Web Content Adaptation System

May H. Riadh
Assistant Professor
Zarqa University, Zarqa, Jordon

Akram M. Othman
Assistant Professor
Amman Arab University for graduate studies
Amman, Jordon

ABSTRACT

In this paper we present a Web Content Adaptation System for mobile devices. The system enables the presentation of Web content by considering the problem of small screen display of mobile computing devices, also independent-device access to web content is considered.

The focus has mainly been on the adaptation of HTML web page content to make it viewable on mobile devices, constraint that no server-side content adjustments are assumed. The adaptation is done by using the re-authoring technique started by parsing the HTML web page and converting it to tree structure. This conversion will separate presentation from content which will be more efficient in dealing with content, then converting to XML document that is a well structured document.

The result is a device independent user interface that could be shown on any device. The output shows TOC that consists of list of hyperlinks, each either the header of the web page or a title of a paragraph or using the first sentence elision as hyperlink and a link to image that will be resized to fit on mobile screen.

A major advantage of this adaptation is to deliver content with multiple versions and XML/XSL transformations to a number of mobile devices and save time and power by eliminate scrolling vertically and horizontally the page content.

General Terms

Web adaptation system, mobile content.

Keywords

Content Adaptation, Re-authoring technique, Computing Device, XML, XSL

1. INTRODUCTION

With the rapid development of wireless communication technology, many users are accessing the internet from mobile appliances, such as notebooks, PDAs, and cellular phones. Many emerging computation paradigms, such as pervasive computing, have been proposed to embrace this blooming portable computation trend. However, mobile devices have various hardware limitations, such as CPU speed, power, memory, and image resolutions. They are also restricted in software support, such as operating system, installed programs, real-time processing capability, and rendering functionality. These ad hoc limitations have become barriers in human-computer interaction. Especially, current internet contents, such as web pages and images, are mainly in the HTML format designed for desktop computers. Without any modification, it is hard to render them properly in most mobile devices [1].

Computing devices such as Personal Digital Assistants (PDAs) and mobile phones have been increasingly used and getting more powerful every day. Although the latest PDAs are even able to display frames, it is still important to adapt the content for these devices in order to provide a satisfactory surfing experience for users. Web content access will not only have to support mobile access, but will also have to deal with other forms of web access such as voice interfaces. [2]

The common assumption was that a web site would always be accessed by a browser found on a personal computer or a laptop. Recent developments in mobile computing software and hardware not only have changed this view, but have also increased the importance of device-independent access to Web content: The ability to access web sites using a wide variety of web devices. [3, 4]

Web content and applications should be generated or adapted for a better user experience. Device independence principles are independent of any specific markup language, authoring style, or adaptation process. So HTML is not a device independent markup language due to its mixture elements of content and presentation. A good device independent application is one where content can be specified in a unified, optimized way on many different kinds of devices. [5, 6]

One way, according to device independence principles, is to use styling languages CSS or eXtensible Stylesheet Language (XSL) to add style and presentation information to content written in XML, and then the web output will be a suitable content format for a mobile browser. [7]

2. Web Content Adaptation System

We build a Web Content Adaptation System that can be used for independent-device access to Web content taking into consideration the types of adaptation, method of adaptation, and adaptation technique. There are several ways to adapt a conventional desktop HTML page to better fit on a device; web page re-authoring taken as an example system that can be of great interest.

The concepts of HTML page parsing is introduced, extract content from the page, and convert these content to XML document, content processing and XSL stylesheet preprocessing to render the new content to be displayed on mobile devices.

To deliver adaptive Web contents on mobile devices, researchers also considered to re-author web pages, which can be done at server side, intermediate side, or client side.

1-. Re-authoring Web pages at server-side. Server-side adaptation provides the Web page author maximum control over content delivery for mobile devices.

2- Re-authoring Web pages at intermediate-side. Proxies typically apply intermediate adaptations. Today, many of web page visualization efforts fall into this category. Without changing the layout of original web pages reduced the size of images which were larger than that of mobile screens and removed media.

3- Re-authoring Web pages at client-side. A client device can use style sheets to format contents in a browser [5]. For instance, the font size of textual contents can be adjusted by users. Together with the above intermediate-side approaches, by storing user's operations with the DOM tree in a profile, the system automatically generated a DOM-tree with branches expanded or hidden based on the user's interest. [8, 9]

This will be used in our system for a device-independent web content adaptation, the system is structured as follows:-

First, a requested to web page in HTML format is made.

Second, an overview of the re-authoring system is given.

Third, the concepts of page parsing, content processing technique and XSL stylesheet pre-processing are presented and discussed.

3. SYSTEM DESCRIPTION

The adaptation system introduces and uses three techniques, page parsing, content processing technique, and render to mobile device that allow mobile users to access web content and the sizes of generated pages will be according to the characteristics of a device that is being used. These techniques overcome the problem of displaying web pages on devices with small displays and memory sizes. The adaptation of content for a web page and rendering the content to mobile devices is performed during the processing stage by the administrator which has full control over the whole system; Figure (1) shows the complete system diagram.

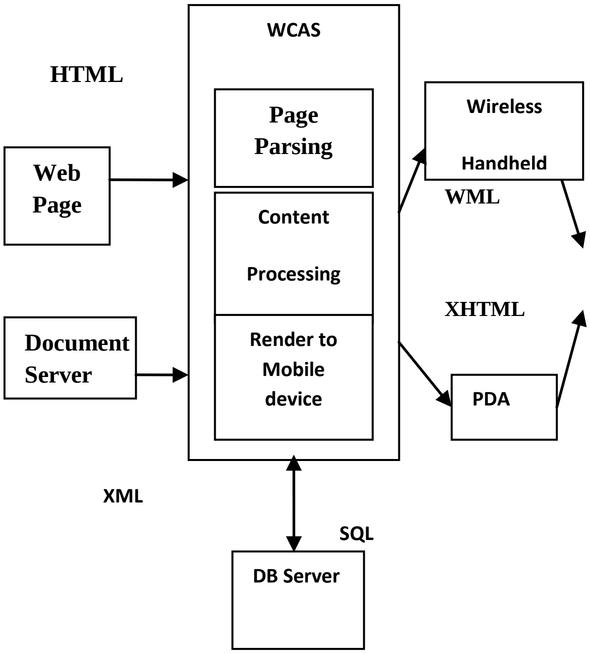


Figure 1: The complete system

4. THE ADAPTATION SYSTEM ARCHITECTURE

The Adaptation system is to analyze the web page, examine its structure, extract content, re-engineer a page, and then display the modified web page on a mobile device. This usually creates a multi-dimensional document structure from the flat two-dimensional web document.

Figure (2) shows the main parts of the adaptation system, theses are:

1- the parse engine part that parses the requested web page in HTML format and extracts the content from the web page.

2- the content re-authoring part that will convert the extracted content from the parse engine part to XML document and apply the content re-authoring processing.

3- the user interface generator that will render the adapted content from the previous part to device-independent type to display the new content on different mobile devices.

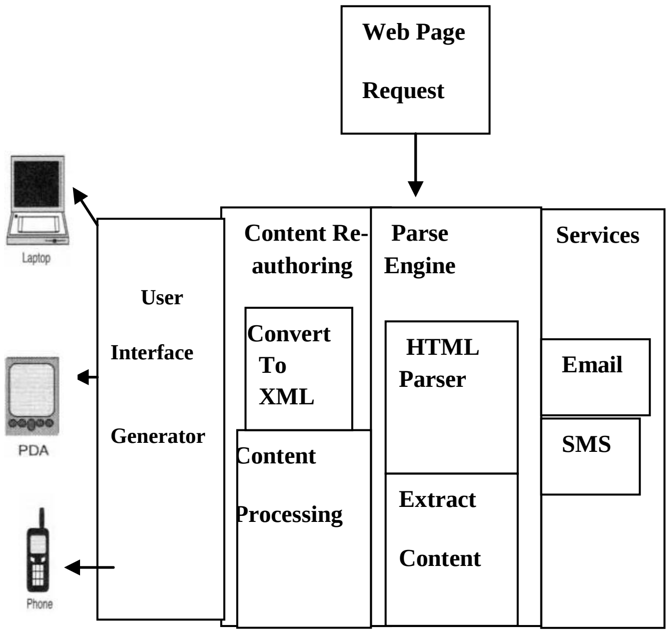


Figure 2: Adaptation system architecture

4.1 The Parse Engine

This part of the system accepts a request from the user to display a web page which is in HTML format on their mobile devices; the parse engine will analyze the web page structure and extract the content from web page as shown below:

4.1.1 HTML Parser

HTML represents a certain range of hypertext information, it is a simple markup language used to create hypertext documents that are platform independent. Since this type of pages is used to be displayed on PC’s screens that do not fit in to mobile devices screen, so it needs to be adapted.

The first step is to enter the web page to HTML parser that will parse it as:

Handles HTML frames, Frameset and Frame tags describe the frame structure in HTML, for each frame whenever the parser encounters a Frameset element, it recursively calls itself to iterate through each Frame element, inserting it into the tree and effectively flattening the structure.

Constructs an Abstract Syntax Tree (AST), which many web pages use extensively to partition pages into separate areas. Starting with the root, labels each node of the AST with a unique identifier and starts with the text part by identifying each <p> tag, i.e. paragraph, then iterates through section headers from <h1> … <h6> if there is any, then retrieves any embedded images so that their size can be determined (as necessary) as shown in Figure (3), this process is done on the client side. [10, 5]

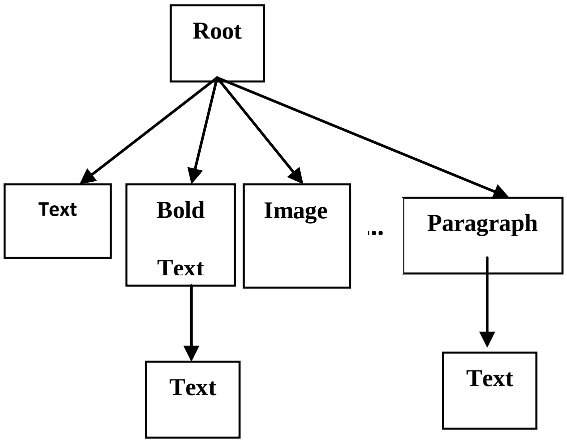


Figure 3: Parse tree of HTML Page

Parsing HTML web page is not an easy task, since most of them (pages) do not have a standard structure, so many tags, logos, tables, banners … etc are difficult to manipulate during parsing, all the previous work done in this field uses the parse engine of their proxy server and not build a parse engine to parse the HTML. Our parse algorithm is shown below.

The parse algorithm for the HTML document (Algorithm 1) is shown:

Algorithm 1

```
Parse (in Q: sentence; G: set of rules; out T: parse tree)
Begin: initialize T to be a tree with root "S";
    Loop while there are non-terminal leaves in T
      Pick a non-terminal leaf L of T;
      Choose a rule in G of the form "L -> RHS"
         If RHS is a word W
           Then choose an occurrence O of W in Q;
             Associate O with L;
               If the order of leaves in T violates the order of
                 Associated words in Q then fail
               End if;
           Else
               Make children for L corresponding to RHS
         End if;
    End loop
    If not all words in Q are attached then fail;
End.
```

4.1.2 Extracting the Content

The second important part of the parse engine is the capability to extract content from AST, and save these contents in a database file to be processed in the next step.

This involves structural analysis of the Abstract Syntax Tree (AST) by exploiting the HTML data structure from the AST.

Once segmented, content extraction can be 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.

Extraction and processing images also play a large part in this process. Classification of these images into various classes has also been attempted.

Clearly there is a need to keep content and presentation separated from the beginning of the information chain for flexibility. Information should stick to the principle of "single content, multiple accesses", i.e. it should be originated in a common form which is automatically interpretable and transformable to different presentations for a wide range of requesting clients with different capabilities. Thus the presentation level must not be fixed, but flexibly interchangeable. In addition, the content, which is the core of information, should consist of data and metadata. Metadata is the description of the actual data, additional information about what the data is representing. Data and metadata should be coupled together at the source and be available as a complete content before any style formatting. [11]

All the HTML tags are deleted, each node in the parse tree will be represented by another user defined tags that will be converted to a record and these records are stored in tables to be accessed later. The description of these records (i.e. the metadata) will be stored in temporary file, which is the structure description for XML data file. The content extraction algorithm is shown below.

4.2 The content extraction algorithm (Algorithm 2) is shown.

Algorithm 2

```
Initialize: Extract (open [start]; closed [ ] ;
While open < > [ ] do
Begin:
  Remove the next state from the left open;
    Call it X;
      If X is a goal then return [success];
        Generate all possible children of X;
         Put X in closed;
             Eliminate children of X already on either open or
           closed, as those will cause loops in the search;
        Put the remaining children of X in order of discovery on
        the left of open:
      End if;
 End.
```

4.3 Content Adaptation

This part accepts the new structured document from the content selection part, converts to XML document, applies re-authoring technique and prepares it to be presented on user-interface device

4.3.1 Converting to XML Document

A meta language is XML, a subset of SGML optimized for Web use. XML is an extensible meta language defining markup languages that describe structured data, not visual presentation.

In this part, the new data record file from the previous step is converted to XML document which is structured and well formed data, XML suits well for transmission of information, as individual documents that can be easily requested and transferred by HTTP between applications or from server to browser.

4.3.2 Content Processing for Re-authoring

The content is extracted from HTML web page and then segmented into classes in XML document will be processed more by the re-authoring technique, and using the client-side adaptation with these specifications.

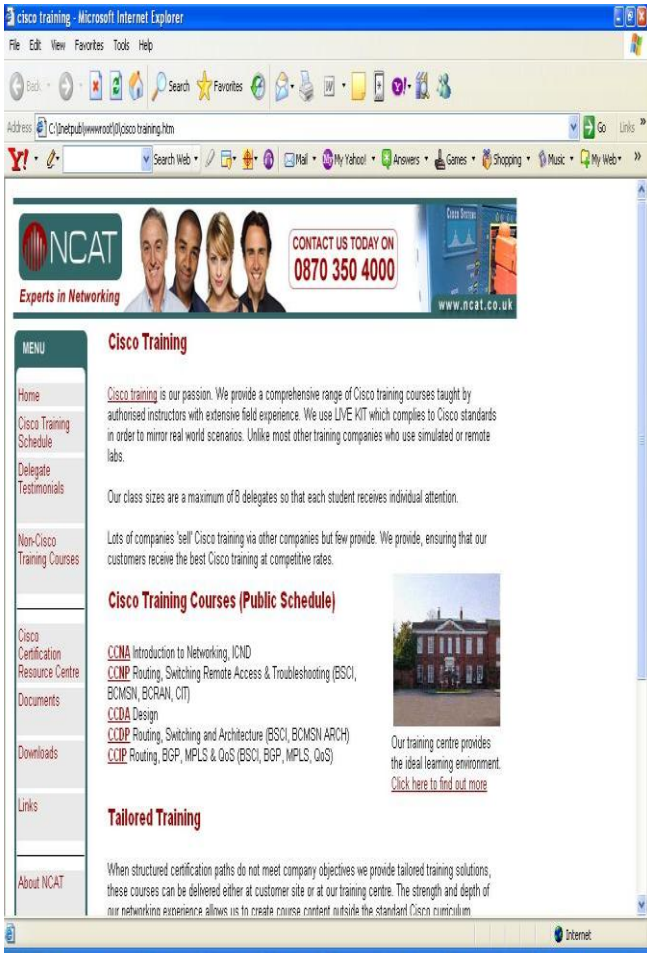
4.4 Table of Content

The document is re-created based on extracted content and a list of *heading* is created. This heading hides a hyperlink, which if selected, can load up details associated with the headline.

So the first display per web page is always a table of content (TOC) with hyperlinks. In this model, there can be any number of abstractions, but practical considerations dictate that any more than two levels are confusing for most users.

TOC techniques provide a very good method for reducing the required display size for structured documents. The contents of each section is elided from the document and the section header is converted into a hypertext link which, when selected, loads the elided content into the browser. The approaches to perform the elision works by keeping only the section headers and eliding all content, with the results looking like a table of contents for a book. Figure (4) shows the HTML page request in PC computer.

With multiple section levels (sections, subsections, sub-subsections, etc.), the approaches to performing the elision works by keeping only the section headers and eliding all content, with the results looking like a table of contents for a book.


cisco training - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address C:\Inetpub\wwwroot\0\cisco training.htm
Go Links
Search Web Mail My Yahoo! Answers Games Shopping Music My Web
NCAT
Experts in Networking
CONTACT US TODAY ON
0870 350 4000
www.ncat.co.uk
MENU
Home
Cisco Training Schedule
Delegate Testimonials
Non-Cisco Training Courses
Cisco Certification Resource Centre
Documents
Downloads
Links
About NCAT
Internet

Cisco Training

Cisco training is our passion. We provide a comprehensive range of Cisco training courses taught by authorised instructors with extensive field experience. We use LIVE KIT which complies to Cisco standards in order to mirror real world scenarios. Unlike most other training companies who use simulated or remote labs.

Our class sizes are a maximum of 8 delegates so that each student receives individual attention.

Lots of companies 'sell' Cisco training via other companies but few provide. We provide, ensuring that our customers receive the best Cisco training at competitive rates.

Cisco Training Courses (Public Schedule)

CCNA Introduction to Networking, ICND
CCNP Routing, Switching Remote Access & Troubleshooting (BSCI, BCMSN, BCRAN, CIT)
CCDA Design
CCDP Routing, Switching and Architecture (BSCI, BCMSN ARCH)
CCIP Routing, BGP, MPLS & QoS (BSCI, BGP, MPLS, QoS)

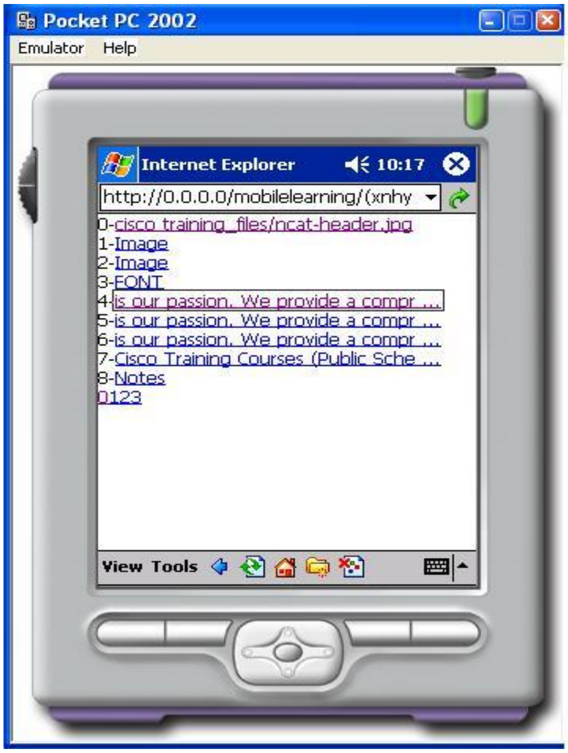
Our training centre provides the ideal learning environment.
Click here to find out more

Tailored Training

When structured certification paths do not meet company objectives we provide tailored training solutions, these courses can be delivered either at customer site or at our training centre. The strength and depth of our networking experience allows us to create course content outside the standard Cisco curriculum

Figure 4: the total HTML page

By using the re-authoring technique the TOC of this web page will displayed on a pocket PC as in Figure (5).



0-cisco training_files/ncat-header.jpg
1-Image
2-Image
3-FONT
4-is our passion. We provide a compr ...
5-is our passion. We provide a compr ...
6-is our passion. We provide a compr ...
7-Cisco Training Courses (Public Sche ...
8-Notes
0123

Figure 5: TOC of HTML page

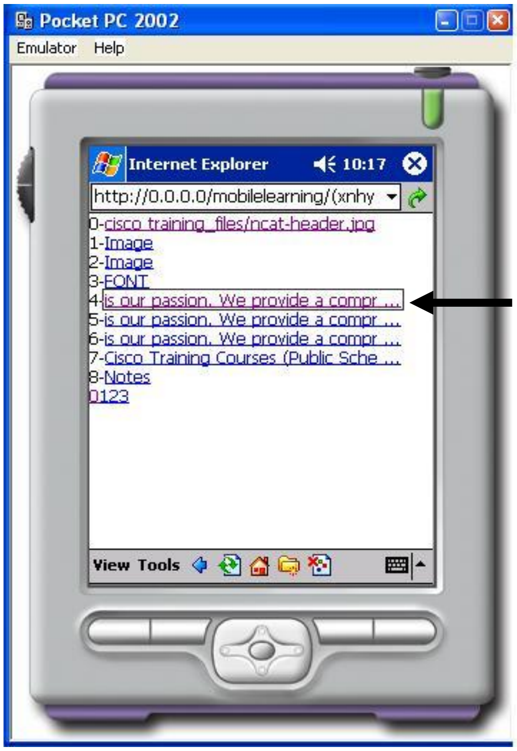
4.4.1 First Sentence Elision

Since most pages have text blocks, even when no section headers are present, first sentence elision can be a good way of reducing required screen area. In this technique, each text block is replaced with its first sentence (or phrase up to some natural break point). and this sentence is also made into a hypertext link to the original text block, as shown in Figure (6-a) the second hyperlink is first sentence elision to the paragraph shown in Figure (6-b).

The main aim of the work is to design and develop a technique to adapt standalone text content for mobile platform. As the platform is mobile so, the size of deployable and resource requirement during execution should be minimum. [12]

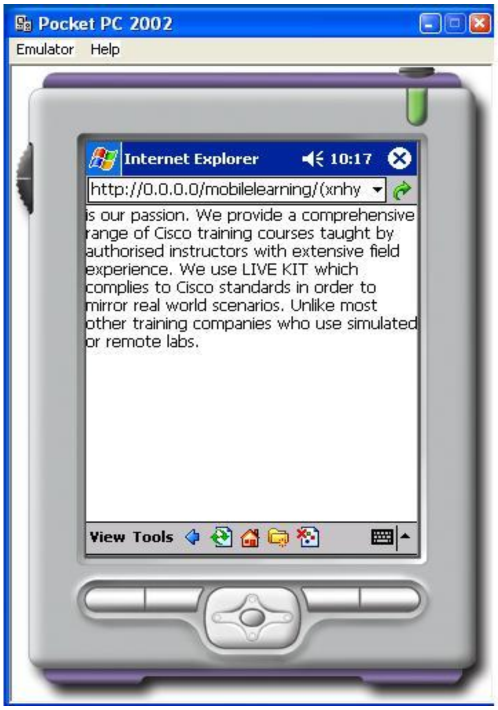
4.4.2 Image Reduction and Elision

Images present one of the most difficult problems for re-authoring, because the decision of whether to keep, reduce, or eliminate a given image should be based on an understanding of the content and role of the image on the page. However, image reduction and elision can be applied without content understanding, as long as users are provided a mechanism by which they can retrieve the original image. The approach taken is to provide a set of techniques which transform images of these types (.bmp, .gif, .jpg) in a page by pre-defined scaling factors (25%, 50%, and 75%), and making the reduced images hypertext links back to the originals [5, 9, 13], as shown in Figure (7-a) the third hyperlink of image and the image reduction shown in Figure (7-b).



0-cisco training_files/ncat-header.jpg
1-Image
2-Image
3-FONT
4-is our passion. We provide a compr ...
5-is our passion. We provide a compr ...
6-is our passion. We provide a compr ...
7-Cisco Training Courses (Public Sche ...
8-Notes
0123

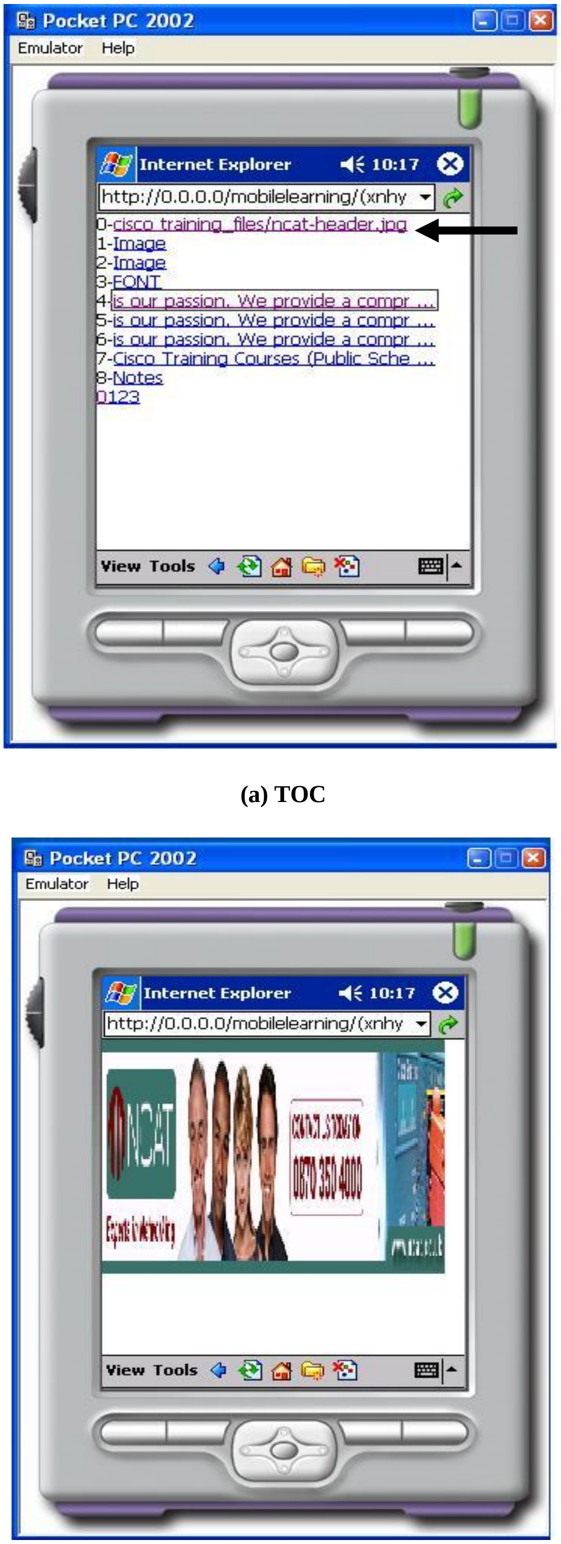
(a) TOC



is our passion. We provide a comprehensive range of Cisco training courses taught by authorised instructors with extensive field experience. We use LIVE KIT which complies to Cisco standards in order to mirror real world scenarios. Unlike most other training companies who use simulated or remote labs.

(b) First sentence elision

Figure 6: Page Content



Pocket PC 2002
Emulator Help
Internet Explorer 10:17
http://0.0.0.0/mobilelearning/(xnhy

0-cisco training_files/ncat-header.jpg
1-Image
2-Image
3-FONT
4-is our passion. We provide a compr ...
5-is our passion. We provide a compr ...
6-is our passion. We provide a compr ...
7-Cisco Training Courses (Public Sche ...
8-Notes
0123

View Tools

(a) TOC

Pocket PC 2002
Emulator Help
Internet Explorer 10:17
http://0.0.0.0/mobilelearning/(xnhy

View Tools

(b) Image display

Figure 7: Hyperlink of the Image

4.5 Understanding the Re-authoring Process

In conclusion, to perform document re-authoring two things are required: a set of re-authoring techniques, and a strategy for applying them. Of the techniques used in the manual re-authoring study, those most important and meanable are the syntactic elision techniques (section outlining, first sentence elision, image elision) and the syntactic transformation techniques (image size reduction, font size reduction). The design strategy learned during the study consists of a ranking of the transformation techniques (i.e., *try this before that*) and a set of conditions under which each transformation or combination of transformations should be applied.

4.6 User Interface Generator

The user interface generator will render the content to the user according to the capability of the mobile device which is defined in the system. The user interface generator begins by identifying the type of device making the request. It then determines the appropriate type of response markup and dispatches it to a markup handler.

4.6.1 Transformation from XML to XSLT

In an XSL transformation, an XSLT processor reads XML data and an XSLT stylesheet which is also an XML document; that is, all instructions of the language are expressed in the form of XML elements.

Based on the instructions the processor finds in the XSLT sheet, it outputs a new XML document.

XSLT is applied to the layout content to meet the device characteristics so if the connected device is some kind of WAP device, the XSLT processor will transfer the XML data into WML content, or if the device is PDA the XSLT transfer it into XHTML content, Although XSLT is designed primarily for XML-to-XML transformations, there is also support for outputting non-XML documents, such as HTML and plain text. [5, 14] XSLT is applied to the layout content to meet the device characteristics.

5. SYSTEM IMPLEMENTATION

To implement the WPCAS system, a design of a mobile application that will reflect the use of mobile device to access web page by using system content adaptation that will let the user with mobile devices access web content documents from their devices. Connecting to a network depends on the type of connection their devices accept (i.e. wireless LAN card, Bluetooth, cradle…etc) and navigate through courses that are dedicated by administrator who is responsible for specifying the web document related to the course by request HTML page, parsing to (AST), extracting the content and converting to XML document, XSL pre-processing and the new content will render to the user mobile device related to the device capability.

Using the visual studio.net 2003 tool to design the system, a visual basice.net and ASP.net as a programming language is used also a JavaScript is used as a scripting language with ASP.net, with the platform of Microsoft Windows Server 2003 with (IIS) to implement the system.

This work is done by using both the server-client adaptation, using the ASP.net with the IIS to perform all the server side work, and the JavaScript language for the client side and the visual basic.net for the CGI interface.

6. CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORK

The following conclusions reached from the implementation of the proposed adaptation system.

6.1 Conclusions

1. Simple content adaptation of HTML web pages with multiple versions and XML/XSL transformation methods for selecting the appropriate presentations and changing image sizes is done and implemented on Pocket PC.
2. Instead of building different webs for different devices, we strongly believe that the right direction is to convert and deliver the same content in different ways to different devices, by using the re-authoring technique for content adaptation.
3. Separating content from presentation which is important for each element in a given web page and generating multi dimensional customized “web” for mobile devices.
4. A table of content (TOC) from HTML web page is generated with a hyperlink that hides the content of the page.
5. First sentence elision technique is used when the text is a big paragraph and not under any heading.
6. Image resize technique is used to show the image on mobile devices, or an ALT text is displayed if the image is not shown.
7. Experiments show that in the vast majority of cases the adaptation system provides the expected results for a range of web pages that are well structured, like the pages that contain texts, images, header and tables but problems are found in the web pages that have too many tags, banners, and links.
8. The result was shown that the adapted content on the reduced display gain good impression on the users approach to retrieve web page on their mobile devices.
9. There is no perfect solution to adapt content for different mobile phones available yet with current technologies, no matter whether a phone is WAP, or i-mode with their different platforms.

6.2 Suggestions for Future work

The following is a suggestion for a future work.

1. Upgrading the system to be used for different types of HTML web pages taking in consideration all the HTML tags to be adapted to mobile devices.
2. Image format conversion: When a mobile phone preferred image format is not available on the server, an image format conversion function should be performed to convert the existing image format to a proper image format particularly for the mobile phone.
3. This research is made on text and image types of content, multimedia content could be added and its content could be adapted in the Infopyramid content selection technique.
4. Summarization of web pages is another approach of web page re-authoring techniques. In this approach, the content is not separated into separate layers; the textual part of the content is summarized using natural language techniques.
5. The service should be fully tested when the server is under heavy load and used by many simultaneous users, to see how long a user needs to wait for the navigation.

7. REFERENCES

[1] Chichang Jou, 2008 A Semantics-Based Automatic Web Content Adaptation Framework for Mobile Devices , Department of Information Management, Tamkang University, WEBIST 2007, LNBIP 8, pp. 230–242, © Springer-Verlag Berlin Heidelberg.

[2] Engin Kirda, , 2002 Engineering Device-Independent Web Services . Doctoral Thesis, Technical University of Vienna.

[3] Bill N. Schilit, Jonathan Trevor, David M. Hilbert, and Tzu Khiau Koh, 2002 Web Interaction Using Very Small Internet Devices. Intel Research, FX Palo Alto, Laboratory, Xerox Singapore, IEEE October 2002.

[4] Hassan Alam and Fuad Rahman1, 2003 Web Document Manipulation for Small Screen Devices: A Review” BCL Technologies Inc.

[5] Sudhir Dixit, and Tao Wu , 2004 Content Networking in The Mobile Internet, John Wiley &Sons,Inc.

[6] Timothy W. Bickmore , Bill N. Schilit , and FX Palo Alto Laboratory 1996 Digestor: Device-independent Access to the World Wide Web , 3400 Hillview Avenue, Bldg. 4 Palo Alto, CA 94304 USA.

[7] Rui Guan 2003 Content Adaptation on Mobile Phones , Master’s Thesis, CTI Technical University of Denmark Kgs. Lyngby.

[8] Rahul Pradhan 2001 Adaptive Multimedia Content Delivery for Scalable Web Servers, Master’s thesis Worcester polytechnic institute.

[9] Timo-Pekka Viljamaa, 2005 Types and Methods of Content Adaptation, T-110.456 Next Generation Cellular Networks.

[10] Kai Hendry 2005 Web Engineering for Mobile Devices , Master’s Thesis, University of Helsinki, Department of Computer Science.

[11] Mohd Farhan, Md Fudzee, and Jemal Abawajy, 2008 A Classification for Content Adaptation System, In Proceedings of iiWAS2008, November 24–26, Linz, Austria, pages 426-429.

[12] Moumita Majumder, Sumit Dhar, and Subrata Debbarma, November 2010 Developing and Simulating a Content Adaptation Tool for Mobile Platform, International Journal of Computer Applications (0975 – 8887), Volume 10, No .3, pages 25-27.

[13] Lin Qiao, Ling Feng, and Lizhu Zhou, 2008 Information Presentation on Mobile Devices : Techniques and Practices, Dept. of Computer Science & Technology, Tsinghua University, Beijing, China , APWeb 2008, LNCS 4976, pp. 395–406, ©_Springer-Verlag Berlin Heidelberg .

[14] Jeff Jurvis, February 2003 Issue Adapt Web Content to Any Device , Use ASP.Net Mobile Controls to generate properly formatted content for any XML language on any device, XML & web service magazine, web site: http://fawcette.com/xmlmag.

8. AUTHORS PROFILE

May H. Riadh with over a career of 26 years in professional and academic field. She work as engineer for about 18 years and a lecturer in Informatics institute for postgraduate studies/ Iraq-Baghdad.

She is now lecturing in computer science department in Zarqa university, Zarqa, Jordan.

Currently she is a member in the Iraqi Computer Society (ICS) and a member in of the Union of Arab ICT Associations (IJMA3).

Dr May has an Bsc. from university of technology in Baghdad in computer engineer, Msc from university of technology in Baghdad in computer science, and Phd in computer science from Informatics institute for postgraduate studies/ Iraq- Baghdad.

Akram. M. Othman with over a career of 30 years in the professional and academic fields. Dr. Othman has completed over 19 professional case studies covering various aspects within the field of Information Technology (IT), and has published numerous academic papers. While Dr. Akram is a renowned expert in the field of IT, he has also held senior posts in research and development in economic studies, science and technology as well as human resources capacity-building. Additionally, Dr. Othman has held several honorary positions on the boards of various IT advancement institutions, from board member to vice president.

Currently he is a member of the Iraqi Science Academy (ISA), President of the Iraqi Computer Society (ICS) and Advisor of the Union of Arab ICT Associations (IJMA3). He is now lecturing with both MIS Department and CIS Department, Amman Arab University for Graduate Studies, Amman- Jordon. .

Dr Akram Othman has an MSc from Baghdad University and doctorate in Computer Sciences from University of Technology - Iraq and a BSc in Mathematical Sciences.