HTML 문서 내에서의 정보 추출에 관한 연구

강현관, °윤종선, 강경진
(hkkang, jsyoon, kjkang)@lgic.co.kr LG 정보통신㈜

Research on Information Extract Process for HTML Document

Hyun-Kwan Kang, Jong-Sun Yoon, Kyoung-Jin Kang
LG Information & Communications, Ltd.

요 약

본 논문에서는 HTML(Hypertext Markup Language)을 사용하여 표현된 인터넷상의 유용한 정보들을 휴대폰에서 사용하기 위한 방법을 생각해 본다. 뉴스 정보나 증권, 교통, 날씨와 같은 정보는 매일 또는 매 시각 정보의 내용은 달라지지만 보통 웹브라우저를 통하여 보여지는 형태는 유사하거나 같은 것을 알 수 있다. 이러한 정해진 형식 내의 정보를 얻어내어 일반 웹브라우저가 아닌 휴대폰이라 불리는 이동 무선 단말기와 같은 곳에 표시 할 수 있는 형태로의 변환을 고려 해보면 대형 화면을 고려하여 만들어진 모든 내용이 변환되어 작은 화면의 저속 통신을 지원하는 일반 단말기로 전달되어져야 한다. 현재는 WAP(Wireless Application Protocol)을 통한 이동 무선 단말기와 Internet 을 연결하는 방안 들이 많이 연구되어지고 있는데 아직도 HTML 로 만들어진 정보를 휴대폰에 나타내기에는 방법이나 결과로 나타나는 품질면에서 부족함을 느낀다. 따라서 기존의 일반적인 방법들과 더불어 고품질의 정보를 제공하는 방법들을 생각하게 되었고, 본 논문에서 이러한 문제들을 해결하는 방법으로 정보 추출과 사용이라는 방법을 통한 변환 방법에 관하여 제안한다.

1. 서론

인터넷이 생활의 중심으로 다가오면서 새로운 기술과 서비스를 통하여 우리는 때와 장소에 상관 없이 인터넷을 접 할 수 있는 순간을 맞이 하게 되었다. 그 기술의 하나가 WAP(Wireless Application Protocol)으로 우리가 사용하는 이동 단말기에 인터넷의 정보를 연결 시켜주는 기술이다. 현재는 WAP 을 사용하기 위해서는 일반 HTML 형식의 문서가 아닌 WML(Wireless Markup Language) 형식의 문서를 사용하여야 하기 때문에 기존의 HTML 형식의 정보들을 WML 형식으로 변환하는 작업이 필요하게 된 것이다. 이러한 작업에 적용하기 위한 첫걸음으로 고정된 형식의 HTML 문서 형식에서 WML 과 같이 이동 단말기로 전송되어 표시 될 수 있는 언어로의 변환이 필요하다.

일반 브라우저에 나타나는 뉴스 사이트의 한 화면은 이동 무선 단말기의 작은 화면에 표시한다면 수십 장의 화면으로 나타내 질 수도 있으며 결과적으로 HTML 문서 하나의 화면 표시 및 수신을 위하여 많은 시간 및 작업이 필요해지는 것이다. 따라서 이러한 상황을 막기 위하여 하나의 화면은 작은 화면으로 나누어야 하며 사용자가 원해서 선택하기 전에는 사용자에게 불필요하게 전달되는 정보가 적거나 없어야 한다.

기존의 HTML 로 작성되어진 문서의 정보를 추출하여 재구성하고자 할 때 필요한 정보나 형태를 기술할 수 있는 방법이 필요해 졌다. 기존의 여러 방법들을 고려해 보면 CGI 에서 사용되는 Perl 과 같은 언어로 작성하여 변환을 시도하거나 C 나 기타 다른 고급 언어를 사용하여 표현 할 수 있지만 고급 언어가 가지는 다양한 기능과 제어성은 강하지만 특정한 일을 하는 경우에는 번거롭고 표현이 어려워지는 문제가 발생한다. 또한 다루고자 하는 언어가 기존의 C 나 C++ 과 같이 변화가 적거나 없는 것이 아니고 계속적으로 버전 업을 하면서 새롭게 변화하고 또한 새로운 기술과 기능을 접목하기 위한 요구가 발생하였다.

<표 1> HTML 내에서의 순수 Text 의 비율(byte)

분류 / 문서	전체 문서	순수 Text	Text 의 비율
A	12934	3025	23.4%
B	8947	1021	11.4%
C	16989	2172	12.8%
D	44148	4514	10.2%
E	11856	2274	19.2%
합계	94874	13006	13.7%

따라서 본 논문에서 제안 하고자 하는 방법은 ML(Markup Language)형태의 언어를 사용하여 HTML 의 문서를 WML 로 변화 시키는 방법을 제안한다. 여기에서 HTML 이나 WML 의 분석 능력은 이미 확보 되어 있다고 가정한다. 즉 ML 의 TAG 및 내용들을 파싱하는 능력을 가지고 있다고 생각될 때 새로운 몇 가지의 TAG 를 가진 IML(Information Markup Language)를 만들게 되었다. 이 언어는 ML 형태의 문서에서 정보 추출을 목적으로 하고 있으며 추출된 정보를 사용자가 원하는 형태로 가공하는 것을 목적으로 한다.

또한 기존의 HTML 에는 전달하고자 하는 일반 내용의 비중이 보통 20% 미만임을 감안할 때 이중에서 무선 이동 단말기에서 불필요한 TAG 나 내용들을 적절하게 검색, 추출하여 새롭게 구성한다면 새롭게 만들어지는 WML 문서의 품질은 IML 을 작성하는 노력에 따라서 증가할 것이다.

2. 정보 추출 언어가 지녀야 하는 기능

정보 추출을 위해서 많은 방법이 존재 할 것이다. 그 중에서도, 본 논문에서 HTML 에서 정보 추출을 하기 위한 방법으로, HTML 문서 상에서 정보 추출 포인트를 이동을 하며 해당 정보를 찾고, 찾아낸 정보를 추출하여 가공한 후 새로운 모습으로 표시하는 기능들을 가져야 하며 그 기능을 요약해 보면 다음과 같다.

- HTML 내에서의 이동기능
- 표시기능
- 정보 추출 기능
- 정보 저장
- 저장된 정보 사용 기능

2.1 HTML 내에서의 이동 기능

많은 TAG 들이 존재하는 HTML 안에서의 이동은 TAG 들을 이용해서 이동을 하게 된다. 즉 정보가 있는 위치는 문서 전체로 보면 처음에서부터 n 번째 TAG 다음 위치, 또는 상대적인 위치에 놓이게 된다. 따라서 정보가 있는 위치는 IML 에서 정의된 "이동 기능 Tag"를 사용하여 쉽게 이동 할 수 있다.

<표 2> 이동 기능 Tag 를 이용한 예

```
<SKIP TAG="/TABLE">
<SKIP TAG="FONT" COUNT=5>
```

위와 같이 표기를 하면 "</TABLE>"을 만날 때 까지 이동을 하고 "<FONT>"를 5 번 만날 때 까지 이동을 한다. 즉 "<SKIP>을 사용하여 전진하는 이동을 처리하게 된다.

<표 3> 이동 기능 Tag

TAG	기능	사용 예
<SKIP>	해당 Tag 다음으로 포인트를 이동 시키다.	<SKIP TAG="FONT" COUNT=5>
<BLOCK>	블록 안에서 반복 적으로 수행 하면서 해당 Tag 를 만나면 블록을 빠져 나간다.	<BLOCK> : </BLOCK TAG = "/TABLE">
<REWIND>	포인트를 처음으로 이동 시키다.	<REWIND>

2.2 표시 기능

ML 자체가 표현을 자유롭게 하는 언어 이기 때문에 표시 자체는 일반 ML 처럼 표시하면 되고 특별히 형식을 필요로 하는 경우에는 출력 형식을 정하여 출력 할 수 있다.

<표 4>표시 기능 Tag

TAG	기능	사용 예
문자열	입력된 내용을 표시한다.	<WML>
<PRINT>	해당 형식으로 표시한다.	<PRINT FORMAT="%02d" VALUE = $(age)>

2.3 정보 추출 기능

문서의 정보를 추출하여 본 논문에서는 Contents 와 Attributes 에서 정보를 추출하고 추출된 정보는 변수 형태로 저장되어 사용되어진다.

<표 5> 정보 추출 기능 Tag

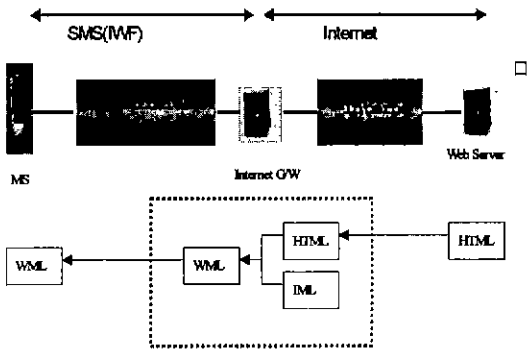
TAG	기능	사용 예
GET	Contents 나 Attributes 정보를 추출한다.	<GET TAG="FONT" KEY=cont> <GET ATTRIBUTE="HREF" KEY=url>

2.4 정보 저장/사용 기능

추출된 정보는 변수 형태로 저장되어서 사용되어진다. 또한 일반 WAP 에서와 마찬가지로 그 변수 정보는 세션이 바뀌기 전까지 그대로 유지된다. 그 사용 예는 2.2 와 2.3 과 같이 <GET>을 사용하여 원하는 정보를 지정한다. Contents 인 경우에는 Contents 바로 앞에 위해 있는 Tag 명을 표시하며 Attribute 인 경우에는 Attribute 명을 표시한다. 저장된 내용은 기본적으로 문자열 성격을 가지며 표시 형식을 변화 시키고자 하는 경우에는 <PRINT>를 사용하여 해당 성격으로 변화 시킬 수 있다.

3 문서 정보 추출 방법

본 논문에서 제안한 언어는 HTML 문서에서 정보를 얻어내어 새롭게 사용하고자 만들어진 언어이다. HTML 문서에서 정보를 추출할 때의 기본적인 동작은 HTML 내에서 정보가 있는 곳으로 이동하여 정보를 추출하고 추출된 정보를 원하는 형태로 만드는 것이다.



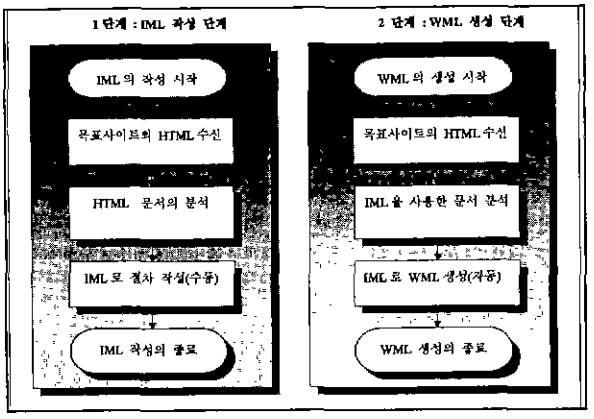
<그림 1> 변환 언어를 사용한 HTML 에서의 WML 로의 변환

1 단계는 벽돌을 만들기 위하여 틀을 만드는 것과 같이 해당 사이트의 기존 HTML 을 분석하여 IML 을 작성하는 단계가 필요하다. 이 과정은 수동이기 때문에 번거로울 수 있으나 필요한 정보들만을 재 배열하기 때문에 보다 양질을 정보를 얻을 수 있다.

1 단계의 내용을 살펴 보면 변환 하고자 하는 사이트를 분석한다. WML 로 변환 했을 때 보여져야 하는 부분과 불필요한 부분을 찾아낸다. 다음은 원 문서인 HTML 안에서 추출할 정보가 있는 위치를 정한다. 다음은 IML 언어를 사용하여 추출된 정보와 절차를 기술하면 IML 이 완성된다.

2 단계는 1 단계에서 만들어진 IML 을 이용하여 HTML 을 WML 로 변환하는 과정이다. 컴퓨터에 컴퓨터 언어로 프로그래밍해 놓으며 그 절차에 따라서 작동 하는 것처럼 IML 에 기술된 순서에 따라서 HTML 에서 정보를 추출하여 새로운 WML 을 만드는 것이다.

먼저 HTTP를 통하여 받아온 문서를 IML 사용하여 분석 변환 한다. IML 에 기술된 절차를 따라서 HTML 을 이동하면서 정보를 추출하고 WML 로 출력하는 과정을 반복하면서 WML 을 완성한다.



<그림 2> IML 생성 및 HTML 에서의 정보 추출 방법

4 장점 및 단점

기존의 방법에서 같이 일괄적으로 변화하는 것이 아닌 중간에서 한 과정을 더 거치면 아래와 같은 장점과 단점이 존재한다.

4.1 장점

IML 과 같은 언어를 사용하면 다음과 장점들이 존재한다.

- HTML 과 같은 문서에서 정보를 얻어 낼 수 있다.

원하는 위치를 정하는 것으로 해당 정보를 얻을 수 있다. 기존 방법을 생각해 보면 정보를 얻는 것은 마찬가지이지만 원하지 않는 많은 정보로 인하여 진짜 알맹이의 가치가 떨어지게 하는 것과는 달리 중요 부분이나 꼭 필요한 부분만을 추출할 수 있는 것이다.

- 얻어진 결과 원하는 형태로 만들 수 있다.

얻어진 정보는 변수 안에 담겨져 있으며 표현 방법의 기술만으로 원하는 모습으로 만들어진다. 즉 WML 로의 변환 뿐만이 아닌 다양한 형태로 자료를 만들 수 있다.

- 단순하고 쉬운 언어로 습득과 사용이 쉽다.

ML 이 가지고 있는 장점 중의 하나가 비교적 배우기 쉽다는 것이다. Tag 와 Attribute 로 이루어져 있는 모습으로 HTML 을 알고 있는 사람 들은 보다 쉽게 언어를 사용 할 수 있다.

- Script 처럼 IML 문서를 변경함으로 새로운 결과를 얻을 수 있다. HTML 의 경우와 마찬가지로 컴파일 과정 없이 텍스트 문서의 수정만으로 원하는 결과를 얻을 수 있다.

4.2 단점

IML 과 같은 언어를 사용하면 다음과 단점들이 존재한다.

- IML 을 작성해야 한다.

변환 절차를 하나하나 기술해 주어야 한다.

- URL 별로 IML 이 존재하여야 한다.

인터넷 내의 정보 모습들이 다르기 때문에 유사하지 않은 사이트의 경우는 각 각의 변화 시키기 위하여 서로 다른 IML 이 존재하여야 한다.

5. 분석

인터넷 상에서 정보를 수집하는 로봇이 등장한지도 많은 시간이 흘렀다. 로봇 역시 정보를 분석하고 추출하는 능력을 가지고 있다. 그러나 정보를 사람의 선별능력 정리하는 기능은 아직 가지고 있지 못하다 그 좋은 예로 로봇을 사용하여 수집한 많은 정보가 있는 검색 사이트 보다도 적은 정보를 가지고 있지만 사람에 의하여 가공되고 정리된 검색 사이트가 더 인기가 있는 것을 볼 수 있다.

위에서 보는 바와 같이 IML 은 단점을 가지고 있다. 가장 큰 단점은 HTML 에서 WML 로 자동적으로 변환되는 것이 아닌 IML 을 통하여 변환된다는 점이며 IML 을 사람이 구축해야 한다는 것이다. 하지만 IML 도 어느 정도는 자동적으로 작성될 수 있을 것으로 본다. 즉 사이트와 관련된 특성 패턴을 가지고 IML 을 자동적으로 생성하는 도구를 만드는 것이다. 하지만 이 과정에서도 어느 정도는 사람의 직관이 들어가는 것이 정보의 품질을 높일 수 있을 것으로 보여 진다.

IML 의 사용은 표현의 자유를 부여하는 것이다. Internet 을 사용하면서 수 많은 정보를 얻을 수 있다 하지만 그 형식은 HTML 이나 또 다른 형태의 모습을 지니고 있다. 이러한 문서 안에서 내가 원하는 정보를 추출하고 그 추출된 결과를 내가 사용할 수 있는 형태로 자유롭게 기술 할 수 있다는 것은 정보 가공의 열쇠를 부여한 것이다. 현시점에서 IML 작성을 보다 쉽게 작성할 수 있는 부분이 과제로 남을 수 있지만 HTML 작성할 때 사용하는 많은 저작 도구들이 등장하는 것처럼 IML 의 저작을 도와 주는 도구의 등장은 시간 문제라고 생각되며 보다 다양한 정보 추출 방법이 추상에서 현실화 될 수 있을 것이다. 나아가 서로 다른 형식의 형 변환 과정에서 WAP 의 WML 을 위한 좋은 변환 도구라고 생각되며 이로 인하여 HTML 로 작성된 양질의 정보가 이동 단말기 사용자에게 전달될 수 있을 것이다.

6. 결론 및 향후 방향

본 논문에서 기술한 IML 은 실제 사용을 위하여 구현되었다. 따라서 사용상의 불편함이나 문제점은 계속적으로 버전업 될 것이다. 본 논문에서 제시한 방법은 상대적으로 정보의 질이 떨어지는 일대일 변환과 고품질의 운용자 정의 변환 중에서 후자쪽을 선택 했다. 그 이유는 모든 정보를 변환하여 이동 단말기로 전달하면, 너무 많은 양의 정보가 전달되어 사용하로 하여금 인터넷 사용을 방해하거나 포기하게 만든다. 따라서 단말기의 용량과 처리 형태를 고려한 모습으로 운용자가 가공하여 전달하는 방법을 채택한 것이다.

현 사회는 하나의 도구만이 존재하는 것이 아니며 여러 도구를 사용하여 더 좋은 결과를 만들 수 있다면 그렇게 하는 것이 효율적일 것이다. 본 논문에서 제시한 방법은 기존의 HTML 을 WML 로 변환하는 여러 방법 중에서 부족하다고 생각되는 부분을 개선하는 용도로 사용될 수 있다. 하지만 기존의 모든 방법이나 도구를 부정하고 독단적으로 사용되는 것이 아닌 서비스의 종류와 형태에 따라서 기존의 방법과 공유하며 발전하기를 바란다.

참고문헌

[1] WAP Forum, Ltd , "WAP Binary XML Content Format", Proposed Version Feb 1999
[2] WAP Forum, Ltd , "WAP WML", Proposed Version Feb 1999
[3] WAP Forum, Ltd , "WMLScript Standard Libraries Specification", Proposed Version Feb 1999
[4] WAP Forum, Ltd , "WMLScript Specification", Proposed Version Feb 1999
[5] Nokia Corporation , "Developer's Guide", Dec 1998
[6] Nokia Corporation , "Getting Started", Dec 1998
[7] Nokia Corporation , "WML Reference", Dec 1998
[8] Nokia Corporation , "WMLScript", Dec 1998
[9] Nokia Corporation , "Developer's Guide", Dec 1998
[10] Unwired Planet, "UP.SDK Getting Started Guide",Jan 1999
[11] Unwired Planet, "UP.SDK Developer's Guide",Jan 1999