3.Modifying Page Content

A relatively unused technique for client-side reformatting of pages is to modify the page content (as opposed to the formatting). This can take the form of contract extraction or summarisation. The reason for this lack of use may be due to the relative computational complexity of altering content as opposed to reformatting it.

3.1DOM-based Content Extraction

An easily implementable technique for content extraction is to alter the generated DOM (Document Object Model) tree[1]. This involves using analytical techniques to identify where relevant content is located in a page. For example, by analysing the percentage of text that is a link in a particular paragraph, it can be identified whether that paragraph contains important information. Depending on the result, the paragraph can be positioned elsewhere in the document, or removed from the DOM tree entirely. Figure 1 shows an example of a page before and after using DOM-based content extraction.

Figure 1 – Before and after of a page using DOM-based content extraction (Images from [1])

An advantage of this technique is that the original structure of the page remains, thus reducing the amount of input necessary to view content. Also, as the technique is independent of target device, it could also be implemented as a proxy service. Paper [1] chose this technique, as they believed it was an optimal method for reducing a website to only its relevant information, without altering the underlying structure of the page.

This technique can prove to be problematic for pages that make heavy use of JavaScript though, as some JavaScript pages modify the DOM themselves and may produce unexpected results if it has been modified externally. This problem is becoming increasingly important to consider as AJAX (Asynchronous Javascript And XML) becomes popular. Another danger of using contact extraction is that it may extract the incorrect data. For example, the method used in paper [1] fails on sites that make heavy use of links in the main content, as shown in figure 12 of said paper.

Assuming that it doesn't class it as irrelevant information, this technique would not hamper the viewing of sites based on Macromedia Flash[12], or other proprietary browser plug-ins. These types of sites have accessibility issues outside the scope of this document, however, and would need to be considered as special cases.

3.2Summarisation

Another technique, covered in the paper *Text compaction for display on very small screens[*5], is to summarise certain data on a page. This technique works by contextually matching certain phrases, words and numbers and providing shorter versions as replacements. The particular implementation in [5] is applied to e-mails and can, for example, replace the text '2nd November 2005' with '2/11/05'. Further reductions can be accomplished by omitting unimportant words from sentences, however, algorithms must be carefully tested so as not to alter the meaning of the content. A particular downfall of this technique is that it is locale-specific (specialised for a particular region) and if employed incorrectly, can actually make content harder to read.

4.Modifying Page Layout

The most common method of altering content to be viewable on devices with small screens is to alter the layout of the page, usually by employing custom CSS. This is the method employed by the Minimo browser[6], pictured in figure 2.

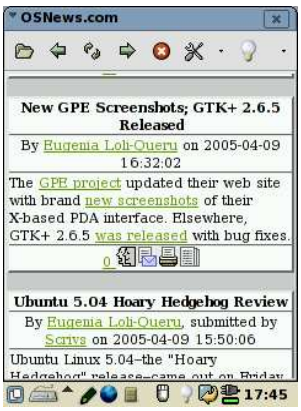

Figure 2 – The Minimo (Mini Mozilla) browser

By examining the CSS that Minimo uses[7], it can be seen exactly what methods it uses to reformat pages. The Minimo CSS works by overriding properties of the style of the page in question by marking those properties as important. It cancels all explicit size settings, removes all padding and margins, removes floating elements, flattens tables, sets the maximum width of all elements to the width of the browser to remove the horizontal scrollbar and highlights all links. This gives websites a 'stacked' appearance, similar to the proprietary method of the Opera[4] browser and is a form of linear transformation. Research papers [9][10] show that it is desirable to eliminate horizontal scrolling to reduce the amount of user interaction required to understand the content of a page.

This technique does not reorganise the content in any order of importance, however, so sites with large amounts of irrelevant content before the main