<a href> 태그 추출을 이용한 웹 문서 구조화

박은주, 임한규
안동대학교 컴퓨터공학과

Web site construction using <a href> tag extraction

Eunju-Park, Hankyu Lim
Dept. of Computer Engineering, Andong National University

요 약

하이퍼텍스트 구조를 가진 웹문서들은 HTML 태그를 사용하여 문서들을 링크한다. 웹사이트가 점점 더 커짐에 따라 새로운 웹사이트를 디자인하고, 일일이 웹 페이지들을 만들고 그들간의 링크를 만들어 다시 갱신하는 작업은 몹시도 지루한 작업이 되었다. 웹문서를 그래프로 표현함으로써, 웹사용자들이 요청한 자료로의 빠른 접근을 가능하게 해 주기 위하여 페이지의 조직을 동적으로 바꾸거나 문서들간의 링크의 구조를 바꾸어 줄 경우 웹사이트 관리를 편리하게 해주는 것이 가능하고, 자동적으로 웹문서들을 재조정하는 것이 가능하다.

본 논문에서는, HTML 문서에서 문서를 연결해주는 태그인 <a href="http://파일명">으로 구성된 태그들을 추출하여, 이 태그들을 방향을 가진 링크로 간주하여 웹 페이지들을 그래프로 표현한다. 이 그래프 구조에서 링크를 재조정함으로써 사이트들에서 페이지의 조직을 동적으로 구성하고, 사용자들에게 편리함을 제공하는 링크를 제시한다.

1. 서론

현재의 웹은 많은 페이지들에서 원하는 정보를 얻기 위하여 정보들을 특정 영역이나 관심 분야로 분류하고, 요약하고 정형화하여 데이터베이스로 만들어 추출(extraction)함으로써 이 문제를 해결하고 있다.[1]

하이퍼텍스트는 비순차적 문서이며, 각각의 노드가 일정량의 텍스트나 다른 자료를 포함하고 있는 방향성 그래프라 할 수 있다. 웹을 노드와 링크를 가진 그래프 구조로 표현하고 링크들을 재조정하여 그래프 구조 자체를 변화시킴으로써 웹 사용자에게 편리함(접속 빈도 증가, 접속 시간의 증가)을 제공하는 것이 가능하다.[2]

본 논문은 웹 문서를 그래프로 표현하기 위하여 HTML 문서의 태그들 중 문서들을 연결하는 태그인 <a href="http://파일명">인 태그들을 추출하여 이 태그들을 그래프에서 방향성을 가진 링크로 간주하여 그래프 구조로 표현한다. 구조화된 그래프 구조에서 그래프의 링크들을 재조정함으로써 웹 사이트내에서 페이지들의 위치를 바꾼다.

웹문서의 재구조화는 그래프 구조로 이루어진 웹사이트에서 링크들만을 재조정함으로써 가능하고, 바뀌는 것은 구조내에서의 페이지들의 위치이다.

2장에서는 웹이 가지는 하이퍼텍스트 구조와 <a href>태그를 추출하고, 추출한 태그들을 그래프로 표현하기 위하여 관련된 연구들을 살펴보며, 3장에서는 그래프 형태로 구현되어진 웹 문서들의 링크를 재조정하는 방법에 대하여 알아본다. 4장에서는 웹문서를 읽어들여 <a href> 태그를 추출하고, 추출된 태그들을 그래프로 표현하는 방법에 대하여 알아보고 구현한다. 5장에서 결론을 맺고 향후 연구과제에 대하여 기술한다.

2. Information Visualization

2.1 하이퍼텍스트

사실상 웹은 그림 1에서 보여주는 것처럼 파일 시스템의 전통구조와 하이퍼텍스트의 복잡한 연결구조인 2개의 환경들 내에서 존재한다. 이 2개의 환경들 내에서 리소스들 사이의 관계들뿐만 아니라 리소스들에서 인터페이스들도 기본적으로 다르다.[3]

하이퍼텍스트는 비순차적 문서이며, 각각의 노드가 일정량의 텍스트나 다른 자료를 포함하고 있는 방향 그래프(Directed graph)라고 할 수 있다. 하이퍼텍스트 문서의 기본적인 정보단위는 노드라고 불리고, 노드는 링크라고 불리는 상호참조기구(cross-reference)에 의하여 연결된다. 특별한 주제에 관해 더 많은 정보를 얻기 위해서 사용자는 노드 중에 존재하는 앵커를 활성화시킨다. 앵커는 노드 내부의 링크의 존재를

나타내는 노드의 일부분이다. 앵커가 활성화되면 컴퓨터 시스템은 즉각 현재 노드를 링크의 목적 노드로 대치하여 화면을 변화시킨다.[4]

2.2 Information Visualization

정보 가시화(information visualization)의 목적은 복잡하고 차원이 많은 정보 데이터(information data)를 이해하기 쉽게 그림이나 도표와 같은 특정한 형식을 이용하여 효과적으로 나타내고 비교하는데 있다. 그동안 제시되어온 다차원 정보 가시화 기법의 대표적인 것으로는 Scatterplots, Perspective Wall, Parallel Coordinates, Graph를 들 수 있다.[5]

이 분야의 초기연구들은 입자의 운동이나 공기의 흐름 같은 물리적 현상들을 시각화하는데에 치중했으나, 최근의 연구들은 데이터베이스나 소프트웨어 구조 같이 훨씬 추상적인 정보들을 시각화하는 데에 초점이 맞춰지고 있다. 이들 구조들은 단지 메모리나 디스크 안에 존재하는 정보일 뿐이지 실제 형태를 가지고 있는 것이 아니다. 이들 구조들에 어떤 형태를 주기 위한 연구와 사람들이 이들 정보구조를 쉽게 접근하고 질의 할 수 있도록 해주는 환경을 만드는 연구들이 많이 진행되고 있다.[6]

2.3 인터넷 정보 추출(Information Extraction)

정보추출은 한 문서에서 그 문서의 중심적 의미를 나타내는 특정 구성요소를 인식하여 추출하는 작업을 의미한다. 정보추출의 예로는 날씨 정보를 제공하는 웹 문서로부터 지역, 날짜, 최고온도, 최저온도, 습도 등의 정보를 뽑아 내거나, 아파트 정보 문서로부터 방의 개수, 매매가, 전세가, 전화번호 등을 추출하는 것을 들 수 있다.

정보추출 에이전트(information extraction agent)는 인터넷 문서에서 원하는 부분 텍스트 정보를 추출해 내는 작업을 수행하며 wrapper라 불리는 추출 규칙을 각 정보소스에 대하여 생성하여야 한다.

우리가 관심을 가지고 있는 정보추출 작업은 인터넷 페이지이며, 대부분의 추출 항목은 시각적으로 구별되는 특성을 가지고 있다.[6]

3. 링크 재조정에 의한 웹문서 재구조화

웹사이트에서 문서의 연결은 <a href> 태그를 사용한다. 링크된 문서들은 문서간 링크를 다시 구성하고자 할 경우 <a>태그를 사용하여 링크설정을 새로이 구성해 주어야만 한다.

그림 2는 본 논문의 전체적인 구조를 보여주고 있다. 구조화하고자 하는 웹사이트를 읽어들인 다음 웹문서들간의 링크관계를 그래프로 표현한다. 이 때 링크의 생성은 HTML 문서에서 문서와 문서를 연결해주는 태그인 <a href>태그의 추출을 사용한다. 그래프로 표현된 웹사이트는 링크들을 재조정하는 과정을 거치므로 웹사이트의 문서들은 모두 동일한 문서들이지만 새로운 연결구조를 가진 웹사이트가 된다.

본 논문에서는 HTML 태그 중 문서들간을 연결해주는 <a href>태그를 이용하여 상하계층을 만들어냄으로써 그래프 형태의 데이터 구조를 만드는 것이 가능하다. 웹 문서를 그래프 형태로 만들면, 데이터에 대한 접근방법이 한 문서에서 처음부터 매번 그 데이터를 찾아가서 문자열을 잘라내고 붙이고 하는 것보다 쉽다.[4]

HTML 태그들로 구성되어진 웹문서는 링크의 생성, 삭제시 <a href> 태그를 사용하여 HTML 문서를 새로 구성해 주어야만 한다. 즉, 웹문서를 재구조화할 경우 수동적인 방법에만 의존한다. 그러나 수동적인 방법은 작은 양의 웹문서들을 관리하는 경우에는 크게 문제가 되지 않으나 많은 양의 웹문서를 가지는 경우에는 링크의 생성, 삭제시 너무나 많은 시간과 노력을 필요로 하게 된다.

그래프로 구조화된 웹문서들은 DB에 저장되어 있는 정보를 사용하여 자동적으로 재구조화하는 것이 가능하다. DB에 저장되어 있는 정보는 상위노드에 대한 링크정보를 가지게 되므로 이 정보를 추가하거나 삭제하는 것으로 웹문서에서 특정한 문서로의 연결이 추가되거나 삭제되어진다.

링크를 재조정하여 웹사이트를 재구조화하였을 경우, 변화되어진 웹문서의 구조에 따라 HTML 문서의 태그들 또한 변화되어져야 한다.

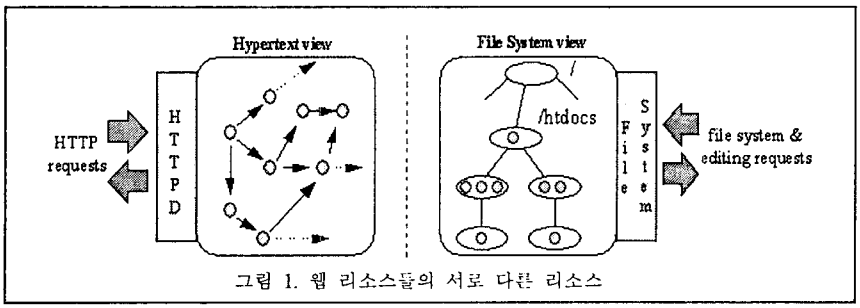

그림 1. 웹 리소스들의 서로 다른 리소스

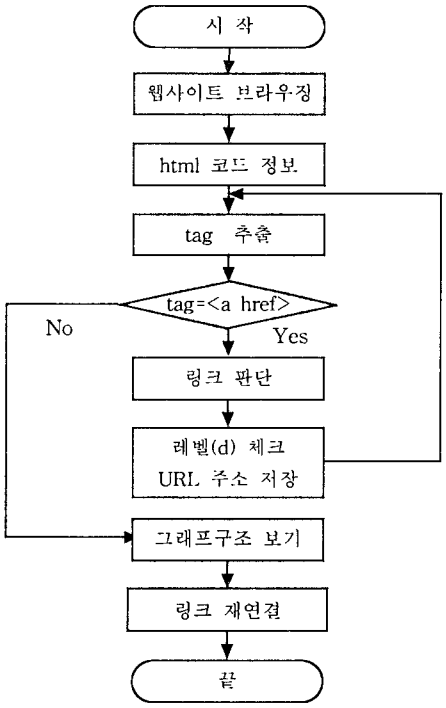

그림 2. 웹사이트 재구조화 처리절차

4. <a href> 태그 추출을 이용한 웹사이트 그래프 표현 구현

일반적인 웹사이트의 시각화는 인터넷상에 존재하는 웹 문서를 보다 시각적이면서 체계적으로 관리할 수 있게 하는 시스템이다. 즉, 웹 문서의 구조를 시각적으로 사용자에게 보여줌으로써, 불필요한 문서 읽기를 하지 않고도 원하는 사이트(site)로 직접 이동할 수 있게 하거나 문서 체계를 쉽게파악할 수 있게 한다.[7]

4.1 <a href="http://파일명"> 태그 추출 구현

본 논문에서 제안하는 웹사이트 그래프화 구현환경은 펜티엄 PC와 MS WINDOWS NT에서 VB 6.0을 사용하였다.

본 논문에서 기술하는 인터넷 정보추출 기법의 대상이 되는 것은 주로 HTML로 작성된 텍스트 형태의 문서이다. <h1>이나 <p>와 같은 대부분의 HTML 태그들은 의미를 가지기보다는 화면에 출력되는 포맷을 지정하기 위한 태그로 정보 추출에 도움이 되지 못한다. [8]

웹사이트를 그래프로 표현하기 위해 HTML 태그들 중에서 시각적으로도 구분이 가능하고, 문서와 문서를 연결해주는 태그인 <a href = "http://파일명.html">으로 시작되는 태그를 추출의 대상으로 삼는다.

본 논문은 한 웹사이트의 링크 재조정을 위해 같은 컴퓨터 안에서의 문서 연결만을 다루므로, 다른 컴퓨터의 파일과의 연결은 DB에 저장하지 않는다.

a.html문서에서 <a href="http://b.html">태그를 만난다면 문서 a에서 b로의 링크가 생성한다. 웹사이트의 모든 문서에 대해서 <a href>태그를 읽어들여 링크를 생성하고, DB에 파일명을 저장한다.

4.2 그래프 표현 구현

웹 문서를 그래프 구현시 깊이 d를 고려한다.

d: 홈페이지의 깊이(홈페이지로부터 얼마나 많은 단계를 가지는가)

그래프 표현시 홈페이지로부터의 깊이 d를 고려하여, 깊이 d를 가지는 페이지들은 깊이 d-1을 가지는 페이지의 하위레벨에 위치하여야 한다.

그림 3에서 보는 바와 같이 웹사이트를 그래프로 표현시 Node 4(노드 id = 4)는 노드 2와 노드 3 양쪽과 연결되어 있다. 노드 2와 1은 노드 3과 같거나 낮은 d를 가지므로 노드 3의 부모 노드가 될 수 있으나, 노드 4는 노드 3보다 낮은 d를 가지므로 노드 3의 부모노드로 간주되어지지 않는다. 그러므로 노드 3과 4 간의 링크는 의미를 가지지 않게 된다.

본 논문에서 웹사이트를 그래프로 표현하는 것은 링크 재구성을 하기 위함이므로, 의미를 가지지 않는

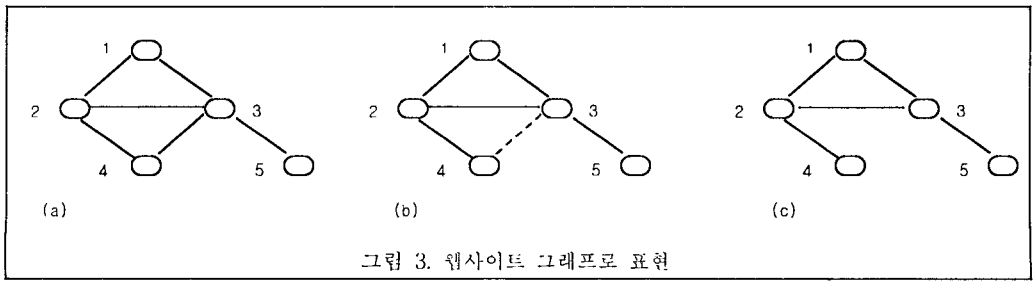

그림 3. 웹사이트 그래프로 표현

아래 태그에 대한 링크 생성은 하지 않는다.

1. 한 문서에서 다른 문서로의 중복 링크생성은 하지 않는다
2. 문서내 연결은 링크를 생성하지 않는다
3. 문서들에 대한 링크생성순서는 DFS 탐색 알고리즘을 적용한다.

각 문서들을 읽어들여 링크를 생성할 때 링크가 추가되는 순서는 그림 4와 같이 깊이우선탐색(depth-first search, DFS)을 적용한다.

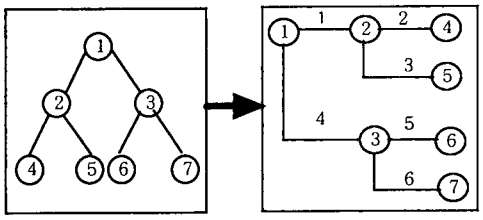

그림 4. 링크 생성 순서

4.3 구현결과

그림 5는 특정 웹문서를 읽어들였을 때 문서의 HTML 코드를 볼 수 있도록 구현한 화면이다. 그림 6은 웹사이트를 그래프로 구조화한 화면이다. 탐색기 형태로 구성되었으며, ⊞를 클릭할 경우 연결된 하위 문서들에 대한 정보를 얻을 수 있다.

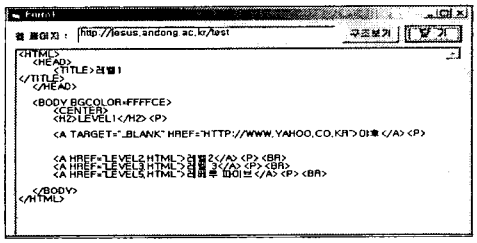

그림 5. 선택한 페이지에서 html 태그를 보여주는 화면

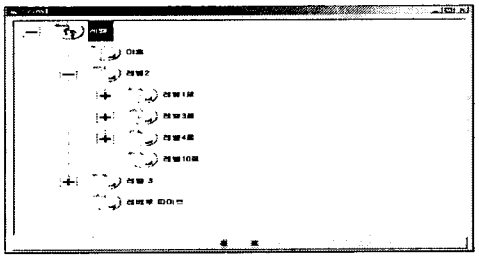

그림 6. 웹문서를 그래프로 구현한 화면

5. 결론

웹문서는 HTML 태그들로 구성되고 문서를 링크시키고자 할 경우 수동적인 방법으로 <a href> 태그를 사용하여 연결시켜주어야 한다. 물론 구조화된 웹사이트들에서 새로운 링크를 생성하거나 링크의 삭제를 원하는 경우에도 여전히 수동적으로 <a href> 태그로 링크시켜야만 한다. 많은 양의 웹문서들을 관리해야 하는 경우 문서들을 연결하기 위한 수동적인 노력들은 상당히 많은 시간을 요구하게 된다.

본 논문은 웹 문서들을 링크와 노드를 가진 방향성 그래프로 정의하고, 문서를 구성하는 HTML 태그들 중 문서를 연결하는 <a href>태그를 링크로 간주하여 웹문서를 그래프로 구조화하였다. 웹문서를 그래프로 표현하기 위하여 HTML문서에서 <a href="http://파일명">으로 구성된 태그들만을 추출하여 DB에 저장하고, 이 정보를 사용하여 그래프로 구현하였다.

웹문서를 그래프로 표현함으로써, 웹사용자들이 요청한 자료로의 빠른 접근을 가능하게 해 주기 위하여 페이지의 조직을 동적으로 바꾸거나, 문서들간의 링크의 구조를 바꾸어 주는 등의 웹사이트 관리를 편리하게 해주는 것이 가능하고, 특히 자동적으로 웹문서들을 재조정하는것이 가능하다.

향후 웹사이트를 재구조화했을 경우 HTML 문서내에서의 tag들을 자동적으로 삽입, 삭제하는 방법과 사용자가 원하는 정보를 얻기 위한 시간을 단축시킬 수 있는 웹사이트 재구조화 방법에 대한 연구가 지속되어야 할 것이다.

[참고문헌]

[1] 조민재, 황수철, 김기태, "웹 문서의 개념 지식과 그래프 구조를 이용한 정보 분류 및 추출 시스템", 한국정보과학회 추계학술 발표대회, 1999.

[2] John Garaofalakis, "Web site optimization using page popularity", IEEE INTERNET COMPUTING, JULY.AUGUST 1999.

[3] D.Ingham et al., "W3Object: Bringing Object-Oriented Technology on the Web," The Web., Vol.1, No.1,pp.89-105.

[4] 전경현, "가상경로 정보를 이용한 하이퍼텍스트 링크 생성 모델", 한국과학기술원 석사학위논문, 1993.

[5] http://www.cwi.nl/InfoVisu/General.html

[6] http://gaston.snu.ac.kr/webtech/vrml/part3/no-te 1.html

[7] 김성운, 김성진, 강현석, "웹 문서 분석/검색 시스템의 설계 및 구현", 한국정보처리학회 춘계 학술발표, 1999.

[8] 최중문, "인터넷 정보 추출 에이전트", 정보과학회지 제18권 제5호, 2000.5.