웹사이트의 구조분석을 위한 소프트웨어 에이전트

서연규 김경중 정윤경 조성배
연세대학교 컴퓨터과학과
{uribyul, kitestar, ygyoung}@candy.yonsei.ac.kr, sbcho@csai.yonsei.ac.kr

A Software Agent to Analyze the Structure of Web Site

Yeon-Gyu Seo, Kyoung-Joong Kim, Yun-Gyoung Jung and Sung-Bae Cho
Computer Science Department, Yonsei University

요 약

인터넷 사용의 급속한 증가로 인해 사용자는 많은 양의 정보들을 웹페이지를 통해서 공유할 수 있게 되었다. 그러나 웹문서들 중에는 구성이 제대로 되어있지 않아 원하는 정보를 얻기 위해 사용자의 많은 행동을 요구하기도하며 존재하지 않거나 변경되기 전의 사이트를 링크 함으로써 인터넷 사용의 효율성을 저하시키는 문서들도 있다. 본 논문에서는 웹사이트의 효율성을 검토하기 위한 방법으로 웹사이트의 구조분석을 위한 에이전트의 구현에 대해 설명한다. 웹사이트 구조분석을 위한 에이전트는 해당사이트와 연결된 문서들의 구조 및 이들의 연결관계를 조사하여 사용자에게 제시함으로써 웹사이트의 구조를 한 눈에 파악할 수 있도록 한다. 이러한 구조분석 에이전트는 웹문서 구조에 기반한 정보검색에 유용하게 사용될 수 있다.

1. 서론

인터넷의 급성장으로 인해 사용자는 필요한 많은 정보들을 웹에서 얻을 수 있게 되었다. 그러나 구성이 제대로 되어 있지 않은 웹문서의 경우 원하는 정보를 얻기 위해선 많은 웹페이지를 방문하도록 요구함으로써 사용자에게 지루함을 유발시킨다. 이는 웹사이트의 평균깊이가 깊거나 웹사이트의 편향성 구조에서 기인하는 문제이다. 잘 구성된 웹사이트의 경우 사용자가 적은 수의 사이트를 방문해도 원하는 정보를 획득할 수 있도록 한다.

비구조화된 문서인 웹문서에서 유용한 정보를 추출하는 방법에 대한 연구가 많이 진행되고 있다. 웹사이트 구조를 표현하기 위해서는 대부분 Wrapper를 이용한다 [1, 2, 3]. Wolfgang [2, 3]은 특정 사이트에서 웹문서를 가져온 후 SGML파서를 통해 문서구조를 획득한 다음 링크 정보를 추출하고 이들의 골격구조를 저장하는 방식을 이용하였다. 그러나 이러한 연구들의 대부분은 정보검색에 편중되어 있으며 웹사이트의 또 다른 중요한 측면인 웹사이트의 효율성에 대한 것은 거의 없다.

웹사이트의 효율성은 관련된 지식 표현과 정보의 저장의 효율성으로 볼 수 있다. 일반적으로 웹사이트는 사용자가 링크를 따라가며 원하는 정보를 얻기 때문에 사이트맵을 제공하지 않을 경우 사이트의 전체적인 모습을 파악하기란 쉽지 않다. 뿐 만 아니라 사이트맵이 제공될지라도 링크된 문서를 따라가며 하나씩 조사하지 않는 한 해당 웹사이트의 전체적인 모습을 파악하는 것은 쉽지 않으며 사이트의 크기가 비교적 큰 검색포탈의 경우에는 전체적인 사이트맵을 구성하기 위해 많은 시간이 소요된다.

본 논문에서는 이를 해결하기 위해 이는 웹사이트의 구조를 분석하는 에이전트를 제안한다. 에이전트는 웹사이트의 구조분석을 위해서 비구조화 데이터를 정형화 형태로 변환하고 태그정보들을 이용하여 구조화한다. 구조화된 문서에서 링크정보들을 추출하여 링크된 문서들을 방문하면서 사이트맵을 구성해나간다. 방문한 사이트에 대한 정보는 파일로 저장되어 다음에 참고할 수 있도록 한다.

본 논문의 구성을 다음과 같다. 2절에서는 웹사이트의 구조분석을 위한 에이전트의 모듈별 기능을 제시하며 3절에서는 구조분석 에이전트를 이용한 웹사이트의 분석결과를 제시한다 그리고 마지막으로 결론을 내린다

2. 웹사이트의 구조분석

웹사이트의 구조분석을 위해 본 논문에서는 비구조화된 웹문서를 정형화된 형태의 HTML로 변환한 후 구조생성기에 보내지면 구조생성기는 그 문서에서 태그정보와 속성들을 추출하고 포함관계를 이용하여 트리구조로 변환한다. 트리가 구성되면 트리에서 다음에 방문할 링크정보들을 추출한 후 BFS방식으로 웹사이트를 방문하면서 같은 과정을 반복한다. 그림 1은 이러한 웹문서의 구조화에 대한 과정을 보여주고 있다.

2.1 웹문서의 정형화된 형태로의 변환

전처리과정은 HTML문서를 well-formed XML형태로 변환하는 작업과 불필요한 태그 삭제작업으로 구성된다. XML은 구조적으로 명확하게 정의된 확장가능한 웹문서 정의언어이다 [1]. HTML의 장점은 웹문서를 손쉽게 작성할 수 있다는 점이다. 하지만 에이전트를 개발하는 입장에서 HTML언어는 적합하지 않다. HTML언어는

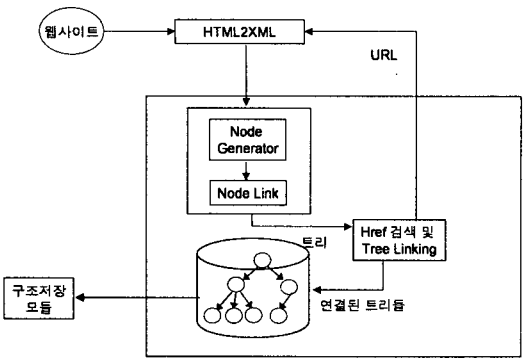

그림 1. HTML에서 구조정보의 추출

SGML상에서 정의되었기 때문에 구조적으로 모호한 점이 있다. 구조적인 모호함으로 인해 파싱작업을 수행하는 것이 쉽지 않다. 이러한 문제점을 해결하기 위해 HTML문서를 구조적으로 명확한 XML 문서로 변환하여 에이전트를 개발하려는 시도가 있다 [2].

HTML을 well-formed XML로 변환하는 작업은 다음과 같다.

1) Non-empty element의 경우 끝 태그를 첨가하여 반드시 닫혀 있도록 한다.
2) Empty element의 경우 태그가 닫히기 전에 '/'를 첨가한다.
3) 속성의 값은 반드시 인용부호로 닫는다.
4) 하나의 태그에 대해 중복 선언된 속성은 하나로 줄인다.

Non-empty element는 끝 태그가 필요한 element이고 empty element는 끝 태그가 필요없는 element이다. 예를 들어 <a> 태그는 Non-empty element에 해당하고 <img>태그는 empty element에 해당한다. 이상과 같은 작업을 통해 생성된 well-formed XML문서는 구조적으로 명확한 장점을 가진다. 구조적으로 명확한 정형화된 XML 문서를 이용하여 손쉽게 파싱작업을 수행할 수 있다. 또한 이미 개발된 많은 XML 응용프로그램을 적용할 수 있다는 장점도 있다. 표 1은 전처리과정에 들어가는 입력 HTML화일이다. 전처리과정을 통해 생성된 정형화된 XML은 표2와 같다.

변환된 문서는 구조분석작업전에 불필요한 태그들을 삭제한다. 주석이나 DTD정보등 구조분석을 하는데 필요하지 않은 정보들을 삭제한다. 이와 같은 작업은 구조분석 과정을 효과적으로 만든다.

2.2 웹문서의 구조화

웹문서의 구조화는 구조에 기반한 정보검색이나 정보획득 등에 사용될 수 있으며 또한 다양한 웹문서를 분류할 수 있는 방법을 제공할 수 있다. 본 논문에서 웹문서의 구조화는 특정 웹사이트의 문서와 링크된 문서들의 구조를 정보화하여 트리구조로 표현하고 이들 문서들의 연결관계를 쉽게 파악할 수 있도록 그래프와 트리구조로 구성하는 것을 말한다. 따라서 웹문서의 구조화는 웹문서의 구조화, 링크정보의 구조화 및 그래프 표현으로 나뉘어진다.

정형화된 HTML형식의 문서에서 구조를 추출하기 위해 태그의 포함관계가 이용될 수 있다. HTML에서 태그는 크게 시작태그와 끝태그를 가지는 non-empty 태그와 끝태그가 없는 empty 태그로 구분되어지는데 non-empty 태그의 경우 시작태그가 나타나고 그 안에

```
<html>
<head><title>hello
<body background=back.gif>
<li> this is a test
<img src="image.gif">
</body>
</html>
```

표 1. 전처리과정

```
<html>
<head><title>hello</title>
</head>
<body background="back.gif">
<li> this is a test </li>
<img src="image.gif"/>
</body>
</html>
```

표 2. 전처리 과정의 출력

다른 태그를 포함할 수 있다. 끝태그가 나타나기 전에 포함되는 태그는 그 태그안에 포함되는 태그가 된다. empty 태그의 경우 끝태그가 없기 때문에 안에 다른 태그를 포함할 수 없다. 이러한 정보를 이용하여 웹 문서를 구조화하기 위해 스택을 사용하였다. 시작태그가 나타나면 스택에 push하고 끝태그를 만날 경우 pop한다. 팝되는 태그는 스택에 top에 위치한 태그의 하위노드로 구성된다. 일반적으로 웹 문서를 트리로 표현할 경우 다차원 트리로 표현되는데 본 논문에서는 다차원 트리를 이진트리로 변환하여 사용한다.

해당 사이트에서 링크된 문서들은 그 사이트와 관련된 내용을 포함하고 있는데 이들 문서들간에는 사이클이 형성될 수 있기 때문에 사이클을 제거한 후 트리로 구성한다. 사이클생성을 방지하기 위해 이전에 방문한 사이트의 경우 하위노드로 연결하고 방문하지 않도록 한다. 링크정보는 구조화된 웹문서에서 추출되는데 Queue를 사용하여 BFS방식으로 방문하도록 하였다. 또한 방문하는 사이트가 해당 사이트의 도메인이 아닌 경우 방문할 수 없도록하여 해당 사이트에 연결된 문서들간의 관계를 (사이트맵) 파악할 수 있도록 한다.

2.3 구조적 데이터 표현

웹문서 구조를 효율적으로 저장하기 위해 속성이 포함되지 않은 단순구조와 속성이 저장된 구조로 나누어 사이트별로 저장하였다. 속성이 포함된 웹문서 구조는 파일로(FileXXX.txt) 저장된다. 또한 서버 내 웹문서들에 대한 속성이 포함되지 않은 단순구조들을 속성이 포함된 구조의 파일명을 묶어 하나의 구조리스트 파일에 저장한다. 이는 서버내의 속성을 제외한 구조만을 확인하고자 할 때 사용하기 위해서이다. 사이트와 속성이 포함된 파일 관계를 나타내기 위해 사이트리스트파일을 생성하고 사이트 URL과 속성이 포함된 화일명을 저장하였다. 이는 URL에 따른 단순구조를 구조리스트를 이용하여 얻어내고 속성이 포함된 구조를 직접 접근하는데 사용된다.

단순구조의 경우 태그사이의 연결관계를 이진트리형태로 표현하고 각각의 태그에 대해 DFS(Depth First Search)방식으로 저장하며 각 태그에 대한 자식태그와 이웃태그의 이름을 나열하였다. 그렇지 않은 경우 특정 표기를 하여 다른 노드와 연결되지 않았음을 명시하였다. 만약 해당태그가 없을 경우 특수문자로 표기하였다.

속성이 포함된 구조의 경우 단순구조와 내용에 따른 속성리스트(attribute list)를 포함하게 된다. 한 노드에 대해 속성이 있을 경우 속성에 대한 이름과 값을 순서대로 표기하여 리스트를 구성한다. 링크정보는 인접리스트를 이용하여 저장했으며 저장된 웹문서간의 구조를 그래프 또는 계층적 구조로 나타내었다.

3. 시스템 구현 및 실험결과

웹사이트의 분석을 위해 여러 웹사이트를 테스트하여 보았다. 그림 2는 테스트한 증권사이트들 중 한빛증권 사이트의 예를 보여주고 있는데 그림에서 방문한 문서들과 특정문서의 구조를 보여주고 있다. 웹문서들간의 연결관계는 그림3 및 4와 같이 그래프 또는 계층구조를 나타나게 된다. 그림 3은 방문한 사이트에서 사이클을 제거하여 생성된 링크의 구조정보를 보여주고 있으며 그림 4는 사이트맵을 그래프로 표현한 그림인데 백워드 링크와 포워드 링크를 표현하기 위해서 링크 끝에 방향표시를 하였다.

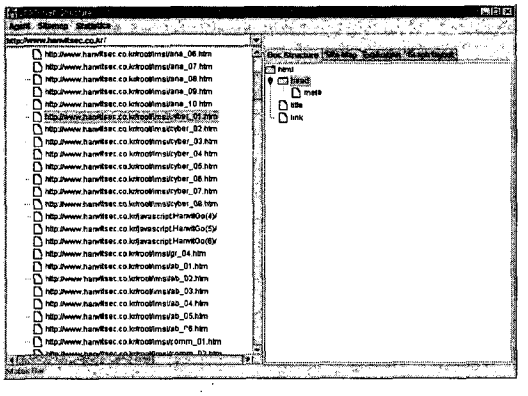

그림 2. 웹문서내의 계층구조

4. 결론 및 향후연구

웹문서의 효율성을 평가하기 위한 방법으로 본 논문에서는 웹사이트 구조분석을 위한 에이전트를 구현하였다. 이를 위해 웹문서 구조 및 사이트맵을 구성하는데 웹문서의 구조를 위해서는 HTML문서를 정형화된 형태로 변환하고 태그들의 포함관계를 이용하여 트리구조를 생성하였다. 그리고 웹문서 구조에서 링크정보들을 추출하고 BFS방식으로 링크된 문서들을 방문함으로써 사이트맵을 구성하여 해당사이트의 전체적인 모양을 그래프나 트리형태로 볼 수 있도록 하였다.

요즘들어 많은 웹사이트의 문서들이 자바스크립트를 이용하고 있는데 현재 시스템은 자바스크립트에 대한 일부만의 처리를 하고있기 때문에 정확한 사이트맵을 구성하기 위해서는 자바스크립트에 대한 정확한 처리가 필요하다 그리고 온탈리지를 이용하여 태그에 의미정보를 표현한다면 본 논문에서 생성한 웹문서 구조는 정보검색 에이전트에서 유용하게 사용될 수 있을 것이다.

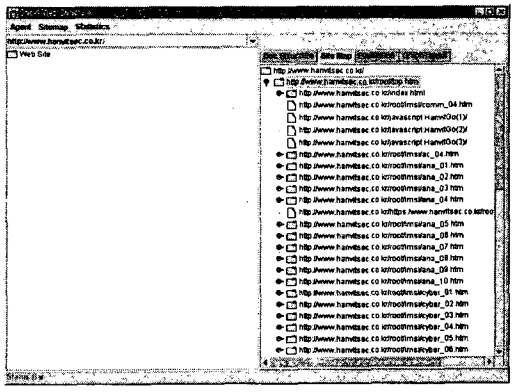

그림 3. 계층구조로 표현한 웹문서 관계

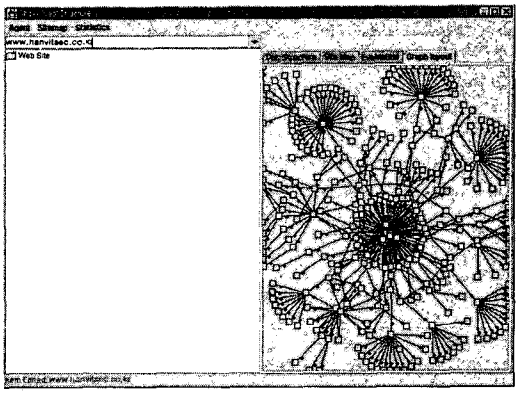

그림 4. 그래프로 표현한 웹문서간 관계

참고문헌

[1] N. Ashish and C. knoblock, "Wrapper generation for semi-structured internet sources," *SIGMOD Record*, Vol. 24, No. 4, pp. 8-15, 1997.

[2] W. May, et. al, "A Unified Framework for Wrapping, Mediating and Restructuring Information from the Web," *Int'l. Workshop on the World-Wide Web and Conceptual Modeling (WWWCM'99)*, LNCS 1727, pp. 307-320, 1999.

[3] W. May and G. Lausen, "Information Extraction from the Web," *Technical Report 136*, Institut fur Informatik, Universitat Freiburg, 2000.

[4] *http://www.w3.org/XML*

[5] H. Ouahid, A. Karmouch, "Converting Web Pages into Well-formed XML Documents," *IEEE Int'l. Conf. on Communications*, Vol.1, pp. 676~680, 1999.

[6] *http://www.ncdesign.org/html/list.htm*