1. INTRODUCTION

The last two decades have seen an explosion of the World Wide Web (WWW). The major way of accessing the WWW is viewing Web pages offered by the website designer using a PC Web browser. However, nowadays, more and more people are using small screen devices such as cell phones to access the WWW. Obviously, it is difficult for users to view whole web pages on a small screen using common browsers. Furthermore, for non-visual browsers which are used by people with visual disability, current web pages also face the challenge of information overload, which makes the browsing very inefficiency. Therefore, for these applications, analysis and reorganization of the web page become inevitable.

For web pages containing a main text (e.g., a news story), the problem is relatively easy, since the main story can be segmented and extracted according to text percentage and area. However, for web pages without main text and serving as “index pages” or “hub pages” (e.g., home page of most websites), analysis and reorganization of such web pages remains a major challenge. Earlier methods merely extract all texts from web pages and reorganized the texts and display them on small screens or transform them into speech for non-visual browsing. Obviously, these methods do not solve the problem of information overload and
make browsing inefficient. Nowadays, much research effort has been paid on segmenting a web page into small blocks and transforming the web page into a sequence of blocks. However, for both methods using text information or visual information, current segmentation performance remains dissatisfactory: Blocks obtained by current methods are often not visually and semantically consistent with the original web pages. Furthermore, current methods just segment web pages into blocks and allow users to skip between blocks, which are still inefficient in browsing.

In this paper, the TB-WPRO (title-block based web page reorganization) method is proposed. The method segments web pages from the designer’s perspective using both visual information and page layout. The method is mainly used for segmentation of “index pages” or “hub pages” as introduced above. The main idea is to extract title blocks from web pages and reorganize them in a hierarchical way. A title block is a block with a title which describes the category of the block and a main content within the category. Compared to current methods, the proposed approach can obtain a promising segmentation result where blocks are visually and semantically consistent with original web pages. Furthermore, the proposed method can filter less important contents such as navigation bars and some advertisements by only considering title blocks, which helps in dealing with the challenge of information extraction.

The rest of this paper is organized as follows: Section 2 provides a brief review of related work. Section 3 and Section 4 presents the page segmentation method and content reorganization methods in TB-WPRO, respectively. Experimental results are given and analyzed in Section 5. Finally, conclusions are drawn in Section 6.

2. RELATED WORK

A variety of methods have been proposed to segment web pages into small blocks. The early way is considering the tag information in the DOM tree such as <P> (paragraph), <TABLE> (table), <UL> (list), <H1>-<H6> (heading), etc. Lin and Ho (2002) partition a page into several content blocks according to HTML tag <TABLE> and then use entropy to distinguish redundant block from the information block. Hattori et al. (2007) propose a hybrid method using content-distances and layout information to segment the webpage. However, the HTML tags do not contain any semantic information and could be misused. A FOM model is proposed by Chen et al. (2001). It treats each object in the webpage as either a basic or composite FOM which describe the object functionality. But the model does not describe the concrete segmentation method. The VIPS algorithm is proposed by Cai et al. (2003) which analyzes the webpage using the vision separator to segment the content into different areas. However, a parameter PDoC (Permitted Degree of Coherence) must be given by the user to get segmentation result. The paper does not tell how to determine proper PDoC to get blocks visually and semantically consistent with the original web page. Baluja (2006) uses a learning method to divide a webpage into 9 parts which the user could select to zoom. Chakrabarti et al. (2004) puts the nodes in DOM tree into a weighted graph, and formulates an appropriate optimization problem on it. The optimization problem is solved by a learning framework. But such machine learning method need a training process and is difficult to implement.

3. WEB PAGE SEGMENTATION

3.1. Overview

In most of the web pages, a block is often organized with a title on the top to indicate the category of the content below. Figure 1 (b) shows an example: the line contains “REAL ESTATE” informs us that the content below tells stories of “real estate”. This pattern is generally adopted by the designers of web pages when they designed the body of the page excluding the navigation bar on the top of the page, the
Map-Based LBSs for Hiking: A Review of Requirements, Existing Solutions, and Future Trends
[www.igi-global.com/chapter/map-based-lbss-hiking/67047?camid=4v1a](www.igi-global.com/chapter/map-based-lbss-hiking/67047?camid=4v1a)

Design and Implementation of Bipolar Digital Signal Acquisition and Processing System Based on FPGA and ACPL-224
[www.igi-global.com/article/design-implementation-bipolar-digital-signal/66060?camid=4v1a](www.igi-global.com/article/design-implementation-bipolar-digital-signal/66060?camid=4v1a)