웹 문서 경량화에 의한 모바일용 콘텐츠 변환 시스템

Contents Conversion System for Mobile Devices using Light-Weight Web Document

김 정 희* 권 훈** 곽 호 영***
Jeong-Hee Kim Hoon Kwon Ho-Young Kwak

요 약

본 논문은 유선용으로 작성된 웹 문서를 모바일용 단말기에서 서비스가 가능하도록 유선 콘텐츠를 모바일 콘텐츠로 변환하는데 목적을 두고 있다. 유선 콘텐츠는 일반적으로 Pop-Up 광고창, 불필요한 이미지, 유용하지 못한 링크들로 꾸며져 있어서 유선 환경에 비해 낮은 대역폭, 메모리, 스크린 크기를 갖고 있는 모바일 단말기상에 프리젠테이션이 어려울 뿐만 아니라 모바일 사용자들에게 직접 콘텐츠에 접근하는 것을 혼란스럽게 하고 있다. 그러므로 유선 웹 문서로부터 유용하고 적절한 콘텐츠를 추출하고 이를 모바일 단말기의 특성에 맞춤식으로 제공하는 요구가 대두되고 있다.

따라서 본 논문은 WAP 2.0과 여기에 채용된 콘텐츠 생성 언어인 XHTML Basic을 기반으로 한 콘텐츠 변환 시스템을 제안하였다. 제안된 시스템의 콘텐츠 변환 방식은 유선 웹 문서를 경량화한 후, 기존의 변환 방식인 필터 방식 변환 규칙을 적용하였다. 그리고 적용된 변환 규칙은 수정과 삭제가 쉽게 처리될 수 있도록 XHTML Basic의 모듈단위 기반을 사용하였으며, 또한 변환의 확장성 및 문서의 유효성을 유지하기 위하여 변환 규칙은 XSLT 기술의 XSL 문서 내에 정의하였다. 또한, WAP 1.X의 Legacy한 서비스와의 연동을 위해 CC/PP 프로파일 및 단말기 헤더 정보를 분석하는 모듈을 갖도록 시스템을 구성하였다.

Abstract

This paper aims to develop a system for converting web contents to mobile contents that can be used on mobile devices. Since web contents generally consist of pop-up ad windows, a bunch of unnecessary images and useless links, it is difficult to efficiently display them on common mobile devices that have lower bandwidth and memory, as well as much smaller screen, than the online environment. It is also troublesome for mobile device users to directly access contents. Thus, there has been a great demand for a new method for extracting useful and adequate contents from web documents, and optimizing them for use on mobile phones.

In the paper, a system based on WAP 2.0 and XHTML Basic, which is a content creation language adopted for WAP 2.0, has been suggested. The system is designed to convert web contents by using the conversion rules of the existing filtering method after making the size of web documents smaller. The adopted conversion rules use the XHTML Basic's module units so that modification and deletion can be carried out with ease. In addition, it has been defined in a XSL document written in XSLT to maintain the extensibility of conversion and the validity of documents. In order to allow it to efficiently work together with WAP 1.X's legacy services, the system has been built in a way that can have modules, which analyze information about CC/PP profiles and mobile device headers.

☞ Keyword : Content Conversion, Mobile, Light-weight, Web Document

1. Introduction

As mentioned above, it is not easy to efficiently display web contents on mobile devices, and mobile device users can not access web contents in a simple way, as they are normally made up of pop-up ad windows, as well as unnecessary images

* 정 회 원 : 제주대학교 시간강사
carina@cheju.ac.kr(제1저자)
** 준 회 원 : 제주대학교 첨단기술연구소 연구원
dreamerz@cheju.ac.kr
*** 정 회 원 : 제주대학교 통신컴퓨터공학부 교수
kwak@cheju.ac.kr
[2005/06/07 투고 - 2005/06/20 1차 심사, 2005/10/13 2차 심사 - 2005/10/28 심사완료]

and links. However, a number of applications have been recently introduced to enable users, who wish to easily access web contents with PDAs or mobile phones, to extract only useful and desired contents from web documents. One of the widely used approaches among many applications is to first remove unnecessary elements (tags, images, etc) from web documents, then simplify web document sources while maintaining the look and feel of documents by modifying certain attributes, such as font size, in a way that can become more legible or by disabling scripts, just like what WPAR[1], Webwiper[2] and Junkusters[3] do. Since these applications, however, are dependant on certain types of website design and are hard coded, inaccurate analysis results may be produced from web documents designed in new ways. Besides, unnecessary web document layouts can be still found on mobile devices[5].

Another approach is, like Opera[4], to reformat contents so that data is recognized in a way that meets the hardware constraints of mobile devices. This method, however, has not been widely adopted for presentation on mobile devices, since it does not remove unnecessary elements.

At last, there is an approach like Suhit et al.[5], which combines the two methods mentioned above. In this approach, content extraction can be controlled, since user interfaces, which allow users to extract contents by choosing from elements that form a web document, are supported. The weakness of the approach is that it is impossible to select and remove elements that have been implemented using the JAVA language as well as elements that are not part of options to choose from for diverse mobile devices.

In this paper, therefore, a WAP 2.0 & XHTML Basic (a content creation language adopted for WAP 2.0)-based content conversion system is proposed. The proposed system first remove unnecessary elements and overlapped layouts from a web documents, and makes the size of the documents smaller into a single layout format. Then, it converts the documents to the XHTML Basic format. Although the existing filtering conversion method is used for its conversion rules, the XHTML Basic's module units are adopted to facilitate the modification and deletion of the conversion rules. In addition, the conversion rules are defined in a XSL document written in XSLT to maintain the extensibility of conversion and the validity of documents. The system is also built in a way that can have modules, which analyze information about CC/PP (Composite Capability/ Preference profiles) profiles[12] and mobile device headers in order to enable it to effectively work in relation to WAP 1.X's legacy services.

2. Related Work

2.1 Conversion

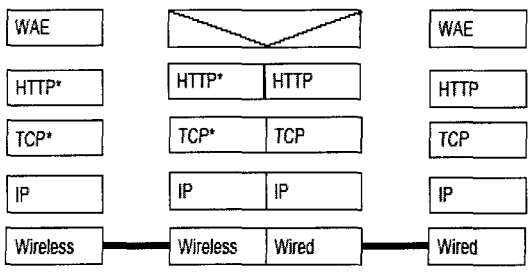
The existing studies on transcoding methods for presenting web documents on the small display screens of mobile devices have been mainly focused on HTML-based research carried out to extract summarized texts or perform conversion that involves manual procedures[6, 7]. Since the performance of mobile devices has been rapidly improved these days, studies on converting directly on mobile devices have also been conducted[8]. In summary, web document conversion can be classified into HTML-based transcoding[6], document summarization & filtering-based transcoding[9], and annotation-based transcoding[6].

The document summarization & filtering-based

transcoding[9] is a semi-automatic method that HTML documents are converted directly by system administrators, using separate filtering. This method, however, has shortcomings in that correct information can not be transferred since the contents of original document are lost and, therefore, authors's intention may be distorted. The annotation-based transcoding method[9] is a framework (IBM WebSphere) for transcoding HTML and text documents with annotations so that they can be efficiently displayed on mobile devices such as Palm OS. However, structural information like XML is hard to be transcoded with this method, and a complicated manual annotation procedure is required for transcoding.

2.2 Wireless Internet Application Protocol : WAP 2.0

WAP 2.0 can be adopted to the various types of mobile devices, and has standards that support compatibility between wire and mobile as shown in Fig.1. Especially, since it has adopted HTTP 1.1 and the XHTML technology for a protocol and markup language, wire and mobile internet-integrated services are accomplished. WAP 2.0 uses XHTML Basic in which the structures of documents and the profiles of mobile devices are well defined as its markup language[10].



〈Fig. 1〉 WAP 2.0 HTTP Proxy

2.3 XHTML Basic

Although there are currently various languages for creating mobile contents, such as WML, mHTML, cHTML, etc, one that has been adopted for the WAP 2.0 environment is XHTML Basic. It consists of the smallest sets of XHTML, and is made up of modules like images, forms, basic tables and objects for each function of elements, for web client devices that can not present all of the XHTML elements, such as mobile phones, PDAs, pagers, settop boxes, etc. Additional modules (ex, scripting, etc), which are needed for diverse user agents, can also be further added in XHTML Basic[11].

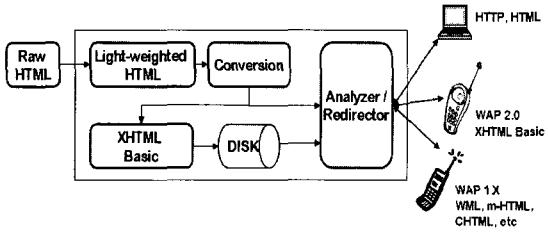
3. Our Approach

The system structure that is proposed to convert web contents to ones for mobile use is described in Fig.2. The system consists of a module that make the size of unstructured and non-standardized web documents smaller, a conversion module that convert the resized-web documents to XHTML formatted documents, and a module that transfers appropriate web documents or pages converted specially for a specific mobile device after analyzing CC/PP, that is the profile of a device, and header information, when a request is received from a WAP 1.X, HTTP or WAP 2.0-based device.

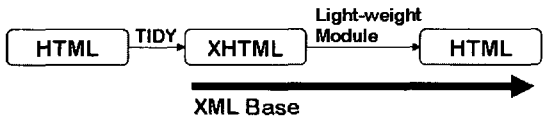
3.1 Module for web documents to lightweight

The resizing module makes the size of unstructured and non-standardized web documents smaller and restructures them as XHTML-based web documents. As shown in Fig.3, the resizing

procedure involves a preprocessing stage in which the HTML format is converted to the XHTML format using TIDY[13], which is a commercial tool. Then, the resizing module receives results from TIDY and restructures web documents into single layout structures. Once web documents are converted to the XHTML format, they can be available on mobile devices, TV and car navigators that can browse XHTML. Useful and appropriate information can also be restructured using the XML technology so that it is optimized for certain types of mobile devices.



〈Fig 2〉 System Structure by Our Approach

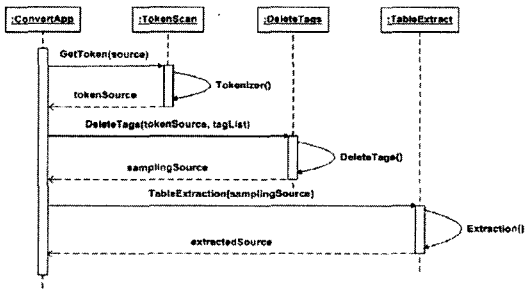


〈Fig 3〉 Steps for light-weight

Adopted rules and UML sequence diagram are presented in Table.1, fig. 4.

The resized documents mean;

- Documents reorganized in a single layout after removing overlapped layouts from web documents.
- Contents without a layout organized as a document in a single layout.
- Documents with the usage location of a layout and "<FORM>" element specified based on XHTML.
- Documents from which elements that are not supported by XHTML Basic have been removed.



〈Fig 4〉 UML sequence diagram about adopted rule

3.2 Contents Conversion Module

There are various methods for converting web contents to mobile contents for use on mobile devices. Most of them, however, carry out 'one-to-one' tag conversion using a language that creates mobile contents through filtering. And, conversion rules such as Delete, Modify and Replace are used for tag conversion. In addition, to improve the efficiency of conversion, web contents are first preprocessed into a transaction language format, then reconverted into a format for a certain device. In other words, if a web document, for example, is to be converted into a format called WML, it is first converted into a middle stage format, then converted into WML, mHTML or cHTML, rather than converting it directly into WML. This is because that different mobile devices have different characteristics, and are dependant on service environments. However, as WAP 2.0-based services have a characteristic that tries to integrate the web and mobile environment and has adopted XHTML Basic, which belongs to the HTML family, for its content creation language, integrated content

〈Table 1〉 Rules for light-weight web page

RULE (1)
IF : fore tag is (Starting Page Layout) and post tag is also (Starting Page Layout)
THEN : (1) insert the (Ending Page Layout) fore post tag
(2) extract the current position (Page Layout) to Single Table
RULE (2)
IF : fore tag is (Starting Page Layout) and post tag is (Ending Page Layout)
THEN : extract the current position (Page Layout) to Single Table
RULE (3)
IF : fore tags is (Ending Page Layout) and
post tags is ((Ending Page Layout) or (Starting Page Layout))
THEN : (1) insert the (Starting Page Layout) fore post tag
(2) extract the current position (Page Layout) to Single Table
RULE (4)
IF : contents block existed between the ((BODY) and (Starting Page Layout))
or ((Ending Page Layout) and (Ending BODY))
THEN : (1) insert the (Starting Page Layout) post (BODY) or fore (Ending BODY)
(2) collect the contents block and insert the (Ending Page Layout)
(3) extract the current contents block to Single Table
RULE (5)
IF : (Starting Page Layout) existed post (FORM components) or
(FORM components) existed post (Starting Page Layout)
THEN : (1) insert the (Starting Page Layout) fore (FORM components)
(2) delete the HTML tag except (FORM components) and contents block
(3) insert the (Ending Page Layout) and (Ending FORM components)
RULE (6)
IF : There is can be delible tags
THEN : (1) delete the tag from left angle bracket to right angle bracket
(2) preserve the contents block from next right angle bracket to left angle bracket
(3) delete the tag from next left angle bracket to right angle bracket

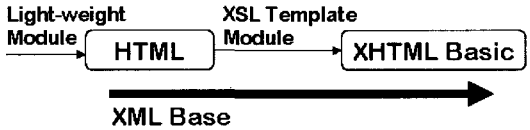
creation without distinction between web and mobile can be achieved as far as conversion rules for XHTML Basic exist, if every content becomes XHTML- oriented since HTML 4.01 has been introduced.

The content conversion method proposed in this paper, therefore, converts web documents to XHTML Basic by applying a XSL template in which conversion rules have been drawn up with over-lapped layouts and other dynamic elements that are not currently supported by XHTML Basic excluded to resized (smaller) web documents. As a result, the structure of resized web documents allows the XSL template to process a tag search algorithm (XPATH) in single layout units. In addition, extensibility and flexibility that XSL feature can be achieved. Table 2 shows conversion rules between the HTML 4.01 and XHTML Basic module.

〈Table 2〉 Modules of XHTML Basic

Preserve Module	structure, text, hypertext, image, object, link, meta information, base,
Replace of Delete module	applet, presentation, edit, bi-directional text, frame, iframe, scripting, style sheet, html specific
Options Module	forms, table

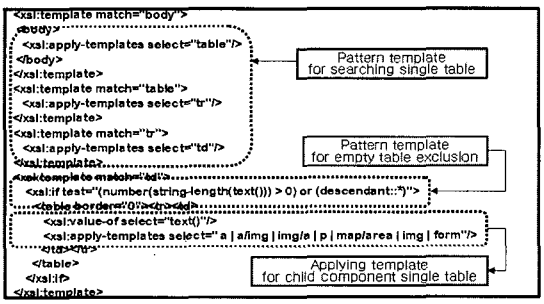
Fig. 5 and 6 show the schematic diagram presenting a conversion process respectively and a XSL document with conversion rules drawn up.



〈Fig 5〉 Steps for contents transformation

3.3 Analysis and Redirect Module

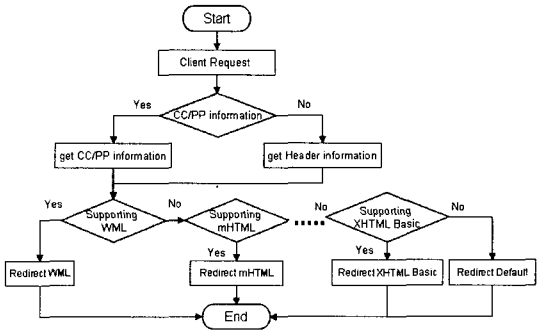
An analysis module for integrated working between WAP 1.X and legacy services processes a function for analyzing the CC/PP and UAProf (User Agent Profile) of a requesting device, when diverse requests are received from a device. The CC/PP suggested by W3C is a transfer context meta-information description model that can describe the elements of the device's transfer environment, and expresses the properties and values of mobile devices based on RDF (Resource Description Framework)[14].



〈Fig 6〉 Example of XSL template for transformation

Device header information, which is a UAProf, is characteristic information unique to each device, and is automatically forwarded to a server on request. In other words, an analysis module refers to CC/PP and UAProf in order to understand the properties of a requesting device, then forwards results to a conversion module.

A conversion module then processes a function for converting to a format that is most appropriate for the requesting device (HTML, WML, mHTML or XHTML Basic document) using results forwarded from an analysis module. Fig. 7 presents a flowchart that describes the process of grasping the properties of a requesting device.



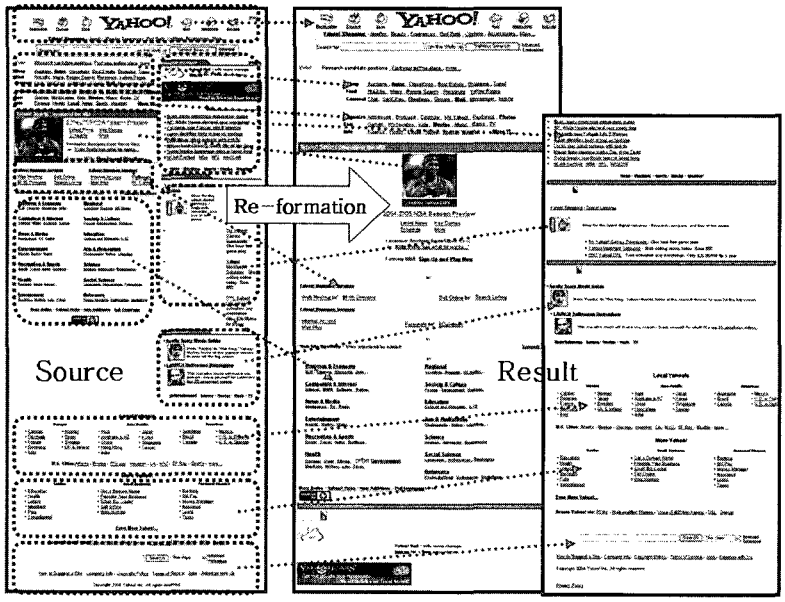
〈Fig 7〉 Flow diagram of analysis and redirect module

4. Implementation

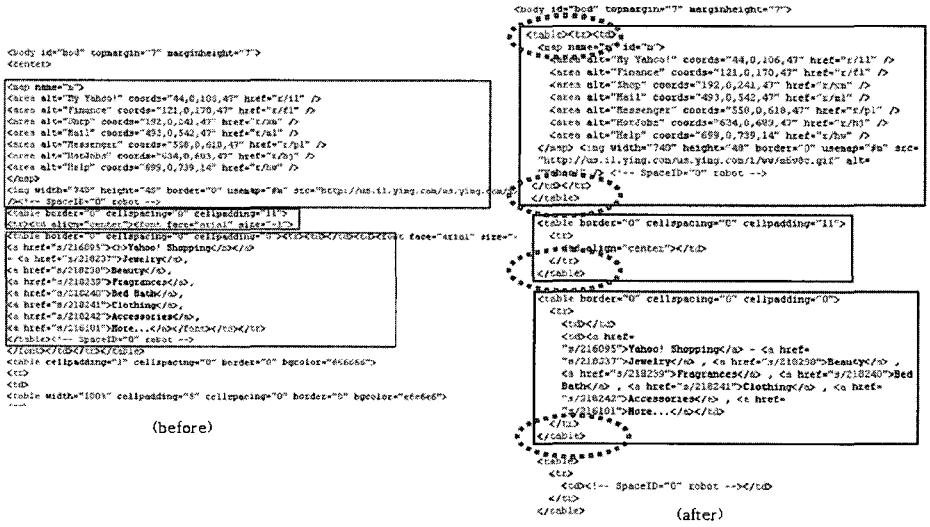
The implementation environment of the system proposed in this paper is described below.

- Operating System : Windows XP professional
- Server daemon : Apache Tomcat 4.1
- Parser : Xerces2 Java Parser 2.6.2
- Hardware : Pentium 4 CPU 2.86GHz, 1.00GB RAM
- Emulator : Openwave SDK 6.1, Nokia Mobile Browser 4.0

Fig. 8 shows the results of a web document from "yahoo.com" restructured into a smaller-sized document. All overlapped layouts were removed and the document has been restructured in single layouts. The content of the restructured document are still in the same browsing order as the original web document. In other words, resized single layouts that have been proposed maintain the same look and feel as contents that might have existed in each overlapped layout. Fig. 9 shows the comparison of HTML source codes between a original web document and a resized (smaller) document. A source code indicated with the dotted circle is the one that has been added and restructured with the proposed single layout method.



〈Fig 8〉 Light-weight web document

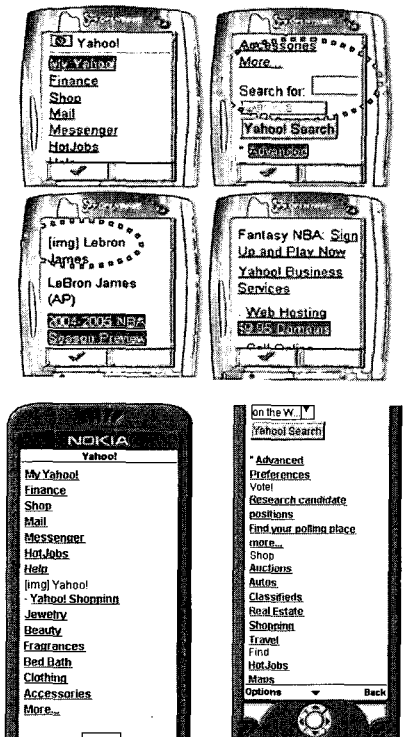


〈Fig 9〉 Light-weight web document source code

Fig. 10 shows the results of the conversion method proposed in this paper displayed on Openwave and Nokia Mobile Browser, which are mobile emulators. In converting images to XHTML Basic's module units, attribute values were displayed at the time of conversion, and "[IMG]" indicates that there were no attribute values found. This is because that conversion can be perfored in conjunction with image conversion tools. As seen from Fig. 10, text-based contents were converted by applying a XSL conversion template that has been proposed for mobile contents. In other words,

HTML web contents can be converted to elements that belong to the XHTML Basic modules.

And, a conversion distortion rate of 0.14% over original contents was found as a result of applying the implemented system to 104 information sites. The distortion was found to result from a TIDY's encoding problem over special symbols, normal-width characters and half-width characters. Thus, the XHTML Basic's module-based XSL conversion template was converted without any distortion, and conversion took 20ms less than when documents were not resized. This is one of the advantages gained from smaller-sized documents. If web documents are resized and kept smaller before requests are made by mobile devices, conversion will require less time.



〈Fig 10〉 Conversion results of yahoo.com in mobile browser emulator

5. Conclusion and Future Work

In this paper, a conversion system, which converts web documents to the WAP 2.0-based XHTML Basic language so that they can be used by the different types of clients in both web and mobile environment, has been proposed. The proposed system makes the size of unstructured and non-standardized web documents smaller, and converts resized documents to the XHTML Basic language. The system also features a function that grasps the properties of mobile devices in order to provide services to the various types of devices. Resized web documents can be made available not only on mobile devices, also on TV, car navigators, etc. Futhermore, since conversion rules can be drawn up in single layouts for converting into XHTML Basic, the XSL tag algorithm has been simplified. The conversion module is given extensibility and flexibility, which are the merits of XSL, as it uses XSL that is the XSLT technology. And, since WAP 2.0-based XHTML Basic module unit conversion rules are applied, the addition and deletion of XSL conversion rules for new mobile devices and languages have been facilitated.

Issues that require further study and problems still to be resolved are as follows:

First, WAP 2.0 and XHTML Basic have not been widely used for the implementation of mobile internet environments today. This is mainly caused by conflicting interests among mobile hardware and software manufacturers as well as those among mobile telecommunication companies. To take advantage of legacy systems that are currently available, diverse XSL template conversion rules that convert into mobile content languages such as WML, mHTML, cHTML and sHTML should be further developed. Second, in order to provide the

dynamic script elements of mobile internet contents, modules that support the affected elements should be added to XHTML Basic. After that, studies on conversion between scripts (Javascript vs WML Script) should be conducted. Third, TIDY that converts HTML to XHTML should be further improved, and a mobile XHTML browser that can display XHTML Basic on mobile devices should be developed. Fourth, a wire/wireless mobile server to which the content conversion system proposed in this paper can be applied must be implemented. Finally, contents conversion has various problems according to mobile environment. This paper emphasizes the conversion that contents distortion is not happened for reusability of web contents in mobile environment. Therefore, It has a problem that the converted contents are listed vertically long after conversion. For this reason, contents omission and condensing like clipping, customization needed for actual service.

Reference

[1] http://sourceforge.net/projects/wpar

[2] http://www.webwiper.com

[3] http://www.junkbusters.com

[4] http://www.opera.com

[5] Suhit Gupta, Gail Kaiser, David Neistadt, Peter Grimm, "Dom-based Content Extraction of HTML Document", In WWW2003 proceeding of the 12 Web conference, pp. 207-214

[6] M. Hori, G. Kondoh, K. One, S. Hirose and S. Singhal, " Annotation-Based Web Content Transcoding", 9th World Wide Web Conference, 2000

[7] T. Bickmore, A. Girgensohn and J. W. Sullivan, "Web Page Filtering and Re-Authoring for Mobile Users", The Computer Journal, Vol. 42, No. 6, pp. 534-546, 1999

[8] n. Milic-Frayling and R. sommerer, "Smart-View : Flexible Viewing of Web Page Contents", World Wide Web Conference, 2002

[9] IBM, Websphere Transcoding Publisher, http://www-3.ibm.com/software/webservers/transcoding/index.html

[10] wapforum, http://www.wapforum.org

[11] XHTML Basic, http://www.w3.org/TR/2000/REC-xhtml-basic-20001219

[12] CC/PP, http://www.w3.org/TR/2004/REC-CCPP-struct-vocab-20040115

[13] HTML Tidy Library Project, http://tidy.sourceforge.net

[14] RDF, http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210

◐ 저 자 소 개 ◑

김 정 희 (Jeong-Hee Kim)
1994년 제주대학교 정보공학과 졸업(학사)
1997년 제주대학교 대학원 정보공학과 졸업(석사)
2005년 제주대학교 대학원 정보공학과 졸업(박사)
1998년~현재 제주산업정보대학 컴퓨터정보계열 겸임교수
2002년~현재 제주대학교 시간강사
관심분야 : XML, Semantic web, Sensor Network. Mobile Computing
E-mail : carina@cheju.ac.kr

권 훈 (Hoon Kwon)
2003년 제주대학교 해양생물공학 졸업(학사)
2005년 제주대학교 대학원 컴퓨터공학과 졸업(석사)
2004년~현재 제주한라대학 시간강사
2005년~현재 제주대학교 첨단기술연구소 연구원
관심분야 : XML, Sensor Network, RFID
E-mail : dreamerz@cheju.ac.kr

곽 호 영 (Ho-Young Kwak)
1983년 홍익대학교 전자계산학과 졸업(학사)
1985년 홍익대학교 대학원 전자계산학과 졸업(석사)
1991년 홍익대학교 대학원 전자계산학과 졸업(박사)
1990년~현재 제주대학교 통신컴퓨터공학부 교수
관심분야 : 객체지향 프로그래밍, 프로그래밍 언어론, Web 응용
E-mail : kwak@cheju.ac.kr