學位論文의 全文索引시스템 설계

A Study on the Design of a Full-Text Indexing System for Thesis

추윤미
(이화여자대학교 문헌정보학과 대학원)

Yoon Mi Chu
(Dept. of Lib. & Inf. Sci., Ewha Womans Univ.)

전문데이터베이스는 원문의 접근가능성과 전문탐색의 장점으로 인해 최근 급속하게 발전하고 있다. 그러나 이제까지 대부분의 전문데이터베이스는 문헌의 구조를 고려하지 않고 본문의 문자열에서 자동추출한 색인어를 대상으로 비통제탐색방법을 사용하여 왔으므로 효율적이고 다양한 검색방법을 적용하기 어려웠다. 본 연구에서는 SGML을 이용하여 문헌을 구조화하고 이를 이용한 색인시스템을 설계함으로써, 문헌구조를 이용한 다양한 검색이 가능하도록 하였다. 이를 위해 논문을 대상으로 하여 문헌의 구조를 분석하고, 주요 문헌요소인 초록, 목차, 본문, 참고문헌의 특성을 색인에 반영하였다. 색인시스템은 문헌요소를 태그와 텍스트데이터로 분석하여 색인하는 일차색인과, 일차색인에 의해 만들어진 문헌요소테이블과 내용데이터파일을 이용하여 주요 문헌요소를 색인한 이차색인으로 구성된다.

I. 서론

全文데이터베이스는 문헌의 원문을 제공하고, 전문검색이 가능하다는 장점으로 인해 최근 급속하게 발전하여 왔으며, 도서관의 주요 정보원으로 자리잡고 있다. 그러나 문헌은 방대한 양의 텍스트로 구성된 비정형데이터이므로, 이를 효과적으로 조직하여 검색하는 데는 많은 어려움이 있다.

주로 전문의 검색을 위해서는 문자열을 자동색인하여 이를 파일로 조직하고 불논리연산기호, 또는 인접연산기호를 이용한 비통제탐색방법을 사용하여 왔다. 그러나 이러한 방법은 문헌의 내용과 구조를 고려하지 않기 때문에 검색효율성이 떨어지고 다양한 검색방법을 적용하기 어렵다.

따라서 문헌을 논리적 계층구조를 가진 문헌요소의 집합으로 파악하는 관점이 대두되었으며, 이러한 관점은 문헌구조를 표현할 수 있는 표준 마크업언어인 SGML이 ISO에 의해 제정됨으로써 더욱 본격화되었다(ISO 8879, 1986).

SGML을 이용하여 문헌의 구조를 마크업하면, 각각의 문헌요소를 독립적인 객체로 볼 수 있기 때문에 이를 이용한 다양한 검색이 가능하다. 또한 문헌중에서 특정 문헌요소를 대상으로 검색할 수 있으므로 빠르고 효율적인 검색이 이루어질 수 있다. 이를 위해서는 검색대상이 되는 문헌요소와 그 특성을 규명하고, 효과적이며 다양한 전문검색에 이용할 수 있는 색인방법을 모색해야 한다. 따라서 본 연구는 SGML을 이용하여 문헌을 구조화하고 이를 기반으로 하는 색인시스템을 설계함으로써, 문헌구조화의 장점을 적용한 전문데이터베이스의 구축에 기초를 제공하고자 한다.

II. 전문데이터베이스의 특성

A. 전문데이터베이스의 장점

全文데이터베이스는 1992년 현재 서지데이터베이스의 거의 두 배에 달하는 급속한 성장을 이루면서 전체 데이터베이스의 47%를 점유하고 있다(Williams, 1993). 이렇게 전문데이터베이스가 급성장한 이유로는 서지데이터베이스에 비해 다음과 같은 두 가지 장점이 있기 때문이다.

①원문의 접근가능성: 이용자는 검색 즉시 적합성에 대한 판단을 내릴 수 있다.
②전문탐색: 전문탐색은 서지검색에 비해 재현율이 월등히 높고, 이용자에게 다양한 접근점을 제공한다.

B. 문헌구조화의 필요성

이러한 장점에도 불구하고 이제까지의 전문데이터베이스는 검색에 있어서 다음과 같이 해결해야 할 문제점을 안고 있다. 첫째, 문헌 전체, 또는 문단, 문장과 같은 문헌의 형식적 요소가 검색단위가 되므로 문헌의 내용에 따라 선택적으로 검색하거나 출력할 수 없다. 둘째, 전문의 비통제탐색방법은 재현율은 높은 반면, 정확률이 낮다. 셋째, ASCII 텍스트형식의 전문은 비텍스트 요소와 특수문자를 표현할 수 없으므로 완전한 전문을 제공할 수 없다. 반면, 이미지형식은 전문의 비통제탐색이 불가능하다.

따라서 이러한 문제점을 해결하기 위해 SGML을 도입하여 문헌의 논리적 구조를 검색에 이용하고 문헌을 구성하는 모든 요소를 통합하고자 하였다. 문헌구조화를 전문데이터베이스에 적용하면,

첫째, 문헌요소가 검색단위가 되므로, 다양하고 융통성있는 검색과 출력이 가능하다.

둘째, SGML 문헌에 마크업된 태그를 데이터베이스의 필드로 이용함으로써 문헌요소의 특성에 따라 제한탐색을 할 수 있다.

셋째, 문헌유형에 따른 특성을 색인과 검색에 이용할 수 있다.

넷째, 문헌요소 간의 관계정보를 이용하여 문헌구조의 브라우징과 문헌 내, 또는 문헌 간 하이퍼텍스트를 구축할 수 있다.

다섯째, 그림이나 도표와 같은 비텍스트 요소를 문헌에 통합할 수 있으므로 완벽한 전문을 이용할 수 있다.

III. 전문색인시스템의 설계

A. 시스템의 특징

본 연구에서 설계한 전문데이터베이스의 색인시스템은 그림 1과 같이 구성된다. 색인시스템의 특징은 첫째, 이미지와 ASCII 텍스트, SGML 텍스트형식으로 데이터를 생성하고, 본문의 페이지이미지와 ASCII 텍스트를 색인시스템과 연결하였다. 둘째, 학위논문을 대상으로 문헌의 구조와 문헌요소의 특성을 분석하여 이를 색인에 반영하였다.

B. 전자본의 생성과정

1. ASCII 텍스트 생성

ASCII 텍스트데이터는 본문의 비텍스트요소는 제외하

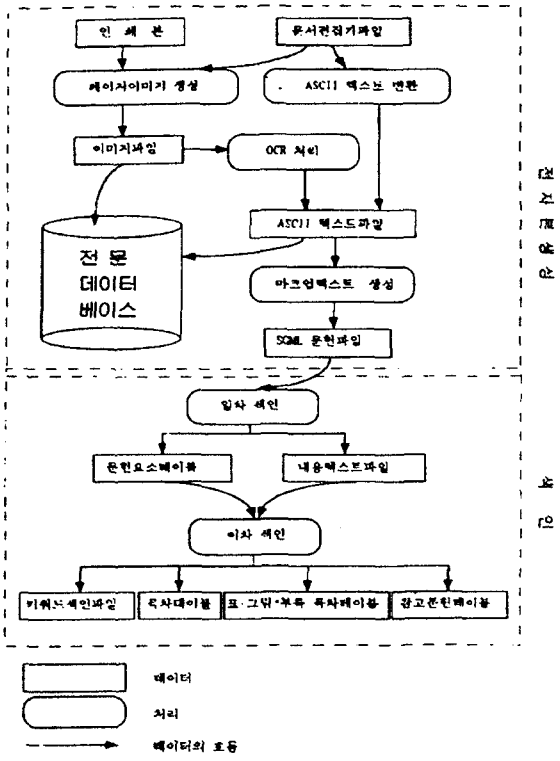

그림 1. 전문색인시스템 구성도

여 생성하며 두 가지로 나뉘어 진행된다.

①인쇄본의 처리 : 스캐닝하여 생성된 페이지이미지파일을 OCR 처리하여 ASCII 텍스트로 생성한다. 이 때 본문 내의 특수문자와 비텍스트 요소는 제외한다.
②문서편집기파일의 처리: ASCII 텍스트형식으로 변환하여 저장한다. 또한 논문의 각 페이지를 이미지파일로 출력하여 페이지이미지파일을 생성한다.

2. SGML 문헌 생성

SGML 마크업을 위해서 학위논문의 구조(그림 2 참조)를 분석하고 이를 DTD로 정의하였다. 다음은 본문 앞부분의 DTD의 예이다.

```
<!-- DTD of thesis -->
<!DOCTYPE thesis
[<!-- ELEMENT     MIN      CONTENT                    -->
<!ELEMENT thesis  - -    (fm, bdy, bm)                >
<!--            Front Matter Elements                 -->
<!ELEMENT fm      - -    (tipg, abs)                  >
<!ELEMENT tipg    - -    (title, author, grad, dept,
                          major?, year, dgr)          >
<!ELEMENT title   - -    #PCDATA                      >
<!ELEMENT author  - -    #PCDATA                      >
...
```

그러나 본문 앞부분의 목차와, 그림, 표, 부록의 목차는 DTD에 포함시키지 않았다. 이는 본문 내에서 태깅된 표

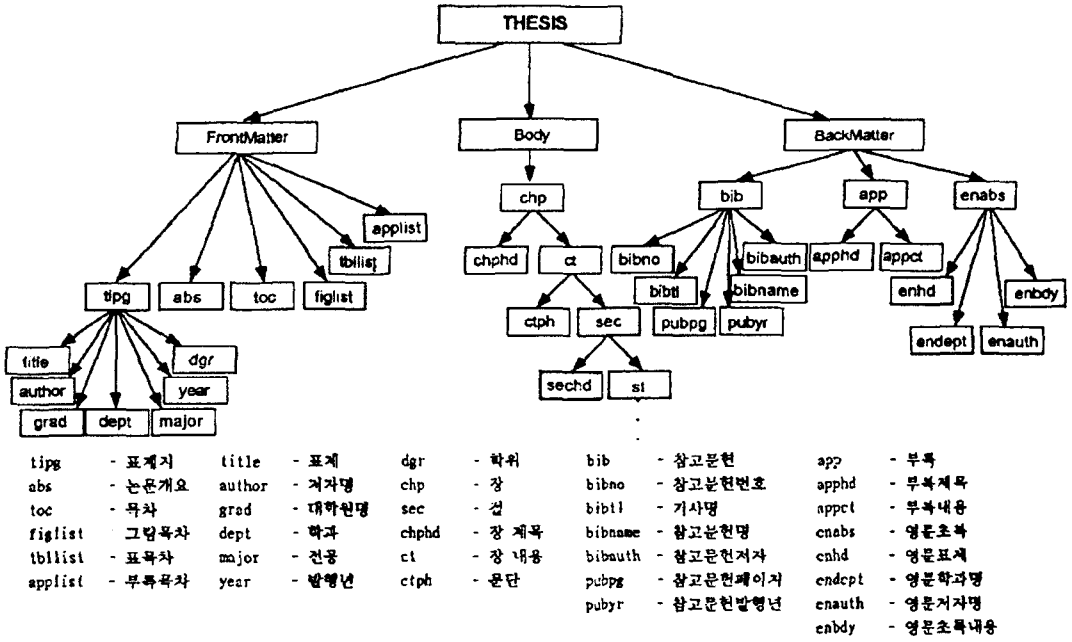

그림 2. 학위논문의 구조

제와 그림, 표, 부록의 제목을 이용하여 색인시스템에서 생성하기 때문이다. 다음의 예는 본문 앞부분의 DI이다.

```
<thesis>
<fm>
<tipg>
<title>개별화 교육을 위한 하이퍼텍스트 저작 시스템의 설계 및 구현</titie>
<author>김태영</author>
<dept>전자계산학과</dept>
<dgr>석사</dgr>
<year>1993</year>
</tipg>
<abs>본 연구에서는 학습자의 인지 스타일에 맞게 학습 구조를 동적으로 구성하는 교육용 하이퍼텍스트 저작 시스템인 NaVil(Navigator for Instruction)를 설계 및 구현하였다. NaVil는 저작 모듈과 학습 모듈로 구성된다. ...</abs>
</fm>
```

C. 색인시스템 설계

1. 색인대상의 특성

전자문헌은 그 구조가 어떻게 표현되느냐에 따라 그 문헌이 가진 정보의 활용에 크게 영향을 준다(長尾眞, 1994:60). 따라서 학위논문의 색인을 위해서 먼저 논문의 검색대상이 되는 각 문헌요소를 규명하고 그 특성을 파악해야 한다. 본 연구에서는 다음과 같이 분석하였다.

①초록: 원문헌의 내용을 요약하여 원문헌을 대신하는 기능과 적합문헌 판정에 도움을 주는 정보검색의 기능이 있다. 초록은 1~2 페이지 정도의 분량에 문헌의 내용을 압축하고 있으므로 표제나 본문에 비해 키워드 추출이 용이하다.

②목차: 문헌의 내용구조 표현하며 초록과 유사하나 좀 더 효율적인 대표정보이다. 따라서 목차를 색인하여 이용자의 문헌 적합성 판정과 목차구조의 검색, 본문검색에 이용할 수 있다.

③본문: 텍스트와 비텍스트요소로 구성되며, 텍스트 내에도 각주와 인용 등의 다양한 요소를 포함하고 있다. 따라서 본문의 디스플레이 시에 이러한 모든 요소를 포함하여야 하며, 이를 위한 각각의 색인이 필요하다.

④참고문헌: 저자가 본문에서 인용하거나 참조한 문헌의 목록으로 본문의 내용과 주제적 연관성을 가지고 있다. 따라서 이용자에게 이차정보로서의 의미를 가지고 있으며, 중요한 정보원이 된다.

2. 색인파일의 구조

가. 일차색인

SGML 텍스트파일을 분석하여 문헌요소테이블과 내용데이터파일로 생성한다(그림 3, 4 참조).

doc no	element no	element name	attr	parent no	child no	next no	page no

doc no - 문헌번호
element no - 문헌요소번호
elementname - 문헌요소명
attr - 속성
parent no - 상위 문헌요소번호
child no - 하위 문헌요소번호
next no - 문헌구조 상의 다음 형제 노드의 문헌요소번호
page no - 페이지번호

그림 3. 문헌요소테이블

doc no	element no	text data

doc no - 문헌번호
element no - 문헌요소번호
text data - 텍스트데이터

그림 4. 내용데이터파일

나. 이차색인

이차색인은 작성된 문헌요소테이블과 내용데이터파일을 이용하여 학위논문의 주요 문헌요소에 대한 색인파일을 생성하는 과정으로 다음과 같이 구성된다.

① 키워드색인파일: 표제와 국문·영문초록, 그리고 목차를 이용하여 키워드를 추출, 도치색인파일을 생성한다(그림 5 참조).

② 목차테이블: 본문의 목차에 신속하게 접근하기 위해 목차테이블을 별도로 생성하였다(그림 6 참조).

③ 표, 그림, 부록목차테이블: 표, 그림, 부록의 브라우징을 위해 각각의 목차테이블을 생성하였다(그림 7 참조).

④ 참고문헌테이블: 참고문헌목록을 본문 내의 인용정보와 함께 테이블로 생성하였다(그림 8 참조).

색인어파일

keyword	docnum	doc pointer

문헌번호파일

doc no	title num	abs num	cont num

keyword - 색인어
docnum - 적합문헌수
doc pointer - 문헌번호파일 포인터
doc no - 문헌번호
titlenum - 표제 발생빈도
absnum - 초록 발생빈도
contnum - 목차 발생빈도

그림 5. 키워드색인파일

doc no	element no	group level	heading	parent no	child no	next no	previous no	page no

group level - 문헌구조상의 레벨(1,2,3...으로 트리의 깊이 표시)
heading - 목차 제목
next no - 다음목차의 문헌번호
previous no - 이전목차의 문헌번호

그림 6. 목차테이블

doc no	element no	heading	table id	ref id*	page no

table id - 표번호
heading - 표제목
ref id* - 참고문헌번호(중복가능)

그림 7. 표목차테이블

doc no	element no	bibauth*	bibtl	bibna me	pub cat	pub pg	pub yr	ref_ele no*

bibauth* - 저자(중복가능)
bibtl - 참고문헌 기사제목
bibname - 참고문헌명
pubcat - 출판물 구분(저널/논문집/단행본/보고서/회의록)
pubpg - 페이지
pubyr - 출판년
ref_ele no* - 본문 내 발생한 참고문헌의 문헌요소번호(중복가능)

그림 8. 참고문헌테이블

IV. 결론

본 연구에서 설계된 전문색인시스템은 다음과 같이 문헌구조를 이용한 다양한 전문검색과 브라우징이 가능하다.

첫째, 문헌요소테이블과 내용데이터파일을 이용하여 탐색어가 포함된 문헌요소를 검색할 수 있으며, 특정 문헌요소만을 검색대상으로 하거나 검색범위를 제한할 수 있다.

둘째, 생성된 표제, 초록, 목차의 키워드색인파일은 키워드탐색의 범위를 표제, 초록, 또는 목차로 선택적으로 제한할 수 있으며, 각 문헌요소에서의 키워드 발생빈도를 이용한 순위매김이나 문헌요소의 중요도에 따른 가중치검색 등 다양한 검색방법을 구현할 수 있다.

셋째, 목차테이블을 이용하여 목차검색과, 검색된 문헌의 적합성 판정, 그리고 목차를 통한 본문의 브라우징이 가능하다.

넷째, 표, 그림, 부록의 목차테이블을 이용하여 본문의 표, 그림, 부록을 브라우징할 수 있다.

다섯째, 참고문헌테이블은 본문의 인용표시와 링크되어 본문에 나타난 참고문헌의 발생위치를 추적하거나 참고문헌을 검색하는 데 이용할 수 있다.

이와 같이 본 연구에서 설계한 전문색인시스템은 문헌의 구조와 특성을 이용한 색인을 통해 다양한 전문검색이 가능하다는 것을 보여줌으로써 문헌구조화를 적용한 전문데이터베이스시스템의 장점을 제시하였다. 앞으로 다양한 문헌유형의 구조와 각 문헌요소의 특성, 전문데이터베이스의 이용자특성분석을 반영한 전문색인 및 검색방법이 연구되어야 할 것이다.

참고문헌

유석종, 고영곤, 최윤철(1995). SGML 한글문서의 논리적 구조에 근거한 자동색인기법에 관한 연구. 『정보관리학회지』, 12(2), 85-101.

長尾眞(1994). 『電子圖書館』. 東京: 岩波書店.

Harman, Donna(1994). Automatic indexing, In Challenges in indexing electronic text and images. ed. by Raya Fidel et. al. Medford, NJ: Learned Information, Inc., 247-264.

ISO 8879(1986). *Information Processing - Text and Office Systems - Standard Generalized Markup Language(SGML)*, International Organization of Standardisation.

Williams, M.(1993). The state of databases today.. In *GALE Directory of Databases*, ed. by Kathleen Young Marcacci, Gale Research Inc., xvii-xxvii.