博士學位 論文

모바일용 콘텐츠 生成을 위한 웹 文書 構造 分析과 變換

濟 州 大 學 教　大 學 院

情 報 工 學 科

金　正　熙

2004年 12月

모바일용 콘텐츠 生成을 위한
웹 文書 構造 分析과 變換

指導教授　郭 鎬 榮

金　　正　　熙

이 論文을 工學 博士學位 論文으로 提出함

2004年 12月

金 正 熙의 工學 博士學位 論文을 認准함

審 査 委 員 長 _____ 印

審　査　委　員 _____ 印

審　査　委　員 _____ 印

審　査　委　員 _____ 印

審　査　委　員 _____ 印

濟 州 大 學 教　大 學 院

2004年 12月

Analysis and Conversion of Web Document Structure for Generation of Mobile Contents

Jeong-Hee Kim
(Supervised by professor Ho-Young Kwak)

A thesis submitted in partial fulfillment of the requirement
for the degree of Doctor of Information Engineering

December 2004

This thesis has been examined and approved.

Thesis director, _____

Thesis director, _____

Thesis director, _____

Thesis director, _____

Thesis director, _____

December 2004

Department of Information Engineering

Graduate School

Cheju National University

Contents

SUMMARY

Analysis and Conversion of Web Document Structure for Generation of Mobile Contents

Jeong-Hee, Kim
Department of Information Engineering
Graduate School of Cheju National University

In this thesis, methods of analysis of web document structure supporting wired internet service in wireless environment and transformation from complex contents for wired internet to mobile contents are suggested.

In addition, analysis model and evaluation standard are developed as a test-bed for proposed two research results based on degree of user preference.

The presented analysis method of web document structure makes the web document in previous stage reformed as XHTML document based on XML. In addition it unifies layout components in overlap pages in web documents, and clarifies the places for use of related components. Also it makes the web documents lightweight by eliminating the dynamic and unnecessary components.

With the proposed analysis method of web document structure, hardware limitation of mobile terminal can be overcome. and browsing view of wired and mobile environments can be retained. In addition, XSL pattern template in the stage of contents transformation can be made simple. For reusing in various mobile devices, web documents are reformed as a light lightweight XHTML documents which is composed of single table.

Next, the proposed content transformation method makes lightweight web documents transformed to XHTML basic documents which is adopted as mobile mark-up

language, and it applies XSL transformed template document including conversion rule based on XHTML basic module to a single table which is the result of the analysis of web document structure.

Finally, it is shown that the proposed methods of content transform and conversion processing retain the consistent browsing view between wired and mobile environment, and make the XPATH complexity in conversion technique lessened quantitatively.

List of Figures

List of Tables

I. 서론

1. 연구 배경 및 목적

현재 웹(World Wide Web)(WWW)은 빠른 속도로 확산되고 있으며, 따라서 거의 모든 정보들은 웹을 통해 얻고 있는 실정이다(S. Chandra, 2000). 이러한 웹 정보들은 HTML(HyperText Markup Language)(HTML)을 통하여 웹 문서로 작성되어 웹 브라우저에 의해 해석되고, PC 모니터를 통하여 최종 사용자에게 전달되고 있으며 최근에는 무선 기술과 인터넷의 통합으로 사용자들은 휴대전화, PDA(Personal Digital Assistants), 인터넷 TV, 웹 패드 등과 같은 PC가 아닌 다양한 스크린 크기를 갖는 모바일 단말기를 통해서 인터넷에 액세스할 수 있게 되었다(B. C. Housel 등 1998).

그러나 모바일 단말기들의 디스플레이 화면의 물리적 크기는 대부분의 기존 웹 페이지가 포함하고 있는 데이터의 양을 지원하지 못하고 있으며, 또한 데이터 입력이 제한적이기 때문에 브라우저의 기능에 제약을 주고 있고(B. N. Schilit 등 2001), 서비스를 위한 모바일용 콘텐츠들은 다양한 서비스 업체들이 모바일 단말기에 특화된 언어를 사용하여 직접 제작하고 있어서 추가 비용이 들고 있다. 따라서 유선망에 접속된 PC를 대상으로 제작된 기존의 웹 콘텐츠를 다양한 디스플레이 크기의 모바일 단말기에서도 사용할 수 있도록 자동으로 변환하여 추가 투자비용 없이 유

선과 모바일 환경에서 웹 서비스를 제공할 수 있도록 하는 기술이 필요하다(B. C. Housel 등 1998),(R. Han 등 1998).

또한 웹 문서 변환 방식에서는 문서의 구조를 파악하기 위해 HTML 태그 분석으로 문서의 구조를 파악(T. W. Bickmore 등 1997), (T. W. Bickmore 등 1999), (D. W. Embley 등 1999)하고 있는데, 이와 같은 방식은 태그 중심의 분석이므로 변환 모듈을 가지고 있는 WAP(WAP Forum) 게이트웨이나 프락시, 또는 서버들은 필터링 과정의 오버헤드를 갖게 되며, 또한 변환 과정이 하드 코딩되어 있어서 유용성이 떨어지고 있다. 그리고 XHTML(XHTML) 모듈 기반 변환 과정은 웹 문서의 구조와 태그 변환 규칙을 XSL(XSL and XSLT) 템플릿으로 정의하고 수행하지만 템플릿 작성이 복잡하고 유선과 모바일 환경의 일관된 뷰를 제공하고 있지 못하다.

그러므로 본 연구에서는 유선망에 접속된 PC를 대상으로 제작된 기존의 웹 콘텐츠가 WAP 2.0(WAP 2.0) 기반의 XHTML Basic(XHTML Basic)언어로 변환되어 모바일 단말기의 소형 디스플레이에서 효율적으로 표현될 수 있도록 다음과 같이 4단계로 연구를 수행하였다.

첫 번째, 현재의 웹 콘텐츠 변환은 대부분 WAP 게이트웨이에서 자동, 또는 수동의 태그 필터링, 또는 XHTML 모듈 기반 과정으로 수행된다. 이러한 일대 일 태그 매핑에 의한 물리적·논리적 태그 필터링 과정은 변환될 웹 문서의 크기에 따라 변환 소요 시간은 WAP 게이트웨이 및 XSL 문서가 떠맡고 있다. 그리고 변환된 콘텐츠는 모바일 기기의 디스플레이 특성 차이로 인해 일관되지 않은 유선과 모바일

브라우징 뷰 등의 문제가 있으며 또한 콘텐츠 생성 언어인 XHTML Basic의 중첩 구조 테이블을 지원하지 않는 문제점들을 갖고 있다. 그러므로 콘텐츠 변환을 단순하게 하면서 유선과 모바일 간 일관된 브라우징 뷰 지원 및 중첩 테이블 제거, 그리고 다양한 모바일 단말기에서 재사용이 가능도록 경량의 웹 문서를 생성할 수 있는 웹 문서 구조 분석 방법을 제안하였다.

두 번째, 웹 구조 분석을 통하여 도출된 연구 결과인 경량의 웹 문서를 WAP 2.0 기반 XHTML Basic으로 변환하는 콘텐츠 변환 방식을 제안하였다. 이는 경량화되고 단일 테이블 구조로 재구성된 웹 문서 구조 분석의 결과에 적용될 변환 규칙이 작성된 XSL 변환 템플릿이다.

세 번째, 광범위한 모든 유선 인터넷 콘텐츠를 모바일 인터넷 콘텐츠로 변환하여 서비스하는 것은 모바일 단말기의 특성상 불가능한 일이지만 모바일 단말기를 이용한 인터넷 사용 분야를 영역별 서비스(Domain-speific)로 구분한 후 해당 서비스 영역으로의 콘텐츠 변환은 상당히 유용하며 콘텐츠 재작성의 비용도 감소시킨다. 따라서 앞의 두 단계에서 제안된 방식의 결과를 분석할 분석 모델과 성능 평가 조건을 작성하였으며, 작성 시 기준 데이터는 유선과 모바일 인터넷 사용 통계 및 이용자들의 선호도를 참고하여 상위 랭크에 해당되는 사이트를 구조 분석과 변환의 대상으로 선정했으며, 성능 평가 기준은 웹 문서 구조 분석과 콘텐츠 변환의 신뢰도, 그리고 변환 모듈의 복잡성으로 정했다.

끝으로 웹 문서 구조 분석과 변환의 구현 결과를 보였으며, 결과의 분석을 통해 문제점들로 제시된 사항들이 제안된 방식에 의해 해결됨을 정량적으로 검증하였다.

2. 연구 내용 및 논문 구성

본 연구는 유선망에 접속된 PC를 대상으로 제작된 기존의 웹 콘텐츠가 모바일 환경의 단말기인 소형 디스플레이에서도 효율적으로 표현될 수 있도록 하기 위해 웹 문서 구조 분석 방법과 이를 이용한 콘텐츠 변환 방법을 제안하였다. 웹 문서 구조 분석은 WAP 게이트웨이에서 HTML을 필터링하는데 소요되는 시간의 오버헤드, 유선과 모바일 환경에서의 일관되지 못한 뷰, WAP 2.0 기반 마크업 언어인 XHTML Basic의 중첩 테이블 미지원 등의 문제와 재사용 측면을 해결하기 위한 방안이며, 콘텐츠 변환은 웹 콘텐츠를 모바일용으로 변환하기 위한 변환 방법이며 이는 구조 분석 결과의 특징인 단일 테이블 기반 변환 기법이며 변환 규칙의 정의는 XSLT 기술의 XSL 변환 템플릿에 정의하였다. 또한 성능 평가를 위해 분석 모델 및 성능 평가 기준을 작성하였으며, 끝으로 제시된 문제점들이 제안한 방법을 사용함으로써 해결됨을 보였다.

콘텐츠 변환 방법에서 이 논문의 범주를 정리하여 Fig. 1에 나타내었다. 모바일 콘텐츠 생성 방식은 크게 전용 마크업 언어(WML, mHTML, cHTML, etc.) 기반, 기존 유선 콘텐츠에서 모바일 콘텐츠로의 변환, 그리고 단말기 중간 포맷(XML, XHTML, 또는 중간 언어)으로 생성하고 XSL 기법을 사용하는 방법들로 구분할 수 있는데 본 연구는 유선 콘텐츠에서 모바일 콘텐츠로의 변환 범주에 해당된다.

본 연구의 전체 구성 및 각 장의 내용은 다음과 같다.

2장의 내용은 선행 연구 및 관련 연구로서 무선 인터넷 기술과 웹 문서 구조 분석, 그리고 유선 콘텐츠에서 모바일 콘텐츠로의 콘텐츠 변환 방법의 필요성 및 변

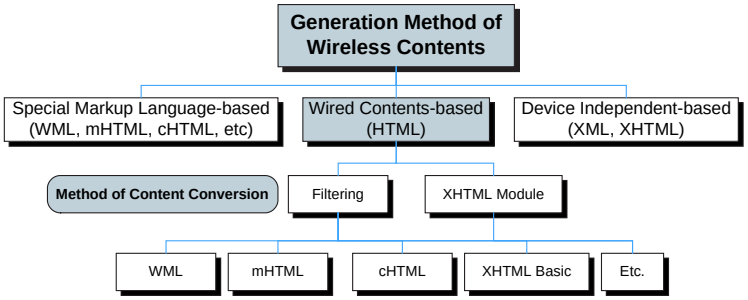


Fig. 1 Category of content conversion method considered in the thesis.

환 방식의 종류와 그 세부 기법들을 살펴보았다. 또한 콘텐츠 생성의 방향에서 현재와 향후 모바일 인터넷 마크업 언어 통합의 발전 방향도 살펴보았다.

3장에서는 웹 문서 구조 분석 방식을 제안하였다. 이는 WAP 게이트웨이 콘텐츠 변환 모듈이 웹 문서를 필터링하기 이전 단계에서 변환할 웹 문서를 XML 기반으로 문법 규칙을 잘 지키면서(Well-Formed) 단일 테이블들로 재구성된 XHTML 문서로 작성한다. 그리고 XHTML Basic에서 지원되지 않는 구성 요소들을 제거하고 문서를 단순하고 경량화 시키게 된다.

4장에서는 3장의 경량화된 XML 기반 문서를 이용하여 WAP 2.0 환경의 XHTML Basic 마크업 언어 기반 문서로 변환하는 콘텐츠 변환 방법을 제안하였다. 제안된 변환 방법은 XSLT 기술을 사용하며 변환 규칙은 XSL 문서로 정의하였다.

5장에서는 3장과 4장에서 제안된 방식의 성능 평가를 위해 분석 모델과 성능 평가 기준을 작성하였다. 광범위한 유선 콘텐츠를 모바일 콘텐츠로의 변환은 사실상 그 이용성 및 유용성이 낮다. 하지만, 모바일 인터넷 사용자의 선호도와 그 선호도에 따른 정보 사이트를 분석하면, 변환할 대상 사이트 및 정보를 선택할 수 있기 때문이다. 그리고 작성된 분석 모델과 성능 평가 기준에 따라 제안된 방법의 성능 평가 및 결과를 분석하여 그 결과 제안된 방식이 문제점으로 제시된 사항들을 해결함

을 보였다.

끝으로 6장에서는 본 연구의 결과 분석과 향후 연구 과제에 대하여 기술하였다.

II. 관련 연구

2장에서는 무선 인터넷, 웹 문서 구조 분석, 유선 콘텐츠에서 모바일 콘텐츠로의 변환 방법, 그리고 모바일 마크업 언어와 관련된 선행 연구들을 살펴보고 이를 요약하였다.

1. 무선 인터넷

1) 무선 인터넷 개요

무선 인터넷이란 말 그대로 전화선이나 전용 회선과 같은 선을 이용하지 않고 무선망(Wireless Network)을 이용하여 Anytime, Anywhere 즉 때와 장소에 관계 없이 자유롭게 인터넷상에 존재하는 정보에 접근하게 하는 유비쿼터스 컴퓨팅 환경을 의미한다(R. Kalden 등 2000), (N. K. Sharma, 2002), (M. Shi 등 2003). Fig. 2(강태규 등 2002)에는 유·무선 네트워크 구조를 보였으며, Table 1(최용길, 2003)에는 무선 인터넷의 특징을 정리하여 나타내었다.

2) 유선 인터넷과의 차이

유·무선 인터넷 환경은 Table 2(윤성일, 2002)와 같이 비교할 수 있으며, 각 구분에 따른 차이에 의해 유선 상에서 제공되는 인터넷 서비스는 모바일 환경에서 유선과 동일하게 서비스 받을 수 없다. 따라서 유선 콘텐츠를 모바일 환경에서 서비스가 되도록 적절한 변환 기술들이 필요하다(WML 1.1, 1999), (Unwired Planet, 1997).

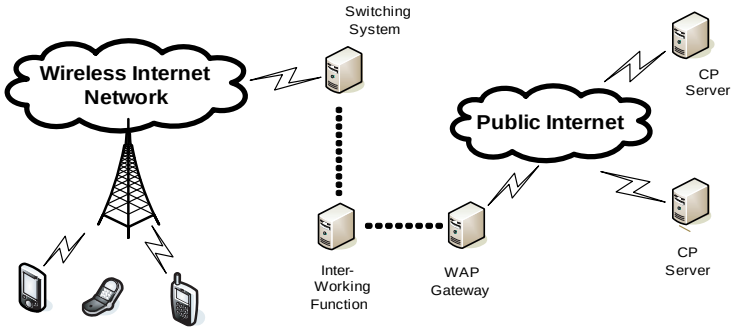


Fig. 2 Wired and wireless network structure.

3) 무선 인터넷 서비스 방식

무선 인터넷 서비스 방식으로는 크게 WAP(Wireless Application Protocol) 프로토콜 기반의 WAP 방식, 그리고 HTTP(Hyper Text Transfer Protocol) 프로토콜 기반의 ME(Mobile Explorer) 방식, 또한 i-Mode(NTT DoCoMo)(i-Mode) 방식으로 구분할 수 있다. 또한 응용 서비스로는 핸드폰 기반의 모바일 인터넷, PDA에 CDMA(Code Division Multiple Access) 칩을 탑재하여 데이터를 직접 전송 받을 수 있는 PDA 및 핸드폰 기반의 모바일 인터넷, 노트북에 CDMA 칩을 탑재하여 무선 인터넷에 직접 접속할 수 있는 노트북 및 핸드폰 기반의 모바일 인터넷으로 발전하고 있다(유영환, 2002), (이동근 등 2002).

Table 1. Feature of wireless Internet.

특징	- 언제 어디서나 접속이 가능 - 멀티미디어 서비스 가능 - 기존 전자제품 및 장비에 대한 원격 조종이 가능
대표적 서비스	- 개인 정보 관리 서비스 - 부가 정보를 제공하는 SMS 서비스 - M-Commerce 서비스
제한 사항	- 무선 이동 통신망의 낮은 대역폭 - 상대적으로 작은 디스플레이 창 - 무선망 사업자간 이해 관계 복잡

Table 2. Comparison between wired and wireless Internet environment.

구분	유선 인터넷	무선 인터넷
전송 속도	56Kbps에서 수백 Mbps	14.4Kbps에서 2Mbps(IMT 2000)
화면	640 × 480 Pixels	128 × 128 Pixels 이하
인터페이스	다양한 입력장치와 출력장치	액정화면, 소프트 버튼 등
통신 에러율	낮음	높음
프로토콜	TCP/IP	TCP/IP, WAP
콘텐츠 형태	HTML	WML, cHTML, mHTML, HDML
접근 형태	양방향	단방향(사용자 콜 필수)
애플리케이션	다양한 추가, 변경이 쉬움	한정되고 추가 및 변경 어려움
저장성	데이터 저장 용이	데이터 저장에 제한

(1) WAP 방식 서비스

WAP은 GSM(Global Standard for Mobiles), TDMA(Time Division Multiple Accwss), CDMA, CDPD(Cellular Digital Packet Data) 등을 포함한 모든 무선 네트워크에 연결할 수 있는 모바일 컴퓨터용 아키텍처로 1997년 Ericsson, Motorola, Nokia, Unwired Planet(현 Phone.com) 등 4개의 회사가 중심이 되어서 결성한 WAP 포럼에서 개발되었다. WAP은 HTTP, TCP/IP(Transmission Control Protocol/Internet Protocol) 등 기존 인터넷 표준의 프로토콜을 사용하지 않아서 HTML을 HDML 또는 WML로 변환하여 상호 교환성 및 화상 표시를 지원하는 게이트웨이가 필요하며, 현재 AT&T, BellSouth, Wireless Data, IBM 등을 포함하여 국내의 SK 텔레콤, LG 텔레콤, 신세기 통신 등 전 세계 200여개 업체들이 WAP 포럼 회원으로 가입되어 있다. (정보통신기술경영연구소, 2000), (소프트뱅크미디어, 2002), (이동근 등 2002). 따라서 Fig. 3(Yong-Woon Kim, 2001)과 같이 무선망과 기존의 유선 인터넷망의 연동을 위하여 둘 사이에 WAP 게이트웨이가 위치하며, 사용자의 단말기와 게이트웨이 사이에서는 WAP에서 정의된 프로토콜로 통신이 이루어지고 WAP 게이트웨이와 유선 인터넷망은 기존의 인터넷 통신 방식인 HTTP로 통신이

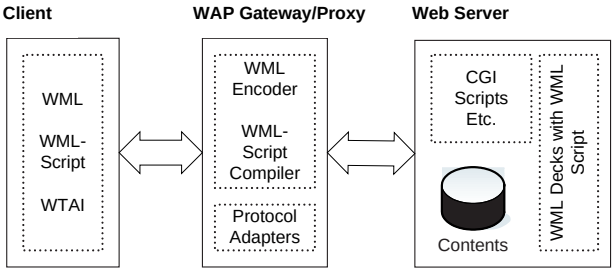


Fig. 3 Model of WAP service.

이루어지고 있다. 즉 사용자가 이러한 WAP 서비스를 받기 위해서 통신 사업자는 WAP 게이트웨이를 구현해야 하고 단말기 업체는 WML로 된 사이트를 볼 수 있는 WAP 브라우저를 제공하여야 하며 콘텐츠 업체(Contents Provider)는 서비스 될 사이트를 콘텐츠 생성 언어인 WML로 구현해야 한다(WAP 2.0-1),(박기현 등 2001).

반면 WAP 1.X의 유선망 연동 어려움으로 인해 WAP 포럼은 무선에 특화된 WML과 프로토콜을 포기하고 XHTML을 채택하여 변형시킨 WML 2.0(WML 2.0)과 SSL(Secure Sockets Layer), TCP/IP에 이르는 기존 유선 인터넷 표준을 지원하는 WAP 2.0 차세대 무선 인터넷의 국제 표준 규격을 2001년에 발표하여 향후 무선 인터넷 서비스 환경을 하나로 통합하는데 앞장서고 있다. 하지만 현재 진행 중인 상태이다(이동근 등 2002), (김신효 등 2002), (박기현 등 2004). Fig. 4(WAP 2.0)는 WAP 2.0 서비스 방식이며, Fig. 5(Yong-Woon Kim, 2001)는 WAP 1.X 프로토콜 스택이다.

(2) ME 방식 서비스

ME는 마이크로소프트사가 퀄컴과 제휴하여 제안한 방식으로 WAP 1.X 모델 방식과는 달리 유선 인터넷의 프로토콜 스택을 그대로 사용하는 현재의 인터넷

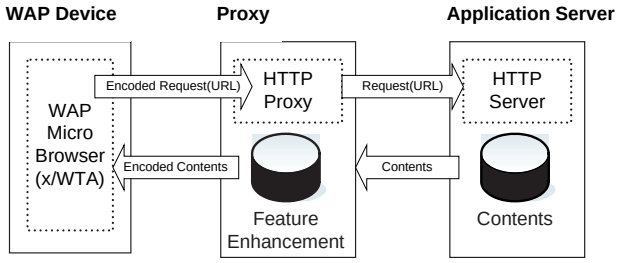


Fig. 4 Proxy model of WAP 2.0 service.

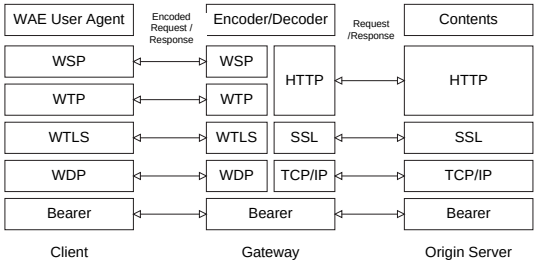


Fig. 5 Protocol stack of WAP 1.X.

표준을 지원하는 브라우저로써, 소형 단말기(Phone, PDA)등에 효율적이며 마이크로소프트의 스팅거(Stinger) 프로젝트의 초기 단계에서 향후 Windows CE 기반의 스마트 폰에 사용됨을 목적으로 하고 있다. 그리고 콘텐츠 생성 언어는 무선 환경의 제약 때문에 기존 유선망의 콘텐츠를 수용할 수가 없어 HTML을 축약한 mHTML(Microsoft)이라는 마크업 언어를 사용하여 서비스를 하고 있다(이동근 등 2002). 즉 ME 방식은 유선 환경의 HTTP 프로토콜을 이용하기 때문에 WAP 게이트웨이와 같이 프로토콜 변환과정을 거치지 않아 WAP 보다 데이터 송수신 시간이 빠르다. Fig. 6은 ME 서비스 모델이며 Fig. 7은 ME 프로토콜 스택이다(소프트뱅크미디어, 2002), (ME).

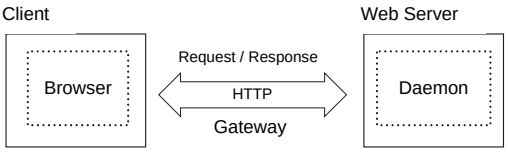


Fig. 6 Model of ME service.

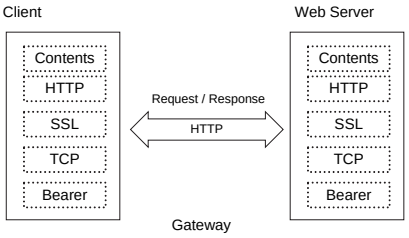


Fig. 7 Comparison of standards of i-Mode, WAP and ME technique.

(3) i-Mode(NTT-DoCoMo)

i-Mode는 1999년 일본의 이동통신 사업자인 NTT-DoCoMo(i-Mode)사가 구축한 독자적인 무선 인터넷 표준이다. 일본의 TDMA 디지털 셀룰러 방식인 PDC(Personal Digital Cellular) 위에서 패킷 통신을 지원하는 기술로서, i-Mode 대응형 단말기를 통해 전자메일과 웹 페이지 서비스를 받을 수 있다. 또한 i-Mode에서는 중간 게이트웨이를 거치지 않고 인터넷상의 콘텐츠를 i-Mode형으로 간단히 변환시킬 수 있으며, HTML 4.0(HTML 4.0)의 서브 셋인 cHTML이라는 마크업 언어를 웹 페이지 작성에 사용하면서 최대 9.6kbps의 전송속도를 지원한다. 또한 향후 WAP 2.0 기반으로 서비스가 전환이 되면 WAP 2.0에 대응한 단말기만 있으면 i-Mode용 콘텐츠와 WAP용 콘텐츠 모두를 열람할 수 있어, i-Mode 서비스는 WAP에 의한 무선 인터넷 단일 표준을 이룰 수 있을 것으로 기대하고 있다(i-Mode, 1998), (최용길, 2003). Fig. 8에서는 i-Mode, WAP, ME간의 특징들을 비교하였다(윤성일 등 2002).

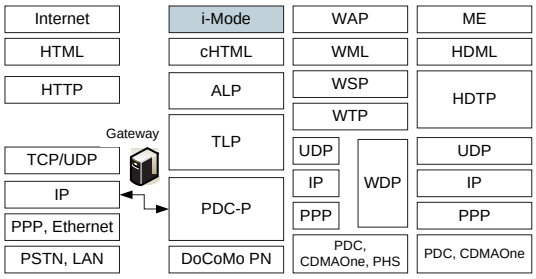


Fig. 8 Comparison of standards of i-Mode, WAP and ME technique.

4) 모바일 콘텐츠 언어

모바일 콘텐츠 제작 언어는 크게 WAP 기반의 HDML과 WML, 그리고 HTML 기반의 cHTML, mHTML, sHTML등으로 구분된다(김경아 등 2002).

WML(Wireless Markup Language)은 WAP Forum에 의해 정의되었으며, XML(김정희 등 2002), (김정희 등 2003), (고혁준 등 2004 등),(XML)에 기반을 둔 마크업 언어이다. WML은 휴대폰, PDA, 양방향 호출기와 같은 모바일 단말기에서 제한된 사용자 인터페이스 특성을 고려하고, 텍스트 기반의 콘텐츠를 제공하기 위하여 만들어졌으며, 태그 기반이면서 텍스트, 이미지, 데이터 입력을 지원한다(김규정, 2002), (WML 1.1).

HDML(Handheld Devices Markup Language)(HDML)은 Unwired Planet(UP)사에 의해 독자적으로 개발되었으며, 기존의 HTML과 달리 제한된 제원을 가진 기기상에서 데이터 표현 및 사용자 상호작용을 가능하게 한다. HDML 문서는 Deck과 Card라는 작동 개념을 사용하고 있다. 이것은 응용 프로그램이 다중의 Card로 구성된 문서를 나타내도록 한다. 논리적인 하나 혹은 다중의 Card의 집합은 하나의

Deck을 이루며 이 Card의 집합을 통해 웹을 탐색한다(Unwired Planet). 우리나라에서는 017, 019에서 기본 언어로 서비스하고 있고 많은 폰들이 HDML을 지원하지만 현재는 WML의 급속한 보급으로 그 사용이 현저히 줄어 들고 있으며, 대신 WML을 기본으로 하여 서비스하는 추세로 바뀌어 가고 있다(소프트뱅크미디어, 2002), (김규정, 2002).

cHTML(Compact HTML)은 i-Mode 환경에서 사용되며, HTML 2.0, 3.2, 그리고 4.0의 부분 집합이다. S-JIS character encoding은 반드시 사용되어져야 하고 이미지는 GIF 포맷만을 지원한다. i-Mode의 단점은 테이블과 프레임, 그리고 JAVA와 다른 스크립트 언어를 포함해서 단말기 브라우저에 표시할 수 없는 것은 지원하지 않는다(i-Mode, 1998), (이정환, 2002).

mHTML(Moblie Hypertext Markup Language)(Microsoft)은 기존의 유선 브라우저에서 사용하던 HTML 콘텐츠들을 모바일 환경에서 사용하기 위해 HTML 3.2의 부분 집합으로 규정지어진 조건에 만족하는 HTML을 의미하며, 서비스 환경은 TCP/IP 상의 표준 HTTP 프로토콜을 사용하고 HTML 3.2와 호환이 되지만 디바이스의 한계로 인해 태그들을 표현할 수 없는 경우 무시한다. 즉 HTML 4.0 태그들과는 호환되지 않는다(ME),(Phone.com).

sHTML(Small Hypertext Markup Language)(sHTML)은 AnyWeb 솔루션 마크업 언어이며, 인터넷 표준 문서인 HTML을 따른다. 기본적으로 javascript, color, applet, animation, ActiveX script, VB script 등을 지원할 수 있으나 모바일 인터넷

Table 3. View of mobile markup languages.

마크업 구분		통신사	브라우저	비고
현재 WAP 1.X	SK-WML	SK, 신세기 통신, LGT	AUR	WML 2.0 지원
	HDML	신세기 통신, LGT	UP 3.1, 4.1	WML 2.0, XHTML 지원
	mHTML	KTF	ME	WML 2.0 지원
	UP-WML	신세기 통신, LGT	UP 3.1, 4.1	WML 2.0, XHTML 지원
	cHTML	NTT DoCoMo	Access	WML 2.0 지원
	MML	J-Phone	MML	WML 2.0,XHTML 지원
향후 WAP 2.0	WML 2.0	WML 1.1 + XHTML Basic(WAP Forum)		
	XHTML Basic	Reformulation of HTML 4.0(W3C)		

단말기의 제약점들 때문에 시스템 부하를 고려하여 제외시키고, 필요한 나머지 정보들만 텍스트 위주로 제공한다(sHTML, 1999).

5) 향후 콘텐츠 제작 언어 전망

모바일 브라우저는 모바일 단말기가 웹에 접속할 때 사용되는 소프트웨어이다. 현재 모바일 브라우저 기술 시장은 Wireless Planet의 UP Browser가 지배하고 있으며, Ericsson과 Nokia, Spyglass, MS등이 자사의 브라우저를 시장에 내놓고 있다(정보통신연구 학술과제, 2002), (김학범, 2002).

Table 3(김기천, 2002)은 현재 국내·외 이동 통신 사업자들이 채용하고 있는 모바일 콘텐츠 제작 언어 현황 및 현재 WAP 1.X의 서비스가 향후 WAP 2.0 기반의 WML 2.0으로 전환될 전망을 보여주고 있다. 그러나 WAP 2.0과 WML 2.0(XHTML Basic(XHTML Basic)의 변형)이 이미 WAP 포럼에서 권고 되었지만 현재 업체간 이해 관계로 인해 WAP 2.0 기반의 서비스는 실현되고 있지 않다(Zionwap).

XHTML Basic은 XML로 된 HTML 4.0의 변형인 XHTML의 부분 Set이면서 확장성과 이식성이 부여된 차세대 콘텐츠 생성 언어로써, 휴대 전화 등 CPU 능력과 메모리가 적은 모바일 단말기용으로 태그를 줄이면서, 무선 인터넷 속도를 개선하

기 위한 통합 콘텐츠 기술언어로 WAP 포럼이 W3C와 공동으로 작성한 권고안이다(이동근 등 2002). 또한 국내 이동 통신사들이 XHTML Basic으로 콘텐츠를 서비스하게 될 때 이를 브라우징하는 브라우저는 각각 NMB, KUN, LION이 될 것으로 보고 있다(KTF 포탈 기획팀, 2003).

2. 웹 문서 구조 분석

HTML이 시각적인 포맷이기 때문에 컴퓨터로 하여금 정보를 처리하게 한다는 측면에서 한계를 가지고 있는데, 웹 콘텐츠 구조 분석은 HTML로 표현된 웹 구조를 분석하게 되면, 분석된 웹 페이지는 다양한 웹 응용 프로그램의 입력으로 사용할 수 있게 된다. 즉 웹으로부터 유용한 정보를 추출하기 위한 목적을 가지고 있는 웹 구조 분석의 결과를 유선과 모바일간 콘텐츠 변환 시 변환 모듈의 입력 데이터로 사용함으로서, 모바일 인터넷용 웹 페이지 작성 단계에서 정보를 추출하거나, 문서의 레이아웃을 참고하여 모바일 단말기용 콘텐츠 문서를 유용하게 생성할 수 있다.

관련 연구들을 살펴보면, 문서 영상 인식을 위해, 텍스트 영역의 계층적인 트리를 파싱하고 논리적인 계층 구조를 추출하여, SGML/XML에 기반을 둔 전자 문서를 생성하기 위한 구문론적인 구조분석 방법(이경호 등 2001), XML 문서로의 자동 변환을 수행하기 위해, 유사한 패턴을 갖는 HTML 문서의 구조를 분석하고, 그에 관련된 경로 정보를 인식하는 방법(오금용 등 2002), 웹 문서 변환에 있어서 컨텍스트를 제공하고 변환의 유연성을 향상시키기 위해, HTML 태그 패턴 분석뿐만 아니라 콘텐츠 정보의 속성 분석을 통하여 실시간 분석으로 웹 문서 변환에 필요한 인

덱스 정보를 추출하는 하거나, 대부분의 웹 페이지가 레이아웃을 위해 다수의 테이블을 사용하고 있다는 점에 착안하여, Table-Layout Based Structure Analysis Algorithm을 사용한 웹 페이지 내용에 해당하는 텍스트, 이미지 등을 모두 중첩된 테이블 속의 한 셀에 포함시킴으로써 원하는 레이아웃을 만들고, 부가적인 선택 정보를 위한 테이블의 일부분을 인덱스로 추출하는 방법(조수선 등 2002), Layout-Forming Tag Analysis Algorithm과 Component Grouping Algorithm을 사용하여 시각적 표현을 주도하고 있는 태그 정보에 대한 구조적인 분석 및 내용 블록 단위의 추출을 수행하여 분리된 블록들의 분류와 재구성 및 인덱스 생성 과정의 방법(신희숙 등 2002), 웹 문서의 문자열, 구조에 대한 사용자의 사전 지식을 상호 반응적으로 사용하여 HTML의 비구조적 데이터들로부터 필요한 정보 묶음을 선택하는 방법(장영건, 2003), 웹 문서를 뷰(텍스트 스트링, 순서화된 그래프)로 추상화하고, 데이터 모델과 연산자를 제공하여 HTML 페이지로부터 정보를 추출하는 방법(정재목 등 2003), 웹으로부터 유용한 정보를 추출하기 위해 전처리 단계에서는 진짜 테이블과 레이아웃 테이블을 구별하고, 여기서 구별되지 않은 테이블은 속성-값 단계에서 구문적으로 구분하며, 또한 테이블 식별을 위한 8개 규칙을 정의하여 HTML 문서로부터 테이블을 식별하는 방법들을 살펴볼 수 있다.

하지만 위 연구들은 모두 WAP 1.X 기반과 WML 1.X 기반으로 유선에서 모바일간 콘텐츠 변환 방법들을 다루고 있으며, 이경호(2001)는 원본 소스가 HTML 문서가 아니라 HTML 영상이며, 또한 오금용(2002)과 김범호(2002), 조수선(2002), 신희숙(2002), 장영건(2003), 정재목(2003)의 연구들은 웹 문서 정보를 유지하기 위해 경로 정보나 인덱스 등 사전 지식들을 필요하며, 또한 웹 문서 전체가 아니라 특정

영역 단위를 대상으로 진행되어, 유선과 모바일 장치간 일관된 브라우징 뷰 지원과 관련된 연구가 미흡하다.

3. 모바일 콘텐츠 변환

모바일 콘텐츠는 모바일 단말기를 이용하여 무선 인터넷을 사용할 때 제공 받을 수 있는 정보들을 말한다. 단순한 텍스트에서 이미지, 동영상, 게임 등 그 종류가 다양하다.

1) 모바일 콘텐츠 생성 방식

모바일 콘텐츠들은 2장 1절 4)항의 모바일 전용 마크업 언어를 사용하여 직접 제작하는 콘텐츠 저작 시점의 모바일 마크업 언어 기반 생성 환경과 유선 인터넷에서 이미 서비스되고 있는 콘텐츠들을 실시간 시점에서 모바일 콘텐츠로 변환하여 생성하는 방식으로 크게 분류할 수 있다(조수선 등 2002).

2) 모바일 콘텐츠 변환 방법

콘텐츠 생성 언어 기준으로 유선 콘텐츠를 모바일 콘텐츠로 변환하는 변환 방법을 정리하면 다음과 같이 3가지로 분류할 수 있는데, 본 연구에서는 각각 HTML에서 WML로 변환하는 변환기를 HTML2WML로, 그리고 HTML에서 WML, cHTML, mHTML, sHTML등으로 변환하는 변환기는 HTML2Multi-Languages, 또는 원본 문서가 XML 기반일 경우의 변환기는 XML/XHTML2Multi-Languages라는 용어를 사용하여 표현한다.

(1) HTML2WML Converter(일대 일 방식)

이는 HTML 웹 페이지를 WML 문서로 변환하는 방식이다. 변환 과정은 웹 문서를 필터링하여 WML 태그와 HTML 태그사이에 태그 매핑 방법을 적용하며 주로 WAP 게이트웨이에서 이와 같은 기능이 처리되도록 하는 상용화된 변환기들이 현재 제공되거나 서비스를 하고 있다.

HTML 필터링 기법의 연구들은 김환근 등 (2000), 박기현 등 (2001), 민영수 등 (2001), 강경용(2002)들이 연구에서 진행되었으며, 이 방식을 구현한 HTML 필터 구현 방법 종류들은 웹 문서의 레이아웃과 콘텐츠를 동시 변환해 주는 HTML Reformatting 방법, 레이아웃을 제외한 웹 문서 태그만 WML 태그로 변환해주는 Tag Converting 방법, HTML과 WML 태그가 혼재되어 있는 문서에서 WML 문서만을 찾는 Web Clipping 방법 등으로 정리할 수 있다.

그러나 이러한 방식들에서의 공통된 점은 WAP 1.X와 WML 1.X 기반의 필터링 방식이며 주요 단점을 정리하면 WML 모바일 단말기 전용이며, 페이지 레이아웃 구성 요소인 “DIV”에 대한 고려가 없고, 변환에 대한 변환 규칙이 하드 코딩으로 프로그램 되어 있다. 또한 텍스트 요약 및 추출에 초점을 맞추고 있어서 이미지 변환 및 변환 시 원본 문서가 변경되어 손실됨으로써 기본적인 변환 기반은 마련됐지만 완벽한 정보 전달이 미비하다. 따라서 웹 저작자의 의도가 변경되어 수작업이 필요시 되는 문제점을 발생시키는 복잡하고 기능이 많은 웹 문서를 처리하는 대안이 절실하다(T. Bickmore 등 1999), (M. Hori 등 2000), (김환근 등 2000), (박기현 등 2001), (민영수 등 2001).

(2) HTML2Multi-Languages(일대 다 방식)

HTML2WML 변환 방법에 의해서 기존의 웹 문서를 모바일 단말의 전용 언어인 WML로 변환하게 되면 WML을 사용하는 단말기에서만 그 서비스를 이용할 수 있다는 단점이 있다. 따라서 WML 언어 이외의 다양한 모바일 전용 언어를 위해 HTML2Multi-Languages는 HTML을 다양한 여러 기종의 모바일 단말기 마크업 언어로 변환하는 방식으로써, 변환하고자 하는 하나의 웹 문서를 WML, 또는 mHTML, cHTML등의 언어로 변환한다. 이 방식에서는 중간 메타언어 기법을 사용하여 원본 문서를 독립적인 중간단계로 변환하고, 여기에 변환하고자 하는 단말기의 특성 및 변환 규칙이 정의된 XSL 문서를 적용시켜 원하는 문서를 생성한다.

HTML2Multi-Languages는 강성천(2000), 윤성일 등 (2002), 양서민 등 (2004), 조승호 등 (2004)의 선행 연구들이 있으며, 이들의 변환 기술은 중간 형태의 마크업 언어(Well-Formed HTML, XHTML, XML, etc.)로 웹 문서를 변환한 후 변환된 중간 형태의 문서를 태그 필터링 방식, 또는 변환 규칙(강경용, 2002)과 선택적인 장치 프로파일 정보(CC/PP : Composite Capability/Preference Profile)(CC/PP)를 포함한 XSL 문서를 이용하여 웹 문서를 요청한 모바일 단말기의 전용 마크업 언어로 변환하고 있다. 그리고 콘텐츠 변환의 효율성 및 응답 속도를 개선하기 위해 중간 형태의 마크업 언어로 변환된 문서를 서버 또는 프락시 서버, WAP 게이트웨이에 캐시 관리자를 두어 또 다른 단말기의 요청 시 최초 캐시된 문서를 사용하도록 한다.

하지만 선행 연구들은 실제 웹 사이트에 사용되는 태그들에 대한 체계적이고 정량적인 분석이 미비하며(강성천 등 2000), 또한 웹의 CGI 및 ASP, PHP, JSP 페이

지들에 대한 고려, 웹 페이지 내의 내용 중에서 반복적으로 입력되는 사항들에 대한 자동 입력 메커니즘과 다양한 멀티미디어 데이터(SIS Color, VOD, Sound 등)의 변환 등의 연구가 요구된다(윤성일 등 2002). 특히 웹 문서에 추가 정보를 기입해야 하는 양서민 등 (2004), 조승호 등 (2004)의 연구는 모바일 단말기 특성에 최적화된 형식으로 변환은 가능하지만 최적화하기 위한 별도의 추가 정보 삽입 시간이 요구되고 있다. 무엇보다도 이러한 변환 방법들은 정량적 성능 평가와 더불어 다양한 단말기로의 변환으로 인해 무선 인터넷 서비스에 대한 기반 기술 표준 제정이 필요하다(양해술 등 2004). 또한 HTML2WML 기반과 동일하게 WAP 1.X와 WML 1.X 기반의 변환이다.

(3) XML/XHTML2Multi-Languages

XML 기반의 원본 문서를 변환하는 XML/XHTML2Multi-Languages와 관련된 선행 연구로는 김경아 등 (2002)이 있으며, 모바일 언어 상호간 변환은 이미 모바일 단말기로 서비스되고 있는 콘텐츠를 또 다른 모바일 단말기로 서비스하고자 할 때 XSL을 적용시켜 콘텐츠를 변환하는 최지원 등 (2002)등의 연구가 있다.

하지만 이들도 HTMLtoMulti-Languages와 같이 WAP 1.X와 WML 1.X 기반이며, 단점으로는 XML 기반이기 때문에 콘텐츠를 저작 시점에서 새로 작성해야 하는 문제와 단말기별 변환이기에 단말기에 최적화하기 위한 별도의 주석 작성 시간의 필요하다는데 있다.

Table 4. Tag types and example of filtering process for HTML tags.

Tag Type	Description & Example
valid	HTML Filter의 변환 처리가 가능하며 단순히 Tag Set만을 이용한 변환 수행, → , <b> → <b/>
validAttributes	HTML Filter의 변환 처리가 가능하며 valid와는 다르게 HTML 태그 내부의 특정 Attributes를 함께 변환함 <a href="x" target="y"> → <a href="x">
validAttributesData	validAttributes와 마찬가지로 HTML 태그 내부의 Attribute를 함께 이용하여 변환하지만 변환된 결과가 WML 태그가 아니고 Content Data로 사용됨 <img src="x" alt="xxx"> → [IMG] - xxx
findEndtag	해당 HTML 태그에 대해서는 Close Tag를 찾아서 Tag Role에 해당하는 Function 수행 <script> xxx </script> → 삭제 <pre>xxx</pre> → xxx(content data)
discard	변환 처리가 불가능하여 태그 삭제

3) 모바일 콘텐츠 변환 세부 기술

(1) 필터링

필터링은 필터에서 수행되는 각 태그에 대한 변환 처리 규칙을 명시한 Rule Set의 Tag Type 정보를 기반으로 한 HTML 태그들이 변환 테이블을 이용하여 태그 변환, 삭제, 치환의 작업 과정을 말한다. 일반적으로 Rule Set은 데이터베이스로 관리된다. Table 4는 Tag Type 및 HTML 태그에 대한 필터링 처리 예이다(M. Metter 등 2000), (강경용, 2002).

태그 변환은 일대 일 변환으로 하드 코딩 또는 확장성을 위해 XSL 문서를 적용하기도 한다. Table 5는 태그 변환표이다(김환근 등 2000).

Table 5. Tag conversion table.

Delete(23)	applet, base, basefont, col, colgroup, dd, dt, font, form, frame, frameset, iframe, li, link, map, noframes, noscript, object, param, scaript, style, textarea, title
Replace(41)	address → i, caption → p, hr → p, html → wml, area → (a,img) ins → u, label → b, center → (p, align="center"), menu → fieldset, cite → i, sub → small, code → pre, dl → p, th → (td, br)
Preserve	a, big, em, head, img, meta, option, pre, small, table, tr, b, br, fieldset, i, input, optgroup, p, select, string, td, uu

Table 6. Modules of XHTML Basic.

Preserve	structure, text, hypertext, image, object, meta information, link, base
Conversion or Delete	applet, presentation, edit, bi-directional text, frame, iframe, scripting, stylesheet, html specific
Option	forms, table

(2) 모듈성 기반의 태그 변환

출력 태그 집합을 XHTML 모듈로 구성하여 변환한다. XHTML 모듈은 비슷한 종류의 태그들을 묶어 놓은 태그 집합의 기본 단위로 XHTML family 문서들은 모듈의 조합에 의해 하나의 마크업 언어로 구성할 수 있어서, 기존의 태그 변환 시스템들의 태그 위주 변환 규칙과 달리 태그 기능별 모듈 기반으로 변환한다. 이 방식은 HTML 4.1(XHTML 1.0)(HTML 4.1)의 모듈과 XHTML Basic의 모듈을 이용한 모듈간의 변환이다. XHTML Basic으로의 변환 모듈은 Table 6을 기반으로 한다(XHTML Basic), (Modularization of XHTML).

(3) Well-Formed HTML 또는 XML, XHTML 기반 중간 언어 기반

선행 연구인 HTML2Multi-Languages와 XML/XHTML2Multi-Languages들이 대부분 사용하는 방법이며, 이는 특정 모바일 단말기의 마크업 언어로 변환하기

위함이 아니라 일대 다의 변환을 목적으로 할 때 웹 문서를 잘 구성된 HTML(Well-Formed HTML)이나 XML, 또는 XHTML 기반의 중간 단계 원본 콘텐츠로 생성한 후 XSL 문서를 적용하여 최종 원하는 모바일 콘텐츠를 생성하게 된다. 중간적 마크업 언어로 원본 문서를 재구성하는 이유는 보다 다양한 모바일 단말기로의 서비스를 지양하기 위함이지만, 이들은 단말기에 종속적인 변환이기 때문에 특별히 정해진 규칙과 표준이 없는 실정이다.

또한 기타 어노테이션 기법을 사용한 주석 시스템(M. Hori 등 2000), (K. Nagao 등 2001), 페이지 맵을 구성하여 관심 영역 단위 변환(송동리 등 2002), 추상화 뷰를 제공하여, 웹 문서를 텍스트 스트링 뷰, 순서 화된 그래프 뷰로 처리하는 방법(정재목 등 2003)들이 있다. 이 방식들은 사용자 중심의 관심 영역 추출로써, 전체 웹 문서를 모바일 단말기에서 브라우징 할 때 유선 콘텐츠와의 일관된 뷰를 제공하지 못하는 단점을 가진다.

(4) 마크업 언어 통합 - XHTML Basic

무선 마크업 언어에서 HDML 언어는 더 이상 지원하는 단말기가 없어서 유럽과 일본은 HDML을 지원하지 않는 WAP 기기를 사용하고 있고, 미국과 캐나다는 WML과 HDML을 지원하는 단말기를 사용하고 있다. Openwave는 HDML을 대체한 새로운 WML을 개발하였고, 그 결과 XML, XHTML등이 HDML을 대체할 것으로 예상하고 있다(H. M. Deitel 등 2002).

또한 WAP 진영에서는 WAP 1.X의 단점인 제한된 대역폭과 비신뢰성을 보완하여 2001년 WAP 2.0이 새롭게 발표하였으며, 이에 맞춰 W3C와 일본의 NTT-DoCoMo는 마크업 언어로 XHTML Basic을 채택하기에 이르렀으며, 국내의 이동통신의 표준 플랫폼으로 WIPI(Wireless Internet Platform For Intero-perability)(변시우 등 2003)가 채택되게 되면 XHTML을 사용하여 유·무선 인터넷의 통합 서비스가 가능하게 되는데(최우영 등 2003), 이러한 무선 인터넷 기술에서 콘텐츠 생성 언어의 흐름은 WAP 2.0 기반에서 개발되는 무선 마크업 언어들이 대부분 XHTML Basic 언어임을 의미하고 있다(Zionwap).

4. 관련 연구의 요약

전술한 바와 같이 무선 인터넷, 웹 문서 구조 분석, 유선 콘텐츠를 모바일 콘텐츠로의 변환, 콘텐츠 생성 언어와 관련된 선행연구들을 살펴보았다. 콘텐츠 변환은 모바일 단말기와 관계된 모바일 언어에 따라 HTML2WML, HTML2Multi-Languages, 그리고 XML/XHTML2Multi-Languages로 구분하였다. 또한 변환 기술로는 Rule Set과 태그 변환표를 참고하여 일대 일 태그 단위로 변환을 수행하는 HTML 필터링 기법, 또는 XHTML 모듈화 기법을 살펴보았다.

HTML2WML로의 콘텐츠 변환의 장점은 변환 과정이 단순하다는데 있다. 하지만 단점은 WAP 게이트웨이에게 필터링 부담을 주며 WML용 단말기만 이용할 수 있으며, 반면 HTML2Multi-Languages로의 콘텐츠 변환의 장점은 다양한 단말기로의 콘텐츠 변환이 이루어지지만 단점은 단말기에 종속적이기 때문에 태그 변환의

Table 7. Summary of related researches.

구분	종류	장점	단점
구조 분석	영상, 특징 영역	- 영상 정보 추출 가능 - 사용자별 서비스 기능	- 그래픽 기반 - Annotation 추가 삽입
콘텐츠 변환	HTML2WML	- 변환과정 단순	- WML 단말기로만 서비스 - 태그 변환의 확장성 결여 - 필터링 필요(오버헤드) - 하드 코딩
	HTML2Multi-Languages	- 다양한 단말기로 서비스	- XSL 추가 작업 필요 - 태그 변환의 확장성 결여 - 필터링 필요(오버헤드) - 하드 코딩
	XHTML 기반	- 확장성 및 기능별 변환	- 모든 모듈 고려한 작업
공 통			- 브라우징 뷰 미흡
모바일 언어	XHTML Basic	- HTML Family - 유선과 모바일 통합	

확장성 결여 및 XSL과 같은 변환 규칙이 모바일 단말기별로 작성해야하는 추가 작업이 필요하다. 또한 공통적으로 WAP 1.X와 WML 1.X 기반과 유선과 모바일간 일관된 브라우징 뷰 지원이 미흡하다. Table 7은 관련 연구들의 요약이다.

또한 모바일 언어 측면에서는 이미 모바일 서비스 환경으로 WAP 2.0이 제안되었고, 그리고 WAP 2.0에서 모바일 통합 마크업 언어로써는 XHTML Basic이 채용되어서 향후 모바일 인터넷 환경을 이끌어갈 전망이다. 이는 기존의 다양한 모바일 마크업 언어를 사용한 콘텐츠 생성을 XHTML 및 XHTML Basic 기반 생성으로 전환을 말하며 유선과 모바일 인터넷 서비스의 통합을 의미하고 있다.

따라서 본 연구에서는 관련 연구의 단점 중 콘텐츠 변환의 확장성과 이식성을 높이고, 또한 유·무선 간 브라우징 뷰를 일관되게 유지하면서 WAP 2.0의 XHTML Basic 기반으로 유선 콘텐츠를 모바일 콘텐츠로 변환하는 방법을 제안하였다.

III. 웹 문서 구조 분석

1. 웹 문서 구조 분석

1) 개요

인터넷에서 중요한 서비스 위치를 차지하고 있는 웹 서비스 정보 소스인 웹 문서는 이를 작성하는 HTML 언어의 쉬운 특징으로 인해 다양한 분야의 다양한 웹 저작자에 의해 생성되고 있으며, 기술(Description) 언어라는 특징으로 표현에만 중점을 두고 있어서 동일한 내용일지라도 웹 문서의 소스 코드는 서로 상이한 결과를 낳고 있다. 즉 표현과 내부 구조가 상이하여 프로세서가 처리하기에 어려울 뿐만이 아니라 원하는 정보를 추출하는 것 또한 어렵게 하고 있다. 그러므로 이러한 비구조적 문제점을 해결하기 위해 웹 문서를 구조적으로 분석하는 것이 필요하다.

또한 모바일 서비스에서, 현재 대부분의 웹 콘텐츠 변환은 환경에 따라 콘텐츠 변환 모듈을 WAP 게이트웨이 또는 서버에서 제공하여 변환 기능을 수행하고 있으며 변환 모듈 입장에서는 무거운 웹 문서 변환 과정에서 많은 오버헤드를 갖고 있다. 그러므로 이를 해결하기 위한 방법은 가급적이면 단순하거나 경량화 된 웹 문서를 넘겨받을 수 있으면 콘텐츠 변환 모듈 입장에서는 그 만큼 콘텐츠 변환 소요 시간을 줄일 수 있게 된다(최은정 등 2001).

따라서 3장은 웹 문서의 비구조적 문제점을 해결하면서 변환 모듈이 넘겨받을 웹 문서를 단순화되고 경량화된 웹 문서로 생성하기 위한 웹 문서 구조 분석 방법을 제안하였다. 제안된 방법은 4장에서 제안하는 콘텐츠 변환 기법의 전처리 기능에 해당되며, 따라서 콘텐츠 변환 모듈의 입력 데이터가 된다. 본 연구의 모바일용 콘텐츠 생성을 위한 웹 문서 구조 분석과 변환은 아래의 매핑 함수로 정의된다.

$$y = XTST(x) \tag{1}$$

식 (1)은 본 연구의 콘텐츠 변환을 나타내며 정리하면 다음과 같다.

- XTST는 단일 테이블 기반 XSL Template
- 정의역 $x = \{HTML\ tag\ lists\}$
- 치역 $y = \{XHTML\ Basic\ modules\}$

반면 변환 시 HTML 구성 요소에 대응되는 원소가 치역에 존재하지 않아서 HTML 구성 요소를 삭제하는 경우가 있음으로 치역의 구성 요소에 이와 관련된 원소 "ϕ" 가 추가된다. 그리고 식 (1)의 정의역 x는 아래에 정의된 식 (2)와 (3)에 의해 전처리된 결과이다.

$$x' = TIDY(html) \tag{2}$$

$$x = STETS(x') \tag{3}$$

식 (2)와 (3)은 본 연구의 웹 문서 구조 분석을 나타내며 정리하면 다음과 같다.

- TIDY는 HTML을 XHTML로 변환하는 과정
- STETS는 태그 순서에 의한 단일 테이블 추출하는 과정

Table 8. Ratio of tags for main page layout of web sites.

관련 요소	TABLE	DIV	LAYER	FRAME
사 용 률	78.8%		0%	21.2%
	95.2%	4.8%		
대상 사이트 (104개)	정보 검색 및 제공(13), 교육(19), 영화(5), 신문(11), 음악(9), 방송(4), 여행(6), 학술(1), 날씨(2), 교통(3), 부동산(2), 도서관(6), 온라인 서점(3), 게임(4), 레포츠(1), 운세(2), 관공서(1), 포탈 정보(1),쇼핑몰(2), 건강(3), 모바일(6)			

2) 웹 문서 구조 분석

경량화된 단일 테이블 구조로 웹 문서를 추출해 내기 위하여 본 연구에서는 한국 인터넷 정보센터(2003)의 모바일 서비스 이용 통계 중 현재 정보 검색과 정보 및 멀티미디어 제공 사이트를 대상으로 초기화면 페이지 레이아웃을 위해 사용되고 있는 “TABLE”, “DIV”, “LAYER”, “FRAME” 태그를 기반으로 Table 8과 같이 그 사용률을 분석하였다.

분석 결과 웹 페이지 레이아웃을 위해 가장 많이 사용되고 있는 요소는 “TABLE”과 “DIV” 태그가 78.8%, 그리고 “FRAME”을 사용한 경우가 21.2%로 나타났다. 또한 “TABLE”과 “DIV”를 사용한 경우 95.2%가 “TABLE”을 사용하고 있었으며, 4.8%는 “DIV”의 “position:absolute”를 사용한 경우로 웹 페이지내의 코드의 순서와 상관없이 그 위치를 결정할 수 있음을 나타낸다. 또한 “position:absolute”를 사용하지 않은 “DIV” 관련 요소는 하나의 논리적 단위이므로 이는 “TABLE” 사용률에 포함되었다. 따라서 웹 문서는 “TABLE”, “DIV”, “FRAME” 관련 요소를 웹 문서 코드 순서에 따라 적절하게 재구성하면 단일 테이블들로 추출할 수 있다.

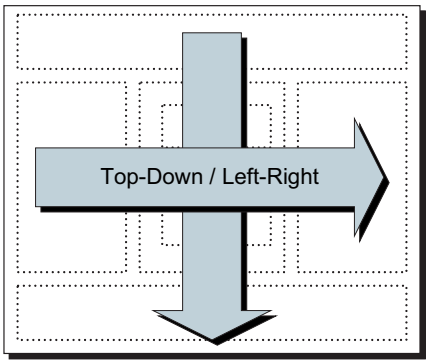


Fig. 9 Proposed concept of browsing view.

2. 제안하는 웹 문서 구조 분석 방법

1) 제안하는 단일 테이블 기반의 구조개요

웹 문서의 페이지 레이아웃 구조는 물리적·논리적인 HTML 태그 코드 흐름에 의해 단일 테이블 기반 구조로 분석해 낼 수 있다. HTML(HTML 4.0) 스펙에 의하면 “TABLE” 태그는 웹 문서 작성 시 기술되는 위치가 페이지내의 위치인 물리적인 특성을 가지며 “DIV”와 “LAYER”과 같은 태그는 좌표 값을 가질 수 있는 논리적 특성으로 페이지내의 위치는 좌표 값 또는 기술되는 위치가 될 수 있다. 또한 Frame은 포함되는 정보가 하나의 URL(Uniform Resource Locator) 정보를 가지면서 웹 브라우저의 화면을 분할시키는 기능을 제공한다. 따라서 이러한 요소들로 구성된 웹 문서를 본 연구에서 제안하는 단일 테이블 추출 방식으로 그 코드를 순서대로 분석하여 경량화시키면 모바일 브라우징이 가능하며, 특히 WAP 2.0의 XHTML Basic 언어의 문법에 준하는 중첩되지 않는 단일 테이블 기반의 중간 단계 콘텐츠가 된다. 제안하는 단일 테이블 추출은 휴리스틱 방식에 의해 웹 페이지 내의 콘텐츠를 포함하고 있는 테이블 또는 콘텐츠 블록들을 중첩되지 않은 단일 테이블들로 추출하는 방식이다.

Fig. 10 Example view web document.

결국 단일 테이블들로 재구성된 웹 문서를 모바일 브라우저에서 브라우징 시 그 기본 원리가 Fig. 9와 같이 웹 페이지의 위에서 아래로, 그리고 왼쪽에서 오른쪽으로 브라우징이 되도록 하여 유선 환경의 브라우징 뷰 순서를 모바일에서도 휴리스틱하게 일관되도록 할 수 있다. 이는 웹 문서 소스 코드의 흐름에 따라 “TABLE”, “DIV”, “LAYER” 태그를 만날 때마다 그 이전 블록까지를 하나의 테이블로 추출함으로써 가능하다. 즉 제안된 방식은 웹 문서내의 다양한 레이아웃을 위한 중첩된 테이블 내의 테이블들도 단일 테이블들의 집합으로 구성할 수 있으며, 또한 테이블을 사용하지 않고 표현된 콘텐츠 블록들까지도 하나의 단일 테이블로 구성할 수 있게 된다.

또한 메인 페이지들 중에서 프레임 구조를 사용하는 경우는 별도의 단일 테이블

로 재구성하지 않고 4장의 콘텐츠 변환 기법에서 프레임의 “SRC” 속성과 “NAME” 속성 값을 참고하여 링크로 구성된 논리적 페이지를 구성하도록 하면 복잡한 페이지 레이아웃을 경량화된 단일 테이블 기반으로 단순화 시킬 수 있다. Fig. 10은 본 연구에서 추출하고자 하는 웹 문서의 예이며 ⓐ-ⓗ는 페이지 레이아웃 테이블이고, ①-⑮는 추출하고자 하는 단위 테이블들이다.

2) 제안하는 웹 문서 구조 분석 방식의 장점

비구조적인 웹 문서를 제안하는 웹 문서 구조 방식에 의해 구조적으로 분석하게 되면 웹 문서가 XML 기반으로 변환이 됨으로, 다음과 같이 XML이 가지는 장점을 그대로 갖는다.

첫째, 웹 문서는 다양한 프리젠테이션을 위해 웹 페이지내의 페이지 레이아웃을 사용한다. 직접적인 정보를 제공하기보다는 화면 프리젠테이션 목적이 강하다. 웹 문서 구조 분석을 수행하면 웹 문서내의 레이아웃 테이블들을 제거하여 직접적인 정보 접근을 제공하게 된다.

둘째, 향후 모바일 콘텐츠용 언어인 XHTML Basic의 미 지원 요소의 문제점을 제거한다. XHTML Basic에서는 중첩 테이블을 지원하지 않음으로, 웹 문서 구조 분석을 수행하면, 콘텐츠의 왜곡 없이 중첩 테이블들의 문서에서 단일 테이블들의 집합으로 구성된 웹 문서가 된다.

셋째, 페이지 레이아웃, 중첩 테이블들의 제거로 재 생성된 웹 문서는 모바일용 디바이스에 적용할 수 있는 경량화된 문서가 되며, 이는 TV, Car Navigation 장치에서 효율적으로 브라우징 및 재사용 된다.

넷째, 다양한 단말기로의 포맷 변환을 제공하는 XSLT 템플릿 적용 시 이중 구

조 템플릿 사용을 최소화 한다. 즉 특정 패턴의 요소를 검색할 때 반복적인 XPATH 검색식을 단순화 시킨다.

끝으로 모바일 콘텐츠 생성 시 단말기에 직접적인 콘텐츠 생성 언어를 사용하지 않고 경량화된 웹 문서를 이용하여 XHTML Basic을 지원하지 않는 모바일 콘텐츠로의 재생성이 가능하다.

3) 제안 방식의 구조 분석 내용

웹 문서의 제작은 3장 1절에서 기술한 바와 같이 HTML 언어를 익힘으로써 대부분 논리적 흐름 없이 기술(Description)적으로 표현하고자 하는 내용들을 담고 있다. 이러한 페이지 작성의 별 어려움 없는 사항은 현재 유선 인터넷의 웹 정보를 모바일에서 구하고자 할 때 변환 과정을 요구하거나 또한 모바일에서의 브라우징을 어렵게 하고 있는데, 이는 앞장에서 살펴본 바와 같이 모바일 환경에서 유선용으로 작성된 콘텐츠를 재사용하고자 하기 때문이다. 따라서 본 연구에서 제안하는 구조 분석 내용은 유선보다 열악한 모바일 환경으로 유선 콘텐츠를 서비스하기 위해서 모바일 환경 또는 브라우저에서는 필요 없거나 지원되지 않는 유선 콘텐츠의 관련 요소들을 텍스트 기반으로 제거하고 단순화 시킨다. 그럼으로써 변환 모듈이 갖는 오버헤드를 감소시키고 유선과 모바일의 브라우징 뷰의 일관성을 지원하게 된다. 처리 내용들은 문법적으로 규칙이 지켜지지 않은 태그들의 사용을 규칙화하게 되고 동적인 요소(스크립트 및 스타일, 이벤트)와 이미지, 그리고 불필요한 태그 요소들을 제거하며, 태그 사용 위치가 다양한 요소들(FORM)의 사용 위치를 명시화 시키고 화면 프리젠테이션용 중첩 테이블들을 단일 테이블화 하게된다. Table 9는 제안하는 웹 문서 구조 분석의 분석 내용이며, Fig. 11, Fig. 12, Fig. 13, Fig. 14는 모

Table 9. Proposed web document structure analysis.

▶ 규칙이 지켜지지 않는 태그들의 규칙화 - 정형화되지 않은 웹 문서의 XML화(XHTML) - 자동화 툴인 TIDY를 이용하여 HTML 문서를 XHTML 문서로 변환 ▶ 태그의 종류 분석 및 구분(토큰화)을 통한 불필요 요소 제거 및 명시화 - 태그 단어의 종류 구분 및 분리 - 동적인 요소 및 불필요 요소 제거 - “<FORM>” 태그 요소를 테이블로 재구성(명시화) ▶ 화면 프리젠테이션용 중첩 테이블 단일화(웹 문서 구조 분석) - 페이지 레이아웃과 중첩 테이블들을 단일 테이블들로 추출 - 단일 테이블 기반으로 “DIV”, “LAYER” 요소들을 처리 - 테이블을 사용하지 않은 콘텐츠 블록들을 단일 테이블로재구성

Fig. 11 Example page with plenty image and page layout.

바일에서 브라우징하기 어려운 유선 콘텐츠 예이다.

3. 제안된 구조 분석의 세부 처리 방법

본 연구의 구조 분석 결과는 HTML의 XML 응용인 XHTML 문서가 되며, 이는

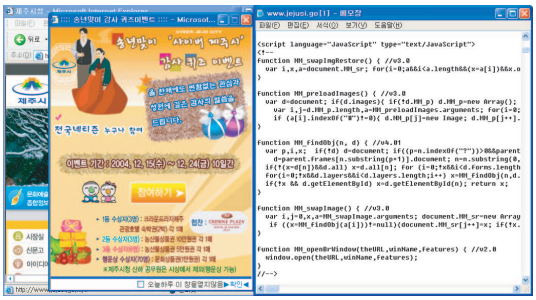


Fig. 12 Example page with dynamic script.

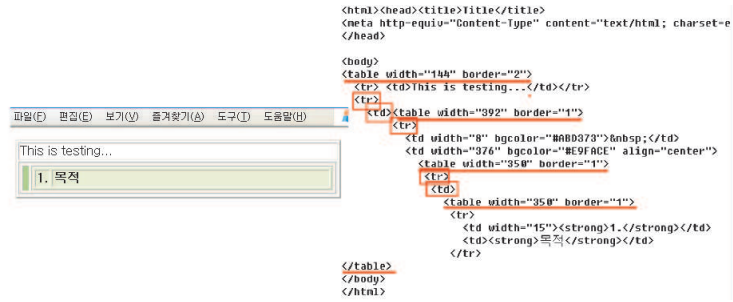


Fig. 13 Mistaken in tag usage.

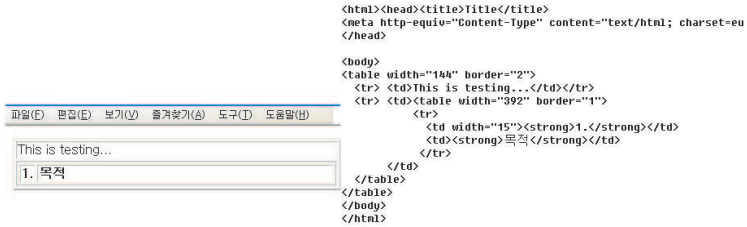


Fig. 14 Corrected tag usage by removal of unnecessary tag in Fig. 13.

모바일 휴대폰 이외의 인터넷이 가능한 TV, Car Navigation 장비 등으로도 서비스가 가능하다. 먼저 제안된 구조 분석 방법은 분석하고자 하는 웹 문서가 비 정형화 및 비 구조화로 인해 규칙이 지켜지지 않은 다양한 오류들을 문서 내에 포함하고 있으므로 이러한 오류들을 수정시킨다. 이는 웹 문서를 XML 응용이 되는 XHTML로 자동 변환시키는 TIDY(Dave Raggett)을 이용하여 Well-Formed한 HTML이 되도록 사전 전처리를 수행함으로써 가능하다. 그리고 TIDY는 HTML 문서를 XHTML 문서로 변환해 주는 기능을 제공하면서도 하나의 문장을 인위적으로 줄 바꿈을 처리하고 있기 때문에 이 부분은 변환 후 캐리지 리턴 된 라인들을 단일 문장으로 보정하도록 한다. 이 과정을 수행한 후 불 필요한 요소들을 제거하고 특정 요소의 사용 위치를 명시화하기 위해 스캐닝 과정을 거치면 웹 문서는 중첩 테이블을 포함하고 있는 Well-Formed한 웹 문서가 되며, 여기서 다음 각 사항들의 처리 방법을 적용하여 단일 테이블들로 구성된 웹 문서를 생성한다.

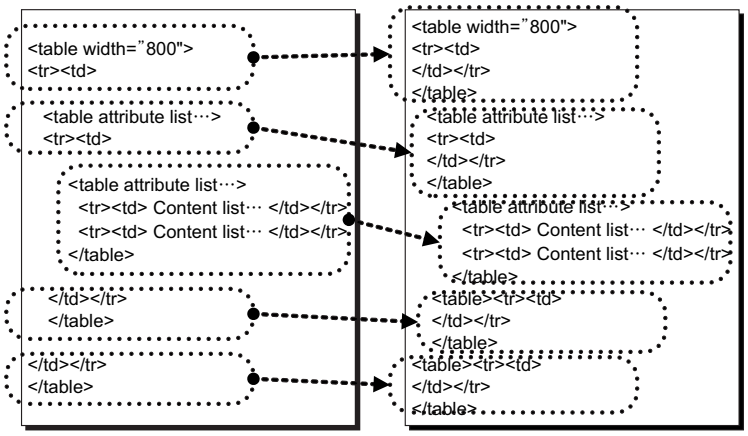


Fig. 15 Concept of single table extraction from web documents.

1) 페이지 레이아웃 테이블, 중첩 테이블 처리

유선 콘텐츠들은 디스플레이 크기에 따라 800×600, 1024×768등의 페이지 레이아웃을 이용하여 생성되고 있다. 이러한 페이지의 크기를 모바일 디스플레이에서 유선 환경과 동일하게 표현되는 것은 불가능하며, 모바일 디스플레이 장치에 표현되기 위해서는 적절한 크기로 재구성이 필요하다. 따라서 모바일 장비의 특성인 화면 해상도(평균 120×160)를 고려하면 유선 콘텐츠들을 세로로 구성하는 것이 적절하다. 따라서 Fig. 15와 같은 유선 콘텐츠의 페이지 레이아웃 테이블 및 중첩된 테이블의 사용은 Table 10의 규칙에 따라 단일 테이블들로 추출한다. 규칙은 1절 2)항의 웹 문서 분석 결과를 기반으로 작성되었으며 기술된 “Page Layout”의 종류는 “TABLE”, “DIV”, “LAYER”이다.

이와 같은 단일 테이블 추출 방식의 장점은 웹 문서에서 콘텐츠를 추출할 때 중첩된 서로 다른 테이블에 속한 콘텐츠들을 잃어버리지 않으면서 추출이 가능하고 단일 테이블 형식으로 추출되므로 정보 접근의 직접성을 제공한다.

Table 10. Rules for single table extraction from web documents.

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

2) 테이블로 간주될 수 있는 요소(DIV, LAYER) 처리

"DIV", "LAYER" 태그는 명시적인 "TABLE" 태그와 같이 하나의 독립된 페이지 레이아웃을 생성시킬 수 있는 기능을 가진 요소이다. 두 개의 요소 중 "LAYER" 태그의 사용은 조사 결과 사용되고 있지 않으므로 "DIV" 태그는 3절의 1)과 같이 "TABLE"에 준하여 처리하도록 하며, 다만 이들의 속성 중에 좌표 값(absolute 속성)을 이용하여 페이지 상의 위치를 결정하는 경우, 본 연구에서는 페이지 하단에 위치하도록 하였다. 이는 좌표 값을 가지므로 유선 웹 브라우저 상에서는 그 위치가 의미가 있지만 소형 단말기 브라우저 상에서는 하나의 "DIV" 요소의 내용을 디스플레이 할 만큼 단말기 디스플레이가 지원되지도 않는다. 또한 "DIV" 요소의 사용은 유선 상에서 채널 또는 베너 광고 등에 자주 사용되는데 이러한 정보는 옵션에 가깝기 때문이다. 그러나 좌표 값 속성을 사용하지 않는 경우에는 현재 코드의 위치를 그대로 갖는다.

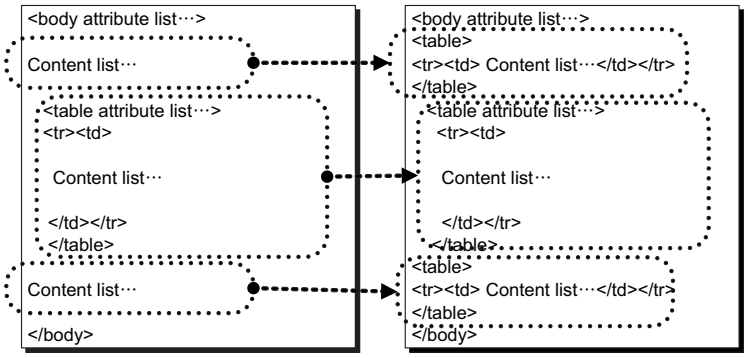


Fig. 16 Concept of single table extraction for contents block.

3) 테이블이 사용되지 않은 콘텐츠 블록들의 단일 테이블화

테이블을 사용하지 않고 표현된 콘텐츠 블록들도 변환의 유연성을 통일하게 하기 위해 하나의 단일 테이블로 구성하도록 한다. 콘텐츠는 어떠한 테이블에도 속하지 않는 경우와 테이블에 속한 경우의 콘텐츠로 구분할 수 있는데, 테이블에 속한 경우는 하나의 "TD"내에 명시적 구분 없이 묵시적으로 그 내용이 분리되는 경우이다. 그 순서가 콘텐츠와 테이블, 또는 테이블과 콘텐츠 순으로 다양하지만 3절 1)의 처리 방법을 이용하여 처리하도록 하며 다만, 웹 문서의 시작과 끝 지점에 Fig. 16과 같은 상황의 콘텐츠 블록들은 하나의 테이블로 구성한다. Table 11은 이러한 콘텐츠 블록들을 추출하기 위한 규칙이다. Table 11의 BODY는 HTML 태그의 "<BODY>"이며, "Starting Page Layout"은 "<TABLE>" 태그이다. 또한 "Ending Page Layout"과 "Ending BODY"는 각각 "</TABLE>" 태그와 "</BODY>" 태그를 의미한다.

Table 11. Rules for single table extraction for contents block.

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

Table 12. Rules for "<FORM>" components extraction.

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)

4) FORM 관련 구성 요소 추출

단일 테이블 추출 시 특별한 경우에 해당되는 "<FORM>" 구성 요소는 클라이언트 사용자와의 상호작용(interaction)을 지원하는 HTML 요소이지만 웹 문서 작성 시 "<FORM>" 요소의 사용은 페이지 디자인에 따라 아주 다양하게 사용되면서 그 사용 위치도 대부분 정해지지 않고 있다. 즉 사용법이 느슨한 이유로 "<TABLE>" 요소 내·외부에 위치가 가능하여 웹 구조 분석을 어렵게 하고 있다. "<TABLE>"을 사용함으로써, 서로 다른 요소(TD)안에 "<FORM>" 관련 요소들이 존재할 수도 있으며, 또한 "<FORM>" 요소 내부에 콘텐츠 블록을 둘 수도 있다. 따라서 본 연구에서는 "<FORM>" 관련 요소만을 하나의 단일 테이블로 추출하고자 한다. 혼합되어 있는 "<FORM>" 관련 요소 이외는 제거하도록 하며, "<FORM>" 내부의 콘텐츠 블록은 관련 요소에 포함하여 함께 처리하도록 한다. Fig. 17은 "<FORM>" 관련 요소들을 재구성 방법을 보이며, Table 12는 추출을 위한 규칙이다.

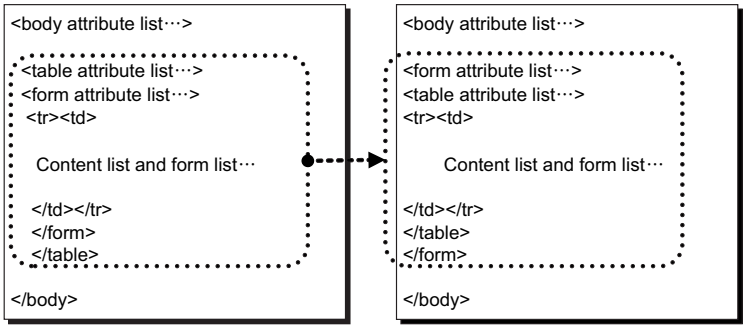


Fig. 17 Extraction concept of "<FORM>" components.

Table 13. Tags list for deletion.

applet, basefont, bdo, center, col, colgroup, del, dir, embed, fieldset, font, hr, i, ins, isindex, legend, noframes, noscript, optgroup, s, script, small, strike, style, sub, sup, tbody, tfoot, thead, tt, u

5) 콘텐츠 변환을 위한 불필요한 요소 제거

이는 XHTML Basic에서 지원되지 않는 HTML 태그 요소(XHTML Basic)들의 제거 과정을 말한다. 콘텐츠 변환 과정에서 처리하여도 무방하지만, 본 연구에서는 순서 없이 사용될 수 있는 불필요한 요소들을 웹 문서 구조 분석 과정의 스캐닝(토큰화) 과정에서 일괄적으로 검색해서 삭제하도록 하였다. 삭제 대상인 태그 리스트들은 Table 13과 같으며, 삭제하기 위한 규칙은 Table 14와 같다.

이상과 같이 비구조적이고 정형화되지 않은 웹 문서를 제안된 분석 방법을 이용하여 단일 테이블 기반의 웹 문서로 추출하게 되면 서로 관련된 정보들을 그 순

Table 14. Rules for tag deletion.

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

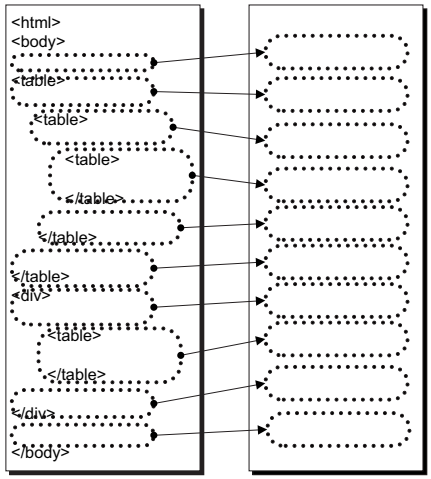


Fig. 18 Simplified diagram of extraction of single table.

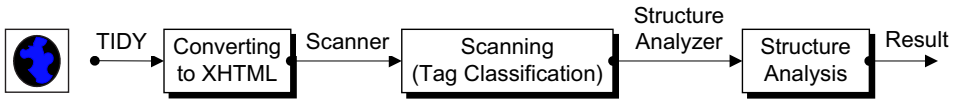


Fig. 19 Steps of web document structure analysis.

서대로 유지하면서 왜곡 없는 콘텐츠 신뢰도를 유지하면서 XHTML Basic의 지원 요소에 적합한 새로운 웹 문서가 추출된다. 새롭게 추출된 웹 문서는 단일 테이블들의 연속된 집합으로써, 경량화·단순화된 특징을 갖으며 4장에 제안될 콘텐츠 변환의 입력 문서이다. 단일 테이블 추출과정을 간략하게 도식화하면 Fig. 18과 같다.

4. 제안된 웹 문서 구조 분석 알고리즘

웹 문서 구조 분석의 결과는 XHTML 스펙을 따르게 된다. 이는 XML 응용 애플리케이션들에게 필요에 의한 접근을 가능하도록 하게 할 수 있으며, 향후 모바일 인터넷 분야 이외의 응용에서 재사용이 가능하다. 본 연구의 웹 문서 구조 분석은 Fig. 19의 순서를 따르며 각 과정은 다음과 같다.

Table 15. Scanning algorithm.

```
mTable = new String[HTML 태그 리스트] // HTML 태그 배열 선언
index = 0, source = ""
while(지정된 URL 문서가 Null이 될 때까지 라인단위로 읽어옴) {
        split = 읽어온 내용을 "< ... >" 단위로 분리
        for( split 길이만큼 ) {
         while(true) {
           tags = split의 부분 문자
           if( mTable[index]를 소문자로 바꾼 내용과 tags가 같다면) {
              if(tags가 삭제 대상 태그) { tags 삭제 }
              else if(tags가 삭제 대상 태그가 아니면)
              { source += tags }
              else if("<FORM>" 태그)
              { source += 단일 테이블로 재구성 }
              else if("특수문자('!', '/')") { source += 특수문자 }
           } else { source += split의 부분 문자 }
         index++;
        }
       index = 0;
    }
  }
```


1) 스캐닝(토큰화)

제안하는 구조 분석을 위하여 스캐닝은 웹 문서내의 태그들을 태그 단위로 분리하도록 한다. 즉 이 과정은 라인 단위로 원문을 읽어 와서 단어의 종류를 구분하여 분리한다. 이 과정에서 문서내의 대·소문자 사용을 소문자 사용으로 일률적으로 변경하며, 또한 불필요 태그 요소들을 삭제하는 기능을 수행한다. 그럼으로써 웹 문서 구조 분석 모듈의 입력 데이터를 작성한다. Table 15는 스캐닝 알고리즘이며, 가장 내부의 if 문에서 HTML 태그의 알파벳 순서대로 비교하여 불필요한 태그 삭제, 그리고 "<FORM>" 태그일 때 새로운 테이블로의 작성, 대·소문자 변환 등을 수행하여 구조 분석에서 단일 테이블 추출이 용이하도록 전처리한다.

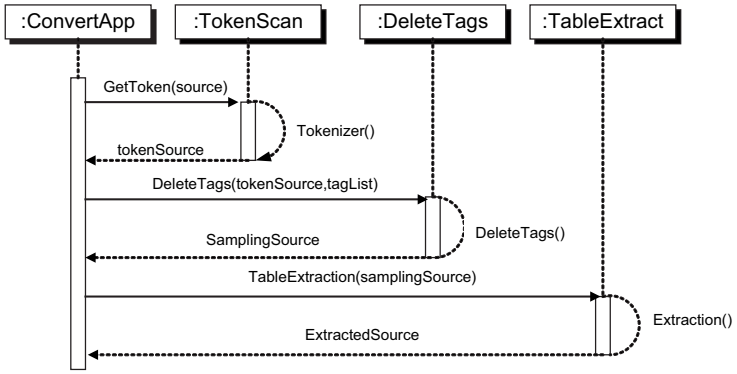


Fig. 20 UML sequence diagram of algorithm.

2) 단일 테이블 추출 알고리즘

웹 문서의 단위 테이블 추출 알고리즘을 Table 16에 나타내었다. 알고리즘은 전술된 바와 같이 페이지 레이아웃, 중첩 테이블, 독립된 콘텐츠 블록, 그리고 “<FORM>” 관련 요소들을 하나의 단일 테이블로 추출하게 된다. 즉 웹 문서의 코드 순서대로 “<TABLE>”, “<DIV>”, “<LAYER>”, “<FRAMESET>” 태그를 식별하여 “<TABLE>”,“<DIV>”,“<LAYER>”와 독립된 콘텐츠 블록, 그리고 “<FORM>” 관련 요소들을 단위 테이블들로 추출한다. “<FRAMESET>”은 단일 테이블 추출 과정에서 생략되는데 이는 차후 제안하는 콘텐츠 변환 과정에서 처리한다.

3) 제안된 웹 문서 구조 분석 알고리즘 UML 순차 다이어그램

웹 문서 구조 분석 알고리즘의 UML 순차 다이어그램을 Fig. 20에서 나타냈으며 Table 14의 규칙과 Table 15, Table 16의 알고리즘 결과는 Fig. 18의 추출된 단위 테이블들의 집합들로 구성된 웹 문서를 도출하게 된다.

Table 16. Algorithm for single table extraction.

```
sr = 스캐닝 결과 파일, input = null, counter = 0,
flag = true, fFlag = true, preSource = null, source = “”
while( ( input = sr을 라인 단위로 읽은 내용) != null ) {
    if( input == “<body” ) { // “body” 시작 부분
        flag = false, source += input }
    else if( (input == (“<table” or “<table>”)) and (counter == 0) ) {
        if(flag == false) {
            preSource=“<table><tr><td>”+preSource+“</td></tr></table>”
            source += preSource
        }
        source += input, counter = counter + 1, flag = true
    }
    else if( (counter == 0) and (flag == false)) {
        preSource += input
    }
    else {
        if(( input == (“<table” or “<table>”)) and (counter != 0)) {
            source += “</td></tr></table>”
            source += input, counter = counter + 1, flag = true
        }
        else if( input == “</table>” ) {
            counter = counter - 1
            if( counter != 0 ) { source += input + “<table><tr><td>” }
            else { source += input }
            flag = true
        }
        else if( input == “<form” ) {
            source += “</td></tr></table>” + input + “<table><tr><td>”
            flag = true
        }
        else if(input == “</form>”) {
            source += “</td></tr></table>” + input + “<table><tr><td>”
            flag = true
        }
        else { source += input; flag = true }
        flag = true
    }
}
print source // 결과 출력
```

IV. 콘텐츠 변환

4장에서는 경량화 된 웹 문서 구조 분석의 결과를 입력으로 받아서 모바일 콘텐츠로 변환하는 콘텐츠 변환 방식을 제안한다.

1. 콘텐츠 변환 방식의 개요

모바일 콘텐츠 변환 방식은 2장에서 언급한 바와 같이 최초 작성된 환경에 따라 그 변환 방법을 달리하면서 현재까지 이어져 오고 있다. WAP 포럼이 WAP 1.X의 차세대 버전인 WAP 2.0을 발표했는데, 이는 향후 모바일 인터넷 서비스의 표준 규격으로 준비되고 있으며, 더불어, 유럽·북미·아시아의 WAP 진영, 마이크로소프트의 ME 진영, DoCoMo의 i-Mode 진영으로 구분되던 무선 인터넷 기술 규격이 하나로 수렴되는 상황을 만들었다. 또한 WAP 2.0은 유선 인터넷과 같이 유연하고 개방적인 구조로 무선 인터넷 망을 운영할 수 있도록 고안되었으며 문자 중심의 기존 WML의 한계를 극복하고자 XHTML Family인 XHTML Basic을 모바일 마크업 언어로 수용하기에 이르렀다(WAP Forum). 하지만 각 이동통신업체들의 현실은 무선 인터넷 프로토콜의 표준과 모바일 인터넷 마크업 언어의 표준을 채택하기까지는 다소 시일이 필요할 것으로 보인다. 그러나 꾸준히 표준에 수렴해 가는 양상을 띠고 있어서, 앞으로 많은 킬러 애플리케이션들이 개발되고 서비스 될 것으로 보고

있다(Zionwap),(KTF 포탈 기획팀, 2003).

따라서 본 연구의 콘텐츠 변환은 바로 이러한 움직임에 맞추고자 향후 모바일 인터넷 마크업 언어의 표준으로 자리 잡을 XHTML Basic에 초점을 두고 기존의 웹 문서를 변환하였다. 그 변환 방식이 적용될 대상은 3장에서 도출된 경량화된 웹 문서인 단일 테이블 기반 문서이며, 여기에 적용될 변환 규칙은 XSLT 기술인 XSL 템플릿 내에 정의하였다.

2. 변환 규칙 정의

콘텐츠 변환에 사용될 규칙들을 정의하기 위해 관련된 마크업 언어들을 Table 17에서 HTML 4.0과 XHTML의 모듈, 그리고 XHTML Basic의 주요 태그들을 살펴보았다. 이는 변환될 원본 웹 문서는 HTML 4.0 스펙을 따르며, XHTML 모듈은 W3C의 XHTML의 모듈화 버전(XHTML 1.1) 스펙이며, 또한 XHTML Basic은 XHTML의 Sub Set이면서 모바일 장치를 위해 권고한 마크업 언어임을 의미한다. W3C에서 권고된 XHTML 모듈화 버전은 HTML 4.0의 구성요소(태그와 속성)들을 기능별로 28개의 모듈로 구분하여 그룹화 했으며, 이렇게 구성된 모듈 중에서 모바일 장치용 마크업 언어로 최소한으로 적용될 수 있는 11개 모듈을 정의하였다.

따라서 본 연구의 콘텐츠 변환은 HTML 4.0 기반 웹 문서에서 XHTML Basic의 각 모듈에 속해 있는 해당 요소로 변환되도록 하는 것이며, 이를 위해 HTML과 XHTML Basic간 태그 및 속성 변환 규칙이 필요하게 된다. 이는 XHTML Basic에

서 지원되지 않는 HTML 태그 구성 요소들은 삭제되거나 유지, 치환되는데 현재 대부분 텍스트 지향이며, 동적인 요소들(Applet, Script, iFrame, Style 등)과 프레임, 이미지 맵 등은 Table 17에서 보는 바와 같이 XHTML Basic에서 지원되고 있지 않다.

1) 태그 변환 규칙

선행 연구들의 콘텐츠 변환 방법인 HTML 필터링은 태그 단위 변환이다. 즉 하나의 태그를 만나면, 태그 변환표에서 이에 해당되는 태그를 찾고 유지, 치환, 삭제의 과정을 변환 규칙에 따라 적절하게 변환을 수행하게 된다. 대부분이 그대로 유지되지만 몇 개의 치환과 삭제가 존재한다. 관계된 정보는 Table 18에 나타내었다.

또한 기존 연구의 XHTML 모듈 기반은 모듈별로 변환 규칙을 XSLT 기술의 XSL 문서로 작성하고, 태그 패턴 매칭의 기반으로 변환하고 있다. 마찬가지로 모듈 기반 콘텐츠 변환도 Table 18의 HTML 필터링 변환처럼 대부분 태그들이 그대로 유지되면서 치환과 삭제가 존재한다.

하지만 두 가지 방식의 차이점은 변환 대상이 되는 태그들은 동일하지만 변환 방법에서 필터링의 하드 코딩과 모듈 기반 XSL의 문서 재 포맷으로 볼 수 있는데, 이는 태그 변환 과정 상 두 가지 변환 방식은 물리적인 측면에서는 동일하지만 XSL 특성 때문에 XSLT 기술의 XSL을 이용한 모듈 기반 변환이 확장성과 문서의 유효성(Validation) 측면에서는 장점을 갖는다. 그러나 XSLT 기술의 XSL 기반은 위에서 기술한 바와 같이 패턴 매칭으로 재 포맷 템플릿 작성 시 중첩된 페이지 레이아웃들이 복잡하게 사용된 정형화되지 않은 웹 문서를 대상으로 템플릿 작성이 이루

Table 17. Comparison between HTML 4.0 and XHTML Basic based on XHTML modularization.

	HTML 4.0	HTML Mod.	Components of Module	XHTML Basic
1	body,head,html,title	Structure	html,head,title,body	body,head,html,title
2	abbr,acronym, address,blockquote, br,cite,code,dfn, div,h1-h6,kbd,p, pre,q,samp,span, strong,var	Text	abbr,acronym,address, blockquote,br,cite, code,dfn,div,em, h1-h6,kbd,p,pre,q, samp,span,strong,var	abbr,acronym,address blockquote,br,cite, code,dfn,div,em, h1-h6,kbd,p,pre,q, samp,span,strong,var
3	a	Hypertext	a	a
4	dd,dl,dt,li,ol,ul	List	dl,dt,dd,ol,ul,li	dl,dt,dd,ol,ul,li
5	applet	Applet	applet,param	
6	b,big,hr,i,small sub,sup,tt	Presentation	b,big,hr,i,small, sub,sup,tt	
7	del	Edit	del,insert	
8		Bi-directional	bdo	
9		Basic forms	form,input label, select,option,textarea	form,input lable, select,option,textarea
10	button,fieldset, form,input,label, legend,optgroup, option,select, textarea	Forms	form,input,select, option,textarea, button,fieldset,label, legend,optgroup	
11		Basic table	caption,table,td,th,tr	caption,table,td,th,tr
12	catpion,col,td,th, table,colgroup, tfoot,thead,tr	Tables	caption,table,td,th, tr,col,colgroup,thead, tfoot	
13	img	Image	img	img
14	area,map	Client-side image map	a&, area,img&, map,object&	
15		Server-side image-map	img&	
16	object,param	Object	object,param	object,param
17	frame,frameset	Frames noframe	frameset,frame, noframe	
18		Target	a&, area&, base&, link&, form&	
19	iframe	iFrame	iframe	
20		Intrinsic events	a&, area&, form&, body&, label&, input&, select&, textarea&, button	
21	meta	Metainformation	meta	meta
22	noscript, script	Scripting	noscript, script	
23	style	Style sheet	style	
24	link	Link	link	link
25	base	Base	base	base
26	기타 Style attribute, Name identification, Legacy Module 존재			

어진다. 따라서 XSLT 기술에서 제공하는 다양한 문서 생성 요소 사용이 필수적으로 요구되거나, 또는 특정 태그 요소 검색식을 표현함에 있어서 웹 문서 전체를 검색 범위로 표현해야 하는 문제점이 있다. 그러나 본 연구에서 제안하는 콘텐츠 변환 기법은 비구조적이고 정형화되지 않은 웹 문서를 분석한 결과인 단일 테이블 기반의 문서를 토대로 변환이 수행되기 때문에 태그 변환상의 패턴 매칭과 검색 범위가 지역적이며 단순하다.

따라서 본 연구의 변환 방식은 기존 필터링 기반의 변환 규칙을 사용하면서 관련 태그 요소들을 XHTML 모듈 기반으로 구분하여 수행한다. 태그 변환 규칙은 Table 19와 같다.

2) 속성 변환 규칙

기본적으로 필터 기반 및 모듈 기반과 동일하게 속성 변환도 태그 변환 과정과 동일하다. 서로 매핑되지 않는 속성들은 삭제하거나, 유사한 기능을 하는 속성으로 치환하며, 매핑되는 속성은 유지시킨다. Table 17과 Table 18을 참고하여 몇 가지 주요 태그 및 속성 변환의 예를 추출하면 Table 20과 같다. 요소들을 구성하는 항목 중에서 “Style”과 “Script Event”들은 삭제하고, 모든 요소들이 최소 구성 속성들로 이루어지도록 속성들을 변환되도록 하였다.

Table 18. Mapping table for conversion based on filtering.

구분	Before	After	Attribute	Before	After	Attribute
유지 (p)	html	html	P	meta	meta	P
	head	head	P	object	object	P
	title	title	P	option	option	P
	body	body	D	param	param	P
	abbr	abbr	P	select	select	P
	acronym	acronym	P	span	span	P
	address	address	D	textarea	textarea	P
	base	base	D	kbd	kbd	D
	blockquote	blockquote	D	label	label	D
	br	br	D	li	li	D
	caption	caption	D	ol	ol	D
	cite	cite	D	p	p	D
	code	code	D	pre	pre	D
	dfn	dfn	D	samp	samp	D
	dd	dd	D	strong	strong	D
	dl	dl	D	th	th	D
	dt	dt	D	tr	tr	D
	form	form	P	ul	ul	D
	input	input	P	var	var	D
	link	link	P	comment	comment	D
	속성 변환 시 참고 사항					
	q	q	속성 중에 “lang”만 유지			
	a	a	- href 속성 값이 html, 사이트 주소가 아닐 때 즉, 스크립트일 때는 콘텐츠 내용만 유지 - 속성 값이 스크립트이고 연결링크가 이미지이면 전체 삭제			
	img	img	- 단순 이미지는 전체 삭제 - 링크 이미지의 이미지는 삭제, 연결 링크는 [IMG]로 표현			
	table*	table	border 속성만 유지, 나머지는 삭제			
	td	td	colspan, rowspan, class만 유지, 나머지는 삭제			
	위 전체 태그 리스트 중에서 속성이 style 또는 스크립트 이벤트들은 삭제					

치환 (R)	Before	After	Attribute
	b	strong	D
	big	strong	D
	em	strong	D
	menu	ul	D
	div	table	위의 *와 동일
	area, map	a	alternative
	frame, frameset	a	alternative
	iframe	a	alternative
	h1 - h6	p	D
	- area, map : a 태그로 치환하고 map name의 인덱스 위치에 삽입 <area shap=“rect” coord=“x1,x2,y1,y2”> → link - frame, frameset : <frame src=“test.html” name=“name” frameborder=“1” ..> → link - iframe : <iframe src=“test.html” ... > → link - h1-h6 : test → test		
삭제 (D)	applet, basefont, bdo, button, center, col, colgroup, del, dir, embed fieldset, font, hr, i, ins, inindex, legend, noframes, noscript, optgroup, s, script, small, strike, style, sub, sup, tbody, tfoot, thead, tt, u		

Table 19. Rules for conversion.

RULE (7)
IF : A equivalence tag existed in XHTML Basic about the source tag
THEN : Preserve the source tag
RULE (8)
IF : A equivalence tag isn't existed or
have no function similar to source tag in XHTML Basic
THEN : Delete the only source tag except contents block
RULE (9)
IF : There is no equivalence tag but,
have function similar to source tag in XHTML Basic
THEN : Replace the source tag to tag in XHTML Basic module

Table 20. Example of attribute and tag conversion.

Tag	Change of Tag
frame	frame의 src, name 속성 값을 추출하여 a 링크로 치환
iframe	iframe의 src, name 속성 값을 추출하여 a 링크로 치환
map	area의 href, alt 속성 값을 추출하여 a 링크로 치환
etc...	...
	Change of Attribute
q	lang 속성만 유지, 나머지는 삭제
a	href 속성만 유지, 나머지는 삭제
img	src 값이 절대 경로인 경우 삭제, [IMG] 표시로 치환
table	border 속성값만 유지, 나머지는 삭제
tr	삭제
td	colspan, rowspan 속성만 유지, 나머지는 삭제
others	속성 항목 중에서 동적인 내용은 삭제

3. 제안하는 콘텐츠 변환 방식

1) 제안하는 콘텐츠 변환 방식의 장점 및 변환 과정

유선 콘텐츠를 모바일 콘텐츠로 변환하기 위해서 본 연구에서 제안하는 콘텐츠 변환 방식은 단일 테이블 기반 콘텐츠 변환 방식이다. 이는 변환 대상인 비구조적이고 정형화되지 않은 웹 문서를 구조적이고, 경량화 되고, 단일화된 웹 문서로 전처리를 수행한 후 구조 분석 된 경량의 웹 문서를 대상으로 콘텐츠 변환을 수행하게 된다. 경량화된 웹 문서의 특징은 단일 테이블들의 집합이며, 테이블들은 중복되거나 복잡한 페이지 레이아웃이 사전에 제거되어 있어서 콘텐츠 변환 규칙 템플릿을 단일 테이블 기반으로 쉽게 작성할 수 있게 한다. 즉 제안하는 콘텐츠 변환 방식을 사용하면 다음과 같은 장점을 얻을 수 있다.

첫째, XSLT 기술의 XSL 변환 기법을 사용한다. 이는 XSL의 특징을 가짐으로 변환 템플릿 작성에 있어서 확장성과 문서의 유효성을 제공한다.

둘째, 변환 템플릿을 적용할 웹 문서가 경량화 되어 있기 때문에 변환 템플릿 작성을 쉽게 할 수 있다. 즉 웹 문서의 페이지 레이아웃, 또는 중첩된 테이블들을 사전에 제거하여 생성된 단일 테이블 기반으로 특정 태그 검색식인 XPATH(XPATH)를 표현함으로, 템플릿을 중첩되지 않게 단순화 한다.

셋째, 태그 또는 속성의 변환은 WAP 2.0 마크업 언어인 XHTML Basic에 기준하여 처리하기 때문에 유선과 모바일 인터넷에서 공통적으로 사용된다.

끝으로 제안하는 콘텐츠 변환 방식은 모바일 콘텐츠 언어의 표준화를 따른 변환 방식이기 때문에 XHTML Basic 또는 WML 2.0을 지원하는 모든 브라우저에서

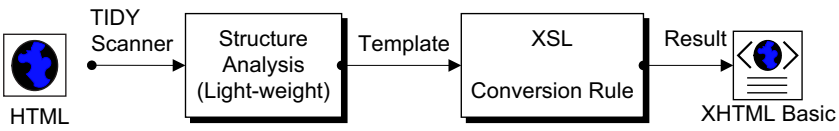


Fig. 21 Steps from web documents structure analysis to contents conversion.

재 변환 모듈 없이 재사용 할 수 있다.

콘텐츠 변환 과정은 먼저 정형화되지 않은 웹 문서를 TIDY를 이용하여 XML 기반 구조인 XHTML로 전처리 한 후, 이를 3장에서 제안한 웹 문서 구조 분석을 통해 경량화되고 Well-Formed한 XML 문서를 생성한다. 그리고 4장 2절에서 정의된 태그변환 규칙이 작성된 XSLT 기술의 XSL 문서를 생성된 XML 문서에 적용하여 XHTML Basic 문서로 변환한다. Fig. 21에 3장에서 웹 문서 구조 분석과 이를 기반으로 하는 콘텐츠 변환까지의 과정을 나타내었다. 변환 과정상의 XSL의 사용은 앞에서 기술했듯이, 변환의 확장성과 문서의 유효성을 제공하기 위함이다. 기존 필터링 기법은 하드 코딩의 문제점이 존재하며, 또한 기존의 XSL 변환기법은 변환 템플릿 작성 시 XPATH 검색식 범위가 문서 전체에 해당된다. 하지만 이러한 단점을 보완하기 위해, 웹 문서를 경량화한 후, XSL을 사용하면 확장성을 제공할 수 있으며 또한 XPATH의 검색식을 지역적으로 단순화하여 변환 할 수 있게 된다.

2) 단일 테이블 기반 콘텐츠 변환 템플릿

XSLT 기술의 XSL 변환 템플릿에는 HTML 태그들을 XHTML Basic에 해당되는 태그들로 변환하는 규칙들을 태그의 패턴 매칭으로 정의하였다. 패턴 매칭의 검색식은 3장에서 도출한 구조 분석된 경량의 웹 문서의 특징을 사용하였다. 경량화된 웹 문서는 XML 기반이며 또한 단일 테이블 기반이므로 정형화되지 않은 웹 문

Table 21. Structure of single table with contents.

```
<html>
<head><title> content </title></head>
<body>
.....content.....
        <table><tr><td> .....content..... </td></tr></table>
        <table>
           <tr><td> ...content model... </td></tr>
           <tr><td> ...content model... </td></tr>
        </table>
        <table><tr><td> .....content..... </td></tr></table>
.....content.....
</body></html>
```

서보다는 단순하게 패턴 매칭으로 특정 태그의 검색식을 정규화 할 수 있다. 즉 단일 테이블이므로 모든 콘텐츠가 Table 21과 같이 "<TABLE> ··· content model ··· </TABLE>"의 연속된 구조로 구성된다.

따라서 태그들을 변환하기 위한 XSLT 기술의 XSL 변환 템플릿을 단일 테이블 기본 구조에 맞춰 패턴 매칭과 변환 규칙에 의거 작성하여 변환 대상의 태그들을 변환한다. 이는 변환할 웹 문서가 여러 개의 테이블 모듈로 추상화 된 것이라 볼 수 있기 때문에 XHTML 모듈인 "TABLE MODULE"를 변환 대상으로 정하고 사용된 태그들을 검색해 나가는 방식으로 변환할 수 있다. 즉 Table 22에서 보여주는 바와 같이 XSL 변환 템플릿의 XPATH 패턴 검색식을 표현할 때 중첩된 테이블들을 모두 검색하는 "//(* or all)"의 수식을 사용하지 않아도 된다. 또한 변환 템플릿에서는 웹 문서 구조 분석과정을 거치지 않는 "FRAME", "IFRAME", "MAP"과 같은 구성 요소들은 XHTML Basic에서 지원되지 않으므로 콘텐츠 변환에서 링크(<a>) 요소로 변환하도록 한다. Table 23은 콘텐츠 변환 템플릿 예이다.

Table 22. XPATH example of conversion template.

Pre Researches	<xsl:template match="/"> ..... <xsl:template match="//table"> <xsl:apply-templates select="tr/td" /> // 모든 테이블을 검색 .....
Proposed Method	<xsl:template match="/html/body/table"> ..... <xsl:template match="table"> <xsl:apply-templates select="tr/td" /> // body의 child인 테이블 검색 .....

3) 모바일 디바이스 프로파일 분석 및 전환기

본 연구의 콘텐츠 변환 결과는 XHTML Basic과 WAP 2.0 기반이므로 이러한 모바일 환경에서는 별다른 추가 작업 없이 변환된 콘텐츠를 서비스 받을 수 있지만 현재 서비스 되고 있는 환경과 연동하여 서비스하기 위해서는 요청한 클라이언트의 정보를 파악하고 적절한 페이지를 제공할 수 있어야 한다. 모바일 디바이스 프로파일 분석 및 전환기는 이러한 기능을 제공하기 위해 요청한 모바일 클라이언트의 헤더 정보를 파악한 후 해당 단말기에 적절한 페이지로 전환하도록 한다. 또한 서비스 제공자 입장에서는 모바일 단말기 특성별로 해당 서비스 페이지를 준비해야만 한다.

Fig. 22에 모바일 단말기 CC/PP 및 헤더 정보를 파악한 후 해당 모바일 단말기의 특성별로 작성된 적절한 페이지로 전환시키는 전환 흐름도를 나타낸다. CC/PP는 W3C에서 제안한 단말기의 전송 환경 요소를 기술할 수 있는 전송 컨텍스트 메타 정보 표현 모델이며 RDF(Resource Description Framework)(RDF) 기반으로 단말기의 속성과 값을 표현한다. 또한 UAProf(User Agent Profile)인 단말기 헤더 정

Table 23. Example of contents conversion template.

```
<?xml version="1.0" encoding="euc-kr"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output encoding="euc-kr" method="html" />
<xsl:template match="html">
<html><xsl:apply-templates select="head |body |frameset" /></html></xsl:template>
<xsl:template match="head">
<head><xsl:apply-templates select="title" /></head></xsl:template>
<xsl:template match="title">
<title><xsl:apply-templates select="text()" /></title></xsl:template>
<xsl:template match="body">
<body><xsl:apply-templates select="table" /></body></xsl:template>
<xsl:template match="table">
<xsl:apply-templates select="tr" /></xsl:template>
<xsl:template match="tr"><xsl:apply-templates select="td" /></xsl:template>
<xsl:template match="td">
    <xsl:if test="(number(string-length(text())) > 0) or (descendant::*)">
    <table border="0"><tr><td>
    <xsl:value-of select="text()" />
    <xsl:apply-templates select="a |a/img |img/a |p |map/area |img |form |text()" />
    </td></tr></table></xsl:if></xsl:template>
<xsl:template match="p">
    <xsl:value-of select="text()" /><br/></xsl:template>
<xsl:template match="a"><xsl:element name="{name()}">
    <xsl:attribute name="href"><xsl:value-of select="@href" />
    </xsl:attribute><xsl:value-of select="." />
</xsl:element><br /></xsl:template>
<xsl:template match="img">
    [img] <xsl:value-of select="@alt" /><br /></xsl:template>
<xsl:template match="area">
    <xsl:element name="a"><xsl:attribute name="href"><xsl:value-of select="@href" />
</xsl:attribute><xsl:value-of select="@alt" /></xsl:element><br /></xsl:template>
<xsl:template match="form"><xsl:element name="{name()}">
    <xsl:attribute name="name"><xsl:value-of select="@name" /></xsl:attribute>
    <xsl:attribute name="method"><xsl:value-of select="@method" /></xsl:attribute>
    <xsl:attribute name="action"><xsl:value-of select="@action" /></xsl:attribute>
    <xsl:value-of select="text()" /><xsl:apply-templates select="input |select" />
</xsl:element></xsl:template>
<xsl:template match="input"><xsl:element name="{name()}">
    <xsl:attribute name="type"><xsl:value-of select="@type" /></xsl:attribute>
    <xsl:attribute name="name"><xsl:value-of select="@name" /></xsl:attribute>
    <xsl:attribute name="value"><xsl:value-of select="@value" /></xsl:attribute>
</xsl:element></xsl:template>
<xsl:template match="select"><xsl:element name="{name()}">
    <xsl:attribute name="name"><xsl:value-of select="@name" /></xsl:attribute>
    <xsl:apply-templates select="option" /></xsl:element></xsl:template>
<xsl:template match="option"><xsl:element name="{name()}">
    <xsl:attribute name="value"><xsl:value-of select="@value" /></xsl:attribute>
<xsl:value-of select="text()" /></xsl:element>
    // .....omission.....//
</xsl:template>
```

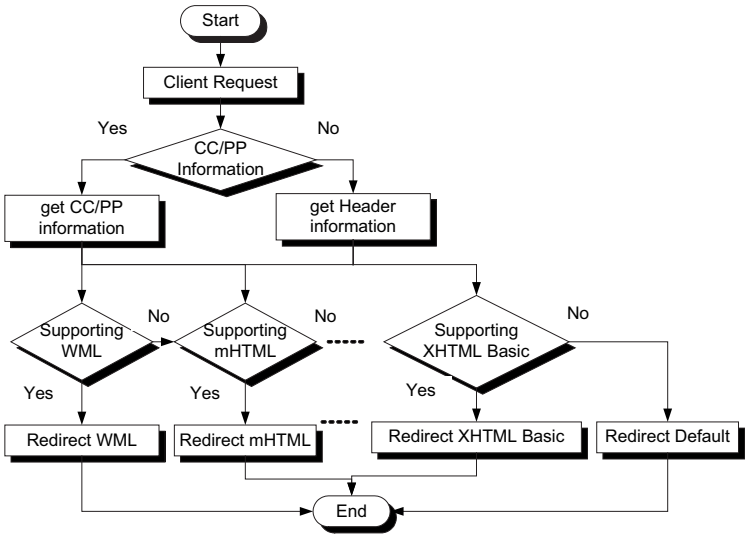


Fig. 22 Flow diagram of analyzer and redirector.

보는 단말기별로 소유하고 있는 고유의 특성 정보이며 요청 시 자동으로 서버에 전송되는 정보이다.

4. 서비스 메타 모델

본 연구의 서비스 메타 모델은 3장에서 제안된 웹 문서 구조 분석과 4장에서 제안한 콘텐츠 변환 방식으로 구성되며 Fig. 23과 같다. 메타 모델의 세부 구성 요소는 XHTML 변환기(XC), 웹 문서 구조 분석기(WDSA), 변환 템플릿 수행기(CTO), 그리고 모바일 디바이스 프로파일 분석 및 전환기(MDHA/R)이다.

XHTML 변환기는 웹 문서를 XML 기반의 문서로 변환하는 역할을 담당한다. 웹 문서는 구조적이지 못하며 또한 비 정형화되어 있기 때문에 이를 분석하기 위해서는 일정한 문서 형식을 갖도록 해야 한다. 따라서 XHTML 변환기는 웹 문서를 XML 기반인 XHTML 형식으로 변환해주는 기능을 담당하며 전용 애플리케이

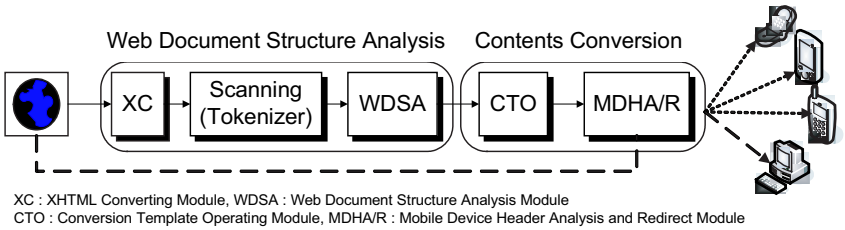


Fig. 23 Service meta model.

션을 작성할 수도 있지만 본 연구에서처럼 TIDY라는 상용 툴을 사용할 수 있다. 또한 스캐닝 작업은 태그들의 종류를 구분하여 분리시켜 주며, 불필요한 요소 제거 및 태그 사용의 명시화 작업을 수행한다. WDSA인 웹 문서 구조 분석기는 XHTML 및 스캐닝 과정의 결과를 입력으로 받아들여 웹 문서를 단일 테이블들의 집합으로 재구성하는 역할을 담당한다. 그리고 변환 템플릿 수행기는 웹 문서 구조 분석기의 결과에 XHTML Basic 형식으로 변환시키는 변환 규칙이 정의된 XSL 변환 템플릿을 적용시켜 유선 콘텐츠를 모바일 콘텐츠로 변환하는 기능을 처리한다. 끝으로 모바일 디바이스 프로파일 분석 및 전환기는 Legacy 환경과 연동하기 위해 클라이언트 헤더를 분석하고 적절한 서비스 파일로 전환시켜주는 역할을 담당한다.

V. 분석 모델 작성 및 성능 평가

5장에서는 본 연구의 성능을 평가하기 위한 분석 모델과 성능 평가 기준을 작성하였으며, 또한 구현 결과를 보였다. 그리고 작성된 분석 모델 기반 상에서 제안된 모바일용 콘텐츠 변환을 위한 웹 문서 구조 분석 변환 모듈의 성능 평가 및 결과를 분석하였다.

1. 분석 모델 작성

이 절에서는 성능을 평가하기 위한 분석 모델과 필요한 전제 조건 및 성능 평가 기준을 작성하였다. 제안된 웹 문서 구조 분석 방법과 콘텐츠 변환 방법을 적용할 분석 모델을 작성하기 위해 향후 모바일 인터넷에서의 제공될 희망 서비스와 이용 용도를 고려하여 현재 상용 서비스 중인 웹 사이트들을 서비스 종류에 따라 분류하고, 분류된 웹 사이트를 대상으로 제안된 방법을 적용하기 위해서 Fig. 24와 Fig. 25의 통계 데이터(한국인터넷정보센터, 2003)를 분석하였다.

Fig. 24는 모바일 인터넷 이용자가 모바일 인터넷을 통해 가장 많이 제공받기를

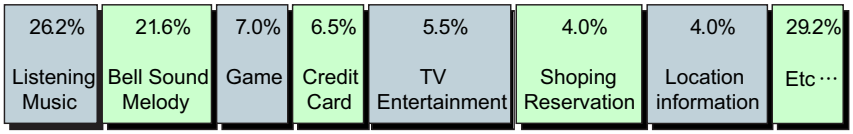


Fig. 24 Preferred services of mobile users (units:%).

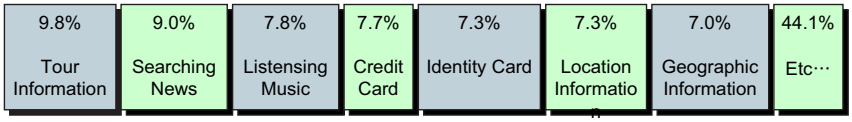


Fig. 25 Excepted preference services of mobile users (units:%).

희망하는 서비스가 “음악듣기(26.2%)”이며, 그 다음으로 “벨소리/멜로디(21.6%)”와 “게임(7.0%)” 등의 순서임을 보여준다. 또한 Fig. 25는 현재 무선 인터넷 비이용자들이 향후 무선 인터넷을 이용하게 될 경우 선호하게 될 서비스 내용으로 “여행/교통 정보 검색(9.8%)”을 가장 많이 지적하고 있으며, 다음으로는 “뉴스 검색(9.0%)”, “음악 듣기(7.8%)”, “신용카드 기능(7.7%)”, “신분증(7.3%)”, “위치정보 확인(7.3%)”, “지리정보 제공(7.0%)” 등임을 보여주고 있다.

Fig. 24와 Fig. 25의 통계 데이터를 분석하여 Table 24에 나타내었다. 분석 결과 멀티미디어와 정보 검색 서비스를 가장 선호하는 희망 서비스로 도출되었다. 따라서 음악 전문 사이트, 동영상 제공 사이트, 여행/교통 정보 제공 사이트 및 뉴스 검색 사이트가 모바일 단말기를 사용하여 이용할 서비스들 중에서 가장 많은 사용자들이 원하는 서비스 종류가 된다. 그러므로 제안한 웹 문서 구조 분석과 콘텐츠 변환은 3장의 Table 8에 나열된 104개의 카테고리 중에서 정보 검색 제공 포탈 사이트 및 멀티미디어 제공 사이트를 웹 구조 분석과 콘텐츠 변환을 수행할 대상 사이트로 선정하였다.

2. 성능 평가 기준

제안된 웹 문서 구조 분석의 성능을 평가하기 위한 기준에서 웹 문서 구조 분석

Table 24. Analysis data of Fig. 24 and Fig. 25.

	모바일 인터넷 사용자 선호도	향후 희망 서비스
서비스 종류 (상위5위)	음악듣기, 벨소리/멜로디, 게임, 신용카드 기능, TV/영화/동영상, 쇼핑/예매/예약 서비스	여행/교통/뉴스 정보 검색, 음악듣기, 신용카드/신분증 기능
분석	상위에 해당되는 서비스는 멀티미디어 서비스에 해당됨	상위에 해당되는 서비스는 정보 검색 서비스에 해당됨

의 성능은 콘텐츠 변환을 위한 전처리 기능에 해당되므로 이는 콘텐츠 변환 모듈의 성능을 평가함으로써 얻을 수 있다. 따라서 웹 문서 구조 분석의 성능 평가 기준은 콘텐츠 변환 성능의 기준에 초점을 둔다.

하지만 콘텐츠 변환의 평가 기준은 다음과 같이 두 가지로 정한다.

첫째, 일반적으로 휴대폰을 이용하여 유선 웹 서비스에 접근하면 대부분은 데스크톱에서의 브라우징 뷰와 휴대폰에서의 브라우징 뷰가 상이하다. 따라서 유선 인터넷과 모바일 인터넷에서의 브라우징 뷰의 일관성 여부를 평가 기준으로 정한다. 유선과 모바일에서의 프리젠테이션의 순서가 서로 일관되는지의 판단은 각각의 환경에서 프리젠테이션 되는 콘텐츠들을 순서적으로 매핑하여 그 순서가 동일한 순서인지를 정량적인 값으로 도출한다. 또한 도출된 수치는 콘텐츠 변환의 신뢰도 및 모바일 단말기의 하드웨어적인 제약성이 극복되었는지의 판단 기준이기도 하다.

둘째, 변환 모듈의 변환 템플릿 작성과 관련된 복잡도를 평가 기준으로 정한다. 이는 XPATH에 표현되는 경로식이며 특정 태그 요소가 존재하고 있는 범위를 지역적 또는 전역적인지를 의미하며 또한 도출된 수치는 웹 문서 구조 분석의 유용성을 의미한다.

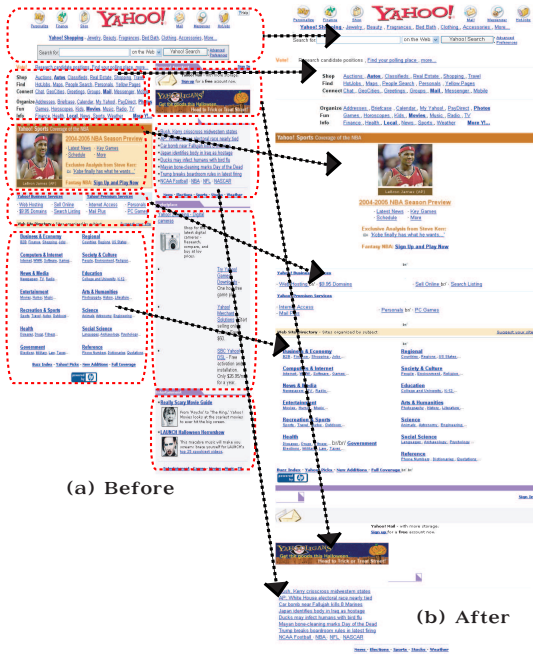


Fig. 26 Results of structure analysis for yahoo.com.

3. 구현 결과

이 절에서는 제안한 웹 문서 구조 분석 결과와 이를 기반으로 한 콘텐츠 변환 방법의 성능을 알아보기 위해 구현 결과를 나타내었다.

1) 웹 문서 구조 분석 구현 결과

먼저 5장 1절에서 작성한 분석 모델에 기준하여 선정된 Table 8에서 정보 검색 및 정보 제공 사이트 중에서 “yahoo.com”의 웹 문서 구조 분석 결과를 Fig. 26에 나타내었다. Fig. 26의 (a)는 Internet Explorer상에 프리젠테이션된 유선 콘텐츠이며, (b)는 제안된 웹 문서 구조 분석에 의해 단일 테이블 기반으로 재구성된 모바일용 콘텐츠를 Internet Explorer상에서 프리젠테이션한 내용이다. 화살표의 의미는 3장의 기본 브라우징 뷰의 개념인 TDLR 순서를 유지하고 있음을 나타낸다.

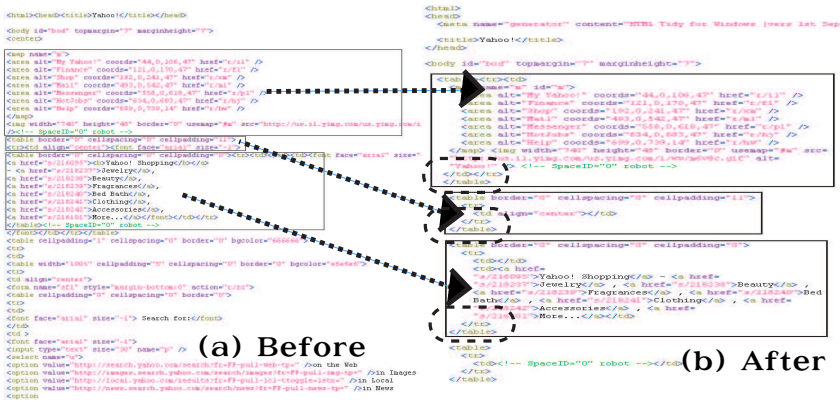


Fig. 27 Results of source code form yahoo.com.

그리고 Fig. 27의 (a)는 유선 콘텐츠의 소스 코드이며, (b)는 (a)를 TIDY를 이용하여 XHTML로 변환한 후 스캐닝 과정 및 제안된 웹 문서 구조 분석 방식을 수행한 후의 소스 코드 결과이다. 단일 테이블 기반 형식의 휴리스틱한 순서로 웹 문서가 재구성되었다. 점선의 동그라미는 웹 문서 구조 분석 과정에서 하나의 단일 테이블을 생성하기 위해 추가 삽입되고 있는 태그 요소를 나타낸다. 또한 (b)에서는 그 대상 코드를 살펴볼 수는 없지만 웹 문서 구조 분석이 완료되면 불필요한 태그 요소(<FONT>) 들이 삭제되었으며, “<FORM>” 태그의 사용 위치 명시화 및 단일 테이블화 등이 수행되었다.

Fig. 26과 동일 방법으로 각각 “amazon.com”과 “buy.com”의 웹 문서를 제안한 구조 분석 방법에 의해 단일 테이블들로 추출한 결과를 Fig. 28과 Fig. 29에 나타내었다. 마찬가지로 TDLR한 휴리스틱 기반으로 콘텐츠 순서가 유지되었다. 그리고 Fig. 30에 나타난 바와 같이 구조 분석 후의 소스 코드는 단일 테이블들로 문서가 재구성되었다.

Fig. 28 Results of structure analysis for amazon.com.

따라서 제안된 웹 문서 구조 분석의 결과는 현재 유선 콘텐츠들을 모바일 콘텐츠로 변환해 나가는 과정에서 XHTML Basic의 미 지원 사항을 해결할 수 있는 포맷으로 변환이 가능하고 또한 불필요한 요소나 명시적 태그 사용의 위치를 지정할 수 있음을 보여 주었다. 하지만 Fig. 28의 (c)에서 보인 바와 같이 변환 시 특수 기호의 변환 왜곡이 발생되었다.

2) 콘텐츠 변환 구현 결과

Fig. 31은 “yahoo.com”의 유선 콘텐츠 일부분이며, Fig. 32는 제안된 웹 문서 구조 분석과 콘텐츠 변환 방법을 수행한 후 Internet Explorer상에 프리젠테이션한 결과이고 Fig. 33은 모바일 에뮬레이터상에 프리젠테이션한 결과이다. 모바일 단말기 상에서의 브라우징 뷰가 유선 환경의 그것과 휴리스틱하게 일관되도록 콘텐츠

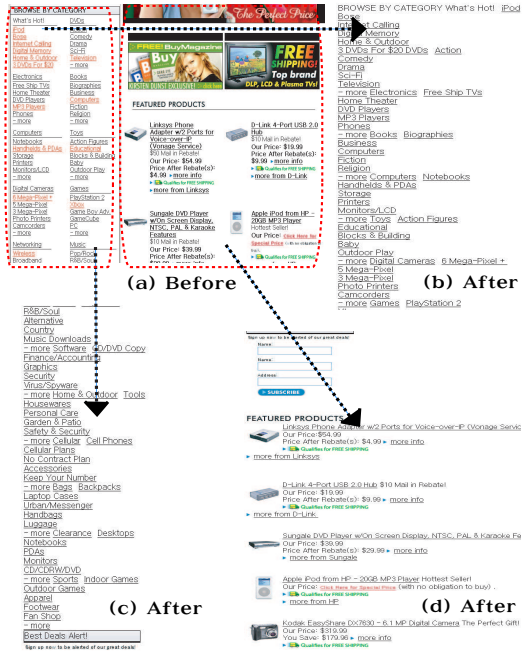


Fig. 29 Results of structure analysis for buy.com.

들이 세로로 재구성되었음을 알 수 있다.

또한 Fig. 31, Fig. 32, Fig. 33과 같은 방식으로 Fig. 34, Fig. 35, Fig. 36 그리고 Fig. 37, Fig. 38, Fig. 39을 나타내었다. 마찬가지로 모바일 단말기 상에서의 브라우징 뷰가 유선 환경의 그것과 휴리스틱하게 일관되도록 콘텐츠들이 세로로 재구성되었다.

4. 성능 평가 및 결과 분석

이 절에서는 제안한 웹 문서 구조 분석 결과와 이를 기반으로 한 콘텐츠 변환 결과를 토대로 본 연구에서 제안된 모바일용 콘텐츠 변환을 위한 웹 문서 구조 분석과 변환에 대한 결과를 분석하였다.

Fig. 30 Results of source code for amazon.com and buy.com.

Fig. 31 A part of yahoo.com in internet explorer.

1) 성능 평가

(1) 브라우징 뷰의 일관성

브라우징 뷰는 유선과 모바일 장치간의 하드웨어적인 차이점으로 인하여 직접 비교할 수는 없다. 하지만 3장에서 제시된 휴리스틱한 브라우징 순서인 TDLR 방식으로 유선 브라우징과 모바일 브라우징간을 비교할 수 있으며 이는 변환전의 콘텐츠 배열 순서와 변환후의 콘텐츠 배열 순서를 서로 순서적으로 비교함으로써 가능하다.

따라서 원본 문서와 타겟 문서와의 프리젠테이션 순서를 비교하기 위해 Table 25의 비교 기준 단위 및 전제 조건과 측정하기 위해 제안하는 Table 26의 알고리즘, 그리고 원본과 타겟 문서에서 텍스트를 추출하기 위한 Table 27의 XSL 템플릿을 이용하여 Table 28의 정량적 수치를 도출하였다.

Table 25. Estimating criterion.

비교 기준	매칭된 글자(단어) 수 / 전체 글자(단어) 수	
	글자 단위 매칭	단어 단위 매칭
전제 조건	모든 공백 제거	공백 한 개를 제외한 모든 공백 제거
	전각 기호가 반각 또는 알 수 없는 기호로 변환되는 상황 에러	

Table 26. Reliability algorithm.

```
// Matching for between the characters
array source[ ] = stream that all empty blank removed between characters
array target[ ] = stream that all empty blank removed between characters
for( i=0; i ¡ source.length(); i++) {
        if(source[i] == target[i]) matched += 1;
}
reliability between the characters = ( matched / source.length() ) × 100

// Matching for between the words
list slist() = stream that all empty blank removed between words
                except one empty blank between words
list tlist() = stream that all empty blank removed between words
                except one empty blank between words
for( i=0; i ¡ slist.count(); i++) {
        if(slist(i) == tlist(i)) matched += 1;
}
reliability between the words = ( matched / slist.count() ) × 100
```


Table 27. XSL template for text extraction.

```
<xsl:template match="/">
<html>
<xsl:apply-templates select="//text()"/>
<html>
<xsl:template match="text()"><xsl:value-of select="."/></xsl:template>
```

Table 28. Estimated value of contents assembling sequence after contents conversion.

대상 사이트 (26개 사이트)	콘텐츠 배열 순서		
	글자 단위 매칭	단어 단위 매칭	전각 기호 사용
amazon.com	99.95%	98.78%	◯
bbc.co.uk	99.84%	99.33%	◯
buy.com	100%	100%	×
cam.ac.uk	99.66%	98.41%	◯
chosun.com	99.33%	97.91%	◯
cornell.edu	99.77%	98.88%	◯
elibrary.com	99.72%	98.54%	◯
harvard.edu	99.83%	99.07%	◯
kr.yahoo.com	97.65%	92.86%	◯
latimes.com	100%	100%	×
lcweb.loc.gov	100%	100%	×
lycos.com	99.68%	98.31%	◯
mobile.org	99.93%	99.65%	◯
music.com	99.81%	98.98%	◯
naver.com	98.09%	94.50%	◯
nytimes.com	99.98%	99.91%	◯
ox.ac.uk	100%	100%	×
stanford.edu	100%	100%	×
umn.edu	99.66%	98.25%	◯
usatoday.com	99.96%	99.80%	◯
vlib.org	99.84%	98.81%	◯
washingtonpost.com	99.98%	99.88%	◯
winamp.com	99.85%	99.27%	◯
wsj.com	100%	100%	×
yahoo.com	97.86%	88.05%	◯
yale.edu	98.59%	92.31%	◯
Average	99.58%	98.13%	

Fig. 32 Conversion result of yahoo.com in internet explorer.

Table 28의 도출된 수치 값에 대한 결과를 분석해보면 거의 대부분이 콘텐츠 왜곡 없이 변환이 수행되었다. 하지만 원본 문서 내에서 하나의 아이템을 구별하기 위하여 일반적으로 사용되고 있는 글머리표 기능의 특수 기호인 전각 문자들이 변환 후 반각 문자 또는 알 수 없는 기호로 변환이 되었으며 이러한 변환 오류에 해당되는 내용을 Table 29에 정리하였다. 그리고 Table 29에 속하는 사항들은 전제 조건에서 제시한 바와 같이 변환 에러로 처리하였으며, 결과적으로 콘텐츠 배열 순서를 상이하게 하여 브라우징 뷰의 일관성을 감소시키는 원인이 되었다. 하지만 이러한 변환 오류는 제안된 웹 문서 구조 분석 및 변환 모듈이 아닌 원본 웹 문서를 XHTML로 변환시키는 전처리 과정을 수행하는 TIDY에서 발견되었으며, 향후 TIDY의 인코딩 기능을 보완하면 브라우징 뷰의 일관성은 도출된 수치 값보다 높아질 것으로 보인다. 따라서 본 연구에서 제안한 방법으로 웹 문서를 분석하고 변환을 수행하게 되면 평균 98.86%의 유·무선간 브라우징 뷰 일관성을 지원하게 되었다.

Fig. 33 Conversion result of yahoo.com in mobile browser.

Table 29. Error characters.

Before	After
^{TM}, ⓒ, …, ·, “, ”, ‘, ’, ↓, Ⓡ, 전각 공백 등 전각 특수 기호	? or non character

(2) 콘텐츠 변환 신뢰도

콘텐츠 변환 신뢰도는 원본 웹 문서의 모든 콘텐츠가 왜곡 없이, 그리고 빠짐 없이 타겟 문서로 변환되었는지를 의미한다. 그러므로 콘텐츠 변환 신뢰도의 측정 기준은 원본과 타겟 간 일대 일 대응 관계와 동일하며, 또한 콘텐츠 변환 신뢰도는 브라우징 뷰의 일관성 검증 방식과 동일하다. 즉 원본과 타겟 간 변환 신뢰도가 존재하지 않는다면 브라우징 뷰를 검증할 수 없기 때문이며, 또한 브라우징 뷰의 일치성이 낮다면 그 만큼 변환 신뢰도를 보장할 수도 없기 때문이다. 따라서 본 연

Fig. 34 A part of amazon.com in internet explorer.

구에서의 콘텐츠 변환 신뢰도는 브라우징 뷰와 동일한 수치 값을 갖게 되며, 평균 98.86%의 콘텐츠 변환 신뢰도를 확보였다.

(3) 모바일 단말기의 하드웨어 제약성 극복

모바일 단말기의 다양한 하드웨어 제약성 중에서 극복하기 위한 사항은 3장에서 전술한 바와 같이 모바일 단말기 디스플레이 환경으로 인해 발생되는 유·무선간 브라우징의 차이점이였다. 따라서 본 연구는 유·무선간 브라우징을 휴리스틱한 방식으로 일치하도록 하였으며, 그 결과는 위에서 도출된 브라우징 뷰의 일관성이다.

(4) 템플릿 작성의 복잡도

템플릿 작성의 복잡도는 중첩된 테이블들과 콘텐츠들을 검색하기 위한 XSL의 경로식인 XPATH의 검색식에 대한 평가이다. 비구조적인 웹 문서는 레이아웃 테이블과 중첩 테이블들이 존재함으로 Table 30의 (a)와 같이 XPATH 검색식에는 반드시 재귀적 특성인 “//(all의 의미)”의 사용이 불가피하다(XPATH). 하지만 제안된 콘텐츠 변환 템플릿은 단일 테이블 기반으로 분석된 웹 문서에 적용함으로 “//” 사용이 필요 없다. 이는 웹 문서 구조 분석 단계에서 중첩된 테이블들이 제거되었기

Fig. 35 Conversion result of amazon.com in internet explorer.

Table 30. Example of searching expression of XPATH.

(a)	(b)
<xsl:template match="//table">	<xsl:template match="table">

때문이다. 따라서 Table 30의 (b)와 같이 템플릿 작성이 가능함으로써 템플릿 작성 복잡도가 감소되었고 이는 웹 문서 분석의 결과가 주는 유용성이라 볼 수 있다.

2) 결과 분석

본 연구에서 제안한 웹 문서 구조 분석과 콘텐츠 변환 방법의 연구 결과는 다음과 같다.

첫째, 비구조적이고 정형화되지 않은 웹 문서를 XML 기반의 중첩되지 않은 단일 테이블 구조로 재구성하였다. 그럼으로써 문서 구조가 지원되지 않더라도 XML

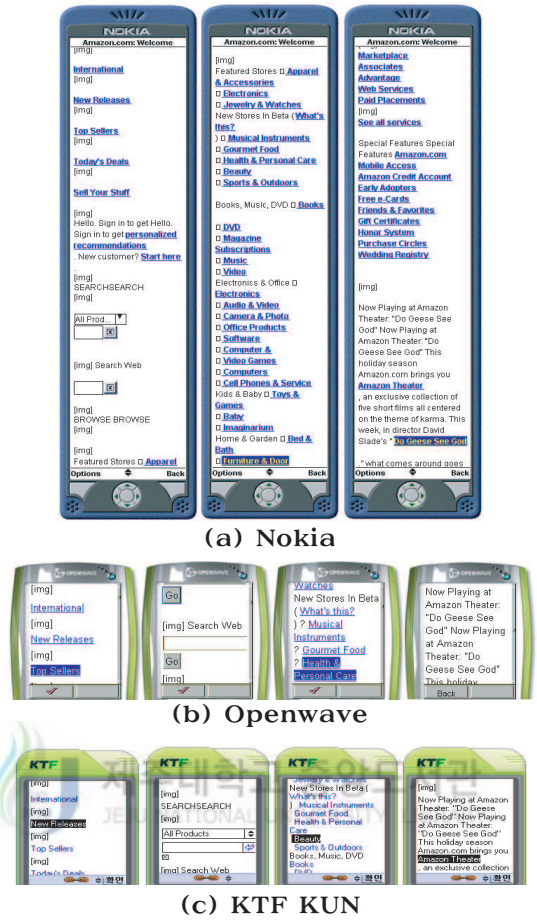


Fig. 36 Conversion result of amazon.com in mobile browser.

응용 애플리케이션들이 기본적인 문서의 구조를 알 수 있기 때문에 더 쉽게 접근이 가능하며, 또한 XML 기반이므로 Legacy 모바일 서비스 및 콘텐츠를 브라우징 할 수 있는 TV, Car Navigation 시스템과 연동이 가능하다.

둘째, 복잡한 레이아웃과 중첩된 테이블들이 사용된 웹 문서를 단위 테이블들의 집합으로 추출하고 변환하여 모바일 환경에서도 유선 환경과 일치되는 휴리스틱한 브라우징 순서를 갖도록 재구성하였으며, 콘텐츠의 변환은 왜곡 없이 변환되어 변환 신뢰도를 확보하였다. 그럼으로써 유선 콘텐츠를 모바일 환경에서 유선 환경과 일치되는 브라우징 뷰를 갖도록 할 수 있으며, 소형 모바일 단말기의 디스플

Fig. 37 A part of buy.com in internet explorer.

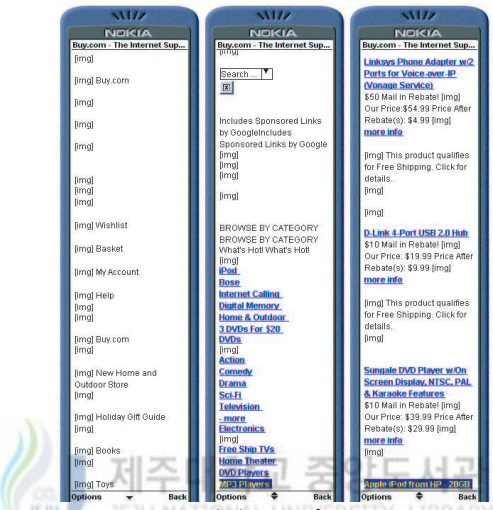
레이 제약성을 극복하였다.

셋째, 웹 문서를 웹 문서 구조 분석에 의해 단위 테이블로 단순화 및 경량화 시켰다. 그럼으로써 변환 모듈의 변환 템플릿인 XPATH 검색식을 단순화 시켰다.

넷째, 콘텐츠 변환의 변환 방식은 기존의 필터링 기반의 규칙을 사용하면서 변환의 확장성과 유연성을 갖도록 XSLT 기술인 XSL 기법을 사용했으며, XHTML Basic의 모듈 기반으로 변환하였다. 그럼으로써 새로운 변환 규칙의 추가 및 삭제가 용이하다.

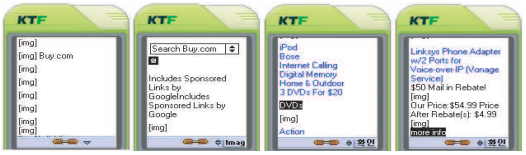
끝으로 제안된 연구 방식을 적용하면 현재 접근되고 있지 않는 유선 콘텐츠들을 모바일 환경으로 서비스 할 수 있다. 따라서 모바일용 콘텐츠 재생성의 비용을 절감시킬 수 있다.

Fig. 38 Conversion result of buy.com in internet explorer.



(a) Nokia

(b) Openwave



(c) KTF KUN

Fig. 39 Conversion result of buy.com in mobile browser.

VI. 결론

본 연구에서는 모바일 콘텐츠 변환을 위해, 웹 문서 구조 분석 방법을 제안하였다. 그리고 구조 분석 결과의 유용성을 입증하기 위하여 구조 분석의 결과를 입력받아 모바일 콘텐츠를 생성하는 콘텐츠 변환 방법을 제안하였다. 또한 콘텐츠 변환 방법의 성능을 평가하기 위한 분석 모델을 작성하였으며, 분석 모델 기반으로 제안된 웹 문서 구조 분석과 콘텐츠 변환 방법의 성능을 기존 연구들과 비교하여 다음과 같은 결과를 도출했다.

본 연구에서 제안한 웹 문서 구조 분석의 결과로는 첫째 웹페이지내의 페이지 레이아웃들을 제거하고 콘텐츠를 포함하는 테이블들만 단위 테이블들로 추출했으며, 추출된 단위 테이블들은 휴리스틱하게 재구성하였다. 둘째, 디스플레이가 작은 모바일 단말기상에 단위 테이블 단위로 콘텐츠 제공 및 모바일 콘텐츠 변환 작업 시 유선 인터넷 콘텐츠의 복잡성을 줄였다. 셋째, XML 기반이므로 다른 XML 기반 응용 애플리케이션들이 목적에 따라 사용할 수 있고, 넷째, 모바일 콘텐츠 생성의 부담을 줄였으며 콘텐츠 변환 모듈의 변환 템플릿 복잡도를 감소시켰다. 다섯째, 웹 문서 구조 분석의 결과는 웹 문서의 콘텐츠를 왜곡하지 않는 콘텐츠 신뢰도를 확보하였으며, 끝으로 유선 인터넷 콘텐츠가 모바일 인터넷에서도 휴리스틱한 브라우징 일관성을 제공하였다.

그리고 제안된 콘텐츠 변환 방법은 단순화되고 경량화된 웹 문서인 유선 인터넷 콘텐츠를 WAP 2.0과 XHTML Basic 기반의 모바일 서비스 환경으로의 변환으로써, 그 연구 결과는 첫째, 기존 HTML 필터링 기반의 태그 단위 변환을 논리적으로 개선한 XHTML Basic 기반 모듈 중 "TABLE MODULE"로 추상화하였고, 둘째, 변환을 위한 변환 규칙 및 변환 템플릿을 작성하였다. 또한 Legacy 모바일 서비스 환경을 위한 서비스 메타 모델을 작성하였다.

본 연구에서 수행한 모바일용 콘텐츠 생성을 웹 문서 구조 분석과 변환의 기대효과 및 활용 방안은 첫째, 웹 문서 구조 분석의 결과는 XML 기반이므로 XML 응용 애플리케이션 접근이 가능하고, 둘째, 향후 WAP 1.X가 WAP 2.0 기반으로 전환이 이루어지면 XHTML Basic이 활성화됨으로 Legacy 웹 문서를 모바일 콘텐츠로의 변환 시 단말기 별 추가 템플릿 생성 작업을 최소화 할 수 있으며, 셋째, 웹 문서 구조 분석 과정상의 XHTML 문서의 생성으로 인터넷 브라우징이 가능한 TV 및 가정용 홈시어터, 또한 CAR Navigation 등의 디스플레이 장치로 웹 페이지를 서비스 할 수 있다. 또한 현재 서비스되고 있는 유선 인터넷 콘텐츠를 모바일 단말기로 서비스가 가능하다.

본 연구에서 개선되어야 할 문제점 및 향후 연구는 다음과 같다.

첫째, 무선 인터넷 환경이 아직은 WAP 2.0과 XHTML Basic으로 구현되는 상황이 미흡한 실정이다. 이는 모바일 하드웨어 및 소프트웨어, 그리고 이동 통신사간의 이해관계 등이 원인이 되고 있으며, 따라서 현 시점에서 적용하기 위해서는 WML, mHTML, cHTML, sHTML등의 모바일 언어로 변환하는 XSL 템플릿 변환

규칙이 추가 되어야 한다.

둘째, 유선 인터넷 콘텐츠의 동적인 다양한 요소들을 제공하기 위해, XHTML Basic에서 해당 요소들을 지원함과 동시에 변환 템플릿에 해당 템플릿 추가 및 이미지 변환 툴과의 연동이 요구된다.

셋째, 테스크탑 화면 기준의 콘텐츠를 모바일 단말기 화면상에 적응적으로 최적화하여 보여주기 위한 작업이 필요하다. 이는 단말기 특성을 분석하고 웹 문서 구조 분석 후 콘텐츠 재구성 작업 시 모바일 헤더파일 특성별로 웹 문서를 재구성함으로써 해결 가능하리라 본다.

넷째, 웹 문서 구조 분석의 측면에서 HTML 문서를 XHTML로 변환하는 TIDY 같은 변환 자동 툴의 개선이 필요하다.

마지막으로 본 연구에서 제안한 모바일용 콘텐츠 생성을 위한 웹 문서 구조 분석과 변환을 상용 서비스에 지원하기 위해서는 유선과 모바일 서비스를 통합하여 제공하는 통합 서버와의 연동이 필요하다.

참 고 문 헌

B. C. Housel, G. Samaras, and D. B. Lindquist, "WebExpress : A Client/intercept Based System for Optimizing Web Browsing in a Wireless Environment," *Mobile Networks and Applications,* vol. 3, no. 4, pp. 419-431, 1998.

B. N. Schilit, J. Trevor, D. M. Hibert, T. K. Koh, "m-Links:An Infrastructure for Very Small Internet Devices," *7th Annual Int'l Conf. on Mobile Computing and Networking,* pp. 122-131, July 2001.

변시우, 변숙은, "효율적인 모바일 비즈니스를 위한 WIPI 플랫폼에 관한 연구," *한국인터넷비즈니스학회논문지* vol. 4, no. 2, pp. 79-93, 2003. 1

CC/PP, "Composite Capability/Preference Profile(CC/PP):Structure and Vocabularies 1.0," *http://www.w3.org/TR/2004/REC-CCPP-struct-vocab-20040115/*

최지원, 김기천, "무선전용 다중 언어의 번역을 지원하는 변환기의 구현," *정보처리학회논문지,* vol. 9, no. 2, pp. 293-296, 2002. 4.

최우영, 허신, "모바일 표준 플랫폼 WIPI를 위한 WAP 2.0 마이크로 브라우저의 설계 및 구현," *정보과학회학술대회,* vol. 30, no. 1, 2003. 4.

최은정, 한동원, 임경식, "무선 인터넷 서비스를 위한 WAP 게이트웨이용 WML 컴파일러의 설계 및 구현," *정보과학회논문지,* vol. 7, no. 1, pp. 165-182, 2001.

최용길, "국내 모바일 무선 인터넷의 현황 및 전망," *인터넷정보학회지,* vol. 4, no. 2, pp. 10-18, 2003. 6.

TIDY, "HTML Tidy Library Project," *http://tidy.sourcefore.net/*

D. W. Embley, Y. S. Jiang, and Y. Ng, "Record-Boundary Discovery in Web Document," *Int'l Conf. on Management of Data(SIGMOD'99),* pp. 467-478, June 1999.

N. K. Sharma, "Enhancing Wireless Internet Performance," *IEEE Communications Surveys & tutorials,* vol. 4, no. 1, pp. 2-15, 2002.

HDML, "Handheld Device Markup Language Spec," *http://www.w3.org/TR/NOTE-Submission-HDML-spec.htm*

한국정보문화센터, "2003년 무선인터넷이용실태조사 최종보고서," 2003. 9.

H. M. Deitel, P. J. Deitel, T. R. Nieto, K. Steinbuhler, "Wireless Internet & Mobile Business How to Program," *Prentice Hall,* 2002.

HTML 4.0, "HTML 4.0 24-Apr-1998 Recommendation," *http://www.w3.org/TR/1998/REC-html40-19980424*

HTML 4.1, "HTML 4.01 24-Dec-1999 Recommendation," *http://www.w3.org/TR/1999/REC-html401-19991224*

i-Mode, "i-Mode : FAQ Frequently Asked Questions about i-Mode and the answer," February 1998.

장영건, "내장 문자와 사전 구조 지식을 이용한 HTMLtoVXML 변환 에이전트 개발," *정보처리학회논문지,* vol. 10, no. 2, pp. 343-350, 2003. 4.

정보통신기술경영연구소, "무선 인터넷 산업 현황 및 발전 전략," *무선산업연구팀,* 2000. 10.

정보통신학술 연구과제, "온라인 컨텐츠 산업 활성화를 위한 컨텐츠 사업자와 망사업자와의 바람직한 관계조사 연구," *정보통신부,* 2002.

정재목, 김형주, "웹 정보의 추출 및 통합을 위한 래퍼 시스템," *정보과학회논문지,* vol. 9, no. 5, pp. 551-559, 2003. 10.

조수선, 이동우, 신희숙, 황치정, "모바일 웹 서비스를 위한 컨텐츠 재작성 기술," *인터넷정보학회논문지,* vol. 3, no. 5, pp. 63-72, 2002. 10.

조승호, 차정훈, "Clipping 기반의 무선 인터넷 사이트 구축용 변환 서버 구현," *정보처리학회논문지,* vol. 11, no. 2, pp. 165-174, 2004. 4.

강경용, "무선 인터넷 서비스를 위한 계층 구조의 Deck를 갖는 HTML Filter의 구현 방안," *컴퓨터산업기술학회논문지,* vol. 3, no. 2, pp. 179-184, 2002. 2.

KFT 포탈 기획팀, "무선 인터넷 컨텐츠 발전," 2003. 8.

K. Nagao, Y. Shirai, and K. Squire, "Semantic Annotation and Transcoding:Making Web Content More Accessible," *IEEE MultiMedia,* vol. 8, no. 2, pp. 69-81, April. 2001.

강성천, 정광수, "이동단말을 위한 적응적 웹 문서 변환," *정보과학회논문지,* vol. 6, no. 6, pp. 635-642, 2000. 12.

강태규, 김도영, 김봉태, "유무선 통합 네트워크에서의 VoIP를 위한 공통 논리 기능 구조 분석," *전자통신동향분석,* vol. 17, no. 5, 2002. 10.

고혁준, 김정희, 곽호영, “XML-RPC 기반의 분산환경 문서관리 시스템 모델,” *한국해양정보통신학회논문지,* vol. 8, no. 2, pp. 394-406, 2004. 4.

김경아, 용환승, “PC 및 PDA 브라우저 지원을 위한 XML 기반의 웹 컨텐츠 개발 사례 연구,” *한국디지털컨텐츠학회논문지,* vol. 3, no. 1, pp. 59-74, 2002. 6.

김규정, “예제로 배우는 무선 인터넷 프로그래밍,” *가메출판사,* 2002. 5.

김기천, “모바일 서비스 기술동향,” *정보처리학회논문지,* vol. 9, no. 2, pp. 17-23, 2002. 3.

김신효, 정병호, “스마트카드를 이용한 무선 인터넷 보안 기술 현황 및 전망,” *정보과학회지,* vol. 20, no 4. pp. 14-20, 2002. 4.

김정희, 곽호영, “서버 간 요청 위임을 고려한 XML 기반 RPC 자원 서비스 시스템 설계 및 구현,” *한국멀티미디어학회논문지,* vol. 6, no. 6, pp. 1100-1110, 2003. 10.

김정희, 곽호영, “Edge-Labeled Graph를 적용한 XML 저장 모델,” *한국해양정보통신학회논문지,* vol. 7, no. 5, pp. 993-1001, 2003. 10.

김정희, 곽호영, “Legacy 데이터베이스를 위한 DTD의 IDREF-ID 속성관계 모델링,” *인터넷정보학회논문지,* vol. 3, no. 3, pp. 31-38, 2002. 6.

김학범, “M-Commerce 보안기술 동향,” *IT Forum Korea,* 2002.

김환근, 강형일, 유재수, 최한석, “HTML-WML 변환기 설계 및 구현,” *데이터베이스연구회논문지,* vol. 16, no. 2, pp. 55-66, 2000. 12.

이경호, 최윤철, 조성배, “문서 영상의 정교한 기하적 구조 분석을 위한 지식베이스 시스템,” *정보과학회논문지,* vol. 28, no. 11, pp. 795-812, 2001. 11.

이동근, 김기조, 임경식, “무선 응용 프로토콜 보안 기술,” *정보과학회지,* vol. 20, no. 4, pp. 58-65, 2002. 4.

이정환, “Mobile Beginner’s Guide,” *삼양출판사,* 2002.

Microsoft, “Microsoft Mobile Explorer 1.0 Specification,” 1999. 3

ME, “http://www.microsoft.com/mobile/”

M. Hori, G. Kondoh, K. Ono, S. Hirose, S. Singhal, “Annotation-Based Web Content Transcoding,” *Porc. of the 9th International World Wide Web Conference,* May 2000.

M. Metter, Dr Robert Colomb, "WAP enabling existing HTML application," *User Interface Conference 2000,* pp. 49-57, 2000

Modularization of XHTML, "http://www.w3.org/TR/2001/REC-xhtml-modularization-20010410"

M. Shi, X. Shen, J. W. Mark, "A light weight authentication scheme for mobile wireless internet application," *WCNC 2003, IEEE Wireless Communications and Networking Confierence,* vol. 4, no. 1, pp. 2126-2131, March 2003.

민영수, 강형일, 유재수, "무선 인터넷을 위한 HTML-WML 변환기 설계 및 구현," *인터넷학회논문지,* vol. 2, no. 2, pp. 37-50, 2001. 6.

오금용, 황인준, "유사 패턴을 갖는 HTML 문서의 XML 자동 변환," *정보처리학회논문지,* vol. 9, no. 3, pp. 355-364, 2002. 6.

Phone.com, "WapGateway, http://www.phone.com"

박기현, 강동우, 권정선, "HTML 필터 기능을 갖춘 WAP 게이트웨이 시스템 구축," *정보과학회논문지,* vol. 7, no. 4, pp. 350-358, 2001. 6.

박기현, 신양모, 주홍택, "WAP 프록시의 구축 및 무선통신 효율을 위한 개선," *정보처리학회논문지,* vol. 11, no. 3, pp. 379-386, 2004. 6.

RDF, "Resource Description Framework," *http://www.w3.org/TR/2004/REC-ref-syntax-grammar-20040210/*

R. Kalden, I. Meirick, M. Meyer, "Wireless Internet Access based on GPRS," *IEEE Personal Communications,* pp. 8-18, 2000. 4

R. Han, P. Bhagwat, "Dynamic Adaptation In an Image Transcoding Proxy for Mobile Web Browsing," *IEEE Personal Communications Magazine,* pp. 8-17, December 1998.

S. Chandra, C. S. Ellis, A. Vahdat, "Differentiated Multimedia Web Services Using Quality Aware Transcoding," *INFOCOM 2000,* pp. 961-969, March 2000.

sHTML, "SHTML Specification, SAMSUNG Electronics Co," 1999

송동리, 황인준, "무선 단말기를 위한 웹 페이지의 자동 재구성," *정보처리학회논문지,* vol. 9, no. 5, pp. 523-532, 2002. 10.

소프트뱅그미디어, "무선 인터넷 백서 2002," *무선 인터넷 백서 편찬 위원회,* 2002.

신희숙, 마평수, 조수선, 이동우, "소형 화면 단말기를 위한 웹 문서 변환 기법," *정보처리학회논문지,* vol. 9, no. 6, pp. 1145-1156, 2002. 12.

T. W. Bickmore, B. N. Schilit, "Digestor : Device Independent Access to the World Wide Web," *6th Int'l World Wide Web Conf.,* pp. 655-663, April 1997.

T. W. Bickmore, A. Girgensohn, J. W. Sullivan, "Web Page Filtering and Re-Authoring for Mobile Users," *The Computer Journal,* vol. 42, no. 6, pp. 534-546, 1999.

Unwired Planet, "Handheld Device Markup Language(HDML) 2.0 Specification," 1997. 4.

WAP 2.0, "Wireless Application Protocol Architecture Version 12," 2001.

WAP 2.0-1, "WAP 2.0 Technical White Paper," *http://www.wapforum.org/*

WAP Forum, "*http://www.wapforum.org/*"

WML 1.1, "Wireless Markup Language(WML) 1.1 Specification," 1999. 1.

WML 2.0, "Wireless Markup Language Spec," *http://www.wapforum.org/what/technical.htm*

WWW, "*http://www.w3.org/*"

XHTML, "XHTML 1.0 Recommendation Second Edition," *http://www.w3.org/TR/2002/REC-xhtml1-20020801*

XHTML Basic, "W3C Recommendation, 19 December 2000," *http://www.w3.org/TR/2000/REC-xhtml-basic-20001219*

XML, "XML 1.0 Third Edition Recommendation," *http://www.w3.org/TR/2004/REC-xml-20041204*

XSL and XSLT, "*http://www.w3.org/Style/XSL/*"

XPATH, "XML Path Language," *http://www.w3.org/TR/1999/REC-xpath-19991116*

윤성일, 송정길, "Mobile 기반의 유무선 플랫폼 통합 컨텐츠 변환기 설계 및 구현," *컴퓨터산업기술학회논문지,* vol. 3, no. 9, pp. 1295-1314, 2002. 9.

유영환, "무선 인터넷을 위한 WPAN 기술," *인터넷정보학회지,* vol. 3, no. 1, pp. 28-33, 2002. 3.

양서민, 이혁준, “컨텐츠 제공자 지정 웹 클리핑 방식의 이동 인터넷 컨텐츠 변환,” *정보처리학회논문지,* vol. 11, no. 8, pp. 35-44, 2004. 2.

양해술, 최민용, 황석형, “무선 인터넷을 위한 컨텐츠 변환 시스템의 설계 및 구현,” *정보처리학회논문지,* vol. 11, no. 5, pp. 1073-1086, 2004.

Yong-Woon, Kim, “Review and Forecast of Mobile Internet Technologies,” *Z-TE FutureTel Co., LTD.* White Paper, 2001. 10.

Zionwap, “http://www.zionwap.net”

국 문 초 록

모바일용 콘텐츠 生成을 위한 웹 文書 構造 分析과 變換

김 정 희
정보공학과
제주대학교 대학원

본 연구에서는 모바일 환경에서 유선용 인터넷 서비스를 지원할 수 있도록 하는 웹 문서 구조 분석과 함께 복잡한 유선용 콘텐츠로부터 모바일 콘텐츠를 생성하는 변환 방법을 제안한다.

또한 제안된 두 가지 연구 결과에 대한 테스트 베드로서 사용자 선호도를 기준으로 성능 평가를 위한 분석 모델과 평가 기준을 작성하였다.

웹 문서 구조 분석은 WAP 게이트웨이 콘텐츠 변환 모듈이 웹 문서를 필터링하기 이전 단계에서 변환할 웹 문서를 XML 기반 문서인 XHTML로 재구성 하며 또한 웹 문서내의 중첩 페이지 레이아웃 구성 요소들을 단일화 하고, 동시에 FORM 관련 요소들의 사용 위치를 명시화 한다. 그리고 동적인 요소들과 불필요한 요소들을 제거하여 웹 문서를 경량화 시킨다.

제안한 웹 문서 구조 분석을 적용하면 모바일 단말기들의 하드웨어 제약성을 극복할 수 있을 뿐만 아니라 유선과 모바일 환경의 브라우징 뷰를 일관성 있게 유지할 수 있고, 콘텐츠 변환 과정의 XSL 패턴 템플릿을 단순화 시킬 수 있다. 또한 웹 문서를 다양한 모바일 장치에서 재사용이 가능하도록 단일 테이블들로 구성된 경량의 XHTML 문서로 생성 시킬 수 있다.

그리고 콘텐츠 변환은 경량의 웹 문서를 WAP 2.0 환경의 모바일 마크업 언어로 채용된 XHTML Basic언어로의 변환 방법이며, 이는 XHTML Basic 모듈성 기반의 변환 규칙을 담고 있는 XSL 변환 템플릿 문서를 웹 문서 구조 분석의 결과인 단일 테이블에 적용한 변환 기법이다.

끝으로 제안된 웹 문서 구조 분석과 콘텐츠 변환 과정을 적용함으로써 유선과 모바일 환경사의의 브라우징 뷰 일관성을 유지함과 동시에 변환 기법상의 XPATH 복잡도를 줄일 수 있음을 정량적으로 검증하였다.

감 사 의 글

돌이켜 보면 오랜 시간을 열심히 공부하지도 않으면서 마치 중요한 일을, 또한 소중한 연구를 하는 모습으로 비춰졌었던 지난날 들이였던 것 같습니다. 그렇게 늘 부족하였던 제가 곽호영 지도교수님의 지도와 배려 덕분에 세번째 사각모를 쓸 수 있게 되었고, 또한 저를 알고 있는 분들께 논문을 내놓을 수 있게 되었습니다. 그래서 지도교수님께 감사의 말씀을 다시 한번 더 올리지 않을 수 없습니다.
교수님, 고맙습니다.

또한 논문 심사와 지도를 위해 열정을 아끼지 않으신 김장형 교수님, 안기중 교수님, 변상용 교수님께도 감사의 말씀을 올립니다. 그리고 바쁘신 일정에서도 먼 길을 마다 않으시고 내려오셔서 꼼꼼하게 지적해주신 원유헌 교수님께도 이 지면을 통하여 감사의 말씀을 올립니다. 또한 이상준 교수님, 송왕철 교수님, 변영철 교수님, 그리고 김도현 교수님께도 보다 학술적인 접근을 바라는 말씀과 충고를 아끼지 않아 주셨음에 감사 드립니다.

이 논문을 완성하기까지 주위의 모든 분들로부터 수많은 도움을 받았습니다. 같은 연구실에서 함께 공부하고 생활한 동료, 늘 웃음 잃지 않으면서 연구실 분위기를 이끄는 인석, 얼핏 보기엔 연구실 큰형 같고 꺼벙하지만 논리와 원리로 무장한 훈, 항상 뒷처리를 해줘야만 결과를 낸다고 투정 부리는 제게 고운말만 쓰면서 사투리 없던 석건, 그리고 이 논문의 구현을 도와주면서 한쪽에선 열심히 축구(FIFA)하던 봉남, 모두에게 감사의 말을 전합니다. 또한 성철 형님과 경복이게도 좋은 결과가 있기를 이 지면을 통해 전합니다. 그리고 어디선가 열심히 프로그램을 코딩하고 계실 행진 형과 기획을 하고 있을 동현에게도 고마움을 전합니다.

그리고 연구실은 다르지만 영도, 강석 선배님, 항상 친구같은 영민과 재경, 제출할 서류를 챙겨줬던 은경, 그리고 정아 선생님을 비롯하여 학과 모든 연구실 식구들, 또한 좋은 벗들의 모임인 일오인, 자칭 월남 스키부대의 영민과 광민, 산업정보대학 컴퓨터 정보계열 교수님들께도 그동안의 배려와 격려에 고마움을 전합니다.

늘 열심히 하려고 했던 저의 뒤에서 항상 첫째 자리를 튼튼하게 지켜주시고 또한 동생을 믿음으로 지켜봐 주신 큰형님과 형수님, 전화 통화할 때마다 잘하라고 말씀해 주셨던 둘째 형님, 그리고 논문의 마지막 편집을 꼼꼼하게 살펴주신 셋째 형님, 너무 많은 빚을 그동안 진 것만 같습니다. 미흡하지만 이 지면으로 감사와 고마움을 대신합니다. 또한 사랑하는 현정과 귀여운 채영, 어렵고 힘들어도 행복한 날들을 위해 늘 함께 할 수 있기를 바랍니다.

끝으로 오늘의 제가 있을 수 있도록 사랑으로 키워 주신 부모님, 겨울이 되어 추위가 심해지면 늘 염려하면서도 제대로 찾아뵙지 못했던 것 같습니다. 근사한 식당에 한번도 모셔보지 못한 마음이 늘 죄송스러웠습니다. 이 작은 결실이 조금이나마 저의 마음을 대신할 순 없겠지만 감사의 마음으로 부모님께 이 학위 논문을 바칩니다. 정말 오래 오래 사시고 항상 건강하십시요. 그리고 늘 고맙습니다.

2004년의 끝자락, 연구실에서

김 정 희