무선 단말기에서의 웹 컨텐츠 변환을 위한 인덱스 추출 방법

김범호[0] 마평수
한국전자통신연구원 인터넷정보가전연구부
{mots, pmah}@etri.re.kr

Index Extraction Method of Web Contents Transcoding System for Small Display Devices

Bumho Kim[0] Pyeongsoo Mah
Dept. of Internet Appliance, Electronics and Telecommunications Research Institute

요 약

기존의 웹 컨텐츠는 유선망에 접속된 PC를 대상으로 제작되었기 때문에 무선환경의 단말기 상의 소형 디스플레이에서 표현되기 위해서는 웹 문서의 변환이 필요하다. 이를 위해 본 논문에서는 웹 컨텐츠를 자동 변환하는 데 있어서 필요한 인덱스를 추출하는 기능을 제공한다. 기존의 방법과는 달리 HTML 태그 패턴 분석뿐만 아니라 컨텐츠 정보의 속성 분석을 통하여 실시간 분석으로 웹 문서 변환에 필요한 인덱스 정보를 추출하는 방식을 제안한다. 웹 페이지에서 인덱스 정보를 추출하고 이러한 인덱스 정보를 브라우저에게 전달함으로써, 웹 문서 변환에 있어서 콘텍스트를 제공하고 변환의 유연성을 향상시킬 수 있다.

1. 서론

오늘날 웹은 엄청난 속도록 확산되기 시작하여 거의 모든 정보들은 웹을 통해 얻을 수가 있게 되었다[3]. 이러한 웹 정보들은 HTML(HyperText Markup Language)을 통하여 웹 문서로 작성되어 웹 브라우저에 의해 해석되고, PC 모니터를 통하여 전달된다. 최근에 무선 기술과 인터넷의 통합으로 이제 사용자들은 휴대전화, PDA, 인터넷 TV, 웹 패드 등과 같이 PC가 아닌 다양한 스크린 크기를 갖는 단말기를 통해서 인터넷에 액세스할 수 있게 됐다[6].

하지만 이와 같은 무선 단말기들의 디스플레이 화면의 물리적 크기는 대부분의 기존 웹 페이지가 포함하고 있는 데이터의 양을 지원하지 못하고 데이터 입력이 제한적이기 때문에 브라우저의 기능에 제약을 받게 되었다[8]. 그러므로 유선망에 접속된 PC를 대상으로 제작된 기존의 웹 컨텐츠를 다양한 디스플레이 크기의 단말기에서 사용할 수 있도록 자동으로 변환하여 추가 투자비용 없이 유무선 환경에서 웹 서비스를 제공할 수 있도록 하는 기술이 필요하다[5, 6].

웹 컨텐츠를 변환함에 있어서의 제약은, HTML 태그는 정보의 시각적 표시 방법만을 나타낼 뿐 XML 태그처럼 정보에 대한 의미를 포함하고 있지 않기 때문에 컨텐츠를 분리하기가 어렵다는 점이다[7]. 그러므로 웹 컨텐츠를 변환하기 이전에 웹 컨텐츠를 분석해 의미 있는 정보를 추출해 내야 한다. 이 때 가장 유용한 정보가 웹 문서의 구조에 대한 정보이다. 보통의 웹 문서는 일정한 구조를 가지고 있기 때문에 웹 문서의 구조를 파악한다면 효율적인 웹 문서 변환을 수행할 수 있다.

웹 문서의 구조 중에서 가장 중요하고 파악하기 쉬운 부분이 메뉴, 게시판, 테이블 등의 인덱스 구조이다. 이 같은 인덱스 구조는 일정한 형식의 컨텐츠가 나열된 형태라는 공통된 특징을 가지고 있다. 이러한 공통된 특징을 바탕으로 웹 컨텐츠에서 인덱스 정보를 추출함으로써 무선 단말기상의 브라우저가 컨텐츠를 표현하기 적당한 형식으로 웹 페이지 형식을 최적화시킬 수 있다.

기존의 웹 문서 변환 방식에서는, 문서의 구조를 파악하기 위해 HTML 태그 분석으로 문서의 구조를 파악한다[1, 2, 4]. 이와 같은 방식은 태그 중심의 분석이므로 컨텐츠 속성을 파악하지 못하므로 인덱스 정보 추출의 정확도가 떨어지게 된다. 또한 웹 문서에서 정보를 추출하는 방식에서는 컨텐츠의 유의미성을 파악하기 위해 HTML 태그뿐만 아니라 컨텐츠도 함께 분석하나, 추출하고자 하는 정보와 관련된 주제를 이용해 컨텐츠를 분석하므로 임의의 웹 문서 구조 파악에 이용하기에는 적절하지 못하다.

따라서 본 논문은, 유선망에 접속된 PC를 대상으로 제작된 기존의 웹 컨텐츠가 무선환경의 단말기 상의 소형 디스플레이에서 효율적으로 표현될 수 있도록 웹 컨텐츠를 자동 변환하는 데 있어서 필요한 인덱스 정보를 HTML 태그 패턴 분석과 컨텐츠의 속성 분석을 통하여 추출하는 방안을 제안하고자 한다.

2. 관련연구

몇몇 시스템들이 무선 단말기의 적합한 웹 문서의 변환을 지원하기 위해 제안되었다. Digestor[1, 2]는 웹 문서들을

자동으로 변환하는 시스템으로 최적의 레이아웃에 초점을 맞추었다. Digestor는 주어진 디스플레이 크기에 잘 보여질 수 있는 문서를 변환하기 위해 몇 가지 휴리스틱 알고리즘을 이용한다. 하지만 태그 패턴 분석과 컨텐츠의 위치 정보를 통하여 변환 정보를 추출하므로 인덱스 정보 추출의 정확도가 떨어지게 된다

Embley의 연구[4]에서는 정보추출을 위해 웹 문서에서 레코드의 경계를 찾는 방법은 제안하였다. 이를 위해 레코드 경계를 찾는 방법들을 위한 휴리스틱들을 제공하고 이 휴리스틱들을 결합한다. 하지만 태그 패턴과 추출하고자 하는 정보와 관련된 주제를 이용해 컨텐츠를 분석하므로 임의의 웹 문서 구조 파악에 이용하기에는 적절하지 못하다.

3. 태그 패턴 분석과 컨텐츠 속성 분석에 기반한 인덱스 추출

3.1. 고려사항

보통의 웹 문서는 일정한 구조를 가지고 있기 때문에 웹 문서의 구조를 파악한다면 효율적인 웹 문서 변환을 수행할 수 있다. 이를 위해 기존의 기술에서는 HTML 태그 패턴 분석을 통하여 문서의 구조를 파악하였으나 전체적인 구조를 파악하기 위해서는 웹 문서의 HTML 태그 분석에 많은 시간이 소요되므로, 웹 문서를 실시간에 자동으로 변환하기 위해서는 몇 가지 중요한 구조 정보를 추출하는 것이 중요하다. 또한 임의의 웹 문서 구조를 파악하기 위해서는 컨텐츠의 속성 분석이 필요하다.

본 논문에서는 웹 문서의 구조 중에서 핵심적인 구조인 인덱스 정보를 추출하는 것을 그 목적으로 한다. 기존의 방법과는 달리 HTML 태그 패턴 분석뿐만 아니라 컨텐츠 정보의 속성 분석을 통하여 실시간 분석으로 웹 문서 변환에 필요한 인덱스 정보를 추출하는 방식을 제안한다. 추출하고자 하는 인덱스의 종류를 다음과 같이 구분하고 그 특징을 인덱스 추출에 적용한다.

	메뉴형 인덱스	게시판형 인덱스	테이블형 인덱스
컨텐츠 속성 태그	일정	다양	일정
컨텐츠의 길이	짧음	비교적 길고 다양함	중간
컨텐츠 길이의 표준편차	적음	큼	중간
컨텐츠 개체 속성	텍스트, 이미지, etc	텍스트	텍스트, 이미지, etc

3.2. 시스템 구조

인덱스 추출의 가본적인 시스템 구조는 다음과 같다. 무선 단말기가 무선망에 연결되어 있으며 인덱스 추출기를 통해서 인터넷 상의 웹 서버와 연결된다. 사용자는 단말상의 웹 브라우저에서 웹 문서 요청을 하면 인터넷을 거쳐 웹 서버에서 웹 문서를 전송하고 인덱스 추출기에서 인덱스 정보를 추출해 웹 문서와 함께 요청한 단말기로 전송한다. 전송된 웹 문서와 인덱스 정보는 웹 브라우저에서 디스플레이 성능에 적합하도록 변환된다.

3.3. 인덱스 추출기

인덱스 추출기는 HTML 태그 트리 생성기에서 생성된 HTML 태그 트리를 입력으로 받아들여 HTML 태그 트리를 탐색한다. 탐색 중에 분리 태그가 나타나면 분리 태그를 추출하여 최종적으로 분리 태그 정보를 추출한다.

분리 태그란 웹 문서를 분석하기 위해 서브 트리로 구분할 때 사용되는 태그를 말한다. 대부분의 웹 문서 형식은 규칙적이므로 인덱스의 구조도 인덱스를 구분해주는 몇 가지의 일정한 태그를 이용해 이루어진다. 이 같은 분리 태그로 태그 트리를 분리함으로써 인덱스 추출의 정확도를 높일 수 있게 된다. 다음은 분리 태그들을 나열한 것이다.

분리 태그 = {<HR>, <TABLE>, <LI>, <MENU>, <Hn>}

분리 태그로 추출된 서브트리는 최소 분리 태그 트리 단위로 추출된다. 최소 분리 태그는 컨텐츠 단위의 태그 분석을 위해 서브 트리를 하나의 컨텐츠를 포함하는 트리로 구분하는 태그이다. 다음은 최소 분리 태그를 기준으로 서브트리를 분석해 하나의 컨텐츠를 포함하는 최소 분리 태그 트리를 추출한다.

*최소 분리 태그 = {
, <TR>, <TD>, <UL>, <OL>}*

추출된 최소 분리 태그 단위로 각 서브트리 S의 태그 분석 점수(TAS: Tag Analysis Score)값과 컨텐츠 분석 점수(CAS: Contents Analysis Score) 값을 구한 후 다음과 같은 공식으로 인덱스 점수(IS: Index Score) 값을 구한다.

$$IS(S) = \alpha \cdot TAS(S) + (1-\alpha) \cdot CAS(S)$$

매개 변수인 □는 태그 분석 점수와 컨텐츠 분석 점수 사이의 비중을 조절하는 변수이다. □ 의 값이 클 경우에는 태그 분석 점수의 비중이 높아지게 되므로 게시판형 인덱스 컨텐츠를 추출하는 경우이다. 반면 □의 값이 작을 경우에는 메뉴형 인덱스 컨텐츠를 추출하는 경우이다.

위의 식에서 구해진 IS 값으로 인덱스 정보를 추출하게 된다. 다음은 HTML 태그 패턴 분석과 컨텐츠 속성 분석을 통하여 TAS값과 CAS값을 구하는 방법에 대해 서술한다.

3.3.1. 태그 패턴 분석

분리 태그로 추출된 서브 트리들 중에서 일관성 있게 반복적으로 나타나는 태그 쌍들과 태그 속성들이 존재할 수 있는데, 이 태그 패턴들을 이용하여 태그 패턴의 정도를 계산한다. 서브 태그 트리를 DFS(Depth First Search) 방식으로 탐색하면서 반복적으로 나타나는 태그들의 일관성을 조사해 태그 패턴과 속성을 추출해낸 후, 태그 분석 점수를 계산한다. 다음은 서브트리 S의 태그 분석 점수(TAS)를 계산하는 공식이다.

$$TAS(S) = \alpha \cdot RPS(T,S) + (1-\alpha) \cdot AS(T,S) \qquad S = \sum_{i=1}^{n} S_i$$

RPS(T, S)는 반복 패턴의 점수(RPS: Repetition Pattern Score)이고 AS(T, S)는 속성 태그의 점수(AS: Attribute Score)이다. 매개 변수인 □는 반복 패턴의 점수와 속성 태그 점수 사이의 비중을 조절하는데 쓰인다.

서브트리 S의 반복 패턴 점수인 RPS(T, S)는 태그 트리에서 일관성 있게 반복적으로 나타나는 태그 쌍들의

반복되는 정도를 점수로 환산한 값이다.

$$RPS(T,S) = \prod_{i=1}^{n} \frac{RP(T,S_i)}{RP(T,S_1)}$$

RP(T, Si)는 반복되는 태그의 리스트이고 (RP(T, Si) / RP(T, S_1))값은 첫 번째 최소 분리 태그 트리의 태그 패턴에 대한 i번째 최소 분리 태그 트리의 태그 패턴의 일치되는 비율이다.

서브트리 S의 속성 태그 점수 AS(T, S)는, 글자의 속성 태그나 단어와 구절에 효과를 주는 태그의 경우에 다음 속성 태그가 나올 때까지 그 속성이 그대로 유지되므로 반복 패턴으로 분석되지 못하는 경우에 속성들의 일관성을 점수로 계산한 것이다. 다음은 속성 태그를 분류한 것이다.

글자 속성 태그 = {, <font size = "1~7">, <font face = "글꼴">, <font color = "RGB">, }
논리적 속성 태그 = {<EM>, <STRONG>, <DFN>, <VAR>, <CODE>, <CITE>, <KBD>, <SAMP>}
물리적 속성 태그 = {<B>, <I>, <TT>, <U>, <S>, <Strike> <BIG>, <SMALL>, <SUB>, <SUP>}

서브트리 S의 태그 속성 점수인 AS(T, S)는 서브 태그 트리 S에서 속성 태그를 비교해 값으로 환산한 값이다.

$$AS(T,S) = \prod_{i=1}^{n} \frac{A(T,S_i)}{A(T,S_1)}$$

위의 식에서 AS(T, S)는. A(T, Si)는 첫 번째 최소 분리 태그 트리의 태그 속성 리스트이고 (A(T, Si) / A(T, S_1))값은 첫 번째 최소 분리 태그 트리의 태그 속성에 대한 i번째 최소 분리 태그 트리의 태그 속성의 일치되는 비율이다.

3.3.2. 컨텐츠 속성 분석

컨텐츠 속성 분석에서는 서브 태그 트리에 포함되어 있는 실제적인 컨텐츠의 다양한 속성을 분석해 컨텐츠 분석 점수(CAS)를 계산한다. 컨텐츠 분석에는 컨텐츠 길이 비교, 컨텐츠 길이의 표준편차 비교, 컨텐츠 속성 비교의 세가지 방법을 조합하여 판단한다. 다음은 컨텐츠 분석 점수(CAS)를 구하는 공식이다.

$$CAS(S) = \alpha \cdot LS(S,C) + \beta \cdot SDS(S,C) + \gamma \cdot AS(S,C)$$

(단, □ + □ + □ = 1)

LS(C, S)는 컨텐츠의 길이 점수(Length Score)이고 SD(C, S)와 A(C, S)는 각각 컨텐츠의 길이의 표준편차 점수(Standard Deviation Score)와 컨텐츠의 속성 점수(Attribute Score)를 나타낸다. 세가지 매개 변수인 □, □, □는 각각 컨텐츠 길이 점수, 컨텐츠 길이의 표준편차 점수, 컨텐츠 속성 점수 사이의 비중을 조절하는 데 쓰인다.

컨텐츠의 길이 비교(LS)는 추출된 각각의 컨텐츠 리스트의 길이를 비교함으로써 유사한 길이의 컨텐츠를 인덱스로 결정하는 방법이다. LS(C, S)는 서브트리 S에서 각 최소 분리 태그 트리의 텍스트 컨텐츠의 길이의 평균값으로 다음과 같은 공식으로 구한다.

$$LS(C,S) = \frac{\sum_{i=1}^{n} L(C,S_i)}{N}$$

컨텐츠의 길이 표준편차(SDS)는 인덱스 추출의 정확도를 높이기 위해 컨텐츠 리스트의 길이의 표준편차를 비교한다. SDS(C, S)는 서브트리 S의 각 최소 분리 태그 트리의 텍스트 컨텐츠의 길이의 표준편차로 다음과 같은 공식으로 구한다.

$$SDS(C,S) = \sqrt{\frac{\sum_{i=1}^{n} (LS(C,S) \quad L(C,S_i))^2}{N}}$$

컨텐츠의 속성 비교는 컨텐츠의 속성을 비교함으로써 텍스트로 이루어진 인덱스뿐만 아니라 다른 개체로 이루어진 컨텐츠를 추출할 때 정확성을 높이는 방법이다. 컨텐츠 속성 점수인 AS(C, S)는 서브 태그 트리 S에서 컨텐츠의 속성을 비교해 값으로 환산한 값으로 구해진다.

$$AS(C,S) = \prod_{i=1}^{n} \frac{A(C,S_i)}{A(C,S_1)}$$

위의 식에서 A(C, Si)는 첫 번째 최소 분리 태그 트리의 컨텐츠 속성 리스트이고 (A(C, Si) / A(C, S_1)) 의 값은 첫 번째 최소 분리 태그 트리의 컨텐츠 속성에 대한 i번째 최소 분리 태그 트리의 컨텐츠 속성의 일치되는 비율이다.

4. 결론

본 논문에서는 유선망에 접속된 PC를 대상으로 제작된 기존의 웹 컨텐츠가 무선환경의 단말기 상의 소형 디스플레이에서 최적의 디스플레이 방식으로 표현될 수 있도록 웹 컨텐츠를 자동 변환하는 데 있어서 필요한 인덱스 정보를 추출하는 기능을 제공한다. 웹 문서의 인덱스 정보를 추출함으로써 웹 문서 변환에 있어서 콘텍스트를 제공하고 변환의 유연성을 향상시킨다. 추출된 인덱스 정보를 이용해 단말기의 디스플레이의 크기에 따라 별도의 인덱스 페이지를 만들거나 인덱스에 해당되는 컨텐츠를 링크시킴으로써 효율적인 네비게이션을 제공할 수 있다.

5. 참고 문헌

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

[2] T. Bickmore, A. Girgensohn, and J.W. Sullivan, "Web Page Filtering and Re-Authoring for Mobile Users," The Computer Journal, 42(6), pp. 534-546, 1999.

[3] S. Chandra, C.S. Ellis, and A. Vahdat, "Differentiated Multimedia Web Services Using Quality Aware Transcoding," INFOCOM 2000, pp. 961-969, Mar. 2000.

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

[5] R. Han and P. Bhagwat, "Dynamic Adaptation In an Image Transcoding Proxy For Mobile Web Browsing", IEEE Personal Communications Magazine, pp. 8-17, Dec. 1998.

[6] 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, 3(4), 1999.

[7] K. Nagao, Y. Shirai, and K. Squire, "Semantic Annotation and Transcoding: Making Web Content More Accessible," IEEE MultiMedia, 8(2), pp 69-81, April. 2001.

[8] B.N. Schilit, et al., "m-Links: An Infrastructure for Very Small Internet Devices," 7th Annual Int'l Conf. on Mobile Computing and Networking 2001, pp. 122-131, July 2001.