학술논문의 내용구조에 의한 전문검색시스템 구현과 성능평가에 관한 연구

A Study on the Implementation and Performance Evaluation of Full-text Information Retrieval System based on Scientific Paper's Content Structure

이두영(Too-Young Lee)*, 이병기(Byeong-Ki Lee)**

목 차

초 록

본 연구는 문헌의 내용구조와 이용자의 정보요구는 밀접한 관련성이 있기 때문에 문헌의 본문을 내용 단위구조로 분할하여 색인한다면 기존의 전문데이터베이스 구축방식에 비해 검색효율을 향상시킬 수 있다는 가설을 설정하고 이를 검증하는데 목적이 있다. 이 가설을 검증하기 위하여 먼저 학술논문의 내용구조 모델을 설정하고, 이 모델을 기반으로 컴퓨터 관련분야 70여편의 학술논문을 대상으로 실험용 전문데이터베이스를 구축한 다음, 이에 대한 검색효율을 측정하여 내용구조 기반 전문검색시스템의 성능을 실험적으로 평가하였다.

ABSTRACT

Conventional full-text information retrieval system has been proved with high recall ratio and low precision ratio. One of the disadvantages of full-text IR system is that it is not designed to reflect the user's information need. It is due to the fact that full-text IR system has been designed based on physical and logical structure of document without considering the content of document. The purpose of the study is to develop more effective full-text IR system by resolving such disadvantages of conventional system. The study has developed new method of designing full-text IR system by using Content Structure Markup Language(CSML) other than conventioanal SGML.

키워드 : 내용구조, 문헌구조화, 전문검색, 전문데이터베이스

* 중앙대학교 문헌정보학과 교수
** 중앙대학교 강사, 공항고등학교 사서교사
■ 논문접수일 : 1998년 12월 1일

1. 서 론

현행 전문검색시스템은 본문에 출현하는 모든 단어를 탐색어로 사용할 수 있기 때문에 접근점이 많다는 장점이 있으나 대량의 문헌이 검색되어 부적합 문헌이 검색될 가능성이 높고, 정보요구 상황이나 목적에 따라서 본문의 특정 부분만을 지정하여 탐색할 수 없다는 단점이 있다. 이는 서지사항이나 디스크립터, 초록 등을 추출하여 시스템을 구축하는 서지 초록형 데이터베이스와 같이 전문을 하나의 독립된 필드로 간주하고 본문 필드에 특정 용어의 출현여부에 따라 탐색하는 거시적인 색인 및 검색방식에서 벗어나지 못했기 때문이다.

따라서 각각의 문헌을 일정한 단위로 구조화하고 단위요소별로 특정 부분만을 지정하여 전문을 탐색하거나 특정 부분만을 탐색할 수 있는 부분검색(passage retrieval)의 필요성이 증대되고 있다. 지금까지 문헌 구조를 전문 색인이나 검색에 이용함으로써 검색효율을 향상시키려는 연구로는 담론 구성요소를 이용한 언어학적 구조화(O'connor, 1980), 텍스트의 의미분석을 이용한 구조화(Hearst & Plaunt, 1993), 통계 및 확률을 이용한 구조화(Mittendorf & Schäuble, 1994 ; Salton, 1993), SGML을 이용한 검색 등이 있다. 그러나 이와같은 문헌의 구조화 방식은 전역수준의 색인을 보완하기 위해 국지적 수순의 개념을 도입하여 색인하거나 문헌의 외형적인 구조에 의존함으로써 이용자가 궁극적으로 관심을 갖고 있는 내용상의 특성을 반영하지 못하고 있다.

이에 본 연구에서는 문헌의 내용구조와 이용자의 정보요구 구조는 밀접한 관련성이 있기 때문에 문헌의 본문을 내용 단위구조로 분할하여 색인한다면 기존의 전문데이터베이스 구축방식에 비해 검색효율을 향상시킬 수 있다는 가설을 설정하고 이를 검증하는데 목적이 있다. 구체적으로 본 논문의 가설을 검증하기 위하여 먼저 우리나라 학술논문의 내용구조 모델을 설정하고, 이 모델을 기반으로 하여 실제로 학술논문을 대상으로 실험용 전문데이터베이스를 구축한 다음, 이에 대한 검색효율을 측정하여 전문검색시스템의 성능을 실험적으로 평가해 보고자 한다.

본 연구에서 사용한 Content Structure Mark-up Language(이하 CSML)라는 용어는 문헌의 내용구조에 의한 단위요소를 표현하기 위한 메타언어를 의미하며, 문헌의 물리적 논리적 구조를 표현하는 종래의 표준범용마크업언어(SGML)와 유사하지만 내용단위요소를 식별할 수 있다는 점에서 차이가 있다. 또한 내용구조 정보를 갖는 전자문헌을 CSML문헌, 내용구조 모델에 기초한 전문데이터베이스 구축 및 전문탐색 기능을 갖는 시스템을 CSML전문검색시스템이라 칭하고자 한다.

2. 문헌의 내용구조에 대한 이론적 배경

2.1 전문검색의 문제점과 부분검색의 필요성

본문내의 모든 단어(기능어 제외)를 탐색어로 사용할 수 있는 전문검색시스템은 주제 색인의 비일관성 문제 해결, 원문 입수의 용이성, 문헌표현물이 아닌 원문에 의한 적합성 판정, 고유명사와 같은 특정 용어에 의한 탐색 등 많은 장점이 있으나 자연언어 시스템이 갖는 본질적인 단점(J. Aitchison and A. Gilchirst, 1987) 이외에 전문데이터베이스의 구조와 관련된 2가지 문제점

이 있다.

첫째, 높은 재현율에 비해 정확률이 현저하게 낮다. 탐색 대상인 전문은 서지, 초록과 같은 문헌대용물에 비해 많은 탐색 어휘를 포함하고 있기 때문에 접근점이 많다는 장점이 있으나 그 만큼 대량의 문헌이 검색되어 부적합한 문헌이 검색될 가능성이 높다. 검색효율의 측면에서 대량의 문헌이 검색되는 현상은 높은 재현율에 비해 정확률이 현저하게 떨어짐으로써 잡음이 많고, 이용자에게 검색된 문헌 가운데 적합문헌을 선별해야 하는 인지적 부담을 준다(Negel Woodhead, 1991).

둘째, 문헌구조에 대한 고려없이 문자열의 단순출현이나 형식적인 구조(예를들면 물리적 구조 혹은 논리적 구조)에 의존하여 탐색하고 있기 때문에 내용과 관련된 정보요구 상황이나 목적에 따라서 특정 부분을 지정하거나 본문의 특정부분만을 탐색할 수 없다. 이는 현행 전문검색시스템이 문헌구조에 대한 고려 없이 본문 전체를 대상으로 하는 전문 색인방식을 취하거나 구둣점이나 문단과 같은 형식적인 구조에 의존하여 탐색하고 있기 때문에 내용과 관련된 이용자의 특정 목적이나 상황에 따라서 본문의 특정 부분을 지정하거나 탐색하는데 한계가 있다. 따라서 이용자의 정보요구상황이나 문헌의 이용 목적을 고려하여 전문 가운데 특정 부분을 단위로 색인하고 탐색할 수 있도록 문헌을 구조화할 필요성이 있다.

2.2 내용에 의한 문헌구조화의 타당성

전문데이터를 구조화함으로써 이용자의 정보요구에 적합한 문헌의 일부를 검색함으로써 검색효율을 향상시키려는 지금까지의 연구는 크게 ①형식적 요소에 의한 문헌 구조화와 ②유사성에 의한 문헌구조화로 대별할 수 있다.

형식적 요소에 의한 문헌의 구조화 방법은 최소문단 병합 · 최대문단 분할방법, 페이지 단위 분할방법, 1000바이트 페이지 단위 분할 방법, 30단어 단위 분할방법 등 다양한 방법이 제기되고 있으나 실험적인 단계에 그치고 있다(J. Zobel and A. Moffat, 1995; Stanfill and D. Waltz, 1992). 그 이외에 형식적 요소를 이용한 문헌구조화 수단으로 SGML, HTML과 같은 마크업 언어를 들 수 있다. 그러나 형식적 요소에 의한 문헌 구조화 방법은 이용자가 궁극적으로 관심을 갖고 있는 내용과 관련된 구조를 반영하지 못하고 있으며, 탐색범위 설정에 널리 이용되고 있는 형식구조의 단위요소인 문장과 문단은 문헌내에서 각 문장 혹은 문단을 식별할 수 있는 요소가 없다는 문제점이 제기되고 있다.

유사성에 의한 부분탐색은 한 문헌 전체를 대상으로 적합성 여부를 판단하는 2분법적 탐색이 아니라 유사성을 측정하여 일정 기준 이상의 문헌 혹은 문헌의 일부를 탐색하려는 것이다. G. B. Salton(1994) 등은 전형적인 벡터공간모델을 이용하여 기존의 질의와 문헌간의 유사성 측정을 확대하여 특정 부분 탐색이나 자동구조분할, 자동요약 등에 대한 일련의 실험을 전개하였다. 또한 M. A. Hearst(1993) 등은 문헌 전체 뿐만아니라 다수의 소주제문을 구분할 수 있는 알고리즘을 개발하여 특정 부분만을 탐색할 수 있는 시스템(Textiling)을 개발하였다. 그 이외에도 문헌을 적합한 부분과 부적합한 부분의 연속적인 확률과정으로 표현한 은닉 마코프모델(HMM; Hidden Markov Model)을 이용한 유사성 탐색도 있다(E. Mittendorf and P. Schäuble, 1994). 그러나 유사성에 의한 전문탐색은 전역수준의 색인을 보완하고 문헌의 일부분에 대한 검

색의 필요성을 제시해 주고 있으나 형식적 요소에 의한 문헌구조화와 마찬가지로 이용자가 궁극적으로 관심을 갖고 있는 내용상의 특성을 반영하지 못하고 있다.

반면에 문헌의 내용에 따른 구조는 텍스트언어학과 인지심리학, 이용자의 전문 이용 형태 등의 관점에서 다음과 같은 특징이 있다. ① 내용구조는 실제 언어 사용 환경에서 의사소통 단위로 작용하는 텍스트 단위와 일치한다. ② 문헌의 유형에 따라서 사회적으로 관습화된 전형적인 내용단위가 존재하며, 단위별로 의미 속성을 갖는다. ③ 텍스트 단위 내용구조는 문헌의 생성 이해과정에서 인지구조로 작용한다. ④ 내용구조는 이용자의 정보요구 상황이나 목적과 밀접하게 관련되어 있다. ⑤ 한 편의 논문이나 문헌은 항상 전체를 통람하는 것이 아니라 정보요구의 상황이나 목적에 따라서 특정 부분만을 필요로 하거나 읽는 순서가 달라지며, 문헌 전체가 아닌 소단위만을 보고 적합성을 판단하는 경우가 있다.

이러한 내용구조의 특징은 기존의 문헌구조화 방식이 갖는 한계점을 극복하고 이용자의 정보요구 상황이나 특성에 따라서 검색할 수 있는 전문검색시스템 구현의 가능성을 제시해 주고 있다.

3. 학술논문의 내용구조 모델 설정과 검증

3.1 학술논문의 내용구조 모델 설정

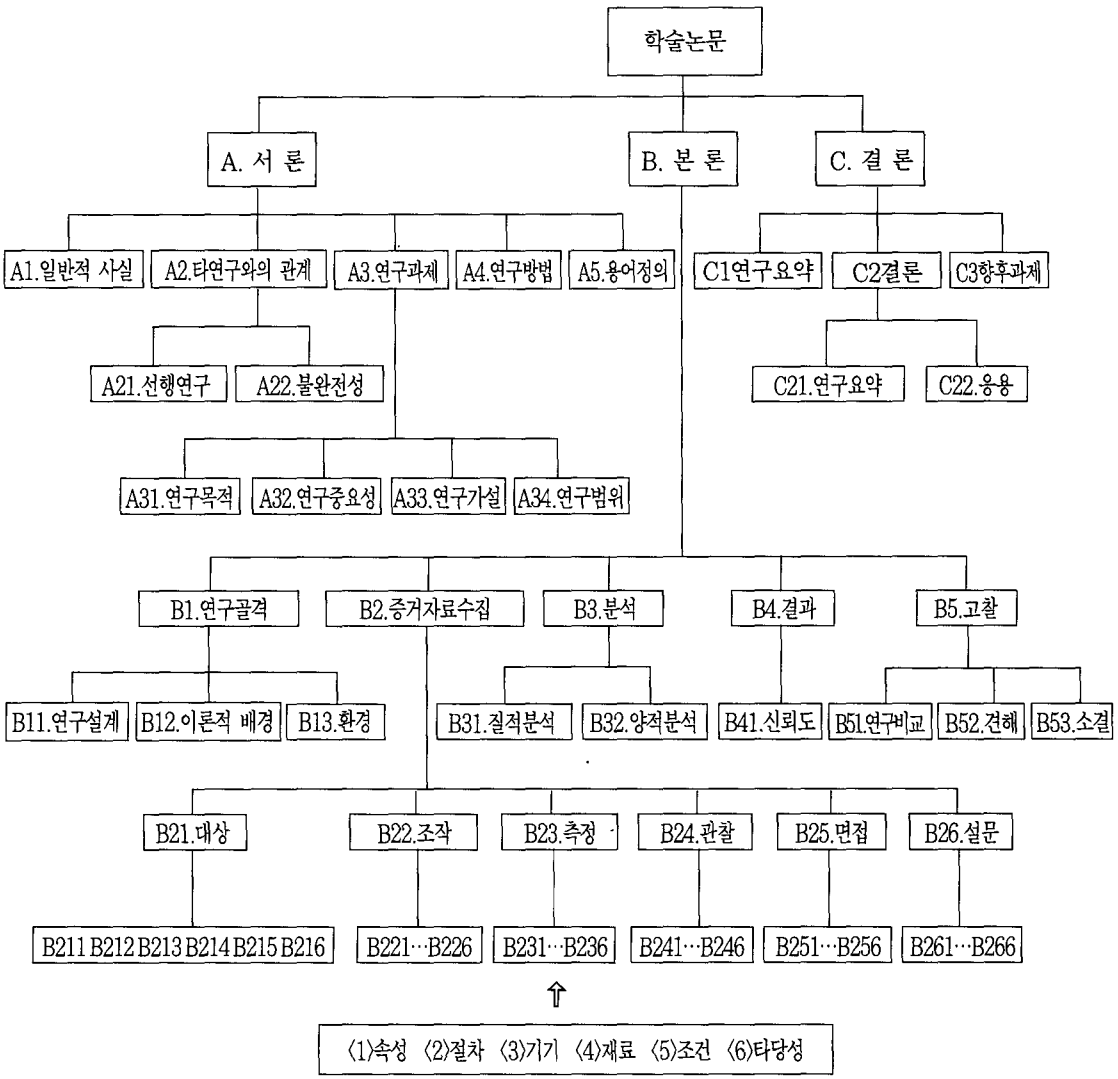
내용구조에 의한 전문검색시스템을 구현하기 위해서는 정보검색시스템 설계자와 이용자간에 공통적으로 인식할 수 있는 표준적인 내용구조 모델이 필요하다. 따라서 학술논문을 대상으로 전문검색시스템을 설계할 때에 적용할 수 있는 내용구조 모델을 설정하고, 이 모델에 대한 이용자의 인지도를 측정해 보고자 한다. Van Dijk(1980)의 상부구조, Van Dijk와 Kintsch(1983)의 스키마구조, Liddy(1991)의 초록구조, Noriko Kando(1992)의 내부구조 등 선행연구에 나타난 학술논문의 모델을 상호 비교하여 공통적인 내용 요소를 추출하고, 이를 국내 학술논문에 직접 적용해 봄으로써 문제점을 수정 보완하여 최종적인 학술논문의 내용구조 모델을 설정하였다.

분석대상 학술논문은 국회도서관에서 발행하는 "정기간행물 기사색인(1996. 1-3)"에 수록된 기사 가운데 특정 분야에 편중되지 않도록 13개 주제 분야별로 각 15건씩 총 195건을 1차적으로 선정하고, 원문을 입수한 결과 기술 소개, 경향분석, 시사 정보, 논평 등 학술 논문의 성격에서 벗어난 15건을 제외한 총 180건을 최종 분석 대상으로 삼았다.

이러한 분석 결과를 바탕으로 불필요한 요소를 삭제, 통합, 분리하여 최종적으로 학술논문의 내용구조 모델을 〈그림 1〉과 같이 설정하였다.

3.2 학술논문의 내용구조 모델 검증

설정한 학술논문의 내용구조 모델에 대한 타당성을 검증하기 위해서 동일한 내용의 학술논문을 다수인에게 분석하게 한후 종합적인 신뢰도를 측정함으로써 얼마나 내용구조 모델을 이용자들이 공통적으로 인식할 수 있는지, 그리고 어느 정도 일치하는지를 측정해 보았다. 국문학, 역사학, 지리학, 교육학, 생물학 분야의 석 박사과정에 재학중인 5명에게 주제분야 및 연구유형이 각각 다른 3편의 학술논문과 설정한 학술논문의 내용구조



⟨그림 1⟩ 학술논문의 표준 내용구조 모델

모델을 제시해주고 적절한 단위요소 태그를 학술논문에 직접 표시토록 하였다.

학술논문의 특정 부분에 대해 피험자들이 부여한 단위요소 태그를 비교하여 5명이 모두 일치하면 5점, 4명이 일치하면 4점, 3명이 일치하면 3점, 2명이 일치하면 2점, 1명이면 1점을 주어 Kaplan과 Goldsen이 제시한 지수를 통해 개인별일치도와 조사문헌별 종합일치도를 측정하여 신뢰도를 조사하였다.

$$\text{개인별일치도}(R_i) = \frac{\sum_{j=1}^{K} e_{ij}}{NK} X100 \begin{cases} N: \text{분석자수} \\ K: \text{부여한 태그 수} \\ e_{ij}: \text{태그별 점수} \end{cases}$$

$$\text{종합일치도}(R_t) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{K} e_{ij}}{N^2K} \; X\ 100$$

그 결과 문헌1,2,3의 종합일치도는 57.2%, 71.6%, 77.2%로 나타났으며, 이를 상관계수로 해석해 보면 문헌1의 신뢰도는 보통수준(R=0.57)이지만 문헌2와 3은 매우 높은 신뢰도(R=0.71, 0.77)를 보이고 있다. 이상의 검증 결과를 볼 때, 본고에서 설정한 학술논문의 내용구조 모델은 타당성이 있으며, 이용자들에게 내용구조모델의 단위요소를 제시해 주면 공통적으로 인식할 수 있고, 특정 내용범주를 설정할 수 있음이 확인되었다.

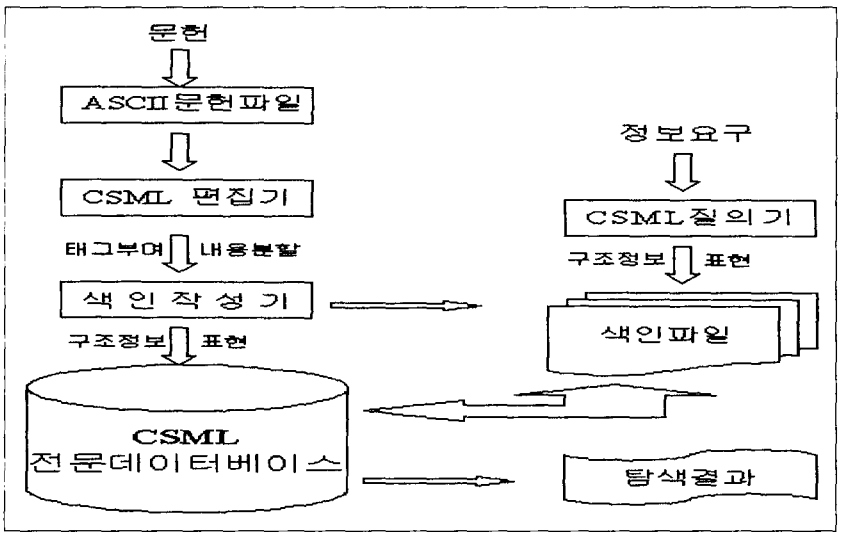
4. 내용구조 기반 전문검색시스템의 설계 및 평가

4.1 시스템 환경 및 구성

본 시스템은 Borland社에서 개발한 Delphi 2.0과 BDE(Borland Database Engine)라는 내장 DBMS를 이용하여 구현하였으며 Window 95 환경에서 작동한다.

본고에서 구현한 CSML 전문검색시스템은 실제 이용환경에서 시스템이 갖추어야 할 제반 요소를 고려한 것이 아니라 최초 연구가설을 검증하는데 필요한 최소한의 환경으로 구현하였다. CSML 전문검색시스템의 전반적인 구성은 〈그림 2〉와 같다.

〈그림 2〉에서 보는 바와같이 본 탐색시스템은 크게 세 부분으로 구성되어있다. 첫째로 텍스트 형태의 파일을 읽어 들여 문서 편집기를 통해 단위요소의 태그를 부여함으로써 확장자 *.csml 파일을 생성하는 부분이고, 두 번째 부분은 전문색인 작성기를 통해 全文索引 파일을 작성하고 CSML파일을 토큰 값으로 저장하여 전문데이터베이스를 구축하는 과정이다. 세 번째 부분은 전통적인 전문탐색 시스템에서 제공되는 논리조합, 거리지정, 논리구조에 의한 제한탐색 이외에 내용구조를 지정하여 탐색할 수 있는 CSML 전문



〈그림 2〉 내용구조 기반 전문검색시스템의 구성도

탐색기이다.

4.1.1. CSML 생성모듈

CSML생성모듈은 스캔닝한 이미지 파일을 OCR처리하거나 ASCII 파일로 저장된 학술논문을 문헌 단위로 호출하여 CSML 문서편집기를 통해 내용구조의 단위요소를 태깅하는 부분이다. CSML 전문데이터베이스에는 내용 단위요소 이외에 서지적 태그가 있으나 최소한의 문헌 식별에 필요한 기본적인 요소로 한정 하였다. 〈그림 3〉은 CSML 문서에서 사용한 서지적 태그와 그 의미를 나타내고 있다. CSML 태그는 제3장에서 설정한 표준 내용구조 모델에서 사용한 코드를 그대로 사용하였다. 태그는 영문자 하나와 아라비아 숫자 코드를 조합하여 구성하였으며, 숫자 코드에서 아라비아 숫자와 자리수는 태그간의 계층적 속성을 갖는다. 예를들어, CSML 문서는 크게 단위요소 A · B · C로 구성되며, 단위요소 A는 A1과 A3이라는 하위요소를 갖고, A3은 A31, A32, A33이라는 하위요소로 구성된다. 태그 부여의 기본 형식은 시작 표시기호(〈)와 종료표시기호(〈/〉)로 표현하였다.

〈DOID〉문헌번호〈/DOID〉
〈TITL〉제목〈/TITL〉
〈JOUR〉수록저널명〈/JOUR〉〈YEAR〉발행년
〈/YEAR〉〈VOLU〉권〈/VOLU〉〈NOUM〉호
〈/NOUM〉〈PAGE〉수록범위〈/PAGE〉
〈CHTL〉1.서론〈/CHTL〉
〈CHTL〉장제목(제1수준 제목)〈/CHTL〉
〈SETL〉절제목(제2수준 제목)〈/SETL〉
〈SUTL〉항제목(제3수준 제목)〈/SUTL〉

〈그림 3〉 서지적 태그와 의미

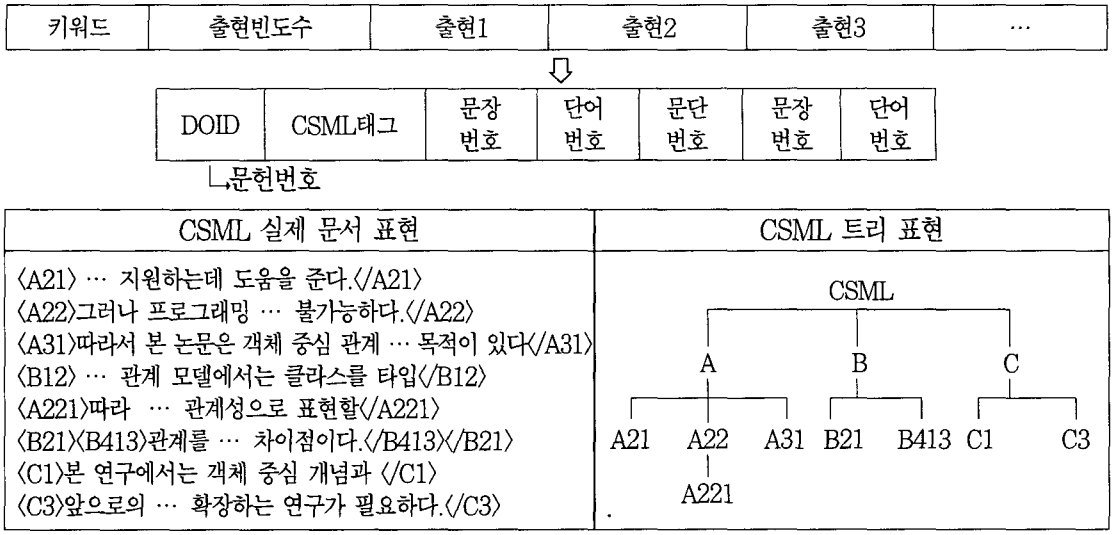
CSML 문서 편집기를 실행하면 편집 화면이 나타나고 파일 버튼을 클릭하여 해당 ASCII 파일을 호출하고, 편집 메뉴를 통해 필요한 부분에 해당 태그를 부여할 수 있다. 편집 버튼을 클릭하면 서지적 태그와 내용구조 단위요소의 태그 전체를 보여주며 해당 부분을 클릭하면 자동적으로 태깅이 되도록 설계하였다.

4.1.2 전문색인 모듈

전문데이터베이스 전체를 차례로 읽어들여 필요한 문자열을 찾아내는 순차 텍스트 검색은 많은 시간이 소요되기 때문에 대부분의 전문검색시스템에서는 문단 번호, 문장 번호등 위치 정보를 갖는 전문색인 파일을 유지하고 있다. CSML 전문검색시스템의 색인작성모듈은 이러한 전문 색인파일의 개념과 내용구조 정보에 의한 위치정보를 갖는 색인구조로 표현된다. 다만 CSML시스템의 전문색인은 문단번호, 문장번호와 같은 위치정보 이외에 내용상의 구조 정보를 유지하고 있다는 점에서 차이가 있다.

내용구조 정보를 갖는 색인파일을 작성하기 위해서는 확장자 *.csml 파일을 불러들여 계층적 의미를 갖는 태그를 노드로 하는 문헌 트리(Document Tree)를 생성할 필요가 있다. 문헌트리는 자노드의 개수가 일정하지 않기 때문에 일반트리 방식으로 구성하였다. 전문색인 파일의 기본구조와 문헌트리 자료 저장구조는 〈그림 4〉과 같다.

전문 색인파일에 수록될 키워드는 한글 자동색인기를 통해 후보 색인어를 추출하고 오류를 반자동으로 수정하는 2단계 방법을 채택하였다. 현재까지는 한글을 대상으로 한 자연어 처리에 있어서 단어의 의미와 문맥에 의해 주제어를 정확하게 추출한다는 것은 불가능 하며, 그 대안으로

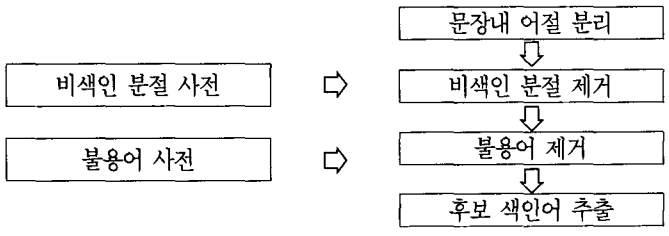


〈그림 4〉 전문 색인파일의 기본 구조

써 어절단위 색인법과 형태소 단위 색인법이 널리 적용되고 있다. 어절단위 색인법은 문헌이나 질의에서 불용어를 제외한 모든 어절들을 색인어 후보로 간주하고 각 어절로부터 색인어의 부분으로써 무의미한 비색인분절(non-indexable segment)을 제거한 나머지 부분을 색인어로 선택하는 방법이다. 형태소단위 색인법은 일반적으로 형태소 해석을 수행함으로써 문장 가운데의 각 어절을 명사, 조사, 부사 등의 형태소 단위로 분리한 후 명사나 명사구를 색인어로 선택하는 방법이다. 본 시스템에서는 어절 단위 색인법을 사용하였으며, 본 시스템의 자동색인기에 의해 후보 색인어를 추출하는 전반적인 과정은 〈그림 5〉와 같다.

각 단계별 색인 과정을 구체적으로 살펴 보면 다음과 같다.

첫째, 문장내 어절 분리 · 우선 띄어쓰기(공간), 마침표, 쉼표를 구분자로 하여 어절을 인식하였



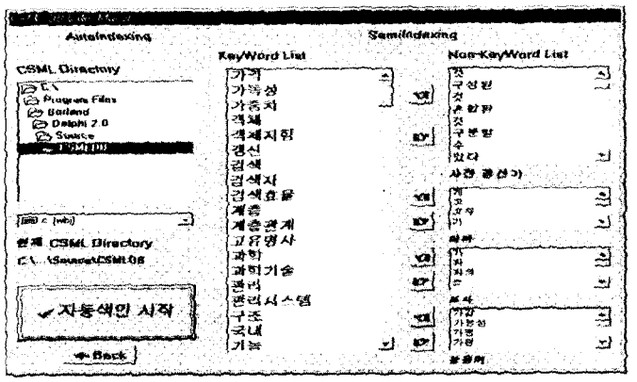
〈그림 5〉 자동 색인 처리과정

다. 띄어쓰기에 있어서 원저자의 문법적 오류에 관계없이 원문에 나타나 있는 그대로를 입력 정보로 삼았다. 둘째, 비색인 분절 제거 · 분리된 각 어절과 비색인 분절사전을 대조하여 일치하는 분절이 있으면 해당 규칙에 따라 일치하는 부분을 제거하였다. 비색인 분절 사전에는 〈표 1〉과 같이 활용형 어미를 포함하고 있는 용언, 체언에 접속되는 조사 혹은 조사 겸 어미 등이 포함되어 있다. 셋째, 불용어 제거. 비색인 분절은 아니지만 키워드로 적합하지 않은 어절을 제외시켰다. 여기에는 주로 대명사, 수사, 감탄사, 부사 그리고 '기존, 제안, 설명, 연구' 와 같이 주제와 관련없는 상용어구를 포함시켰다. 비색인 분절 색인이나 불용어 색인의 규칙에 적용없이 키워드로 선정된 어절([명사]+[명사]+ · · ·)은 각 명사 뿐만아니라 복합명사([명사]+[명사] · · ·)도 키워드에 포함시켰다.

어절 단위에 의한 한글 자동색인은 비색인 분절을 절단할 때 오류가 발생할 수 있으며, 특히 의미가 동일한 복합명사의 띄어쓰기 문제를 적절히 처리하지 못하기 때문에 문서내에 많은 복합명사들이 포함되어 있을 경우 검색 효과가 저하되는 문제점을 지니고 있다. 따라서 자동색인기의 처리 결과(후보 색인어)를 수작업으로 확인하여 수정하는 반자동 색인을 병행 하였다. 자동색인과 반자동색인의 처리 과정을 보여주는 윈도우

〈표 1〉 자동색인과정에 사용된 사전의 유형 및 사례

사전의 유형	내 용	요 소	적 용 규 칙
비색인 분절 사전	용언에 접속되는 어미	른, 인, 라, 도록, 운, 될, 해, 며, 어서, 시, 고자, 기, 진, 던, 한, 게, 어, 여, 다, 된, 는데, 으며, 고, 는, 할, 냐, 면 등	[어간]+[어미활용]⇨[제거] 어미활용에 해당하는 분절이 포함된 어절은 키워드에서 제외.
	체언에 접속되는 조사	에서의, 에서는, 으로부터, 으로써, 으로는, 로부터, 과의, 으로서, 만큼, 처럼, 에서, 이란, 에도, 로써, 로서, 으로, 부터, 마다, 이나, 보다, 의, 을, 에, 를, 과, 와, 는, 은, 는, 이, 가 등	[체언]+[조사]⇨[체언] 각 조사의 문자수를 세어서, 그 수 만큼 각 어절의 뒤에서부터 읽어 나가면서 비교하여 일치하는 부분을 제거, 나머지부분은 키워드로 선정 함.
불용어 사전	• 대명사, 수사, 감탄사, 부사 등 • 비주제적 명사 • 의존 명사 • 숫자, 특수문자	이, 그, 저, 이것, 그 것, 이것, 저것, 그는, 첫째, 둘째, 셋째, 넷째, 할, 수, 바, 것, 숫자, 특수문자, 경우, 예, 해석, 가능성, 구성,그러나, 그리고, 그런데, 그리하여 등	[어절]=[불용어사전]⇨ 제거
	• 복자음 어절	않, 많, 앉 등	[복자음 어절로 시작]⇨ 제거



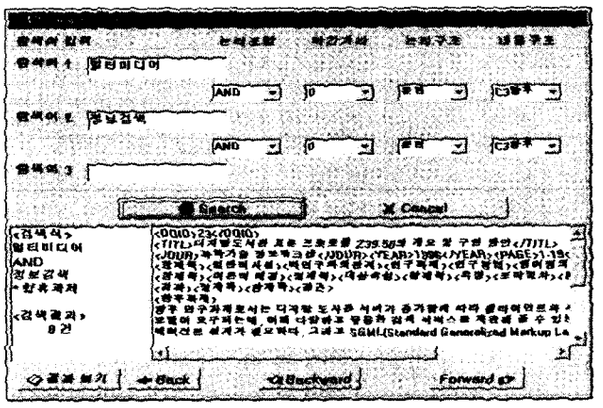
〈그림 6〉 전문색인 실행 Windows

는 〈그림 6〉과 같다.

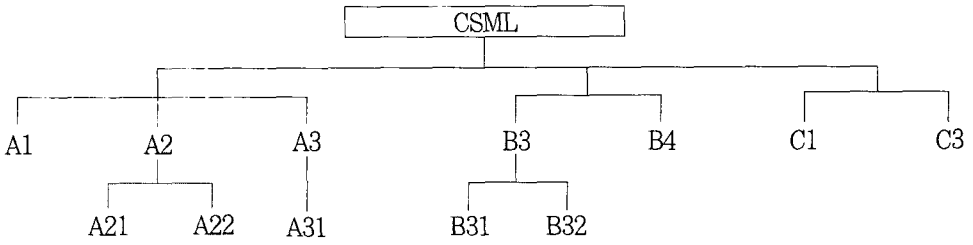
〈그림 6〉에서 보는 바와 같이 자동색인 과정의 처리결과를 사용자가 직접 확인하여 변경할 수 있도록 설계하였다. 키워드리스트 혹은 비키워드리스트를 확인하고 방향키(⇦⇨)를 통해 상호 이동시킴으로써 오류를 수정할 수 있다. 특히 키워드리스트에서 비키워드리스트(⇨)로 이동할 때는 대화상자를 통해 비분절색인 사전 및 불용어 사전을 갱신할 수 있도록 설계하여 동일한 작업을 반복하지 않도록 고려하였다.

4.1.3 전문탐색 모듈

전문탐색 모듈은 탐색어와 해당 조건을 입력받아 적합한 문헌, 문장, 문단, 내용 단위를 찾아내서 결과를 디스플레이하는 부분이다. 이 모듈은 GUI 방식으로 질의 대화상자를 통해 탐색어와 해당조건을 입력할 수 있으며, 〈그림 7〉과 같이 논리조합(AND, OR, NOT), 어간거리지정, 논리구조지정 이외에 내용 구조를 지정할 수 있도록 설계하였다. 이 모듈에서는 탐색어 입력 부분에 찾고자 하는 용어를 입력하고, 지정하고자 하는



〈그림 7〉 전문탐색모듈 실행 Window



〈그림 8〉 내용 범위에 따른 탐색 영역 지정

탐색 범위에 대화상자를 클릭하면 지정할 수 있는 메뉴상자가 자동으로 열리고 마우스 혹은 방향키를 이용하여 조건을 지정할 수 있다.

본 시스템의 전문탐색모듈은 기존의 전문탐색과 내용구조 정보를 검색에 이용함으로써 보다 다양한 검색을 지원할 수 있도록 설계하였다. 본 시스템의 전문탐색 기능은 크게 기본탐색과 복합탐색으로 나눌 수 있다. 기본탐색 기능은 A라는 용어가 특정 수준의 제목에 포함된 문헌 검색, A라는 용어와 B라는 용어가 동시에 출현한 문헌 검색, A · B가 동일한 문장 혹은 문단에서 출현하는 문헌 검색, A · B가 동일한 내용 단위내에서 출현한 문헌 혹은 부분 검색 등과 같이 문헌 전체 혹은 특정 범위를 지정하여 탐색하는 방법을 말한다. 내용 단위에 의한 범위 지정은 계층적 속성에 따라서 〈그림 8〉에서와 같이 단위요소 A2를 지정하면 A2 이하의 모든 단위요소(A21,A22)를 지정한 효과가 있다. 최상위 내용 단위인 CSML로 지정하면 내용구조를 전혀 지정하지 않은 상태와 동일하게 된다. 복수의 탐색어가 동일한 내용 범주에서 출현한 부분이나 문헌을 찾고자 할 때에는 각 탐색어에 동일한 내용 단위를 지정하면 된다.

복합탐색 기능은 어간거리 지정은 물론 논리구조와 내용구조 정보를 종합적으로 이용하는 검색기법을 말한다. 이는 〈탐색어A와 B가 동일한 문단내에 출현하고 통계적 기법(B322)을 적용한 문헌 검색〉, 〈소제목에 탐색어A를 포함하고 향후 과제(C3)에 B라는 용어를 포함하고 있는 문헌 검색〉, 〈탐색어 A와 B가 동일한 문단에서 출현하는 문헌의 결론(C)부분을 출력〉과 같이 각 제한 탐색간의 조합에 의한 탐색이 가능하다.

4.2 CSML 전문검색 시스템의 성능평가

4.2.1 평가방법

본 시스템의 성능을 평가하기 위해서 컴퓨터공학과 문헌정보학 분야의 학술잡지 기사를 실험장서군(test collection)으로 삼아 내용구조 정보를 갖는 CSML전문데이터베이스를 구축하였다. 컴퓨터공학과 문헌정보학 분야를 실험장서군으로 선정한 이유는 두 분야에서 사용하고 있는 탐색언어간에는 높은 유사성이 있으나 적합문헌의 성격에는 많은 차이가 있어서 적합문헌 판단에 도움이 되기 때문이다. 선정된 학술잡지는 정보관리학회지 제13권 제1 · 2호(1996년 6월, 12월) 전체 22건, 과학기술정보 워크샵(KOSTI '96 proceeding)에서 英文記事를 제외한 國文記事 32건, 정보과학회지에서 16건으로 총 70건을 대상으로 하였다. 정보과학지에서 선정한 16건은 적

절한 수의 적합 문헌과 부적합 문헌을 전문데이터베이스에 수록하기 위한 방안으로 본 실험에서 설정한 탐색질의문에 포함된 용어를 하나 이상 표제 혹은 목차에 갖고 있는 논문 기사를 선정한 것이다.

본 연구에서 사용한 탐색질의문은 실험장서군에 포함된 문헌의 용어 출현패턴과 내용구조를 고려하여 실험에 적합한 탐색과제문 5개를 연구자가 직접 작성하고, 이로부터 탐색질의문을 구성하였다. 실험장서군으로부터 객관적인 적합문헌 리스트를 결정하기 위한 방안으로 5개의 탐색과제문을 주제전문가에게 제시하였으며, 탐색과제문으로부터 키워드를 추출하고 불연산기호로 조합하여 최종 탐색질의문을 작성하였다. 각각의 탐색과제문은 아래와 같다.

탐색과제 1: 멀티미디어 환경에 있어서 정보검색에 관한 향후연구 과제에 대한 자료

탐색과제 2: 자동색인 기법이 검색효율에 끼치는 영향에 대한 실험 결과 자료

탐색과제 3: 이용자 연구에 있어서 조사 방법에 관한 자료

탐색과제 4: 정보검색의 기술적 측면에 대한 연구 동향

탐색과제 5: 정보검색 측면에서 본 디지털도서관의 사례

각각의 탐색과제문에 대한 적합성 판정의 일관성 및 객관성을 확보하기 위해서 4명의 주제전문가로 하여금 실험장서군에 포함된 모든 문헌을 검토하여 각각의 탐색과제에 적합한 문헌 혹은 적합한 부분을 판단토록 하고, 이를 종합하여 적합 문헌 리스트를 생성하였다. 주제전문가로부터 적합성 여부를 표시한 실험 대상 문헌을 수합한 후 분석자간에 두명 이상이 적합하다고 표시한 문헌만을 최종 적합문헌으로 삼아 분석자간의 불일치가 큰 문헌은 제외시켰다. 그리고 N. C. North 등이 제안한 분석자 상호간의 일치도와 상호간의 평균일치도를 이용한 복합 신뢰도계수를 통해 적합성 판단에 대한 타당성을 측정하였다. 분석자 상호간의 일치도와 복합 신뢰도 계수를 구하는 공식은 다음과 같다.

$$\text{복합 신뢰도 계수} = \frac{N(\text{분석자 상호간의 평균일치도})}{1+(N-1)(\text{분석자 상호간의 평균일치도})}$$

$$\text{분석자 상호간의 일치도} = \frac{2M}{N_1 + N_2}$$

$\begin{cases} M = 2\text{명의 분석자간에 일치한 적합문헌수} \\ N_1 = \text{분석자1이 표시한 적합문헌수} \\ N_2 = \text{분석자2가 표시한 적합문헌수} \end{cases}$

주제전문가들이 탐색과제문 ①에 대해 적합문헌(혹은 부분)을 판정한 결과와 그 결과에 따른 분석자간의 상호 일치도 및 복합신뢰도 계수 산정값은 〈표 2〉와 같다.

〈표 2〉에서 '분석자별 적합문헌 판정 결과' 항목에 있는 숫자는 각 분석자들이 적합하다고 판단한 문헌의 고유번호이고, 괄호안의 숫자는 해당 문헌의 특정 부분이 적합한 내용임을 표시하는 문단번호이다. 그리고 '적합문헌 리스트' 항목은 각 분석자들이 표시한 적합문헌중에서 2명 이상이 일치하는 문헌번호를 나열한 것이다. 상호일치 행렬표는 분석자 상호간의 유사성 계수를 산정하여 표시한 것으로 이에 대한 산술평균이 상호간평균일치도이다. 이 상호간의 평균일치도는 복합 신뢰도 계수 산정에 사용된다. 이상과 같은 과정을 탐색과제문②-⑤에도 동일하게 적용하여 적합문헌 판정 및 그 신뢰도를 산출하였다.

그 결과 탐색과제①에 대한 적합문헌은 실험장서군 70편중에서 11편으로 나타났으며, 분석자 상호간의일치도는 분석자 A와 B, 그리고 B와 C

〈표 2〉 탐색과제①의 적합문헌 판정 및 신뢰도

탐색과제	분석자	분석자별 적합 문헌 판정 결과			적합문헌 리스트
1	A	1(25), 2(65-69), 3, 23(61), 26(118), 37(18), 41, 56(34-36), 53 59(15), 63(43-44), 64(52-55)			2(65-69) 3 23(61-63) 26(118) 37(18) 41 56(34-36) 58(26) 59(15) 63(43) 64(52-55)
	B	2, 3, 23, 41, 56, 58(26), 63(43), 64			
	C	2(65-69), 3, 9(15), 23(61-63), 26(118-119), 28(53), 37(18) 56(34-36), 58(26), 63(43)			
	D	2(65), 4(26), 26(118), 37(18-19), 56, 58, 59(15), 63(43), 68(72)			
상호일치 행 렬 표		A B C	상호간 평균 일 치 도	복합 신뢰도 계수	
D C B		0.57 0.47 0.56 0.67 0.70 0.70	0.61	0.86	

※ ( )안의 숫자는 문단번호에 의한 범위표시임
탐색과제①: 멀티미디어 환경에 있어서 정보검색에 관한 향후연구 과제에 대한 자료

간에 0.70으로 높게 나타났다. 그리고 상호간의 평균일치도는 0.61, 복합신뢰도 계수는 0.86을 보임으로써 매우 높게 나타났다. 탐색과제문② (자동색인 기법이 검색효율에 끼치는 영향에 대한 실험결과 자료)에 대한 적합문헌 판정 및 신뢰도 측정 결과는 최종 적합문헌이 14편으로 나타났고, 분석자 B와 D 그리고 C와 D간의 일치도가 다소 떨어지는 경향을 보이고 있으며, 상호간의 평균일치도는 0.52, 복합신뢰도 계수는 0.81로 나타났다. 탐색과제문③(이용자 연구에 있어서 조사방법에 관한 자료)에 대한 적합문헌 판정 및 신뢰도 측정 결과는 최종 적합문헌이 10편으로 나타났고, 상호간의 평균일치도는 0.50, 복합신뢰도 계수는 0.80을 보이고 있다. 탐색과제문④(정보검색의 기술적 측면에 대한 연구동향 자료)에 대한 적합문헌 판정 및 신뢰도 측정 결과는 최종 적합문헌이 12편으로 나타났고, 분석자간의 상호일치도가 다른 탐색과제문에 비해 낮고 특히 A와 B간의 상호일치도는 0.30으로 매우 저조함으로써 상호간의 평균일치도는 0.45, 복합신뢰도 계수는 0.76을 보임으로써 다른 탐색과제문에 비해 다소 떨어지는 경향을 보이고 있다. 탐색과제문⑤(자동색인 기법이 검색효율에 끼치는 영향에 대한 실험결과 자료)에 대한 적합문헌 판정 및 신뢰도 측정 결과는 최종 적합문헌이 14편으로 나타났고, 분석자 B와 D, C와 D간의 상호일치도가 저조함으로 인해 상호간의 평균일치도는 0.49에 그치고, 복합신뢰도 계수는 0.79로 나타났다.

적합문헌 판정 및 신뢰도를 측정해 본 결과 복합신뢰도 계수는 탐색과제문이나 특정 분석자간의 상호일치도에 따라서 최소 0.76에서 최대 0.86에 이르기 까지 다소 차이를 보이고 있다. 그러나 탐색과제문①의 경우에는 0.86, 탐색과제문②의 경우에는 0.81, 탐색과제문③의 경우에는 0.80, 탐색과제문④의 경우에는 0.76, 탐색과제문⑤의 경우에는 0.79로써 North가 제시한 타당성의 기준치(0.70) 보다 모두 높게 나타났다. 따라서 각 탐색과제문에 대한 적합문헌 리스트는

〈표 3〉 실험에 사용한 탐색질의문

탐색문	불리언 탐색질의문*제한탐색(표제 문장 문단 내용구조)
질의1	멀티미디어 AND 정보검색 *C3
질의2	검색효율 AND 자동색인 *A3
질의3	이용자 AND 조사 *A4, *B25, *B26
질의4	정보검색 AND 기술 *A2, *B12
질의5	정보검색 AND 디지털 도서관 *B21

※ 별표(*)는 내용구조에 의한 탐색조건 설정 표시임

높은 객관성을 유지하고 있음을 알 수 있다.

한편, 앞에서 설정한 탐색과제문으로 부터 키워드를 추출하고 〈표 3〉와 같이 불연산기호를 이용한 탐색질의문을 작성하였다.

탐색질의문에 포함된 용어의 다양성이 검색 결과에 미치는 영향을 최소화하기 위해 동의어 및 관련어, 영문 표기, 복합어 등의 변수를 탐색시에 반영하였다. 예를들면, 탐색질의문①의 경우에 '멀티미디어 OR 멀티 미디어 AND *정보검색* OR *정보 검색'과 같이 확장 불리안 탐색문을 통해 탐색어의 다양성을 수용하였다.

탐색질의문의 논리연산에서 논리적(AND)은 지정한 범위내에서 두 용어가 동시에 출현하는 문헌 집합을 의미하며, 논리화(OR)는 일반적인 탐색문에서와 같이 두 용어중 어느 것이라도 만족하는 문헌 집합을 의미한다. 각 탐색질의문마다 5가지 유형-무제한 탐색, 장·절제목 탐색, 문장 탐색, 문단 탐색, 내용단위탐색-으로 구분하여 탐색을 실행하였다.

여기서 무제한 탐색은 아무런 제약없이 탐색어가 문헌 어디에라도 출현하는 문헌을 검색하며, 표제·장·절제목탐색은 표제(TITL), 장(CHTL), 절(SETL), 항(SUTL)의 제목에 탐색어를 동시에 포함하고 있는 문헌을 검색한다. 그리고 내용단위탐색은 지정한 내용단위요소의 범위내에서 두 용어가 동시에 출현하는 문헌을 검색한다. 본 시스템의 평가를 위한 탐색에서는 탐색과제문에 내포된 내용구조적 요소를 고려하여 탐색질의문에 내용단위를 지정하였다. 탐색과제①(멀티미디어 환경에 있어서 정보검색에 관한 향후 연구과제에 대한 자료)은 단위요소 C3(향후과제)를 지정하여 탐색하였고, 탐색과제②(자동색인 기법이 검색효율에 끼치는 실험결과)는 A3(연구과제)를 지정하고, B4(결과)부분을 디스플레이하였다. 탐색과제③(이용자 연구에 있어서 조사방법에 관한 자료)은 A4(연구방법), B25(면접), B26(설문)을 각각 지정하여 탐색하였다. 탐색과제④(정보검색에 있어서 기술적 측면에 관한 연구 동향)는 A2(타연구와의 관계), B12(이론적 배경)부분을 지정하여 탐색하였다. 탐색과제⑤(정보검색 측면에서 본 디지털 도서관의 사례)는 B21(대상)을 지정하였다.

4.2.2 평가 척도

평가척도는 정보검색시스템의 효율척도로 가장 널리 사용되고 있는 정확률과 재현율을 사용하였다. 본 연구의 시스템 평가는 내용구조 기반 전문 탐색 기법이 다른 탐색 기법에 비해 적합 정보의

손실 없이 어느 정도의 정확률, 즉 잡음을 감소 시킬수 있는가를 측정하는데 목적이 있기 때문에 각 탐색질의문별로 잡음률과 누락률을 산출하였다.

전통적인 정보검색시스템인 경우에 문헌의 속성 즉, 서지 정보를 대상으로 문헌 레코드를 검색하기 때문에 전체 적합문헌수와 검색된 적합문헌수를 파악하면 검색효율의 측정이 가능하다. 그러나 전문을 대상으로 탐색하는 경우에는 항상 문헌 단위로 검색하는 것이 아니라 장, 절, 문장 혹은 내용단위로 검색하고, 적합성을 판단할 수 있기 때문에 단순히 해당 문헌의 검색 여부와 그 수량만으로는 그 성능을 평가하기가 어렵다. 전문을 대상으로 하는 경우에는 적합 문헌이라는 개념보다는 색인 대상물(indexable item) 혹은 의미단위체로 구분하여 검색 효율을 측정할 필요성(Jennifer Rowley, 1994)이 제기되고 있는 것은 이를 잘 설명해 주고 있다. 따라서 본 연구에서는 문헌을 탐색단위별로 구분하여 정보요구에 따라서 특정 부분만을 검색할 수 있는 능력을 측정해 보았다. 특정 부분만을 검색할 수 있는 능력의 측정에는 부분정확률이라는 새로운 척도를 이용하였다. 부분정확률의 산출식은 아래와 같으며, 산출식에서 단위수는 문장탐색인 경우에는 문장의 수, 문단탐색인 경우에는 문단의 수, 내용구조 탐색인 경우에는 내용단위요소의 수를 각각 의미한다.

$$\text{부분 정확률}(\%) = \frac{\text{적합한 부분에 포함된 단위수}}{\text{검색된 총 단위수}} X\ 100$$

예를 들어 탐색질의문 'a AND b *문단제한'인 경우에 탐색어 a와 b를 동시에 포함하고 있는 문단이 10개인데 그중에서 정보요구를 만족시키는 문단이 5개 였다면 부분정확률은 50%가 된다.

4.2.3 평가 결과

탐색질의 ①-⑤까지의 탐색유형별 총검색문헌수, 정확률, 재현율, 잡음률, 누락률 등은 〈표 4〉에 정리 되어 있다.

〈표 4〉에서 표제 · 장 · 절제목 탐색, 문장 탐색, 문단 탐색 항목에 포함된 문헌번호와 탐색건수는 "정보검색"과 "멀티미디어"를 표제내에 동시에 포함하고 있는 문헌, 문장내에 동시에 포함하고 있는 문헌, 문단내에 동시에 포함하고 있는 문헌을 각각 의미하며, 내용구조 탐색은 "정보검색"과 "멀티미디어"라는 단어가 C3(향후과제) 범위내에서 동시에 출현하는 문헌을 나타내고 있다. 그리고 탐색결과 문헌은 前項에서 제시한 탐색과제별 적합문헌리스트와 대조하여 적합문헌과 부적합문헌을 선별하고, 정확률, 재현율 등을 산출하였다.

〈표 4〉에서 보는 바와 같이 탐색범위가 넓으면 넓을수록 즉, 제목 · 문장 · 문단 · 무제한으로 탐색 범위를 확대 하면 할수록 평균 총검색문헌수(8건, 17.6건, 28.2건, 38.4건)와 평균재현율(32%, 58.4%, 80.6%, 94%)은 크게 상승하고 있으나 평균 정확률은 47.7%, 36.9%, 33.4%, 27.4%로 현저하게 떨어지고 있다. 이는 자연언어 탐색에 있어서 레코드의 길이가 길어지면 길어질수록 높은 재현율에 비해 정확률이 떨어진다는 기존의 입장을 그대로 입증해 주고 있다.

무제한 탐색의 경우에 평균재현율은 94%로 거의 모든 적합 문헌을 검색할수 있었으나 잡음률이 72.5%에 이르러 대규모 전문데이터베이스를 대상으로 탐색할 때 순위부여 알고리즘이나 추가 필터링 기법을 적용하지 않는한 이용자에게 대량의 검색결과로부터 적합문헌을 선별해야 하는 인지적 부담이 될 수 있음을 보여 주고 있다.

〈표 4〉 탐색질의별 측정결과

탐색유형 / 결과	표제·장·절제					문 장					문 단				
	①	②	③	④	⑤	①	②	③	④	⑤	①	②	③	④	⑤
전체 적합 문헌수	-	-	-	1	-	1	2	2	2	2	3	2	3	2	2
부분 적합 문헌수	5	6	-	-	-	7	9	3	1	2	7	11	6	5	3
부적합 문헌수	3	3	-	6	-	8	17	6	19	7	12	27	13	34	11
총검색 문헌수	8	9	-	7	-	16	28	11	22	11	22	40	22	41	16
정 확 률 (%)	62.5	66.6	-	14.2	-	50.0	39.2	45.4	13.6	36.3	45.4	32.5	40.9	17.0	31.2
재 현 율 (%)	45.4	42.8	-	8.3	-	81.8	78.5	50.0	25.0	57.1	90.9	92.8	90.0	58.3	71.4
잠 음 률 (%)	37.5	33.4	-	85.8	-	50.0	60.8	54.6	86.4	63.7	54.6	67.5	59.1	83.0	68.8
누 락 률 (%)	54.6	57.2	-	91.7	-	18.2	21.5	50.0	75.0	42.9	9.1	7.2	10.0	41.7	28.6

탐색유형 / 결과	무 제 한					내용구조 태그 제한				
	①	②	③	④	⑤	①	②	③	④	⑤
전체 적합 문헌수	3	2	4	2	2	3	2	4	2	2
부분 적합 문헌수	8	11	6	9	4	6	9	5	8	3
부적합 문헌수	23	30	21	48	19	2	3	2	3	2
총검색 문헌수	34	43	31	59	25	11	14	11	13	7
정 확 률 (%)	32.3	30.2	32.2	18.6	24.0	81.8	78.5	81.8	76.9	71.4
재 현 율 (%)	100	92.8	100	91.6	85.7	81.8	78.5	90.0	83.3	71.4
잠 음 률 (%)	67.6	69.8	67.8	81.4	76.0	18.2	21.5	18.2	23.1	28.6
누 락 률 (%)	-	7.2	-	8.4	14.3	18.2	21.5	10.0	16.7	28.6

표제·장·절제목 탐색에 있어서 탐색질의①은 정확률이 62.5%, 탐색질의②는 66.6%, 그리고 탐색질의④는 14.2%로 탐색질의문간에 편차가 크고, 탐색질의문 ③과⑤의 경우에는 단 1건의 적합문헌도 검색하지 못함으로써 탐색어의 특정성과 저자의 장 절 구분패턴이 검색효율에 큰 영향을 끼치고 있음을 알 수 있었다. 탐색질의문① 과 ②만을 대상으로 할 경우에는 정확률이 각각 62.5%, 66.6%로 비교적 높은 수준을 보이고 있으나 67.8%의 누락률을 보여 효율적인 탐색으로 보기에는 어려움이 있었다. 또한 탐색질의④의 정확률은 다른 탐색질의에 비해 현저하게 낮게 나타남으로써 "기술"과 같은 일반적인 용어가 탐색문에 들어갈 때에는 정확률이 떨어 진다는 R. A. Love(1985)의 주장을 재확인 시켜 주고 있다.

문장 문단탐색은 탐색질의①~⑤까지 모두 무제한탐색(평균정확률 27.46%)에 비하면 정확률이 각각 36.9%, 33.4%로 나타나 다소 상승하는 효과를 보이고 있으나 잡음률이 63.1%, 66.6%

〈표 5〉 부분탐색 실험결과

결과 / 질의	탐색유형	단위별 총검색건수	검색된 적합단위수	부 분 정확률(%)
탐색질의1	문 장	16(48)	8(12)	25.0
	문 단	23(54)	8(15)	27.7
	내용단위	11(11)	9(9)	81.8
탐색질의2	문 장	28(67)	11(32)	47.7
	문 단	40(58)	13(21)	36.2
	내용단위	14(14)	11(11)	78.5
탐색질의3	문 장	11(35)	5(18)	51.4
	문 단	22(41)	9(23)	56.0
	내용단위	11(11)	9(9)	81.8
탐색질의4	문 장	22(46)	3(8)	17.3
	문 단	41(72)	7(18)	25.0
	내용단위	13(13)	10(10)	76.9
탐색질의5	문 장	11(28)	4(16)	57.1
	문 단	16(38)	5(15)	39.4
	내용단위	7(7)	5(5)	71.4

※ ()안의 숫자는 문헌내 단위수를 나타냄

에 이르러 적합문헌을 선별해야 하는 인지적 부담이 그대로 남아있다. 논리적 구조에 의한 탐색(문장 문단)은 무제한 탐색에 비하면 6-9%정도의 정확률 상승효과를 보이고 있으나 적정수준이라 보기는 어려웠다.

반면에 내용구조에 의한 탐색의 평균정확률은 78%로써 표제, 문장, 문단, 무제한탐색과 비교하면 각각 30.3%, 38.4%, 44.6%, 50.6%씩 높게 나타나고 있으며, 평균재현율은 81%로써 무제한 탐색(94%) 보다는 낮으나 제목(32%), 문장(58.4), 문단(80.6) 탐색 보다는 높은 결과를 보이고 있다. 여기서 내용구조에 의한 탐색은 잡음률과 누락률을 최소화함과 동시에 높은 수준의 정확률(평균 78%)과 재현율(평균 78)을 보이고 있음을 알 수 있다.

한편 문헌단위에 의한 탐색이 아니라 정보요구에 적합한 특정 부분만을 탐색하는 부분탐색에 대한 성능 비교는 〈표 5〉과 같다. 〈표 5〉에서 보는 바와 같이 평균 부분정확률이 문장인 경우에 39.7%, 문단인 경우에 36.8% 그리고 내용단위인 경우에는 78%로써 문장, 문단에 비하면 38.8%, 41.2%까지 상승하는 효과를 보였다. 탐색질의별 문장, 문단, 내용구조에 의한 부분탐색 결과를 비교해 보면 내용구조에 의한 탐색이 정보요구에 따른 특정 부분탐색에 있어서 다른 탐색 수단에 비해 월등함을 알 수 있다. 이는 표제

장 · 절제목, 문장, 문단과 같은 논리적인 요소를 이용한 탐색에서는 각 문장 혹은 각 문단을 의미내용에 따라서 식별할 수 있는 요소가 없기 때문에 잡음이 많다는 Ingwersen(1996)의 지적을 뒷받침해 주고 있다.

이상의 탐색 실험에서 얻은 결과를 요약하면 다음과 같다. 첫째, 표제, 문장, 문단과 같은 논리적 요소에 의한 탐색은 무제한 탐색에 비해 6-9%정도의 정확률 향상을 보였으나 적정수준에는 이르지 못하였다. 둘째, 내용구조에 의한 제한 탐색은 다른 탐색 수단에 비해 잡음을 최소화함과 동시에 적합정보의 손실없이 많은 적합 문헌을 검색할 수 있었으며, 내용단위와 관련된 특정 부분을 탐색하는데 있어서는 다른 탐색수단에 비해 월등하였다. 셋째, 표제 · 문장 · 문단 · 무제한으로 탐색의 범위가 넓어지면 넓어질수록 총검색문헌수, 재현율은 높아지지만 정확률은 현저하게 떨어진다. 이는 한편의 글이나 문헌이 갖고 있는 내용단위구조는 의사소통에 작용하는 인지적 구조이자 정보요구 상황을 반영하고 있기 때문에 이를 고려한 전문검색시스템은 검색효율을 향상시키고, 다양한 접근점을 제공할 수 있다는 최초 연구가설의 타당성을 뒷받침하고 있다.

5. 결 론

본 연구는 전문검색시스템의 검색효율을 향상시키기 위한 방안으로 문헌의 외형적 구조에 의한 종래의 시스템보다는 문헌의 내용을 단위로하는 전문데이터베이스 표현 방법이 더 효과적일 것이라는 가설하에 문헌의 내용구조에 의한 전문검색시스템을 설계하고, 이에 따라 전문데이터베이스를 구축하여, 검색효율을 평가한 바 그 결과를 요약하면 다음과 같다.

첫째, 표제, 문장, 문단과 같은 논리적구조에 의한 종래의 전문탐색 기법은 문헌 전체를 대상으로 하는 무제한 탐색에 비하면 6-9% 정도의 정확률 향상을 보였으나 잡음률이 52.3, 63.1%, 66.6%에 이르러 적정수준의 정확률에는 미치지 못하였다.

둘째, 내용구조에 의한 탐색은 평균정확률이 78%로써 표제, 문장, 문단, 무제한탐색과 비교하면 각각 30.3%, 38.4%, 44.6%, 50.6%씩 상승하는 효과를 보였고, 평균재현율은 81%로써 무제한탐색(94%) 보다는 낮으나 제목(32%), 문장(58.4), 문단(80.6) 탐색 보다는 높게 나타났다.

셋째, 정보요구에 따라서 특정 부분만을 탐색할 수 있는 평가척도인 부분정확률에 있어서는 내용구조에 의한 탐색이 문장, 문단 탐색에 비해 38.3%, 41.2%까지 상승하는 효과를 보였다.

넷째, 표제, 문장, 문단, 무제한 탐색으로 탐색범위가 넓어지면 넓어질수록 총검색문헌수(평균 8, 17.6, 28.2, 38.4건)와 재현율(평균 32, 58.4, 80.6, 94%)은 높아지지만 정확률은 평균 47.7, 36.9, 33.4, 27.4%로 점차 떨어진다.

이러한 연구결과는 한편의 글이나 문헌이 갖고 있는 내용단위구조는 문헌의 이용행위나 정보요구 상황을 반영하고 있기 때문에 이를 고려한 전문검색시스템은 검색효율을 향상시키고 접근방법의 다양성을 제공할 수 있다는 최초 연구가설의 타당성을 입증해 주고 있다. 따라서 전문데이터베이스를 구축하거나 전문검색시스템을 설계할 때에는 정보검색의 효율성 측면에서 내용 단위 구조화를 고려할 필요가 있다. 이는 현재 전자문헌의 유통, 관리에 널리 이용되고 있는 SGML, HTML, XML과 같은 서식구조를 배척하려는 것이 아니라 다양한 탐색력을 갖는 전문검색시스템

구현을 위해 내용구조를 수용할 필요가 있음을 강조하려는 것이다.

또한 내용단위구조 정보를 갖는 전문검색시스템은 문헌의 내용에 따른 탐색어의 위치 지정, 검색된 문헌의 특정 내용만을 골라서 읽을 수 있는 횡적 브라우징, 내용 단위요소를 노드로 하는 하이퍼텍스트 시스템, 내용상의 위치에 따른 가중치의 부여 등 다양한 활용이 기대된다. 그러나 본 연구는 학술논문만을 대상으로 실험하였기 때문에 향후 다른 문헌의 유형에 대한 내용구조모델을 개발할 필요가 있고, 내용구조를 자동적으로 인식할 수 있는 알고리즘에 대한 연구가 병행되어야 할 것이다.

참 고 문 헌

加藤藤澤. "大規模データベース用テキストサーチマシンの開發." 1991年情報學シンポジウム 豫告集. pp.97-106.

神門典子(Noriko Kando). "SGML 文書による全文データベースのための文法的處理を用いた論理構造の變換手法." 學術情報センタ紀要, 第7號(1995), pp.1-12.

———. "情報メデアの構造: 傳達內容の分析と利用." Library and Information Science, No.30, 1992. pp.1-19.

———. "構成要素かテゴリを用いた原著論文の 內部構造分析." 情報處理學會研究報告, 92(32), 1992. pp.39-46.

石塚 英弘. "SGMLと全文データベース." 情報管理, 37(2), (May 1994). pp.149-159.

猪瀨博他. "文獻の論理構造に基づく全文データベース檢索システムの開發研究." 科學研 究費研究成果報告書, 1993. pp.1-158.

Aitchson, J. and Gilchirst A. Thesaurus *construction*. 2nd. ed., *London* : Aslib, 1987.

Allen, Bryce. "Propositional analysis : a tool for library and information science research." Library and Information Science Research, V.11, 1989. p. 235-246.

Fuller, S. S. Schema theory in the representation and analysis of text. University South. Calif., PH.D thesis, 1984.

Guthrie, J.T. "Location information in documents." Reading Reserch Quartiy,23(2) 1988. pp. 178-199.

Hearst, M. A. and Plaunt, C. "Subtopic *structuring for full-length document* access," Proc. SIGIR 93, Association for computing machinery. pp.59-68.

Ingwersen, Peter. "Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory." Journal of Documentation, 52(1). 1996.

Kintsch, W. and Van Dijk, T.A. "Toward a model of text comprehension and

production." Psychology Review, 85(5), 1978. p.363-394.

Liddy, Elizabeth Duross. "The discourse level structure of empirical abstracts : an exploratory study." IPM, 27(1) 1991 p. 55-81.

——. The discourse-level structure of natural language texts : an exploratory study of empirical abstacts. Unpublished Ph.D. dissertation, Syracuse Uni. School of Information studies, Syracuse, NY, 1988.

Love, A. R. "Precison in Searching the Full-Text Database-ACS Journal Online." In: National online meeting proceedings-1985. Medford : Learned Information, 1985. pp.273-282.

Mittendorf, E. & Schäuble, P., "Document and passage retrieval based on Hidden Makov model. In Proceedings of the ACM Int. conference on R&D in IR(SIGIR), 1994. pp.318-327

Moffat, A., Sacks-Davis, R., Wilkinson, R. and Zobel, J. " Retrieval of partial documents." TREC-2 Proceedings, 1993.

O'connor, John. "Answer-passages retrieval by text searching." Journal of the American Society for Information Science. 31(3), 1980. p.227-239.

Oddy, R. N. "Toward of the Use of Situation Information in Information Retrieval." Journal of Documentation, 48(2), 1992. p.131-141.

Rowley, Jennifer "the Controlled Versus Natural Indexing Languages Debate Revisited: a Perspective on Information Retrieval Practice and Research." JASIS, 20(2), 1994. p.110.

Rumelhart, D. E. Notes on a schema for stories:representation and understanding.

New York : Academic Press, 1975. p. 211-236.

Salton, G. Allan, J. and Buckley, C. "Approaches to passage retrival in full text information systems." Proc. SIGIR-93, Association for Computing Machinery, New York,1993, p. 49-58.

Salton, G., Buckley C. and Singhal, A., "Automatic analysis, theme generation and summarization of machine-readable texts." Science, 264, 1994, p.1421-1426.

Salton, G. Automatic text processing : the Transformation analysis and retrival of information by computer. MA : Addison Wesley Publishing Company, 1989.

Salton, G. and Singhal, A. Automatic text theme generation and the analysis of text structure. technical report TR94-1438, Computer Science Department, Cornell University, 1994.

Van Dijk, Teun A. Macrostuctures : an Interdisciplinary study of global structures in discourse, interaction, and cognition. Hillsdale : Lawrence Erlbaum, 1980.

Van Dijk, Teun A. Kintsch, Walter, Strategies of discourse comprehension. New York : Academic Press, 1983.

Woodhead, Negel. "Hypertext and hypermedia theory and application". Wilson : Sigma Press, 1991.

Zobel, J. and Moffat, A. "Efficient retrieval of partial document". IPM 31(3) 1995. pp.361-377.