Efficient Text Identifier for Mobile Web Browser

*Leonardo Juniti Nomoto, **Chang-Su Kim,
School of Electrical Engineering, Korea University

Abstract

Mobile devices are being widely used to access Internet contents. However, most available web pages are designed for desktop computers and consequently it is inconvenient to browse large web pages on mobile devices with small screen. Text identification is a process to extract texts from the body of a web page, which are then displayed in a comfortable way for reading. In this paper, we propose a text extraction scheme and discuss its implementation.

1. Introduction

Recent advances in mobile communications have brought new generations of mobile devices that are capable of rendering web contents as desktop computers. There are many techniques for rendering web pages in mobile web browsers [1-9]. Most of them focus on how to fit rich web contents into small screens.

Mobile devices with small screen severely restrict textual presentation. In our work, we attempt to extract desired text contents from a certain area of the web page, so that they can be later displayed seprately on small displays in a visually pleasant way.

Initially, a parameter should be identified: the point position of the web page is identified inside the text area. This position can be obtained by a user click on the screen. Then the algorithm takes tree steps to extract the text. The first step is to search for the deepest and smallest node in the document object model (DOM) tree that contains the input position. The second step determines a node or a group of nodes that contains all texts related to one paragraph, line, title from the web page, since the node from the first step usually does not contain all desired texts. Finally, if the text is an article with more than one paragraph, by using semantics analysis and geometric analysis, we propose extract the whole text.

2. Background

2.1. Web Page Position vs Screen Position

First, when a web page is opened in a web browser and the page does not fit on the size of the screen because its size is larger than the screen size, some part of theweb page is hidden as shown in Figure 1.

The point position parameter is obtained from the web page area and not from the screen area. This position consists of two parameters x and y, where x is the horizontal position and y is the vertical position. The (0,0) point reference denotes the upper left position.

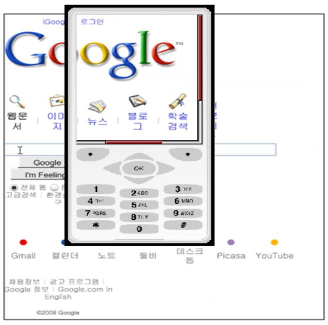
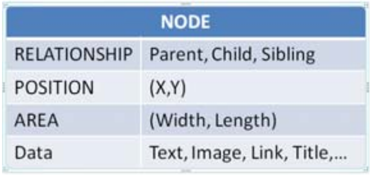


Figure 1. Screen display vs Web Page Size

2.2. Web Page Structure

When a web page is opened in a mobile web browser, its HTML parser generates the HTML DOM tree that contains a group of DOM nodes. In this method, we consider a web browser compatible with at least the DOM Level 1 [9].

The most significant information is that each node contains the information of relationship between other nodes, position on the page, area, and data (text, picture, link, title, etc), as shown in Figure 2.



NODE	
RELATIONSHIP	Parent, Child, Sibling
POSITION	(X,Y)
AREA	(Width, Length)
Data	Text, Image, Link, Title,...

Figure 2. Significant information in a DOM node [9].

3. Algorithm

3.1 Deepest Node Searching

Given the input point position (x,y), the algorithm should search into the DOM tree starting from the top node. If in any step the node contains the position (x,y) and does not have children, stop the algorithm and store the node in reference. In this way, we select the deepest node containing the position (x,y). Figure 4 demonstrate the result of our simulation in obtaining the deepest node.

3.2 Neighbour node analysis

After searching the deepest node containing the position (x,y), the next step is to determine if there is previous or following nodes on the tree that contains complimentary information to the deepest node. The proposed algorithm tracks the siblings of the deepest node to check if they have the same kind of information or at least they are text nodes. In this way, we determine which sibling contains the beginning of the paragraph and which one contains the end of the paragraph.

Figure 5 shows an example in extracting whole paragraph extracted from an article. Usually, at this point a paragraph can be determined to be contained inside one node, the parent node of the deepest node. This node contains the group of nodes from the start node until achieving the finish node. Lately, when the extraction is required, only this sequence of nodes will be verified. The parent node is used here as a reference, because probably contains the shape and position of where the paragraph is located on the web page.

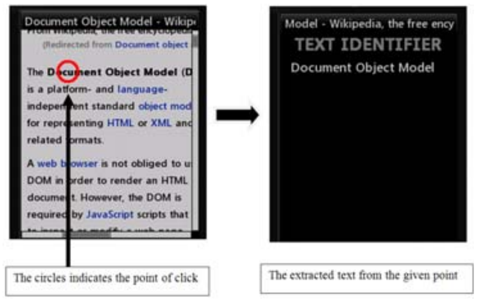
3.3 Semantic and Shape analysis

Finally, we extract whole text considering all paragraphs above and below the selected one. We observe that if one paragraph is extracted, the previous and following paragraph shapes tend to have some similarities.

Therefore, we search for the previous sibling of the node obtained from the second step. Check it and its children nodes to verify if it contains a paragraph. If a paragraph node is found, then search the previous sibling. Keep searching until there is no more siblings or no paragraph node is found.

4. Experimental Results

The first step was implemented successfully, and always returns the deepest node to the given position. The second step is now being developed and some experiments demonstrated that more rigorous semantic analysis may be needed since in some web sites the whole paragraph cannot be extracted correctly. One example is from the wikipedia web pages, where there are too many links to refer to other contents. In such a case, it become very difficult to know which is the parent node for the paragraph.



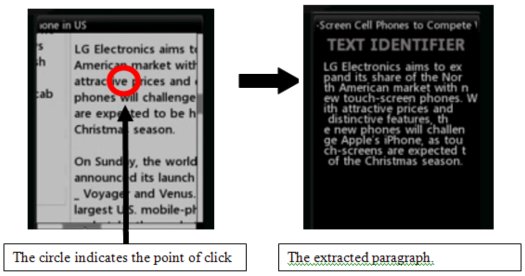
Document Object Model - Wikip
(Redirected from Document object
The Document Object Model (D
is a platform- and language-
independent standard object mod
for representing HTML or XML and
related formats.
A web browser is not obliged to u
DOM in order to render an HTML
document. However, the DOM is
required by JavaScript scripts that

Model - Wikipedia, the free ency
TEXT IDENTIFIER
Document Object Model

The circles indicates the point of click

The extracted text from the given point

Figure 4. Deepest node searching result



one in US
s
h
ab
LG Electronics aims t
American market with
attractive prices and
phones will challenge
are expected to be h
Christmas season.
On Sunday, the world
announced its launch
_ Voyager and Venus.
largest U.S. mobile-ph

Screen Cell Phones to Compete
TEXT IDENTIFIER
LG Electronics aims to ex
pand its share of the Nor
th American market with n
ew touch-screen phones. W
ith attractive prices and
distinctive features, th
e new phones will challen
ge Apple's iPhone, as tou
ch-screens are expected t
of the Christmas season.

The circle indicates the point of click

The extracted paragraph.

Figure 5. Whole paragraph extracted

5. Conclusion

In this work, we proposed a method for text extraction from web pages. The text extraction can increase the capabilities of mobile web browsers to browse web content. More reliable extraction scheme is currently being developed.

6. References

[1] L. Chittaro, “Visualizing Information on Mobile Devices,” Computer, vol. 39, 2006, pp. 40 – 45.

[2] T. Maekawa, T. Hara, S. Nishio, “Two Approaches to Browse LargeWeb Pages Using Mobile Devices,” Mobile Data Management-MDM `06, May 2006,pp. 52.

[3] Y. Arase, T. Hara, T. Uemukai, S. Nishio, ”Nine-Button Web Browsing System for Cellular Phone Users,” Innovations in Information Technology, Nov. 2006, pp. 1-5.

[4] Y. Arase, T. Maekawa, T. Hara, T. Uemukai, S. Nishio, “A Web Browsing System based on Adaptive Presentation of Web Contents for Cellular Phones,” ACM International Conference Proceeding Series, vol. 134, 2006, pp. 86 – 89

[5] P. J Timmins, S. McCormick, E. Agu, C. E. Wills, “Characteristic of Mobile Web Content,” 1st IEEE Workshop on Hot Topics in Wireless Systems and Technologies – HOTWEB `06, Nov.2006, pp. 1-10.

[6] Y. Chen, W. Y. Ma, H. J. Zhang, “Detecting Web Page Structure for Adaptive Viewing on Small Form Factor Devices,” International World Wide Web Conference, 2003, pp. 225-233.

[7] M.S. Castelhano, P. Muter, “Optimizing the reading of electronic text using rapid serial visual presentation”, Behaviour & Information Technology, vol. 20, 2001, no. 4, pp. 237-247.

[8] Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices,”

[9] "Document Object Model (DOM) Level 1 Specification," http://www.w3.org/TR/REC-DOM-Level-1/, version 1.0, Oct. 1998.