全文데이터베이스의 특성과 정보검색성능

조 명 희*

<목 차>

Ⅰ. 서 론

1. 全文데이터베이스의 발전배경

1970년대초 이래로 상업기관들이 온라인데이터베이스를 제공하기 시작하였으며 그 숫자는 지난 십년동안에 십여배 이상 증가하였다.[1] 이처럼 온라인데이터베이스가 양적으로 성장한 것 이외에 정보검색서비스의 내용과 형태면에 있어서도 발전을 거듭하여 왔으며, 이는 상업적으로 데이터베이스를 제공하는 기관들의 계속적인 경쟁과 과학기술의 발전에 힘입은 바 크다. 오늘날의 데이터베이스 발전에 있어서 가장 고무적이고 주목할만한 사실은 全文데이터베이스의 출현이다. 많은 양의 텍스트를 저장할 수 있는 컴퓨터테크놀리지의 발전으로 全文데이터베이스가 구현된 것이다. 종래의 서지데이터베이스가 서지사항이나 초록등 문헌의 대용물을 입력하여 데이터베이스를

* 이화여자대학교 도서과학과 강사

1) 1979년에 전세계적으로 400개에 불과했던 온라인데이터베이스수가 1989년 7월 현재에는 4,245개에 달하고 있다 Directory of Online Databases (July, 1989)

만들었던 것에 비하여 全文데이터베이스는 문헌의 기사내용을 모두 컴퓨터에 수록하여 이용자가 서지사항뿐 아니라 문헌의 全文까지도 검색할 수 있도록 만들어졌다. 따라서 이용자는 不用語를 제외한 모든 본문의 출현단어로 本文을 탐색할 수 있으며, 本文이외에도 표, 참고문헌, 각주와 삽화명등도 검색할 수 있다.

현재 신문기사, 뉴우스통신정보, 법률관계자료, 연감류, 백과사전, 디렉토리 등의 전문(全文)이 여러 온라인탐색서비스 밴더들을 통해 검색될 수 있다. 1973년 여러해동안의 시험과정을 거쳐 법률정보검색시스팀인 NEXIS가 전문서비스를 시작한 이래 주로 법률정보데이터베이스가 全文데이터베이스의 주종을 이루었으나 최근에는 위에 열거한 이외에도 단행본, 비지니스잡지, 과학잡지등의 全文데이터베이스가 상당수 생산되고 있다. 이러한 데이터베이스들은 신문기사를 제공하는 NEXIS, 법률정보를 제공하는 LEXIS등 주제영역별 정보검색서비스를 이용하여 검색하거나 DIALOG, BRS, ORBIT 등 일반적인 대규모 검색시스팀을 통하여 제공되고 있다. 이외에도 DOW JONES NEWS/RETRIEVAL, VU/TEX, WESTLAW, NEWSNET등이 현재 全文검색서비스를 제공하는 대표적인 시스팀들이다. 뉴우스통신사정보나 신문기사 그리고 디렉토리같은 참고자료들은 일반적으로 축약형의 간단한 정보를 제공하며 작은 공간에서 가능한한 많은 정보를 전달할 수 있다. 이들은 일반적으로 서지데이터베이스에서 찾아볼 수 있는 초록문과 유사하며 따라서 연구자들도 기존의 서지데이터베이스의 검색에 대한 연구결과나 성능의 측정방법등을 이와같은 요약문 형태의 全文데이터베이스에 적용시킬 수 있다고 여겨왔다. 그러나 이러한 간결한 형태의 全文데이터베이스들은 진정한 의미의 全文(full-text)이라고 볼 수 없으며 미국아카데믹백과사전(Academic American Encyclopedia), 타임지(Times), Harvard Business Review와 같이 기사와 책전체의 완전내용을 온라인으로 검색할 수 있는 것이 본격적인 全文데이터베이스라고 말할 수 있다.[2] 따라서 본격적인 全文데이

2) W.A. Katz, *Introduction to Reference Work, V.Ⅱ. Reference Services and Reference Processes.* 5th ed. (N.Y.: McGraw-Hill Book Co., 1987), p.129.

터베이스의 특징을 감안하여 검색효율을 향상시킬 수 있는 독자적이고 체계적인 연구가 필요하게 되었다.

全文데이터베이스의 발생은 전자출판(electronic publishing)과 밀접한 관계를 가지고 있다. 새로운 전자미디어의 등장으로 출판이 종래의 지면상에 인쇄하는 것으로부터 전자매체를 사용하여 처리하고 저장하며, 이용자의 요구에 따라 디스플레이 시키거나 인쇄출력시키는 형식을취하게 되었다.[3] 이를 통해 수작업인쇄시의 반복적인 조판작업과 기계작업이 생략될 수 있고 시스팀에는 부산물로써 텍스트가 남게되므로 많은 양의 출판이 가능하며 동시에 노동력절약을 도모하므로 현재는 여러신문과 출판물들이 이러한 컴퓨터화된 출판시스팀을 이용하고 있다. 이처럼 인쇄물제작을 위한 효율적인 출판기술의 부산물로 기계가독형의 텍스트가 나타나게 되었으며[4] 이로부터 全文데이터베이스가 유래된 것이다. 이러한 컴퓨터인쇄이외에도, 컴퓨터의 저장시설의 용량이 증가되었고, 이러한 하드웨어의 가격이 하락되었으며 디지탈기술의 발달로 인하여 제5세대 컴퓨터의 등장, 또한 인간색인자의 인공색인작업에서 비롯되는 비지속성문제와 경비문제를 해결할 수 있다는점등이[5] 데이터베이스 생산자측에서 全文데이터베이스의 개발을 서두르는 요인이 되었다.

가까운 日本의 경우 1986년에 이미 26개의 출판사가 「日本電子出版協會」라는 전문기관을 설립하고 전자출판의 기술상의 문제, 요금, 저작권, 표준화등에 대한 의견수렴과 정보교환을 도모하고 있다.[6] 이를 통하여서도 데이터베이스 최전선에서는 전자출판을 통한 全文데이터베이스의 미래의 발전이 구체화되고 있음을 알 수 있다. 굳이 Lancaster의 예측, —즉 21세기 초까지, 기존잡지중 약 25퍼센트가 전자화되며, 1990년까지 참고도서의 약 25

3) C. Reedijk, *Large Libraries and New Technological Development*, (N.Y.: Saur, 1984), p.105.
4) S.E. Terrant. "Computers in Publishing," *ARIST* 15(1980), 202.
5) D.C. Blair and M.E. Maron, "An Evaluaion of Retrieval Effectiveness for a Full-Text Document-Retrieval System," *Communication of the ACM*, 28, (March, 1985), 289.
6) 田屋裕之, 電子メテイアと圖書館, (東京 : 勁草書房, 1989), p.46.

퍼센트, 21세기 이후에는 약 50퍼센트까지 전자화가 증가할 것이며 연구보고서의 약 절반가량이 1995년까지 전자화될 것이라는,[7]—을 굳이 빌지 않더라도 전자신문, 전자도서관, 전자사무실, 전자시대, 종이없는 사회(paperless society)등이 우리에게 이미 전혀 생소한 단어는 아니다.

국내에서는 1975년 한국과학기술정보센터가 온라인시스템 TECHNOLINE을 개발한 이래 주로 연구소들을 중심으로 온라인정보검색서비스가 본격적으로 이루어지게 되었다. 국내의 데이터베이스의 이용현황은 기업자체내에서 제작하여 내부용으로 주로 이용하는 경우와, 디렉토리나 서지, 수치등 기존인쇄물자료를 기계가독형태로 제작하여 데이터베이스화하여 서비스하는 경우를 들 수 있으며, 해외데이터베이스를 수입하여 보유하면서 일반에게 서비스하는 경우와 해외데이터뱅크와 연결하여 온라인검색을 제공하는 경우 등이 있다. 이중에서 산업연구원부설 산업기술정보센타는 자체제작된 데이터베이스 17종, 해외데이터베이스 17종을 각각 보유하고 있어 국내기관중 가장 많은 데이터베이스를 보유하고 서비스를 제공하며 또한 17개의 해외데이터뱅크를 연결하여 서비스하고 있다.[8] 국내 최초의 全文데이터베이스로는 동아출판사의 원색세계대백과사전 31권이 1989년에 나온 것을 들 수 있으며, 이는 아직 일반이용자의 접근을 허용치 않고 있다. 이외에도 대법원 판례데이터베이스가 全文으로 入力중이다.[9]

2. 연구의 필요성과 목적

현재 국내에서 온라인 全文데이터베이스를 이용할 수 있는 방법은 해외데이터뱅크의 화일을 이용하는 것인데, 서지데이터베이스의 활발한 이용에 비하여 별반 이용이 없는 편이다. 미국의 경우 全文 시스템인 NEXIS가 가장 이용이 많은 데이터베이스 중 하나로 꼽히고 있는 것에 비하여 볼 때, 우리

7) F.W. Lancaster, "The Future of the Library in the Age of Telecommunications," *In Changing Information Concepts and Technologies*, (N.Y.: Knowledge Industry, 1982), p.149.

8) 사공철, 구자영, 김석영, 과학기술문헌정보론 (서울 : 구미무역출판부, 1989), pp.373~374.

9) 남상식, "신문자료 데이터베이스에 관한 고찰" 정보관리학회지, 제6권 1호, (1989), 69.

나라에서 전문데이터베이스의 잠재적인 수요도 분명 있으리라고 생각된다. 출력건수가 많아지고 이에 따른 이용자의 비용부담이 커지고, 부적합자료가 많아짐에 따르는 적합정보선택의 혼란야기와 경비문제등 全文탐색에 따르는 단점도 있으나, 탐색중개자의 이에 대한 적극적인 소개나 안내가 미흡했던 점도 全文데이터베이스 이용상황에 영향을 끼쳤으리라 추정된다. 또한 국내의 데이터베이스 환경이 초보단계에 머므르고 있는것도 원인이 된다. 그러나 구미나 日本의 활발한 데이터베이스제작과 특히 全文데이터베이스 정보산업, 전자출판의 성장가능성, 국내의 국가전산망계획을 감안하여 볼때, 국내의 全文데이터베이스제작과 서비스환경에 변화가 오리라는 것을 예측하기는 어렵지 않으며 이에 대비한 연구의 필요성은 절실하다고 느껴진다. 따라서 본 연구는 全文데이터베이스의 특성을 파악하고 효율적인 검색방법을 조사하여 全文데이터베이스의 선택과 서비스에 효과적으로 대처하기 위한 노력의 일환으로 시도된 것이다.

3. 연구의 범위와 방법

본고에서는 새로운 정보매체로써, 주제접근의 새로운 가능성으로써 부각되고 있는 全文데이터베이스서비스의 국내외적인 동향과 발전이 어느정도까지 이르렀나를 살피고, 효율적인 이용방안, 그리고 현 全文데이터베이스의 이용자 분석과 도서관이나 정보센터에서 全文데이터베이스의 수용등, 미래지향적인 정보 서비스를 어떻게 효율적으로 이용할 수 있나 하는 관점에서 고찰하였다. 이에 따라, 全文데이터베이스의 출현과 발전, 온라인全文데이터베이스의 현황과 서비스의 특징, 全文데이터베이스의 주제색인 방법과 검색성능검토, 全文데이터베이스서비스에 대한 이용자반응과 도서관의 수용문제, 여러가지 문제점과 금후의 전개방향등을 살펴보았다.

본 연구는 문헌조사를 중심으로 행하였으며, 全文데이터베이스 연구가 최근 몇년내에 활발하여졌고 기술개발에 민감히 영향을 받는 분야이므로 가능한한 최근 자료까지를 참조하려고 노력하였으며, 특히 국내의 현황파악을 위하여는 문헌으로 보고된 것이 거의 없으므로 해당 프로젝트 담당자들과의

직접면접방법을 통하여 정보를 입수하였다.

Ⅱ. 온라인 全文데이터베이스의 국내외 서비스 현황

온라인검색시스템을 통해서 탐색할 수 있는 상업적인 全文데이터베이스의 수는 지난 수년동안 급성장을 하여 왔다. 全文서비스 제공에 매우 적극적인 BRS의 경우 1986년 당시 총 122개의 보유화일중 22개, 1989년 현재로는 162개의 화일중 41개가 全文으로 구성되어 있다.[10]

본장에서는 국내외에서 현재 가동중인 중요 全文데이터베이스 및 검색시스템들의 현황을 파악하고 각각의 탐색특징을 中心으로 살펴보고자 한다. 편의상 법률정보검색시스템, 신문과 뉴우스정보검색시스템, 정기간행물, 참고도서류로 나누어 조사하였다.

법률관계의 데이터베이스는 가장 최초로 全文을 제공하였으며, 일찌기 全文데이터베이스의 주종을 이루었었다. 특히 LEXIS와 WESTLAW는 법률분야에서 쌍벽을 이루는 시스템들이다. LEXIS는 일반 데이터베이스인 NEXIS를 생산하는 Mead Data Central사가 제공하는 시스템으로써, 1960년대의 실험과정을 거쳐서 1973년부터 서비스를 제공하고 있으며 자체 개발한 검색프로그램을 사용하고 있다. 이 시스템은 주제명표목이 없으며, 데이터베이스에 색인을 부여치 않고 全文온라인을 제공하고 있다.[11] 탐색영역은 일반적인 패턴을 따르며 서명, 저자명, 법률자료의 本文에 출현한 단어를 가지고 자료의 全文을 탐색할 수 있다. 탐색향상을 위하여 불리언논리와 인접탐색기법(proximity operators)을 사용한다. 이 시스템은 미국연방법원과 각주(州)법원의 판례, 미합중국의 법령집, 대법원 판결문등을 수록하고 있으며, 현재 미국전역의 데이터베이스중에서 가장 많이 이용되는 것들 중 하나로 꼽히고 있다.[12] LEXIS 서비스는 전용선을 통하여 英國에서도 이용되

10) *BRS Database Catalog* 1989, (Latham: BRS Information Technologies) pp.3~5.
11) J.A. Sprowl. "WESTLAW VS LEXIS: Computer-Assisted Legal Research Comes of Age," *Program*, 15(1981), 135.
12) W.A. Katz, op. cit., p.115.

고 있다. WESTLAW는 법률자료출판사인 West Publishing Co.가 1975년부터 제공하고 있는 시스템이며, 원래는 사건명, 인용사항, 주제명, West사가 부여한 서두문(headnotes)등을 수록한 색인시스템이었으나, LEXIS사와의 경쟁력을 강화하고 원문을 곧바로 이용하고자 하는 이용자요구가 반영되어 1978년부터는 全文검색서비스도 아울러 제공하고 있다.[13] 위의 두 가지 시스템은 동일한 자료를 상당히 중복수록하고 있으나 검색프로그램에는 명확한 차이점이 있다. 즉 LEXIS는 全文만을 수록하며 자연어탐색(free-text searching)만을 허용하며, WESTLAW는 全文과 동시에 여러가지 대용물들(surrogates)도 제공하고 있는 것이다. 두 시스템 모두 불리언논리와 인접탐색기법을 사용할 수 있도록 설계되어 있다. 이 두 시스템의 검색결과를 비교하는 연구들이 계속적으로 이루어지고 있으며 이용의 편리성과 경비문제에 초점이 맞추어지고 있다.[14] 이외에도 판례법, 성문법과 같은 각종 법률과 법규, 판례와 판결문, 기타 법률관계 정보를 제공하는 全文검색시스템으로 영국의 EUROLEX, 독일의 JURIS(Juristisches Information System), 미국의 JURIS(Justice Retrieval and Inquiry System)등이 있다.

신문, 통신정보, 뉴우스잡지의 全文이 상업온라인을 통하여 광범위하게 탐색될 수 있다. 현재 활발히 운영되고 있는 몇몇 시스템을 중심으로 살펴보겠다.

New York Times Information Service는 신문과 시사지기사의 全文검색 시스템으로써 선구자적노력을 해왔으나, 1983년 이후 Mead Data Central사가 제공하는 NEXIS가 New York Times Bank를 인수하여 운영하고 있다. 따라서 현재 New York Times 全文온라인은 NEXIS가 제공하고 있으며, 이 시스템은 이외에도 여러가지 신문과 시사잡지의 全文을 제공하고 있다. 일간지의 全文데이터는 매일갱신되어서 NEXIS를 통하여 이용할 수 있으며 美國, 日本, 中國, 유럽지역으로부터의 통신서비스가 매일 갱신되어 제공되고 있다. 따라서 이시스템을 통하여 거의 모든 세계적인 사건정보에 접할

13) C. Tenopir *Retrieval Performance in a Full-Text Journal Article Database* (Ph. D. Dissertation, Uni. of Illinois, 1985), p. 9.

14) S.E. Larson, "Computer Assisted Legal Research," *ARIST* 15(1980), 258~259.

수 있다.[15] 이 시스템은 여러 온라인全文서비스중 가장 발전성이 있고 전망이 밝다고 말할 수 있으며, 신문이외에도 20여개의 정기간행물을 온라인으로 제공하고 있다. 또한 NEXIS를 통하여는 브리태니커백과사전과 몇몇 참고자료와 약간의 정부간행물등도 全文검색이 가능하다. 그러나 참고자료는 CRT상으로만 全文가독할 수 있으며 저작권문제와 연관지어 출판사가 인쇄출력을 허용치 않고 있다.[16] NEXIS의 텍스트들은 불리언연산자와 단어인접 탐색기법을 사용하여 자연어본문탐색(free-text searching)을 할 수 있으며, 통제 색인어는 첨가되어 있지 않다. 그러나 통제어휘의 몇몇 잇점들을 자동적으로 제공할 수 있도록 소프트웨어내에 약간의 통제어휘체제를 지니고 있다. 즉 단어가 3개 이상의 문자를 가지게 되면 단어의 복·단수와 소유격형태가 자동적으로 탐색되며, 영미의 철자법차이와 중국어를 로마자화 할 때의 여러가지 철자변형들이 자동적으로 탐색되고 모든 동등관계어휘(equivalent)들이 검색된다. 이용자가 행정부기관 명칭의 두문자를 입력시키면 완전한 명칭이 검색되며, 그 반대의 경우도 가능하다.[17] NEXIS와 LEXIS는 관련된 탐색단어를 포함하는 문장만도 검색할 수 있으며, 이때 탐색단어들이 터미널의 스크린상에서 더 밝게 돋보여 눈에 띄게 된다.[18]

The Dow Jones News/Retrieval Service (DJN/R)는 1970년대 중반에 일반온라인 서비스를 시작하였으며 1989년 현재 DJN/R서비스는 뉴우스, 全文검색서비스, 산업과 기업에 관한 정보, 산업과 기업에 관한 통계 및 예상정보, 주식시장정보, 비지니스서비스, 쇼핑등 일반서비스로 구분지을 수 있다. 이중 全文검색서비스는 주로 1984년 이후에 시작된 서비스로서 The Wall Street Journal, Washington Post, Barron, DOW JONES NEWS의 全文기사검색을 할 수 있으며, 이외에도 미국내의 여러가지 경제관계 출판물의 全文을 온라인으로 제공한다. 이들은 모두 90초내의 최신 비지니스나 금융정보를 제공하고 있는데 뉴스기사나 주식정보를 원문으로 제공하며, 이러한

15) C. Tenopir "Full-Text Databases," *ARIST* 19(1984), 218.
16) W.A. Katz, op. cit., p.131.
17) C. Tenopir *Retrieval Performance...*, p.11.
18) C. Tenopir "Newspaper Online," *Library Journal*, 109(Mar., 1984), 452~453.

자료는 실제로 DOW JONES의 인쇄물 서비스에 앞서 온라인으로 이용할 수가 있을 정도로 신속한 것이다. 데이터는 시스템에 90일 동안만 보관하고 오프라인으로 보내지지만 선정된 내용은 계속 보관한다. 국내에서는 DJNR과 독점계약을 맺은 중앙일보사의 JOINS를 통하여 이 서비스를 이용할 수 있다.[19]

VU/TEXT Information Services는 7개의 일간지의 全文을 온라인으로 제공한다. 이 시스템은 全文과 통제어휘의 질을 높이며 탐색과정을 편리하게 하는 장치로써 마스터통제어휘집과 색인어필드를 제공하고 있다.

InfoGlobe Search Service는 최초의 온라인 신문인 카나다의 Toronto Globe and Mail지의 全文온라인을 제공한다. 1977년부터 현재까지 자연어 본문탐색기법으로 검색할 수 있으며, 통제어휘는 첨가되어 있지 않다. 신문의 발행 당일에 온라인으로 이용할 수 있다. 이외에도 News Net System (News NET)은 신문과 USA Today지등 약 250종 이상의 산업계의 뉴우스레터지에 대한 全文을 제공하고 있다.

日本의 朝日新聞데이터베이스는 1986년에 日本 최초로 신문의 全文서비스를 개시하였다. 이 신문의 조간과 석간은 NELSON이라는 컴퓨터시스템에 의해서 제작되고 있으므로 모든 기사는 NELSON에 저장되고 있으며 이를 데이터베이스로 이용하는 것이다.[20] 이 데이터베이스에서는 자연어 탐색을 보강하기 위하여 분류된 키워드도 병용하여 사용할 수 있다. 즉 800여개의 주제와 약 200여개의 국명이 있어서 작은 디스크러스 역할을 하고 있다. 이것을 기사마다 그 주제에 맞추어 부여하고 있으며 이 키워드로 주제범위를 제한한 후 자연어로 검색하면 어느정도 잡음이 방지되는 것이다. 자유어와 키워드는 각각 단독으로 사용하거나 또는 조합하여서 사용하여도 무방하다. 이외에도 日本에서는 컴퓨터신문제작(CTS)의 부산물로 日本經濟新聞과 讀賣新聞의 全文데이터베이스인 NEEDS(Nikkei Economic Electronic Data-

19) JOINS (Joong-ang Online Information & News Service)는 중앙일보사가 운영하는 종합정보시스템으로써 발족하여 1989년 9월부터 서비스를 제공하고 있다. JOINS *Catalog*. (서울 : 중앙일보 중앙 경제신문 데이타뱅크국, 1989)

20) 이관기, 신문기사 정보은행—뉴스데이타베이스 이론과 실제—(서울 : 우정출판사, 1986), p. 99.

base System)와 YOMIDAS(Yomiuri Shimbun Database System)가 최근 全文서비스를 가동하고 있다.

정기간행물류로는 먼저 Harvard Business Review Online (HBRO)을 들 수 있으며 비지니스와 경영쪽의 종합정보를 포함하는 잡지인 Harvard Businessie Revw(HBR)을 온라인화한 것으로 1982년부터 온라인으로 全文을 제공하고 있다. John Wiley & Sons사의 Electronic Publishing Division이 HBR의 온라인자료를 배포하고 있으며, 여기서 HBR을 다섯가지 부분으로 나누어 놓고 있다. 즉 서지사항, 초록, 추출정보(회사명, 기관명, 생산품과 봉사내용, 지리적위치, 회사종류), 全文, 색인어의 다섯 부분이다.[21] 효율적인 탐색을 위하여는 불리언연산자보다 인접 또는 위치연산자를 사용할 것이며, 동의어와 특정한 자연어를 사용하고, 적절한 영역까지 탐색을 제한시키며, 자연어탐색과 통제언어탐색기법을 조합하여 사용할 것을 권하고 있다.[22] 각 全文레코드들은 인쇄잡지의 내용과 정확히 일치하며, 모든 텍스트, 그래픽제목, 참고문헌이 포함되나, 그래픽자체는 온라인화일에 포함되지 않고 있다. HBRO는 BRS와 DIALOG를 통하여 이용할 수 있다.

미국화학회(American Chemical Society)의 잡지 18종이 ACS Journals Online이라는 명칭으로 1983년 이후 BRS를 통하여 서비스되고 있다. 이 시스템은 과학분야 잡지의 全文온라인의 실현에 앞선 실험연구에서 선구자적 역할을 해왔다. 이 데이터베이스는 1980년 이후 현재까지 잡지의 全文을 수록하고 있으며 검색작업을 용이하게 하기 위하여 문절마다 순차번호를 부여하고 있으며, 참고문헌, 각주, 초록이 함께 검색될 수 있다.[23]

The International Research Communications System (IRCS)의 Medical Science Online Database는 IRCS의 32개의 인쇄 의학잡지에 상응하는 내용을 全文온라인으로 제공하는 시스템이다. IRCS시리즈인 이 잡지들은 의

21) John Wily & Sons, Inc. *HBR/Online: User Guide.*(N.Y.: John Wiley & Sons, Inc. 1982), pp.4~5.
22) Ibid., pp.28~29.
23) S.W. Terrant, "Online Searching: Full Text of ACS Primary Journal," *Journal of Chemical Information Computer Science.* (Nov., 1984) : 230~235.

학과 생의학분야의 모든 분야를 포괄하고 있다.[24] 각 기사는 약 1,000 단어 정도로 목적, 방법론, 결과부분으로 구성되며, 이 각각은 독립적으로 탐색할 수 있다. 따라서 전형적인 잡지기사 탐색보다는 요약형 자료나 백과사전의 항목탐색과 매우 유사하다. 모든 표와 삽화명도 온라인화일에 수록된다. 인쇄잡지에서 기사가 수록되는 잡지의 주제로 논문의 주제를 대략 분류할 수 있다. 이러한 주제분류코드는 온라인화일에서 대략적인 색언어 역할을 하게 된다.[25]

Comprehensive Core Medical Library (CCML)는 25개의 기본적인 의학관계 매뉴얼, 참고도서, 텍스트북과 10개 이상의 기타 의학출판물들로 구성된 패키지이다. 자료는 인쇄물출판 당일에 온라인으로 접근가능하고 의학참고텍스트들은 정기적으로 갱신된다.[26] 이 시스템은 핵심적 의학자료를 계속적으로 수록할 계획이다. 그림, 사진, 그래픽등은 아직 全文온라인화일에 포함되지 않고 있으며, 1983년부터 BRS를 통하여 탐색이 가능하다.

Mead Data Central사가 제공하는 MEDIS는 미국의학협회등 여러기관들이 출판하는 50개 이상의 의학관계 잡지의 全文데이터베이스이다.[27] 이 시스템은 주로 의학전문가와 전문의가 많이 이용한다.

Magazine ASAP와 Trade & Industry ASAP 모두 Information Access Co. (IAC)가 생산해낸 데이터베이스이며, 여기에는 여러 출판사가 출판한 잡지 약 130여종의 全文이 수록되어 있다. 즉 여러가지 대중잡지로부터 수준높은 전문적인 잡지에 이르기까지 다양한 종류가 포함된다. 탐색자는 자연어로 本文을 탐색하거나 IAC에 부가된 주제명 표목을 이용하여 검색할 수 있다.[28] 비적합자료검색을 최소화하기 위하여 단어인접탐색기법과 동일문단내에 출현하는 단어의 탐색으로 제한시키는 것이 효율적이다.[29] 이들은

24) L.R. Garson and M. Stanley, *User's Manual Primary Journal Database ACS Full-Test File.* (Washington, D.C.: American Chemical Society, 1983), p.295.
25) International Research Communications System, "IRCA Speeds Publication in Powerful Database Link," *IRCS Medical Science,* 11(1983), 189~192.
26) W.Katz, op. cit., p.135.
27) J.M. Homan, "End-User Information Utilities in the Health Sciences," *Bulletin of the Medical Library of Association,* Vol. 74, No.1, (1986), p.34.
28) W. Katz, op. cit., 132.

1984년 이후 Dialog에 수용되어 서비스되고 있다.

참고도서류로써 몇몇 종류가 全文온라인서비스 되고 있으며 브리태니카백과사전은, NEXIS를 통해 제공된다. 이외에도 참고도서로써는 Academic American Encyclopedia, Directory of Graduate Research, Mental Measurement Yearbook, Encyclopedia of Associations, Peterson's Guide to College and Universities, Marquis' Who's Who 등이 BRS와 Dialog를 통하여 全文서비스 된다. 국내에서는 두산그룹산하의 정보통신팀이 1989년 10월에 「동아원색세계대백과사전」 31권 전권을 全文데이터베이스로 내어 놓았는데, 이것은 백과사전을 컴퓨터출판으로 출판한 연후에 그 시스팀을 全文데이터베이스로 活用한 것이다. 현재는 두산그룹의 임직원만을 이용자집단으로 하여 퍼스널컴퓨터를 이용한 가정서비스를 하고 있으며, 항목명만을 색인어로 구성하여 메뉴식으로 제공하므로 항목별로 本文탐색을 할 수 있다. 그러나 本文의 자연어탐색은 실용화되지 않고 있으며, 따라서 접근점은 인쇄자료와 동일하다. 현재로써는 자연어탐색을 할 수 있도록 시스팀을 개선시키려는 구체적인 계획은 없으나 이같은 백과사전 정보서비스를 일년간 시범운영한 후 미비점을 보완하여 내년경부터 일반인에게도 서비스를 확대할 것을 고려중이다. 이것이 실현되면 일반이용자는 퍼스널컴퓨터와 전화를 서로 연결시켜주는 모뎀을 설치하여 가정에서 이 서비스를 이용할 수 있을 것이다.[30] 또는 DACOM이나 KIETLINE등이 두산측과 계약을 맺어 일반서비스를 제공하는 방법도 있을 것이다.

Ⅲ. 전문데이터베이스의 자연어시스팀과 검색성능

1. 통제언어색인과 자연어색인

全文데이터베이스의 서비스가 행해지고 있으나, 어떠한 방법이 가장 효율

29) Dialog Information Services, Inc., "Now Available: Magazine ASAP, Trade and Industry ASAP," *Chronology*, 12(May, 1984), 107~109.

30) 「두산그룹 사무개선본부 정보통신팀」의 담당자들과의 면담을 통하여 이 사실이 밝혀졌음 (1989. 10. 17)

적이고 경제적인가에 대하여는 아직도 합의에 이르지 못하고 있다. 본고에서는 기존의 서지데이터베이스의 검색방법으로 사용하여 왔던 통제언어색인과 자연어색인의 여러가지 측면을 비교하여 보고 全文탐색시 검색효율을 향상시키기 위한 여러가지 방법들을 검토하여 보고자 한다.

통제언어색인은 첫째, 용어간의 어의적 문제를 효율적으로 해결하며, 어의적 관계를 이용하여 일차적인 탐색어이외에 다른 용어를 탐색어로 추가하므로써 탐색영역의 확장이 가능하다. 즉 어의적으로 관계있는 용어들을 모아줌으로써 망라적인 탐색을 가능하게한다. 둘째, 스코프노트(scope note)를 이용하여 색인자와 이용자 모두에게 명확한 안내를 한다. 스코프노트란 색인어의 이용에 대하여 간략하게 지시하는 문장이며 의미가 매우 애매하거나 용도를 한정시킬 필요가 있는 용어에 대하여 추가하며 SN이라는 표시기호를 사용하기도 한다. 세째, 계층구조와 상호참조를 통하여 관련개념을 잘 식별할 수 있다. 네째, 특정개념을 항상 같은 용어로 색인하므로 일관성있는 주제표현을 제공한다. 즉 동의어, 유사동의어, 근사동의어 등을 통제하여 주므로써, 색인과정과 검색과정에서의 주제분산을 방지할 수 있다. 그러나 통제언어색인으로 全文검색을 할 경우, 통제언어는 자연어에 비하여 용어의 특정성이 덜하므로 주제의 구체적 표현이 어려우며, 시스팀을 효과적으로 이용하기 위해서는 통제어휘집에 익숙해야 하며, 따라서 색인자와 이용자측의 기술과 경험이 필요하다. 이는 정보전문가에게는 적합하나 일반이용자에게는 불편한 점이라고 할 수 있다. 또한 통제언어색인은 계속적 갱신을 하므로써 전문주제분야의 어휘변화를 반영시켜야 하며, 색인작성자의 인력경비와 통제어휘집의 작성을 위한 정보분석과 이의 입수를 위한 시간지연 등을 문제점으로 들 수 있다.[31]

자연어색인은 경제적이고 자동화된 운영체제이며, 탐색기법이 용이하여 이용자가 복잡한 통제어휘체계에 익숙해져야 할 필요가 없다. 그리고 本文을 신속히 자동적으로 입력시킬 수 있을 뿐만 아니라 정보입수시에도 시간

31) E. Perez, "Text Enhancement-Controlled Vocabulury vs. Free Text—," *Special Libraries*, Vol.73, No.3, (July, 1982), 186.

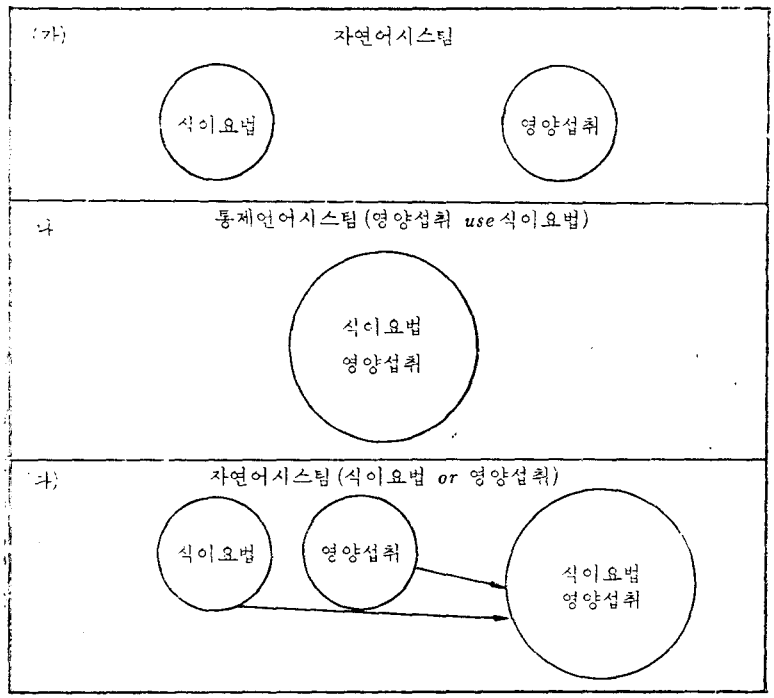
지연이 거의 없다. 또한 자연어시스팀은 검색시스팀내에서 정보유실이나 비효율적인 포맷의 변화가 거의 없으며 모든 정보를 검색하여 볼 수 있다. 또한 정보를 분석하거나 색인, 초록등 가공을 하지 않으므로 특정성을 잃지 않으며, 인간색인자가 필요치 않으므로 입력과정에서 색인자의 주관적인 오류가 개입되지 않는다는 장점이 있다. 뿐만 아니라 단어가 본문에 나타나는 즉시 자연어(free-text)시스팀에 수록되어 단어수록에 시간지체가 없으므로 통제어휘집에 아직 수록되지 않은 새로운 표현과 색인어도 탐색할 수 있다는[32] 잇점이 있다. 자연어시스팀의 가장 큰 문제점으로 지적되는 것은 본문에 나타나는 모든 단어는 동일한 검색가치를 가지는 접근점이 되므로 실제로 이 단어가 주제어가 아닌 비적합문헌도 검색되는 결과를 가져오며 따라서 검색의 효율성이 떨어지게 된다는 점이다. 이외에도 적절치 못한 단어의 조합, 부정확한 용어관계등으로 인하여 잡음을 일으켜 정확율을 저하시킬 수 있다. 자연어시스팀을 효율적으로 이용하기 위하여 이용자는 동의어의 중요성을 인식하고, 구문(synax) 지식도 있어야 하며, 따라서 탐색자는 어느정도 지적인 부담을 안게 된다. 경험이 적은 탐색자가 본문출현단어만 가지고 탐색하는 데에는 문제가 있다. 탐색자는 全文탐색시 많은 단어통정을 행해야 하며 문헌에서 사용되는 다양한 용어로 인하여 중요한 것을 빠뜨리게 되거나 또는 많은 비적합 문헌을 검색하게 되기도 한다.

그림 1은[33] 자연어시스팀과 통제언어시스팀을 비교하고 있다. 특정 데이터베이스에 "자양물"이란 주제를 취급하는 100여개의 초록이 있다. 이중 "식이요법"이란 단어는 50개의 문헌에서 나타나며 "영양섭취"란 단어는 또 다른 50개의 문헌에서 나타나고 있다. 그림 1(가)에서 보듯이 자연어시스팀은 "식이요법"과 "영양섭취" 두가지로 문헌을 구분한다. 그러나 통제언어시 스팀은 두 단어를 동의어로 취급하여 100개의 문헌을 함께 취급하고 있다(그림 1(나)). 자연어시스팀은 그림 1(다)에서처럼 '식이요법 *or* 영양섭취'라는 논리조합을 이용하여 두 가지 주제의 특성성을 유지하면서 동시에

32) W. Katz, op. cit., p.96.

33) F. W. Lancaster *Vocabulary Control for Information Retrieval*, 2nd ed. (Arlington, Vir.: Information Resources Press, 1986), p.164.

〈그림 1〉 자연어시스템과 통제언어시스템 비교



통제언어시스템이 행하는 것과 같은 정도의 포괄적인 탐색을 제공하고 있음을 알 수 있다. 이처럼 자연어시스템에서 이용자는 통제언어시스템에서 취할 수 있는 수준의 망라성을 얻음과 동시에 필요하다면 여러개의 세분된 주제구분을 유지시켜 특정성을 잃지 않고 문헌을 입수할 수도 있음을 알 수 있다. 물론 이것은 초록을 대상으로 한 제한된 범위의 조사이므로 여기서 보여준 자연어시스템의 망라성과 특정성이 전적으로 全文데이터베이스에 적용된다고 말할 수는 없을 것이나, 적어도 자연어시스템의 유연성과 더 많은 가능성을 엿볼 수는 있다고 생각된다. 자연어색인은 개념이라기 보다는 용어자체이므로 용어색인이라고도 하며, 반면에 통제언어색인은 색인대상이 개념이므로 개념색인이라고도 한다. 자연언어색인은 또한 색인자가 통제어휘집을 참고하지 않고 문헌으로부터 또는 자신의 지식을 바탕으로 하여 임

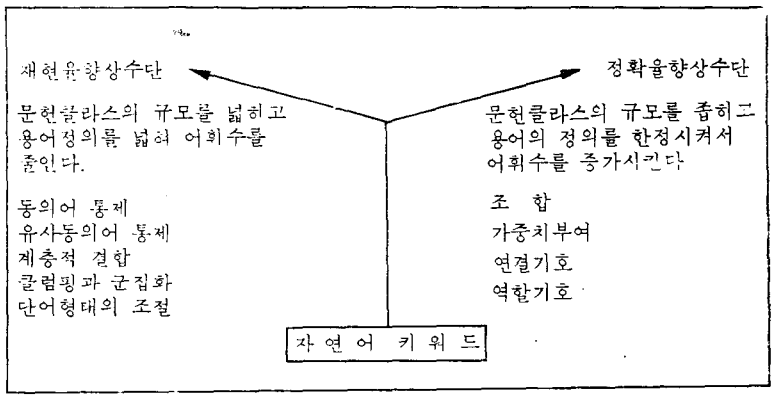
의로 색인어를 선택한다는 의미에서 자유색인이라고도 한다.[34]

2. 자연어시스템의 검색효율성에 영향을 미치는 요인

全文데이터베이스가 자연어시스템을 사용할 경우 검색효율성의 지배요인으로는, 여러부가적인 방법의 도입여부, 本文과 더불어 표제, 초록등의 부가가치적영역(value added field)의 입력여부와 또는 검색작업시의 탐색전략 탐색자의 자질, 탐색출력형태등을 들 수 있다. 이외에도 이용자와 시스템의 상호작용이나 처리과정에서 발생할 수 있는 인간적 실수등도 효율성에 영향을 미치는 요인이 될 수 있다. 자연어시스템의 효율성을 높이기 위해 사용되는 부가적인 방법은 本文에서 출현한 단어를 기본적 색인어로 보고 여기에 재현율이나 정확율을 향상시키기 위한 방법을 도입하는 것이다. Cranfield 척도에 따라 全文데이터베이스의 검색효율성을 평가하였을 때 일반적으로 자연어탐색이 통제언어를 사용하는 것보다 높은 재현율과 낮은 정확율을 보여주는 것으로 나타났다.[35),36),37),38)] 따라서 적합 문헌이 많이 검색되기는 하나 동시에 비적합문헌도 많이 검색된다는 문제를 안게 되는 것이다. 全文데이터베이스의 재현율을 유지하면서, 단점으로 나타난 정확율을 향상시키는 방법이 여러가지로 모색되고 있다. 표 1[39)]에서 보는 바와 같이 조합, 가중치부여, 연결기호나 역할기호 등으로 검색에 부가가치를 부여하여 정확율을 개선시키는 연구들이 이루어지고 있다. 그러나 이러한 방법들은 아직도 일반적인 실용화가 되지 못하고 있으며, 획기적인 자동언어향상방안이 없는

34) 정영미, 정보검색론, (서울 : 정음사, 1987), p. 54.
35) C. Tenopir *Retrieval Performance...*, p. 163.
36) Jung Soon Ro, "An Evaluation of the Applicability of A Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. I. On the Effectiveness of Full-Text Retrieval," *JASIS* Vol. 39, No. 2, (1988), 73~78.
37) Jung Soon Ro, "An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. II. On the Effectiveness of Ranking Algorith ms on Full-Text Retrieval," *JASIS*, Vol. 39, No. 3 (1988), 147~160.
38) 남영준, 全文데이터베이스의 檢索效率性分析(중앙대학교 대학원 석사학위논문, 1987), p. 42.
39) F.W. Lancaster, *Information Retrieval Systems: Characteristics, Testing, and Evaluation* (N.Y.: John Wiley, 1968), p. 85. 최수연, "색인언어의 효율성 측정에 관한 연구(1)," 정보관리연구, Vol. 11, No. 5. (1978), 148에서 재인용.

〈표 1〉 재현율과 정확율의 향상방법



한 기존의 全文데이터베이스 생산자들이 이러한 장치들의 추가로 인한 경비부담을 기꺼워하지는 않을 것이다. 그러므로 탐색자들은 現在의 시스팀상황하에서 全文탐색을 위한 최대의 방법을 모색하여야 한다. 단어인접탐색기법과 단어절단법과 같은 기존의 탐색기법은 全文탐색시 검색효율성을 향상시킬 수 있다. 또한 단어출현빈도나 단어위치데이터등이 사용되기도 한다. 현재 全文데이터베이스를 가장 잘 탐색할 수 있는 방법으로 불리언연산자보다는 단어인접탐색을 사용할 것과 주요개념을 탐색하기 위하여 여과(filter) 역할을 하는 초록이나 통제어휘탐색을 병용할 것을 추천하고 있다.[40]

全文데이터베이스에서는 本文에 나타난 형태 그대로의 단어가 도치화일을 구성하는 색인어가 되므로 같은 의미라도 여러개의 다른 단어로 표현할 수 있다. 따라서 실제 질문으로부터 추출해낸 탐색어와 본문의 단어가 그대로 일치하기가 어려우므로 탐색어리스트 작성시에 하나의 탐색어와 관련된 단어를 모아 전부 탐색어로 포함시켜야 효율을 높이게 된다.

온라인 텍스트의 가독성과 미적인 외양은 인쇄자료에서처럼 중요하며 검색효율에도 영향을 미친다. 디스플레이 기능에 따라 이용자의 수용성이 다

40) C. Zuga, "Full Text Databases: Design Considerations for the Database Vendor," *In Proceedings of the 7th International Online Information Meeting*. (Oxford: Learned Information, Ltd. 1983), pp. 427~434.

르게 나타나며, 검색효율에 있어 탐색전략만큼이나 중요한 요인으로 작용한다. 효과적인 디스플레이 방법으로 탐색어를 더욱 눈에 띄게 하고 각 탐색어에 대해 볼 수 있는 텍스트의 양을 정할 수 있어야 하며, 필드마다 명확하게 명칭을 붙일 것을 권하고 있다.[41] 또한 탐색자가 문장에서 탐색어를 대강 훑어보고 특정문헌을 더 많이 가독하기 위해서 어떤 지점에서라도 멈추어 볼 수있도록 디스플레이 장치가 용이해야 한다. 자동색인시 과학기술문헌의 자연어가 인문사회과학문헌의 자연어보다 더 많이 색인어로 지정되는 경향이 있으며, 따라서 주제가 자연어검색시스템의 성능에 커다란 변수가 될 수도 있다.[42] 즉 인문과학분야를 全文化하여 자연어 키워드 시스템을 사용할 경우 언어의 추상성, 애매성, 함축성 때문에 검색효율을 크게 기대하기가 어려우리라고 본다.[43]

3. 탐색용디소러스

자연어검색시스템에서 검색효율을 높일 목적으로 사용되는 소프트웨어로서 탐색용디소러스(searching thesaurus)를 들 수 있다.[44] 이는 자연어디소러스나[45] 사후통제어휘집(post-controlled vocabulary)[46]이라고 불리워지기도 하며, 전통적인 디소러스와 비교하여 볼때, 탐색용디소러스는 탐색시의 탐색도구로만 사용되며 입력시 어휘를 표준화시키는데 사용하지는 않는다. 또한 동의어, 유사동의어, 어형변화, 어의적관련어등을 모아주는 정도로 대략적인 구성으로 되어 있다는 점이 다르다. 탐색용디소러스의 作成方法은 몇개월동안 탐색자가 시스템을 사용토록 놓아두고, 시스템이 그동안 개개 탐색자가 행한 탐색전략을 저장토록 하여, 이를 분석하여 탐색용디소러스를 만들기 위한 자료로 삼는다. 또는 도서관시스템 전문가가 그들의 경험에 비추어 필

41) *Ibid.*
42) M.E. Rowbottom and P. Willett, "The Effect of Subject Matter on the Automatic Indexing of Full Text," *JASIS*, Vol.33(3), (1981), 141.
43) 최수연, "색인언어의 효율성 측정에 관한 연구(2)," Vol.11, No.5 (1978), 189.
44) E. Perez, op.cit., p.189.
45) F. 윌프리드 랑카스터, 정보검색시스템, 제 2 판, 윤구호·김태승역(서울 : 구미무역, 1985), pp.314~316.
46) F.W. Lancaster, *Vocabulary Control...*, pp.165~169.

요하다고 생각되는 사항이나 일반적인 참고과정에서 중요하게 취급되는 사항들을 모아 수작업으로 만들 수도 있다. 이처럼 기본적인 디소러스가 만들어지면 이것을 탐색전략에 이용하도록 하며, 그후 탐색자들이 추가로 개발한 탐색전략을 모아 디소러스를 계속 확장시켜 나가는 방법을 취한다. 결과적으로 여러가지 탐색에 유용하게 사용된 개념그룹들을 조합하므로써 디소러스는 점점 확장될 수 있을 것이다. 탐색용디소러스는 수작업으로 만들거나 또는 入力된 탐색문을 일정기간 저장시킨 후 出力시켜서 연결된 탐색어들을 하나의 관련어 집단에 속하도록 처리하는 기계적인 방법이 있다. 따라서 일반 디소러스처럼 인쇄물형태나 기계가독형 데이터베이스로 생산될 수 있다. 탐색자가 자연어시스팀을 이용할 때마다 동일한 개념그룹을 반복적으로 형성해야 하므로 지적인 부담을 가지게 되며, 또한 탐색자가 특정주제에 대한 포괄적인 탐색을 위하여 필요한 모든 단어들을 생각해내기도 쉽지 않다. 자연어검색시스팀에서 이 탐색용디소러스를 함께 병용하면, 동의어와 유사어 변형된 철자들, 약어, 관련어등의 참조가 나와 있으므로 자연어가 가지는 용어의 특정성을 유지하면서 동시에 포괄적인 탐색이 가능하다.

4. 혼합형시스팀

全文데이터베이스를 자연어검색과 작은 규모의 통제어휘검색을 혼용하여 두가지 검색방법의 잇점을 취하는 실용적인 절충형시스팀,[47] 또는 혼합형(hybrid)시스팀이 있다[48] 이 경우 수백개의 단어로 구성된 비교적 포괄적인 개념의 통제어휘가 시스팀의 전반적인 기본구조를 구성하게 된다. 문헌은 하나 이상의 포괄적인 색인어나 표제나 본문에 출현한 자연어로 색인된다. 따라서 탐색시 자연어는 특정성을 높여주며, 통제어휘는 망라성을 높여주게 된다. 이처럼 제한된 통제어휘와 자연어를 함께 사용하면 높은 검색효율을 얻을 수 있다. 이러한 시스팀이 경제성도 있는 것으로 나타났다.[49] 즉 주제코드가 별 어려움없이 대략적으로 주어질 수 있으며, 색인자가 기억할 수

47) E. Perez, op. cit., p. 190.
48) F. Lancaster, *Vocabulary Control...*, p. 175.
49) *Ibid.*, p. 176.

있을 정도의 주제코드 수로써 충분하므로 어휘리스트를 매번 조사할 필요가 없으므로 경제적으로 이용될 수 있다. 이러한 대략적인 색인접근을 통하여 탐색자는 비적합자료까지도 훑어보아야 하는 수고를 덜 수 있게 된다. 많은 신문들과 New York Times지는 全文검색시 이용할 수 있는 독립된 레코드 영역에 일반표목으로 구성된 작은규모의 통제어휘시스팀을 주고 있다. 이 색인어들은 아주 세밀하지는 않으나 그들 상호간에 불리언논리조합을 사용하거나 자연어와 함께 조합하여 사용될 때 자연어탐색만을 행할때의 약점이었던 일반적이고 폭넓은 개념검색까지를 제공한다. 이처럼 작은 규모의 통제어휘집을 자연어 시스팀과 병용하는 신문데이터베이스들로는 美國의 Chicago Sun-Times가 있으며, 약 150여개로 구성된 통제어휘집을 가지고 있다. Star & Tribune지는 500여개 New York Times지는 약 950여개로 구성된 통제어휘집을 가지고 있으며,[50] 日本의 朝日新聞은 800여개의 주제코드와 200개의 國名을 합쳐 1,000여개의 작은 디소러스를 가지고 자연어 탐색시에 혼합 사용하여 검색성능을 향상시키고 있다.[51]

Ⅳ. 全文데이터베이스의 이용자 분석

1. 이용자 요구

여러가지 상업적인 全文데이터베이스를 실험대상으로 하여 이용자의 요구사항이나 반응을 조사한 연구들이 있으며, 이들의 결과를 결집하여 시스팀에 반영시키는 것도 시스팀향상의 한 수단이라고 볼 수 있다.

백과사전의 온라인화는 영국과 미국등을 비롯한 몇몇 국가에서 실시하고 있는데 실시초기에는 출판사의 경계의 대상이 되었다. 백과사전의 온라인화가 실용단계로 접어들면 현재까지 종이로 만들었던 기존 백과사전의 판매실적이 떨어질 것을 우려했던 까닭이다. 따라서 초창기에는 학교나 도서관 이외의 이용자에게만 이용을 제한시켜 놓았으나, 종이로 만든 백과사전에 비

50) E. Perez., op. cit., p.191.
51) 이관기, op.cit., p.99.

하여 풍부한 액세스가 가능하기 때문에 점차 영역이 확대되어 가는 추세이다. 한편 종이로 인쇄된 백과사전도 존재가치가 있기 때문에 자유화를 단행해도 공존할 수 있을 것이라는 의견이 제기되고 있다.[52]

온라인잡지의 이용자들은 이를 꼭 인쇄잡지의 대체로써 사용하는 것이 아님이 밝혀졌다. 이용자들은 全文을 온라인으로 가독하지 않으며 서지서비스의 이용처럼 서지정보를 찾고 그것이 적합한지 여부를 판단하기 위하여, 즉 탐색향상수단으로 주로 이용한다. 그 다음에 그들은 서고에 가서 원문을 찾아 읽는다.[53] 이것은 부분적으로는 本文의 팩시밀리전송의 불완전함과 터미널의 빈약한 화질수준으로부터 기인된 것이기도 하다. 즉 이용자는 온라인잡지에 그래픽이 수록되지 않은 점에 실망하며, 인쇄잡지에는 그래픽이 수록되므로 인쇄물과 온라인全文 모두를 사용하는 경향이다. 과학자들은 그들의 연구분야에서 특정잡지 서명이 빠져있거나, 가장 최신 잡지호수가 아직 들어와 있지 않은점에 대하여 불만을 가지는 것으로 나타나며, 이러한 것은 全文데이터베이스 서비스를 확대시키므로써 요구에 부응할 수 있을 것이다.[54]

법률분야의 全文검색시스템의 이용자는 탐색결과가 적합자료에 더하여 비적합자료도 많이 검색될지라도 이를 망라적인 완전한 탐색이 이루어졌다는 표시로 받아들이며, 혹시 놓쳐 버릴 수도 있는 자료까지도 검색할 수 있다는 점에서 全文데이터베이스를 기꺼이 선호하고 있는 것으로 나타났다.[55]

의학분야의 全文데이터베이스 MEDIS의 이용자에게 출력정보의 적합성여부를 물은 결과 절반정도가 긍정적인 반응을 보였으며, 약 23퍼센트 정도는 부정적으로 응답하였다. 또한, 25퍼센트는 무응답이었는데 이중 약 절반정도는 출력정보의 적합성 여부에 관계없이 탐색방법이 편리하다고 지적하였다.[56] 또한 의학분야의 全文데이터베이스인 IRCS에 대하여 의학분야의

52) 한국전자통신연구소. 산업사회에서 정보사회로. (대전 : 한국전자통신연구소 ; 1985), pp. 68~69.

53) K.J. Winkler, "Chemical Society Offers Full Text of Scholarly Journals by Computer," *The Chronicle of Higher Education*, (June, 1983), 23~24.

54) C. Tenopir, "Full-Text Databases," p. 224.

55) S.E. Larson and M.E. Williams, op. cit., pp. 258~259.

56) M. Collen and C.D. Flagle, Full-Text Medical Literature Retrieval by Computer, *Journal of American Medical Association*. Vol. 15, (1985), 2768~2774.

전문적인 종사자 및 연구자들을 대상으로 행한 조사에서는 검색자료의 약 50 내지 70퍼센트가 부적합했다는 반응을 나타냈다.[57] 이처럼 같은 의학주제분야 일지라도 데이터베이스의 종류나 이용자집단에 따라 다른 반응을 보이고 있음에 유의해야 할것이며, 이용자의 요구에 영향을 미치는 요인에 대한 연구가 더욱 필요하다고 사려된다.

특허全文데이터베이스를 이용하는 특허조사자들은 온라인 특허정보에 만족해하며 本文에 나타나는 자연어로 탐색하는 것을 선호하는 경향이다. 특허全文데이터베이스 이용자들은 유용성을 충분히 인식하고 더 많은 특허全文온라인을 요구하고 있다.[58]

全文데이터베이스 탐색시 가장 두드러진 장점은 이용자가 적합성판단을 즉시 할 수 있다는 점이다. BRS는 全文탐색시 탐색단어의 출현빈도를 보여주며, 탐색단어를 포함하는 문단만을 인쇄출력하도록 서비스하고 있는데, 과학자들은 이러한 방법으로 적합성여부를 판단하는 것은 너무 성급한 것이라는 반응을 보이고 있다.[59] 대규모의 데이터베이스에 대한 연구가 더 행하여져 가장 좋은 디스플레이 방식을 발견할 필요가 있다.

온라인잡지나 특허자료를 全文데이터베이스로 이용하는 이용자가 일반적으로 만족을 나타내고 있으며, 따라서 全文데이터베이스의 시장성은 충분히 있다고 볼 수 있다. 과학자들은 최신자료와 많은 잡지서명을 검색하기를 원하고 있으나, 과학자의 全文데이터베이스 이용이 아직은 인쇄잡지를 대체할 정도까지 이르지는 못하고 있다. 그러나 몇가지 기술적 문제점이 보완되고 경제성을 갖추게 되면, 이러한 과학자의 요구를 감안하여 볼때 인쇄물 대신 全文데이터베이스를 상당히 이용하게 되리라는 것을 예측할 수 있다. 현재 우리가 알고 있는 全文데이터베이스 이용자에 대한 지식은 기존 탐색시스템

57) J. Franklin, M. Buckingham, and J. Westwater, "Biomedical Journals in an Online Full Text Database: A Review of Reaction to ESPL," *In: 7th International Online Meeting, London,* (1983), pp. 6~8.

58) D. Stein, et al., "Full Text Online Patent Searching: Results of a USPTO Experiment," *In Proceedings of the Online 1982 Conference.* (Weston, C.T.: Online Inc.,) 1982, pp. 289~294.

59) C. Tenopir, "Full-Text Databases," p. 224.

상에서 이용되는 全文데이터베이스에 대한 이용자반응을 집약한 것이 대부분이다. 이처럼 경험에 의한 연구뿐만 아니라, 더 조직적이고 체계적인 이용자연구가 이루어져 시스템에 반영되어야 하리라 생각된다.

2. 도서관에서 全文데이터베이스의 활용

도서관이나 정보센터에서 이용자에게 全文서비스를 제공하는 방법은 대략 두 가지로 나눌 수 있다. 첫째, 도서관이 BRS나 DIALOG, ORBIT 등 일반네트워크에 가입하여 이들이 보유하고 있는 全文데이터베이스를 온라인으로 이용하거나, 全文專用네트워크인 LEXIS나 NEXIS등을 통하여 온라인서비스를 이용할 수 있다. 현재 이용할 수 있는 온라인全文서비스에 대하여는 이미 제Ⅱ장에서 언급한바 있다. 이들을 이용할 경우 대부분의 온라인全文데이터베이스가 1970년대 중반이후 자료를 全文데이터베이스화 하였으므로 소급자료처리문제와 컴퓨터이용시간이 길어지므로 이에따른 이용자의 경비부담을 고려해야 한다. 뿐만 아니라 全文만을 수록하고 있는지 또는 통제어휘나, 초록, 서명등이 추가된 부가가치 全文데이터베이스인지를 확인하고 복합적인 접근점을 제공하는 全文서비스를 선택하는 것이 유리하다. 온라인全文데이터베이스의 이용시 다운로딩(downloading)이 가능한지 여부를 확인하고 가능하다면 이를 이용하여 도서관 자체의 데이터베이스를 구축 하는 것이 경제적이다. 그러나 때로는 다운로딩을 허용하지 않거나, 심지어는 디스플레이만 허용하고 인쇄출력조차 허용치 않는 시스템도 있으니,[60] 사서나 탐색자는 시스템의 규정을 사전에 잘 알고 있어야 함은 물론이다. 도서관에서 全文데이터베이스를 제공하는 두번째 방법은 패키지형 전자출판물인 광디스크와 CD-ROM형태로 제작된 全文데이터베이스를 구입하여 이용하는 것이다. 이러한 축적매체는 마이크로 컴퓨터를 이용하여 도서관에서 이용할 수 있으며, 온라인통신망을 사용할 때 소요되는 많은 통신비를 절약할 수 있고, 이용할 때마다 매번 비용을 지불치 않으므로 비용을 절감시킬 수 있다. 그뿐만 아니라 기존의 온라인全文데이터베이스에서 축적이나 전송이 용

60) W. Katz. op. cit., p. 131.

이치 않았던 그래픽, 사진, 도표등의 화상정보의 수록이 가능하며 질적으로 더 우수하게 제공된다.[61] CD-ROM 全文데이터베이스들은 자체검색시스팀을 별도의 플로피디스크에 수록하여 제공하거나 광디스크에 全文과 함께 수록하여 제공하며 자기 테이프에 비하여 훨씬 신속히 데이터베이스를 검색 할 수 있다.[62] 이용자가 인쇄백과사전을 이용할 경우는 항목검색만이 가능하나 CD-ROM 전자백과사전에서는 항목검색은 물론 문장, 문단, 인접단어검색까지도 가능하다. 그러나 全文온라인에 비하여 갱신이 신속히 이루어지지 못하여 최신정보검색에 문제가 있으므로 온라인으로 최신정보를 제공하는 併用型으로 이용할 수도 있다.[63] 또한 하드웨어와 소프트웨어의 표준화가 이루어지지 못하였으므로 데이터베이스 종류마다 다른 드라이버를 사용하여, 검색시스팀도 다르고 따라서 이용자는 CD-ROM 종류마다 검색시스팀을 익혀야하는 어려움이 따르게 된다.[64] 全文데이터베이스는 일반적으로 여러개의 디스크에 수록되므로 디스크의 이용이 번거로우며 여러개의 디스크를 사용하여 검색해야 하는 것이 이용자에게는 불편한 점이다. 또한 온라인全文정보검색서비스가 빈번한 도서관에서는 CR-ROM 全文데이터베이스를 설치하여 운영하여도 경제성이 있으나 데이터베이스가 자주 사용되지 않거나 온라인全文利用이 많지 않은 도서관에서는 설치를 고려해야 한다. 최초의 구입설치비에는 추후의 서비스비용까지 가산되어 있으므로 구입비가 비싸지게 되며, 이용이 많지 않은 경우 경제성을 잃게 되는 것이다. 이와같은 문제점들은 계속적인 기술개발로 인하여 현재 부분적으로 해결이 되고 있으며, 머지않아 기술적인 문제는 상당부분 해결이 되리라고 예측된다.

현재 광디스크와 CD-ROM 형태로 제작되어 도서관에서 제공될 수 있는 全文데이터베이스들은 다음과 같다 : Grolier 사가 발행한 전 21권의 미국아카데믹백과사전, 미국화학회의 잡지들, Datetek Corp. 이 생산해 내는 10개

61) 노정순, "美國情報産業의 變化", 국회도서관보 24권 2호(1987), p. 12.
62) W. Katz. op. cit., p. 145.
63) S. Shimbori, "The New Information Media CD-ROM," J. of IPM, Vol. 30, No. 1, (April, 1987), 김종회번역, "새로운 정보매체 CD-ROM," 도서관문화, Vol. 29, No. 4, (1988), 190.
64) 노정순, op. cit., p. 13.

의 신문, 미국국립미술관(National Gallery of Art)의 미술품소장작품과 미술관 안내정보, 日本의朝日新聞, 平凡社大百科事典, 最新科學技術用語辭典. 이외에도 유럽 EC는 유럽지역의 규격정보 全文데이터베이스를 CR-ROM에 수록하는 프로젝트를 추진하고 있다.[65]

국내의 경우, 도서관이나 정보센터의 CD-ROM 드라이버의 보유현황은 아직 미미한 편이며,[66] 全文데이터베이스의 이용을 위하여 쓰이기보다는 LC MARC의 이용이나 Index Medicus등 목록데이터나 서지정보의 이용에 주로 사용하고 있으며, 따라서 열람용 보다는 도서관 내부의 사서들의 자료정리용으로 주로 이용되고 있는 것으로 나타났다. 포항공대에서는 미국아카데믹백과사전을 CD-ROM 全文데이터베이스로 제공하고 있다.[67] CD-ROM 全文데이터베이스를 도서관에서 활용하는 문제는 아직 정확히 평가할 수 없지만, 도서관이나 정보센터에게 있어서 이것이 커다란 가능성을 지닌 자원이라는 사실에 이의를 제기할 수는 없다고 본다.

V. 전망과 결론

현재까지 나타나고 있는 全文데이터베이스의 여러가지 문제점을 종합하여 보면 다음과 같다. 첫째, 全文데이터베이스의 온라인시스팀들은 그림, 사진 그래픽등 도해자료들을 적절히 제공할 수 없다는 기술적인 한계점이 있으며 둘째, 온라인탐색시스팀밴더들은 全文데이터베이스에 적용시킬 탐색기법을 신속히 개발하고 있지 않으며, 서지데이터베이스에서 사용하고 있는 불리언 논리, 도치색인, 단어인접탐색 등에만 주로 의존하고 있다. 全文데이터베이스와 서지나 문헌대용물형태의 데이터베이스 사이에 검색성능의 차이가 있는지 여부와 全文데이터베이스에 적용시킬 가장 좋은 탐색방법에 대한 합의

65) 田屋裕之, op. cit., pp.42~43.

66) 이화여대·한림대·연세대의과대학·과학기술대학·포항공대 도서관과 표준연구소의 자료실이 CD-ROM 드라이버를 갖추고 있다.

67) FAXON과 EBSCO의 CD-ROM 담당자면접결과 현재로서는 국내의 CD-ROM 全文데이터베이스 시장이 개척단계에 있는 것으로 나타났다. (1989.9.29)

가 이루어지지 않고 있다. 세째, 全文검색에 사용되는 자연어색인이 일반적으로 높은 재현율을 보여 검색의 포괄성은 보여주나 정도율이 떨어져 비적합문헌도 많이 검색된다는 문제가 있다. 네째, 이에 따라서 컴퓨터이용시간이 길어지므로 이용자에게는 경비부담이 추가된다. 한 예를들면, 통제언어 탐색에 비하여 자연어탐색은 컴퓨터 이용시간이 세배정도 더 소요되는 것으로 나타났다.[68] 다섯째, 全文데이터베이스의 이용자에 대한 연구는 대부분 선정된 작은 집단을 대상으로 하여 기존의 탐색시스팀의 서비스 능력에 대한 반응조사들이 대부분이므로 아직 정확한 이용자요구 파악이 미흡한 상태이다. 이상과 같은 여러 문제점들이 현존하고 있으나 컴퓨터기술의 급속한 발달로 도해자료의 수록 문제는 근시일내에 해결이 가능할 것으로 전망되며, 全文과 문헌대용물의 검색성능에 대한 비교연구도 진행되고 있다. 또한 경비문제는 全文데이터베이스가 점차 많이 사용됨에 따라 하락하게 될 것이나 이러한 문제에 대한 부분적인 해결책으로 밴더들이 일과후 시간에 부과하는 값싼요금을 이용하여 全文레코들을 다운로드(download)시켜 필요한 全文을 탐색하거나 많이 이용되는 데이터베이스는 CD-ROM 全文데이터베이스형태로 구입하여 사용하는 방법이 있다.[69] 그러나 다운로드를 이용할 경우는 저작권문제가 개입되는지 여부를 고려해야 한다. 미국의 경우 온라인 처리비용이 매년 21퍼센트씩이나 감소하고 있으므로[70] 비용문제는 전망이 그리 어두운 것만은 아니다. 앞으로 全文데이터베이스의 폭넓은 시장개척을 위하여 더 광범위한 폭으로 이용자조사를 확대할 필요가 있다. 일반이용자와 주제전문가는 그 반응과 요구가 각각 다를 수 있으므로, 일반이용자와 주제전문가 모두를 조사집단에 포함시켜야 할 것이다. 연구자들은 기존의 탐색서비스능력에 대한 반응조사 뿐만 아니라, 이용자가 어떠한 탐색이나 디스플레이 방법을 원하고 있는지에 대하여 조직적이고 체계적인 조사연구의 필요

68) D.F. Hersey, et al., "Free Text Word Retrieval and Scientist Indexing: Performance Profiles and Costs," *Journal of Documentation*, 27, (Sept., 1971), 167~183.

69) W. Katz, op. cit., p.130.

70) 이우범, "전자미디어의 개발과 미래도서관의 역할," 국회도서관보. 제 25권 6호(1988.11.12), 42.

성을 인정하고 있다. 全文검색시스템 설계자들과 생산자들은 이러한 연구결과들을 효율적인 시스템 운영에 반영시켜야만 할 것이다.

오늘날의 데이터베이스 환경에 있어서 가장 발전적인 현상은 全文온라인 데이터베이스의 실현이라고 볼 수 있다. 일찌기 1960년 Swanson은 이미 컴퓨터에 의한 본문탐색(text searching)이 인간색인자가 행하는 전통적인 검색보다 더 훌륭하다는 실험결과를 발표하였으며,[71] 이 연구는 全文검색의 가능성에 대한 선구자적 연구로서 평가받고 있다. 이어 1970년에 Salton은 全文자동검색(automatic full-text searching)에 관한 일련의 실험결과를 매우 낙관적으로 보고하였다.[72] 이러한 실험적인 모색단계를 거쳐 지난 수년동안에는 全文데이터베이스들이 상업탐색시스템을 통하여 본격적으로 제공되어 왔으며, 점점 더 많은 수의 全文데이터베이스들이 실용화될 것으로 전망된다. 국내의 데이터베이스계는 아직 정적인 상태를 벗어나지 않고 있으나, 몇몇 신문사들이 이미 컴퓨터인쇄를 계획하고 있으므로 이것이 실행되면, 부산물로 全文데이터베이스가 형성되어지게 될 것이며, 실제로 이에 대한 예비작업으로 JOINS 등이 활약을 시작하고 있다. 국내의 全文데이터베이스시장은 신문의 全文化가 가장 먼저 활성화될 것으로 예측된다. 이러한 시점에서 본고에서는 全文데이터베이스의 발전유래와 현재 상업적으로 제공되는 국내외 서비스의 현황을 살펴보고 자연어시스템의 잇점과 검색성능을 향상시키는 방안, 분야별 전문데이터베이스 이용자의 요구, 도서관과 정보센터에서 全文데이터베이스의 활용방안등 미래지향적인 정보서비스에 대하여 고찰해 보았다. 본 논문은 全文데이터베이스에 대한 기초적 연구이며, 앞으로 全文데이터베이스의 검색을 향상시키는데 반영시킬 수 있는 실험적연구들이 행하여질 것과 특히 한글의 全文데이터베이스화에 대비하여 한글문헌을 대상으로 하는 자연어탐색분야의 연구가 다양하게 이루어질 것을 기대한다.

71) D.G. Swanson, "Searching Natural Language Text by Computer," *Science* 132, 3434 (Oct, 1960), 1099~1104.

72) G. Salton. "Automatic Text Analysis," *Science* 168, 3929(April, 1970), 335~343.

On the Characteristics and Information Retrieval Performance of Full-Text Databases

Myung-Hi Cho

Abstracts

Appearance of full-text online is the most encouraging phenomenon during the development of databases. The full-text databases of today is derived from by-product of electronic publication of printed materials. Now, there are also some movements toward electronic production of documents in Korea although not powerful.

The present study is designed to examine the characteristics and effective retrieval method of full-text databases now commercially available through various vendors.

The outline of this paper is as follows:

First, background and present situation of existing full-text database services through national and worldwide are examined.

Second, free-text searching system of full-text databases is compared with controlled vocabulary system. The factors influencing on free-text retrieval performance, searching thesaurus, and hybrid or compromising system, which is using limited controlled vocabulary in conjunction with natural language for the enrichment needed for practical operation of the system, are examined.

Third, user demands through the analysis of preceding studies on various types of full-text databases are recognised.

Fouth, application of CD-ROM full-text database to the libraries and information centers is examined as prospective resources for them.

Finally, some problems and prospect of full-text databases are presented.