문헌의 내용단위구조에 의한 전문검색시스템의 타당성 고찰*

A Study on the Feasibility of Full-Text Information Retrieval System Based on Document Content Structure

이 병 기(Byeong-Ki Lee)**

목 차

초 록

전통적인 전문검색시스템은 전문색인방식을 취하고 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. 따라서 본고에서는 전문데이터베이스의 본문을 내용단위로 구조화해야 할 이론적 타당성을 검토하였으며, 선행연구에 나타난 학술논문의 내용단위구조를 비교 분석하여 내용단위요소를 추출하고, 실제 한국어로 작성된 180여건의 학술논문에 적용 검토함으로써 표준적인 학술논문의 내용구조 모델을 개발하였다. 그 결과 문헌의 내용단위구조는 이용자의 정보요구 상황이나 목적과 밀접하게 관련되어 있기 때문에 전문데이터베이스 구축이나 전문검색시스템의 설계시에 내용단위로 구조화해야할 필요성을 도출하였다.

ABSTRACT

In these days the online full-text database are increasing, but 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. Therefore, the purpose of the study examined feasibility of document content structure in full-text IR system by resolving such disadvantages of conventional system. 180 journal articles have been analyzed to find common structure of document content and finally general model of the structure of journal articles were developed. The result shows that have relation to between user's cogntive schema structure, user's information need and contents structure of document. Thus it is concluded that full-text IR system need to be designed by using document content structure in order to meet user's information need more effectively.

* 이 연구는 중앙대학교 대학원 박사학위 논문을 부분 축약한 것임.
** 중앙대학교 강사, 공항고등학교 사서교사
접수일자 1998년 3월 18일

1. 서 론

최근들어 출판 과정의 전자화, 고밀도 대용량 기록 매체의 등장, 저렴한 통신료 등으로 인해 2차 정보 뿐만아니라 원문 자체를 수록한 전문데이터베이스(full-text database)가 급증하고 있다. 초기의 전문데이터베이스는 주로 법령 및 판례, 신문기사와 같이 본문의 길이가 짧거나 정형화된 텍스트만을 대상으로 하였으나 점차 단행본, 잡지기사, 학술논문 등 다양한 문헌 형식에 적용되고 있으며, 온라인 전문검색시스템은 물론 CD-ROM과 같은 오프라인방식의 전문데이터베이스 및 인터넷의 클라이언트/서버 환경을 이용한 全文 情報의 이용이 보편화 되고 있다. 특히, 최근에는 디지털도서관에 대한 관심이 고조 되면서 전문데이터베이스의 구축 방안과 전문탐색기법에 대한 연구가 활발히 진행되고 있다.

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. 다시말해서 기존의 전문검색시스템은 문헌 내용의 구조적 특성을 고려 하지 않고, 문헌 전체를 대상으로 단어의 출현여부나 출현빈도 혹은 문헌의 외형적인 구조에만 의존하여 탐색하고 있다. 따라서 이용자의 정보요구에 해당되는 본문의 특정 부분만을 지정하여 탐색함으로써 검색효율을 향상시킬 수 있는 방안에 대한 연구의 필요성이 제기되고 있다.

이에 본 연구에서는 현행 전문검색시스템의 문제점을 해결하기 위한 방안으로 문헌의 본문을 일정한 단위로 구조화해야 하며, 특히 이용자의 정보요구 및 全文의 브라우징 행위를 고려한 내용단위구조 기반 전문검색시스템의 타당성을 검토하고 실제 전문검색시스템을 설계할 때 필요한 표준적인 내용단위구조 모델을 설정하는데 그 목적이 있다.

우선 인간의 의사소통 단위를 연구하는 텍스트언어학과 정보의 인지과정을 다루는 인지 심리학적 이론을 바탕으로 문헌의 내용단위구조가 어떠한 특성이 있으며 全文의 이용이나 의사소통 과정에서 어떻게 작용하고 있는지 고찰하고, 이용자의 정보요구와 내용단위구조간에 관계를 규명함으로써 내용단위구조를 전문검색시스템에 고려해야 할 타당성을 밝히고자 한다. 또한 문헌의 내용단위구조에 의한 전문검색시스템을 구현하기 위해서는 이용자들이 공통적으로 인식할 수 있는 표준적인 내용구조 모델이 필요하다. 따라서 학술논문을 대상으로 내용구조 모델을 개발코자 한다.

문헌의 구조적 특성을 이용하여 전문데이터베이스를 구축하거나 전문검색시스템을 구현하려는 지금까지의 연구는 대부분 문헌의 외형적인 구조의 적용에 그치고 있으며, 정보내용이 갖고 있는 구조의 이용에는 이르지 못하고 있다. 문헌의 구조적 특성을 전문데이터베이스 구축이나 탐색에 적용하려는 연구로는 Macleod(1990)의 "장·절, 문장, 문단구조에 의한 정보검색에 관한 연구"와 Blake(1994)의 "SGML을 통한 구조적 질의

에 관한 연구"가 있다. 또한 Dewire(1994)는 "논리적구조를 이용한 전문데이터베이스 구축 과정에 관한 연구"를 수행한 바 있다. 국내에서는 유석종(1995)의 "논리적 구조를 이용한 색인에 관한 연구"와 이택경(1996)의 "SGML 검색엔진에 관한 연구" 등이 있다. 이와같이 문헌구조를 전문데이터베이스 구축이나 탐색에 이용하려는 연구는 국내외를 막론하고 대부분 문헌의 외형적인 구조에 의존하고 있으며, 문헌의 내용구조를 이용한 전문검색시스템 개발에는 이르지 못하고 있다.

내용단위구조에 의한 전문검색시스템의 구현(CSML; Content Structure Mark-up Langeuag Full-text Information system)과 이에 대한 검색효율 측정에 관한 연구는 차후에 발표할 예정이다.

2. 전문검색시스템의 문제점

2. 1 전문색인의 구조와 탐색

전문데이터베이스는 신문기사나 법원의 판결문, 잡지기사, 교과서, 백과사전과 같은 1차 정보원의 원문이나 데이터 전체를 수록한 데이터베이스를 말하며, 전문탐색은 이러한 전문데이터베이스를 대상으로 각 문헌의 본문에 포함되어 있는 모든 단어(보통 기능어는 제외)를 탐색어로 사용하여 전문이나 전문의 일부를 탐색할 수 있는 기법을 말한다.

전문데이터베이스의 파일구조 및 탐색방식은 기존의 서지데이터베이스와 매우 유사하다. 보통 전문데이터베이스 역시 문헌파일과 도치파일로 구성되며, 문헌파일에는 표제, 저자, 디스크립터, 초록, 본문 등의 필드로 세분된다. 다만 전문데이터베이스에는 본문필드가 존재한다는 점 그리고 도치파일에는 문헌번호 뿐만아니라 본문에서 추출한 각 단어의 위치정보가 관리됨으로써 디스크립터가 아닌 본문에 포함되어 있는 모든 단어를 통해 검색할 수 있고, 블리언 연산자 이외에 인접지정, 순서 지정, 어간거리지정, 문장제한, 문단제한등 다양한 위치연산자(proximity or positional operator)가 사용된다는 점에서 차이가 있다.

전문탐색 기능은 KMP, Boyer-Moore와 같은 매칭 알고리즘을 통해 키워드 문자열과 텍스트 문자열을 순차적으로 검사하는 문자열 탐색(string search)이나, 전문데이터를 일정한 방식으로 압축한 요약파일을 작성하고 이를 대상으로 탐색하는 패턴탐색(加藤藤澤, 1991)도 있으나, 일반적으로 대규모 전문데이터베이스인 경우에는 전문색인을 통해 탐색이 이루어 진다. 전문데이터베이스는 일반적으로 고속 접근과 문단, 문장탐색을 위해 4가지 위치정보 즉, 용어가 있는 문헌번호, 문헌내의 문단 순차번호, 문단내의 문장 순차번호, 문장내의 단어 순차번호를 이용하여 색인을 작성한다

이와같이 본문내의 모든 단어를 이용하여 탐색할 수 있는 전문검색시스템은 주제색인의 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; A Gilchirst, 1987) 이외에 전문데이터베이스의 구조적 표현과 관련된 2가지 문제점이 있다.

첫째, 높은 재현율에 비해 불필요한 문헌이 과다 출력됨으로써 정확률이 현저하게 떨어진다는 점이다. 탐색 대상인 全文은 서지, 초록과 같은 문헌대용물에 비해 많은 탐색어휘를 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(N. Woodhead, 1991). 전문검색시스템의 검색효율을 측정한 C. Tenopir(1988)와 J. S. Ro(1988)및 D. C. Blair(1990) 등의 연구에 의하면 전문검색시스템의 검색효율은 데이터베이스의 규모, 탐색자의 배경지식, 주제, 언어, 자료의 유형 등 여러가지 변수에 의해 영향을 받고 있으나, 레코드의 길이가 길어지면 길어질수록 자연언어 탐색에서는 높은 재현율에 비해 정확률이 현저하게 떨어진다는 문제가 있음을 지적하고 있다.

현행 전문검색시스템의 두 번째 문제점은 문헌구조에 대한 고려없이 문자열의 단순출현이나 형식적인 구조(예를들면 물리적 구조 혹은 논리적 구조)에 의존하여 탐색하고 있기 때문에 내용과 관련된 정보요구 상황이나 목적에 따라서 본문의 특정부분만을 탐색할 수 없다. 한 편의 논문이나 문헌은 항상 전체를 통람하는 것이 아니라 정보요구의 상황이나 목적, 문헌 유형에 따라서 특정 부분만을 필요로 하거나 읽는 순서가 달라지며(Andrew Dillon; J Richardson, 1988), 문헌 전체가 아닌 특정 부분만을 보고 적합성을 판단하는 경우가 많다. 이는 이용자가 필요로 하는 부분만을 탐색, 제공할 수 있는 정보검색시스템이 이용자의 정보요구를 보다 더 효율적으로 충족시켜 줄 수 있다는 가능성을 제시하고 있다.

그러나 현행 전문검색시스템은 문헌구조에 대한 고려 없이 본문 전체를 대상으로 하는 전문색인방식을 취하거나 구둣점이나 문단과 같은 형식적인 구조에 의존하여 탐색하고 있기 때문에 내용과 관련된 이용자의 특정 목적이나 상황에 따라서 본문의 특정 부분을 지정하거나 탐색하는데 한계가 있다. 이는 메타데이터를 추출하여 시스템을 구축하는 서지 초록형 데이터베이스와 같이 전문을 하나의 독립된 필드로 간주하고 본문 필드에 특정 용어의 출현여부에 따라 탐색하는 거시적인 색인 및 검색기능에서 벗어나지 못했기 때문이다(影浦 峡, 1990). 따라서 이용자의 정보요구상황이나 문헌의 이용목적에 따른 브라우징 행위을 고려하여 전문을 일정한 단위로 지정하여 탐색하고 이용할 수 있는 전문검색시스템의 필요성이 제기되고 있다.

그동안 전문검색시스템의 문제점을 해결하기 위한 방안으로 탐색시소러스, 어근절단, 위치연산자, 제한탐색 등 다양한 기법이 제시되고 있으나 전문데이터베이스의 본문을 일정단위로 나누고 구조정보에 의한 색인을 통해 검색효율을 향상시키려는 연구는 크게

① 형식적구조에 의한 문헌구조화와 ② 유사성에 의한 문헌구조화로 나눌 수 있다.

여기서 형식구조에 의한 문헌구조화는 외형적으로 식별이 가능한 요소(예를들면, 페이지, 구둣점, 레이아웃, 글자크기, 장 절의 제목, 도표, 문단, 문장 등)를 이용한 탐색기법을 말하고, 유사성에 의한 탐색은 용어의 출현 빈도를 연산하여 일정단위로 구분하고 이를 탐색에 이용하려는 방법을 말한다. 형식구조와 유사성에 의한 문헌의 구조화와 탐색방식의 특징을 개략적으로 고찰함으로써 두 기법만으로는 위에서 지적한 전문검색시스템이 갖는 문제점을 해결하는데 한계가 있고, 새로운 형태의 문헌구조화 기법이 필요함을 제시코자 한다.

2. 2 형식적구조에 의한 문헌구조화

한 문헌을 형식적 구조로 구분하여 전문데이터베이스를 구축하려는 방안으로 널리 활용되고 있는 것은 물리적구조와 논리적구조가 있다. 물리적구조를 정보검색시스템에 적용하려는 연구는 특정 글자체 혹은 크기가 다른 단어 등 저자의 강조 표시 어구에 높은 가중치를 부여하는 색인시스템이 있으나 주로 특정 편집 시스템이나 응용 시스템에서 사용하고 있으며, 일정 기준 이상의 크기를 갖는 문단은 둘로 나누고 일정 기준 이하의 문단은 합치는 최소 최대 문단 기법, 1000바이트를 한 페이지 단위로 분할하고 이를 단위로 탐색하는 방법(J. Zobel; A Moffat, 1995), 통계적 기법을 이용하여 30단어를 단위로 구조화하는 방법(C. Stanfill; D. Waltz, 1992)등 인위적인 물리적 단위를 전문탐색에 적용하려는 시도가 있으나 실험적인 단계에 그치고 있다.

이와는 달리 논리적구조는 표제, 장, 절, 문단, 도표와 같이 서식에 의한 단위요소로 구조를 파악하려는 것으로써 현재 전자문헌의 형식구조에 의한 전문데이터베이스 표현수단으로 가장 널리 이용되고 있으며, 표준범용마크업언어(SGML)로 표준화 되어 있다. 또한 디지털문헌의 구조적 표현하면 SGML을 의미할 정도로 보편화되어 있기 때문에 여기에서는 SGML에 의한 전문탐색 방안을 구체적으로 살펴봄으로써 형식적구조에 의한 탐색의 한계점을 밝히고자 한다.

마크업(mark-up)이란 말은 원래 출판분야에서 편집자가 서체나 활자 크기 혹은 도표 위치 등에 관한 지시 사항을 원고상에 기록하는 것을 뜻하였으나 현재는 문헌의 구조를 표시하는 태그를 뜻한다. 마크업에는 레이아웃, 폰트와 같이 물리적인 구조를 표현하는 절차적마크업과 장 절 주기 문단과 같이 문헌의 서식구조를 중심으로 표현하는 기술적 마크업이 있다. 그 중 SGML은 기술적마크업 언어로써 전자문헌의 논리구조 즉, 어떤 부분이 제목이고, 어떤 부분이 章이고 章題인지 혹은 도표이고 도표 제목인지 등을 어떤 시스템이라 하더라도 동일하게 인식할 수 있도록 각 요소에 태깅하는 방식을 정한 "국제규격(ISO 8879, 1986)을 말한다.

기존의 전문데이터베이스는 문헌구조에 대한 정보를 갖고 있지 않음으로써 효율적이지 못하고 다양한 검색을 제공하지 못한다는 문제의식하에 SGML을 이용한 전문탐색 기

법이 제기되고 있으며, SGML을 이용한 전문데이터베이스 구축과 탐색엔진에 대한 연구가 활발하게 진행되고 있다.

SGML을 이용한 전문검색시스템은 문헌구조에 대한 고려없이 문헌 전체를 대상으로 한 전문탐색에 비하면 SGML의 단위요소에 따른 탐색 범위지정이나 특정 단위요소에 출현하는 용어에 더 높은 가중치를 부여함으로써 어느정도 검색효율을 향상 시킬 수 있음이 밝혀지고 있다(野來道子, 1994). 그러나 이 논리적구조는 앞서 언급한 전문검색시스템의 문제점을 해결하는데 있어서 다음과 같은 한계점을 갖고 있다.

첫째, 이용자가 궁극적으로 관심을 갖고 있는 내용과 관련된 구조를 반영하지 못하고 있다. 논리적 구조는 내용에 의한 단위 구분이 아니라 저자가 내용을 전개하는데 사용한 서식 구조에 불과하다. 다시말해서 서론, 본론, 결론, 연구방법, 연구목적, 결과, 고찰등과 같은 내용상의 기능과 역할에 의한 구조가 아니다. 따라서 문단, 문장, 제목, 도표 등과 같이 서식 구조상의 구분은 가능하지만 '연구 목적 부분에 A라는 용어가 포함된 문헌을 탐색하라'와 같이 이용자가 궁극적으로 관심을 갖는 내용을 단위로한 범위 지정은 불가능 하다는 것이다.

둘째, 탐색범위 설정에 널리 이용되고 있는 형식구조의 단위요소인 문장과 문단은 문헌내에서 각 문장 혹은 문단의 역할 및 자질을 식별할 수 있는 요소가 없다. 이용자가 관심을 갖는 특정 부분을 지정하기 위해서는 단위간에 식별이 가능해야 한다. 예를들면 '서론 문단' '고찰 문단' '배경 문단'등으로 구분할 수 있어야 특정 문단을 지정하여 탐색할 수 있으나 SGML은 각 문장이나 문단을 식별할 수 있는 요소를 갖고 있지 않다. 그 이외에도 "문헌마다 논리적 구조가 통일되어 있지 않기 때문에 일관성 있게 탐색범위를 지정하기가 어렵다는 문제가 있다"(高須順宏, 1993). 예를들어, 유사한 내용을 단일 항목으로 종합하여 기술하는 저자가 있는가 하면 여러 항목으로 구분하여 기술하는 저자도 있으며, 그 구분의 계층 수준 또한 다양하기 때문에 이용자가 탐색 범위를 설정하는데 어려움이 있다.

이와같이 논리적 구조는 ISO 8879 (SGML)에서 밝히고 있는 바와 같이 서로 다른 시스템간에 효율적으로 문서를 유통, 관리, 처리하는데 주요 목적이 있기 때문에 이용자가 관심을 갖고 있는 실제 내용구조와는 거리가 있다. 따라서 논리적 구조는 내용과 관련된 이용자의 정보요구를 수용하는데 한계가 있으며, 정보검색에 적합한 구조라기보다는 문서의 표준적인 유통을 위해 개발된 것이라 말할 수 있다. 그동안 논리적구조를 전문검색에 응용하려는 기존의 연구는 SGML을 이상적인 문헌구조화 방안으로 본 것이라기 보다는 문서의 유통과정에서 이미 구현된 SGML 파일을 그대로 전문데이터베이스의 구축에 활용해 보려는 시도로 이해할 수 있을 것이다.

2. 3 유사성에 의한 문헌구조화와 부분탐색

유사성에 의한 부분탐색은 한 문헌 전체를 대상으로 적합성 여부를 판단하는 2분법

적 탐색이 아니라 유사성을 측정하여 일정 기준 이상의 문헌 혹은 문헌의 일부를 탐색하려는 것이다(G. Salton; J. Allan; C. Bukley, 1993). 문헌의 일부분을 탐색, 제공함으로써 이용자의 정보요구를 보다 더 효과적으로 충족시킬 수 있다는 가정하에 인접한 수개의 문장을 단위로하는 부분탐색이 제기된 이후 다수의 유사성 탐색기법이 제기되고 있다.

문헌의 특정 부분에 대한 명칭은 연구자들사이에 subpart, fragments, subtopic, segmentation, subtext, passage, texttiling 등 다양한 용어를 사용하고 있으나 그 의미에는 큰 차이가 없다. 유사성에 의한 부분 탐색은 대부분 문헌에 포함된 각 용어의 출현 패턴에 따른 가중치 용어 집합에 의해 문헌의 부분을 텍스트 벡터로 표현하고 벡터 산을 통해 유사성을 산출하여 기준치 이상의 문헌이나 문헌의 일부를 탐색한다. G. B. Salton(1991)등은 전형적인 벡터공간모델을 이용하여 전통적인 유사성 측정기법을 확대하여 특정 부분탐색이나 자동구조분할, 자동요약 등에 적용하려는 일련의 실험을 전개하였다.

또한 M. A. Hearst(1993)등은 全文을 응집력있는 多文段單位(multiparagraph)로 자동으로 분할함으로써 문헌 전체뿐만아니라 소주제문을 탐색할 수 있는 알고리즘을 개발하고 이를 텍스트타일링(Text Tiling)이라 하였다. 여기서 응집력이란 통일된 의미단위로써 Hearst는 어휘의 연결성을 이용하여 구분하고 있다. 텍스트타일링 알고리즘은 2단계 과정을 거쳐 수행된다. 우선 텍스트를 3-5문장으로 블록화하고, 인접한 두쌍의 블록을 비교하여 유사도를 측정한다. 블록간의 유사도는 다음과 같은 코싸인 계수를 통해 산정하고 있다.

$$sim(b_1,b_2)=\frac{\sum_{t=1}^{n} W_{t,b1}\,W_{t,b2}}{\sqrt{\sum_{t=1}^{n} W_{t,b1}^2 \sum_{t=1}^{n} W_{t,b2}^2}}$$

여기서 t는 블록내의 용어의 범위, Wt,b1은 블록 b1에서 용어 t에 할당된 가중치를 의미한다. 또한 측정한 유사도를 그래프로 표현하여 최고점과 최저점을 체크한다. 여기서 최고점은 가장 높은 유사도 측정치로써 인접한 블록간에 응집력이 있음을 나타내고, 최저점은 역으로 가장 낮은 유사도 측정치로써 이를 통해 텍스트타일간의 경계를 정한다. 문장을 중심으로 토큰화(tokenization)하여 블록으로 삼은 뒤 두 블록간의 유사성을 측정, 일정 기준 이상의 유사성이 있으면 동일한 소주제문, 유사성이 없으면 이질적인 소주제문으로 간주하고 있다.

Elke Mittendorf와 Peter Schüble(1996)은 문헌을 일정 단위로 구조화하기 위한 방안으로 문헌을 확률과정(stochastic process)으로 정의하고 문헌이 생성되는 두가지 확률과정을 제시하고 있다. 첫 번째 확률과정은 특정 질의에 적합한 분절문을 생성하는 것이고, 두 번째 확률과정에서는 특정 질의와 독립적인 분절문(segments)을 생성하는 과정으로 구분하여 문헌 구조를 모형화하였다. 이러한 두가지의 확률과정에 의한 분절문의 생성은 은닉마코프모델(HMM; Hidden

Markov model)에 기초하고 있으며 확률과정에 의해 분절문마다 확률값이 부여되고 확률값의 크기에 따라 분절문을 검색하려는 것이다.

이상과 같이 유사성에 의한 문헌구조화 및 탐색방법은 문헌 전체가 아닌 특정 부분에 대한 중요성을 인식시켜 주고 있으며, 문헌부분별 적합성 랭킹의 가능성을 제시해주고 있으나 어휘의 발생빈도나 확률과정에 의존함으로써 문헌의 내용구조를 반영하지 못하고 있으며 내용구조에 따른 이용자의 정보요구를 수용하는데 한계가 있다.

형식적구조나 유사성에 의한 부분탐색의 특징을 살펴본 바 이러한 탐색기법만으로는 정보요구의 상황이나 목적에 따라서 본문가운데 특정 내용만을 지정하여 탐색하는데 한계가 있음을 확인하였다. 이러한 문제점을 해결하기 위해서는 문헌의 본문 가운데 특정 부분간에 식별이 가능한 단위로 구조화하여 이용자의 정보요구와 관련된 부분만을 지정하고 탐색할 수 있는 표현 수단이 필요함을 알 수 있다. 이는 SGML과 같은 형식구조와 유사성에 의한 탐색기법을 전적으로 부정하는 것이 아니라 이용자의 정보요구를 최대한 반영할 수 있도록 전문데이터베이스를 다양하게 표현할 필요가 있음을 제기하는 것이다.

3. 전문검색과 문헌 내용단위 구조와의 관계

3. 1 내용단위구조의 인지적 측면

한 편의 글이나 문헌에는 전달하려는 정보 내용을 갖고 있으며, 이 "정보내용은 처음부터 끝까지 한 덩어리로 되어있는 것이 아니라 소단위 내용들이 유기적으로 연결되어 전체를 이룬다"(김영채, 1995). 이와 같이 한편의 글 혹은 문헌의 내용을 일정한 소단위로 구분하고 소단위 내용들이 이루는 관계양상을 내용구조라 할 때 전체 내용을 어떤 소단위로 구분하고 소단위간의 관계를 어떻게 표현할 것이냐하는 문제는 내용구조 분석에 있어서 핵심적인 요소이다. 내용에 따른 소단위 설정이나 이 소단위간의 관계 양상은 분석 목적 및 방법에 따라서 다양하다.

M. Halliday와 R. Hasan(1976)은 글의 종류에 관계없이 일반적인 문장을 최소 내용단위로 보았고, 설명적인 글을 분석한 B. Meyer와 W. Kintsch등은 인간이 어떤 정보내용에 대해 진위를 판정할 수 있는 최소 단위인 명제를 대상으로 내용구조를 분석하고 있으며, 이야기를 대상으로 분석한 D. E. Rumrlhart는 어휘적 범주를 분석단위로 삼기도 하였다(이삼형, 1993). 소단위간의 관계표현에 있어서도 원인-결과, 비교-대조, 질문-대답, 문제-해결, 신정보-구정보, 부분-전체, 레마-테마등 무수히 많은 부분간의 관계를 설정하고 있다(Jiri Janos, 1979). 이와 같이 분석의 방법이나 목적에 따라서 다양한 소단위 설정이나 관계 표현이 있으나 전통적으로 언어학 분야에서는 문장을 내용분석의 기본단위로 설정하고 있다. 그러나 문장중심의 단위 설정이나 이들간의 관계 양상을 전문데이터베이스의 내용구조화 수단으로 적용하기에는 다음과 같은 한계점이 있다.

첫째, 전달코자 하는 내용 혹은 메시지의 의미는 문장의 구조나 단어들에 의해 전달되는 것이 아님에도 불구하고 "문장중심 내용구조는 언어적 표현에만 의존함으로써 화용론적인 측면과 사회 문화적인 측면이 고려되지 않았다는 점이다(G. Brawn; G. Yule, 1983). 즉, 실제 언어사용 상황에서 이루어지는 의사소통 과정은 문장들이 모여서 이루는 더 큰 단위에 의해 이루어 지기 때문에 문헌의 전체 내용을 의사소통 단위로 분석할 때에는 문장 이상의 단위 설정이 필요한 것이다.

둘째, 내용구조를 분석하기 위해서는 소단위들간의 관계를 설정해야 하는데, 문장수만큼 단위간의 관계 표현이 많아지기 때문에 실제 분석과 적용이 어렵다. 실제로 각 문장들간에 관계 분석이 가능하다 하더라도 수많은 관계 표현을 이용자에게 제시해야 하기 때문에 전문검색시스템의 인터페이스로도 적합하지 않다.

따라서 전문데이터베이스의 내용구조화 수단으로 사용하기 위해서는 실제 언어사용 환경에서 작용하는 언어단위와 이들간의 관계를 설정할 필요가 있으며, 전문검색시스템의 인터페이스로도 적절해야 할 것이다. 실제 언어사용 환경에서 작용하는 단위와 관계표현에 관한 연구로는 정보내용의 소단위를 텍스트로 설정하고 텍스트간의 관계를 통해 내용구조를 분석하려는 텍스트언어학(textlinguistic)과 글의 생성 및 이해에 대한 인지과정에 관심을 갖는 인지심리학적 관점을 들 수 있다.

3. 1. 1 텍스트단위 내용구조

정보내용의 소단위를 텍스트로 설정하고 텍스트간의 관계를 통해 내용구조를 분석하려는 텍스트단위 내용구조를 이해하기 위해서는 먼저 문자로 쓰여진 기록물이라는 일상적인 의미를 벗어나 내용단위로서의 텍스트에 대한 개념과 출현 배경을 먼저 살펴 볼 필요가 있다.

언어의 본질적인 기능은 의사 소통에 있다. 기록에 의해 일정한 내용을 전달하려는 문헌 또한 언어의 본질적인 기능에서 벗어나지 않는다. 그러나 前項에서 지적한 바와 같이 기존의 언어학에서는 문장을 가장 큰 언어적 단위로 분석함으로써 문장 이상의 언어적 기능을 규명하는데 한계가 있었다. 다음의 예문은 이러한 사실을 단적으로 보여 주고 있다.

> 예문 "탐색자들이 수행한 탐색 성과를 탐색 전략, 탐색 노력, 탐색결과 변수별로 살펴 보면 다음과 같다."

위의 예문을 읽고 우리는 이 글이 의미 전달을 위한 완전한 단위가 아님을 알 수 있다. 뒤에 무엇인가가 덧붙여질 것임을 느끼게 되기 때문이다. 위의 예문 한 문장만으로는 완전한 의사소통이 불가능하고 분석될 수 없음을 보여 주고 있다. 이와 같이 인위적이며 고립된 문장에 대한 분석에 치우친 기존 언어학에 대한 반성과 "인간이 실제 환경에서 사용하는 언어단위는 문장 이상의 텍스트라는 문제의식하에 1970년 초반부터 텍스트언어학이 등장하였다."(윤석민, 1993)

텍스트언어학자들간에도 텍스트에 대한 개념 정의는 매우 다양하다. 다수 문장의 응집된 연쇄, 생성자와 수용자의 의도에 따라 언어적으로 완결된 언어단위, 주제를 가지고 정리된 이어진 글 등 통일되어 있지 않으며 (이영수, 1993), 국내에서도 텍스트라는 말 대신에 이야기, 文章群, 문장떼, 글월 등 다양한 용어로 사용되고 있다. 또한 문장의 연속체가 하나의 텍스트가 되기 위한 요건 즉, "텍스트성(textuality)-응집구조, 응집성, 의도성, 수용성, 상황성, 상호텍스트성, 정보성"-을 들어 텍스트의 개념을 설명하기도 한다 (R. A. Beaugrande; W. U. Dressler, 1984). 텍스트가 갖추어야 할 요건에 의하면 텍스트란 언어적 형식(응집구조, 응집성) 뿐만 아니라 텍스트의 생성자와 수용자의 심리적 측면(의도성, 수용성) 그리고 사회적 상황성 등 의사 소통에 필요한 모든 요인이 복합적으로 작용하고 있음을 알 수 있다.

이와 같이 텍스트에 대한 개념은 명확하게 정의하기 어려우나 공통점을 추출해 보면 텍스트란 문장의 집합체로써 의미적으로 완결되고 일정한 의사소통 기능을 갖는 내용단위라 할 수 있다. 텍스트단위 내용구조는 기존 언어학처럼 고정된 길이의 언어 단위로 분리하거나 문장의 구문 규칙을 규명하는 등 언어의 정적인 체계와 구조에 대한 이론이 아니라 인간이 의사소통을 목적으로 활용하는 자연적인 단위에 관심을 갖는다.

한편, 특정 문헌을 읽다보면 어떤 부분은 역사적 배경, 어떤 부분은 문제제기 혹은 가능한 해결책 등 문헌의 세부 내용이 맡고 있는 역할이나 기능에 따라서 소단위로 구분할 수 있으며, 이러한 소단위를 텍스트단위로 설정하고, 텍스트간의 관계를 통해 내용을 분석할 수 있음이 제기되고 있다(Patrik Wilson, 1978). 예를들어 학술논문인 경우에는 '연구방법에 해당하는 텍스트', '연구 목적에 해당하는 텍스트', '결과에 해당하는 텍스트' 등으로 구분할 수 있으며, 신문기사의 경우에는 누가에 해당하는 텍스트, 언제에 해당하는 텍스트 등으로 구분할 수 있으며, 이들간의 관계 즉, 〈목적-방법-결과〉, 〈누가-언제〉와 같은 방법으로 내용구조를 표현할 수 있다는 것이다.

T. Van Dijk(1980)는 학술보고서를 대상으로 내용을 분석한 결과 (배경-문헌리뷰-목적-방법-결과-결론)과 같이 기능, 역할에 따라서 텍스트단위 내용구조화를 시도하였으며, 문헌유형 혹은 장르에 따라서 특정 역할을 담당하는 고유 텍스트단위요소가 포함되어 있다는 결론을 제기하면서 이를 문헌의 상부구조(superstructure)라 하였다. Niroko Kando(1992) 또한 특정 역할을 담당하는 각 부분을 내용구성요소라하고 내용구성요소간의 관계를 기능구조라 하여 테스트단위에 의한 내용구조를 제시하고 있다. 그 이외에도 Meyer(1980)는 최상위구조, Kintsch(1983)는 거시구조라는 표현으로 텍스트 단위에 의한 내용구조를 논증하고 있다.

이러한 연구는 공통적으로 문헌의 전체 내용을 텍스트단위로 구분할 수 있으며, 각각의 텍스트는 특정 역할과 기능을 갖고 있기 때문에 상호 텍스트간에 식별이 가능하다는 것이다. 또한 문헌의 유형에 따라서 포함

되어야 할 전형적인 텍스트단위 혹은 텍스트 단위구조가 존재한다는 것이다. 현재 텍스트 언어학 분야에서는 학술논문, 신문기사, 광고문, 매뉴얼 등 특정 유형의 문헌을 대상으로 전형적인 텍스트 단위 내용구조 분석이 활발하게 전개되고 있다.

텍스트단위와 관련된 또 다른 문제로써 문장 이상의 내용단위로 인식하고 있는 문단 혹은 단락과의 관계 설정이다. 문단은 주제의 일부, 하위 개념을 집중적으로 펼치는 일련의 문장들로 엮어진 조직체로서 그 형식이 명확히 구획된 글 속의 글로 정의하고 있다. 그렇다면 앞서 언급한 텍스트 단위와 단락은 어떻게 다른가를 규명할 필요가 있다.

대부분의 전문검색시스템에서는 'A라는 용어와 B라는 용어가 동일 문단내에서 출현하는 문헌을 검색하라' 와 같은 문단탐색기능이 있으며, 하이퍼텍스트시스템의 링크 수단으로도 널리 사용되고 있기 때문에 그 차이점을 규명할 필요가 있다. 또한 부분적 텍스트가 어디에서 시작하고 어디에서 끝나는가 하는 문제는 문헌을 텍스트 단위로 분할하여 구조화하는데 있어서 중요한 과제이기도 하다.

앞서 언급한 텍스트의 조건만 갖추었다면 「불이야!」와 같은 단일 어구도 텍스트가 될 수 있으나 보통 다수의 문장이 모여 텍스트 단위를 이룬다는 점에 대해서는 언급한 바 있다. 그러나 텍스트와 문단의 경계에 있어서는 문장(S)이 모여서 텍스트(T)를 이루고 두 개 이상의 텍스트가 모여 문단(P)을 이루며, 문단이 모여 節과 章, 궁극적으로는 가장

〈표 1〉 문단 구분과 각 문단의 역할과 의미

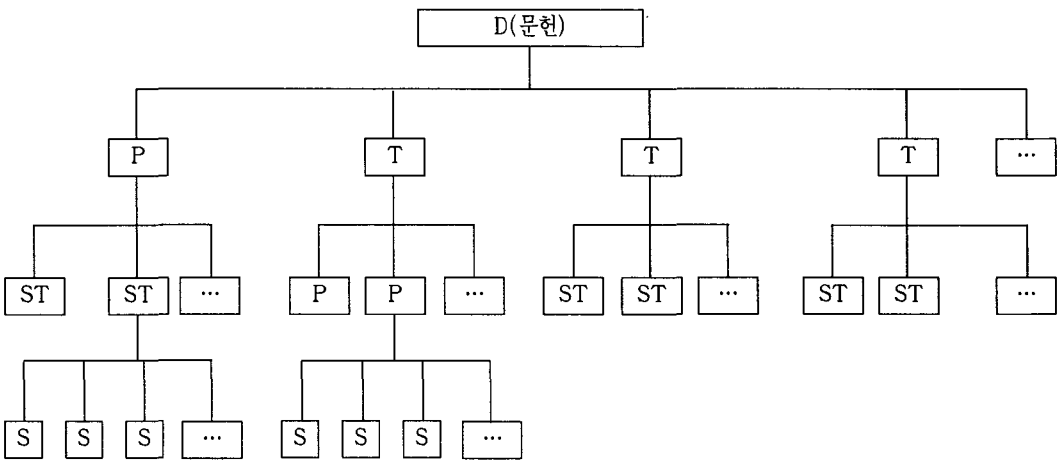
문단 번호	각 문단의 역할 및 의미 내용	문단 번호	각 문단의 역할 및 의미 내용
1	일반적 설명	15	과거의 연구결과 종합
2	(전체)문제제기		
3	예비지식의 제공	16	연구 실험행위 문제제기
4	예비지식의 제공	17	실험 방법 설명
5	4의 사례	18	17의 예
6	4의 사례	19	저자의 제1실험 절차
7	예비지식의 제공	20	도표 보는법 해설
8	서론	21	19의 결과
9	과거의 실험1	22	저자 제2의 실험 설명
10	9의 결과	23	22의 절차
11	9의 결과	24	22의 결과
12	9의 결과	25	저자 제3의 실험 설명
13	과거의 실험2	26	25의 결과
14	과거의 실험3	27	저자 실험의 종합
		28	고찰

큰 텍스트인 문헌을 형성한다는 견해(전병선, 1987)와 문단은 텍스트 보다는 작지만 문장 보다는 큰 의미 구성체(김종인, 1995)로 보는 견해가 있다. 그러나 문단은 보통 새 줄을 잡아서 들여쓰는 외형적 특성으로 구분하고 있으나 일상적인 글에서 저자의 문단 기준을 명확하게 설정하기 어렵고, 실제 표기에 있어서도 통일되어 있지 않다.

따라서 텍스트는 외형적인 요소나 길이에 관계없이 한 문단이 하나의 텍스트가 될 수 있고, 또는 2개이상의 텍스트도 될 수 있으며, 2개 이상의 문단이 모여 하나의 텍스트가 될 수 도 있다. 하나의 예로 히로코 서지하라(杉原寬子, 1985)가 28개 문단으로 구성된 학술논문을 분석하여 각 문단의 역할과 의미내용을 제시한 〈표 1〉을 보면 쉽게 이해할 수 있다.

〈표 1〉에서 보는 바와 같이 3~7번째 문단은 모두 「예비 지식」에 대한 정보를 제공하려는 단일 텍스트로 간주할 수 있으나 외형적으로는 5개의 문단으로 구분되어 있으며, 8~15 번째 문단 역시 「과거의 실험 및 연구결과」를 종합 정리한 부분으로써 똑같은 목적과 기능을 갖는 단일 텍스트로 간주할 수 있으나 외형적으로 8개의 문단으로 구분되어 있다. 일반적인 문단의 개념과 텍스트 내용단위가 서로 일치하는 경우도 있으나, 저자의 표현 방식이나 습관에 의존적일 수밖에 없는 문단과 텍스트 단위간에는 차이가 있으며, 문단은 들여쓰기와 같은 외적 요소에 의해 쉽게 식별할 수 있다는 장점이 있으나 각각의 문단을 의미단위로 식별할 수 있는 요소가 없다.

이상에서 밝힌 문장, 문단, 텍스텍간의 관계를 도식화하면 〈그림 1〉과 같다.



S(sentence):문장 P(paragraph):문단 T(text):텍스트
ST(sub-text):하위텍스트 D(documents):문헌

〈그림 1〉 문장, 문단, 텍스트간의 관계도

3. 1. 2 내용구조에 의한 인지적접근

문헌에 의한 의사소통은 저자인 생성자와 저자가 생성한 글, 수용자라는 3가지 요소를 바탕으로 이루어 진다. 인지심리학이나 언어심리학자에 의하면 문헌의 유형에 따라서 존재하는 텍스트단위 내용구조는 글의 생성이나 이해과정에 있어서 인지적 스키마로 작용하고 있음을 밝히고 있다. 본 절에서는 텍스트단위 내용구조가 문헌의 생성, 이해과정에서 작용하는 인지적 역할을 규명함으로써 전문데이터베이스를 내용단위로 구조화해야 할 이론적 바탕을 제시하고자 한다.

문헌을 생성하고 이해한다는 것은 글과 참여자들간에 이미 공유하고 있는 사전지식을 매개로한 지속적인 상호작용 및 타협의 과정이다(Beaugrande; Dressler, 1984). 여기에서 사전지식은 일종의 스키마로써 프레임, 스크립트, 플랜이라고도 하며 글의 생성과 이해 과정에서는 텍스트단위구조가 스키마로 작용을 한다. 인지구조론자에 의하면 "스키마는 인지 과정에서 반드시 요구되는 세상사에 대한 지식 구조, 경험의 총체로써 스키마의 동원 없이는 문헌의 생성이나 이해 및 활용이 불가능하다는 것이다.(Anderson, 1984). P. W. Thorndyke(1977)는 글의 유형에 따라서 다양한 종류의 스키마, 예를 들면 보도문 스키마, 역사문 스키마, 과학논문 스키마 등이 존재하며, 이 스키마는 텍스트단위 내용구조로써 글의 생성과 이해 방법에 대한 인지구조로 작용하고 있음을 밝히고 있다.

그는 실제 이야기를 대상으로 한 텍스트단위 구조모델을 제시하고, 이야기를 읽을 때 그가 제시한 구조를 회상하면서 읽으면 이야기 정보를 그 문법 구조에 잘 적응시킴으로써 이해와 기억을 촉진시킬 수 있음을 실증적으로 입증하고 있다.

또한 Kintsch(1978)는 과학연구보고서를 대상으로 텍스트단위 내용구조를 제시하였고, M. Just와 P. Carpenter(1980)는 설명문이나 논설문의 내용 구조를 분석하여 글의 내용 조직유형이 글의 생성이나 이해에 영향을 끼치는 인지적 스키마구조임을 밝히고 있다.

또한 인간의 언어이해 모형은 크게 상향식(bottom-up)과 하향식(top-down) 그리고 상호작용식(interactive)모델이 있다(Frank Smith, 1994). 상향식모델은 내용을 이해함에 있어서 작은 단위의 언어(음운, 철자, 단어 지각)에서 시작하여 점진적으로 보다 큰 언어적 단위(단어→구→절→문장)로 확대되고 나아가 전체 의미가 형성된다는 입장이다. 그러나 언어이해는 글에 제시되어 있는 단어들의 의미를 결합하여 문장의 의미를 해석하고, 문장의 의미를 결합하여 글 전체의 의미를 해석하는 단순하고 순차적인 과정이 아니라는 비판과 더불어 하향식 모델이 등장하였다. 하향식 모델은 주제에 대한 배경지식이나 글의 내용구조에 대한 스키마구조를 통해 이해가 된다는 것으로 두 모델을 혼합한 상호작용 모델과 아울러 널리 수용되고 있다. 이는 하향식이든 상호작용식 모델이든 간에 스키마구조가 내용 이해에 영향을 끼치고 있음을 알 수 있으며, 내용의 조직 형태가 스키마 구조로 작용하고 있다는 것이다. 또한 글의 생성자와 수용자는 현실적인

언어 세계와 관습적인 내용구조를 문헌에 반영함으로써 의사 소통이 이루어진다는 모델(Clare Beghtol, 1986)을 제시 함으로써 커뮤케이션에 있어서 내용구조의 중요성을 강조한 연구도 있다.

결국, 문헌의 유형에 따라서 어떠한 역할과 기능을 갖는 텍스트단위가 포함되어야 한다는 텍스트단위 내용구조가 관습화되어 있기 때문에 글의 생성이나 이해, 기억과정에서 인지적 스키마구조로 작용하고 있으며, 이러한 구조를 활용하면 글이나 문헌의 이해, 기억, 전달 과정을 촉진 시킬 수 있음을 시사하고 있다.

3. 2 내용단위구조와 정보요구의 관계

3. 2. 1 정보요구에 대한 인식의 변화

정보검색시스템의 궁극적인 목적은 이용자의 정보요구에 적합한 문헌을 제공하는데 있기 때문에 정보요구 및 적합성의 개념은 정보검색 분야에 있어서 핵심적인 영역이다. 특히, 전문검색시스템은 서지데이터가 아닌 원문이 수록되어 있고, 문헌 전체 뿐만아니라 문헌의 부분 탐색이 가능하기 때문에 적합성 판단 기준이나 정보요구에 대한 인식은 전문데이터베이스의 표현 및 인터페이스 설계에 있어서 중대한 영향을 끼친다.

전통적인 정보검색시스템에서는 질의와 문헌간의 주제어 일치에 의존하여 적합한 문헌을 검색하여 왔다. 그러나 적합성이란 단순히 시스템 내부의 효율적인 매칭에 의해서 결정되는 시스템 현상이 아니며, 적합성 판단은 주제 이외에 다른 요인에 의해 영향을 받고 있음을 밝힌 C. Cuadra & R. Katter(1967)와 A. Rees & G. Schultz(1967)의 연구를 계기로 이용자중심, 인지적 관점으로 전환하게 되었다. 이용자중심의 인지적 관점에 의하면 적합성은 주제어간의 일치에 의해서 결정되거나 고정된 것이 아니라 이용자의 상황이나 목적에 따라 결정되는 역동적인 개념임을 밝히고 있다. 이는 주제에 의존한 적합성 기준과 실제 이용자의 적합성 판단 기준간에는 많은 차이가 있으며, 주제 이외에 적합성 판정에 영향을 끼치는 요인 및 판단 과정을 규명하려는 실증적 연구가 활발하게 이루어 지고 있다.

3. 2. 2 정보요구와 내용단위구조

이용자 중심의 인지적 상황론적 접근을 통해 적합성 판단 기준을 실증적으로 규명한 연구에 의하면 정보요구의 구조는 문헌의 내용구조와 밀접하게 관련되어 있음이 밝혀지고 있다. 적합성 판정에 영향을 끼치는 요인에는 여러 가지가 있으나 '정의에 필요한 정보', '방법론에 대한 정보' '향후 연구과제에 대한 정보'와 같이 문헌의 내용구조와 밀접하게 관련되어 있다는 것이다.

Brace Allen(1988)은 참고면담에 있어서 어떤 형태의 질문이 이용자로 하여금 정보요구를 잘 표현할 수 있도록 하는가를 규명하기 위해서 내용 구조적 질문, 서지적 질문, 자유형 질문으로 구분하여 실험하였다. 그 결과 "내용구조적 질문이 다른 유형의 질문에 비해 유용한 키워드를 많이 추출해 냄으로써 구조적 질문이 유용성이 있다는 결론을 도출하고 있다. 또한 R. N. Oddy(1992)등은

'텍스트단위 내용구조와 이용자의 정보요구 상황은 밀접하게 관련되어 있기 때문에 텍스트 구조 정보를 이용하면 이용자의 정보요구 상황을 반영한 정보검색 시스템 구축이 가능함을 실험적으로 입증하고 있다. 이들에 의하면 이용자의 적합성 판단의 기준으로는 주제관련 기준과 비주제적 기준(예, 특정 저자, 언어등)이외에 과제지향 혹은 상황적 기준이 크게 작용하고 있으며, 텍스트단위 내용구조는 이용자의 정보 요구 상황 혹은 목적을 표현하는데 있어서 유용한 방안임을 밝히고 있다.

그리고 T. K. Park(1993)는 교육학, 사회학, 인류학등 교수 및 석 박사 과정의 학생 10명을 대상으로 검색결과를 제시하고 적합성 판정에 영향을 끼치는 요인을 규명하는 실험을 통해 정보요구를 구조화하였다. 그 결과 적합성 판정에 영향을 끼치는 요인을 크게 3가지 범주(내적요인, 외적요인, 내용적 요인) 22가지 요소로 제시하고 있다. 내적 요인은 이용자 자신의 사전 경험, 교육적 배경에 따른 영향이고, 외적 요인은 개개인의 연구조사 상황에 의한 영향으로써 연구조사의 질적 수준, 연구의 목적, 이용가능성, 연구의 최종 생산물 등의 요인을 들고 있다. 그리고 내용적 요인으로는 문헌을 이용하려는 동기, 목적에 따른 영향으로써 '정의에 필요한 정보 배경 정보 방법론 연구과제의 구조' 등의 요소를 들고 있다. 또한 Park는 정보가 학술문제해결에 이용되는 과정을 밝혀 문제해결 과정에 따라서 적합한 문헌을 검색할 수 있는 정보검색시스템의 가능성을 제시하고 있다. 학술적인 문제의 구조는 주제적인 성분과 비주제적인 성분으로 나눌수 있으며 비주제적인 성분중에 하나로 텍스트 기반 내용단위구조와 밀접하게 관련된 「연구단계」요소를 제시하고 있다. J. T. Guthrie(1988)는 일반적인 독해과정과 정보집합체 혹은 특정 문헌으로부터 목적에 부합하는 정보를 찾아내는 탐색 과정간에는 차이가 있다는 전제하에 전자문헌으로부터 특정 부분의 정보를 찾아내는 인지과정 모델을 제시하고 있다. 이 인지모델에 의하면 목표확인-범주설정-정보추출-통합-반복과정을 거쳐 정보를 이해하며, 그 중 '범주설정'은 정보요구 상황에 따라서 특정 부분의 정보내용을 추출하려는 것으로 텍스트단위 내용구조와 높은 상관관계가 있음을 밝히고 있다.

이상과 같이 적합성 판단기준이나 인지과정을 밝히고 있는 연구에 의하면 정보요구는 주제 뿐만아니라 특정 상황 혹은 목적과 관련된 요소를 포함하고 있으며, 특정 상황이나 목적은 문헌의 내용구조와 밀접하게 관련되어 있음을 알 수 있다. 이는 정보검색시스템 특히, 전문검색시스템의 설계에 있어서 중요한 의미를 갖는다. 이는 내용단위구조를 탐색문에 반영할 수 있는 전문검색시스템은 그렇지 않은 시스템에 비해 이용자의 만족도를 향상시킬 가능성이 많기 때문이다. 최근에 P. Ingwersen(1996)은 이용자의 정보요구와 정보검색시스템의 정보내용에 공통적으로 적용되는 다중표현(polyrepresentation)을 이용한 정보검색모델을 제시하였다.

4. 학술논문의 내용단위구조의 비교 및 설정

텍스트단위 내용구조는 문헌의 유형에 따라서 전형적인 구조가 존재하고, 인지적 스키마로 작용하고 있으며, 이용자의 정보요구와도 밀접하게 관련되어 있음을 살펴 보았다. 그러나 텍스트단위 내용구조에 의한 전문검색시스템을 구현하기 위해서는 정보검색시스템 설계자와 이용자들간에 공통적으로 인식할 수 있는 표준적인 내용구조 모델이 필요하다. 따라서 본장에서는 학술논문을 대상으로 전문데이터베이스를 구축할 때에 적용할 수 있는 내용단위구조 모델을 개발함으로써 표준적인 내용단위구조모델 설정의 가능성과 이용자의 인지도를 측정해 보고자 한다.

4. 1 학술논문의 내용단위구조 모델 비교

학술논문을 대상으로 내용을 구성하고 있는 단위요소와 이들간의 관계구조를 제시하고 있는 대표적인 모델로는 Van Dijk(1980)의 상부구조와 Van Dijk와 Kintsch(1983)의 스키마구조, Liddy(1991), Noriko Kando(1992)를 들 수 있다.

Van Dijk(1980)는 한 편의 글 속에는 3가지 층위구조 즉, 미시구조, 거시구조, 상부구조가 존재하며, 그 가운데 상부구조는 글 전체의 내용를 조직하는데 필요한 일련의 범주로 구성된 추상적인 스키마라 정의하고 있다. 이러한 상부구조의 대표적인 형태로써 연구보고서의 문제-해결구조를 들고 있다. 연구보고서는 결론과 그 결론의 정당화 그리고 문제제기와 해결에 해당하는 텍스트로 구성되며, 각각의 텍스트는 그 하위 텍스트를 갖는다는 구조이론을 제시하면서 관찰, 설명, 가설, 예언, 테스트, 결론, 실험, 구성, 피험자, 연구조건, 결과, 토론, 해답등과 같은 내용단위요소를 모형화하고 있다.

또한 Dijk는 Kintsch(1983)와 더불어 글의 이해와 생성에 관한 인지모델을 제시하고 스키마구조의 역할을 강조하는 과정에서 학술논문의 내용구조를 규명하고 있다. 스키마구조를 형성하고 있는 대표적인 글의 형태로 이야기, 논설문, 학술 논문 등을 들면서 아래와 같이 구체적인 학술논문의 내용단위구조 모델을 제시하고 있다.

학술논문 스키마
- 서론(실험연구)
- 연구환경
  - 시간
  - 장소
  - 선행문헌
  - 연구목적
- 방법
- 결과
- 고찰

또한 Liddy(1991)는 현행 정보검색 시스템이 특정 용어나 어구를 중심으로 검색이 이루어 지기 때문에 용어가 나타내고 있는 개념이 문헌내에서 어떠한 역할이나 기능을 하는지 혹은 개념간의 관계를 표현하지 못한다는 문제점을 지적하고 문헌의 내용구조가 갖

는 유효성을 지적하고, 실증적 연구를 통해 학술논문의 내용단위구조 모델을 제시하고 있다. Liddy는 우선 초록전문가들이 초록을 작성할 때 어떠한 인지적 구성요소가 작용하고 있는가를 조사하고, ERIC과 PsycINFO 데이터베이스로부터 각각 150건, 126건의 샘플 초록을 추출하여 구성요소의 역할 자질을 표현하는 단서어를 근거로 초록의 구성요소를 분석하rh 내용단위구조 모델을 도출하였다. Liddy의 내용단위 구조 모델은 목적-방법-결과-결론등의 1차수준, 타연구와의 관계-가설-연구과제-대상-절차-고찰-의의 등으로 구성된 2차수준, 독립변수-종속변수-표본-조건-재료 등으로 이루어진 3차수준으로 단위요소를 구분하여 제시하고 있다.

그리고 Noriko Kando(1992)는 문헌에 의해 전달되는 내용에는 일정한 내부 구조를 갖고 있으며, 이러한 구조를 이루는 단위 요소를 통해 검색효율을 향상시킬 수 있다는 전제하에 일본어 학술논문의 구조를 분석하고 있다. 문제-문제해결에 필요한 증거-해답을 최상위구조로 하고 60여개의 하위요소를 갖는 내용단위구조 모델을 설정하고 있다.

이상과 같이 선행연구에 나타난 학술논문의 내용구조 모델을 비교 분석한 결과 ① 특정 주제분야를 대상으로 분석했다는 점 ② 이용자의 정보요구를 반영할 수 있을 정도로 구체적이지 못하다는 점 ③ 내용과 직접적으로 관련된지 않은 단위요소가 많다는 점 ④ 내용구조 단위요소간에 명확한 구분이 어렵다는 문제점을 발견할 수 있었다.

따라서 학술논문의 내용구조를 전문데이터베이스 구조화 수단으로 사용하기 위해서는 이용자들이 공통적으로 적용할 수 있는 새로운 형태의 모델을 개발해야 할 필요성이 제기 되었다. 표준적인 내용단위구조 모델을 개발하기 위한 방안으로 우선 선행연구에 나타난 모델의 단위요소를 다음과 같은 원칙에 의거 종합 추출하였다. 그리고 추출한 내용단위요소는 국내의 학술논문에 직접 적용해 보고, 문제점을 수정 보완함으로써 최종 적으로 내용단위구조모델을 설정하였다.

첫째, 전문을 분석대상으로 하였으며 가장 구체적으로 제시된 Noriko Kando의 모델을 기본으로 상하위 개념을 고려하여 여타의 모델에서 제시한 단위요소를 통합, 배치 하였다. 둘째, 가능한한 일반적이고 간결한 용어로 단위요소의 명칭을 전환하였다. 셋째, 기존의 서지데이터베이스 혹은 SGML과 같은 논리적 구조에 의한 전문데이터베이스를 통해서 접근이 가능한 단위요소는 문헌의 내용과는 직접적으로 관련이 없기 때문에 제외시켰다. 넷째, 용어상으로 차이가 있으나 같은 내용일 경우에는 대표성이 있는 단위요소로 통합하였다.

4. 2 학술논문의 내용단위구조 모델 설정

4. 1項에서 추출한 내용단위구조의 요소를 한국어로 작성된 학술논문에 직접 적용해 봄으로써 단위요소의 적용가능성과 단위요소간의 관계가 적절하게 설정되어 있는가를 검토해 보고자 한다. 그리하여 단위요소의 출현여부와 출현패턴에 대한 실증적 분석자료를 바탕으로 각 단위요소의 영역을 분명히 하고, 표준적인 내용구조모델을 설정하고자

한다.

4. 2. 1 분석대상 및 방법

분석대상은 국회도서관에서 발행하는 「정기간행물 기사색인(1996. 1-3)」에 수록된 기사 가운데 특정 분야에 편중되지 않도록 13개 주제 분야별로 각 15건씩 총 195건을 선정하였다. 주제 분야별로 15건씩 추출함에 있어서 동일한 학술지에 게재된 기사는 단위요소의 출현이나 출현패턴이 유사하기 때문에 주제 분야내 세분 항목별로 정보원이중복되지 않도록 선정하였다. 이와 같이 선정한 195건의 원문을 입수한 결과 기술 소개, 경향 분석, 시사 정보, 논평 등 학술 논문의 성격에서 벗어난 15건을 제외하고 총 180건을 최종 분석 대상으로 삼았다.

분석대상 논문의 전문 가운데 표제, 초록, 목차, 키워드, 고유 번호와 같은 前部 그리고 부록, 참고문헌, 요약 등으로 이루어진 後部는 논문의 내용과는 직접으로 관련이 없기 때문에 이들은 제외하고 순수 내용으로 구성된 本體部 만을 분석 범위로 삼았다.

내용구조 단위요소를 실제로 부여할 때에는 원문에 나타나 있는 문장 혹은 문단, 장절의 구분 방식에 관계없이 내용상으로 식별이 가능하면 단위요소코드를 부여하였다.

하나의 겹문장(복문)이라 하더라도 단위요소의 식별이 가능한 경우에는 이를 구분하였다. 구성요소의 식별은 역할어구, 문장과 문장 혹은 문단과 문단간의 접속 관계, 텍스트 언어학적 형식 등을 기준으로 판정하였다. 가능한한 구체적인 단위요소, 즉 최하위 단위요소부터 순차적으로 부여하였다. 앞부분에서 출현한 단위요소를 다시한번 '예시, 구체화, 강조' 등의 논리 전개에 의거 동일한 단위요소가 여러곳에서 발생하는 경우에는 같은 단위요소 태그를 부여 하였다.

4. 2. 2 내용단위구조의 분석결과

서론부, 본론부, 결론부로 나누어 각 내용단위요소별로 단위요소의 적용여부, 출현율과 출현패턴을 〈표 2〉 〈표 3〉과 같은 방법으로 조사 분석 하였다.

분석 결과를 요약하면 다음과 같다. 첫째, 주제 분야 혹은 연구 유형에 편중된 단위요소가 다소 있었으나 대부분은 공통적으로 적용할 수 있었다. 둘째, 4.1項에서 추출한 분석용 단위요소중에는 실제 문헌에 적용할 수 없거나 다른 단위요소와 식별이 곤란한경우가 있었다. 셋째, 분석용 단위요소만으로는 내용구분이 불가능하거나 단위요소간의 계층적 속성이 불분명한 경우가 있었다. 넷째, 단위요소의 출현패턴은 상위수준으로 올라갈수록 공통성이 많았고, 하위수준으로 내려갈수록 단위요소간의 반복, 생략, 치환등 불규칙인 경향을 보였다.

첫째, 4.1項에서 추출한 내용구조의 단위요소중에서 실제 문헌의 내용에 적용할 수 없거나 다른 단위요소와 의미상으로 식별이 모호한 요소는 삭제하거나 하나의 단위요소로 통합하였다. 둘째, 상위개념 단위요소와 하위개념 단위요소간에 구분이 않되거나 계층적 속성이 실제 문헌의 내용과 일치하지 않는 경우에는 하나로 통합하거나 계층관계를 재조정하였다. 셋째, 4.1項에서 추출한 단위요소 이외에 새로운 단위요소가 필요하거나 기존

〈표 2〉 서론부의 내용 단위요소 출현율

주제영역 / 단위요소	A	B	C	D	E	F	G	H	I	J	K	L	M	합계	%
서론 章題 출현	13	14	13	12	14	14	13	15	11	14	15	13	13	174	97
서론內 소제목 구분	-	-	-	-	-	3	-	1	-	-	2	-	1	7	4
A1 주지의 사실	7	9	7	7	8	8	9	9	11	8	3	4	5	95	53
A11 일반적 견해	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
A2 타연구와의 관계	13	14	11	14	13	12	13	15	14	15	14	12	13	173	96
A21 선행연구(인용)	10	12	10	11	11	12	11	14	13	14	13	12	12	155	86
A22 선행연구의 불완전성	8	12	10	10	8	9	11	12	11	14	13	10	11	139	77
A3 문제점 명료화	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
A31 연구 중요성 의의	6	7	7	9	5	6	4	3	3	7	2	11	10	80	44
A32 연구 착수의 동기	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
A33 저자의 입장	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
A4 연구 과제	13	14	13	14	14	14	13	15	14	15	15	13	13	180	100
A41 문제제기	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
A411 가설	1	1	2	3	4	3	-	2	1	-	1	2	1	21	12
A4111독립변수	2	-	2	-	-	3	-	-	-	-	-	3	-	10	4
A4112 종속변수	2	-	2	-	-	3	-	-	-	-	-	3	-	10	4
A412 연구목적	13	14	13	14	14	14	13	15	14	15	15	13	13	180	100
A42 연구범위	10	11	9	4	12	11	12	10	8	7	10	13	12	129	72
A421 연구방법	12	12	13	14	14	14	12	14	12	15	15	13	13	173	96
A422 검토고찰 항목	1	-	-	-	-	-	-	1	-	-	1	1	-	4	2
A423 결과 결론	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
A424 논문의 구성	-	-	3	-	-	-	-	-	2	-	-	-	-	5	3
A5 용어 정의	-	-	2	-	-	-	-	-	-	2	-	3	-	7	4

* 주제영역 표시 A:정치 행정 B:법률 C:경제 D:산업 농산업 E:사회 노동 F:교육 G:문화 예술 H:문학 어학 I:철학 종교 J:역사 지리 K:순수과학 L:의학 약학 M:공학 기술

의 단위요소를 세분할 필요성이 있는 경우에는 새로운 단위요소를 신설하였다.

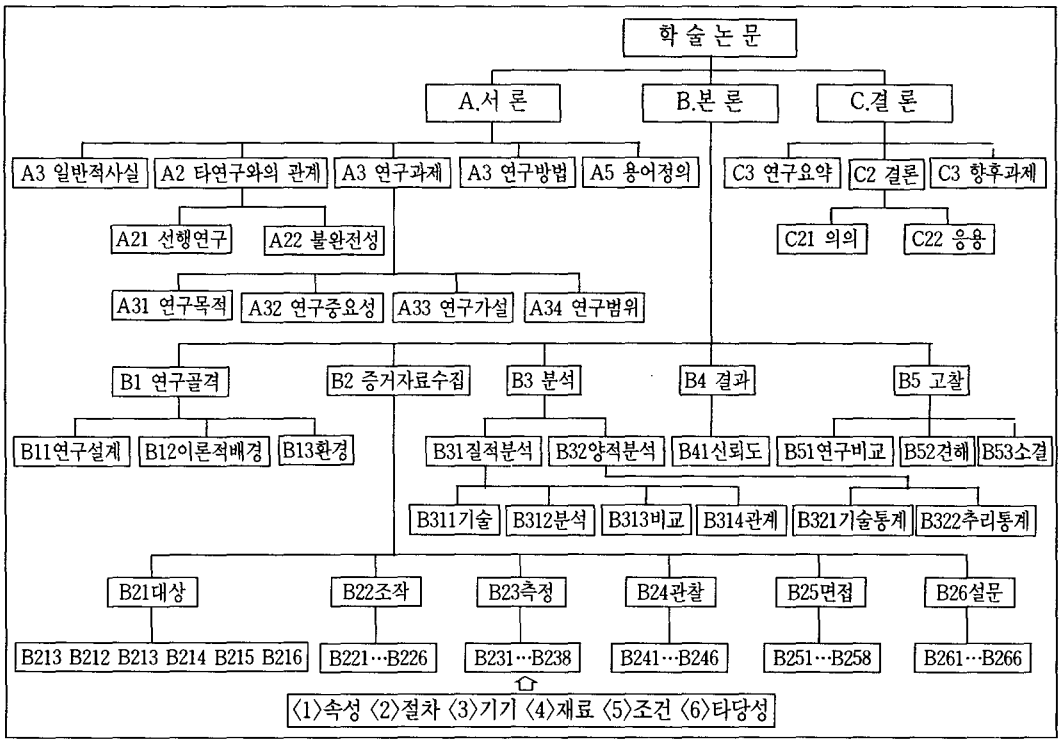
이상과 같은 원칙에 따라서 설정한 학술논문의 내용구조 모델은 〈그림 2〉와 같다.

〈그림 2〉에서 보는 바와 같이 학술논문의 내용구조 모델은 전체적으로 서론, 본론, 결론 등 3단구성으로 이루어져 있고, 본론부분에는 주로 증거자료의 수집과정 및 증거자료의 분석과정 그리고 그 결과에 해당하는 단위요소들로 구성되어 있다. 최하단 부분에 있는 네모 칸 속의 내용(속성, 절차, 기기, 재료, 조건, 타당성)은 B21(대상)에서부터

〈표 3〉 서론부 내용 단위요소의 출현 패턴

주제영역 / 단위요소 출현패턴	A	B	C	D	E	F	G	H	I	J	K	L	M	합 계	%
A1-A2-[A3-A4]-(A5)	5	4	2	3	2	3	3	1	1	3	-	3	4	34	19
A1-----[A3-A4]-(A5)	-	-	-	-	1	1	-	-	-	1	-	1	-	2	2
A1-A2----A4-(A5)	-	5	4	4	5	3	6	8	11	6	3	-	-	55	31
A1--------A4-(A5)	2	-	1	-	-	1	-	-	-	-	-	-	-	4	2
A2-[A3-A4]-(A5)	1	3	4	6	3	1	2	2	2	3	2	7	6	42	23
A2----A4-(A5)	5	2	1	1	3	4	2	4	-	2	8	2	3	37	21
														176	98

*주제영역 표시 A:정치 행정 B:법률 C:경제 D:산업 농산업 E:사회 노동 F:교육 G:문화 예술 H:문학 어학 I:철학 종교 J:역사 지리 K:순수과학 L:의학 약학 M:공학 기술



〈그림 2〉 학술논문의 표준 내용구조 모델

B26(설문)까지 공통적으로 적용되는 단위요소를 간략하게 표현하기 위한 것으로써 B251은 면접-속성, B252는 면접-절차와 같이 전개된다.

4. 2. 3 내용구조 모델의 설정 및 검증

학술논문의 내용단위구조를 전문검색시스템에 적용하기 위해서는 설정한 표준모델이 과연 이용자들이 정보탐색상황에서 인지할 수 있는가, 또 얼마나 공통적으로 인식하고 있는가 하는 문제는 매우 중요하다. 왜냐하면 구현코자 하는 내용구조 기반 전문검색시스템의 인터페이스를 통해 정보요구에 따른 내용범위를 지정할 수 있느냐하는 문제와 직결되기 때문이다. 따라서 본 논문에서 개발한 내용구조 모델의 타당성을 검증해 볼 필요가 있다. 학술논문의 내용구조 모델의 타당성을 검증하기 위한 방안으로 동일한 내용의 학술논문을 다수인에게 분석케 한후 종합적인 신뢰도를 측정함으로써 얼마나 내용구조 모델을 공통적으로 인식할 수 있는지, 그리고 어느 정도 일치하는가를 측정해 보았다. 검증 과정을 구체적으로 살펴보면, 우선 모 고교 교사중에서 국문학, 역사학, 지리학, 교육학, 생물학 분야의 석 박사과정에 재학중인 5명에게 주제분야 및 연구유형이 각각 다른 3편의 학술논문과 〈그림 2〉에 나타나 있는 학술논문의 내용구조 모델을 제시해주고 적절한 단위요소 태그를 학술논문에 표시토록 하였다. 5명의 피험자들에게는 내용구조 모델의 전체적인 구조와 각각의 단위요소에 대한 의미를 상세하게 설명해 주었으며, 장 절 문단 등 원저자의 경계구분에 관계없이 피험자가 생각하는 내용구분에 따라서 단위요소를 부여토록 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 Goldsen(1964)이 제시한 指數를 통해 개인별일치도와 조사문헌별 종합일치도를 측정하여 신뢰도를 조사하였다. 측정 결과는 〈표 4〉에 제시되어 있으며, 개인별일치도와 종합일치도를 산출한 공식은 아래와 같다.

개인별일치도와 종합일치도를 상관계수로 해석해 보면 문헌1의 신뢰도는 보통수준(R=0.57)이지만 문헌2와 3은 매우 높은 신뢰도(R=0.71, 0.77)를 보이고 있다. 대상문헌 1이 문헌2·3에 비해 일치도가 저조한 것은 조사연구 혹은 실험연구는 단위요소 B2(증거자료 수집)부분이 패턴화 되어 있음에 비해 문헌연구는 내용구분이 다소 불분명하기 때문인 듯하다. 각 문헌별 개인별일치도의 표준편차는 각각 3.4, 5.7, 2.7로 피험자의 주제분야는 내용구조 분석에 큰 영향을 끼치지 않고 있음을 알 수 있었다.

이상의 검증 결과를 볼 때, 본 장에서 설정한 학술논문의 내용구조 모델은 타당성이 있으며, 이용자들에게 내용구조모델의 단위

$$\text{개인별일치도}(R_i) = \frac{\sum_{j=1}^{k} e_{ij}}{NK} \times 100 \begin{cases} N\text{;분석자수} \\ k\text{;부여한 태그 수} \\ eij\text{;태그별 점수} \end{cases}$$

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

〈표 4〉 내용구조 모델의 검증결과

일치도 / 문헌	개인별 일치도(%)					종합일치도(%)
	A	B	C	D	E	
문헌1	63.6	66.3	68.1	57.2	60.0	57.2
문헌2	76.3	61.8	70.9	78.1	70.9	71.6
문헌3	78.4	81.5	75.3	73.8	76.9	77.2

요소를 제시해 주면 공통적으로 인식할 수 있고, 이를 통해 특정 내용범주를 설정할 수 있음이 확인되었다.

5. 결 론

전통적인 전문검색시스템은 문헌의 구조에 대한 정보를 갖고 있지 않거나 표제, 문장, 문단, 도표와 같은 형식적인 요소에만 의존하여 색인하고 탐색하기 때문에 과도한 문헌이 검색되어 부적합 문헌이 검색될 가능성이 높고, 내용과 관련된 이용자의 정보요구를 반영할 수 없다는 문제점이 있다. 이러한 문제점을 해결하기 위한 방안으로 전문데이터베이스의 본문을 내용단위로 구조화해야 할 이론적 타당성을 검토하였으며, 선행연구에 나타난 학술논문의 내용단위구조를 비교분석하여 내용단위요소를 추출하고, 실제 한국어로 작성된 학술논문에 적용 검토함으로써 표준적인 학술논문의 내용구조 모델을 개발하였다. 그 결과를 요약하면 다음과 같다.

첫째, 문헌의 내용단위구조는 실제 언어사용 환경에서 작용하는 의사소통 단위로써 특정 부분이 맡고 있는 역할이나 기능에 따라서 구분할 수 있다.

둘째, 문헌의 유형에 따라서 사회적으로 관습화된 전형적인 내용단위 구조가 존재하며, 또한 내용단위구조는 문헌의 생성 및 이해과정에서 인지적 스키마로 작용한다.

셋째, 항상 문헌 전체를 통람하는 것이아니라 이용자의 정보요구 상황이나 목적에 따라서 부분만을 필요로 하는 경우가 있으며, 정보요구 상황이나 목적은 내용단위구조와 밀접하게 관련되어 있다.

넷째, 학술논문을 대상으로 실험한 결과 표준적인 내용단위구조 모델 설정이 가능하며, 내용단위구조 요소를 이용자들이 식별할 수 있기 때문에 정보검색시스템의 인터페이스로 적용이 가능하다.

따라서 전문데이터베이스를 구축하거나 전문검색시스템을 설계할 때에는 정보검색의 효율성 측면에서 내용단위 구조화를 고려할 필요가 있다. 이는 全文情報의 유통, 관리에 매우 효과적인 것으로 알려진 SGML과 같은 논리적 구조의 배척이 아니라 다양한 탐색력을 갖는 전문검색시스템의 구현을 위한 정보표현의 다양성을 강조하려는 것이다.

내용단위구조 정보를 갖는 전문검색시스템은 문헌의 내용에 따른 탐색어의 위치 지정, 검색된 문헌의 특정 내용만을 골라서 읽

을 수 있는 횡적 브라우징, 내용상의 위치에 따른 가중치 부여 등 다양한 활용이 기대된다.

참 고 문 헌

윤석민, 1989. 국어의 텍스트 언어학적 연구 시론. 석사학위논문. 서울대학교 대학원.

유석종, 1995. "SGML 한글문서의 논리적 구조에 근거한 색인기법에 관한 연구". 정보관리학회지 12(2) : pp.85-101

윤석민, 1993. "RST와 국어의 텍스트 분석". 텍스트연구회 편. 텍스트언어학. 서울:서광학술자료사

이영수, 1993. 텍스트 구조의 언어학적 분석. 석사학위논문. 중앙대학교 대학원.

이삼형, 1993. 설명적 텍스트의 내용구조 분석 방법과 교육적 적용 연구. 박사학위논문. 서울 대학교 대학원.

이택경, 1996. "하이퍼미디어 시스템을 위한 Hytime엔진 및 SGML검색엔진의 개발", 정보과학회 논문지23(8) : 882-896

加藤藤澤, 1991. "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 : 97-106

高順淳宏, 1993. "SGMLと全文データベース." 情報の科學と技術 13(12) : 1089-1097.

野來道子, 1994. "段落を對象とした日本語全文データベースの檢索." Library and Information Science 31 : 79-94.

神門典子, 1995. "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7 : 1-12

———, 1992. "情報メデアの構造: 傳達內容の分析と利用." Library and Information Science, 30 : 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) : 39-46.

影浦 峽, 1990. "文獻の論理構造を考慮した全文探索シズテム." 學術情報センタ紀要3.

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

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

———, 1988. "Text structures and the user-inter-mediary interaction." RQ, 27(4): 535-541.

———, 1990. "Knowledge organization in an information retrieval task." Information Processing and Management 26(4), 1990 : 535-542.

Beaugrande, Robert de. 1994. Text production : toward a science of composition. Norwood, New Jersey : ABLEX.

Begthol, Clare. 1986. " Bibliographic classification theory and text linguistics : aboutness analysis, intertextuality and the cognitive act of classifying documents." Journal of Documentation 42(2) : 84-113.

Blair, D. C. 1990. "Fulltext informmation retrieval." IPM., 26(3) : 437-447.

Blake, G. E. 1994. " Text/Relational Database Management Systems: Harmonizing SQL and SGML." Proc. of the International Conference on Application of Database

Brawn, G. & Yule, G. 1983. Discourse Analysis. Cambridge : CPU.

Daniels, P. J. 1986. "Cognitive models in information retrieval: an evaluative review." Jour. of Documentation 42(4) : 272-304.

Dewire, D. T. 1994. Text Management. New York : McGraw-Hill

Dillon, Andrew. 1991. "Readers' s models of text structures : the case of academic articles." International Journal of Man-Machine Studies 35 : 913-925.

Guthrie, John and Irwin S. 1987. " Distinctions between reading comprehension and locating information in text." Journal of Educational Psychology 79(3) : 220-227.

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

Halliday, M and Hasan, R. 1976. Cohesion in English. London : Longman.

Hearst, M. A. and Plaunt, C. 1993. "Subtopic structuring for full-length document access," Proc. SIGIR '93, Association for computing machinery : 59-68.

ISO 8879-1986. Information processing-Text and Office Systems-Standard Generalized Markup Language(SGML).

Janos, Jiri. 1979. "Theory of Funtional Sentence Perspective and its Application for the Purposes of Automatic Extracting." IPM 15 (1) : 19-25.

Kintsch, W. and Van Dijk, T.A. 1978. " Toward a model of text comprehension and production." Psychology Review 85(5) : 363-394.

Kirsch, I. Guthrie J. 1984. "Adult reading practices for work and leisure." Adult Education Quarterly 34(4) : 213-232.

Liddy, Elizabeth Duross. 1991. " The discourse level structure of

empirical abstracts ; an exploratory study." Information Processing & Management27(1) : 55-81.

Macleod, Ian A. 1990. "Storage and retrieval of structured documents." Information & Processing Managements, 26(2) : 197-208.

Mittendorf, E and Sch uble, P. 1996. " Documents and Passage Retrieval Based on Hidden Markov Model." In ACM SIGIR Conference on R&D in Inf. Retrieval : 318-327.

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

Ro, J. S. 1988. "An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval I." JASIS39(2) : 73-78.

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

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

Smith, Frank. 1994. Understanding reading : a psycholinguistic analysis of reading and learning to read. Hillsdale : Lawrence Erlbaum Associates.

Stanfill, C and Waltz, D. 1992 "Statistical Methods, Artificial Intelligence and Information Retrieval." In: P. Jacobs. ed., Text Based Intelligent System : Current Reserch and Practice in Information Extraction and Retrieval. Lawrence Erlbaum Associates.

Tenopir, C. 1985 "Full Text Database Retrieval Performance." Online Review9(2) : 149-164

Tenopir, Carol & Ro, Jung Soon. 1990. Full text database: new directions in information management. NY : Greenwood Press.

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

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

Wilson, Patrick. 1978. "Some fundamental

concepts of information retrival." Drexel Library Quarterly 14(3) : 10-24.

Woodhead, Negel. 1991. Hypertext and Hypermedia Theory and Application. Wilson : Sigma Press.

Zobel, J. and Moffat, A. 1995 "Efficient Retrieval of Partial Document." Information & processing Management 31(3) : 361-377