VoiceNews: XSL을 이용한 웹 컨텐츠 변환기법

김원철° 황인준
아주대학교 정보통신전문대학원 정보통신공학과
{wc323°, ehwang}@ajou.ac.kr

Transformation scheme of web contents using XSL

Woncheol Kim° Eenjun Hwang
The Graduate School of Information and Communication, Ajou University

요 약

무선 단말기의 보급과 네트워크 기술의 발전은 무선 단말기를 이용한 인터넷 접속을 증가시키고 있다. 그러나 대부분의 웹 페이지들이 데스크탑에 최적화 되어 있어 무선 단말기를 이용하여 사용자가 원하는 부분에 접근하기까지 반복적인 스크롤링을 해야하는 불편한 점이있다. 기존의 대부분 연구들이 웹페이지를 요약하는 기법을 제안하였지만, 대부분의 웹 페이지들은 한 페이지에 세분화된 섹션과 많은 내용을 담고 있기 때문에 제한된 화면과 입력장치를 가진 무선단말기에 대한 최적화된 해결책이라고 할 수 없다. 이런 문제점을 해결하기 위해 본 논문에서는 웹의 뉴스 페이지내의 뉴스의 섹션을 추출하고, 무선 환경에 적합하도록 VoiceXML형태로 변환해 주는 기법을 제안한다. 본 논문에서 제안된 기법을 통해 사용자는 무선 단말기의 각종 단점을 극복함과 동시에 뉴스에서 선호하는 섹션의 맞춤형 뉴스 서비스를 제공받을 수 있다.

1. 서 론

무선 단말기의 대중화와 무선 인터넷 기술의 발전으로 인해 무선 단말기를 통한 뉴스 페이지 접속이 크게 증가하고 있다. 그러나 대부분의 뉴스 페이지 구성은 데스크톱의 넓은 화면에 최적화되어 있기때문에 상대적으로 작은 화면과 제한된 인터페이스를 가진 무선 단말기를 통한 원활한 검색이 어렵다. 뉴스 페이지내 기사 검색을 위해서는 많은 스크롤링이 요구되며, 주제별 섹션으로 세분화되어있는 구조의 전체적인 이해가 어렵다. 이러한 단점을 극복하기 위해 뉴스 페이지나 일반적인 웹 페이지를 무선 단말기에 제공하기 위한 연구들이 있다. Digestor[1]는 페이지의 구조 변형과 문장의 제거를 이용하여 웹 페이지를 요약하였고, Pda++[2]은 화면에 줌 형식의 인터페이스를 제공함으로써 작은 화면을 가진 무선 단말기에 웹 컨텐츠를 제공해주었다. WEST[3]는 각 페이지를 포커스(focus)와 문맥(context)방식을 이용해서 작은 화면에 카드 형식의 인터페이스를 사용하였다. Power Browser[4]는 웹 페이지를 STUs(semantic textual units)로 나누어서 요약한 후 무선 단말기에 3 단계로 나누어서 보여주었다. 그러나 이러한 연구들은 일반적으로 웹 페이지의 일정 부분의 내용을 제거하거나 요약하여 제공함으로 본래 페이지가 가지고 있는 의미 손실에 대해 고려하지 않았고, 대부분의 뉴스 페이지들이 한 페이지에 세분화된 섹션과 많은 내용을 포함하고 있기 때문에 제한된 인터페이스를 가진 사용자가 만족할 만한 수준의 서비스라고 할 수 없다. 예를 들면 자동차를 운전하는 사용자의 경우, 무선 단말기를 통해 웹 컨텐츠를 이용하려고 할 때, 운전을 하며 작은 단말기의 인터페이스를 통해 뉴스 컨텐츠를 이용할 수 없다. 따라서 웹 컨텐츠를 텍스트로 전달하는 것이 아닌 VoiceXML로 전달해 준다면 사용자는 운전을 하더라도 자신이 원하는 웹 컨텐츠를 음성으로 전달 받을 수 있다.

본 논문에서는 기존의 연구에서 주로 사용되는 방법인 웹 페이지를 무선 단말기 화면 크기에 맞게 형태를 축소하여 단말기 화면에 보여주는 방법이 아닌, 뉴스렛 기법을 이용하여 뉴스 섹션을 추출하고 XSL을 통해 VoiceXML형태의 음성 서비스 제공기법을 제안한다. 제안된 기법을 이용하여 무선 단말기의 화면 크기를 극복함과 동시에 사용자는 선호 분야에 대한 뉴스 서비스를 제공 받을 수 있다.

2. 관련 연구

무선 단말기의 보급과 이용으로, 무선 단말기를 통해 웹 페이지를 효과적으로 제공해주기위한 다양한 연구들이 진행중이다. 진행되는 연구들을 살펴보면, 첫번째 방법은 웹에 존재하는 웹 페이지들 중 무선 단말기를 위해 제공할 필요가 있는 페이지를 단말기가 수용할 수 있는 페이지로 재작성하는 것이다. 일반적인 재작성 방법은 WML(Wireless Markup Language)[5]을 이용한다. 그러나 이와 같은 방법은 데스크톱과 무선 단말기를 위한 페이지를 각각 별도로 준비해야하는 커다란 부담을 가지게 된다. 두번째 방법은 현존하는 웹 페이지의 구조와 동일한 모형으로 축소하고 변형하는 방법이다. 이와 같은 방법을 이용한 연구들을 살펴보면, 주석기반의 시스템(Annotation-based system)[5]은 웹 페이지의 내용을 주석처리함으로 웹 페이지의 요약 및 제거를 가능하게 하였다. 하지만 이런 방법은 WML기반 재작성 방법과 유사하게 현존하는 모든 페이지에 주석처리를 해야하는 부담을 가지게 된다. 웹 페이지를 자동으로 요약한 후 모바일 단말기에 전달하는 연구들로는 서론에서 언급한 Digestor, Pda++, WEST, Power Browser와 같은 연구들이 있다. 그러나 이런 연구들은 무선 단말기의 문제점을 해결하기 위해 단지 웹 페이지의 내용을 축소하여 작은 화면을 가진 무선 단말기에 제공해왔다. 보다 유연하게 웹 페이지정보를 단말기에 제공하기 위해 웹 페이지 요약기법만이 아닌 음성을 통해서 제공할 수 있는 기법이 필요하다.

3. 뉴스 섹션 추출 기법

뉴스의 각 섹션을 추출하기 위해 본 논문은 HTML 엘리먼트내의 링크의 존재와 뉴스렛 내의 키워드들에 대한 가중치를 부여함으로 각 페이지 내의 섹션 부분을 구분한다. 뉴스렛 추출에 대한 알고리즘은 그림 1과 같다.

```
NewsletPartition(p)
 parse HTML page
 construct Parse tree
 Queue Parse Tree
 while( Queue is not empty)
 if (top element in Queue has a child with at least k links)
     push all the children of top element to Queue
  else
  declare it as a newslet
  if ( user click is included in this newslet)
    return newslet
  end if
 end if
```

그림 1. 뉴스렛 추출 알고리즘

사용자가 웹 페이지를 검색할 때, 본 논문에서 제안한 시스템은 웹 문서를 분석한 후 각 HTML 요소들을 추출하고 파스트리를 만든다. 파스트리의 각 노드를 큐에 저장한다. 다음단계는 큐로부터 노드 엘리먼트를 추출한다. 자식 노드의 링크가 임의의 값보다 크면 독립적인 의미를 포함하고 있다고 판단하고 그 노드를 뉴스렛으로 결정하고, 뉴스렛 큐에 저장한다. 그러나 임의의 수보다 링크를 적게 가지고 있을 경우 그 노드는 부모노드의 토픽에 종속된다고 본다.

그림 1의 알고리즘을 통해 뉴스렛을 추출한 후, 뉴스렛이 어떤 뉴스 섹션에 포함되는가를 판단해야한다.

```
<?xml version="1.0" encoding="UTF-8"?>
<news>
    <article>
        <section name="politics">
            <items>
                <title>Mideast 'road map' on Powell's agenda</title>

    <link>"/2003/WORLD/meast/05/10/powell.mideast/index.html"
</link>
                <content>U.S. Secretary of State Colin Powell arrived
in Israel on Saturday evening, where he will discuss the newly released
"road map" for peace with Israeli and Palestinian leaders. Powell is
scheduled to confer Sunday with Israeli Prime Minister Ariel Sharon and
new Palestinian Prime Minister Mahmoud Abbas, popularly known as
Abu Mazen. </content>
            </items>
              <items>
                <title> First Bali attack trial adjourned</title>

    <link>"/2003/WORLD/asiapcf/southeast/05/12/bali.bomb/index.htm
l" </link>
                <content> DENPASAR, Indonesia (CNN) -- The first
trial of a key suspect in the October 12 Bali bombings, which only began
Monday morning, has been adjourned for a week after defense lawyers
lodged legal objections.
                </content>
            </items>
    .....
    .....
     </article>
</news>
```

그림 2. 섹션별로 구분한 XML문서

일반적으로 뉴스 정보는 시간에 따라 분류가 되고 고정된 날짜의 뉴스 페이지는 여러 개의 하위 목록으로 연결되어 있다. 하위 목록들은 또 다른 하위 페이지로 반복적으로 연결되어 있다. 이와 같이 뉴스 사이트를 보면 정치, 경제, 스포츠 등의 여러가지 섹션으로 나누어져 있으며, 스포츠는 다시 야구, 축구, 농구 등의 여러 목록으로 연결되어 있다. 따라서 뉴스렛의 섹션을 판단하기 위해 기존의 뉴스 페이지에서 추출한 각 섹션의 대표 키워드들과 비교하여 섹션의 유사도를 측정해야한다. 본 연구에서는 기존연구에서 주로 사용하는 Luhn's Keyword cluster기술과 TF/IDF를 이용하여 키워드 가중치를 계산하여 뉴스렛의 섹션을 판단한다. 뉴스렛의 섹션이 판단되면 문서를 재구성하고 XSL처리를 하기위해 섹션별로 구분한 새로운 문서를 만들어야 한다. 그러나 현재 대부분의 뉴스페이지들이 HTML문서로 제작되어 있다. HTML로 제작된 문서는 평면적이고, 화면 표시 중심적 구조를 가지고 있어 데이터의 의미적인 정보를 직접 지정할 수 없고 데이터의 계층 구조 또한 표현 할 수 없다. 의미있는 정보 표현을 위해, 문서의 구조와 의미에 관한 정보를 표현 해주는 XML형식의 문서로 변환이 필요하다. 따라서 각 섹션에 대한 정보를 이용하여 XML문서를 작성한다. 그림 2은 각 섹션별로 구분된 XML문서를 보여준다

4. 뉴스 컨텐츠 전달 기법의 전체 구조

VoiceXML을 통한 뉴스전달 기법은 캐쉬의 문제와 모듈 삽입의 용이한 클라이언트와 서버 사이의 프록시 서버에 위치하게 된다. 그림 3는 전체적인 구조를 나타낸다. 무선 단말기의 사용자가 웹 페이지를 요청하였을 때 처리 순서는 다음과 같다.

1. 무선 단말기 사용자가 웹 페이지를 요청하였을때, 추출엔진은 사용자의 선호 사항에 대한 정보와 단말기에 대한 정보를 파악한다. 이때 웹 서버로부터 응답메시지를 받은 프록시 서버는 사용자를 파악하고 사용자 선호사항 데이터베이스로부터 저장되었던 정보를 추출한다. 그런 후 HTML문서를 뉴스렛 기반의 XML문서로 변환하기 위해 사용자 정보와 HTML문서를 HTML 파서에 전달한다. HTML 파서 모듈이 문서를 전달 받았을 때, 파서는 문서내의 배너와 그림을 제거하고, 문서를 각 노드로 분리하고 파스 트리를 생성한다. 뉴스렛을 만들기 위하여 Link analyzer 모듈이 파스 트리를 생성후 링크 정보를 추출한다.
2. Newslet Divider 모듈은 뉴스 페이지내에서 뉴스의 단위를 파악하기 위해 파스 트리를 이용과 알고리즘을 통해 뉴스렛을 추출하고, 추출된 뉴스렛들을 Content Generation 모듈에 전달한다. Content Generation 모듈은 Newslet Divider 모듈에 의해 뉴스의 각 섹션을 XML문서로 변형한다. 그림 2와 같이 웹 페이지를 뉴스렛 기반의 XML문서로 만듦으로, XSL Template의 적용을 가능하게 한다.
3. XSL Filter는 미리 작성된 XSL을 VoiceXML로의 변환을 이용하여 무선 단말기 장비의 제약을 완화 시킨다.

위와 같은 처리순서를 이용한 VoiceNews시스템은 웹 페이지를 VoiceXML로 변환하고 불필요한 배너와 그림등의 정보를 제거함으로 무선 단말기의 작은 화면에 제약사항을 완화시켰다.

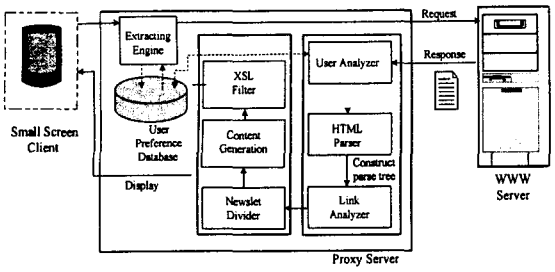

그림 3. VoiceNews 시스템 전체구조

5. 실험 및 분석

본 논문은 웹 브라우징 기능을 가지는 PDA와 유사한 환경을 제공하기위해 모바일 에뮬레이터를 이용하여 구현하였고 이를 이용하여 성능 평가를 하였다. 뉴스 페이지로부터 키워드를 추출하고 사용자가 뉴스 페이지내에서 클릭한 링크와 연관된 뉴스렛을 추출하고 사용자의 선호하는 뉴스섹션에 관한 정보를 데이터베이스에 저장하는 것은 데스크톱환경에서 Windows XP 환경에서 개발되었다. 자바 J2SE1.4.1 Swing의 SAX API 이용하여 구현하였고, 추출된 사용자의 선호사항을 이용하여 모바일 에뮬레이터에 보여주는 것은 임베디드 비주얼베이직 3.0을 이용하여 구현하였다. 링크에 연결되어있는 링크앵커 키워드를 추출하기 위해 Swing이 제공하는 DTD(Document Type Definition) 기반의 파서를 이용한다. XSL 기반의 VoiceNews Translator를 구현하기 위해 썬마이크로 시스템의 J2SE 1.4.1을 기반으로 작성하였다. 주요 패키지는 사용자 인터페이스를 위해 썬마이크로 시스템의 Swing을 사용하였고 JAXP를 사용하였다. JAXP는 DOM, SAX, 그리고 XSLT를 보다 쉽게 지원해 준다. XSL는 XML문서를 다른 문서로 변환해줄 수 있는 언어로서 문서를 변환할 경우 DOM이나 SAX를 사용하는 것 보다 사용이 용이하다. 본 연구에서는 XML문서를 VoiceXML로 변환을 위해 사용하였다. 자바(Java)언어로 개발되었기 때문에 이식성과 플랫폼 독립성을 제공한다.

그림 4는 뉴스 콘텐트를 음성형식으로 제공해주기 위한 변환기이다. 이 변환기는 미리 정의한 XSL Template에 의해 뉴스 콘텐트를 VoiceXML형태로 바꾸어주며, 적용하는 XSL의 Template이 다르면 새로운 VoiceXML문서를 제공해 줄수 있는 장점이 있다. 이 모듈은 Content Generation에 적용된다. Content Generation에서는 만들어진 XML문서를 이용하여 XSL Template을 통해 VoiceXML로 만들만들어 주는 작업을 한다. 그림 4의 왼쪽 창은 뉴스페이지내에서 각 섹션에 대한 정보를 보여주고 각 섹션에 대한 정보로는 신문 기사의 타이틀, 기사에 연결하기 위한 URL, 그리고 기사의 내용으로 구분되어 있다. 그림 4의 오른쪽 창은 적용될 XSL을 보여준다. XSL은 웹 페이지의 구성스타일을 정의하는 방법을 제공한다. XSL은 XML과 함께 쓰이기 위한 목적으로 만들어졌다. XSL은 XQL(Extensible Query Language)과 함께 새로운 문서를 만들고 이미 존재하는 문서의 내용을 조정하며, 문서들을 정렬하고 표현하기 위한 XML을 다루는 방법을 제공한다.

XSL을 이용해 문서 타입이나 한개의 문서에 따라 정의된 XSL을 이용해서 각 타입에 따른 요소들이 어떻게 보여질지 정의 할 수 있다. XSL에서 각각의 요소 타입에 대한 표현정보를 탬플릿으로 정의해서 첨가 할 수 있다. XSL 태플릿은 필터 또는 반복되는 데이터에 대한 정렬, 데이타 값에 따라 어떻게 보여줄지에 대한 결정할 수 있다.

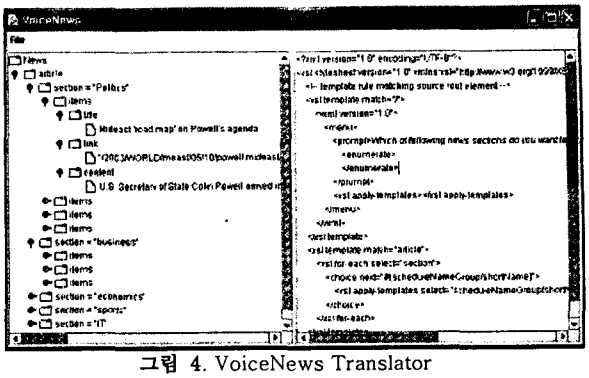

그림 4. VoiceNews Translator

6. 결론 및 향후 연구 과제

무선 단말기 사용자의 인터넷 접속이 증가하면서 점차 웹 페이지는 유선 사용자뿐 아니라 무선 사용자의 요구를 만족시켜야할 필요성이 증가하였다. 이런한 필요성으로 인해 현존하는 웹 페이지들이 각각의 기기에 맞게 표현되어야 하지만 무선 단말기는 데스트톱에 비해 제한된 인터페이스를 가지므로 사용자에게 사용의 불편함을 가져온다. 따라서 본 논문에서 무선 단말기 사용자의 요구를 만족시키고 편리한 인터페이스를 제공하기 위하여 VoiceXML을 이용한 뉴스전달기법을 제안하였다. 특히 무선단말기에 맞는 페이지의 별도 준비없이 현존하는 웹 페이지를 통하여 서비스를 제공하게 되었다. VoiceXML을 통한 뉴스전달기법을 이용해서 무선이란 환경의 낮은 대역폭의 제한을 완화시켰고 작은 화면의 제한된 인터페이스에 대한 문제를 극복하였다.

향후 계획으로는 사용자의 선호사항을 분석한 후에 무선 환경에서 뉴스 사이트만이 아닌 일반적인 사이트에도 웹 컨텐츠를 적용시키는 것이다.

7. 참고 문헌

[1] T. W. Bickmore, B. N. Schilit, " Digestor: Device-*independent* Access to the World Wide Web," Proceedings of the 6th international World Wide Web Conference, Santa Clara, CA, 1997

[2] B. B. Bederson, J. D. Hollan, " Pda++: A Zooming Graphical Interface for Exploring Alternate Interface Physics," ACM Symposium on user Interface Software and Technology, 1994.

[3] S. Bjrk, L.E. Holmquist, J. Redstrm, I. Bretan, R. Danielsson, J. Karlgren, and K. Franzn, " WEST: A Web Browser for small Terminals," ACM Sysmposium on User Interface Software and Technology, 1999.

[4] Buyukkokten, H. Garcia-Molina, and A. Paepcke, T. Winograd, " Power Browser: Efficient Web Browsing for PDAs," Proceedings of CHI' 2000, ACM Press, Amsterdam, 2000.

[5] Wap Forum, White paper(Wireless Internet Today Overview), June, 2000, http://www.wapforum.org/