Web Document Manipulation for Small Screen Devices: A Review

Hassan Alam and Fuad Rahman[1]
BCL Technologies Inc.
fuad@bcltechnologies.com

Abstract

Web browsing using small screen handheld devices is becoming more and more common. There has been a realization over the last couple of years that handheld devices are becoming much more than Personal Digital Assistants (PDAs), as were originally called, that they are here to stay and are about to become direct competitors to laptop and desktop computers. There were two principal shortcomings that prevented the widespread adoption of these devices in the past. The first was the absence of workable connectivity and network access with good speed. The second problem was its relatively tiny display area. Since the incorporation of wireless technology inside the handheld computers and 3.5G systems offering high network speeds, the first problem is being aggressively addressed. The second problem is a different type of problem and requires manipulation of web pages to create a different paradigm for serving and browsing web pages. Researchers have been working to solve this problem and this review presents a concise summary of the state of the art of the research related to web page manipulation for small screen devices.

1. Introduction

Web page manipulation is increasingly becoming a very important part of document analysis and recognition field. Aside from the fact that web documents pose many fascinating research questions to researchers working in the area of document analysis, and often challenge the traditional document model, there are huge commercial implications of this study. Within a decade of its inception, web documents have become an integral part of our daily lives. It has opened up a window of immeasurable opportunities for information exchange, sharing, connectivity, and interaction among diverse groups of people.

However, the concept of a web page, as is traditionally accepted and adopted, is aimed at desktop devices. Especially in the mid-nineties, the cheaper high-resolution monitors have aided in the rich and very sophisticated web page designs. On top of that, integration of multimedia objects, such as audio, video, games, use of flash technology, java scripts, image maps, and the Cascading Style Sheets (CSS) have made the web a kaleidoscope of very interesting but complex content and layout [45].

In recent times, there has been an explosion of handheld and wearable devices capable of web browsing. Due to their very nature, these devices have very small display screens and viewing web pages on these small screens is anything but practical [33]. Researches have started to address this problem by manipulating web pages so that that they are easily viewed on these devices [38].

This paper presents a review of approaches for web page manipulation for viewing on small screen devices. Other aspects of web page analysis, such as classification of web pages, or information retrieval techniques for web pages etc. are not part of this review.

2. Why Manipulate Web Pages?

There is a growing demand for viewing web pages on small screen devices. Mobile viewing allows keeping in touch with the rest of the world while on the move. So web page re-authoring can be of great interest. Another motivation is to summarize web content to help in rapid viewing, as time is a very important commodity and the amount of information available these days makes it impossible to browse through entire web sites. Often there is a demand for alternative browsing, such as the use of voice ([10],[46]). Most email browsing software now support HTML pages, which make the problem of universal email accessibility a problem of web page manipulation. Last, but not the least, web page manipulation is often required in order to extract content and transfer it to other formats, such as PDF and others. In this scenario, web page manipulation supported with document analysis techniques is the best choice.

3. Low Level Manipulation of Web pages

Low level processing is the core of the web page manipulation solution. This involves structural analysis of the web by decomposition and subsequent processing. In this approach, the web document is initially decomposed into constituent segments exploiting the HTML data structure [42]. Once segmented, content extraction can be

[1] Corresponding author

attempted by classifying these segments into various classes, such as image, text, story (large contiguous chunk of text), titles, side bars, tables, top bars, advertisements and so on [43]. When the classification of each segment is known, segments of specific classes can be merged using a set of rules.

Extraction and processing images also play a large part in this process. It has been shown that a large fraction of images found on the web contain textual information which is never repeated [34]. Researchers have been actively working on means to extract and recognize this text embedded into these images ([4],[37]). Classification of these images into various classes has also been attempted [26].

On a slightly higher level, detection of specific structures, such as tables, within web pages has been an active research problem ([20],[27],[41],[47]). Using forms within the narrow display area is another challenge [32].

4. What are the Options for Web Page Manipulation?

There are many ways in which web pages can be manipulated. This section outlines these options.

4.1 By Hand

This was the first approach adopted by early systems. The idea was that content will be maintained separately for access by different devices, such as full HTML for desktop and laptop devices, HTML 3 or under for PDA devices and WAP or other formats for cell phones. It was very quickly realized that the cost of maintaining separate content bases can be enormous, and it was never successfully implemented on a large scale. Some news sites, such as the BBC (www.bbc.co.uk), still have text only and PDA only versions. But the content is essentially scaled down by a huge factor and in no way represents a good solution.

4.2 Transcoding

Transcoding is an automatic solution that attempts to either replace the source of full HTML web pages of tags that are not supported by the target browser or creating approximations by creating closely resembling output [30]. The output is simply re-flowed in the natural order of the source document. Since this is an automated solution, it does not require keeping and maintaining separate content bases, and the conversion can be dynamic and in real time.

Annotation based transcoding is also promising, and initial research has been very encouraging ([5],[23],[24],[25]). In order to support transcoding, an active proxy design is very important ([7],[35]). Some transcoding solutions ([28],[29]) try to exploit the web page structure to generate more accurate output. Researchers have also tried to exploit semantics to leverage some linguistic knowledge of the web page content ([39],[40]) while transcoding web pages.

4.3 Automated Re-authoring

There are some major disadvantages of transcoding. The re-flow is in the natural order of the original document, so less important content, such as top bars, left bars, advertisements etc., needs to be browsed before finding the more important content. In addition, this tries to push the complete content of a web page to the small screen device. These devices are usually low on memory and often use a slower network speed. The alternate solution is to analyze the document, extract its structure, extract content, re-engineer a page and then display the document. This usually creates a multi-dimensional document structure from the flat two-dimensional web document. There are three major approaches of automated re-authoring. The next sections elaborate on this.

4.3.1 Table of Content

In this approach, the document is re-created based on extracted content and a *heading* is created. This heading hides a hyperlink, which if selected, can load up details associated with the headline [31]. So the first display per web page is always a table of content (TOC) with hyperlinks ([8],[44]). In this model, there can be any number of abstractions, but practical considerations dictate that any more than two levels is confusing for most users ([11],[12],[13],[14]). Some attempts include the use of a web digestor that can explore the structure of a web page to detect 'blocks' of information and use transcoding techniques to re-flow the content [9]. Other attempts include approaches to identify content and then either label the blocks of content or convert them to image thumbnails, which in turn are connected to the details of the relevant content ([50],[51],[52]).

4.3.2 Summarization

Summarization of web pages is another approach of web page re-authoring. In this approach, the content is not simply separated into separate layers; the textual part of the content is summarized using natural language techniques [6]. Obviously this works best with web pages that are predominantly textual in nature, but it is possible to get decent results with most web pages. Specific approaches targeted at email processing/viewing [21], presentation/navigation of information retrieval results [36], financial news delivery [49] and others have been reported. Attempts to use context in summarizing web pages is also being explored [22].

4.3.3 Hybrid

In this case, TOC and summarization approaches are combined [3]. The web page is still separated into multiple layers, but then each layer is summarized using natural language technique. The output of the first page of a web site, as in the case of the TOC approach, is still as table of headings. This approach relies heavily on the accuracy of the extraction of the web page structure, but is easy to browse for the user.

5. Web Content Delivery Issues

Web document analysis techniques put additional pressure on the web delivery infrastructure. The web pages are now not directly served as before, the servers and load balancers need to decide how to make this whole process optimum. Researchers have been paying due attention to these issues. Some of the most interesting issues include streamlining web server Quality of Service (QoS) by adaptive content delivery [1], maximizing resources while delivering web images [16], improving server overload behavior [2], resource optimized delivery of web images [48] and quality aware transcoding ([15],[17],[18],[19]).

6. Commercial Web Summarizer

This review will be incomplete without mentioning the web summarizers that are available commercially. They include Copernic® (http://www.copernic.com/), Sinope® (http://www.sinope.nl/en/sinope/index.html), Subject Search (http://www.kryltech.com), Inxight (www.inxight.-com), and Pertinence (http://www.pertinence.net/index_en.html), among others. Additional information about them can be found from their web sites.

7. Further work

The study reported here is limited by the space restriction. We are working on a more extensive version of this review to include detailed discussions of the various techniques and their comparative analysis.

8. Conclusion

This paper has presented a very concise review of the research being carried out in manipulating web documents for accessing and viewing on small screen devices. It is neither exhaustive nor is aimed to be. It is hoped that this will offer a starting point of future research in this area by pointing to important works that are being conduced at the present. It is also hoped that the WDA workshop will act as a forum to offer opportunities to discuss these issue in greater detail.

References

[1] T. F. Abdelzaher and N. Bhatti. Web Server QoS Management by Adaptive Content Delivery. Computer Networks, Elsevier, 31(11-16), 1999, pp. 1563–1577.

[2] T. F. Abdelzaher and N. Bhatti. Web content adaptation to improve server overload behavior. Proceedings of the 8th International World Wide Web Conference, Toronto, Canada, 1999, pp. 465–499.

[3] H. Alam, R. Hartono. A. Kumar, Y. Tarnikova, A. Rahman and C. Wilcox. Web Page Summarization for Handheld Devices: A Natural Language Approach. Proceedings of 7th International Conference on Document Analysis and Recognition (ICDAR'03). In press.

[4] A. Antonacopoulos, D. Karatzas, and J. O. Lopez. Accessing textual information embedded in internet images. Proceedings of the SPIE Internet Imaging Conference, San Jose, California, U.S.A., 24-26 January 2001, pp. 198–205.

[5] C. Asakawa. Transcoding System for the Non-Visual Web Access - Annotation-based Transcoding. Technology and Persons with Disabilities Conference, 2001.

[6] A. Berger and V. Mittal. OCELOT: A System for Summarizing Web Pages. Proceedings of the 23rd Annual Conference on Research and Development in Information Retrieval (ACM SIGIR), pp. 144-151, 2000.

[7] H. Bharadvaj, A. Joshi, and S. Auephanwiriyakul. An Active Transcoding Proxy to Support Mobile Web Access. The 17th IEEE Symposium on Reliable Distributed Systems October 20 - 23, West Lafayette, Indiana, 1998.

[8] T. Bickmore, A. Girgensohn and J. W. Sullivan. Web page filtering and re-authoring for mobile users. The Computer Journal, 42(6), Oxford University Journal, 1999, pp. 534–546.

[9] T. Bickmore, and B. Schilit. Digestor: Device-Independent Access To The World Wide Web. Proceedings of the 6th International World Wide Web Conference, Santa Clara, California, U.S.A., pp. 655-663, April, 1997.

[10] M. Brown, S. Glinski and B. Schmult. Web page analysis for voice browsing. First International Workshop on Web Document Analysis (WDA2001), Seattle, Washington, USA, September 8, 2001.

[11] O. Buyukkokten, H. Garcia-Molina and A. Paepcke. Text Summarization of Web pages on Handheld Devices. Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL), Pittsburgh, PA, U.S.A., 2-7 June, 2001.

[12] O. Buyukkokten, H. Garcia-Molina and A. Paepcke. Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. The 10th International WWW Conference (WWW10). Hong Kong, China - May 1-5, 2001.

[13] O. Buyukkokten, H. Garcia-Molina and A. Paepcke. Accordion Summarization for End-Game Browsing on PDAs and Cellular Phones. Human-Computer Interaction Conference 2001 (CHI 2001). Seattle, Washington - 31 March-5 April, 2001.

[14] O. Buyukkokten, H. Garcia-Molina, A. Paepcke and T. Winograd. Power Browser: Efficient Web Browsing for PDAs. Human-Computer Interaction Conference 2000 (CHI 2000). The Hague, The Netherlands - April 1-6, 2000.

[15] S. Chandra, C. Ellis and A. Vahdat. Differentiated Multimedia Web Services Using Quality Aware Transcoding. IEEE Conference on Computer Communications, InfoCom 2000, Tel-Aviv, Israel, March 26 - 30, 2000.

[16] S. Chandra, A. Gehani, C. Ellis and A. Vahdat. Transcoding Characteristics of Web Images, Multimedia Computing and Networking (MMCN'01), San Jose, CA, SPIE - The International Society of Optical Engineering., M. Kienzle and W. Feng (eds), pages 135-149, vol. 4312. 2001.

[17] S. Chandra, C. Ellis and A. Vahdat. Application-Level Differentiated Multimedia Web Services Using Quality Aware Transcoding. IEEE Journal on Selected Areas in Communications

(JSAC)- Special Issue on QOS in the Internet, Vol 18, No 12, Dec 2000.

[18] S. Chandra, C. Ellis, and A. Vahdat. Differentiated Multimedia Web Services Using Quality Aware Transcoding, Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer And Communications Societies, INFOCOM, March 2000.

[19] S. Chandra, C. Ellis, and A. Vahdat. Multimedia Web Services for Mobile Clients Using Quality Aware Transcoding. Proceedings of the Second ACM/IEEE International Workshop on Wireless and Mobile Multimedia, August 1999.

[20] H.-H. Chen, S.-C. Tsai, and J.-H. Tsai. Mining tables from large scale html texts. In Proc. 18th International Conference on Computational Linguistics, Saabrucken, Germany, July 2000.

[21] S. Corston-Oliver. Text compaction for display very small screens. Proceedings of Automatic Summarization Workshop, the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL), Pittsburgh, PA, U.S.A., 2-7 June, 2001.

[22] J. Delort, B. Bouchon-Meunier, and M. Rifqi. Web Document Summarization by Context. The Twelfth International World Wide Web Conference, Budapest, Hungary, 20-24 May 2003.

[23] M. Hori, R. Mohan, H. Maruyama and S. Singhal. Annotation of Web Content for Transcoding. E3C Note. http://www.w3.org/TR/annot.

[24] M. Hori, G. Kondoh1, K. Ono1, S. Hirose1, and S. Singhal. Annotation-Based Web Content Transcoding. 9th International World Wide World Conference, Amsterdam, May 15 - 19, 2000.

[25] M. Hori, Annotation of Web Content for Transcoding. Semantics for the Web - Dagstuhl Seminar. Schloss Dagstuhl International Conference And Research Center For Computer Science, 2000. http://www.semanticweb.org/events/dagstuhl-2000/mhori.pdf.

[26] J. Hu and A. Bagga. Categorizing images in web documents. Proceedings of 10th Document Recognition and Retrieval Conference, Electronic Imaging Conference Series, 22 - 24 January, Santa Clara, California, USA, SPIE Vol. 5010, pages, 136-143, 2003.

[27] M. Hurst. Layout and language: Challenges for table understanding on the web. Proceedings of the 1st International Workshop on Web Document Analysis, pages 27-30, Seattle, WA, USA, September 2001.

[28] Y. Hwang, J. Kim and E. Seo. Structure-Aware Web Transcoding for Mobile Devices. IEEE Internet Computing. In press.

[29] Y. Hwang, E. Seo and J. Kim. WebAlchemist: A Structure-Aware Web Transcoding System for Mobile Devices. Proc. Mobile Search Workshop, Honolulu, Hawaii, May 2002.

[30] Y. Hwang, C. Jung, J. Kim, and S. Chung. WebAlchemist: A Web Transcoding System for Mobile Web Access in Handheld Devices. Proceedings of the Mobile Computing Data Management, Denver, Colorado, August 2001.

[31] M. Jones, G. Marsden, N. Mohd-Nasir, and G. Buchanan. A site based outliner for small screen Web access. Proceedings of the 8th World Wide Web conference, pages 156--157, 1999.

[32] O. Kaljuvee, O. Buyukkokten, H. Garcia-Molina and A. Paepcke. Efficient Web Form Entry on PDAs. The 10th International WWW Conference (WWW10). Hong Kong, China - May 1-5, 2001.

[33] H. Kanis and P. Vrieze, Course material for of Telematics, offered by Leo Remijn at Tilburg University in the fall of 2000. http://www.niii.kun.nl/~pauldv/telematics/ind-ex2.php.

[34] T. Kanungo, C. H. Lee and R. Bradford. What fraction of images on the web contain text?. Proceedings of the Web Document Analysis Workshop (WDA01), Seattle, USA, 8 September, 2001, pp. 43-46. Online proc. at http://www.csc.liv.ac.uk/~wda2001/.

[35] M. Kylänpää and I. Heino. Personalized Transcoding Proxy - an Approach to Mobile Web Access. The 9th International World Wide Web Conference, Amsterdam, May 15 - 19, 2000.

[36] S. Lok and M. Kan. Employing Natural Language Summarization and Automated Layout for Effective Presentation and Navigation of Information Retrieval Results. The 12th International World Wide Web Conference, Budapest, Hungary, 20-24 May 2003.

[37] D. Lopresti and J. Zhou. Locating and recognizing text in WWW images. Information Retrieval, Kluwer, 2, 2000, pp. 177–206.

[38] S. Maes and T. Raman, Position paper for the W3C/WAP Workshop on the Multi-modal Web, September, 2000. http://www.w3.org/2000/09/Papers/IBM.html.

[39] K. Nagao, Y. Shirai, and K. Squire. Semantic Annotation and Transcoding: Making Web Content More Accessible. IEEE Multimedia, Vol. 8, No. 2, pages 69-81, 2001.

[40] K. Nagao. Semantic Transcoding: Making the World Wide Web More Understandable and Usable with External Annotations. International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet (SSGRR), 2000.

[41] G. Penn, J. Hu, H. Luo, and R. McDonald. Flexible web document analysis for delivery to narrow-bandwidth devices. In Proc. 6th International Conference on Document Analysis and Recognition (ICDAR01), pages 1074-1078, Seattle, WA, USA, September 2001.

[42] A. F. R. Rahman, H. Alam and R. Hartono. Content Extraction from HTML Documents. Proceedings of the Web Document Analysis Workshop (WDA01), Seattle, USA, 8 September, 2001, pp. 7-10. Online Proc. at http://www.csc.liv.ac.uk/~wda2001/.

[43] A. F. R. Rahman, H. Alam and R. Hartono. Understanding the Flow of Content in Summarizing HTML Documents. Proceedings of Document Layout Interpretation and its Applications Workshop (DLIA01), Seattle, USA, 9 September, 2001. Online proc. at http://www.science.uva.nl/events/dlia2001/program/index.html.

[44] A. F. R. Rahman, H. Alam, R. Hartono and K. Ariyoshi. Automatic Summarization of Web Content to Smaller Display Devices. Proceedings of the 6th Document Analysis and Recognition Conference (ICDAR01), Seattle, USA, 10-13 September, 2001, pp. 1064–1068.

[45] A. Rahman and H. Alam. Challenges in Web Document Summarization: Some Myths and Reality. Proceedings of the Document Recognition and Retrieval IX Conference, Electronic Imaging Conference, Santa Clara, California, U.S.A., 21-22 January, SPIE 4670-27, 2002.

[46] H. Takagi and C. Asakawa. Web Content Transcoding For Voice Output. Technology And Persons With Disabilities Conference 2002.

[47] Y Wang and J. Hu. A machine learning based approach for table detection on the web. Proceedings of the 11th World Wide Web Conference (WWW2002), Hawaii, U.S.A., 7-11 May, 2002, pp 242–250.

[48] Y. Wu, D. Lepresti. Resource-optimized delivery of web images to small-screen devices. Proceedings of 10th Document Recognition and Retrieval Conference, Electronic Imaging Conference Series, 22 - 24 January, Santa Clara, California, USA, SPIE Vol. 5010, pages, 144-156, 2003.

[49] C. Yang and F. Wang. Automatic Summarization for Financial News Delivery on Mobile Devices. The Twelfth International World Wide Web Conference, Budapest, Hungary, 20-24 May 2003. In press.

[50] H. Zhang. Adaptive content delivery: A new research in media computing. Proceedings of Multimedia Data Storage, Retrieval, Integration and Applications Workshop (MDSRIA), Hong Kong, January 13-15, 2000.

[51] H. Zhang, J. Chen and Y. Yang. Adaptive delivery of HTML contents. 9th World Wide Web Conference, Amsterdam, Netherlands, May 15-19, 2000, pp. 284–289.

[52] H. Zhang, J. Chen and Y. Yang. An adaptive web content delivery system. Proceedings of International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2000), Trento, Italy, 28-30 August, 2000. Online Proc. at http://ah2000.itc.it/.