

# 모바일 웹 서비스를 위한 콘텐츠 재작성 기술

## Contents Re-authoring Technologies for Mobile Web Services

조 수 선\*  
Soo-Sun Cho

이 동 우\*\*  
Dong-Woo Lee

신 희 숙\*\*\*  
Hee-Sook Shin

황 치 정\*\*\*\*  
Chi-Jung Hwang

### 요 약

최근 모바일 인터넷을 위한 각종 단말 기기와 함께 관련 인프라의 구축이 매우 빠른 속도로 진행되고 있다. 이에 따라 모바일 인터넷 상의 여러 가지 서비스 모델이 등장하고 있으며 기존의 방대한 유선 웹 콘텐츠 자원을 무선 웹에서 재사용할 수 있는 방법들이 다양한 각도로 연구되고 있다. 본 논문에서는 현재까지 발표된 각종 관련 기술들을 정리하여 각각의 장단점을 비교해 본다. 이를 위해 콘텐츠 저작 시점, 저작 장소 및 콘텐츠 가공 정도에 따라서 기술을 분류해 보고 그 중에서도 콘텐츠 자동 재작성 기술에 초점을 두어 자세히 살펴본다. 또한 본 논문에서는 기존의 콘텐츠 자동 재작성 기술을 보완하여 선택정보 압축 기능을 추가한 새로운 콘텐츠 재작성 방법을 제안한다.

### Abstract

Recently, construction of infrastructure for mobile internet has been progressed very rapidly including various mobile terminals. Therefore, many service models on mobile internet have been appeared and the methods to reuse the existing huge wired web contents have been studied with many-sided ways. In this paper we classify the related technologies published up to now and make a comparison among their strengths and weaknesses. For this purpose, technologies are classified according to authoring time, authoring place, and processing level of contents. The automatic contents re-authoring technology is examined in particular and we propose a new contents re-authoring method in which the function of compressing selected information is added in order to complement the existing contents re-authoring methods.

## 1. 서 론

무선 통신의 급격한 성장세와 함께 모바일 웹 서비스는 점차 필수적인 것으로 여겨지고 있다. 그러나 사용자의 요구수준은 높는데 반해 아직까지 원활한 모바일 웹 서비스가 제공되지 못하는 것도 사실이다. 그 이유 중 하나는 기존에 대량으로 구축되어 있는 웹 콘텐츠가 대부분 데스크탑 단말을 위한 유선 웹 서비스용인데 이에 대응하는 모바일 웹 서비스용 콘텐츠가 다양하게 준비되지 못했기 때문이다.

웹 콘텐츠를 단말의 종류에 따라 재사용 또는 재작성하는 방법은 여러 가지 각도에서 시도되어 왔다. 가장 간단하게는 별도의 변환 시스템 없이 단말의 종류별로 웹 콘텐츠를 각기 따로 만드는 것이고 가장 자동화된 방법으로는 데스크탑용 웹 콘텐츠를 무선 단말용으로 자동 변환하는 것이다. 그 외에 하나의 내용 콘텐츠에 여러 종류의 스타일 시트를 적용하여 단말에 적합한 디스플레이를 지원하는 방법과 화면 확대/축소 기법으로 단말의 화면 내에서 네비게이션을 지원하는 방법도 있다.

본 논문에서는 이와 같은 다양한 무선 단말 지원 기법을 각각 살펴보고 서로의 장단점을 비교해보면서 특히 콘텐츠 자동 재작성 기술에 대해 더욱 자세히 조사해 보고자 한다. 또한 기존의 여러 가지 콘텐츠 재작성 기법에서 부족한 점을 찾아내고 이를 보완하기 위한 선택정보 압축형 재

\* 정 회 원 : 한국전자통신연구원 정보가전연구부 선임연구원 scho@etri.re.kr

\*\* 비 회 원 : 한국전자통신연구원 정보가전연구부 연구원 hermes@etri.re.kr

\*\*\* 비 회 원 : 한국전자통신연구원 정보가전연구부 연구원 hsshin8@etri.re.kr

\*\*\*\* 비 회 원 : 충남대학교 컴퓨터과학과 교수 cjhwang@ipl.cnu.ac.kr

작성 기법을 제안한다. 이 기법은 오늘날의 HTML 문서가 거의 대부분 중첩된 테이블을 사용하여 레이아웃을 구성한다는 점에 착안하여 설계되었으며 부가 정보를 리스트 박스 형태로 재배치하고 Main Body에 해당하는 주요 정보는 한눈에 파악할 수 있도록 하여 가독성을 높이면서도 정보의 손실은 방지하였다.

이후 본 논문은 다음과 같이 전개된다. 1절 서론에 이어 2절에서는 단말에 독립적인 웹 서비스 지원 기술 전반에 대해 살펴보고 3절에서는 콘텐츠 자동 제작성 기술을 좀 더 자세히 소개한다. 4절에서는 본 논문에서 제안하는 선택정보 압축형 콘텐츠 제작성 기법을 소개하고 5절에서는 결론과 함께 앞으로의 연구방향을 정리한다.

## 2. 단말 독립적인 웹 지원 기술

### 2.1 모바일 웹 브라우징 환경

오늘날, 웹 정보를 획득하기 위한 데스크탑 컴퓨터에서의 브라우징 모델은 풍부한 사용자 인터페이스(디스플레이, 키보드, 포인팅 디바이스 등)와 컴퓨팅 자원(CPU, 메모리, OS 등) 및 높은 대역폭의 네트워크를 사용함으로써 빠른 성장을 거듭해왔다. 그 중에서도 특히 넓은 디스플레이는 웹 상의 풍부한 콘텐츠를 브라우징하는데 기여한 바가 크다. 반면에 날로 수요가 늘어나는 소화면의 인터넷 단말은 이와 같은 풍부한 웹 콘텐츠를 브라우징하기에는 많은 제약이 따른다.

모바일 인터넷 단말을 두 종류로 구분하면 PDA와 셀룰라폰으로 나눌 수 있다. 이들 두 종류의 디바이스는 무선 네트워크를 위해 CDMA/TDMA/GSM 등을 사용하는 공통점을 가지고 있으나 화면의 크기 및 해상도에는 현저한 차이가 있다. PDA는 보통 320×240 이상, 높게는 640×480 해상도를 지원하지만 셀룰라폰은 96×32 또는 96×72 정도를 지원한다.

이 두 부류의 모바일 인터넷 단말에서 채택하는

웹지원 기술 또한 디스플레이, CPU 등의 차이에 따라 다른 방향으로 발전해 왔다. PDA에서는 화면 크기와 해상도가 높아짐에 따라 기존의 HTML 콘텐츠를 그대로 지원하는 방향으로 발전해 가고 있고 셀룰라폰에서는 소규모 단말용으로 개발된 HDML, WML, cHTML 등으로 저작된 별도의 콘텐츠를 사용하는 경향이 지배적이다. 하지만 많은 PDA 디스플레이가 아직도 320×240 이하의 해상도를 사용하고 있고 PDA와 폰의 경계가 불분명해지면서 셀룰라폰에서도 디스플레이를 점차 확대하여 HTML의 일부를 지원하고 있는 움직임도 뚜렷하다. 본 논문에서는 이와 같은 모바일 웹 브라우징 사용 환경에서 특히 HTML 콘텐츠의 변환 및 제작성(re-authoring)을 통한 소규모 단말의 지원 기술에 주목하고 이에 관련된 현재의 기술 현황을 정리하고자 한다. 또한 기존의 방법에 추가하여 선택정보 압축형 콘텐츠 변환 방법을 소개한다.

### 2.2 모바일 단말용 웹 콘텐츠 지원 기술

다양한 단말을 지원하기 위한 웹 콘텐츠 서비스에는 여러 가지 지원 기술이 개발되어 있지만 어느 한가지가 다른 방법에 비해 월등히 우위에 있지는 않고 서로 보완하여 전체 시스템을 이루는 경우가 많다. 이들 지원 기술을 몇 가지 기준으로 구분하여 살펴보면 전체 기술을 이해하는데 도움을 줄 수 있다. 그 기준들은 1) 콘텐츠 저작 시점, 2) 콘텐츠 저작 장소 (서버, 프락시 서버, 또는 클라이언트), 3) 콘텐츠 가공 정도 (가공 없음, 요약, 취소선택, 재구성) 등이 될 수 있다. 각 기준별로 현재의 기술들을 구분하여 설명하면 다음과 같다.

#### (1) 콘텐츠 저작 시점에 따른 분류

다양한 모바일 단말을 지원하기 위한 웹 콘텐츠의 저작 시점을 기준으로 기술을 분류하는 것은 수동 또는 자동의 기준을 적용하는 것과 유사하다. 즉 단말의 특성에 맞게 각각 콘텐츠를 미리

작성하여 제공하는 방법 (device-specific authoring) 과 단말의 특성에 맞게 콘텐츠를 자동 재작성하는 방법 (automatic re-authoring)으로 구분할 수 있다. 단말에 적합한 별도의 웹 콘텐츠를 미리 만들어 두고 서비스하는 방법과 하나의 HTML 콘텐츠로부터 각 단말용 콘텐츠를 실시간으로 재작성하는 방법으로 대별되는 것이다.

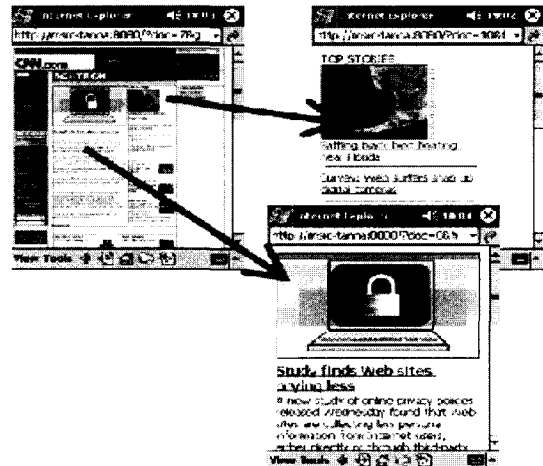
첫 번째 방법은 사용자 만족도가 가장 높은 웹 페이지를 보여줄 수 있지만 단말에 따라 별도의 콘텐츠를 저작해야 하는 수고가 필요하다. 대부분의 웹 콘텐츠가 데스크탑 PC를 대상으로 하여 만들어져 있으므로 모바일 단말용 콘텐츠는 따로 저작해야 하는 어려움이 있고 따라서 모바일 웹 액세스는 상당히 제한된 사이트에서만 지원될 것이다.

두 번째 방법에서는 자동으로 HTML 콘텐츠를 변환 및 재작성하기 위한 다양한 연구들이 진행되어 왔고[1,2,4,9] 상업적인 제품으로 나와 있는 것도 있다[6,8]. 이 방법은 여러 가지 시도에도 불구하고 여전히 자동 재작성된 HTML 또는 비HTML (HDML, WML, cHTML 등) 콘텐츠의 질이 떨어지는 어려움이 있다. 대략적인 과정은 HTML 문서를 파싱하여 구문 또는 의미 정보를 분석 및 추출하고 이로부터 단말의 환경에 맞게 정보를 변환하고 콘텐츠를 재구성하는 것이다.

(2) 콘텐츠 저작 장소에 따른 분류

모바일 단말용 웹 콘텐츠 지원 기술은 콘텐츠 저작 장소에 따라 구분할 수도 있다. 서버측 기술로는 (1)의 분류에서 첫 번째 방법인 device-specific authoring이 이에 해당한다. automatic re-authoring 방법은 프락시 서버를 이용할 수도 있고 클라이언트 단말에서 변환 및 재작성이 이루어질 수도 있다. 대부분의 경우 단말의 컴퓨팅 파워 및 네트워크 성능의 한계로 프락시 서버를 이용하고 있다.

automatic re-authoring 방법이 프락시 서버를 기반으로 한다면 순수히 클라이언트 단말을 기반으로 제공되는 대표적인 기술이 Pad++[3]와 같은 client-side navigation이다. 이 방법은 웹 페이지의 어떤



(그림 1) SmartView HTML콘텐츠 변환방법

부분이라도 마치 돋보기를 들이대듯이 줌 인, 줌 아웃하여 유연하게 볼 수 있게 한다. 이 기술은 축소된 초기 화면에서 어느 부분이 원하는 내용을 포함하고 있는지 확인하기가 어려워 여러 번의 시행착오적인 줌 인, 줌 아웃을 해야 하는 경우가 많다.

Pad++ 시스템은 웹 페이지의 논리적 분석 및 분류 없이 화면의 불특정 파트를 확대/축소하는 방법이지만 SmartView[7]에서는 HTML 페이지의 배치를 분석하고 논리적 섹션으로 파티션한다. 이 분류된 섹션들은 사용자의 선택에 따라 독립적으로 확대되어 보여진다. 이 방법은 내용을 논리적으로 분류된 섹션으로 나누어 볼 수 있다는 점에서 Pad++보다 향상된 기술이라고 할 수 있다. 아래 그림 1은 SmartView의 HTML 콘텐츠 액세스 방법을 보여준다.

(3) 콘텐츠 가공 정도에 따른 분류

모바일 단말용 웹 콘텐츠 지원 기술은 콘텐츠의 가공 정도에 따라 구분될 수 있다. 데스크탑용 웹 콘텐츠가 가공 없이 그대로 제공되는 것은 Pad++ 시스템이 대표적이다. 화면의 확대/축소를 위해 화면 전체를 이미지로 가공하기는 하지만 사용자에게 보여지는 부분이 달라지는 것은 아니다.

대량의 웹 페이지 정보를 적당한 형태로 요약

하는 방법은 Power Browser[4] 등의 연구가 있다. 소규모 디스플레이의 한계로 전체 페이지를 모두 보여줄 수 없으므로 하나 또는 그 이상의 페이지들로 나누어 적당히 요약하여 보여주는 것이다. 이 방법은 웹 페이지의 원래 저작 의도대로 내용이 표현되지 않을 수 있는 위험이 크다. 모바일 단말을 위한 automatic re-authoring 방법은 대부분 웹 페이지의 내용을 축소 변환하므로 이 방식을 사용한다고 볼 수 있다.

웹 페이지 정보를 적당한 형태로 요약하는 방법이 서비스 제공 시스템에 의해 미리 지정된 방식이라면 사용자의 선택에 의한 방식은 필터링[2]이다. 이 기술은 사용자의 관심에 따라 웹 페이지에서 원하는 내용을 키워드나 정규 표현식 또는 추출 명령어 등을 사용하여 필터링하는 것이다. 이 기술은 프락시 서버를 통해 구현될 수도 있고 클라이언트의 디스플레이 관리를 통해 구현될 수도 있다. 하나의 소스로 여러 개의 타겟 디스플레이를 제공하는 automatic re-authoring 방법의 장점을 가지지만 자동으로 필터링되는 내용이 얼마나 원래 웹 페이지의 내용을 왜곡하지 않고 표현하는가의 정도가 기술의 관건이 된다.

이상과 같이 모바일 단말용 웹 콘텐츠를 지원하기 위한 다양한 기술들은 각각의 장단점을 가지고 있으므로 목표로 하는 서비스 및 시스템에 적합한 기술을 서로 융합하고 필요에 따라 취사선택해야 한다.

### 3. 콘텐츠 자동 제작성 기술

2절에서 살펴본 각종 모바일 웹 콘텐츠 지원 기술 중 본 논문에서는 콘텐츠 자동 제작성 기술에 초점을 두고 더욱 자세히 살펴보고자 한다. 콘텐츠 가공 정도에 따른 분류로는 요약 기술에 해당되고 콘텐츠 저작 장소에 따른 분류로는 프락시 서버를 이용하는 방법에 해당된다. 이 기술의 대표적인 연구는 Digester 시스템[1]에서 이루어졌는데 3,188개의 웹 페이지를 대상으로 한 분석을 통

하여 각종 디자인 휴리스틱스를 얻어내고 제작성 기법들을 개발하였다. Digester 시스템은 크게 3가지의 콘텐츠 제작성 기법을 사용한다. 각각은 다음과 같다.

- **Outlining** : 웹 페이지가 여러 개의 섹션으로 구성되어 있을 때 각 섹션의 제목만을 따로 모아서 리스트를 만들고 이것을 초기 화면으로 한다. 각 섹션별 구체적인 내용은 리스트의 해당 제목에 링크로 연결된 별도의 웹 페이지로 만들어 사용자의 선택에 따라 액세스된다. 이 방법은 사용자 매뉴얼 등 소책자 형태의 웹 페이지에 적합하고 섹션의 깊이에 따라 어떻게 리스트를 구성할 것인지는 옵션으로 처리된다.
- **First Sentence Elision** : 내용이 긴 웹 페이지의 경우 섹션의 제목 대신 첫번째 문장이 그 섹션의 내용을 잘 나타내는 수가 있다. 또 섹션의 제목이 없는 경우에도 첫 번째 문장으로 대신하는 것이 유리하다. 이 때 적용하는 방법은 **Outlining**과 마찬가지로 첫 번째 문장을 뽑아서 초기 화면을 만들되 섹션의 구체적인 내용은 별도의 페이지로 링크되게 한다.
- **Image Reduction and Elision** : 데스크탑용의 화려한 웹 페이지는 대부분 많은 수의 이미지를 포함하고 있고 이는 모바일 단말의 소규모 화면에서는 제대로 표현될 수가 없다. 이미지 변환시 고려할 사항은 크기를 줄이더라도 이미지 정보를 유지할 것인가 아니면 전체 이미지를 제거할 것인가 하는 것이다. Digester는 크기를 원하는 만큼 (25%, 50%, 75%) 줄여주거나 첫 번째 이미지를 제외한 나머지를 모두 제거하는 방법 등을 제공한다. 이 때 줄어들거나 제거된 이미지의 원본은 링크로 연결된 별도의 페이지로 제공된다.

Digester 시스템은 또 다른 논문[2]에서 더욱 개선된 제작성 기법들을 소개하는데 여기에는 indexed segment transform, table transform 등이 포함되어

있다. 전자는 논리적으로 구분이 가능한 웹 페이지 내의 요소들을 찾아서 이들을 기준으로 세그먼트로 나누는 것이다. 이 때 타겟 디스플레이의 크기에 맞게 세그먼트를 조절하는 기술이 필요하다. 또 후자는 상-하, 좌-우의 순서로 테이블 내의 각 셀들을 풀어서 리스트로 변환하는 방식이다.

WebAlchemist[9]는 위와 같은 Digester의 기본 변환 방법을 보완하고 또 selective elision transform 등 몇가지 방법을 추가하여 총 5개의 변환 방법을 정의한 후 이들을 적용하는 최적의 순서를 개발하였다. 즉 자동 생성되는 타겟 웹 페이지의 복잡도를 기반으로 양호도를 정하고 양호도가 높게 나올 수 있도록 적용되는 순서를 정하는 것이다. 이 연구는 개발된 방법을 CNN, Yahoo 등 유명한 웹 사이트에 적용하여 그 결과를 34명의 사용자들로부터 얻은 반응을 기초로하여 정성적으로 평가하기도 하였다.

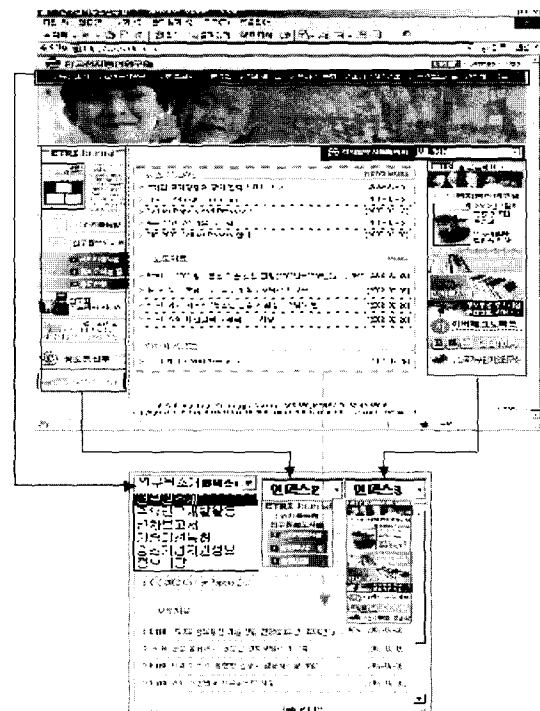
#### 4. 선택정보 압축형 자동 재작성

##### 4.1 개발 동기 및 개요

3절에서 콘텐츠 자동 재작성을 위한 여러 가지 기법들이 소개되었는데 이들 각각을 더욱 향상시키고 추가적인 기법들을 개발하여 콘텐츠 재작성을 위해 서로 융합하여 적용함으로써 더욱 높은 효과를 얻을 수 있다. 본 논문에서는 이를 위해 대부분 테이블로 구성되어 있는 현재의 HTML 페이지에 적용하기 위한 새로운 콘텐츠 재작성 기법을 제안한다. 현재 HTML4.0으로 작성된 대부분의 웹 페이지들은 레이아웃을 위해 여러 개의 중첩된 테이블들을 사용하고 있다. CSS 등 스타일 정보를 위한 별도의 스크립트가 마련되어 있지만 테이블을 사용하여 디스플레이 형태를 만드는 것이 더욱 보편화되어 있다. 즉 웹 페이지의 내용에 해당하는 텍스트, 이미지 등을 모두 중첩된 테이블 속의 한 셀에 포함시킴으로써 원하는 레이아웃을 만들어 내는 것이다. 본 연구는 이와 같이 테이블의 셀들로 구성된 HTML 페이지를 정

보의 손실 없이 소규모 디스플레이 단말에 적당하게 재작성하는 것을 목표로 한다.

3절에서 소개한 HTML 콘텐츠 자동 재작성 기술들은 대량의 텍스트 또는 이미지 정보를 포함하는 HTML 페이지들을 대상으로 하지만 복잡한 테이블 구조에 대한 고려는 무척 미흡하다. 특히 많은 웹사이트의 초기 화면은 텍스트 정보는 많지 않으면서 복잡한 테이블의 중첩 구조를 포함하고 있다. 이 테이블 중 해당 웹 페이지의 주요 내용에는 관련이 없지만 사용자의 부가적인 선택을 위한 부분이 있다. 아래 그림2의 상단 및 좌우측의 3개의 영역이 이 부분이다. 본 논문에서는 이와 같은 부가적인 선택정보를 위한 테이블의 일부분을 인덱스로 추출하여 최종적으로는 리스트박스 형태로 재작성하는 기법을 제안한다[5]. 이 기법을 제2절의 분류에 따라 구분해보면 콘텐츠 저작 시점에 따른 분류로는 콘텐츠를 미리 저작하여 제공



(그림 2) 선택정보 압축형 기법으로 재작성된 HTML 콘텐츠의 화면 예

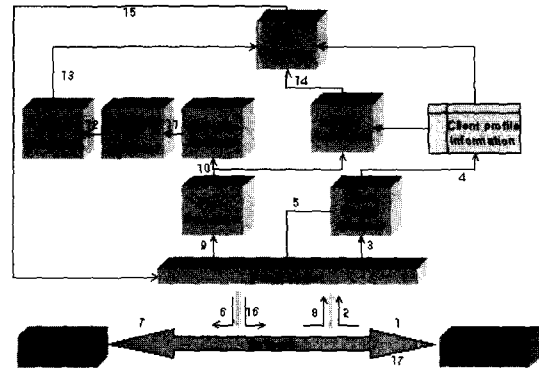
하는 방법이 아닌 필요에 따라 자동 제작성하는 기법이고, 콘텐츠 저작 장소에 따른 분류로는 프락시 서버를 이용하는 방법이며, 콘텐츠 가공 정도에 따른 분류로는 요약이나 필터링이 아닌 전체 정보의 보존 및 재배포 기술에 해당된다.

#### 4.2 요소 기술 및 처리 단계

그림 2는 ETRI 홈페이지를 선택정보 압축형 기법으로 제작하였을 때의 결과 화면을 보여 준다. 원래의 웹 페이지는 복잡하게 중첩된 테이블 구조로 이루어져 있다. 따라서 이 전체 페이지를 제한된 크기의 모바일 단말에 그대로 디스플레이하기 위해서는 단말 화면 크기에 맞게 축소하여 제공한 후 사용자가 선택한 부분을 확대하는 client-side navigation 방법을 적용할 수 있다. 또 간단하게 원래의 크기 그대로 디스플레이 한 후 상하, 좌우 스크롤바를 이용하게 할 수도 있지만 이 방법은 불편한 인터페이스이므로 가능한 피해야 한다.

본 논문에서 제안하는 선택정보 압축형 제작성 기법은 대부분의 웹 페이지에서 포함하고 있는 부가적인 선택을 위한 테이블의 일부 영역을 인덱스 리스트 박스로 재구성하는 것이다. 이 기법을 통해 웹 페이지의 본문에 해당하는 주요 내용만 디스플레이하고 나머지 부분은 거의 모두 리스트 박스로 간단하게 줄여줄 수 있다. 이 기법을 앞장에서 소개한 다양한 기법들에 추가하여 적용한다면 더욱 세련되고 간략하게 제작성된 HTML 콘텐츠를 얻을 수 있을 것이다. 이 기법은 프락시 서버에서 구현되는 것이 가장 바람직하다.

그림 3은 선택정보 압축형 제작성 기법을 구현하기 위한 서버 시스템들과 그 처리 순서를 나타낸 것이다. 프락시 서버에 구현되는 콘텐츠 제작성 시스템은 네트워크를 통해 웹 서버 및 웹 브라우저에 연결되고 주요 처리 과정은 전처리 단계, 구조분석 단계, 요소분류 단계, 인덱스 생성 단계, HTML생성 단계 등으로 구성된다.



(그림 3) 선택정보 압축형 제작성을 위한 시스템 구성

##### 4.2.1 전처리 단계

콘텐츠 변환에 적절한 형태로 HTML을 정제하는 단계이다. HTML을 파싱하여 well-formed HTML로 만들고 Document Object Tree를 생성한다. 일반적인 웹 문서는 태그 규칙이 잘 지켜지지 않은 것이 대부분이므로 태그 분석 후 규칙에 어긋난 부분을 수정하는 과정이 필요하다. 이 과정을 거쳐 well-formed HTML이 생성되면 구문 분석하여 Document Object Model로 변환한다. DOM (Core) Level 2의 기본 API를 포함하고 HTML의 표현에 편리한 API가 추가된 DOM (HTML) Level 2의 모델을 결과로 생성한다. 이로써 다음 단계인 구조 분석 알고리즘에 적용할 수 있는 데이터 구조가 만들어지게 된다.

##### 4.2.2 구조 분석 단계

전처리 단계에서 생성된 Document Object Tree를 입력으로 하여 분석 알고리즘을 적용한다. 분석 알고리즘을 통하여 웹 문서를 semantic 특성에 따라 구성요소 단위로 분리하고 다음 단계에서 유사한 요소들끼리 Grouping할 수 있도록 함으로써, 웹 문서의 구조적인 분석을 수행한다. 본 연구에서는 Table-Layout Based Structure Analysis Algorithm을 사용하여 데이터를 분석한다. 이 알고리즘은 웹 문서가 콘텐츠 내용의 유사성을 가지는 그룹 단위로 분리가 가능하다는 점을 이용

한 것이다. 이 알고리즘이 적용되어 구성 요소들이 적절히 분석되면 그림 2에서와 같이 상단, 좌측, 우측의 3개의 요소 그룹이 인덱스로 추출되어 리스트박스 형태로 재작성될 수 있다.

Table-Layout 기반의 분석이 가능한 것은 HTML을 작성하는 개발자가 명확한 내용 전달을 위해서 의미상의 차이를 가지는 내용에 대해서 레이아웃 및 구조적 태그를 사용하여 시각적인 분리가 이루어지도록 디자인하기 때문이다. 이 단계에서는 TABLE 태그를 이용하여 2차원적인 구조가 분석되고 다음 단계에서 각각의 TABLE 조각들의 위치 정보 등을 활용하여 유사한 조각들의 Grouping/Categorizing이 이루어질 수 있도록 한다. 분석 결과로는 element node class를 leaf node로 가지는 Component Block Object Tree가 생성된다.

#### 4.2.3 요소 분류 단계

구조 분석 단계를 통하여 생성된 Component Block Object를 하나씩 분리해내어 포함하는 콘텐츠의 특성에 따라서 Image Index/Text Index/Main Body로 분류하고 새로 생성될 HTML에서 표현될 위치를 결정한다. 그림2에서 상단에 위치한 Component Block은 Text Index로, 좌우측에 위치한 Component Block은 Image Index로, 가운데 것은 Main Body로 분류되었음을 알 수 있다.

Component block의 content info 값으로부터 콘텐츠의 특성이 일정한 반복패턴을 가지는 index 형태인지, 일반적인 text/image가 섞여서 나타나는지를 검사함으로써 Selected Index 또는 Main Body로 분류할 수 있다. Selected Index로 분류되면 Image Index 또는 Text Index로의 재분류 과정을 거친다. Main Body로 분류되면 전체 width 확인 후 재배치된다.

#### 4.2.4 인덱스 생성 단계

Text 및 image Index의 생성은 Index로 분류된 Component Block을 table object tree 형태로 입력 받고 여기서 인덱스 정보를 Tag & Contents Pattern

Algorithm을 통하여 추출해낸다. 이를 위해 HTML 문서들을 작성할 때 나타나는 몇 가지의 일반적인 규칙들을 이용하여 콘텐츠의 경계가 되는 태그를 추출하여 추출된 태그를 기준으로 콘텐츠를 분리하고 분리된 콘텐츠들 중에서 적절한 인덱스 콘텐츠를 골라낸다.

인덱스 정보가 이미지로 구성되어 있는 경우에는 자바 스크립트와 폼 태그를 이용하여 인덱스를 생성하게 된다. 이때 image의 개수, image files, index id 등의 정보를 가지는 Index Info class 객체를 이용한다.

#### 4.2.5 HTML 재작성 단계

HTML 4.0 표준을 따르는 문서를 생성한다. 앞 단계에서 분석, 변환된 문서를 재정렬 순서에 따라서 배치하고 인덱스 및 자바 스크립트 부분을 추가, 변환된 이미지를 제공하도록 HTML을 생성한다. 재작성되는 HTML의 기본 양식은 다음과 같다.

```

<HTML>
<HEAD>
<TITLE></TITLE>
<SCRIPT>
  → 자동 생성된 자바 스크립트 파일을 첨부한다. Image
  Index가 생성되는 경우 추가된다.
</SCRIPT>
</HEAD>
<BODY>
  → Selected Index 또는 Main Body로 분류된 Component
  Block을 BODY 태그 안에 붙인다.
<SELECT>
<OPTION>
  → Text Index의 수만큼 select list form을 생성하고 각
  각의 value 값을 Option 태그로 적절히 배치한다.
  <OPTION>
</SELECT>
<TABLE>
<TR>
<TD>
  → Body로 분류된 Component 각각을 하나의 TABLE
  TD의 value로 포함하여 정렬한다. 이때 새로 생성되
  는 전체 테이블의 width는 client profile에 나타난
  디스플레이 성능정보에 따라서 정한다.
</TD>
</TR>
</TABLE>
</BODY>
</HTML>
    
```

Image Index를 생성해야 할 경우, 자동 생성된 자바 스크립트 파일을 첨부하고, Text Index는 그 개수만큼 Select Form을 생성한다. 이때 Java Script Generator와 마찬가지로 INDEX info class 정보를 이용하여 Tag를 자동 생성한다. Image/Text Index의 위치는 전체 Layout에서 상단 첫째 라인에 위치하고, 그 아래로는 Main Body로 분류된 Component Block들을 table 내의 주어진 위치에 붙여 나간다.

## 5. 결론 및 향후 연구

본 논문은 크게 두 가지 목적으로 작성되었다. 첫 번째는 모바일 웹 콘텐츠 지원 기술 전반에 대한 고찰과 현재의 기술 수준을 알아보고자 하는 것이고 두 번째는 이 기술들 중에서 콘텐츠 자동 제작성에 대한 또 하나의 시도를 소개하기 위한 것이다. 첫 번째 목적을 위해 2, 3절에 걸쳐 기술 전반을 소개하였다. 2절에서는 1) 콘텐츠 저작 시점, 2) 콘텐츠 저작 장소, 3) 콘텐츠 가공 정도에 따라 모바일 단말용 웹 콘텐츠 지원 기술 전반을 분류하고 장단점을 비교해보았다. 3절에서는 웹 콘텐츠 자동 제작성 기술에 대하여 더 자세히 살펴보았다. 4절에서 두 번째 목적에 해당하는 선택정보 압축형 자동 제작성 기법을 소개하였다. 이 기법은 웹 페이지의 같은 위치에 반복적으로 나타나는 부가정보에 해당하는 테이블 구성 요소를 인덱스로 분류하고 이를 HTML의 리스트 폼 태그로 변환하여 그림 2의 아래쪽 화면과 같은 리스트 박스를 생성하는 것이다.

3절에서 소개한 기존 연구들이 소규모 화면을 가진 단말용으로 웹 콘텐츠를 자동 제작성할 때, 정보를 요약 또는 추출함으로써 정보의 보존이라는 측면에서 취약함을 보이는 반면, 본 논문에서 제안하는 방법은 부가 정보를 리스트 박스 형태로 재배치하고 Main Body에 해당하는 주요 정보는 한눈에 파악할 수 있도록 하여 가독성을 높이면서도 정보의 손실은 방지하였다. 이 방법은 오늘날의 HTML 문서가 거의 대부분 중첩된 테이블을

사용하여 레이아웃을 구성한다는 점에 착안하여 설계되었고 따라서 HTML문서를 분석하고 그 구성 요소를 분류하는 과정에서 TABLE 태그를 집중적으로 분석 및 분류의 대상으로 하고 있다.

앞으로의 과제는 현재 프로토타입으로 구현되어 있는 선택정보 압축형 자동 제작성 기법을 더욱 구체적으로 설계하고 구현하여 성능을 향상시킴과 동시에 더욱 안정적으로 적용될 수 있도록 해야한다. 더불어 제안된 기법이 기존의 콘텐츠 자동 제작성 기법들과 비교 실험되는 것이 필요하다. 프로토타입을 실행한 결과인 그림 2를 보면 부가 정보는 간략하게, 그리고 주요 정보는 한눈에 표현된 본 기법의 우수성이 쉽게 확인될 수 있지만 얼마나 많은 웹 페이지를 대상으로 할 수 있는지는 다른 기법들과 비교 실험되어야 하기 때문이다. 본 논문에서는 지금까지의 웹 콘텐츠 제작성 기술들을 소개하면서 이들 각각을 더욱 향상시키고 추가적인 기법들을 개발하여 콘텐츠 제작성을 위해 서로 융합하여 적용함으로써 더욱 높은 효과를 얻을 수 있음을 주장하였다. 이는 기존의 다양한 기술들과 본 연구에서 제안한 선택정보 압축형 자동 제작성 기법 중 어느 하나가 절대적으로 우수하다고 결론 내리기 힘들다는 뜻이다. 즉 웹 콘텐츠의 구성 및 종류에 따라 각각의 방법은 모두 장단점을 가지고 있으므로 어떤 경우에, 어떤 종류의 웹 페이지에 적용될 때 더 우수한 성능을 보이는지 각각의 방법들을 비교 분석해 보는 것이 의미가 있을 것이다.

## 참고 문헌

- [1] Bickmore T. and Schilit W., Digester: device-independent access to the world wide web, In Computer Networks and ISDN Systems, vol. 29, no. 8, pp. 1075~1082, 1997.
- [2] Bickmore T. et al., Web Page Filtering and Re-Authoring for Mobile Users, In The Computer Journal, vol. 42, no. 6, pp. 534~546, 1999.



- [3] Bederson, B. and Hollan, J., Pad++: A zooming graphical interface for exploring alternate interface physics, In Proc. ACM User Interface Software and Technology 94, pp. 17~26. 1994.
- [4] Buyukkokten O. et al., Power Browser: Efficient Web Browsing for PDAs, In Proc. The ACM Conference on Computers and Human Interaction 2000 (CHI00), 2000.
- [5] ETRI, 지능정보단말용 웹 콘텐츠 변환기 상세 설계서, 2002.
- [6] IBM, WebSphere, <http://www.software.ibm.com/webservers/>.
- [7] Milic-Frayling N. and Sommerer R., SmartView: Flexible Viewing of Web Page Contents, In Proc. World Wide Web Conference 2002 (CD-ROM), 2002.
- [8] Nagao K., Semantic Transcoding: Making the World Wide Web More Understandable and Usable with External Annotations, Unpublished manuscript, <http://citeseer.nj.nec.com/nagao99semantic.html>, 1999.
- [8] Spyglass, Prism2.0, <http://www.spyglass.com>.
- [9] Whang Y., Jung C., Kim J., and Chung S., Web Alchemist: A Web Transcoding System for Mobile Web Access in Handheld Devices, In Proc. ITCom 2001, 2001.

◎ 저자 소개 ◎



**조 수 선**

1987년 서울대학교 계산통계학과 졸업(학사)  
1989년 서울대학교 대학원 계산통계학과 졸업(석사)  
1989년~1994년 (주)웅진미디어 CBE개발부 연구원  
1994년~현재 : 한국전자통신연구원 정보가전연구부 선임연구원  
관심분야 : 모바일 웹지원 기술, 웹 콘텐츠 분석 및 이해  
E-mail : scho@etri.re.kr



**이 동 우**

1995년 경북대학교 전자공학과 졸업(학사)  
1997년 경북대학교 대학원 전자공학과 졸업(석사)  
1997년~2001년 (주)현대전자 전장사업본부 대리  
2001년~현재 : 한국전자통신연구원 정보가전연구부 연구원  
관심분야 : 지능정보 단말, 멀티모달 브라우저  
E-mail : hermes@etri.re.kr



**신 희 숙**

1998년 경북대학교 컴퓨터공학과 졸업(학사)  
2000년 포항공대 대학원 컴퓨터공학과 졸업(석사)  
2000년~2001년 (주)데이콤 종합연구소 연구원  
2001년~현재 : 한국전자통신연구원 정보가전연구부 연구원  
관심분야 : 웹 문서 변환 기술, 마크업 언어, 무선 인터넷 응용 기술 등  
E-mail : hsshin8@etri.re.kr



**황 치 정**

1975년 서강대학교 수학과 졸업(학사)  
1979년 서강대학교 대학원 수학과 졸업(석사)  
1985년 Univ. of Connecticut 전산학과 졸업(석사)  
1987년 Univ. of Connecticut 전산학과 졸업(박사)  
1987년~1988년 Univ. of Connecticut 객원교수  
1988년~현재 : 충남대학교 컴퓨터과학과 교수  
관심분야 : 이미지 프로세싱, 패턴 인식 등  
E-mail : cjhwang@ipl.cnu.ac.kr