierarchy of STUs. For example, the elements within a list are considered to be item STUs nested within a list STU. Similarly, elements in a table, or frames on a page, are nested. Hiding the nested STUs finally completes macro summarization.

Note that this macro level summarization does not require special formatting at the Web sources. This freedom from intrusion is a significant advantage of our approach over schemes that rely on pages to be specially structured for PDAs. Reference [3] provides more detail on how STUs are extracted from pages, and how they are ordered into a hierarchy.

1.2 Micro-Level Summarization

Once we break up the Web page into STUs and organize them into a hierarchy, we need a convenient and efficient way to display each STU. We call this step Micro-Level Summarization.

We have explored five methods for micro-level summarization, and performed user testing to learn how effective each of them are in helping users solve information tasks on PDAs quickly. All of the methods we tested retain our macro-level accordion browser approach of opening and closing large structural sections of a Web page. However, the methods differ in how they summarize and progressively reveal the STUs at the micro-level.

Every method we tested displays each STU in several states. The information for each state is prepared quite differently in each method. All displays are textual. That is, none of the STUs displays images. They work as follows:

- *Incremental:* Each STU is revealed gradually in three states; the first line, the first three lines and the whole STU.
- *All:* This display method shows the text of an entire STU in a single state. No progressive disclosure is enabled.
- *Keywords:* The third method displays in its first state the 'important' keywords that occur in the STU. We show all of the keywords on the display, even if they extend beyond a single line and wrap down to additional lines. The second state shows the first three lines of the STU. The third state shows the entire STU.
- *Summary:* This method consists of only two states. In the first state the STU's 'most significant' sentence is displayed. The second state shows the entire STU.
- *Keyword/Summary:* This method combines the previous two methods. The first state shows the keywords. The second state shows the STU's most significant sentence. Finally, the third state shows the entire STU.

There are of course many other ways to mix keywords, summary sentences, and progressive disclosure. However, in our initial experience, these 5 schemes seemed the most promising, and we hence selected them for our experiments. Also note that in all of these methods, only one state is used if an entire STU happens to fit on a single line. Similarly, if an STU consists of only one sentence, the most significant sentence is the entire STU and there are no additional state transitions.

Conclusion

Our user experiments showed that a combination of keyword extraction and text summarization gives the best performance for discovery tasks on Web pages [4]. For instance, compared to a scheme that does not summarize, we found that for some tasks our best scheme cut the completion time by a factor of 3 or 4. Our overall results suggest that summarization approaches are key to successful PDA-based user interactions with the World-Wide Web. Information task completion times and input effort can be significantly reduced if summarization techniques of various forms are employed.

References

[1] O. Buyukkokten, H. Garcia-Molina, A. Paepcke, and T. Winograd, Power Browser: Efficient Web Browsing for PDAs, In Proc. of the Conf. on Human Factors in Computing Systems, CHI'00, 2000, pp. 430-437.

[2] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke, Focused Web Searching with PDAs, In Proc. of 9th Int. World-Wide Web Conf., 2000, pp. 213-230.

[3] O. Buyukkokten, H. Garcia-Molina, A, Paepcke, Accordion Summarization for End-Game Browsing on PDAs and Cellular Phones, , In Proc. of the Conf. on Human Factors in Computing Systems, CHI'01, 2001.

[4] O. Buyukkokten, H. Garcia-Molina, A, Paepcke, Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices", In Proc. of 10th Int. World-Wide Web Conf., 2001.

[5] O. Kaljuvee, O. Buyukkokten, H. Garcia-Molina, and A. Paepcke, Efficient Form Entry on PDAs, In Proc. of 10th Int. World-Wide Web Conf., 2001.