문서의 효율적인 검색을 위한 HTML 문서 변환 시스템

김수희, 장대용
호서대학교 컴퓨터공학부
shkim@office.hoseo.ac.kr

HTML Document Conversion System for Effective Retrieval of Text Document

Su-Hee Kim, Dae-Yong Chang
Division of Computer Engineering, Hoseo University

요 약

이 연구에서는 텍스트 문서를 웹에서 HTML 문서 형태로 효율적으로 검색할 수 있는 변환 시스템을 개발하였다. 웹상에 사용자가 원하는 부분만을 HTML 문서 형태로 제공하도록 문서의 논리적인 구조를 파악하며 그 구조에 대한 정보와 각 논리 단위에 해당하는 부분의 범위 정보를 저장할 수 있도록 관계형 데이터베이스 스키마를 개발하였다.

개발한 시스템은 문서의 목차 테이블을 자동으로 구축하고 목차 테이블의 각 항목에 하이퍼링크를 설정한다. 문서를 웹에서 검색하기 위한 첫 화면은 목차 테이블이며, 그 중 한 항목이 클릭되면 그 항목의 내용이 제공되고, 만약 그 하위에 속하는 항목들이 있다면 그들에 대한 링크를 역시 제공한다. 이러한 방법으로 한 문서의 전체를 그 논리 구조에 따라 사용자가 원하는 대로 검색할 수 있다.

이 시스템은 멀티미디어 타입의 문서를 하이퍼미디어 문서 형식으로 변환할 수 있도록 확장하여 보완 개발될 수 있고, 장래에 전자 출판과 전자 도서관에 응용될 수 있다.

1. 서론

인터넷의 급격한 발전으로 일상 생활에서부터 전문적인 연구에 이르기까지 필요한 정보를 찾기 위해 많은 사람들이 인터넷을 이용하는 추세에 있다.

문서를 웹상에서 검색하기 위해서는 브라우저가 인식할 수 있는 문서 형태를 갖추어야 한다. 그 대표적인 문서 형태가 HTML 형식의 문서이다. 문서의 크기가 크고 그 구조가 간단하지 않을 때는 이를 웹에서 검색이 가능하도록 HTML 형태의 문서로 변환하기가 용이하지 않다. 그리고 크기가 큰 문서 전체를 웹에 올려 검색하는 경우에는 그 검색속도가 매우 느려질 가능성이 있다.

이 연구에서는 텍스트 문서를 웹에서 효율적으로 검색하기 위해 문서 전체를 로딩하지 않고, 사용자가 원하는 부분만을 HTML 문서 형태로 제공하는 시스템을 개발하고자 한다. 이러한 변환 시스템을 개발하기 위해, 문서의 논리적인 트리 구조를 분석하고 파악하며 그 구조에 대한 정보와 각 논리 단위에 해당하는 부분의 범위 정보를 저장하기 위해 관계형 데이터베이스 스키마를 고안하여 활용하고자 한다.

2. 문서의 구조

2.1 문서의 트리 구조

우리가 사용하고 있는 문서의 물리적인 구조 형태는 여러 가지가 존재하고 있다. 하지만, 문서의 논리적인 구조는 계층구조를 이루고 있다. 교과서나 이와 유사한 부류에 속하는 책은 공통적으로 다음의 형식을 따르고 있다. 가장 상위 단계에 목차가 있고, 그 다음으로 장(chapter), 그 하부 단계로는 절(section)이 있다. 또한 한 절의 하위단계에 그 하위 절(subsection)이 오는 다단계 구조가 존재한다. 이를 트리 구조로 나타내면 다음과 같다.

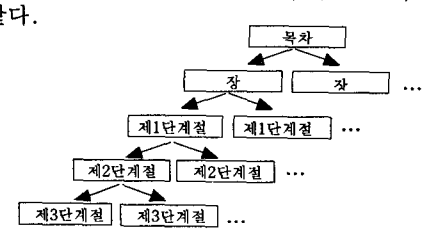

그림 1. 문서의 논리구조

2.2 문서에서 장을 표현하는 유형

문서에서 장을 나타내는 형태는 여러 가지가 있다. 이는 사용하는 언어 문화권마다 제각기 다른 형태를 나타내고 있다. 보편적으로 많이 쓰여지고 있는 장의 형태는 대체로 다음과 같다.

● 일반적인 장의 유형
- 제 1장, 제 2장, 제 3장 -
- 1장, 2장, 3장 -
- 1., 2., 3. -
- Chapter 1, Chapter 2, Chapter 3 -
- Chap 1, Chap 2, Chap 3 -

지금까지 교과서나 이와 유사한 부류에 속하는 책에서 많이 사용하는 장의 표현에 대해 살펴보았다. 문서를 하이퍼텍스트 문서로 변환하기 위해서는 이러한 장의 일반적인 표현 형태에서 어떤 규칙을 발견하여 문서 내부에서 특정 문장이 어떤 장의 제목이라는 것을 인식하는 것이 필요하다. 지금까지 살펴본 일반적인 장의 유형의 예들을 다음과 같이 정규 표현식으로 나타낼 수 있다[1].

```
Digit0 ::= '0' |' 1' |' 2' |' 3' |' 4' | '5' |' 6' |'7' |' 8' |' 9'
Digit1 ::= '1' |' 2' |' 3' |' 4' |' 5' | '6' |' 7' |' 8' |' 9'
Chap_Expression ::= (' ')*('제')(' ')*Digit1 Digit0 (' ')*'장'(' ')*
   |(' ')*('Chapter'|'Chap') (' ')*Digit1 Digit0* (' ')*
   |(' ')* Digit1 Digit0* '.' (' ')*
```

여기서 ε 는 빈 문자열(empty string)이다.

2.3 문서에서 절을 표현하는 유형

앞서 살펴본 장의 표현과 마찬가지로 문서에서 절을 나타내는 형태도 여러 가지가 있다. 그리고 절 내부에 절이 있는 이른바 다단계의 하위 절이 존재할 수가 있다. 다단계의 하위 절에 대한 유형은 다음에 다루기로 하고, 일반적으로 많이 사용되는 단일 단계의 절을 표현하는 유형은 다음과 같다.

● 일반적인 절의 유형의 예
- (1), (2), (3), (4), (5)-
- 1), 2), 3), 4), 5)-

이와 같은 절의 유형은 다음과 같은 간단한 정규 표현식으로 나타낼 수 있다.

```
Start_Sec ::= ' (' |ε          End_Sec ::= ' )'
Sec_Expression ::=
    Start_Sec(' ')* Digit1 Digit0*(' ')* End_Sec(' ')*
```

2.4 문서에서 다단계의 절을 표현하는 유형

일반적으로 문서는 장과 한 단계의 절만으로 이루어져 있지는 않다. 사람들이 표현하고자 하는 내용이 점차 많아지면서 그 내용에 내재하는 논리 구조가 복잡해지고 또한 각 구조의 체계적인 유지를 위해 절 안에 절이 있는 다단계의 절이 존재하게 된다. 이러한 다단계의 절의 표현도 각 문화권마다 다소 다르지만 일반적으로 많이 사용되는 표현은 다음과 같다.

● 일반적인 다단계 절의 유형
- 1.1, 1.2, 1.2.1, 1.2.2 -
- 1-1, 1-2, 1-2-1, 1-2-2 -

이를 다음과 같은 정규 표현식으로 나타낼 수 있다.

```
Multi_Sec_Expression ::= (Digit1 Digit0* '-')*Digit1 Digit0*
        |(Digit1 Digit0*'.')*Digit1 Digit0*(' ')*
```

3. 문서를 HTML 문서로 변환

문서 전체가 HTML 문서로 변환된 결과를 생성하고자 하는 것이 아니라, 사용자가 원하는 영역에 대해 동적으로 HTML 문서를 생성하여 제공하고자 한다.

이 절에서는 대상 텍스트 문서를 어떠한 형식으로 웹 브라우저를 통하여 검색하게 될 것인가를 논의하고, 이러한 인터페이스를 제공하기 위해 필요한 데이터베이스 스키마를 개발하고자 한다.

3.1 사용자 인터페이스

개발하고자 하는 변환시스템은 먼저 문서의 목차 테이블을 자동으로 구축하고 목차 테이블의 각 항목에 하이퍼링크를 설정한다. 목차의 내용은 원래 문서의 논리 구조를 구성하는 장과 절 등에 대한 제목들이다.

이러한 목차 테이블이 웹에서 검색하는 첫 화면이 되며 그 중 한 항목이 클릭되면 그 항목의 내용이 제공되고, 만약 그 하위에 속하는 절들이 있다면 그 절들의 제목을 하이퍼텍스트화하여 제공한다. 이러한 인터페이스로 사용자는 브라우저를 통해 텍스트 문서를 그 논리 구조에 따라 HTML 문서 형태로 문서 내부를 항해할 수 있다.

3.2 데이터베이스 스키마 개발

웹상에 문서 전체를 로딩하지 않고, 사용자가 원하는 부분만을 제공하기 위해서는 각 문서의 구조에 대한 정보와 각 논리 단위에 해당하는 부분의 범위 정보가 필요하다. 이들을 저장하기 위해 관계형 데이터베이스 스키마를 개발한다.

문서의 일부를 HTML 문서로 동적으로 표현하기 때문에 입력 문서에 대한 다양한 정보가 필요하다. 기본적으로 문서를 담고 있는 파일 이름과 저장되어 있는 장소의 정보가 필요하다. 그리고 문서의 특정 부분을 화면에 제공하기 위해 그 부분의 시작과 끝에 대한 정보가 필요하다. 이를 위해 개발한 데이터베이스 스키마의 일부는 표 1~표4와 같다[2].

속성 이름	속성 기능
BOOKNO	문서의 고유 번호
NAME	문서의 이름
FILENAME	문서를 저장한 파일 이름과 절대 경로
기본키	BOOKNO

표 1. BOOK 테이블의 스키마

속성 이름	속성 기능
BOOKNO	문서의 고유 번호
KUBUN	장, 절인가를 판단하는 정보
STARTNO	해당 제목이 문서에서 시작되는 위치
LINECOUNT	해당 제목이 문서에서 끝나는 위치
CONTENT	목차에서 보여줄 제목
기본키	BOOKNO, STARTNO

표 2. MOKCHA 테이블의 스키마

속성 이름	속성 기능
BOOKNO	문서의 고유 번호
STARTNO	해당 제목이 문서에서 시작되는 위치
LINECOUNT	해당 제목이 문서에서 끝나는 위치
CONTENT	장의 제목
기본키	BOOKNO, STARTNO

표 3. CHAPTER 테이블의 스키마

속성 이름	속성 기능
BOOKNO	문서의 고유 번호
CHAPSTARTNO	상위 단계의 장이 시작하는 위치
STARTNO	해당제목이 시작되는 위치
LINECOUNT	해당 제목이 끝나는 위치
CONTENT	절 제목
기본키	BOOKNO, STARTNO

표 4. SECTION1 테이블의 스키마

SECTION1 테이블 스키마는 제 1 단계의 절에 대한 정보를 저장한다. 이와 유사하게 제 2단계, 제 3단계의 절에 대한 정보를 저장하기 위한 스키마를 개발할 수 있다.

4. 문서 변환 시스템 개발

2절에서 개발한 정규 표현식을 이용하여 문서의 논리 구조를 파악하는 모듈을 Visual C++로 구현하였다[3,4]. 문서 구조를 파악한 후 각 구조의 시작과 끝을 3절에서 개발한 데이터베이스 스키마를 이용하여 Oracle 데이터베이스에 저장한다[5]. 이 정보를 이용하여 Windows NT에서 IIS를 이용한 ASP를 통해 각 논리 구조 간에 하이퍼링크를 설정한다. ASP에서 Oracle 데이터베이스를 연동하여 검색된 결과를 HTML 태그를 사용하여 동적인 HTML 문서로 표현한다. 이를 구현하기 위해 Visual Basic Script를 사용하였다[6,7].

개발한 시스템에 대한 구조는 그림 2와 같다.

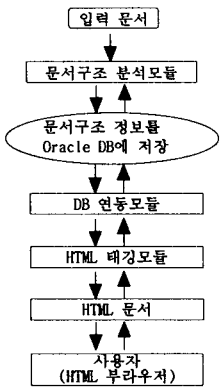

그림 2. HTML 문서 변환시스템의 구조

5. HTML 문서의 검색

개발한 시스템으로 몇 개의 문서를 대상으로 변환하여 본 결과, 논리 구조를 이용한 목차 생성과 각 논리 구조에 대한 하이퍼링크가 정확하게 구현되었음을 확인할 수 있었다. 그림 3은 한 문서의 목차를 익스플로러 브라우저를 통해 볼 수 있는 화면이다.

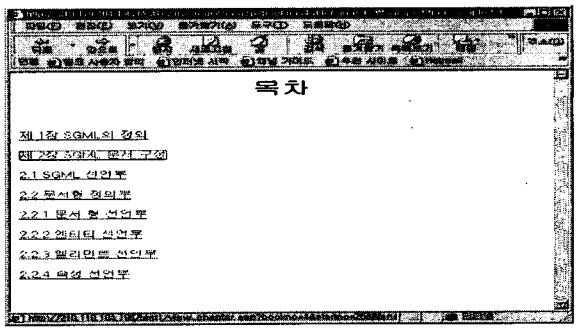

그림 3. 웹상에서 한 문서의 목차를 검색

6. 결론 및 향후 연구 방향

이 논문에서는 전자 문서를 인터넷 환경에서 효율적으로 검색할 수 있는 방법을 제공하기 위해, 일반 텍스트 문서를 HTML 문서 형태로 접근하는 시스템을 개발하였다. 주로 책이나 논문과 같은 문서 타입에서 대중적으로 많이 사용하는 문서 구조를 대상으로 하였다.

이 시스템을 확장하여 멀티미디어 타입의 문서를 하이퍼미디어 문서 형식으로 변환할 수 있도록 보완하여 개발하고자 한다.

참고 문헌

[1] Thomas W. Parsons, Introduction to Compiler Construction, Computer Science Press, 1992.

[2] Elmasri, Navathe, Fundamentals of Database Systems, Third Edition, Addision-Wesley, 2000.

[3] 이상엽, Visual C++ Programming Bible Ver.6.x, 영진출판사, 1998.

[4] Pandolfi, Oliver, Wolski, MFC4 바이블, 송호중 역, 대림출판사, 1997.

[5] 김종근, 홍준호, 송건철, Oracle Bible Ver. 8.x, 영진출판사, 1999.

[6] Fransis, Kauffman, Llibre, Sussaman, Ullman, Beginnin Active Server Pages 2.0, 주병진, 남대우 역, 정보문화사, 1999.

[7] Fransis, Fedorov, Harrison, Homer, Murphy, Sussaman, Smith, Wood, Professional Active Server Page 2.0, 박기성 역, 정보문화사, 1999.