웹 사이트의 목차 디렉토리 생성 에이전트

유남현○ 양현택 김원중
순 천 대 학 교 컴 퓨 터 과 학 과
{hyun43, yht, kwj}@cs.sunchon.ac.kr

The Agent for Construction of The Directory of Contents on Websites

Nam-Hyun Yoo○ Hyun-Teak Yang Won-Jung Kim
Dept. of Computer Science, Sunchon National University

요 약

현재 대부분의 사람들은 자신의 원하는 정보를 찾기 위하여 먼저 사이버공간을 이용한다. 웹 브라우저를 이용하여 망망대해와 같은 거대한 정보의 바다인 인터넷을 항해하는 것이다. 그러나 인터넷에 등록되는 정보의 양은 기하급수적으로 증가하고 있어, 우리가 원하는 정보를 짧은 시간에 찾기가 점점 어려워지고 있다. 따라서 사용자가 웹에서 정보를 탐색할 때, 웹 사이트에서 여러 단계의 복잡한 계층구조들을 탐색하는데 많은 시간을 소비하지 않고, 본인이 관심과 흥미를 가지고 있는 유용한 정보를 쉽게 선택할 수 있는 방법이 필요하다. 본 논문에서는 사용자가 검색하고자 하는 웹사이트의 계층적 목차 디렉토리(Hierarchical Directory of Contents)를 자동으로 생성하는 에이전트(HDOCA)에 관해 연구하였다.

1. 서론

우리가 어떤 미디어에 정보를 표현하여 저장한다는 것은 미래의 사용자가 있다는 것을 전제로 한다. 따라서 정보를 표현할 때는 그것을 필요로 하는 사용자가 전체 정보의 윤곽을 정확히 파악할 수 있도록 하며, 또한 저장된 정보의 어느 부분이라도 손쉽게 접근하여 획득할 수 있도록 하여야 한다.

90년대 초에 시작된 웹의 등장은 정보검색 환경에 일대 변혁을 가져왔다. 멀티미디어 환경에서 하이퍼링크(Hyper Link)를 이용한 웹의 정보검색 방법은 오늘날의 인터넷 혁명을 가져오는데 가장 큰 기여를 하였다. 그러나 인터넷에 등록되는 정보의 양이 기하급수적으로 증가하고 있어, 우리가 원하는 정보를 짧은 시간에 찾아내기가 점점 더 어려워지고 있다. 즉, 아이러니 하게도 웹에 제공되는 정보의 양이 많아지면 많아질수록 필요로 하는 정보를 획득하기가 어려워지고 있는 것이다. 그 원인 중에 하나는 오늘날의 인터넷 성공을 가져온 하이퍼텍스트 기법의 정보탐색 기법이 대량의 정보집단에서 정보탐색 요구를 만족시키는데는 많은 제약점을 가지고 있기 때문이다.

본 연구는 1999년도 순천대학교 공모과제 학술연구비에 의하여 연구되었음

정보검색을 위하여 웹 사이트에서 하이퍼링크를 따라 여러 단계의 복잡한 계층구조들을 탐색하는데는 많은 시간과 노력이 필요하다[1,3].

따라서 이러한 하이퍼텍스트 기법의 제약점을 보완할 수 있는 정보표현 방법이 필요하다. 목차(The Directory of Contents)는 책(Book)에서 처음 사용된 방법으로 책의 전체 내용을 파악할 수 있도록 하고, 페이지를 함께 표시하여 원하는 부분에 곧 바로 접근할 수 있도록 도와주는 정보표현 방법에 있어서 가장 일반적인 방법이다. 그러나 하이퍼링크 기법을 사용하는 웹 사이트들의 대부분은 책의 목차와 같이 전체적인 윤곽을 보여 주지 못하고, 특정의 문서에 직접적으로 접근할 수 있는 편의성을 제공하지 못하고 있다. 물론, 원하는 세부적인 정보의 정확한 URL 주소를 알고 직접 접근할 수도 있지만, 매우 긴 URL 주소를 기억하여 사용한다는 것은 하이퍼링크의 계층구조를 따라 가는 것보다 훨씬 어려울 것이다.

본 논문에서는 사용자가 검색하고자 하는 웹사이트의 계층적 목차 디렉토리를 자동으로 생성하는 에이전트(HDOCA: Hierarchical Directory Of Contents construction Agent)에 관해 연구하였다. 자바 프로그램과 애플릿을 사용하여 구현된 목차 디렉토리 생성 에이전트는 탐색하고자 하는 웹사이트의 내용을 확대/축소

(Expand/Contract)가 가능한 일반적인 파일 디렉토리 구조와 동일한 형태로 표현한다. 따라서 사용자는 손쉽게 웹 사이트의 콘텐츠 내용을 전체적으로 파악할 수 있고, 원하는 정보를 복잡한 하이퍼텍스트 링크의 구조를 따라가지 않고서도 찾아낼 수 있어 정보검색에 소요되는 시간과 노력을 대폭 감소시킬 수 있다.

2. 유용한 정보표현 기법

정보를 이용하는 사용자의 효율적인 정보검색과 보다 빠른 정보접근을 위한 많은 유용한 정보표현 기법들이 존재한다. 그것은 책의 목차처럼 매우 오래 전부터 사용되어 오고 있는 것도 있고, 하이퍼텍스트 기법처럼 최근에 사용되기 시작한 것도 있다. 그러나 유용한 정보표현 기술 중에는 그것의 사용이 미디어의 특성에 좌우되기도 한다. 예를 들어, 전자적인 마우스 클릭 기능이 적용될 수 없는 책에는 하이퍼텍스트 기술의 사용이 제한적일 수 밖에 없다. 이러한 정보표현 기법에는 책에 사용된 기법들, 하이퍼텍스트, 신문에 사용된 기법들, 인용 기법, 재사용 기법, 결합 기법, 인덱스 기법, 질의 기법, 직접/간접/멀티 디스플레이 기법 등이 있다[2, 4].

본 논문과 관련된 정보표현 기법들은 다음과 같다.

① 책에 사용된 목차, 인덱스, 페이지 번호, 장, 절 등의 개념은 가장 오래되고 가장 진보된 정보표현 기술 중의 하나이다. 또한 우리에게 가장 친밀하며, 어떤 미디어를 이용하여 정보를 표현하든지 사용되어야 할 기법들이다.

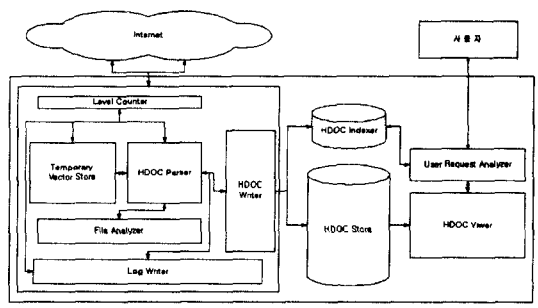
② 하이퍼텍스트 표현 기법은 오늘날의 인터넷 혁명을 이끈 가장 중요한 기술이며 앞으로도 없어서는 안될 기법이다. 그러나 하이퍼텍스트는 저자 인텐시브(Intensive)한 경향이 있으며, 오늘날의 웹과 같은 대량의 정보집단에서 정보탐색요구를 만족시키는데는 많은 제약점을 가지고 있다. 따라서 책의 목차 등과 같은 보완 기법이 필요하다.

③ 신문은 동시에 많은 기사를 함께 표현하여 짧은 시간에 많은 기사를 읽어야 하는 바쁜 독자들을 위한 훌륭한 표현기법을 가지고 있다. 또한, 현재의 인터넷 정보검색에서 해결하여야 할 핵심기술로 부상하고 있는 정보분류(Information Clustering)를 잘 표현하고 있다. 즉, 우리는 신문이 머리기사, 정치, 경제, 문화 관련 기사를 어떤 부분에 어떻게 게재하는지를 잘 알고 있다.

(4) 인용은 새로운 어떤 것을 만들어 낼 때, 공식적으로 인정되고 있는 어떤 기초에 근거하기 위한 메카니즘으로 학문의 세계에서 오랫동안 중요한 역할을 하였으며, 현재 웹에서의 링크는 사이버스페이스에서의 인용이라고 할 수 있다.

3. 구현

본 논문에서 연구한 계층적 목차 디렉토리 생성 에이전트(HDOCA)의 시스템 구조는 [그림1]과 같다. 시스템은 자바 프로그램과 애플릿을 사용하여, 인터넷 익스플로러 5.x와 윈도우2000 환경에서 개발하고 있다. 시스템은 HDOC Viewer, HDOC Parser, HDOC Writer , File Analyzer, Temporary Vector Store 등으로 구성된다. HDOC Parser는 웹 로봇처럼 사이트를 항해(Navigation)하면서 해당 사이트의 정보를 HDOC Store 및 HDOC Indexer에 저장 하며, HDOC Viewer는 사용자의 요구를 받아 HDOC Indexer를 검색한 후, 자료가 존재하는 경우 HDOC Store에 저장되어 있는 해당 사이트의 정보를 사용자에게 디스플레이 하여 준다.



[그림 1] HDOCA 시스템 구조

(1) HDOC Parser

HDOC Parser는 웹 사이트를 항해하면서 사이트의 계층적 정보와 HTML 문서들의 정보를 추출하여 저장한다. 기존 검색 엔진은 내용을 대상으로 검색하여 정보를 추출하지만 HDOC Parser는 웹 사이트의 루트 파일이 위치한 디렉토리를 기준으로 하여 디렉토리 구조를 파악한 후, HTML 문서들의 링크를 추적하여 디렉토리 구조와 매치 시킨다. 그리고 각 파일들과 연관된 파일들의 정보를 추출하는 방법으로 문서의 내용 정보 및 계층적 디렉토리 구조를 추출한다. HDOC Parser가 실행되는 순서는 다음과 같다.

① 웹 사이트의 루트 파일이 위치한 디렉토리를 기준으로 하여 루트파일의 연결된 링크를 너비 우선 탐색(Breadth First Search) 기법으로 디렉토리 구조를 추출하여 계층구조의 높이와 함께 저장한다.

② 파일내의 링크들을 추출, 각 파일들을 이미 저장해 놓은 디렉토리 구조와 연결 구조를 생성한다. 연결 링크가 중복된 경우 디렉토리 구조를 기준으로 비교 분석하여 저장하거나, 또는 삭제하여 파일들의 구조를 정리한다. 연결된 링크가 아닌 디렉토리의 직접적인 접근은 사용하지 않았다.

③ HTML 파일에 연결된 이미지, 오디오, 비디오 등과 같은 기타 파일들은 타입정보와, 파일 이름, 사이즈 정보들을 해당 파일에 포함시킨다.

(2) HDOC Writer

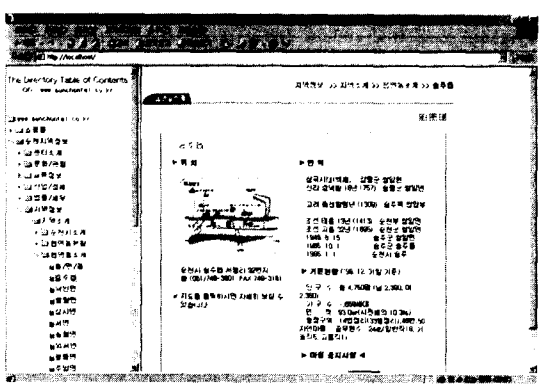
HDOC Writer는 HDOC Parser에서 넘겨준 정보를 HDOC Viewer가 디스플레이 할 수 있도록, HDOC Indexer와 HDOC Store에 [표1]의 형태로 저장한다.

다. <HDOC>와 </HDOC>는 문서의 처음과 끝을 나타

```
1: <HDOC>
2: <DN>
3: <BD NAME=name LEVEL=level>
4: <F NAME=name, TYPE=type, SIZE=size(k)>
5:       <COMMENT>comment</COMMENT>
6:       <LF NAME=name, TYPE=type, SIZE=size(k)>
7:       <LF NAME=name, TYPE=type, SIZE=size(k)>
8: </F>
9: </DN>
10: </HDOC>
```

[표 1] HDOC 파일 구조

내며, <DN>은 해당되는 디렉토리의 이름을 나타낸다. <BD>는 Base Directory의 약어로서 기준 디렉토리의 이름과 단계(Level)을 나타내며, 최상위는 '0' 이다. <F>는 파일의 정보를 표현하는 것으로서 NAME은 실제 파일의 이름, TYPE은 파일의 형태를 나타낸다. TYPE에서 H(HTML or Text), I(Image), M(Audio /Video), E(Execute File or CGI), O(Undefine Type), L(Link)를 나타낸다. TYPE이 링크인 경우 파일 사이즈와 형태는 'null'로 정의되며, NAME에는 해당 URL이 표시된다. SIZE는 그 파일의 크기를 나타내며 단위는 KB이다. <COMMENT>는 HTML문서의 <TITLE>이나 본문의 일부를 추출한 것이다. HDOC Indexer에는 연관된 주제어와 검색된 URL이 저장된다.



[그림 2] HDOCA 구현 화면

(3) HDOC Viwer

사용자가 질의한 내용을 분석하여 HDOC Indexer에서, 해당 정보가 존재하는 경우 HDOC Store에서 정보를 추출하여 사용자에게 보여주는 애플릿이다. [그림 2]는 시제품(Prototype)으로 구현된 결과 화면으로서 왼쪽에는 검색하고자 하는 웹 사이트의 정보내용을 계층적 목차 디렉토리를 보여주는 HDOC Viewer이며, 검색된 결과를 클릭하면 오른쪽에 해당 HTML 문서가 디스플레이 된다. ㅗ,ㄷ는 Expand/Contract를 나타내며, H는 해당 정보가 HTML 문서임을 나타낸다. 또한 화면에는 나타나지 않았지만, I는 이미지, M은 오디오/비디오 파일, E는 실행파일이나 CGI 파일, O는 정의되지 않은 형식의 파일, L은 외부 서버로 연결된 링크를 나타낸다.

4. 결론

본 논문에서는 사용자가 웹에서 정보를 검색할 때, 하이퍼텍스트 링크를 계속적으로 따라 가면서 소비하는 많은 시간과 노력을 감소시키기 위하여 계층적 목차 디렉토리 생성 에이전트에 대해 연구하였다.

사용자가 계층적 목차 디렉토리를 사용하면, 웹 사이트 전체 정보에 대한 전체적인 윤곽을 쉽게 파악할 수 있고, 원하는 정보를 복잡한 하이퍼텍스트 링크의 구조를 따라가지 않고서도 빠르게 접근할 수 있다. 예를들어 [그림2]의 웹 사이트에서 사용자가 승주읍의 인구를 알고자 한다면, 기존의 브라우저에서는 www.sunchontel.co.kr의 첫 페이지- 순천지역정보 - 지역정보 - 지역소개 - 읍면동 소개 - 승주읍의 순서로 하이퍼텍스트의 링크를 계속 따라 가야만 한다. 그러나 계층적 목차 디렉토리를 사용하면 컴퓨터에 대해 약간의 지식만 가지고 있는 사용자라면 단 한번의 마우스 클릭만으로 그 정보를 찾을 수 있을 것이다.

앞으로의 연구과제는 디렉토리 구조뿐만 아니라 요약(Summarize) 및 분류(Clustering) 정보도 함께 표현 할 수 있도록 시스템을 확장하고, 온라인상에서 실시간으로 수행될 수 있도록 하는 것이다.

5. 참고 문헌

[1] E. Kandogan and B. Shneiderman, "Elastic Windows: A Hierarchical Multi-Window World-Wide Web Browser," Proceedings of the 10th annual ACM Symposium on User Interface Software and Technology, 1997.

[2] Golovchinsky, G and M.H. Chignell, "The Newspaper as an Information Exploration Metaphore," Information Processing & Management 33 (5), 1997.

[3] Nation, D., Plaisant, C., Marchionini, G., Komlodi, A., "Visualizing Websites using a Hierarchical Table of Contents Browser: WebTOC," University of Maryland Technical Report, 1997.

[4] 양현택, 박나연, 김원중, "HCI를 위한 다중 디스플레이와 웹 정보 검색", 한국정보과학회 '00춘계학술발표논문집, 제 27권1호, 2000.