SGML을 이용한 문헌의 구조화 및 텍스트 검색에 관한 연구

Document Structuring and Text Retrieval Using SGML

오민경, 정영미
연세대학교 문헌정보학과

Min-Kyung Oh, Young-Mee Chung
Dept. of Library and Information Science, Yonsei Univ.

본 논문에서는 SGML(Standard Generalized Markup Language)을 사용하여 텍스트 검색시스템을 구축하였다. SGML은 개괄적 마크업언어로서 문헌을 문헌요소라는 객체 단위로 이루어진 것으로 보고 이러한 문헌요소간의 관계를 표현하므로, 텍스트 검색시스템에서 SGML을 이용하면 문헌을 구조화할 수 있고 전문(full text)을 효율적으로 조직하고 검색하는 것이 가능하다.

1 서 론

본 논문에서는 SGML을 도입하여 문헌을 작성하고 이에 기반한 텍스트 검색시스템을 구축함으로써 개괄적 마크업의 유용성을 살펴보았다.

SGML문헌은 DOS상에서 일반 편집기(editor)를 사용하여 작성하였으며 영문은 ASCII 코드로, 한글은 표준완성형으로 표현하였다. 구문분석기는 ARCSGML 1.0을 이용하였고, 텍스트 검색시스템은 C언어를 가지고 구축하였다.

2 이론적 배경

2.1 개괄적 마크업으로서의 SGML

문헌에 글자모양이나 글자크기와 같은 포매팅과 관련된 지시어를 포함하는 절차적 마크업(procedural markup)과는 달리 개괄적 마크업은 문헌을 논리적 단위로 쪼개서 내용만을 저장할 뿐 처리에 관한 정보는 저장하지 않는다. 개괄적 마크업을 사용하면 문헌을 편집하고 수정하고자 할 경우 개개의 문헌을 일일이 수정할 필요없이 문서양식(style sheet) 화일과 연결시킴으로써 원하는 형식으로 출력이 가능하고 시스템이나 응용프로그램이 상이해도 SGML문헌을 읽어들일 수 있다.

이러한 개괄적 마크업에 대한 유용성이 부각되자 ISO/IEC/JTC1/SC18에서는 1986년에 SGML(ISO 8879 -- Standard Generalized Markup Language)를 제정하게 되었고 이것은 1988년에 개정되었다.

2.2 SGML문헌의 구조

SGML문헌은 문헌선언부(Document Declaration), 문헌유형정의부(DTD; Document Type Definition), 실제문헌부(DI; Document Instance)로 이루어진다.

1) 문헌선언부

SGML 문헌선언부는 문헌을 작성하고 응용시스템을 실행시키기 위한 환경을 설정하는 부분으로, SGML문헌의 처음에 기술된 문헌선언부에 의해서 이 문헌을 특정 시스템하에서 처리할 수 있는지를 알 수 있게 한다.

2) 문헌유형정의부

문헌유형정의부는 크게 문헌요소선언, 속성정의리스트선언, 그리고 엔티티선언으로 나뉜다. 문헌요소선언에 의해 각 문헌요소의 개괄적 식별기호(generic identifier: 일종의 태그명)를 정의하고 각 문헌요소가 문헌에 나타나는 순서도 정의한다. 속정정의리스트는 각 문헌요소에 명시되는 속성을 정의하고 이러한 속성값의 범위, 속성값의 초기값을 명시한다. 엔티티선언은 긴 문자열을 대치하는 짧은 문자열을 선언하거나 자판기에 의해서 입력할 수 없는 문자를 선언하는 데 사용된다.

3) 실제문헌부

ISO 8879에서 정의한 SGML문헌에 대한 마크업 기술 방법과 문헌유형정의부에 따라서 작성된 SGML문헌으로, 각 문헌요소가 문헌을 구성하는 단위가 된다. 각각의 문헌요소는 시작과 끝에 시작태그(<문헌요소명>)와 끝태그(</문헌요소명>)를 표시함으로써 구별된다.

3 SGML을 이용한 텍스트 검색시스템 구축

본 연구에서 설계한 텍스트 검색시스템은 SGML문헌 생성모듈, 색인작업모듈, 검색모듈로 이루어져 있다.

3.1 SGML문헌 생성모듈

SGML문헌을 생성모듈에서는 문헌유형정의부와 실제문헌부를 작성한다. 본 시스템에서 생성된 SGML문헌의 예를 살펴보면 다음과 같다. <그림 1>에서 article은 front와 body로 이루어져 있으며 front는 하나의 title, 2명 이상의 저자, 그리고 서지사항으로 이루어 진 것을 알 수 있다. <그림 2>에서 보듯이 article, front 문헌요소는 내용으로 문헌요소를 포함하고 있으며 title, author, citation이 실제 데이타(#PCDATA)를 포함하고 있다.

```
<!DOCTYPE article [
<!element (article) -- (front, body)>
<!element (front)  -- (title,author+,citation)>
<!element title, author, citation) -- (#PCDATA)>
<!element (body)   -- (chap+)>
<!element (chap)   -- (h, ps, cont*)>
<!element (cont)   -- (sec+)>
<!element (sec)    -- (h, ps)>
<!element (ps)     -- (p+)>
<!element (h, p)   -- (#PCDATA)>
]>
```

<그림 1> 문헌유형정의부

```
〈article〉
〈front〉
〈title〉
멀티미디어 정보시스템 플랫폼
〈/title〉
〈author〉
황규영
〈/author〉
〈citation〉
정보과학회지, 제10권 제5호 (1992), pp.5-9.
〈/citation〉
〈/front〉
.....
```

<그림 2> 실제문헌부

3.2 색인작성모듈

색인작성과정에 의해 SGML문헌의 텍스트 부분과 문헌요소 태그를 분리하고, 문헌요소 테이블과 텍스트 화일을 작성하게 되는데 문헌요소 테이블의 각 레코드는 문헌요소명, 상·하위 문헌요소(또는 텍스트화일)에 대한 포인터, 그리고 하위문헌요소의 개수(또는 텍스트 행수)를 포함한다. 그리고 텍스트 내용은 텍스트 화일에 따로 저장된다. 따라서 텍스트 내용에 접근하고자 하는 경우 문헌요소 테이블을 통해

서만 가능하다. 문헌요소 테이블과 텍스트 화일의 구조를 살펴보면 다음과 같다.

<그림 3>에서 element[3]은 문헌요소명이 title이고 텍스트 화일에 대한 포인터는 0을 가리킨다.

[0]	article	-1	1	64
[1]	front	0	2	3
[2]	title	1	0	1
[3]	author	1	1	1

<그림 3> 문헌요소 테이블의 레코드 구조

[0] 멀티미디어 정보 시스템 플랫폼 <-- element[2]
[1] 황규영 <-- element[3]]

<그림 4> 텍스트 화일의 예

3.3 검색모듈

1) 검색명령문

명령문은 명령어, 접두코드, 연산기호로 이루어지며 사용되는 명령어는 다음과 같다.

find : 탐색명령어
display : 출력명령어
combin : 검색결과집합(탐색문번호) 조합명령어
save : 검색된 텍스트 내용을 화일로 저장하기 위한 명령어
create : 검색된 문헌요소 포인터리스트를 화일로 저장하기 위한 명령어
read : create로 저장된 검색결과리스트를 읽어 들이기 위한 명령어
quit : 탐색을 종료하자 하는 경우에 사용되는 명령어

접두코드는 탐색어, 탐색 문헌요소 범위, 출력 문헌요소 범위, 검색결과집합을 지시하기 위하여 만든 코드로서 다음과 같다.

ST(Search Term) - 탐색할 용어를 지정하기 위해 사용되는 접두코드이다.
SDE(Search Data-Element) - 탐색하고자 하는 문헌요소 범위를 지정하는데 사용하는 접두코드이다.
SE(Search Element) - SSE와 비슷하나 SSE보다 상위인 문헌요소를 지정하는데 사용한다.
DE(Display Element) - 출력의 범위를 지정하는데 사용하는 접두코드이다.
DDE(Display Data-Element) - DE와 비슷하나 문헌요소명이 DE보다 하위인 문헌요소를 지정한다.
SS(Search Set) - 탐색문번호를 지정하기 위해 사용되는 접두코드이다.

연산기호(operator)는 탐색어나 검색결과집합을 조합하고자 하는 경우에 사용된다. 본 시스템에서 피연산자는 2개만 사용가능하고 연산기호로는 교집합(&)과 합집합(|), 그리고 차집합(!)의 세 가지 연산기호가 사용된다.

2) 탐색문 테이블과 검색결과리스트 구조

탐색문 테이블은 명령문을 저장하고 검색결과 개수와 검색결과리스트에 대한 포인터를 저장하고, 검색결과리스트는 검색된 문헌요소 테이블에 대한 포인터값을 저장한다. 탐색문 테이블과 검색결과리스트의 구조는 다음과 같다. 첫번째 탐색문에 의해 검색된 텍스트 내용을 읽어 들이고자 하는 경우 search[0]이 지시하는 검색결과리스트부터 검색결과 개수만큼 읽어들인다. 즉 result_list[0]부터 3개의 리스트 값을 읽어 들이면 된다. 이렇게 하여 읽어 들인 14, 19, 53이 지시하는 element[4], element[19], element[53]의 텍스트 화일의 내용을 읽어 들여 화면에 출력한다.

	ST	SDE	SE	DE	DDE		
[0]	멀티미디어	h	chap	-1	-1	0	3
[1]	객체	p	sec	-1	-1	3	3
[2]	·	·	·	·	·	·	·

검색결과리스트에 대한 포인터
검색결과 개수

<그림 5> 탐색문 테이블의 구조

[0]	14
[1]	19
[2]	53
·	·

문헌요소 테이블에 대한 포인터임

<그림 6> 검색결과리스트의 구조

3) 텍스트 검색기능

여기서는 문헌요소 제한검색, 문헌의 구조를 이용한 검색 및 검색결과의 저장과 검색결과의 조합과 같은 검색기능을 제공하는데 자세한 예를 살펴보면 다음과 같다.

(1) 문헌요소 제한검색

· '멀티미디어'라는 용어와 '데이터베이스'라는 용어가 소제목(h)에 있을 경우에만 검색해 내어라.

find ST='멀티미디어' & '데이터베이스'
SDE=h

(2) 문헌의 구조를 이용한 제한검색

· '데이터베이스라는 용어가 절(sec)의 소제목(h)에 출현하는 경우의 절의 소제목을 검색해 내어라

find ST='데이터베이스' SDE=h SE=sec

(3) 출력범위의 지정

· '표준'이라는 용어가 장(chap)의 소제목(h)에 출현하는 논문(article)의 제목(title)을 검색하여라.

find ST='표준' SDE=h SE=chap
DE=article DDE=title

(4) 검색결과의 조합

· 탐색문 3과 4를 하나의 검색결과집합으로 만들어라.

combine 3|4

(5) 검색결과의 저장

· 검색된 탐색문 1의 텍스트 내용을 화일로 저장하여라.

save multi.dat SS=1

· 검색된 탐색문 1의 포인터리스트를 화일로 저장하여라

create multi.lst ss=1

(6) 저장된 검색결과화일을 읽어 들이기

· 포인터리스트를 저장한 multi.lst 화일의 포인터리스트를 읽어 들여라.

read multi.lst

(7) 탐색범위만 지정하고 탐색어가 없는 경우

· 목차화일을 만들어라

find ST='*' SDE=h
save content.dat SS=1

4. 결론

본 논문에서는 SGML을 사용하여 문헌을 논리적으로 구조화시키고 텍스트 검색시스템을 구축하였다. 이 시스템에서는 표준적 방식으로 문헌을 작성하여 상호교환할 수 있을 뿐만 아니라 전문(full text) 데이터베이스를 조직할 때 문헌의 논리정보를 이용할 수 있다. 특히 문헌요소를 사용한 색인시스템을 만들 경우 검색시 문헌요소 제한탐색을 가능하게 할 뿐만 아니라 문헌내부의 계층적 구조와 문헌요소간 연관성을 반영한 복잡한 탐색전략도 가능하게 한다.

참 고 문 헌

고승규, "SGML 응용시스템을 위한 SGML 파서 및 API 설계," 연세대학교대학원 석사학위논문, 1993.

이택경, "SGML 문서의 논리적 구조에 근거한 정보검색 시스템에 관한 연구," 연세대학교대학원 석사학위논문, 1994.

현득창, "SGML Parser를 이용한 SGML Document Editor의 구현에 관한 연구," 광운대학교대학원 석사학위논문, 1991.

홍은선, "한글기능을 갖는 SGML 파서와 에디터의 설계에 관한 연구," 광운대학교대학원 박사학위논문, 1992.

Bradley, Neil, "SGML Concepts," *Aslib Proceedings*, Vol.44 No.7/8 (1992), pp.25-28.

Cover, Rubin, "Standard Generalized Markup Language ISO 8879: 1986(SGML): Annotated Bibliography and List of Resources," Ver. 2.0, available from ftp://ftp.ifi.uio.no/pub/ SGML/bibliography.

Goldfarb, Charles F., *SGML Handbook*. (New York : Oxford Univ. Press, 1990)

Macleod, Ian A., "Storage and Retrieval of Structured Documents," *Information Processing & Management*, Vol.26 No.2 (1990), pp.197-208.

Waite, Mitchell and Prata, Stephen, *Waite Group's New C Primer Plus*. 2nd Ed., (Carmel : SAMS Publishing, 1993)

Wright, H., "SGML Frees Information," *Byte*, Vol.17 No.6 (1992), pp.279-286.