본문 데이타베이스 연구에 관한 고찰과 그 전망

노 정 순*

〈목 차〉

1. 서 론

1.1. 본문 DB검색분야의 변화

온라인 본문(Full Text) DB가 소개된 이래로 지난 몇년 동안 온라인 본문 DB검색분야에는 수많은 변화가 일어났다. 우선 양적으로 본문 DB와 그 탐색서비스, 소프트웨어의 숫자는 놀라울 정도로 증가했다.

1989년 정보검색 목적으로 대중이 온라인으로 사용할 수 있는 DB수는 4062개로써, 이 중 본문 DB는 34%를 점유하는 것으로 보고되고 있다. 이

* 한남대학교 도서관학과

것은 1984년의 총 2453 DB 중 18%와 비교할 때, 2배의 점유율을 보여주며, 숫자적으로는 3배에 이르고 있다.[1]

본문 DB가 수적으로 증가한 것 뿐만이 아니라 본문 DB를 온라인으로 탐색할 수 있도록 검색시스팀을 제공하는 탐색서비스도 증가하여, 서지(Bibliographic) DB 탐색을 제공하는 기존의 대부분의 탐색서비스가 본문 DB 탐색을 제공할 뿐만 아니라, 본문DB 탐색만을 제공하는 본문전용 검색서비스도 생겼다.[2] DIALOG나 BRS와 같은 기존 서지 DB 탐색서비스에서도 본문 DB의 비중이 점점 높아져서, 1988년에는 각각 54개(전체 290 DB중 18.6%)와 34개(138개 중 25%)의 본문DB를 제공하고 있다.[3] 이 비중이 앞으로 더욱 높아질 것이라는 것은 "1990년대의 DIALOG의 주요 목표는 본문 DB의 숫자를 늘이는 것"이라는 DIALOG의 로저 서미트(Roser Summit) 사장의 이야기에서도 알 수 있다.[4]

온라인벤더서비스의 탐색시스팀을 통해 본문 DB를 탐색하는 이외에 독자적인 본문검색시스팀을 운용할 수 있도록 본문검색소프트웨어도 상품화되었다. 각종 컴퓨터기종에 맞는 소프트웨어를 선정할 수 있도록 판매되는 패키지만도 50개를 넘어섰다.[5] 독자적인 컴퓨터와 소프트웨어를 갖출 능력이 없는 도서관에서는 마이크로개인용컴퓨터(personal computer)와 CD-ROM 드라이버만을 갖추고 자체도서관에서 배치(Batch)탐색을 할 수 있도록 본문 DB가 탐색소프트웨어와 함께 CD-ROM에 담겨 배포되기도 하는데, 1988년 CD-ROM에 수록되어 판매되고 있는 전문 DB는 35개가 넘는다.[6]

본문 DB나 탐색서비스, 소프트웨어패키지, CO-ROM의 이와 같은 숫자

1) *Directory of Online Databases.* (N.Y : Caudra/Elsevier) 1984년판과 1989년판. 점유율은 무작위추출하여 계산한 것임. *Full Text Databases*, by Carol Tenopir & Jung Soon Ro (Westport, Conn. : Greenwood, 인쇄중)에서 재인용.

2) Data Time과 Vu/Text 등

3) Carol Tenopir & Jung Soon Ro, *Full Text Databases*(Westport, Conn. : Greenwood) 출판중.

4) *Ibid.*

5) Robert Kimberley, *Text Retrieval: a directory of software* (Aldershot ; Brookfield, VT : Gower, 1987).

6) Richard A. Bowers, ed. *Optical Publishing Directory, 1988.* Third ed. (Medford, NJ : Learned Information, Inc., 1988).

적인 증가 외에도 컴퓨터기술상의 변화 또한 주목된다. 지난 30년 동안 온라인정보검색용 서지 DB 혹은 본문 DB의 표준화일구조는 도치화일이었고, 컴퓨터에 축적되는 정보는 오직 문자나 숫자정보 뿐이었다. 그러나 광디스크에 의한 그림정보축적, 패러럴컴퓨터와 같은 하드웨어기술과 화일구조의 발달은 본문탐색을 전통적인 DB구조나 소프트웨어, 시스팀의 한계를 뛰어넘게 할 것으로 기대된다. 이와 같은 본문 DB의 양적 증가와, 탐색을 위한 하드웨어 발달, 인공지능기술 개발, 그림정보를 위한 비디오디스크의 등장 등은 본문탐색분야의 새로운 방향을 약속한다.

1.2. 연구의 목적

과학적인 연구란 자연현상 사이의 추측된 관계에 관한 가정적인 주장(명제, 진술)을 체계적이고 통계적으로, 실험에 근거하여, 비판적으로 조사하는 것이다.

본문탐색을 학문으로 연구한다는 것은 본문탐색시스팀을 이루고 있는 요소 즉, 변수들 사이의 관계를 서술하고 설명하며, 설명된 현상에 근거하여 이론을 정립하고, 더 나아가서는 현상 즉, 변수관계를 예측하고 효과적으로 통제하기 위한 것이다. 본문 DB검색에서 가정될 수 있는 변수들의 관계에는 "어떤 것"들이 있을까? 그동안 행하여졌던 본문 DB 검색에 관한 과학적 연구들을 토대로 앞으로 예측해 볼 수 있는 현상은 "어떤 것"들이 있을까? 본문정보검색상의 많은 변화와 컴퓨터기술발달은 변수관계를 가정·예측하는데 "어떤" 영향을 끼칠 것인가? 이 "어떤 것"을 규명하는 것이 본문 DB검색분야에서 앞으로 수행하여야 할 연구과제일 것이다. 본고에서는 이들 질문에 대한 대답을 찾아봄으로써 본문 DB검색분야의 가능한 연구과제와 그 방향을 제시하고자 하는 데 그 목적이 있다.

2. 본문 DB와 본문 DB검색시스팀

본문 DB란 신문기사나 법원의 판결문, 잡지기사, 교과서, 백과사전과 같

은 1차정보원의 텍스트나 데이터 전부를 레코드로 수록한 일종의 일차정보 DB이다.

오늘날 대부분의 본문 DB는 랭카스타(Lancaster)가 이야기한 첫 단계 전자출판형태이다.[7] 서지 DB가 책자형태의 색인지나 초록지의 전자형태대용물인 것과 같이, 본문 DB는 책자로 된 책이나 잡지의 전자출판문이다. Wall Street Online이나 Harvard Business Review Online처럼 한 종류의 신문이나 잡지만을 대상으로 그 신문이나 잡지의 독자적인 DB로써 존재하기도 하고, Magazine ASAP처럼 100여종의 잡지를 통합하여 하나의 DB로 생산되기도 한다.

본문검색시스팀이란 이와 같은 전자형태의 본문 DB를 대상으로 각 문헌본문內의 모든 문장에 있는 모든 단어를 탐색어(=색인어)로 사용할 수 있도록 설계한 자연어검색시스팀이다. DB생산자나 시스팀벤더입장에서 보면, 본문 DB는 전산화된 출판시스팀의 부산물로 얻어질 뿐만 아니라, 별도의 색인작업 없이, 문장內의 모든 단어를 색인어로 사용하기 때문에 수동색인시의 인건비 문제나 색인자간의 비일괄성 문제를 해결해 주는 장점이 있다. 그러나 이용자입장에서 보면 본문검색시스팀은 다양한 접근점을 제공하는 장점은 있으나, 색인언어의 높은 망라성 때문에 정확률이 낮고 검색비용이 많이 드는 검색결과를 가져올 수도 있다. 이러한 단점을 보완하기 위해 본문검색시스팀은 서지 DB검색시스팀과는 다른 탐색특성을 제공하는데, 용어위치제한탐색(proximity searching), 단어출현빈도표, 동의어/단어철자통제, KWIC Display기능 등이 그것이다.[8]

본문 DB검색시스팀과 본문 DB는 매우 다양하여 1989년 통계에 의하

7) F. Wilfred Lancaster, "Electronic Publishing : Its Impact on the Distribution of Information," *National Forum* (Summer 1983) : 3.

8) Carol & Ro, *Full Text Databases* 3章에서 자세히 설명. 용어위치제한탐색이란 복합개념을 이루는 단일개념을 AND로 연결할 때 그 출현위치에 따라 탐색을 제한하는 것으로, 문단제한탐색(Paragraph Searching : 같은 문단內에 나타나는 문헌만 검색), 문장제한탐색(Sentence Searching : 같은 문장內에 나타나는 문헌만 검색), 인접단어 제한검색(Adjacent Words Searching : 인접하여 나타나는 문헌만 검색)등이 있다. BRS에서는 각각 SAME, WITH, ADJ 연산자를, DIALOG에서는 (S), (nN)과 같은 연산자가 사용된다. 이들 연산자를 용어 위치제한 연산자(proximity operator)라 칭한다.

면[9] 16개의 주요시스템에서 1700여종의 본문 DB가 온라인으로 탐색가능하다. 이들 DB는 법률 DB, 정부간행물 DB, 소식지 DB, 신문기사 DB, 통신문 DB, 뉴스레터 DB, 참고도서 DB, 학술잡지 DB, 취미교양잡지 DB로 대별하여, 각각 고유의 특성과 이용대상자, 이용목적 등을 갖는다.[10] 각각의 검색시스템은 이들 각각의 본문 DB를 대상으로 고유의 탐색기능과 특성을 제공한다. 본문 DB는 500 page 이상의 참고도서류에서부터 몇 백 단어로 이루어진 뉴스요약통신문류에 이르기까지, 그 주제와, 문체, 길이, 구조가 매우 다양하며, DB종류에 따라 이용대상자와 탐색전략도 다르리라고 예상되기 때문에, DB에 따라 특정 탐색기능과 탐색특성, 탐색전략의 개발이 요구되는 것이다.

3. 본문 DB검색 연구에 관한 분석

법률DB이나 신문뉴스본문DB를 제외하고, 본문DB는 지난 몇년 사이에 급증하였기 때문에 본문DB에 관한 연구는 아직 광범위하게 수행되지 못하고 초기단계에 있다. 현재까지 본문 DB에 관한 연구문헌은 ① 본문시스템의 타당성과 성능에 관한 이용자 의견조사(user study)나, ② 본문 DB검색의 효율성이나 효과에 관한 평가, ③ 본문검색의 효율이나 효과를 올리기 위한 시스템의 수정, 이 세 분야에 집중되어 있다.

이용자 조사연구는 ① 본문 DB의 시장성에 관한 조사, ② 탐색성능에 관한 이용자의 의견, ③ 탐색효율을 높이고 이용자에게 친숙한 시스템을 설계하기 위한 DB와 시스템 설계상의 주의사항에 관한 문헌으로 대별된다. 그러나 이들 문헌은 체계적이고 과학적인 연구의 결과라기보다는 관찰이나 경험을 토대로 한 것들이다.

과학적 연구는 본문 DB의 검색효율을 평가하는 연구와 검색효율 향상을

9) Ruth M. Orenstein, ed., *Fulltext Sources Online: for Periodicals, Newspapers, Newsletters & Newswires.* (Needham Heights, MA : BiblioData, 1989).

10) Tenopir & Ro, *Full Text Databases.* 1章에서 자세히 설명.

위한 시스템 수정분야에서 주로 이루어졌으나, 대부분이 현장(field)실험이라기보다는 실험실(laboratory)실험에 의해 연구된 것이 특색이다.

3.1. 본문 DB의 가능성을 예측한 연구

본문 DB가 실용화되기 이전, 본문 DB의 가능성과 타당성은 자연어시스템이나 자동초록색인에 관한 몇몇 연구에서 예측되었다. 이들 연구는 ① 초록문이나 제목에서의 자연어탐색시스템이나 ② 본문을 사용한 자동색인시스템이 수동통제시스템보다 결코 못하지만은 않다는 가정 하에 실험되었다.

샐튼(Salton)의 연구[11]는 제목이나 초록문을 사용한 자연어시스템과 통제어시스템 사이에는 현격한 차이가 없다는 것을 증명하였고, 킨(Keen)과 맥길(McGill)의 연구[12]는 이 결과를 확인시켜 주었다. 올리브(Olive) 등의 연구는[13] 통제어시스템은 제목을 사용한 자연어시스템보다 재현율은 증가시키나 정확률에서는 차이가 없다는 것을 보여주었다.

그러나 다른 연구들은 자연어시스템이 통제어시스템에 비해 더 낫다는 것을 밝히고 있다, 두번째 크랜필드연구와 파커(Parker), 에이치슨(itchison)등의 연구[14]는 자연어시스템이 통제어시스템 보다 효율이 더 좋을 뿐만 아니라 경제적인 이점도 가지고 있다고 한다. 클리버든(cleverdon)은 제목이나 초록문에서의 자연어탐색이 통제어검색보다 정확률에서는 별 차이가 없으나, 재현율에서는 높은 결과를 가져온다고 밝힌다.[15] 마키(Markey)는 자연어탐

11) Gerard Salton, "The Evaluation of Computer-Based Retrieval Systems," in *Automatic Information Organization and Retrieval*, ed. by Gerard Salton (New York, NY : McGraw Hill, 1968), pp. 280~349.

12) Michael E. Keen, "The Aberystwyth Index Language Test," *Journal of Documentation* 29(1) (March 1973) : 1~35.
Michael McGill, *An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems* (Syracuse, NY : Syracuse University, 1979)

13) G. Olive, J.E. Terry and S. Datta, "Studies to Compare Retrieval Using Titles with That Using Index Terms," *Journal of Documentation* 29(2) (June 1973) : 108~191.

14) C.W. Cleverdon, J. Mills and E.M. Keen, *Factors Determining the Performance of Indexing Systems*. 2 vol. (Cranfield, England : College of Aeronautics, 1966).
J.E. Parker, "Preliminary Assessment of the Comparative Efficiencies of an SDI System using Controlled or Natural Language for Retrieval," *Program* 5 (1979) : 26~34.
T.M. Aitchison et al, *Comparative Evaluation of Index Language. Part II: Results* (London, England : The Institution of Electrical Engineers, 1970).

15) C.W. Cleverdon, *A Comparative Evaluation of Searching by Controlled Language*

색은 통제어탐색에 비해 높은 재현율을 그러나 낮은 정확률을 가져온다고 보고했다.[16] 그러나 찰톤(Charton), 로우레트(Rowlett), 캐로우(Carrow), 헨즐러(Henzler), 캘킨스(Calkins), 바이른(Byrne)의 연구[17]는 자연어든 통제어든 어느 하나가 독자적으로는 완전한 검색을 제공하지 못하고, 통제어와 자연어를 통합하여 사용하는 것만이 가장 좋은 검색결과를 가져온다고 밝히고 있다.

윌리암스(Williams), 스미스(Smith), 맥길(McGill)의 연구[18]는 또한 자연어탐색과 통제어탐색 두 시스팀 각각에 의해 검색된 문헌 사이에는 중복되는 문헌이 별로 없다, 즉 두 탐색방법은 각각 고유의 문헌을 검색한다는 것을 밝히고 있다.

이상의 연구는 비록 초록문이나, 제목을 사용한 자연어탐색에 대한 연구결

---

and Natural Language in an Experimental NASA Data Base (Washington, DC : National Technical Information Service, 1977)

16) Karen Markey, Pauline Atherton, and Claudia Newton, "An Analysis of Controlled Vocabulary and Free Text Search Statements in Online Searches," *Online Review* 4(3) (1982) : 225~236.

17) Barbara Charton, "Searching the Literature for Concepts," *Journal of Chemical Information and Computer Science* 17(1977) : 45~46.

Russell J. Rowlett Jr., "Keywords vs. Index Terms," *Journal of Chemical Information and Computer Science* 17(1977) : 192~193.

Deborah Carrow and Joan Nugent, "Comparison of Free-Text and Index Search Abilities in an Operating Information System," in *Information Management in the 1980s: Proceedings of the American Society for Informatino Science 40th Annual Meeting: Sep. 26-Oct. 1, 1977*(White Plains, NY : Knowledge Industry Publicatons, 1981), pp. 131~138.

Rolf G. Henzler, "Free of Controlled Vocaularies : Some Statistical User-Oriented Evaluations of Biomedical Information Systems," *International Classification* 5(1) (1978) : 21~26.

Mary L. Calkins, "Free Text or Controlled Vocabulary? A Case History Step-By-Step Analysis... Plus Other Aspects of Search Strategy," *Database* 3 (1980) : 53~67.

Jerry R. Byrne, "Relative Effectiveness of Titles, Abstracts, and Subject Headings for Machine Retrieval from the COMPENDEX Services," *Journal of the American Society for Information Science* 26(4) (1975) : 223~229.

18) Martha E. Williams, "Analysis of Terminology in Various CAS Dafa Files as Access Points for Retrieval," *Journal of Chemical Information and Computer Sciences* 17 (1977) : 16~20.

Linda C. Smith, "Selected Artificial Intelligence Techniques in Information Retrieval Systems Research"(Ph. D. dissertation, Syracuse University, 1979).

McGill, *An Evaluation.*

과를 보여주지만, 본문이 초록문이나 제목보다는 문헌의 내용을 더 완전하고 정확하게 표현한 대표적인 표현물이라는 것을 생각한다면, 초록이나 제목을 사용한 연구결과를 본문에 적용시켜 예측하는 것이 무리는 아닐 것이다.

본문을 사용한 자동색인에 관한 연구 역시 본문에서 축출된 색인어가 지닌 주제어로서의 가치를 예견하고 있다. 이 분야의 연구는 주로 1) 수동색인에 의해 배정된 색인어 중 몇%가 자동색인으로 얻어지느냐에 관한 연구,[19] 2) 정확률과 재현율에 의해 수동색인과 자동색인을 평가한 연구,[20] 3) 본문을 사용하여 자동색인이론들을 비교 평가한 연구,[21] 이 셋으로 대별된다. 이들 연구의 결과는 자동초록으로 축출된 본문의 단어들은 검색효율을 저하시키지 않는다는 것이다. 연구결과는 연구방법에 따라 약간씩 다르나, "수동색인과 다르지 않다.[22] "초록을 사용한 자동색인보다 낫다.[23] "제목을 사용한 자동색인보다 낫다.[24] "타당성이 있다"[25] "큰 가치가 있다"[26]는 등 대체로 긍정적인 결론을 내리고 있다. 이들 결론은 초록문이나 제목에 나타나는 단어보다는 본문에 나타난 단어가 더 주제어로써 적합하다는 것을 보여준다.

3.2. 이용자조사 연구

이용자 조사연구는 온라인본문탐색의 시장가능성에 대한 조사를 목적으로

19) S. Artandi and E.H. Wolf, "The Effectiveness of Automatically Generated Weights and Links in Mechanical Indexing," *American Documentation* 20(1969) : 198~201.
Miranda Lee Pao, "Automatic Text Analysis Based on Transition Phenomena of Word Occurrences," *Journal of the American Society for Information Science* 29 (May 1978) : 121~124.
Weinberg, "Word Frequency."

20) Salton, "The Evaluation."

21) Fred J. Damerau, "An Experiment in Automatic Indexing," *American Documentation* 16 (1965) : 283~289.
John M. Carroll and Robert Roeloffs, "Computer Selection of Keywords Using Word-Frequency Analysis," *American Documentation* 20 (July 1969) : 227~233.

22) S. Artandi, "Computer Indexing of Medical Articles," *Journal of Documentation* 25 (1969) : 214~223.
Artandi and Wolf, The Effectiveness."

23) G. Salton, *The SMART Retrieval System: Experiments in Automatic Document Processing* (Enflewood Cliffs, NJ. : Prentice Hall, 1971).

24) H.S. Heaps, *Information Retrieval: Computational and Theoretical Aspects* (New York : Academic Press, 1978), pp. 269~280.

25) Pao, "Automatic Text."

26) Wolfgang K.H. Sager and Peter C. Lockemann, "Classification of Ranking Algorithms," *International Forum On Information and Documentation* 1(4) (1976) : 12~25.

수행되었다. 미국화학학회의 ACS 잡지본문 DB의 온라인탐색 타당성에 관한 연구, Elsevier Science Publisher의 ESPL 잡지본문 DB의 시장성조사연구, 패겔(Pagell)의 Magazine ASAP DB에 대한 특성 조사, 테노피어(Tenopir)의 Magazine ASAP의 이용과 잠재적 이용자에 관한 조사 등이 그것이다.

ACS 연구[27]는 온라인탐색경험이 없는 최종이용자(End-User)들이 한두 시간의 훈련으로 온라인본문검색을 만족스럽게 수행할 수 있으며, 본문탐색은 다른 탐색방법보다 강력하다는 긍정적인 반응을 보이고 있으나, 온라인 잡지에 그래픽과 같은 그림정보가 빠진 것, 탐색비용이 비싼 점도 지적하고 있다. 이용자들은 본문 DB를 인쇄물잡지를 소장하는 대용품으로 보는 것이 아니라 탐색과정의 진보·향상으로 보고 있었다. 이러한 연구결과는 다른 후속 연구에서도 확인되고 있으며, 이외에도 Elsevier 연구는[28] 문헌에서 사용된 실험방법이나 참고문헌을 조사하기 위해 본문탐색이 사용되며, 본문 DB를 가지고 온라인으로 즉석에서 검색된 문헌의 적합성판단을 할 수 있다는데에 큰 가치를 부여하고 있다. MEDIS 연구에서는[29] 이용자의 탐색경험이 늘수록 탐색시간이 짧아진다는 사실을 부연하고 있다. 패겔(Pagell)의 연구는[30] 전자형태의 잡지 DB인 Magazine ASAP에는 그래픽만 빠져 있는 것이 아니라, 잡지기사도 선택적으로 선별되어 수록되어 있음을 발견하고,

27) Kay Durkin et al, "An Experiment to Study the Online User of a Full Text Primary Journal Database," in *Proceedings of the 4th International Online Information Meeting, London, Dec. 1980* (Oxford, England : Learned Information, Ltd., 1980), pp. 53~56.

Seldon W. Terrant, Lorrin R. Garson, and Barbara E. Meyers, "Online Searching Full Text of American Chemical Society Primary Journals," *Journal of Chemical Information and Computer Science* 24 (1984) : 230~235.

28) J. Franklin, M.C. Buckingham, and J. Westwater, "Biomedical Journals in an Online Full Text Database ; a Review of Reaction to ESPL," in *Proceedings of the 7th International Online Information Meeting, London, Dec. 1983* (Oxford, England : Learned Information, Ltd., 1983), pp. 407~410.

29) Morris F. Collen and Charles D. Flagle, "Full-Text Medical Literature Retrieval by Computer," *Journal of the American Medical Association* 254(19) (Nov. 15, 1985) : 2768~2774.

30) R. Pagell, "Searching Full-Text Periodicals : How Full is Full?" *Database* 10 (Oct. 1987) : 33~38.

R. Pagell, "Searching IAC's Full-Text Files : It's awfully Confusing," *Database* 10 (October 1987) : 39~47.

인쇄잡지 대용물로써 사용되기에 적절치 못함을 지적하고 있다. 역시 Magazine ASAP의 사용도를 연구한 테노피어(Tenopir)는[31] 다른 문헌에서 과소평가된 browsing기능을 지적한다. 본문은 주제에 대한 주변정보를 얻을 목적으로 문헌의 본문을 browsing하는데 주로 사용되며, 이용자들의 온라인 탐색 후 도서관에 가서 인쇄잡지로 본문을 찾는 대신 온라인으로 다운로드(Download)하는 것을 선호한다는 것을 발견했다. 만일 온라인탐색에서의 비용문제만 해결된다면 이 browsing하고 Downloading하는데 본문을 사용하는 것이 본문 DB의 주요 사용목적이 될 것이라고 지적하고 있다.

웨이저스(Wagers)는 BRS에서의 Magazine ASAP를 사용하여, 문헌의 특성과 주제분야, 탐색용어, 통제언어의 특정성이 본문 DB사용결정에 어떤 영향을 끼치는지를 조사하였다.[32] 연구결과 문헌이 어떤 주제에 대해 일반적으로 다루고, 일반적인 제목을 갖고, 그렇게 초록·색인이되었지만, 직접 관련된 특정사항에 대해서는 특별히 언급할 때 ; 색인어가 너무 제한적이어서 어떤 탐색개념을 표현하는 여러 형태의 용어를 포착하지 못할 때 ; 한 분야에서 여러가지 다양한 토픽이 매우 세부적으로 다루어졌고 그렇게 제목붙여지고 색인되었기 때문에 탐색시 계층구조의 상위개념을 고려해야 할 경우 ; 탐색자들은 본문탐색을 선호한다고 밝히고 있다. 또한 탐색자들이 관련 문헌을 망라적으로 검색하고 싶거나, 특정사실이나 질문에 대한 대답을 문헌에서 직접 찾고 싶을 때 본문탐색을 선호하였다.

탐색자의 탐색행위에 관한 연구는 효과적인 user-friendly system을 설계하기 위한 기초조사이다. 마치오니니(Marchionini)는 CD-ROM에 실린 백과사전을 사용하여 초보자의 본문DB 탐색행위를 연구하였다.[33] 국민학교 학생을 대상으로 상급자가 더 탐색을 성공적으로, 경제적으로 수행하며, 초

31) Carol Tenopir, "Users and Uses of Full Text Databases," in *Proceedings of the International Online Meeting, London, December 1988*(Oxford, England : Learned Information, 1988), pp.263~270.

32) R. Wagers, "The Decision to Search Databases Full Text," in *Proceedings of the 10th International Online Information Meeting, 1986* (Oxford, England : Learned Information, Ltd., 1986), pp.93~107.

33) Gary Marchionini, "Information-Seeking Strategies of Novices Using a Full-Text Electronic Encyclopedia," *Journal of the American Society for Information Science* 40(1) (1989) : 54~66.

보자는 Interactive 탐색전략[34]을 사용하며, 대부분의 이용자들은 시스템의 디폴트(defaults)를 따른다고 밝히고 있다.

3.3. 본문 DB의 검색효율에 관한 연구

3.3.1. 법률 DB

본문 DB에 관한 체계적이고 과학적인 연구는 본문검색시스템의 효율을 평가하는 연구에서 시작되었다. 이들 검색효율에 관한 평가연구는 대부분이 비교평가로써, 본문검색의 효율을 초록이나 제목에서의 자연어탐색이나 디스크립터에 의한 통제어 탐색의 효율과 비교하고 있다. 물론 이들 연구는 본문 DB 중 가장 역사가 오랜 법률분야에서 제일 먼저 시작되었고 또한 가장 활발하게 수행되었다.

대표적인 연구로는 미국의 LITE 시스템에 관한 평가연구, 미국 BAR 재단연구, 영국 옥스포드 실험, RESPONSA 프로젝트, 펠스(Fels)의 연구, 블레어(Blair)와 마론(Maron)의 연구 등이다. 이들 연구의 공통성은 Responsa와 블레어(Blair)와 마론(Maron)의 연구를 제외하고, 법령집의 권말색인을 사용한 수동탐색과 비교하여 본문 DB 탐색을 평가한 점이다.

연구의 결과는 연구에 따라 크게 차이를 보인다. 수동 권말색인검색과 비교할 때, LITE 시스템에서는 본문탐색이 정확률과 재현율이 모두 더 좋았다.[35] 미국 BAR 재단연구에서는 재현율에서는 본문탐색과 권말색인탐색이 모두 비슷하게 잘 수행됐으나 정확률에서는 수동권말색인 탐색이 본문탐색보다 2배나 좋았다.[36] 옥스포드실험에서는 본문탐색이 수동탐색보다 재현율은 높으나 정확률이 낮은 결과를 가져왔다.[37] 펠스(Fels)의 연구에서는 Mooers 율이라는 효율측정기준을 사용하여 수동탐색이 본문탐색보다 우수함을 보여준다.[38]

34) Interactive 탐색전략이란 스크린에 나타나는 탐색결과를 보고 원하는 결과를 얻을 때까지 탐색을 수정하면서 수행하는 탐색행위에 관한 전략.

35) Richard P. Davis, "The LITE System." *Judge Advocate General Law Review* 8(6) (Nov./Dec. 1966) : 6~10.

36) W.B. Eldridge, "An Appraisal of a Case Law Retrieval Project," in *Computer and the Law*, ed. by David Johnston (Kingston, 1968).

37) Colin Tapper, *Computer and the Law* (London : Weidenfield and Nicloson, 1973).

38) Eberhard M. Fels. "Evaluation of the Performance of an Information Retrieval Sys-

한편 RESPONSA 연구와 블레어(Blair)와 마론(Maron)의 연구는 비교평가방법을 사용하지 않고 절대평가에 의해 본문 DB검색효과를 평가하고 있다. RESPONSA 연구는 본문검색의 재현율을 100%로 할 때 정확률은 34%이다, 즉 모든 관련문헌을 모두 검색하기 위해 탐색어를 계속 추가하거나 탐색을 확대할 때 비관련문헌이 66%나 검색된다는 것을 밝혔다.[39] 블레어(Blair)와 마론(Maron)의 연구는 본문 DB검색이 높은 정확률과 낮은 재현율을 갖는다고 결론내리고 있다.[40]

이와 같이 연구에 따라 그 결과가 크게 다른 이유로는 각각의 연구에서 사용된 DB의 크기, 탐색질문의 종류(특정성)과 크기, 탐색시스팀 등 연구방법상의 차이 때문으로 생각된다. 색인언어의 망라성이 클수록 재현율은 높지만 정확률은 낮은 검색결과를 가져온다는 종래의 이론에서 생각하면, 본문 DB는 서지 DB에 비해 색인언어의 망라성이 크고, DB의 크기가 커질수록 망라성이 커지기 때문에, 큰 규모의 본문 DB에서의 검색은 높은 재현율과 낮은 정확률을 초래할 것으로 예측된다. 그러나 블레어(Blair)와 마론(Maron)의 연구결과는 표면적으로는 이러한 인식에의 대도전으로 보이지만 자세히 분석해보면 연구방법상의 특성에서 기인된 것 임을 알 수 있다.

블레어(Blair)와 마론(Maron)이 언급한 대로, 높은 정확률과 낮은 재현율이라는 우리의 기대(다른 연구결과)와는 상반된 결과를 초래한 첫째이유는, DB의 크기가 크면 검색된 문헌수도 많아질 것이다는 탐색자들의 생각 때문이었다. 이 믿음 때문에 탐색자들은 탐색문에 평균 4~5개의 개념을 AND로 연결하여, 축소탐색을 했던 것이다. 즉 정확률과 재현율의 반비례관계를 생각할 때, 이 연구에서는 의도적으로 높은 정확률을 위한 탐색으로 탐색식을 만들었기 때문에 재현율은 그만큼 저하되었다고 이해된다. 그러나 DB의 크기가 커질수록 많은 문헌이 검색될 것이라는 생각은 샐튼(Salton)에 의

tem by Modified Mooers Plan," *American Documentation* 14 (1963) : 28~34.

39) Aaron M. Schreiber, "Computerized Storage and Retrieval of Case Law without Indexing ; the Hebrew Responsa Project," *Law and Computer Technology* 2 (Nov. 1969) : 14~21.

40) D.C. Blair and M.E. Maron, "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System," *Communications of the ACM* 28(1985) : 289~299.

해 반박된 바 있다.[41]

높은 정확률과 낮은 재현율이 초래된 또 다른 이유는 법률분야에서는 주제를 탐색하는데 모든 관련문헌에서 사용됐을 법한 주제표현어(키워드)를 모두 예측하기가 더 어렵다는 것이다. 특정사건의 이름 대신 "사건"이라고만 부른다든지, 사람이름 대신 "피고"라는 단어를 사용한다든지가 그 예이다. 실제로 한개의 주제개념을 탐색하는 데는 3개의 동의어가 사용됐는데, 검색되지 않은 관련문헌을 조사한 결과 재현율을 100%로 하기 위해서는 26개의 다른 동의어를 더 사용했어야 했다고 밝히고 있다. 만일 이 연구에서 모든 동의어를 모두 OR로 연결하여 사용하여 관련문헌을 하나도 남김없이 모두 검색했다면(즉 재현율=100%) 이 시스팀의 정확률은 몇 %였을까? RESPONSA 연구가 "34%"라고 그 대답을 준다.

또 다른 분석에서[42] 블레어(Blair)는 적합성판정척도도 검색효과에 영향을 준다는 사실을 밝히고 있다. 위의 연구에서 적합성 판정은 "아주 관련있다" "관련있다" "약간 관련있다" "관련없다"는 4개의 척도를 사용하였다. 처음 3개에 해당되는 것을 관련문헌으로 간주하여 정확률과 재현율을 산출한 결과 정확률은 79%, 재현율은 20%였다. 그러나 이 적합성 판정 중 "아주 관련있다"라고 판정된 문헌만을 적합문헌으로 간주했을 경우엔 재현율이 20%에서 48.2%로 올라가고, 정확률은 79%에서 18.2%로 내려간 것을 발견했다. 즉 적합성 판결척도를 달리할 경우 본문 DB검색은 반대로 높은 재현율과 낮은 정확률을 가져왔다.

3.3.2. 학술잡지 DB

법률분야에서의 연구와는 달리 학술잡지 DB에 관한 대부분의 연구는 본문탐색을 디스크립터탐색과 비교평가하고 있다. 여기서 사용된 DB는 1 page짜리 비평기사부터 긴 잡지기사에 이르기까지 다양하다.

초기의 연구는 주로 짧은 잡지기사의 본문을 가지고 행하여졌다. 스완슨(Swanson)의 연구는[43] 본문탐색은 수동색인시스팀보다 좋다, SMART

41) Gerard Salton, "Another Look at Automatic Text-Retrieval Systems," *Communications of the ACM* 29(7) (1986) : 648~656.
42) David C. Blair, "Full Text Retrieval : Evaluation and Implications," *International Classification* 13(1) (1986) : 18~23.
43) Don Swanson, "Searching Natural Language Text by Computer," *Science* 132 (Oct.

system[44]은 본문탐색이 초록문탐색보다 우수하다고 밝히고 있다. 반대로 허세이(Hersey)의 SSIE 연구에서는 본문탐색보다는 주제코드를 사용한 수동색인시스팀의 평균 재현율이 30~40%, 정확률은 15~20%나 높다고 보고되어 있다.[45] 그러나 스완슨은 평균 1000 단어로 된 리뷰기사, SMART 시스팀은 평균 1380 단어로 된 ADI 기사, SSIE 연구는 한 페이지짜리 연구프로젝트를 사용하여 조사하였기 때문에, 이들 연구결과를 잡지본문 DB에 적용시키는 데는 조심스럽다. 긴 잡지기사를 대상으로 후속연구들이 수행되었다.

클리브랜드(cleveland)는 본문 DB 탐색을 독자적으로 사용하는 것을 다른 탐색과 섞어 사용할 때와 비교하였다.[46] 본문탐색은 초록탐색이나 제목·참고문헌탐색보다는 좋으나, 이들 탐색을 독립적으로 하는 것보다는 제목과 초록과 참고문헌을 함께 사용하여 탐색하거나, 초록과 참고문헌을 함께 사용하여 탐색하는 것이 정확률과 재현율면에서 월등함을 발견하였다.

스타인(Stein) 등은 탐색어로 사용되는 내용의 주요어는 본문의 어디에 나타나느냐를 연구했다.[47] 본문탐색으로 검색된 문헌의 87%는 본문 중 결론부분과 서술부분만 가지고 탐색하여 얻을 수 있다고 밝혔다.

테노피어(Tenopir)의 연구는[48] 본문탐색은 제목탐색, 초록탐색, 디스크립터탐색에 비해 높은 재현율과 낮은 정확률을 가져오며, 각각의 탐색은 다른 탐색으로부터는 얻을 수 없는 고유의 관련문헌을 검색한다고 밝히고 있다. 그러나 이 연구는 BRS 시스팀에서 본문 DB의 낮은 정확률을 우려하여 AND 연산자 대신으로 사용하기를 권고하는 "SAME"연산자를 사용하여, 탐

1960) : 1099~1104.

44) G. Salton and M.E. Lesk, "Computer Evaluation of Indexing and Text Processing," *Journal of the Association of Computing Machinery* 25 (Jan. 1968) : 8~36.

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

46) Donald B. Cleveland, Ana D. Cleveland, and Olga B. Wise, "Less Than Full Text Indexing Using a Non-Boolean Searching Model," *Journal of the American Society for Information Science* 35 (Jan./Feb. 1984) : 19~28.

47) D. Stein et al, "Full Text Online Patent Searching ; Results of a USPTO Experiment," in *Proceedings of the Online '82 Conference, Atlanta, Nov. 1982* (Weston, CT. : Online Inc., 1982), pp. 289~294.

48) Carol Tenopir, "Retrieval Performance in a Full Text Journal Article Database"(Ph. D. dissertation, University of Illinois, 1984)

색어들이 같은 문단內에 나타나는 문헌만을 검색하도록 본문검색을 축소하여 문단검색을 하였다.

애보트(Abbott)와 스미스(Smith)는 화학관계 본문 DB인 ACS Journal Online (CFTX)을 서지 DB인 CA Search (CHEM)와 비교하여 평가하고 있다.[49] 탐색을 먼저 서지 DB에서 수행한 후 본문 DB에서 다시 수행하였다. 본문 DB탐색시에는 서지 DB탐색시의 탐색문을 그대로 사용하지 않고 필요한 경우 수정하였다. 연구결과 서지 DB사용시 사용된 탐색전략이 본문 DB탐색에서도 적합하지는 않으며, 서지 DB의 정확률(53%)이 본문 DB (36%)보다 높았다. 또한 문단연산자에서 문장연산자, 인접연산자로 탐색을 제한시킬수록 정확률은 좋아지지만, 적합문헌을 놓치는 것도 많아지므로 지나치게 제한하지 않도록 경고하고 있다. 즉 문단제한(SAME)탐색에서 검색된 적합문헌의 50%이상이 문장제한(WITH)탐색에서는 검색되지 않았다. 또한 서지 DB에서 검색된 적합문헌과 본문 DB에서 검색된 적합문헌사이에는 중복도도 낮다는 것이 발견되었다.

같은 본문DB를 사용한 러브(Love)와 가슨(Garson)의 연구는 높은 정확률을 위한 탐색(Brif 탐색)에서의 본문 DB탐색 사용을 긍정적으로 보여준다.[50] 50% 이상의 정확률을 갖는 탐색이 전체탐색중 88%이고, 100% 의정확율을 보인 탐색만도 전체탐색의 23%나 됐다고 보고하고 있다.

3.3.3. 일반교양잡지 DB

일반교양 취미잡지 DB는 일반독자를 대상으로, 특히 개인적 혹은 학교관련 정보를 원하는 사람이나 학생에게 유용한 DB이다. 문체나, 주제, 길이는 잡지에 따라 다르지만, 일반적으로 전문용어나, 초록문, 각주, 참고문헌을 포함하지 않는다는 것이 학술전문잡지 DB와의 차이점이다.

일반잡지본문 DB에 대한 연구는 주로 Magazine ASAP와 Trade & Indu-

49) John P. Abbott and Charles R. Smith, "Full-Text and Bibliographic ACS Databases: Rivals or Companions?" in *Proceedings of the 6th National Online Meeting, 1985* (Medford, NJ : Learned Information, Inc., 1985), pp.5~9.

50) Richard A. Love and Lorrin R. Garson, "Precision in Searching the Full-Text Database-ACS Journals Online," in *Proceedings of the 6th National Online Meeting, 1985* (Medford, NJ : Learned Information, Inc., 1985), pp.273~282.

stry ASAP database를 대상으로 수행되었다. Magazine ASAP는 100여 일반취미잡지의 기사를, Trade & Industry ASAP는 120여 산업경제잡지의 뉴스와 기사를 모은 DB로써, 모두 DIALOG, BRS, NEXIS를 통해 탐색이 가능하다.

테노피어(Tenopir)는 DIALOG 탐색시스템을 통해 Magazine ASAP를 본문탐색할 때 사용가능한 4가지 탐색전략의 효율성을 조사하였다.[51] 탐색주제를 이루고 있는 두개 이상의 단일개념을 합쳐 복합개념으로 표시하는데 그 개념들이 본문 어디에서든지 나타나기만하면 그 문헌을 검색하라(AND 연산자), 같은 문단에 나타날 때만 검색하라((S)연산자), 10단어 이상 떨어져 나타나면 안되고 그 이내에 나타날 때만 검색하라((10N)연산자), 다섯 단어 이내에 존재할 경우만 검색하라((5N)연산자), 이 4가지의 탐색연산자의 비교가 그것이다. 물론 탐색전략을 확장시킬수록 재현율이 높고 정확률은 낮아져, AND연산자는 다른 연산자보다 높은 재현율과 낮은 정확률을 초래하였다. 그러나 가장 제한적인 전략인 (5N)연산자는 문단제한(S)연산자에 비해 재현율 뿐만 아니라 정확률까지도 더 낮은 탐색을 수행하였다. 즉 지나친 제한은 많은 관련문헌을 검색하지 못한다고 결론짓고 있다. 가장 비용적으로 효율적인 탐색은 문단제한이었지만, 문단으로 제한했을 경우 AND로 검색된 관련문헌의 47.9%는 검색되지 못했다. 다양한 잡지들이 한데 모인 DB에서는 잡지의 종류(성격)으로 탐색결과를 제한하는 방법이, 필요없는 문헌이 검색되는 것(false drops)을 방지하는 좋은 방법이라고 암시하고 있다.

슈(Shu)는 Trade & Industry ASAP 본문 DB를 DIALOG와 BRS에서 각기 탐색하여 그 결과를 평가함으로써, 본문 DB탐색시스템이 본문 DB검색의 효율성에 끼치는 영향을 분석하였다.[52] DIALOG의 (S)와 (10N)이

51) Carol Tenopir, "Search Strategies for Full Text Databases," in *Proceedings of the 51st Annual Meeting of the American Society for Information Science, Atlanta, GE, October 1988* (Medford, NJ : Learned Information, 1988), pp.80~86.

52) Man Evena Shu, "Retrieval Performance of Proximity Operators (Full Text Databases): A System Comparison—DIALOG and BRS," in *Proceedings of the 52nd Meeting of the American Society for Information Science, Washington, DC, November 1989* (Medford, NJ : Learned Information, 1989), in press.

BRS의 SAME과 WITH 연산자와 비교되었다. 문단제한연산자 (S)와 SAME은 두 시스템에서 똑같이 문단(paragraph)을 인지하므로 두 시스템간에 차이가 없다. 그러나 문장(sentence)에 대한 인지는, BRS에서는 문장 끝을 나타내는 구둣점인 마침표(.)를 사용하여 문장단위를 인지하나, DIALOG에서는 마침표를 무시하여 처리하기 때문에 문장을 인지할 수 없다. 그러므로 DIALOG에서는 (nN)이라는 연산자를 사용하여 두 단어가 n개 이내의 단어를 사이에 두고 떨어져 있을 경우 검색하라는 명령으로 문장제한을 대신한다, 연구결과 DIALOG가 데이타를 더 자주 update하여 최신성을 유지함에도 불구하고 BRS의 WITH가 DIALOG의 (10N)보다 높은 정확율을 가왔다. 잡지의 종류나 기사의 종류에 대한 값도 DB제작시 마련하여 두었다가, NOT 연산자를 사용하여 잡지나 기사의 종류로 탐색결과를 제한한다면 정확율을 더 향상시킬 수 있다고 제안하고 있다.

3.4. 검색효율 향상을 위한 시스템 수정 연구

이용자조사연구가 온라인 본문DB탐색의 가능성을 긍정적으로 보여주고 있지만, 본문검색효율에 관한 실험연구는 본문탐색이 주의깊게 수행되어야 한다는 것을 보여준다. 보다 많은 본문 DB가 온라인으로 사용 가능하게 되자 많은 이용자들은 탐색경험을 바탕으로 본문 DB의 탐색효율을 증진시키기 위한 여러가지 방법을 제안하였다.

밀스테드(Milstead)와 덕키트(Duckitt), 페레즈(Perez), 스프로울(Sprowl) 등은 본문탐색에서 용어통제를 위한 간단한 통제어휘집을 사용할 것을 제안하고 있다.[53] 이러한 용어통제는 신문이나 뉴스레터 본문DB에서 실용화되고 있는데, 250여 뉴스레터에 대한 본문 DB 탐색시스템인 NEWSNET 시스템의 경우엔 뉴스레터의 주제에 따라 두 자리 수의 주제코드를 사용하고 있고, 뉴욕타임즈 본문DB는 950용어, Star & Tribune (미니아폴리스)지는 500여 용어, Sun Time(시카고)은 약 150여 용어로 된 용어집을 사용한다.[54]

53) Jessica Milstead, "Indexing the News," in *Proceedings of the American Society for Information Science 43rd Annual Meeting, 1980* (White Plains, NY : Knowledges Industry Publications, Inc., 1980), pp. 149~151.
Pauline Duckitt, "The Value of Controlled Indexing Systems in Online Full Text

토시그노트(Tousignaut)는 본문탐색에 "facet 색인" 방법을 응용하는 것을 소개하고 있다.[55] 약품에 관한 본문 DB에서 본문을 이루고 있는 각각의 facet 단위(즉 성분/함량, 투약, 적응증, 용량, 용법, 부작용, 알레르기, 유효기간 등)로 본문을 나누어 코드化한 후 탐색시 이 코드번호로 탐색을 제한하는 방법이다.

이러한 간단한 통제어휘집을 사용하는 이외에도 본문검색효율을 증진시키거나 이용자가 사용하기 쉬운 시스템을 설계하는데 고려되야 할 사항들도 시스템전문가들로 부터 제안되었다.

잭슨(Jackson)은[56] 터미널에 탐색을 쉽게하기 위한 기능키(Function keys) 설치, 메뉴시스템, 탐색어의 동의어나 변형철자(복수/단수, 영/미철자 등)에 대한 자동탐색, 단어출현빈도수에 의한 적합성순위부여(Ranking)나 문헌클러스터링(clustering)등의 자동언어처리를 제안하고 있다.

주가(Zuga)는 본문 DB를 만들때 본문 DB 탐색의 "필터"역할을 할 수 있도록 초록문이나 통제어탐색도 제공하며, 문장제한연산자나 문단제한연산자 외에 더 많은 용어위치제한(proximity)연산자를 제공하라고 DB 밴더에게 권고 한다.[57] 또한 읽기 쉽고 보기 좋은 스크린디스플레이기능도 강조한다. 디스플레이에 대해서는 제닝스(Jennings) 역시 언급하고 있는데, 특히 글자형과 크기도 달리하여 눈에 쉽게 뛸 수 있도록 하라고 권한다.[58]

Databases," in *Proceedings of the 5th International Online Information Meeting, 1981* (Oxford, England : Learned Information, Ltd., 1981), pp. 447~453.

Ernest Perez, "Text Enhancement : Controlled Vocabulary vs. Free Text," *Special Libraries* 73(3) (July 1982) : 183~192.

James A. Sprowl, "WESTLAW vs. LEXIS : Computer Assisted Legal Research Comes of Age," *Program* 15(3) (July 1981) : 132~141.

54) Perez, *op. cit.*, p. 191.

55) Dwight R. Tousignaut, "Indexing : Old Methods, New Concepts," *The Indexer* 15(4) (1987) : 197~204.

56) Lydia Jackson, "Searching Full-Text Databases," in *Proceedings of the 7th International Online Information Meeting, 1983* (Oxford, England : Learned Information, Ltd., 1983) : 419~425.

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

58) E. Judson Jennings, "Sam, you made the window too small," in *Proceedings of the 8th National Online Meeting, New York, May 1987* (Medford, NJ : Learned Infor-

전문가들이 경험에 의해 본문 DB검색의 효율향상을 위한 여러 방안들을 제안한 외에도 체계적이고 과학적인 실험연구가 이 분야에서도 수행되었다. ① Ranking 알고리즘을 사용하여 탐색시스팀을 수정 보완하는 것과, ② 불리언탐색에 메뉴시스팀을 통합시키는 것, ③ 컴퓨터 Architecture 와 레코드 구조를 바꾸는 것 등에 관한 연구가 그것이다.

3.4.1. 랭킹알고리즘의 응용

랭킹알고리즘이란 하나의 탐색질문으로 검색된 문헌들을 그 질문과 (각각의) 문헌사이의 유사도에 의해 순위를 매기는 연산방식이다. 이 알고리즘을 수행하는데는 ① 문헌에 나타난 용어의 가중치, ② 질문에서 사용된 용어의 가중치, ③ 질문과 문헌사이의 유사도측정공식이 필수적으로 필요하며, 이 세 요소를 각각 어떻게 산출하느냐에 따라 수 많은 랭킹알고리즘이 가능하다.

본문 DB 탐색에서 이러한 랭킹알고리즘을 이용하여 탐색효율을 향상시키려는 노력은 세이거(Sager)와 로크만(Lockemann)에 의해 수행되었다.[59] 19가지의 문헌용어가중치산출방법이 한 종류의 백터유사도법과 함께 사용됐을 때 각각의 재현율과 정확률을 측정하였다. 독일어로 된 법률 DB를 사용한 이 실험의 결과는 대체로 알고리즘을 사용했을 때가 알고리즘을 사용하지 않았을 때 보다 좋은 탐색결과를 가져왔으나 절대빈도에 의한 알고리즘에서는 오히려 사용하지 않은 것보다 못하였다.

노르웨이어로 된 법률 본문DB를 사용한 NORIS 프로젝트는 랭킹알고리즘의 효율성 뿐만 아니라 본문DB 탐색에 적합한 탐색전략에 대한 연구도 수행하였다.[60] 세이거와 로크만의 연구에서 사용된 것과 같은 질문가중치산출방법과 백터유사도공식을 사용한 조건 아래에서 두 가지의 문헌용어가중치 산출방법 즉 절대빈도방법과 binary 방법의 효율성을 조사하였다. 두 방법 모두 알고리즘을 사용하지 않은 것보다 좋았으며 특히 binary 방법이 절대빈도방법보다 더 좋은 알고리즘으로 나타났다.

그러나 백터유사도공식이란 Boolean 탐색에서 탐색어들이 AND로 연결

mation Inc., 1987), pp. 197~203.

59) Sager and Lockemann, "Classification."

되든 OR로 연결되든 같은 유사도값을 주기때문에, Boolean 탐색을 사용하는 실제 본문 DB 탐색시스팀에서는 백터유사도공식이 랭킹알고리즘으로 부적당하다는 판단 아래, 노정순은 유사도공식대신 fuzzy-set 이론을 사용한 랭킹알고리즘의 효율성을 조사하였다.[61] 영문잡지기사 DB를 대상으로 29가지의 문헌용어가중치방법을, 한 조건의 질문가중치방법과 함께 사용했을 때, 그 29가지 랭킹알고리즘의 효과를 비교하였다. 모든 알고리즘은 알고리즘을 사용하지 않은 본문탐색에 비해, 재현율에 대한 아무런 손상없이, 정확률을 향상시킬 수 있음이 발견됐다. 또한 29개의 알고리즘 중에서 어떤 탐색목적에서나 탁월한 알고리즘은 존재하지 않으며, 재현율 수준과 탐색전략에 따라 그 상대적 성능은 다르다고 밝힌다.

용어의 출현빈도수에 의한 랭킹알고리즘에 대한 연구 외에도 확률이나 언어학적 수단을 가미한 알고리즘에 대한 연구도 수행되었다. 번스타인(Bernstein)과 윌리암슨 (Williamson)은[62] 확률과 언어학적 수단을 사용하여, 본문을 이루는 각 문단과 질문의 유사도를 측정하는 랭킹알고리즘을 개발하였다. 테스트한 결과, 전체 질문의 85~95%의 질문에서 1순위의 문단만 가지고도 그 해답정보를 얻을 수 있음을 발견하였다. 이 알고리즘에서 문단의 유사도는 질문에서의 빈도수, 문단에서의 빈도수, 전체 DB內의 빈도수, 문단의 길이, 문단에서의 용어의 위치에 의해 계산된다.

3.4.2. Boolean과 메뉴시스팀의 통합

최근 온라인탐색에 대한 End-User의 관심이 점점 높아지자 대부분의 서지 온라인 상업시스팀은 두 종류의 이용자 즉 최종이용자(End-User)와 탐색전문가(Intermediate)가 메뉴시스팀과 명령어시스팀을 선택적으로 사용할 수

60) Jon Bing, "Text Retrieval in Norway," *Program* 15(3) (July 1981) : 150~162.
Jon Bing and Knut Selmer, *A Decade of Computer and Law* (Oslo, Norway : Norwegian University Press, 1980).

61) Jung Soon Ro, "An Evaluation of the Applicability of Ranking Algorithms to Improving the Effectiveness of Full Text Retrieval" (Ph.D. dissertation, Indiana University, 1985).

62) L.M. Bernstein and R.E. Williamson, "Testing of a Natural Language Retrieval System for a Full Text Knowledge Base," *Journal of the American Society for Information Science* 35(4) (1984) : 235~247.

있도록 탐색시스템을 제공한다. 최종이용자에게 사용하기 쉽고 편리한 메뉴시스템을 위해 본문DB분야에서도 Business Research Corporation과 같은 DB생산자는 메뉴식이든 명령어식이든 벤더에 따라 가공처리되기 편리하도록 두 형태로 DB를 생산한다. Harvard Business Review도 두 형태로 출판되는 본문DB의 하나이다.

겔러(Geller)와 레스크(Lesk)는[63] 도서관온라인목록시스템과 뉴스통신본문시스템에서 메뉴와 명령어방식을 비교하였는데, 이용자들은 목록에서는 명령어식을 뉴스통신본문에서는 메뉴식을 선호한다는 것을 발견했다. 두 시스템에 대한 선호도는 DB에 대한 이용자의 지식의 정도에 따라 달랐다.

보크만(Bochmann)이 비데오텍스에서 메뉴시스템과 명령어시스템을 병합하여 사용한 것을[64] 시작으로 본문DB탐색에서도 두 시스템을 병합하는 것이 소개되었다. 와터(Watter)등은[65] 불리언연산자를 사용하여 명령형으로 사용하는 시스템에서 언제 어디에서든지 메뉴식으로 전환할 수 있고, 또한 메뉴식이나 명령어식으로 탐색하는 도중 어디서라도 문헌본문을 직접 접근할 수 있도록 하는 실험시스템을 소개하고 있다. 이 시스템은 법조문이나 규칙, 공업표준에 관한 문헌에서와 같이 체계적구조로 이루어진 문헌의 본문DB에서 효과적이라고 밝힌다. 이런 시스템에서는 탐색할 주제분야에 대해서는 잘 아나 DB자체의 구조에 대해서는 익숙치 못할 경우에 메뉴식으로 시작하고, DB자체의 내용에 대해 잘 알 경우에는 불리언탐색으로 시작하라고 권고하고 있다.

3.4.3. 컴퓨터·레코드 구조 수정

도치화일이 온라인정보검색에서 주로 사용되는 대표적인 화일구조임에도 불구하고, DB의 크기가 증가함에 따라 전통적인 컴퓨터에서 도치색인화

63) V.J. Geller and M.E. Lesk, "User Interfaces to Information Systems : Choice vs. Commands," in *Proceedings of the 6th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1983, pp. 130~135.

64) G.V. Bochmann, J. Gecsei, and E. Lin, "Keyword Access in Telidon-an experiment," in *Videotex 82: International Conference and Exhibition on Videotex, Viewdata and Teletext* (Northwood Hills, England : Online Conferences, 1982), pp. 345~357.

65) C.R. Watters et al, "Integration of Menu Retrieval and Boolean Retrieval from a Full-Text Database," *Online Review* 9(5) (1985) : 391~401.

일은 처리속도면에서 문제를 제기한다. 여기에 스탠필(Stanfill)과 카알(Kahle)[66], 샐튼(salton)과 버클리(Buckley)[67]등은 대형 병렬(Parallel) 컴퓨터의 일종인 Connection Machine (CM)을 이용한 자연어탐색시스템의 성능에 대해 조사하였다. CM은 개당 16384개의 Processor를 지닌 모듈 1~4개로 구성되며 각 Processor는 4096바이트의 메모리와 1-bit-wide 연산논리단위를 가지고 있다. 하나의 모듈은 하나의 마이크로제어장치(controller)에 의해 제어되며 이 마이크로제어장치는 호스트(host)컴퓨터에서 지시하는 macroinstruction을 실행한다. 따라서 4개의 모듈을 지닌 CM은 한개에 하나씩 그러므로 4개까지 독립적인 프로그램을 돌릴 수 있다. CM을 이용한 자연어검색시스템에서는 각각의 Processor에 1~3개의 문헌을 축적시켜 두고, 하나의 명령을 지시하면 그 명령은 각 모듈을 통해 모듈내의 모든 Processor에서 개개의 Processor에 담긴 서로 다른 데이타로 같은 작업을 동시에 수행하기 때문에 각 문헌에 있는 단어의 존재여부를 탐색하는 것을 매우 빠르게 수행한다.

스탠필(Stanfill)과 카알(Kahle)은 1개의 모듈(즉 16384 processor)로 된 CM을 가지고 31993개의 문헌(총량 18Mb)의 자연어탐색을 수행한 결과, 전통적인 씨리얼컴퓨터에서 도치화일을 이용했을 때 보다 신속하고 정확률·재현율이 더 좋은 검색효과를 얻었다. 25개의 탐색용어가 Boolean 연산자로 연결된 1개의 질문을 실행하는데 0.004초가 걸리고 20,000단어로 된 질문을 실행하는데는 0.295초가 걸렸다. 4개의 모듈(총 65536 processors)로 된 full-size CM으로는 128,000 문헌(총 71Mb)을 같은 속도로 처리할 수 있다는 계산이다.

샐튼(Salton)과 버클리(Buckley) 또한 CM을 사용한 자연어검색효과를 보통 씨리얼컴퓨터에서 4가지 백터유사도공식에 의한 알고리즘을 사용했을 때와 비교 평가하였다. CM은 씨리얼컴퓨터에서 가중치를 전혀 사용하지 않

66) Craig Stanfill and Brewster Kahle, "Parallel Free-Text Search on the Connection Machine System," *Communications of the ACM* 29(12) (1986) : 1229~1239.

67) Gerard Salton and Chris Buckley, "Parallel Text Search Methods," *Communications of the ACM* 31(2) (1988) : 202~215.

은 시스템보다는 좋은 검색효율을 주지만, 보통컴퓨터에서 4가지의 백터유사값에 의한 알고리즘을 사용했을 때 보다는 나쁜 검색결과를 주었다. 또한 이 연구는 보통 씨리얼컴퓨터에서와 마찬가지로 CM에서도 가중치를 사용하여 적합성 피드백(relevance feedback) 탐색을 수행할 때가 가중치를 사용치 않은 적합성 피드백방법보다 좋은 탐색결과를 가져온다고 밝힌다.

이상의 연구에서 본 바와 같이 본문검색에 관한 연구과제는 본문검색의 성능은 어떠하며, 어떻게 하면 본문검색효율을 높일 수 있을가 하는 두 가지로 집약된다. 본문검색의 성능이 어떠한가의 과제는 구체적으로 효율과 효과면에서 다른 검색방법과 비교하여 평가하고 있다. 최근의 연구경향은 본문검색의 효율을 향상시키는 쪽으로 관심이 집중되고 있다. 법률쪽을 제외하고는 본문 DB는 이제 막 온라인으로 널리 알려지기 시작했으므로 이 분야의 연구는 본문검색에 대한 어떤 결론을 내리기에는 부족하며, 향상된 탐색기법과 시스템을 개발하기 위해 보다 많은 연구가 기대된다.

4. 본문 DB검색에 관한 연구전망

본문검색을 수행하는데 필요한 요소들로는 DB와, 탐색시스팀, 사용기관, 이용자, 탐색자, 탐색질문 등을 들 수 있다. 이들 요소들은 무엇으로부터 영향을 받으며, 이들 요소들 사이에는 어떤 상관관계가 있으며, 특히 이들 요소들은 본문검색의 결과에 어떤 영향을 끼치는가? 이에 대한 대답을 규명하여 본문검색을 통제하려는 것이 본문검색에 대한 연구의 목적일 것이다.

본문검색에 사용되는 요소들은 각각 여러 특성을 지닌다. 즉 본문 DB는 크기, 언어, 주제, 자료형태에 따라 다양하며 ; 이용자나 탐색자는 학력, 경력, 주제배경이 각각 다를 것이며 ; 탐색질문의 목적이나 질문의 특수성도 다양하며 ; 탐색을 수행하는 기관의 탐색정책이나 탐색시스팀도 각기 다를 것이다. 이들 탐색요소의 서로 다른 특성은 본문탐색의 결과에 어떤 영향을 끼치며 보다 나은 탐색을 위해서 어떻게 통제되어야 하는가가 과학적연구의 과제이다.

4.1. 변수 "Database"에 관한 연구

본문 DB는 DB의 크기, 주제, 언어, 문체, 자료의 종류에 따라 다양하다. 이들 특성은 본문탐색에 어떤 영향을 주는가?

DB크기가 증가하면 색인언어의 포괄성 증대로 본문탐색은 다른 탐색에 비해 높은 재현율, 그러나 낮은 정확률을 가져온다고 가정된다. 그러나 블레어(Blair)와 마론(Maron)의 연구는[68] 비교적 큰 규모의 DB를 사용하여 정반대의 결과인 높은 정확률과 낮은 재현율을 보고하고 있다. 이러한 연구결과는 DB가 커질수록 검색된 문헌수가 많아질 것이라는 이용자들의 생각때문에 큰 DB를 사용하여 탐색할 때는 이용자가 읽을 수 있을 만큼의 문헌만을 검색하기 위해 여러 탐색개념들을 AND연산자를 이용하여 연결(interserting)함으로써 탐색문헌을 핵심문헌만으로 제한한 때문인 것으로 분석되었다. 그러나 DB가 커질수록 검색된 문헌수가 많아질 것이라는 이용자들의 가정은 샐튼(Salton)에 의해 반론되었다.[69] 또한 이러한 상반된 연구결과는 법원의 판결문이라는 DB의 주제나 형태적 특성때문이기도 한 것으로 분석되었다. DB의 크기가 검색결과에 어떤 영향을 주느냐의 연구과제는 서로 다른 주제분야의 대규모 DB를 가지고 앞으로 연구되어야 할 것이다.

DB의 주제는 검색효율에 뿐만 아니라 본문탐색을 향상시키는데 사용되는 랭킹알고리즘의 효과에도 영향을 끼칠 수 있다. 자연과학에서 보다는 인문사회과학에의 주제접근이 더 어렵고[70] 사회분야와 자연분야 사이에는

68) Blair & Maron, *op. cit.*

69) Salton, "Another Look at ..."

70) Pierce Butler, "The Research Worker's Approach to Books-The Humanist," in *The Acquisition and Cataloging of Books*, edited by William M. Randall (Chicago : University of Chicago Press, 1940).

D.W. Langridge, *Classification and Indexing in the Humanities* (London : Butterworth, 1976).

Walter S. Achtert, "Abstracting and Bibliographical Control in the Modern Languages and Literature," in *Access to the Literature of the Social Sciences and Humanities*, ed. by Robert A. Colby and Morris A. Gelfand (Flushing, N. Y. : Queens College Press, 1974).

D.J. Foskett, "Problems of Indexing and Classification in the Social Sciences," *International Social Science Journal* 23(2)(1971) : 244~255.

용어의 출현빈도분포도 다르다고 한다.[71] 즉 두 주제분야에서 용어는 다양성과 분포도가 다르기 때문에 본문검색의 효율향상을 위해 적용될 수 있는 알고리즘 또한 다를 것이다. 어떤 알고리즘이 각각의 주제에 가장 적합할 것인가?

DB를 이루는 문헌의 종류 또한 본문검색의 결과에 영향을 줄 수도 있다. 신문이나 전문학술잡지, 교양잡지, 백과사전, 단행본 등은 각기 길이가 다르고, 용어의 동질성과 난해도(명료도)도 다르고, 이용목적과 이용대상도 다르기 때문이다. 단어의 분포도 다를른지 모르겠다. 이처럼 각기 다른 특성은 각기 다른 탐색전략이나 랭킹알고리즘, 탐색시스팀, 디스플래이 형태를 요구할 것이다.

이처럼 본문 DB의 크기와 문헌의 길이, 용어의 동질성과 난해도, 용어의 분포도가 검색결과에 어떤 영향을 주는가 ; DB의 주제영역에 따라 가장 적합하게 적용할 수 있는 알고리즘은 무엇인가 ; 문헌의 길이나 용어의 동질성과 난해도, 용어의 분포도에 따라 가장 적합한 탐색전략이나 탐색시스팀, 디스플레이 형태는 무엇일 것인가에 대한 연구가 앞으로 계속되어야 할 연구과제가 될 것이다. 이들 연구는 주제가 다르고 크기가 다르고 문헌의 특성이 다른 두개 이상의 DB를 비교연구함으로서 가능하다.

4.2. 변수 "탐색시스팀"에 관한 연구

대부분의 상업 온라인시스팀에서 사용하고 있는 표준탐색시스팀의 특성은 불리언논리와, 용어위치제한연산자, 도치색인화일의 사용이다. 같은 DB로 탐색을 수행해도 탐색에 사용된 탐색시스팀에서 화일을 어떻게 구조하였고, 어떤 S/W를 사용하여 어떤 탐색편의을 제공하고, 문헌의 내용을 어떻게

Maurice B. Line, "Concluding Considerations," in *Access to the Literature of the Social Sciences and Humanities* (Flushing, N.Y. : Queens College Press, 1974).

Stephen E. Wiberley, "Subject Access in the Humanities and the Precision of the Humanist's Vocabulary," *Library Quarterly* 53(Oct. 1983) : 420~433.

71) Mary E. Rowbottom and Peter Willett, "The Effect of Subject Matter on the Automatic Indexing of Full Text," *Journal of the American Society for Information Science* 33 (May 1982) : 139~141.

화면에 디스플레이하느냐에 따라 탐색결과는 달라질 수 있다.

도치색인은 전통적인 씨리얼컴퓨터에서 검색을 위해 사용되는 표준화일구조이다. 그러나 최근의 컴퓨터기술은 데이타량이 증가함에 따라 씨리얼컴퓨터에 만족하지 못하고 새로운 컴퓨터개발에 박차를 가하고 있다. 이들 개발된 비전통적컴퓨터中 정보검색을 향상시킬 수 있는 하드웨어의 구조에 대한 조사도 발표되었고[72] 이 중 병렬컴퓨터의 일종인 CM을 사용하여 자연어탐색을 하였을 때의 높은 검색효과가 보고되어 있다.[73] 그러나 또 다른 연구는 같은 CM을 사용하여 향상된 검색효과는 Parallelism에서 기인한 것이 아니라 시행방법에서 기인한 것이라고 한다.[74] 이 두 실험에서는 DB의 주제도 다르고, 본문이 아닌 대행물(例 : 초록문)을 사용하여 실험되었으므로 본문를 사용하여 병렬컴퓨터가 본문검색을 향상시킬 수 있는지의 연구가 수행될 수 있다. 또한 다른 비전통적인 컴퓨터나 다른 화일구조가 본문검색에 이용될 수 있는지의 연구도 계속 될 것이다.

탐색시스템의 탐색성능 또한 탐색전략과 탐색결과에 영향을 준다. 불리언 연산자 AND와 문단제한연산자는 서로 다른 검색결과를 가져온다.[75] 용어위치제한연산자 중 문단제한과, 문장제한, 인접단어제한연산자 또한 각기 서로 다른 탐색결과를 가져온다.[76]

Parsing rule이 다를 경우 (용어제한연산자의 기능과 불용어처리가 달라지며 따라서) 검색결과도 달라진다. 본문검색에서는 용어위치제한탐색을 위해 어떤 용어가 어떤 문헌의, 몇 번째 문단의, 몇 번째 문장에서, 몇 번째로 나타나는 단어인가를 도치색인화일 만들 때 표시해 준다. 문장內에 나타나는

72) Lee A. Hollaar, "Unconventional Computer Architectures for Information Retrieval," *Annual Review of Information Science and Technology*, 14 (1979) : 129~151.
……, "The Utah Text Retrieval Project," *Information Technology: Research and Development* 2 (1983) : 155~168.

73) Stenfill & Kahle, *op. cit.*

74) Salton & Buckley, *op. cit.*

75) Jung Soon Ro, "An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. I. On the Effectiveness of Full-Text Retrieval," *Journal of The American Society for Information Science* 39(2) (1988) : 73~78.

76) Tenopir, "Search Strategies ..."

단어의 순서값을 주는데는 Parsing rule에 따라 불용어는 제외시키기도 하고(BRS), 불용어를 계산하기도 한다(DIALOG). 즉 "School of Library Science……"로 시작된 문장에서 Library는 BRS에서는 2번째로, DIALOG에서는 3번째로 나타나는 단어로 취급된다. 본문검색에서는 DIALOG에서 사용되는 방법이 더 좋은 검색결과를 가져온다고 밝히고 있다.[77] Parsing rule에 따라 문장제한연산자의 기능이 어떻게 달라지고 또 검색결과에도 어떤영향을 끼치는가에 대한 연구는 DIALOG에서는 문장끝을 표시하는 구둣점을 blank로 취급하기 때문에 DIALOG의 문장제한연산자(nN)은 BRS의 문장제한연산자(WITH) 보다 나쁜 탐색결과를 가져온다고 밝히고 있다. 이들 테스트된 연산자 이외에 여러 시스템에서 사용되고 있는 다른 연산자들은 탐색에 어떤 영향을 주는가 ; 본문 DB검색을 향상시키기 위해서는 블리언연산자, 문단제한, 문장제한, 인접단어제한연산자 외에도 다른 같은 용어위치제한연산자가 더 필요한가에 대한 연구가 필요하다.

시스팀에서 자동언어처리에 대한 편의시설을 제공하는 것 또한 본문검색을 향상시키기 위한 하나의 방법일 것이다. 자동언어처리로는 변형용어탐색을 위한 용어절단(truncation) ; 변형철자탐색을 위한 string 탐색 ; 동의어안내 혹은 동의어 자동탐색 ; 단어의 출현빈도에 의한 문헌의 가중치, 랭킹, 클러스터링 ; 자연어질문에 따라 탐색전략을 작성하여 탐색을 수행하는 인공지능기법 등이 있다.

이중 변형철자나 변형용어탐색을 위한 용어절단이나 string 탐색기능을 제외하고는 본문 DB의 자동언어처리는 아직은 실험단계에 있다. 문헌內에 나타나는 단일어를 대상으로 가중치를 계산하는 몇가지 방법만이, 백터유사도공식과 함께 혹은 블리언탐색상황에서, 본문탐색을 향상시키기 위한 알고리즘으로 테스트되었다. 기타 다른 가중치와 다른 랭킹알고리즘을 사용한다면, 통계적인 기법 외에 언어학적이나 의미론적인 알고리즘을 사용한다면, 혹은 병렬컴퓨터에서 랭킹알고리즘을 사용한다면, 본문검색에 어떤 영향을 끼칠 것인가?

77) Jung Soon Ro, *An Evaluation ...*

대부분의 랭킹알고리즘에서 용어의 출현빈도를 측정할 때, 특정 개념을 반복적으로 사용하는 대신으로 사용되는 조동사나 대명사와 같은 대체용어의 출현빈도는 계산하지 않고 있다. 대체용어가 가르키는 본 용어를 컴퓨터가 구별하는데 어렵기 때문이다. 그러나 최근 초록문에서 대체용어를 구별하여 그 대체용어가 가르키는 용어의 출현빈도에 대체용어의 빈도를 포함시키는 알고리즘을 테스트하였다.[78] 본문 DB에서 대체용어의 빈도를 본 용어의 빈도에 포함하여 랭킹알고리즘을 만드는 것이 검색결과에는 어떤 영향을 줄 것인가?

인공지능시스팀은 서지 DB검색분야에서 탐색훈련을 받지 않은 최종이용자가 직접 탐색을 할 수 있도록 도와주는 중개시스팀으로 기능한다. 지금까지 개발·실용화되어 있는 중개시스팀은 자연어로 된 질문을 이해하고, 탐색전략을 세워, 탐색문을 만들고, 데이타뱅크를 선택·연결하고, DB를 선택하여, 세워 둔 탐색문을 수행하여, 탐색결과를 기계가독형태로 출력(Downloading)하는 기능을 주로 수행한다. 최근 연구개발된 INSTRUCT와 같은 인공지능시스팀은 텍스트에서 불용어를 제외한 후, 단어의 어근을 절단하여(stemming) 같은 어원을 지닌 단어들을 합하여 단어의 출현빈도를 산출한 후 용어의 가중치를 부여하고, 탐색시 용어의 가중치에 기초된 랭킹알고리즘을 사용하여 탐색결과를 질문과의 적합도에 따라 순서대로 디스플레이 한 후 이용자의 적합성 피드백(Relevance Feedback)에 의해 탐색문을 다시 수정하는 기능을 수행한다.[79] 한 연구는 이처럼 랭킹알고리즘과 이용자 피드백에 의해 탐색을 수정할 수 있는 인공지능시스팀인 INSTRUCT의 성

78) Elizabeth Liddy et al, "A Study of Discourse Anaphora in Scientific Abstracts," *Journal of the American Society for Information Science* 38(4) (1987) : 255~261.

79) Jan G. Hendry, Peter Willett, and Frances E. Wood, "INSTRUCT : A Teaching Package for Experimental Methods in Information Retrieval. Part I. The Users' View," *Program* 29(3) (July 1986) : 245~263.

……, "INSTRUCT : A Teaching Package for Experimental Methods in Information Retrieval. Part II. Computational Aspects," *Program* 20(4) (Oct. 1986) : 382~393.

Alina Vickery, Helen Brooks, and Bruce Rovinson, "A Reference and Referral System using Expert System Techniques," *Journal of Documentation* 43(1) (Mar. 1987) : 1~23.

능을 중개시스템(PLEXUS)과 비교평가하였는데,[80] 중개시스템에서 세운 탐색전략을 가지고 INSTRUCT에서 탐색할 때 INSTRUCT가 가장 좋은 탐색결과를 가져온다고 밝히고 있다.

인공지능시스템이란 탐색자의 탐색행위와 탐색전략에 근거하여 만들어진 전문가시스템이다. 본문 DB의 종류에 따라 이용자대상이 다르고, DB의 탐색목적에 따라 다른 탐색전략이 요구되기 때문에, 다양한 이용대상자(탐색자)와 다양한 탐색목적에 대비하여 적절히 사용될 수 있는 본문 DB검색용 인공지능시스템 개발에 대한 연구가 필요하다.

서지 DB와는 달리 본문 DB검색에서는 본문의 내용을 스크린에 디스플레이할 수 있어야 하기 때문에 읽기 쉽게 화면을 만들고, 사용하기 편리하도록 디스플레이 절차를 만드는 것이 매우 중요하다. 탐색어가 출현하는 본문의 앞뒤 내용을 보여주기 위해 얼마나 많은 분량의 본문을 보여줄 것인가? Scanning 명령어는 텍스트필드에서만 작동하게 할 것인가? 다른 필드 탐색에서도 작동하게 할 것인가? Window스케닝 도중 본문의 특정부분으로 되돌아가 그 부분을 디스플레이 하기 위해서는 어떻게 디자인하는 것이 가장 편리하고 효율적인가? 그래픽이나 삽도같은 그림정보를 광디스크에 보관할 경우 온라인 Text와 광디스크 그림 사이를 어떻게 연결하여 왔다 갔다 할 것인가?

본문 DB에서 본문을 한 개의 필드에 모두 집어넣어 두는 것보다는 chapter(章)別로, section heading별로 나누어 각기 다른 필드에 저장한다면 검색에 어떤 영향을 끼칠 것인가? 이 경우 필요한 다른 연산자나 탐색 편의시설은 무엇일까? 간단한 통제어휘집을 사용하여 본문탐색의 결과를 보완하는 것이 좋다고 하는데 이런 어휘집은 모든 종류의 본문 DB에 항상 도움을 주는가? DB의 종류에 따라 사용된 문체와 언어적 특성이 다르기 때문에 이러한 통제어휘집을 사용하여 본문탐색을 보완할 수 있다는 아이디어는 특정 DB에만 국한되는 것은 아닐까? 문헌의 주제에 따라 이들 통제어휘집은

80) Stephen Wade et al, "A Comparision of Knowledge-Based and Statistically-Based Techniques for Reference Retrieval," *Online Review* 12(2) (1988) : 91~108.

어떻게 구조할 것인가? 이상의 질문들이 탐색시스팀에 대한 연구과제들일 것이다.

4.3. 변수 "탐색자"와 "탐색절차"에 관한 연구

서지 DB탐색에서 탐색자의 주제에 대한 전문성이나, 교육, 훈련기간, 경력, DB구조에 대한 지식의 정도, 개성, 비용에 대한 의식, 태도, 인식과 같은 특성이 탐색행태와 탐색결과에 어떤 영향을 미치는지에 대해서는 여러 가지로 보고되었다.[81] 본문탐색은 대부분 최종이용자를 대상으로 제공되고 있기 때문에 본문 DB는 이용자의 특성에 대해 더 많이 고려해야 한다. DB의 주제에 대해 전문지식이나 훈련시간이 본문탐색결과에 어떤 영향을 주는지에 대해서는 연구가 되었으나[82] 다른 특성에 대한 연구는 아직 없다. 최종이용자가 사용하기 편리한 user-friendly 시스팀을 만들기 위해서는 최종이용자는 왜, 어떻게 본문 DB를 사용하는지에 대한 연구도 필요하다. 또한 현행 본문시스팀의 제한점인 탐색비용과 탐색성능, 디스플레이문제 등을 극복한다면 잠재적인 이용자는 어떻게 본문 DB를 사용할 것인가?

본문 DB탐색자가 정보를 찾고 탐색하는 행태(Behavior)에 관한 연구 또한 user friendly 시스팀 디자인에 필수적이다. "사람들은 탐색시스팀에 대해 어떤 정신적모델을 갖고 있는가. 이러한 모델은 전자적정보시스팀을 경험한 후 어떻게 변하는가. 이용자가 전자시스팀에 적합한 정신적 모델을 만드는 것을 돕고 탐색전략을 잘 세우도록 돕기 위해서는 어떤 개념적 모델이 필요한가?"[83] 이는 마치오니니(Marchionini)가 제기한 연구 가능한 연구질문이다.

4.4. 변수 "탐색질문"에 관한 연구

서지 DB탐색에서와 마찬가지로 본문 DB탐색 역시 이용자에 의해 탐색

81) 이에 대한 포괄적인 Review는 노정순, "우리나라 온라인 탐색환경과 탐색자의 탐색행위에 관한 연구"「도서관」43(6) (1988, 11/12월호) : 33~61 참조.
82) Collen & Flagle, *op. cit.*
Marchionini, *op. cit.*
83) Marchionini, *op. cit.*, p. 65.

요청된 탐색질문의 특성에 따라 영향을 받을 수 있다. 본문 DB에서 본문필드를 탐색할지 아니면 다른 필드를 탐색할지의 여부는 문헌과 질문의 깊이와 특정성을 보고 결정할 수 있다고 하였다.[84]

탐색질문의 특정성 뿐만 아니라 사용목적 또한 검색결과에 영향을 준다. 핵심문헌 몇 개만 읽고 그 주제분야의 요점만 파악하고 싶을 때는 (간략탐색을 위해서는) 통제어필드탐색이 바람직하고, 학위논문이나 연구논문을 위해 모든 관련문헌이 다 필요할 경우에는(포괄탐색을 위해서는) 본문탐색이 바람직하다. 본문탐색에서 랭킹알고리즘을 적용할 때 탐색목적에 따라 랭킹알고리즘의 효율이 달라진다고 한다. 탐색목적에 가장 알맞는 랭킹알고리즘은 어떤것일까?

4.5. 변수 "탐색효율"에 관한 연구.

서지 DB에서 정보검색시스팀을 평가하는데 사용되는 척도는 여러 학자들에 의해 수많은 방법들이 제안되었다. 그 중에서도 가장 많이 사용되어 온 것은 정확률(Precision)과 재현율(Recall)일 것이다. 본문검색시스팀을 평가하는데도 정확률과 재현율은 가장 타당한 평가기준이 될 것인가?

DB의 크기가 커짐에 따라 점점 재현율을 측정하는데 어려움이 수반되었다. 재현율측정을 위해서는 검색되지 않은 문헌 중에서 관련된 문헌이 얼마정도나 되는지를 알아야 한다. 즉 DB내 모든 문헌에 대해 질문과의 적합성을 조사해야 하는데, DB의 크기가 수 만件 이상이 됨에 따라 모든 관련문헌수를 조사하는 것은 불가능해졌다. 그러므로 상대재현율(Relative Recall)이나 통계적인 기법을 대신 사용하고 있지만 이것은 개념부터 절대재현율과는 크게 다르다.

정확률 또한 본문 DB검색 평가기준으로 사용하는데 어려움이 보고되었다. 본문 DB는 서지 DB에 비해 몇 배나 많은 문헌을 검색하기 때문에 검색된 모든 문헌의 적합성여부를 판정하는 일 또한 무리이다. 수많은 탐색문헌을 판정하는데 드는 비용과 시간문제로, 한사람이 적합성판정을 혼자서

84) Wagers, *op. cit.*

수행한 연구가 적지않다. 적합성 판정은 판정자에 의해 영향을 받는다.

정확률과 재현율은 적합성판정에 근거한다. 적합성판정에 영향을 주는 요인들에 대해서는 사라세빅(Saracevic)이 정리를 하였지만[85] 본문 DB검색에서 적합성판정 Scale을 다르게 했을 때 정확률과 재현율이 어떻게 달라지는지는 블레어(Blair)에 의해 보고되었다.[86]

서지 DB에서 사용한 적합성판정이나, 정확률·재현율 대신으로 사용할 수 있는 본문 DB검색용 평가기준은 어떤 것일까? 앞으로 연구해야 할 가장 중요한 과제인지도 모르겠다.

5. 결 론

본문 DB에 관한 연구는 본문 DB탐색에 필요한 각종 요소들 사이의 관계를 이해하고, 예측하며, 나아가서는 보다 나은 검색을 위해 이들 요소들을 통제할 수 있도록 하기 위한 것이다.

지금까지의 본문 DB에 관한 연구는 본문검색의 효율성을 평가하는 연구에서 시작하여, 시스템을 향상시키려는 연구로 진행되어 왔다. 시스템 향상을 위해서는 화일구조와, 탐색/디스플레이 성능, 인공지능기술에 기초한 효과적인 이용자인터페이스, 자동언어 처리와 같은 탐색시스템에 관심이 집중되었다.

본문 DB는 앞으로 더욱 빠른 속도로 이용될 것이다. 본문 DB에 관한 체계적인 연구는 이제 시작되었다. 이 분야의 중요한 발전은 가까운 시일에 컴퓨터 기술발달과 함께 올 것이다. 본문 DB는 최종이용자가 관심을 가질 것으로 기대되므로, 이 분야에 대한 보다 많은 연구는 이용자가 온라인 본문 DB을 어떻게 받아들이며, 사용하며, 탐색하는가를 밝힐 것이다. 이들 연구결과는 본문탐색시스템의 디자이너나 DB생산자에 의해 가장 효과적이

85) T. Saracevic, "RELEVANCE : A Review of and a Framework for the Thinking on the Notion in Information Science," *Journal of the American Society for Information Science* 26 (Nov./*Dec.* 1975) : 321~343.

86) Blair, "Full Text Retrieval : ..."

고 효율적인 시스템을 만드는데 사용될 것이다. 본문탐색에 사용되는 요소간의 상호관계에 관한 연구는 본문검색의 이론과 실제에 전문적인 체계를 제공할 것이다.

(접수일자 '89. 10. 14)

Future and Directions for Research in Full Text Databases

Jung Soon Ro*

Abstract

A Full text retrieval system is a natural language document retrieval system in which the full text of all documents in a collection is stored on a computer so that every word in every sentence of every document can be located by the machine. This kind of IR System is recently becoming rapidly available online in the field of legal, newspaper, journal and reference book indexing. Increased research interest has been in this field.

In this paper, research on full text databases and retrieval systems are reviewed, directions for research in this field are speculated, questions in the field that need answering are considered, and variables affecting online full text retrieval and various role that variables play in a research study are described.

Two obvious research questions in full text retrieval have been how full text retrieval performs and how to improve the retrieval performance of full text databases. Research to improve the retrieval performance has been incorporated with ranking or weighting algorithms based on word occurrences, combined menu-driven and query-driven systems, and improvement of computer architectures and record structure for databases.

Recent increase in the number of full text databases with various sizes,

* Asssistant Professor, Hannam University.

forms and subject matters, and recent development in computer architecture artificial intelligence, and videodisc technology promise new direction of its research and scholarly growth. Studies on the interrelationship between every elements of the full text retrieval situation and the relationship between each elements and retrieval performance may give a professional view in theory and practice of full text retrieval.