구조분석 에이전트를 사용한 웹사이트의 평가

정윤경, 조성배
연세대학교 컴퓨터과학과
ygyoung@candy.yonsei.ac.kr sbcho@candy.yonsei.ac.kr

Web Site Evaluation Using Structure Analysis Agents

Yun-Gyoung Chong Sung Bae Cho
Computer Science Department, Yonsei University

요 약

인터넷이 보편화되면서 그에 따른 정보량도 급증하고 있다. 웹문서량이 많아짐에 따라 웹문서를 구조를 이용하여 저장, 분석하는 연구가 활발히 이루어지고 있다. 본 논문에서는 웹사이트를 사용자가 평가하기 위해 계층적인 웹문서들의 관계를 사이트맵으로 구성하며 웹문서 내의 계층구조를 추출, 저장하고 그래픽적으로 표시하였다. 이를 위해 웹문서 내의 계층구조를 위해 W3기관의 공용으로 사용되는 Tidy라이브러리를 이용하여 URL에 대한 HTML문서를 얻고 이를 XML로 변환하였다. 변환된 XML 결과로 이진트리를 구성하고 계층구조를 표현하였다. 웹문서들의 사이트맵은 그래프형식과 계층구조형식으로 표현했는데, 그래프형식을 이용하여 사이트맵의 연결구조를 파악할 수 있게 하였으며, 계층구조를 이용하여 웹문서간의 계층구조에 따른 정보를 얻을 수 있었다. 사이트맵을 구성하기 위해 URL들의 구조를 인접리스트로 저장하였으며, 방향성 그래프형식을 이용하였다. 또한 웹문서 구조를 계층적으로 구성하기 위해 웹문서의 그래프형식에 대해 BFS(Breadth First Search)방식을 이용했다. 또한 계층적 사이트맵을 이용한 평가항목을 이용하여 증권사이트에 대해 실험하였다. 실험을 통해 본 시스템이 웹사이트 평가에 유용성함을 입증하였다.

1. 개요

많은 사람들이 정보를 얻기 위해 인터넷을 사용한다. 하지만 방대한 웹사이트를 방문하고 웹문서를 일일이 찾아야 하는 번거로움이 있기 때문에 자신이 원하는 정보를 얻기는 쉽지 않다. 몇몇 웹사이트 매니저는 사용자가 정보를 쉽게 찾을 수 있도록 사이트의 구조를 표현한 사이트맵을 제공하고 있지만, 이는 한 서버에 대한 정보를 제공할 뿐이다. 따라서 사용자가 정보를 쉽게 찾기 위해 서버에 각각에 대한 시스템적, 웹구조적인 성능평가가 요구된다. 본 논문에서는 웹문서를 이용하여 구조적으로 표현하며 웹구조를 그래픽적으로 표현함으로써 웹서버의 사이트맵을 구성하고 몇가지 항목을 이용하여 웹서버를 평가한다. 또한 웹문서의 구조를 추출하여 웹문서 내의 구조도 분석하려한다.

2. 관련 연구

2.1 XML

인터넷에 대한 관심이 급증하면서 웹상에서 데이터를 쉽게 수집하고, 검색하기 위한 표준 데이터 포맷에 대한 연구가 활발하다[1,2]. 이중 웹사이트 구조적인 데이터를 표현하는데 XML이 제시되고 있다. 이는 XML의 몇 가지 장점 때문이다. XML은 서버와 플랫폼 등에 누구나 사용할 수 있고 ,또 어디서나 사용될 수 있다. ASCII에 근간을 두고 있으므로 XML도 서버나 플랫폼, 운영체제들에 관계없이 사용될 수 있다. 모든 데이터는 XML로 저장될 수 있으며 저장된 XML데이터는 XML 파서를 사용해 손쉽게 읽어들이고 변경할 수 있다. XML은 객체지향 기술의 중요한 바탕을 구성한다. XML은 확장성이 있는 계층적인 구조를 가진 언어이기 때문에 COM,CORBA, 자바,C++등으로 만들어진 어떠한 객체들도 표현할 수 있다. XML이 모든 종류의 데이터에 적용될 수 있는 유연성을 가지고 있기 때문에 모든 계층의 어플리케이션에도 서로 다른 장점을 가지며 적용될 수 있다.

2.2. 구조 분석 에이전트

웹문서 구조를 이용한 연구가 활발히 이뤄지고 있는데, 하나는 웹구조를 이용하여 검색하는 시스템이다. 이는 웹문서의 연결 구조를 이용하여 사용자가 입력한 내용을 서버에서 해당 웹문서를 찾고 입력한 내용과 관련있는 내용을 보내주는 시스템이다. 웹문서의 연결구조를 사용자 인터페이스 관점에서 이용한 시스템의 경우 웹문서에서 웹문서간의 연결 관계를 추출하여 사이트맵을 구성하고 화면에 그래픽적으로 표시하여 웹서핑하는데 이용하기도 한다[3,4].

3. 시스템

본논문에서는 웹구조를 평가하기 위해 구조 분석 에이전트를 정의하였다. 구조분석 시스템 구성도는 그림 1과 같다. 서버에 접속하여 웹문서를 불러오면 시스템은 HTML로 구성된 웹문서를 보다 정형화된 XML로 변환시킨 후 이를 이용하여 문서의 계층 구조를 생성한다. 생성된 구조는 사이트와 구조가 저장되어 주기적으로 업데이트하게 되며, 인터페이스를 이용하여 화면에 보이거나 원하는 포맷형식으로 출력하게 된다. 웹문서 계층 구조 생성시 뽑아낸 웹문서 간의 연결관계를 저장하여 사이트맵을 구성한다. HTML문서를 XML로 변환하기 위하여 W3에 제공되는 Tidy 클래스를 이용하였으며 웹문서들의 링크에 대해서 구조 생성후 링크와 자바스크립트 언어의 HTML 문서 호출에 대한 웹문서를 다시 호출한다[1].

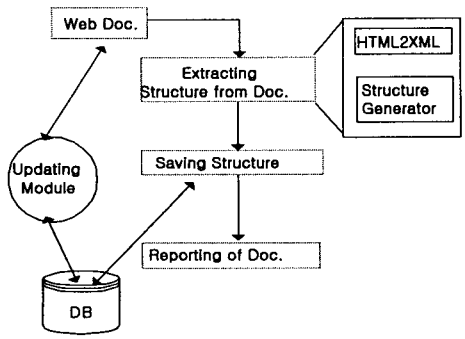

그림 1. 구조 분석 시스템.

본 연구에서는 서버에 대해 웹문서 구조를 저장하며 웹문서간의 연결관계를 이용하여 사이트맵을 구성 저장하고 계층적인 사이트맵을 이용하여 사이트에 대한 평가항목을 정의한다. 또한 사용자가 서버 주소를 입력하면 시스템은 서버의 사이트맵과 평가항목을 사용자에게 전달한다. 그림 2는 시스템 구성도이다.

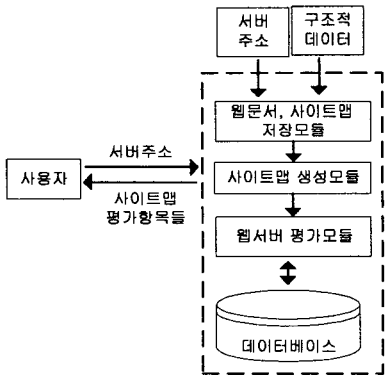

그림 2. 시스템 구성도.

시스템은 구조데이터와 서버 주소를 받아 웹문서의 구조데이터와 웹문서간의 연결관계를 저장한다. 시스템은 사이트맵을 구성하고 이를 이용하여 서버를 평가하고 평가한 값을 저장한다. 사용자가 서버 주소를 입력하면 시스템은 서버에 대한 사이트맵과 평가값들을 사용자에게 전달한다.

3.2 구조적 데이터

웹문서의 구조를 계층 구조로 나타내기 위해 HTML문서를 XML 형식으로 변환하고 이진트리를 이용하여 계층적 구조로 나타내었다. 웹문서에서 얻은 계층 구조는 사이트별로 웹문서에 대해 저장하였다. 각 웹문서에 대해서 속성이 포함되지 않은 단순구조와 속성이 저장된 구조로 나누었다. 속성이 포함된 구조는 서버주소에 대해 파일로(FileXXX.txt) 저장되며 한 서버에서 생성된 웹문서들의 경우 하나의 파일에 단순구조들을 같이 저장시켰다. 사이트와 속성이 포함된 파일 관계를 나타내기 위해 사이트리스트파일을 생성하며 URL과 URL에 따른 속성이 포함된 화일명을 저장하여 URL에 따른 구조를 파악할 때 이용하였다.

단순구조(Simple List)는 태그사이의 연결관계를 나타냈다. 저장순서는 각각의 태그에 대해 DFS(Deth First Search)방식에 운행을 따르며 내용(tagName)과 각 내용에 대한 하부(child)연결관계와 동등(neighbor)연결관계로 저장되었다. 속성이 포함된 구조(List)는 단순구조와 내용에 따른 속성 리스트(Attribute List)를 포함했다. 속성리스트는 다시 속성명과 속성값으로 이루어진다. 구조저장방식에 사용된 데이터 형식은 그림 3과 같으며 이를 이용한 저장포맷은 그림 4와 같다.

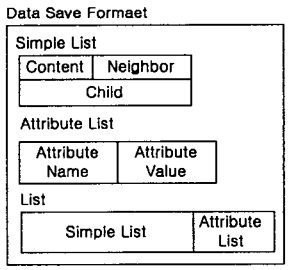

그림 3. 구조표현을 위한 데이터 포맷

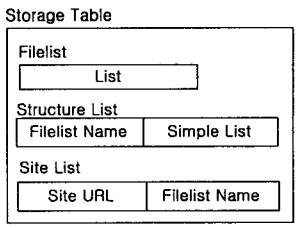

그림 4. 데이터 저장 포맷

웹문서는 연결구조의 구조를 인접리스트를 이용하여 표현하여 저장하고 저장된 연결관계를 이용하여 웹구조를 그래프 또는 계층적 구조로 나타냈었다. 그림 5는 웹문서의 연결관계를 인접리스트 형태로

나타낸 것이다.

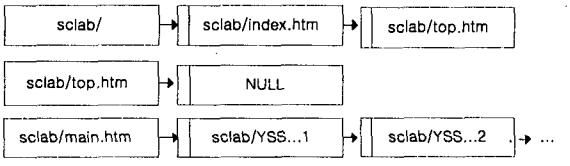

그림 5. 인접리스트

3.3 사이트 맵

인접리스트형태로 저장된 웹문서간의 연결관계를 계층 구조와 그래프 형식의 사이트맵으로 구성했다. 이를 위해 인접리스트를 BFS로 운행하며 이진트리를 구성한 뒤 계층구조의 사이트맵을 구성하였다.

그래프 형식의 사이트맵은 같은 레벨의 경우 동심원에 위치시키는 표현 방식[5]을 사용했다. 방향성을 표시하여 계층관계를 표시했으며 URL들에 대해서 작은 네모로 나타냈다. 서버의 시작 위치는 빨간색으로 표시했으면 각각 네모를 더블클릭하면 각각의 URL이 나타나도록 처리하였다.

그림 6은 한 서버의 사이트맵을 계층구조로 표현한 예이며, 그림 7은 URL에 대한 웹문서 구조 예이다. 그림 8은 한 서버에 대해 그래프 형식의 사이트맵의 예를 보여준다.

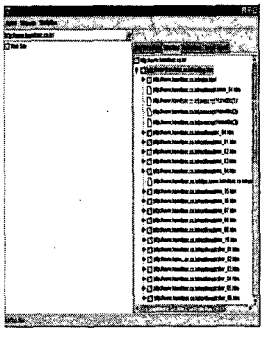

그림 6. 계층적 사이트맵의 예.

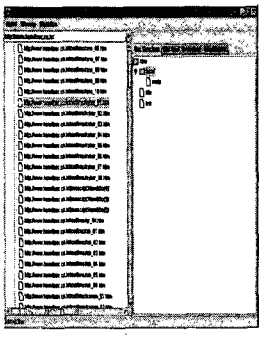

그림 7. 웹문서 구조예.

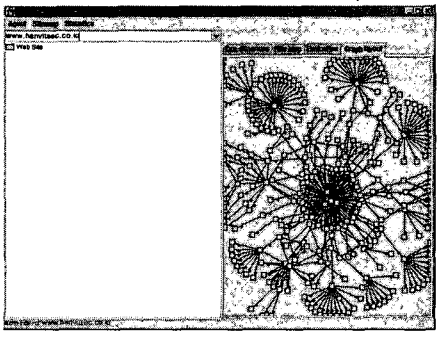

그림 8. 그래프 형식 사이트맵의 예.

4.실험

실험은 증권사이트 9개 사이트에 대해 실험하였다. 평가항목으로 계층적 사이트맵에서 평균깊이와 평균너비를 구하여 사용했다. 평균깊이 작을수록 사용자는 서버의 첫 위치에서 원하는 정보가 있는 문서에 빨리 도달하게되며 평균너비가 작을 수록 데이터가 고르게 분포되는 특성이 있다.

사이트명	평균깊이	평균너비
www.getmore.co.kr	2.046512	1.95
www.dws.co.kr	2.061151	1.940678
www.shinyoung.com	2	1.971429
www.shcyber.com	1.784615	1.784615
www.hanastock.co.kr	2.244898	1.926480
www.hanvitsec.co.kr	2.62069	1.615087
www.etrade.co.kr	2.157459	1.862769
www.webtrade.co.kr	2	1.928515
www.korearb.com	2.09009	1.909091

표 1.

표 1에서 보는 봐와 같이 9개의 증권사이트는 평균깊이 1.7~2.1사이에 분포하였고, 평균너비의 경우 1.78~1.99사이에 분포하였다.

5. 결론

우리는 웹사이트를 평가하기 위해서 웹서버의 문서구조 및 사이트맵을 저장하였다. 이진 트리를 이용하여 웹문서를 저장하였고 인접리스트를 이용하여 사이트맵을 저장하였고 사용자가 시각적으로 평가할 수 있도록 계층적, 그래프 형식의 사이트맵으로 표현하였다. 또한 평가항목을 이용하여 증권사이트에 대해 실험하였다. 실험을 통해 본 시스템이 웹사이트 평가에 유용함을 보였다.

향후 웹사이트를 보다 객관적으로 평가하기 위한 평가항목을 구성하고 구조를 이용한 내용적인 측면의 평가방법도 필요하리라 본다.

Refrenece

[1] O. Liechti, M. J. Sifer and T. Ichikawa, "Structured Graph Format: XML Meta for Describing Web Site Structure," Computer Networks and ISDn Systems, Vol. 30, pp. 11~21, 1998.

[2] http://www.w3.org/XML/

[3] http://www-db.standford.edu/lore.

[4] http://www.inxight.com/

[5] I, Herman, "Graph Visualization and Navigation in Information Visualization: A Survey," IEEE Transactions on Visualization and Computer graphics, Vol. 6, No. 1, pp. 24~43, 2000.