국내 문자정보 데이터베이스의 색인에 관한 연구

Development of an Indexing Model for Korean Textual Databases

정영미(Young-Mee Chung)*

목 차

1. 서 론
2. 색인언어의 검색 성능에 관한 기존의 연구
3. 국내 텍스트 데이터베이스의 색인 현황 및 문제점
4. 텍스트 색인언어의 검색 성능에 관한 실험
5. 텍스트 데이터베이스의 색인 및 탐색 모형
6. 결 론

초 록

본 연구에서는 국내 텍스트 데이터베이스의 색인언어 및 색인기법에 관한 현황을 분석하고, 3개의 텍스트 데이터베이스를 대상으로 하여 자연언어 색인과 통제언어 색인의 검색 성능을 평가하는 실험을 수행하였다. 조사결과 국내 텍스트 데이터베이스의 대부분이 자연언어 색인 방식을 사용하고 있었으며, 검색 실험에서는 적절한 탐색전략을 사용하는 경우 자연언어가 통제언어보다 검색 성능이 우수한 것으로 평가되었다. 색인현황에 관한 조사와 검색 성능의 실험 결과에 근거하여 국내 텍스트 데이터베이스를 위한 효율적인 색인 모형을 제시하였다.

ABSTRACT

The indexing languages and techniques were surveyed for Korean textual databases, and retrieval effectivenesses of two indexing languages were evaluated in an online searching experiment. It was found that most of the Korean textual databases surveyed employ natural language indexing by either an automatic or a manual method, and that natural language indexing may outperform controlled language indexing if appropriate search strategies are employed.

이 논문은 1994년도 한국학술진흥재단의 공모과제 연구비에 의해 연구되었음.

* 연세대학교 문헌정보학과 교수

■ 논문접수일 : 1996년 4월 19일

1. 서 론

1.1 연구의 목적 및 배경

1990년대에 들어 정보검색은 온라인 데이터베이스 및 데이터뱅크의 증가, CD-ROM 데이터베이스의 급증, 저렴해진 통신료, 인터넷의 대중화 등으로 인해 더욱 보편화되고 있다. 국내에서도 컴퓨터 통신망의 확장과 개인용 컴퓨터의 대중화로 인해 데이터베이스의 이용이 급속히 확산되고 있음을 볼 수 있다. 그러나 정보검색 환경이 아무리 발전한다고 해도 이용자가 필요로 하는 정보를 담고 있는 질이 좋은 데이터베이스의 공급이 없이는 이용자의 정보요구를 만족시킬 수는 없다. 특히 텍스트 데이터베이스의 경우에는 텍스트의 내용을 얼마나 정확하게 색인어로 표현하느냐가 정보검색의 성패를 좌우하는 주요한 요인이 되고 있다.

전세계적인 데이터베이스 디렉토리인 Gale Directory of Databases 1995년 7월호에 의하면 현재 공중접근이 가능한 데이터베이스의 수는 모두 9,887개이며, 이 가운데 온라인 데이터베이스가 5,467개이고, 나머지는 CD-ROM이나 디스켓 등에 소장된 데이터베이스로 나타나 있다. 전체 데이터베이스의 수는 1995년 1월의 8,776개에 비하면 6개월 사이에 무려 1,111개가 증가한 것이다. 1995년 1월호에 실린 데이터베이스를 정보 유형별로 분석한 결과를 보면 문자정보 데이터베이스가 전체의 72%를 차지하고 있으며, 이를 다시 세분하면 서지정보가 26%, 전문정보가 49%, 디렉토리정보가 23%, 특허/상표정보가 1%, 사전정보가 1%로 나타나 있다(Williams, 1995). 특히 주제 색인어를 통해 정보의 내용을 적절히 표현함으로써만 성공적인 정보검색이 가능한 서지/초록 데이터베이스(1,827개) 및 전문 데이터베이스(3,462개)는 모두 5,289개로서 전체 데이터베이스의 60%에 달하고 있음을 알 수 있다.

1990년대에 들어 국내에서 제작되는 데이터베이스의 수도 급증하여 1993년 말 현재 700개에 이르고 있다(알기 쉬운 한국의 데이터베이스 편람, 1994). 이 편람에서는 데이터베이스들을 분야별로 분류해 놓고 있는데, 이들 가운데 주제 색인이 필요한 텍스트 데이터베이스는 주로 일반(신문기사), 인문사회과학, 과학기술 분야 및 경제/산업/무역과 비즈니스 분야의 일부 데이터베이스로서 그 수는 매우 적은 것으로 나타나 있다. 본 연구에서 분석대상으로 삼은 산업기술정보원의 〔데이터베이스 총람〕 1993년판에서는 국내 제작 데이터베이스 598개를 정보유형별로 구분하고 있다. 크게는 사실 데이터베이스와 참고 데이터베이스로 나누고 사실은 수치, 문장수치, 디렉토리, 전문 등으로, 참고는 서지, 리퍼럴 등으로 다시 세분하고 있다. 여기에서 주제 색인이 필요한 텍스트 데이터베이스는 초록을 포함한 서지 데이터베이스와 전문 데이터베이스인데 실제로 상당수의 단순한 문자정보 데이터베이스가 서지 데이터베이스로 잘못 분류되어 있음이 발견되었다.

앞으로 데이터베이스 수가 증가하고 국가의 정보화가 진전됨에 따라 질이 좋은 데이터베이스에 대한 요구와 사용하기 편리한 인터페이스 시스템에 대한 요구가 급증할 것은 자명

한 사실이다. 현재 국내 제작 데이터베이스 가운데 텍스트 데이터베이스의 수가 상대적으로 적은 것은 텍스트 데이터베이스에 대한 인식 부족과 가공기술의 부족 때문인 것으로 보인다. 그러나 전세계적으로 전문 데이터베이스가 급증하여 이미 전체 데이터베이스 수의 40%에 육박하고 있는 사실을 감안할 때 국내에서도 전문 데이터베이스의 비중이 점차 커질 것으로 전망되며, 따라서 일반적인 상용 DBMS로는 처리하기 힘든 텍스트 정보의 효과적인 색인에 대한 요구가 커질 것으로 예상된다.

따라서 본 연구의 목적은, 데이터베이스의 양적인 성장도 중요하지만 이보다 더 중요한 것은 고품질의 데이터베이스와 검색시스템을 제공하는 것이라는 전제하에, 국내 텍스트 데이터베이스의 가공실태를 분석한 후 텍스트 데이터베이스를 위한 효과적인 색인 및 탐색 모형을 제시함으로써, 국내 정보검색을 활성화시키며 데이터베이스 산업을 발전시키는 데 있다.

1.2 연구의 내용 및 방법

본 연구는 크게 두 부분으로 구성된다. 첫 번째 부분은 국내 데이터베이스의 색인언어 및 색인기법에 대한 현황 분석으로서, 연구 시작 당시 발간되어 있던 〔데이터베이스 총람〕 1993년판을 기초 자료로 사용하였다.

먼저 〔데이터베이스 총람〕에 수록된 데이터베이스들 가운데 서지 및 전문 데이터베이스로 분류된 데이터베이스를 일차적으로 가려낸 결과 대략 300여 개에 달하였다. 그러나 데이터베이스의 내용을 검토한 결과 단순한 문자정보 데이터베이스가 서지정보로 잘못 분류된 경우가 상당수에 달하였고, 서지/전문 데이터베이스의 경우에도 주제 색인이 되어 있지 않은 데이터베이스들이 많이 있었다. 또한 특정한 도서관이나 정보자료실 소장도서의 목록정보만을 수록한 서지 데이터베이스나 신간도서의 서지 데이터베이스는 국립중앙도서관 데이터베이스 외에는 표제로부터 키워드가 색인어로 추출된 경우라도 제외하였다. 그리고 매일경제신문 DB와 같이 주제 색인은 되어 있는 경우라도 디스크립터가 아니라 소수의 분류코드(또는 주제 항목)에 의해 색인되고 탐색시 분류코드 메뉴에 의해 접근하는 데이터베이스도 제외되었다. 이 외에 아직 완성되지않은 데이터베이스(예 : 한국학술진흥재단 DB), 외부에 공개하지 않는 데이터베이스(예 : 삼성종합기술원 DB), 뱃치로만 제공되는 데이터베이스(예 : 한국화학연구소 DB), 직접 방문하기가 어려운 기관의 데이터베이스(예 : 한국해양연구소 DB)는 대상에서 제외하였는데, 이 결과 실제로 분석대상이 될만한 텍스트(서지/초록 및 전문) 데이터베이스는 수십 개에 불과하였다.

다음 단계에서는 선정된 데이터베이스를 제작하는 기관들을 직접 방문하여 색인담당자와의 면담을 통해 색인언어와 색인기법에 대한 정보를 수집하였다. 이 과정에서 〔데이터베이스 총람〕에 수록되지 않은 데이터베이스들이 분석대상에 추가되었으며, 연구 목적에 부적합한 데이터베이스는 제외되었다. 데이터베이스의 색인 작업이 데이터베이스 유통기관에 의해 수행되는 경우에는 유통기관을 방문하거

나 유통기관의 정보검색시스템을 직접 탐색하여 정보를 수집하였다.

본 연구의 두번째 부분은 색인언어에 대한 검색성능 평가 실험으로 구성되어 있다. 원래의 의도는 국내 텍스트 데이터베이스를 대상으로 자연어 색인어(자유키워드)와 통제어 색인어(디스크립터)를 각각 탐색어로 사용하여 검색한 결과를 분석하는 것이었는데, 앞 단계에서의 색인언어 분석 결과 자연어와 통제어가 동시에 색인어로 부여되어 있으며, 동시에 실험 목적에 부합되는 데이터베이스를 찾을 수가 없었다. 따라서 국내에서 가장 널리 이용되고 있는 DIALOG 시스템의 데이터베이스(ERIC과 The Washington Post)와 DIALOG 데이터베이스를 대체할 만한 CD-ROM 데이터베이스(INIS)를 검색실험 대상 데이터베이스로 선정하였다. 검색실험 평가는 단순히 두 색인언어의 검색성능을 비교하는 것이 아니라 검색결과를 심층적으로 분석하는 것이기 때문에 탐색질문의 수는 모두 10개로 한정하였다.

마지막으로 위와 같이 국내 텍스트 데이터베이스의 색인현황에 대한 조사와 색인언어의 검색 성능에 대한 실험 결과를 토대로 하여, 자연언어 색인과 통제언어 색인이 갖고 있는 장점을 극대화하고 단점을 극복할 수 있는 색인 및 탐색 모형을 제시한다.

2. 색인언어의 검색 성능에 관한 기존의 연구

정보검색시스템의 성능은 일반적으로 경제성, 신속성, 검색효율의 세 측면에서 평가되고 있다. 이 가운데 경제성과 신속성은 정보기술의 발전과 비례하여 향상되고 있음에 비해 검색효율은 그렇지 못하다. 검색효율은 이용자가 요구하는 수준의 정보서비스를 제공하는 시스템의 능력을 측정하는 것으로서, 이것은 데이터베이스에 소장된 특정한 정보로 안내하는 색인의 질과 탐색자의 탐색전략에 의해 좌우된다. 특히 텍스트 데이터베이스의 경우 텍스트의 내용을 축약적으로 표현하는 색인어를 어떻게 선택하느냐에 따라 이용자가 원하는 정보가 검색되기도 하고 검색되지 않기도 한다. 즉, 텍스트 데이터베이스 탐색에서는 색인언어의 검색 성능이 시스템의 검색효율에 영향을 미치는 가장 주요한 요인이 되고 있다.

색인언어는 색인어 선택시 용어에 통제가 가해졌는지의 여부에 따라 자연언어와 통제언어로 구분한다. 텍스트 데이터베이스에서 텍스트에 나타난 형태 그대로가 색인어가 되는 경우를 자연언어 색인라고 하며, 같은 의미로 자유텍스트(free text) 색인 또는 자유텍스트 탐색이란 용어를 사용하기도 한다. 여기에서 자유텍스트는 표제, 초록, 또는 전문(full text)이 되는데, 색인의 대상이 전문 데이터베이스인 경우에는 전문 탐색이란 용어를 사용한다. 반면에 통제언어 색인은 주제명표나 시소러스와 같은 통제어휘를 사용하여 색인어 선택시 통제를 가하는 색인을 말하며, 온라인 정보검색에서는 주로 시소러스나 시소러스 형태의 주제명표가 통제어휘로 사용된다.

자연언어 색인과 통제언어 색인의 장단점에 대한 논의는 통제언어 색인인 분류색인의 문

제점을 해결하기 위해 표제어색인이 출현하였던 19세기부터 시작되었으며, 컴퓨터가 정보검색에 이용되기 시작한 1960년 이래 본격화되었다. Rowley는 자연언어와 통제언어 색인에 대한 논쟁을 네 시대로 나누어 각 시대의 특징을 기술하고 있는데 이를 요약하면 다음과 같다(Rowley, 1994).

제 1시대는 통제언어 시대로서 분류표나 주제명표목표와 같은 통제어휘가 단행본 색인에 사용되었다. 제 2시대는 컴퓨터가 정보검색에 도입된 1960년대부터 1970년대 전반기에 걸친 시기로서 자연언어 색인과 통제언어 색인의 검색 성능을 비교하는 주요한 실험적 연구들이 수행되었으며, 실험의 결과는 대체적으로 자연언어가 통제언어 만큼 좋은 검색결과를 가져온다는 것이었다. 또한 접근점의 수도 검색 성능에 영향을 미치는 중요한 요인임이 밝혀졌다. 제 3시대는 1970년대 중반부터 시작되었으며 실험실이 아닌 실제 정보서비스를 이용한 검색 실험이 수행되었다. 이 시기의 연구들은 대부분이 특정한 환경에서 수행된 사례연구로서 실험결과를 일반화하기에는 무리가 있지만 대체적으로 자연언어와 통제언어를 함께 사용하는 것이 가장 이상적임을 제시하고 있다. 또한 접근점의 수가 검색 성능에 영향을 미치는 요인임이 재확인되었다. 1980년대 후반에 시작된 제 4시대에는 정보서비스의 실제 이용자를 대상으로 하여 두 색인언어의 효용성을 비교하는 연구들이 수행되고 있다. OPAC이나 CD-ROM 시스템과 같이 이전과는 다른 환경, 이용자 편의 인터페이스, 온라인 탐색 전문가시스템 등은 새로운 관점에서 색인언어를 평가하도록 하는 요인이 되고 있다.

자연언어 색인과 통제언어 색인의 장단점은 Dubois(1987), Aitchison과 Gilchrist (1987), Lancaster(1991)에 의해 잘 비교되고 있다. 〈표 2〉는 Aitchison과 Gilchrist가 두 색인언어의 장단점을 비교한 것을 요약한 것이다.

그러나 지금까지 수행된 이러한 검색실험 결과의 타당성이나 일반성에 대한 의문도 많이 제기되었다. 특히 Lancaster는 검색 성능의 차이는 색인언어의 통제 여부보다는 레코드의 길이에 의해 발생한다고 지적하면서, 많은 연구자들이 자연언어 색인과 통제언어 색인을 비교하는 실험에서 레코드의 길이를 통제하는 데 실패했음을 지적하고 있다(Lancaster, 1991). 즉, 실험에 주로 사용된 서지데이터베이스에서 레코드의 길이는 초록의 길이를 의미하며, 자연언어 색인에서는 초록이 길수록 접근점이 많아지므로 초록의 길이가 긴 레코드를 대상으로 두 색인언어를 비교했을 때 접근점이 많은 자연언어 색인이 높은 재현율을 가져오리라는 것은 매우 타당한 가설이라는 것이다. 또한 실험대상 레코드가 전문 레코드인 경우, 전문은 상당한 잉여정보를 갖기 때문에 탐색자가 선택한 표현을 포함할 가능성이 크고 따라서 재현율이 높아지리라는 것도 쉽게 가정할 수 있다.

위의 가설을 입증한 1980년대의 연구 가운데 주목할 만한 것으로는 Tenopir(1984)와 Ro(1988)의 연구가 있다. Tenopir가 Harvard Business Review 온라인 데이터베이스를 대상으로 하여 전문(문단), 초록, 통제어휘를 탐색한 결과 재현율은 각각 73.9%,

〈표 2〉 자연언어와 통제언어의 비교(Aitchison & Gilchrist)

자연언어	통제언어
〈장점〉	〈단점〉
-높은 특정성은 정확률을 높임. 인명 등 특정한 용어 검색에 뛰어남.	-특정성의 부족.
-망라성은 재현율을 높임. 표제만을	-망라성의 부족. 자연언어 수준의 망라성은
-대상으로 한 경우에는 적용안됨.	엄청난 색인비용 요구.
-최신성. 새로운 용어의 즉각 사용.	-최신성이 떨어짐.
-저자의 용어 사용. 색인자에 의해 잘못 해석될 염려 없음.	-색인자에 의해 잘못 해석 가능.
-탐색자가 자연언어를 사용함.	-탐색자가 인공언어를 배워야 함.
-입력 비용이 낮음.	-입력 비용이 높음.
-데이터베이스간 자료의 교환 용이.	-용어의 불일치로 자료교환 어려움.
〈단점〉	〈장점〉
-탐색자에게 지적인 수고가 요구됨. 특히 많은 동의어를 갖는 용어의 경우 문제가 심각함.	-시소러스화일 활용시 탐색자의 부담을 덜어줌.
-구문적 문제. 부정확한 용어의 조합으로 인한 부적합 문헌의 검색.	-복합어나 다른 장치로 구문적 문제를 해결함.
-망라성으로 인한 정확률의 저하.	-지나친 망라성으로 인한 정확률의 손실이 없음.

19.3%, 28.0%였으며, 정확률은 18.0%, 35.6%, 34.0%로 나타났다. 같은 데이터베이스를 대상으로 한 Ro의 연구에서는 전문, 문단, 초록, 통제어휘의 경우 재현율은 각각 83.7%, 61.3%, 18.4%, 21.5%, 정확률은 각각 14.5%, 36.7%, 58.7%, 66.7%로 나타나 있다. 두 연구 모두 자연언어 탐색에서는 레코드의 길이가 길어질수록 재현율은 증가하고 정확률은 감소하였다. 초록을 대상으로 한 자연언어 탐색과 통제언어 탐색을 비교하면 두 연구 모두 재현율은 통제언어가 다소 높았으며 정확률은 Ro의 연구에서는 통제언어가, Tenopir의 연구에서는 자연언어가 다소 높은 것으로 나타나 있다. 두 실험이 공통적으로 제공하는 결론은 자연언어의 경우 레코드의 길이가 길어질수록 재현율이 크게 증가하며 반면에 정확률은 크게 감소한다는 것이다. 또한 Tenopir의 연구를 비롯하여 제 3시대의 많은 연구들이 제시한 중요한 결론은 자연언어 탐색이나 통제언어 탐색만으로는 이용자의 정보요구에 적합한 문헌을 다 찾아낼 수 없으며, 두 가지 탐색방법을 다 사용하는 경우 가

장 많은 적합문헌을 검색할 수 있다는 것이다. 이러한 사실은 DIALOG를 비롯하여 대부분의 주요한 온라인 정보서비스들이 자연언어 탐색과 통제언어 탐색 방법을 모두 제공하는 주된 이유가 되고 있다.

제 3시대의 연구들은 비록 실제 온라인 환경에서 수행되었지만 대부분이 특정한 분야의 데이터베이스를 대상으로 하여 특정한 시스템 하에서 실험이 이루어졌기 때문에 하나의 실험 결과를 모든 분야로 일반화하기는 어려운 점이 있으며, 또한 주제 분야의 특성에 따른 색인언어의 성능상의 차이를 발견하기 위해 각기 다른 실험결과들을 그대로 비교하는 것도 문제가 있다. 즉, ERIC 데이터베이스를 대상으로 한 연구(Markey 등, 1980), 법률 데이터베이스를 대상으로 한 연구(Blair and Maron, 1985), CANCERNET 데이터베이스를 대상으로 한 연구(Henzler, 1978), COMPENDEX와 ENVIRONLINE을 대상으로 한 연구(Calkins, 1980), 비즈니스 분야의 데이터베이스를 대상으로 한 연구(Tenopir, 1985 ; Ro, 1988) 등이 제시한 실험 결과는 각각 그 특정한 분야에 적용되는 것이지 다른 분야의 데이터베이스에까지 확장하여 적용하기는 어렵다는 것이다. 다시 말해 이러한 실험들은 각기 다른 환경에서 수행되었기 때문에 주제 분야의 언어적인 특성이라든가 기타 환경적인 특성이 검색 성능에 있어 어떠한 영향을 미치는지를 밝혀내지 못하고 있다.

최근에 수행된 주목할 만한 연구로는 Fidel(1991)의 연구가 있다. 이 연구에서는 자연언어 탐색과 통제언어 탐색을 병행하는 전문적인 탐색자들의 탐색어 선택과 관련된 행태를 조사함으로써 두 가지 탐색이 상호보완적이라는 사실을 입증하고 있으며, 어떻게 보완하는 것이 바람직한가를 제시하고 있다. 이 연구에서는 탐색어 선택에 영향을 미치는 요인들을 정보요구 관련 특성(높은 재현율/정확률), 데이터베이스 관련 특성(시소러스의 유무, 복수 데이터베이스 탐색의 필요성), 탐색자 관련 특성(개인적 성향)의 세 가지 범주로 나누어 조사하였는데, 데이터베이스 관련 특성이 가장 큰 영향을 미치는 것으로 나타났다. 이 연구가 제시하는 중요한 결론은 특정한 데이터베이스와 관련된 색인언어의 질이 탐색어 선택에 가장 큰 영향을 미친다는 것이다. 즉, 탐색자들은 시소러스와 통제언어 색인의 질이 만족스러울 때에는 통제언어를 탐색어로 사용하며, 복수 데이터베이스의 탐색이 필요한 정보요구에 대해서는 자연언어 탐색을 선호한다는 것이다. 따라서 쉽게 사용할 수 있는 고품질의 시소러스와 복수 데이터베이스 탐색시 도움이 되는 언어변환 장치의 개발이 매우 중요함을 지적하고 있다.

3. 국내 텍스트 데이터베이스의 색인 현황 및 문제점

〔데이터베이스 총람〕의 일차 분석 결과에 의해 직접 방문한 기관 수는 31개였으나, 이 가운데 연구 목적에 부적합한 데이터베이스를 제작하는 기관을 제외한 결과 모두 21개 기관이 제작하는 데이터베이스들과 데이터베이스 유통기관인 천리안에 의해 자동색인되는 데이

터베이스들이 분석 대상이 되었다. 즉, 주로 단행본이 아닌 정보자료에 대해 서지정보 이외에 초록이나 요약, 또는 전문을 수록한 텍스트 데이터베이스로서, 제작기관이나 유통기관에 의해 키워드가 색인어로 부여되어 있는 데이터베이스가 최종적으로 선택되었으며, 국립중앙도서관의 한국문헌정보 DB는 국가표준 DB이므로 예외적으로 포함하였다. 각 데이터베이스에 대해서는 제작기관명, 전송망, 검색소프트웨어, DB 정보유형, DB 내용, 검색방법(메뉴/키워드), 색인언어(자연어/통제어), 색인기법(자동색인/수작업색인), 색인/탐색 시 문제점 등을 조사하였다. 유통기관에 의해 추가로 색인되는 경우는 괄호 안에 유통기관 표시를 하였다.

〈표 1〉은 국내 텍스트 데이터베이스들의 색인 현황을 각 제작/유통기관별로 요약한 것으로서, DB명, 색인기법, 색인언어, 통제어휘(통제언어 색인인 경우)의 순서로 기록하였다.

국내 최대의 데이터뱅크인 천리안이 제공하는 수백 개의 데이터베이스 가운데 본 연구의 조사대상이 되는 텍스트 데이터베이스들은 천리안 최상위 메뉴에서 7, 8, 11, 12, 13,

〈표 1〉 텍스트 데이터베이스의 색인 현황

제작/유통기관	DB명	색인기법	색인언어(통제어휘)
국립중앙도서관	한국문헌정보	수작업	자연어
국민경제연구소 (천리안)	경제정책정보 자동	수작업 자연어	통제어(리스트)
국회도서관	국회회의록색인 국감회의록색인 정기간행물기사색인 국외간행물색인 학위논문목록	수작업 수작업 자동+후통제자연어 자동+후통제자연어 자동+후통제자연어	통제어(리스트) 통제어(리스트)
농촌진흥청	농업기술종합정보 (다수)	수작업 자동	통제어(리스트) 자연어
대한무역진흥공사	해외시장정보(다수)	자동 자연어	
생산기술연구원	KAILIS 정보자료	수작업	자연어
산업기술정보원	과학기술문헌정보	수작업	자연어
연구개발정보센터	과학기술문헌 DB군 (그룹1-그룹8)	수작업 자동	-자연어 -통제어 (과학기술용어시소러스) 자연어

중앙일보사	JOINS	기사자동+후통제	통제어 (JOINS 시소러스)
총무처전자계산소	법령정보 민원정보 ; 도서정보	자동 수작업	자연어 자연어
코스모정보통신	정보통신동향	수작업	자연어
한국경제신문사	한경기사	자동+후통제	자연어
한국건설기술연구원	기술문헌정보	수작업	자연어
한국무역협회	문자정보 DB군 (19개 DB)	수작업	자연어
한국방송공사	방송자료	수작업	통제어 (KINDS 시소러스, 사전류)
한국법률정보시스템	판례정보 법령정보	수작업 자동	자연어 자연어
한국법제연구원	대한민국법령	자동+후통제	자연어
한국보건사회연구원	인구/보건/사회 문헌정보	수작업	통제어 (인구보건사회용어집)
한국언론연구원	KINDS 신문기사정보	자동	자연어
한국전자통신연구소	주간기술동향 외 (6개 DB)	수작업 자동	자연어 자연어
한국통신	사내연구정보	수작업	자연어
천리안	다수(본문 참조)	자동	자연어

14, 18번에 속해 있다. 천리안이 제공하는 텍스트 데이터베이스는 크게 세 가지 유형으로 나뉜다. 첫째는 자체적으로 데이터베이스 검색시스템을 운영하는 기관이 자관의 데이터베이스와 시스템을 천리안을 통해서도 제공하는 경우로서 각 시스템의 고유한 색인어와 검색방식을 통해 정보를 검색한다. 즉 천리안은 통신망만을 제공하는 경우로서 한국언론연구원의 KINDS DB, 한국법제연구원의 대한민국법령 DB, 농촌진흥청의 농업기술종합정보 DB, 국회도서관의 데이터베이스 등이 여기에 속한다. 둘째는 데이터베이스와 시스템이 부분적으로 제공되는 경우로서 데이터베이스 검색방식은 천리안이 제공하는 메뉴와 키워드 방식이 되며, IP가 제공하는 색인어 이외에 천리안의 자동색인시스템에 의해 추출된 색인어가 추가된다. 여기에 해당되는 데이터베이스로는 불교방송의 불교자료실 DB, 대한무역진흥공사의 해외시장정보 DB, 코스모정보통신의 정보통신동향 DB가 있다. 셋째는 천리안을 통해 제공되는 대다수의 데이터베이스들이 해당되는 경우로서 IP들은 천리안이 정한

양식에 따라 데이터를 제공하며 색인과 검색 방식은 천리안에서 전담하는 경우이다. 중앙일보를 제외한 각종 일간지의 기사 DB와 공업진흥청의 표준화소식 등이 이에 해당된다.

천리안에서 키워드 탐색이 가능한 텍스트 데이터베이스에는 다음과 같은 것들이 있다. 7번의 뉴스/날씨/스포츠 분야에서는 조선일보, 동아일보, 한국일보 등 대다수의 종합일간지와 전문지/잡지의 기사 전문 DB들을 전문에 출현한 키워드에 의해 탐색할 수 있다. 8번의 교육/문헌/취업 분야에서는 문헌정보 항목 아래에서 국립중앙도서관, 광주중앙도서관, 국내잡지정보, 정보통신동향, 불교자료실, 경영/경제문헌정보 DB들을 탐색할 수 있다. 11번의 증권/금융/보험/부동산 분야에서는 부동산뱅크, 대한주택공사 분양정보, 주식전략 핫라인 등의 데이터베이스를 키워드에 의해 탐색할 수 있다. 12번의 기업/무역/세무/물가/인물 분야에서는 KOTRA 해외시장정보 DB들과 노무정보, 세무정보 DB를 키워드로 탐색할 수 있다. 13번의 과학/기술/규격/법률/상표 분야에서는 특허정보 DB와 대한민국법령 DB, 표준화소식 등을 탐색할 수 있다. 14번의 여행/문화/가정/의학 분야에서는 여강출판사의 민간한방요법 관련 DB들과 의학전문정보 관련 DB들(의학논문, 간호학학위논문초록 등)을 탐색할 수 있다. 18번의 공공/지역/농수산/이산가족 분야에서는 농업기술종합정보 관련 DB들과 농림수산정보 DB를 탐색할 수 있다.

위의 분석결과를 보면 천리안을 포함하여 22개 기관 가운데 특정한 시소러스나 용어집을 기반으로 하여 통제어 색인을 수행하는 기관은 연구개발정보센터, 중앙일보사, 한국보건사회연구원의 세 곳에 불과하고, 국민경제교육연구소와 농촌진흥청은 주제명을 메뉴방식으로 제공하고 있다. 국민경제교육연구소의 경우 주제명들은 17개의 대주제 아래 모두 138개의 소주제로 세분되어 있으며, 농촌진흥청의 경우도 11개의 대주제가 각각 다시 십여 개씩의 소주제로 세분되어 있다. 농촌진흥청은 메뉴방식에 의해 검색한 레코드들을 키워드에 의해 다시 축소탐색을 할 수 있도록 하거나 또는 바로 키워드에 의한 탐색을 하도록 하고 있다. 국회도서관은 국회회의록색인과 국감회의록색인에 대해 사항명을 통제어 색인어로 부여하고 있다.

나머지 기관들은 모두 자동색인이건 수작업 색인이건 간에 자연어, 즉 자유키워드를 색인어로 주고 있다. 자동색인을 수행하는 기관은 천리안 이외에 국회도서관, 대한무역진흥공사, 연구개발정보센터, 중앙일보사, 총무처전자계산소, 한국경제신문사, 한국법률정보시스템, 한국법제연구원, 한국언론연구원, 한국전자통신연구소 등이다. 이 가운데 국회도서관, 중앙일보사, 한국경제신문사, 한국법제연구원은 자동색인을 한 다음에 후통제 작업을 통해 색인어를 조절하고 있다. 연구개발정보센터는 통제어 색인어 이외에 수작업에 의해 색인자가 임의로 선정한 자연어 색인어를 식별어 필드에 추가하고 있으며, BASIS 검색시스템의 자동색인 방식에 의해 제목, 초록, 주제명(통제어), 식별어(자연어) 필드로부터 키워드를 추출한다.

한국전자통신연구소는 자연어 색인의 단점을 보완하기 위해 동등어 처리를 하고 있으며

자동색인은 표제에 대해서만 적용하고 있다. 동등어 처리란 색인시 색인어로 선정된 키워드에 대해 동등어사전을 이용하여 동의어를 추가함으로써 탐색시 자동으로 탐색어 확장이 되도록 하는 것이다. 국회도서관은 표제를 대상으로 하여 조사처리프로그램에 의해 자동색인을 한 다음 수작업으로 필요한 색인어를 추가하는 등의 후통제를 하고 있다. 한국법제연구원은 불용어사전에 의해 불용어를 제거한 다음 사람이 판단하여 색인어를 추출하는 방법을 사용하고 있으며, 현재 작업 중인 시소러스가 완성되면 통제어 색인을 병행할 계획을 가지고 있다.

한국경제신문사와 중앙일보사는 신문기사 전문을 대상으로 자동색인을 하고 있다. 한국경제신문사는 불용어를 제거한 후 조사나 용언의 어미를 수록한 후절어사전을 이용하여 색인어 후보를 선택한 다음 사람이 최종적으로 색인어를 선택한다. 중앙일보사는 다른 자동색인시스템과 마찬가지로 형태소분석을 통해 불용어를 제거한 후 명사사전을 이용하여 명사를 추출한 다음, 약어, 유행어, 신조어 등 다양하게 표기되는 용어는 시소러스화일을 이용하여 비디스크립터를 디스크립터로 변환시킨다. 다음 단계에서는 가중치 부여에 의해 색인어를 선정하고, 마지막으로 수작업에 의해 불필요한 색인어를 삭제하거나 필요한 색인어를 추가한다(한상길, 1994). 언론연구원은 단어사전에 수록된 정보에 근거하여 최장일치 방식에 의해 어절을 분리하고, 색인어인지 불용어인지를 판단한 다음 조사와 어미 제거를 통해 키워드를 추출한다.

천리안의 자동색인 시스템도 다른 시스템과 마찬가지로 주로 형태소 분석에 의존하고 있으며, 불용어 제거, 조사나 어미 제거, 조사가 제거된 키워드 등록, 사전을 이용한 복합어 분리의 순서로 색인 작업이 수행된다. 천리안은 형태소분석 시스템 이외에 구문분석 시스템도 운영하고 있으나 처리시간 및 시스템 부하가 너무 크기 때문에 실제로 사용하고 있지 않는 것으로 보인다.

앞에서 분석한 바와 같이 현재 운영되고 있는 자동색인 시스템들은 거의 형태소 분석에 의존하고 있으며, 중앙일보사의 자동색인 방식이 약간의 구문분석을 병행하고 있는 것으로 보고되어 있다. 주로 형태소 분석에 의존하는 자동색인은 재현율을 높일 수는 있지만 상당수의 부적합문헌을 검색하게 된다. 다음은 실제로 천리안을 통한 키워드 탐색 결과 나타난 문제가 되는 사례들이다.

1) 생수 · 판매 : 두 탐색어가 문맥에 상관없이 한 기사에 들어 있기만 하면 되므로 전혀 관련 없는 히로뽕 관련 기사가 검색됨.(예문 : "히로뽕을 생수에 타서…/…히로뽕 판매…")

2) 생수 · 시판 : 시판은 판매와 동의어이므로 1)의 탐색문과 유사한 검색결과를 기대하였으나 그렇지 못함.

3) 신도시 + 주거환경 : 신도시가 포함된 용어열은 다 검색되므로 '신도시스템' 등 관련 없는 용어가 포함된 기사가 검색됨.

4) 도서관 : 기사 중에 도서관이란 단어가 나오면 무조건 검색되므로 관련 없는 기사가 다수 검색됨. 예를 들어 유명인사의 방한기사에서 "그가 …도서관을 방문할 것이다."라는 문장이 나오는 경우 등이 검색됨.

위의 사례들은 다음과 같은 자동색인의 기

본적인 문제점을 대표하고 있다. 즉, 첫째, 두 개 이상의 탐색어가 출현한 문맥을 고려하지 못한다는 점, 둘째, 동의어 처리가 안 되므로 관련 레코드가 다 검색되지 않는 점, 셋째, 용어열 탐색으로 인해 의미상 관련 없는 용어가 탐색어로 간주되는 점, 넷째, 탐색어가 텍스트의 주제어가 아닌 경우도 검색되는 점 등이다. 여기에서 첫번째 문제점은 탐색시 인접연산 기능의 사용을 통해 어느 정도 해결할 수 있으며, 두번째 문제는 한국전자통신연구소의 경우처럼 색인어에 동의어를 미리 추가하여 탐색시 자동확장되게 하거나, 탐색시 동의어를 탐색어로 추가함으로써 해결 가능하다. 탐색어 확장을 위해 용어절단 기능도 사용할 수 있다. 네번째는 용어의 출현위치에 따른 가중치 부여 방식이나 기사의 주제 분류에 의한 주제어 판별에 의해 어느 정도 해결 가능할 것으로 보인다. 그러나 세번째 문제의 원인이 되고 있는 용어열 탐색은 조사가 붙은 명사나 띄어쓰지 않은 복합명사의 탐색을 위해서 불가피한 탐색 방식으로서 이로 인해 발생하는 오류는 현재로는 한글이 안고 있는 근본적인 문제로 보인다.

현재 텍스트 데이터베이스 탐색을 위해 운용되고 있는 검색시스템들 가운데 자연어 탐색의 문제 해결에 도움이 될 기능들을 부분적으로나마 제공하고 있는 시스템들은 다음과 같다. 한국전자통신연구소의 ETLARS 시스템은 동등어 처리, 인접연산, 용어절단, 키워드 열람 기능을 갖고 있으며, 연구개발정보센터의 KRISTAL 시스템은 키워드 열람, 인접연산, 용어절단 기능을 제공한다. 한국무역협회의 KOTIS 시스템과 한국언론연구원의 KINDS 시스템도 키워드 열람, 용어절단 기능이 있다. 중아일보사의 JOINS 시스템은 내장된 시소러스를 이용한 용어 확인 기능과 하위어 포괄검색 기능을 제공하며, 동의어사전 화일을 이용하여 비디스크립터를 디스크립터로 자동변환하는 기능도 제공한다.

앞에서 분석한 데이터베이스의 대다수가 수작업에 의한 자연어 색인 방식을 채택하고 있다. 그러나 수작업에 의해 자연어를 색인어로 주는 경우에도 역시 문제가 있다. 특히 동일한 개념을 나타내는 다른 용어들이 별개의 색인어로 부여되거나 동일한 용어가 다르게 표현되는 등 색인어 선정의 일관성 결여 문제는 이미 지적된 바 있다(정영미, 1993).

실제로 각 기관을 방문하여 담당자를 면담한 결과, 수작업색인의 경우에는 색인자가 어느 정도 책임감을 가지고 색인 작업을 하고 있었으나 대부분이 통제어휘에 근거한 색인이 아니기 때문에 색인자의 성실성이 데이터베이스 색인의 질을 좌우하는 것으로 나타났다. 자동색인의 경우는 색인어 선정의 정확성 보다는 재현율을 높이는 색인어 선정이 우선되고 있었다. 특히 색인 대상이 되는 데이터베이스의 주제가 전문적인 분야가 아닌 경우 이런 경향이 더욱 두드러진다.

시소러스를 사용하여 통제어를 색인어로 부여하는 경우 자동색인의 문제나 수작업에 의한 자연언어 색인의 문제를 대부분 해결할 수 있다. 그러나 앞에서 분석한 대로 현재 세 기관에서만이 제대로 시소러스(또는 용어집)를 활용하고 있고, 산업기술원이 자연어 색인을 하되 과학기술시소러스와 전문용어사전을 참고하고 있다. 한국법제연구원은 현재 시소러

스를 개발중이며, 농촌진흥청이 시소러스 화일을 이용한다고는 하지만 분명치가 않다. 이외에 한국경제신문사와 한국언론연구원의 시소러스가 몇 기관에서 자연어나 통제어 색인어 선정에 참고되고 있는 것으로 나타나 있다.

4. 텍스트 색인언어의 검색 성능에 관한 실험

4.1 실험 가설

색인언어의 검색 성능에 관련된 기존 연구의 분석 결과 본 연구의 실험을 위해 주목한 세 가지 중요한 사실은 첫째, 자연언어 색인과 통제언어 색인이 상호보완적으로 사용됨이 바람직하다는 점, 둘째, 기존의 연구들은 동시에 여러 다른 분야의 데이터베이스를 실험 대상으로 삼지 않았기 때문에 실험결과는 특정한 주제 분야에만 국한될 수 있다는 점, 셋째, 탐색어 선택에 영향을 미치는 가장 중요한 요인은 색인언어와 관련된 것이라는 점이다.

본 연구에서는 주제 분야별 데이터베이스의 색인/탐색 모형을 설계하는 것이 원래의 목적이므로 인문/사회과학, 자연/응용과학, 일반 분야의 데이터베이스를 각각 대상으로 하여 자연언어 색인어와 통제언어 색인어를 탐색어로 한 검색 실험을 수행하였다. 이 실험에서 선택한 가설들은 다음과 같다.

가설 1. 자연언어 색인어와 통제언어 색인어는 탐색 대상 주제 분야와 상관없이 상호보완적인 검색 결과를 가져올 것이다.

가설 2. 인문/사회과학, 자연/응용과학, 일반의 세 분야는 각기 언어의 특정성이 다르기 때문에 다른 검색 결과를 보일 것이다.

가설 3. 전문 데이터베이스인 신문기사 데이터베이스는 초록만을 포함하는 데이터베이스와는 다른 검색 결과를 보일 것이다.

4.2 실험 환경

본 연구의 검색 실험에서 원래는 앞에서 분석한 국내 텍스트 데이터베이스들을 실험 대상으로 하려고 했으나 이미 밝혀진 대로 자유키워드와 디스크립터가 동시에 동등한 색인어로 부여된 데이터베이스를 찾기가 어려웠다. 연구개발센터의 KRISTAL 시스템이 자연어(수작업 및 자동 색인)와 통제어(수작업 색인)를 모두 색인어로 주고 있었으나 시스템이 가공하여 제공하는 8개의 데이터베이스들의 주제가 각각 정보산업, 신소재, 항공재료, 에너지, 전기, 원자력, 해양환경, 해사기술로서 모두 자연/응용과학 분야이기 때문에 실험 목적에 적합하지 않았고, 또한 데이터베이스의 크기가 소규모이며 아직 제대로 구축되어 있지 않은 상태여서 실험에 사용할 수가 없었다.

따라서 본 실험에서는 검색시스템으로는 DIALOG과 SilverPlatter CD-ROM 시스템을 사용하였고, 탐색 대상 데이터베이스로는 인문/사회 분야에서는 ERIC(Educational Resources Information Center), 자연/응용과학 분야에서는 INIS(International Nuclear Information System), 일반에서는 신문기사

데이터베이스인 The Washington Post를 선택하였다. 데이터베이스 선정을 위하여 우선 DIALOG에서 온라인 시소러스를 제공하는 데이터베이스를 탐색한 결과(?ss sf=online thesaurus) 모두 27개의 데이터베이스가 검색되었다. 처음에는 자연/응용과학 분야 데이터베이스로 Energy Science & Technology를 선택하였으나 탐색비용의 절감을 위해 이 데이터베이스의 하부 화일인 INIS CD-ROM판으로 대신하였다. INIS는 International Atomic Energy Agency가 제작하는 데이터베이스로서 온라인 데이터베이스는 STN Int'l에 의해 제공되고 있다. 신문기사 데이터베이스의 경우에는 온라인 시소러스가 제공되고 있는 것이 없으므로 우선 DIALOG의 PAPERS 데이터베이스로 들어가서 디스크립터가 부여되어 있는 화일들을 탐색하였는데, PAPERS에 포함되어 있는 54개의 신문 화일 가운데 48개가 디스크립터를 갖고 있었으며 이 가운데 The Washington Post를 실험 데이터베이스로 선택하였다.

4.3 실험 내용 및 결과의 분석

데이터베이스의 탐색은 탐색전문가에 의해 수행되었으며, 탐색 질문과 불논리 탐색문은 ERIC과 WP는 탐색전문가에 의해, INIS는 해당 주제 분야의 연구원들과 탐색전문가에 의해 작성되었다. 자연어 탐색문(탐색문-1) 작성에는 사전 등을 참고하였고, 통제어 탐색문(탐색문-2) 작성에는 해당 시소러스(ERIC 시소러스와 INIS 시소러스)를 참고하였다. 신문기사 데이터베이스의 경우에는 자연어에 의해 검색된 레코드에 부여된 디스크립터를 참고하였다. 탐색문에서는 검색되는 레코드의 수가 너무 많아지는 것을 막기 위하여 출판년도를 제한하였다. 질문의 수는 ERIC에 대해 3개, INIS에 대해 4개, WP에 대해 3개로 하여 모두 10개로 한정하였다. 질문의 수를 많이 하지 않은 이유는 자연언어와 통제언어의 검색효율(재현율과 정확률)을 측정한 실험은 이미 기존의 연구에서 충분히 수행되었고, 이 실험의 목적은 검색효율의 측정이 아니라 검색 결과의 심층적인 분석에 있기 때문이다.

검색된 레코드의 적합성 평가는 ERIC과 WP는 탐색전문가에 의해, INIS는 질문을 작성한 연구원들에 의해 수행되었다. 적합성의 정도는 다섯 수준으로 나누어 0%, 25%, 50%, 75%, 100%와 같이 백분율로 표시하였고, 이로부터 각 탐색문에 대해 검색된 레코드들의 적합성 평균을 산출하였다. 정확률과 재현율의 측정에 있어서는 적합성이 50% 이상인 레코드를 적합레코드로 간주하였다. 검색 성능에 대한 종합적인 척도로는 정확률과 상대재현율 이외에 각 색인언어에 의해 검색된 고유한 적합레코드 수를 사용하였다. 상대재현율은 각 질문에 대해 데이터베이스 내 전체 적합레코드의 수를 산출하기 어렵기 때문에 자연언어와 통제언어에 의해 검색된 레코드의 합을 전체 적합레코드 수로 간주하여 계산하였다.

ERIC 데이터베이스의 탐색

〈질문 1〉 학생들의 여가생활에 대한 교육 (py)=1992)

탐색문-1 : student? and leisure? and (education or teaching)

탐색문-2 : (students and leisure()time and (eduation or teaching)) / de

〈질문 2〉 청소년의 음주, 흡연 및 약물중독이 청소년 범죄에 미치는 영향(py〉=1990)

탐색문-1 : 1. adolescen? or juvenile? or youth or teenager or boy? or girl?
2. drinking or alcohol() abuse or smok? or cigarette? or tobacco? or drug()abuse
3. crime? or offence or guilt? or misdeed? or wrongdoing or misconduct? or delinquency
4. 1 and 2 and 3

탐색문-2 : 1. (adolescents or children) / de
2. (substance()abuse or drinking or alcohol()abuse or smoking or drug() abuse) / de
3. delinquency / de
4. 1 and 2 and 3

〈질문 3〉 TV 광고가 어린이의 의식과 태도에 미치는 영향

탐색문-1 : 1. (TV or television) and advertis?
2. child? or boy? or girl? or youngster? or juvenile?
3. awareness or consciousness or attitude or behavior
4. 1 and 2 and 3

탐색문-2 : 1. ((television and advertising) or television()commercials) / de
2. children / de
3. (awareness or consciousness or attitudes or behavior) / de
4. 1 and 2 and 3

〈질문 1〉의 통제언어 탐색(탐색문-2)에서는 자연언어 탐색에 비해 검색된 레코드의 수가 너무 적어서 시소러스의 용어관계를 이용한 재탐색을 실시하여 보았다. 자연언어 탐색에서 검색된 적합레코드들을 분석해 본 결과 이들 중 40%가 'leisure education'을 디스크립터로 가지고 있었기 때문에 'leisure time'의 관련어인 'leisure education'을 'leisure time'과 'education'의 두 디스크립터 대신 사용하였다. 그 결과 재탐색에서 검색된 레코드의 수는 모두 8개였으며, 두 탐색전략에서 공통되는 레코드는 한 개(부적합레코드)뿐이었다. 또한 추가로 검색된 8개의 레코드 가운데 50% 이상의 적합성을 보인 레코드는 단 한 개에 지나지 않아서 관련어를 이용한 탐색의 효율성에 의문이 제기되었다. 또한 'leisure time and education' 대신 'leisure education'을 탐색어로 선택하였을 때 전혀 다른 검색 결과를 가져온 것은 통제언어 탐색에서 디스크립터 선정에 상당한 주의가 필요함을 시사하고 있다.

〈질문 2〉에 대한 자연언어 탐색에서는 탐색자가 질문을 탐색문으로 변환하는 과정에서 처음에 청소년 범죄에 해당하는 용어인 'de-

linquency' (ERIC 시소러스의 디스크립터)를 포함시키지 않았다. 그 결과 검색된 레코드의 수는 41개, 이 가운데 적합레코드가 10개로서 적합성 평균은 14%, 정확률은 24%에 불과하였다. 2차 탐색에서 'delinquency'를 탐색어로 추가한 결과 검색레코드의 수는 2배로 늘고 적합레코드의 수도 2배 이상 늘어난 것을 볼 수 있다. 같은 자연언어 탐색에서도 탐색 개념을 탐색어로 어떻게 변환시키는가가 검색 성능에 영향을 미침을 알 수 있다. 질문 2에 대한 탐색에서는 자연언어 탐색이 적합성이 높은 레코드를 검색하는 능력이 뛰어난 것으로 나타났다. 즉, 자연언어 탐색에서는 전체 43개의 적합레코드 가운데 100%의 적합성 판정을 받은 12개의 레코드를 모두 검색하였으나 통제언어 탐색에서는 3개만이 검색되었다.

〈질문 3〉에 대한 탐색에서는 자연언어 탐색이 통제언어 탐색에 비해 2배 정도의 적합레코드를 검색하였으며, 고유한 적합레코드도 전체 적합레코드의 반 이상을 차지하고 있음을 볼 수 있다. 상대재현율도 87%에 달하고 정확률도 다른 자연언어 탐색에 비해 매우 높은 편이었다. 자연언어 탐색에서는 검색되었으나 통제언어 탐색에서 검색되지 않은 레코드들을 분석해 본 결과 대부분의 레코드가 'behavior'나 'attitude', 또는 'children'을 디스크립터 필드가 아닌 초록 속에 포함하고 있었다. 자연언어 탐색에서는 초록이나 제목 속의 용어뿐만 아니라 디스크립터 필드에 있는 용어까지도 탐색어가 되므로 통제언어 탐색에 비해 접근점이 훨씬 많아지고, 결과적으로 검색되는 문헌이 많아지는 것으로 보인다.

〈표 3〉은 ERIC 데이터베이스에 대한 각 질문의 탐색 결과이며, 〈표 4〉는 세 질문에 대한 검색 성능을 평균낸 것이다.

INIS 데이터베이스의 탐색

10″

〈질문 4〉 질산용액에서 이산화우라늄 분말의 용해 (py〉=1992)

탐색문-1 : (nitric acid) and (uranium oxide*) and dissolution

〈표 3〉 ERIC 탐색 결과

	질문-1		질문-2		질문-3	
	자연	통제	자연	통제	자연	통제
검색레코드 수	86	11	88	43	80	25
적합레코드 수	20	3	29	19	41	20
고유 적합레코드 수	17	0	22	12	27	6
적합성 평균(%)	23	25	24	28	46	64
정확률(%)	23	27	33	44	51	80
상대재현율(%)	100	15	67	44	87	43

〈표 4〉 ERIC 탐색의 검색 성능 평균

	적합성평균(%)		정확률(%)		상대재현율(%)		고유적합레코드	
	자연	통제	자연	통제	자연	통제	자연	통제
질문 1	23	25	23	27	100	15	17	0
질문 2	24	28	33	44	67	44	22	12
질문 3	46	64	51	80	87	43	27	6
평균	31	39	36	50	85	34	22	6

탐색문-2 : ((nitric acid) and (uranium oxides) and dissolution) in DE

〈질문 5〉 원격조정 운반차 (py〉=1992)

탐색문-1 : 1. remot* and control*
2. car* or bus* or truck* or automobil* or vehicle*
3. 1 and 2

탐색문-2 : ((remote control) and vehicles) in DE

〈질문 6〉 고온 무기 흡착소재 개발 (py〉=1992)

탐색문-1 : 1. (high temperature) or (inorganic ion exchange*)
2. adsorbent*
3. 1 and 2

탐색문-2 : 1. temperature-range-0400-1000-k in DE
2. inorganic-ion-exchangers in DE
3. adsorbents in DE
4. (1 or 2) and 3

〈질문 7〉 처분에서 방사성 핵종의 흡착 : 분광학적 연구(py〉=1993)

탐색문-1 : 1. disposal
2. radionuclide*
3. sorption or spectroscopy or spectrometry
4. 1 and 2 and 3

탐색문-2 : ((waste disposal) and radiosotopes and (sorption or spectroscopy)) in DE

〈질문 4〉의 경우 질문을 구성하는 개념들이 질산용액이나 이산화우라늄과 같이 다르게 표현될 수 없는 매우 특정한 개념들이기 때문에 자연언어 탐색문이나 통제언어 탐색문이 큰 차이가 없음을 볼 수 있다. 특히 Silver Platter CD-ROM 시스템은 별도의 인접연산기호를 사용하지 않고 구(phrase) 탐색을 할 수 있으므로 디스크립터와 조합된 용어(구)의 표현이 똑같게 된 것이다. 탐색 결과 전체 검색 레코드의 수는 자연언어가 다소 많으며 적합 레코드의 수는 자연언어가 2개가 많았다. 통제언어에서 검색되지 않은 2개의 적합레코드를 분석해 본 결과 이 두 레코드에서는 탐색

어가 모두 초록필드에만 나타나 있었다. 이들 레코드는 세 탐색어가 모두 디스크립터로 주어졌다면 검색될 수 있었던 것으로서 통제언어 색인 작성에 있어서 망라성 부족이거나 수작업 색인에 있어서의 오류로 인해 발생한 문제로 보인다. 특히 질문 4의 경우 상대재현율은 100%, 정확률은 50%로서 통제언어에 비해 정확률이 그다지 낮지 않은 결과를 보인 것은 특정한 개념을 탐색어로 변환하는 데 있어서 인접연산 기능의 적절한 사용이 검색의 정확성을 향상시키는 주요한 도구가 될 수 있음을 입증하고 있다.

〈질문 5〉에 대한 탐색 결과는 매우 특이하다. 자연언어 탐색이 통제언어 탐색에 비해 거의 10배에 가까운 많은 수의 레코드를 검색하였으며, 통제언어 탐색에서 검색되지 않은 적합레코드 또한 많이 검색해 내었다. 먼저 자연언어 탐색에서 59개의 적합레코드가 더 검색된 원인을 알아보기 위해 이들 레코드를 조사한 결과 대부분의 레코드에서 '원격조정'과 '운반차'의 두 개념 중 하나가 초록필드에만 출현하였음을 알 수 있었다. 디스크립터인 'vehicle' 대신 이의 하위개념어인 'truck'이나 'bus'가 출현하였기 때문에 검색된 레코드는 극소수였다. 'remote'와 'control'은 초록에 인접하여 출현한 경우가 많았지만(remote controlled, remotely controlled 등) 서로 떨어져 출현한 경우도 많이 있었다. 따라서 'vehicle'이 초록에 출현한 경우 이를 디스크립터로 부여했더라면 검색되었을 레코드들이 여러 개 있었다. 또한 다른 질문에 비해 자연언어 탐색 결과 부적합레코드의 수가 엄청나게 많이 검색된 원인을 찾아 본 결과, 용어절단 표시가 된 'car'나 'bus'를 용어열의 일부로 포함하는 전혀 다른 의미의 많은 단어들이 탐색어로 간주되었기 때문이었다. 특히 'car'의 경우 carry와 이의 변형(carried, carrying)을 포함한 레코드가 가장 많았고, 이외에 care-, cardio-, carbon-, cartridge, card, Carlson(인명) 등 다양한 단어들이 발견되었다. 'bus'의 경우는 소수였지만 business, bus(컴퓨터 용어) 등이 있었다. 따라서 다른 용어의 일부가 될 수 있는 탐색어는 용어절단 기능을 사용할 때 절단되는 글자 수를 한정한다든가, 또는 절단표시 없이 단수형과 복수형을 모두 탐색어로 사용하는 것이 바람직함을 알 수 있다.

〈질문 6〉에 관한 탐색문에서 '고온'과 '무기'에 해당하는 용어를 논리연산기호 'and'가 아니고 'or'로 조합한 이유는 무기는 고온에서만 사용 가능하므로 서로 대체할 수 있는 개념으로 보았기 때문이다. 만일 'or' 대신 'and'를 사용할 경우 대부분의 적합레코드가 누락될 것으로 예상되었으며 실제 검색 결과를 분석한 결과 이 사실이 확인되었다. 〈질문 6〉의 탐색에서 검색된 적합레코드들은 대부분이 '고온'보다는 '무기'의 개념을 갖고 있었기 때문에 결과적으로 자연언어와 통제언어 탐색에서 동일한 탐색어(inorganic-ion-exchangers)를 사용한 것과 같은 셈이었는데도 불구하고 통제언어 탐색에서는 누락된 적합레코드가 많이 있었다. 여기에서도 자연언어 탐색에서만 검색된 적합레코드들은 주로 두 탐색 개념 가운데 하나만이(주로 inorganic-ion-exchangers) 디스크립터 필드에 나타나 있었다. 만일 초록이나 표제 필드에 출현한 '무기'나

'흡착소재'의 두 개념이 모두 디스크립터로 부여되었더라면 이 적합레코드들은 통제언어 탐색에서도 검색되었을 것이다.

〈질문 7〉의 경우 자연언어 탐색과 통제언어 탐색에서 검색된 레코드들이 모두 적합레코드로서, 자연언어 탐색의 정확률이 100%라는 이례적인 결과를 보여주고 있다. 특히 자연언어 탐색에서는 통제언어 탐색에서 검색하지 못한 고유한 적합레코드를 55개나 검색하였다. 이들을 분석해 본 결과 상당수의 레코드가 'radionuclide-migration'을 디스크립터로 포함하고 있었다. INIS 시소러스에는 'radionuclide'는 디스크립터인 'radiosotopes'의 동의어로 나와 있으며, 실제로 'radiosotopes'보다는 광의의 개념으로 사용된다. radionuclide-migration은 absorption, desorption, diffusion 등 방사성 물질의 이동 전반을 나타내는 개념들을 포함하므로 탐색 개념인 '방사성 물질의 흡착'을 정확히 표현하지는 않는다고 보고 탐색자는 이를 디스크립터로 선정하지 않았다. 그러나 검색된 레코드들의 분석에서는 이 디스크립터가 탐색어로 유용할 수 있음을 시사하고 있다.

〈표 5〉는 INIS 데이터베이스에 대한 각 질문의 탐색 결과이며, 〈표 6〉은 네 질문에 대한 검색 성능을 평균낸 것이다.

The Washington Post 데이터베이스의 탐색

〈질문 8〉 환경오염 관련 기사 (pm=9503)

탐색문-1 : 1. air or atmospher? or environment? or water
2. pollution? or contamination?
3. 1 and 2

탐색문-2 : ((air or environmental or water)()pollution) / de

〈질문 9〉 북한의 NPT 관련 기사 (py〉=1994)

탐색문-1 : 1. north()korea
2. (non()proriferation()treaty) or npt

탐색문-2 : 1. korea()north / de
2. (nuclear()weapons or "arms

〈표 5〉 INIS의 탐색 결과

	질문-4		질문-5		질문-6		질문-7	
	자연	통제	자연	통제	자연	통제	자연	통제
검색레코드 수	32	23	233	25	46	26	65	23
적합레코드 수	16	14	84	25	16	7	65	23
고유 적합레코드 수	2	0	59	0	10	1	55	13
적합성 평균(%)	39	49	30	85	33	34	87	84
정확률(%)	50	61	36	100	35	27	100	100
상대재현율(%)	100	88	100	30	94	41	83	29

〈표 6〉 INIS 탐색의 검색 성능 평균

	적합성평균(%)		정확률(%)		상대재현율(%)		고유적합레코드	
	자연	통제	자연	통제	자연	통제	자연	통제
질문 4	39	49	50	61	100	88	2	0
질문 5	30	85	36	100	100	30	59	0
질문 6	33	34	35	27	94	41	10	1
질문 7	87	84	100	100	83	29	55	13
평균	47	63	55	72	84	47	31.5	3.5

control and disarmament") / de
3. treatise / de
4. 1 and 2 and 3

〈질문 10〉 핵무기 감축 관련 최근기사 (pm〉 9504)

탐색문-1 : 1. nuclear()weapon?
2. reduction? or elimination? or disarmament?
3. 1 and 2

탐색문-2 : (nuclear()weapons and "arms control and disarmament") / de

〈질문 8〉의 경우 환경오염의 하위개념인 대기오염과 수질오염을 탐색문에 포함시킴으로써 보다 많은 적합레코드를 검색해 내었다. 특히 통제언어 탐색에서 검색된 14개의 적합레코드 가운데 6개가 대기오염이나 수질오염에 해당하는 디스크립터에 의해 검색되었다. 이것은 시소러스를 이용한 통제언어 탐색시 계층관계를 이용한 탐색어의 확장에 의해 검색효율이 달라질 수 있음을 보여주는 사례이다. 〈질문 9〉와 〈질문 10〉은 자연언어 탐색문에서 인접연산 기능을 적절히 사용함으로써 검색의 정확성을 향상시킬 수 있음을 보여주고 있다. 〈질문 9〉는 북한의 NPT라는 제한된 주제를 표현한 것으로서 특히 NPT라는 매우 특정한 개념을 포함하고 있다. 이 경우 예상한 대로 자연언어 탐색이 훨씬 높은 재현율(91%)을 보이면서도 정확률(68%) 또한 매우 높은 결과를 보이고 있다.

〈표 7〉은 WP 데이터베이스에 대한 각 질문의 탐색 결과이며, 〈표 8〉은 세 질문에 대한 검색 성능을 평균낸 것이다. 〈표 9〉는 ERIC, INIS, WP 데이터베이스에 대한 검색성능의 평균을 비교한 것이다.

위의 실험결과를 종합적으로 평가하면 다음과 같다. 먼저 ERIC에 대한 탐색에서 정확률과 상대재현율 값은 질문에 따라 차이가 있으나, 평균적으로 정확률은 통제언어가 자연언어에 비해 14%가 높은 반면 상대재현율은 자연언어가 통제언어보다 51%가 높다. 검색해낸 적합레코드의 수는 자연언어가 통제언어의

〈표 7〉 WP 탐색 결과

	질문-8		질문-9		질문-10	
	자연	통제	자연	통제	자연	통제
검색레코드 수	57	19	100	19	52	21
적합레코드 수	20	14	68	19	15	11
고유 적합레코드 수	6	0	56	12	4	0
적합성 평균(%)	33	70	71	95	33	44
정확률(%)	35	74	68	100	29	52
상대재현율(%)	100	70	91	25	100	73

〈표 8〉 WP 탐색의 검색 성능 평균

	적합성평균(%)		정확률(%)		상대재현율(%)		고유적합레코드	
	자연	통제	자연	통제	자연	통제	자연	통제
질문 8	33	70	35	74	100	70	6	0
질문 9	71	95	68	100	91	25	56	12
질문 10	33	44	29	52	100	73	4	0
평균	46	70	44	75	97	56	22	4

〈표 9〉 ERIC, INIS, WP에 대한 검색 성능 비교

	ERIC		INIS		WP		평균	
	자연	통제	자연	통제	자연	통제	자연	통제
정확률(%)	36	50	55	72	44	75	45	66
상대재현율(%)	85	34	84	47	97	56	89	46
고유적합레코드	22	6	31.5	3.5	22	4	25	4.5

2.5배에 달하며, 고유한 적합레코드의 수도 거의 4배에 달하고 있다. 결과적으로 ERIC 탐색에서는 자연언어가 통제언어에 비해 정확률은 다소 떨어지지만 적합레코드를 검색하는 능력은 월등히 우수한 것으로 나타났다.

INIS에 대한 탐색 결과도 ERIC과 유사하다. 자연언어 탐색이 통제언어에 비해 정확률은 17%가 낮으나 상대재현율은 37%가 높다.

적합레코드 검색에 있어서도 자연언어가 통제언어에 비해 무려 9배에 달하는 고유한 레코드를 검색하였다. 특히 자연언어 탐색어의 부적절한 절단으로 인해 부적합레코드를 많이 검색한 질문 5를 제외했을 때에는 두 언어가 정확률에서는 거의 차이가 없으며 상대재현율에서는 40% 정도의 차이를 보이고 있다.

WP 탐색에서는 정확률은 통제언어가 31%, 상대재현율은 자연언어가 41% 높게 나타나서 정확률의 차이가 ERIC과 INIS에 비해 상대적으로 큼을 볼 수 있다. 그러나 자연언어 탐색은 상대재현율 평균이 97%로서 뛰어난 검색 능력을 보이고 있다.

3개의 데이터베이스에 대한 검색 성능을 종합한 결과 정확률은 통제언어가 66%로서 자연언어보다 21% 높게, 그리고 상대재현율은 자연언어가 89%로서 통제언어보다 43% 높게 나타나 있다. 또한 자연언어가 통제언어에 비해 5배에 달하는 고유한 적합레코드를 검색해냈으며, 통제언어에 의해 검색된 고유한 적합레코드의 수는 미미하였다. 따라서 데이터베이스에 따라 차이가 있지만, 탐색자가 대략 검색된 레코드 3개 중 2개, 또는 2개 중 1개의 부적합레코드를 감수할 수만 있다면 자연언어 탐색이 유리함을 보여주고 있다.

결론적으로 ERIC, INIS, WP에 대한 검색실험에서 질문 주제의 특정성과 탐색전략에 따라 각 질문에 대한 탐색 결과에는 다소의 변화가 있었지만 전체적으로는 세 데이터베이스가 유사한 양상의 검색 성능을 보이고 있다. 따라서 이 실험과 같이 소규모의 제한된 실험 결과를 가지고도 가설을 검증할 수 있었다. 만일 세 데이터베이스가 각기 다른 양상의 검색 성능을 보였더라면 이 실험의 가설들을 증명하기 위하여 더 많은 탐색 질문과 데이터베이스를 대상으로 한 실험이 요구되었을 것이다.

먼저 자연언어와 통제언어 탐색이 탐색 대상 주제와 상관없이 상호보완적일 것이라는 가설 1은 두 가지 사실에 의해 검증하였다. 첫째는 검색된 고유한 적합레코드의 수에 의한 것으로서 자연언어 탐색에 의한 고유 적합레코드의 수는 평균 25개인 반면 통제언어는 4.5개에 불과하였다. 또한 자연언어가 통제언어에 비해 훨씬 많은 수의 고유한 적합레코드를 검색한 것은 데이터베이스에 상관없이 나타난 결과였다. 따라서 고유한 적합레코드를 검색하는 능력에 있어서는 상호보완적이라고 보기는 어렵고 통제언어 탐색이 자연언어 탐색을 보완하는 위치에 있다고 봐야 할 것이다. 둘째는 정보요구의 특성에 따른 검색능력에 있어서의 상호보완성으로서 각 언어에 의해 높은 정확률과 높은 재현율을 원하는 탐색이 가능한가를 측정하는 것이다. 실험 결과 자연언어 탐색은 높은 상대재현율을, 통제언어 탐색은 높은 정확률을 가져왔으므로 재현율이 높은 탐색에는 자연언어를, 정확률이 높은 탐색에는 통제언어를 사용할 수 있으며, 이 점에서 두 언어는 상호보완적이라고 할 수 있다. 따라서 가설 1은 부분적으로는 옳다고 할 수 있다.

주제 분야에 따른 검색 성능의 차이(가설 2)에 관해 살펴보면 각 색인언어와 관련된 검색 성능의 양상은 유사하나 정확률과 상대재현율 값의 범위에 있어서는 데이터베이스 간에 차이가 있다. 정확률과 재현율을 합친 복

합척도를 검색효율의 척도로 사용했을 때 자연언어에서는 검색 성능의 차이가 그다지 크지 않았다. 그러나 통제언어에서는 WP, INIS, ERIC의 복합척도 값이 각각 131%, 119%, 84%로서 ERIC이 다른 데이터베이스에 비해 비교적 낮음을 볼 수 있다. 자연언어와 통제언어를 통틀어서 정확률에 있어서는 INIS가 가장 높은 값을, ERIC이 가장 낮은 값을 보인 것은 주제 분야 언어의 특정성을 어느 정도 반영한 것으로 보인다. WP에서도 탐색 질문들이 특정성이 높은 개념을 포함하였기 때문에 비교적 높은 정확률을 보인 것으로 판단된다. 특히 INIS의 경우 예상밖으로 자연언어의 정확률이 높게 나온 것은 전문 용어가 탐색어로 많이 사용되었기 때문인 것으로 보이며 이 사실은 주목할 만하다. χ^2 검증 결과 통제언어의 경우 자연언어와는 달리 세 데이터베이스가 유의한 차이(α=0.05)를 보였으므로 가설 2는 기각되지 않았다.

텍스트 유형에 따른 검색 성능의 차이가 있는지(가설 3)를 보기 위하여 초록을 포함한 ERIC 및 INIS의 두 데이터베이스와 전문을 포함한 WP의 검색 성능을 비교한 결과 자연언어와 통제언어에 있어 각기 유사한 양상을 보였다. χ^2 검증 결과 두 가지 텍스트 유형은 자연언어와 통제언어의 경우 모두 유의한 차이(α=0.05)를 보이지 않았으므로 가설 3은 기각되었다.

5. 텍스트 데이터베이스의 색인 및 탐색 모형

국내 텍스트 데이터베이스의 색인현황 분석과 색인언어의 검색 성능 실험을 통해 밝혀진 주요한 사실은 다음과 같다.

첫째, 국내 텍스트 데이터베이스에 대한 주제 색인은 수작업에 의한 자연언어 색인이 주류를 이루고 있다. 자동색인의 경우 주로 형태소 분석과 불용어 제거 방식을 사용하고 있으며, 색인 대상 텍스트는 제목, 초록이나 요약, 전문이 되고 있다.

둘째, 국내 텍스트 데이터베이스에서 자동색인에 의해 추출된 자연어를 탐색어로 사용했을 때 부적합한 레코드가 검색되는 사례가 많았으며, 이러한 사례들을 분석한 결과 부적합한 문맥으로 인한 오류, 동의어 처리의 미흡으로 인한 오류, 용어열 탐색으로 인한 오류, 레코드의 핵심 주제어가 아닌 색인어로 인한 오류 등의 네 경우가 파악되었다.

셋째, 검색 실험 결과 주제 분야에 관계없이 자연언어는 높은 재현율을, 통제언어는 높은 정확률을 보였다. 자연언어의 높은 재현율은 탐색어와 동의어가 되는 디스크립터 필드의 키워드까지 탐색 대상이 됨으로 인해 가능하였다.

넷째, 검색 실험 결과 자연언어와 통제언어는 상호보완적이라기 보다는 오히려 통제언어 색인이 자연언어 색인을 보완하는 결과를 보였다. 단, 자연언어 색인의 경우 적절한 탐색전략의 수립을 통해 검색 성능을 높일 수 있다.

다섯째, 통제언어 색인에서는 망라성 부족

이나 적절한 색인어 부여의 실패가 부적합한 레코드 검색의 한 원인이 되었다.

여섯째, 통제언어 색인에서 계층관계에 의한 탐색확장은 유용할 수 있으나 연관관계는 그다지 효과적이지 못하다.

일곱째, 다른 용어의 일부가 될 수 있는 색인어를 탐색어로 사용할 때에는 용어절단 기능의 사용에 주의가 필요하다.

본 연구에서 수행된 검색 실험 결과 통제언어가 자연언어에 비해 정확률은 높았으나 재현율과 고유 적합레코드의 비율이 상당히 낮았기 때문에 통제언어만을 이용한 탐색은 상당수의 적합레코드를 검색하지 못하는 치명적인 문제를 안고 있다. 반면에 자연언어의 낮은 정확률은 탐색자가 훑어보아야 할 레코드의 수를 가중시킬 뿐이기 때문에 레코드의 순위부여 알고리즘을 사용할 수 있다면 문제가 되지 않는다. 순위부여 장치가 없더라도 자연언어 탐색의 평균 정확률이 45%라는 사실은 탐색자가 검색된 2개의 레코드 가운데 1개의 부적합 레코드를 감수할 수만 있다면 큰 문제가 아닐 수 있음을 시사하고 있다.

결론적으로 자연언어 탐색이 보인 검색 성능상의 우수함뿐만 아니라 시소러스를 이용한 통제언어 색인의 비용을 고려할 때 자연언어 색인이 통제언어 색인보다 효율적이라고 할 수 있다. 그러나 자연언어 탐색은 탐색전략에 따라 그 결과가 크게 달라질 수 있으므로 탐색문 작성에 상당한 주의가 필요하다. 특히 동의어 처리가 검색 결과에 가장 큰 영향을 미치므로 색인작업시 동의어 사전을 이용하여 색인어를 추가하거나 또는 탐색시 탐색용 시소러스(동의어 사전 화일)를 활용한 탐색확장이 바람직하다.

따라서 본 연구에서 제안하는 텍스트 데이터베이스의 색인 및 탐색 모형은 다음과 같다.

(1) 우리말 시소러스 개발 비용과 색인전문가에 의한 색인 비용을 고려할 때 효과적인 자동색인시스템을 이용한 자연언어 색인이 바람직하다.

(2) 자동색인시스템이 불완전할 때는 온라인 방식의 후통제 작업을 도입하여 색인어를 조정하는 것이 필요하다.

(3) 미리 구축된 동의어 사전 화일을 이용하여 색인시 자동적으로 동의어를 색인어로 추가하거나, 또는 탐색시 탐색용 시소러스를 사용하여 자동으로 탐색어 확장이 되도록 한다.

(4) 자연언어 탐색시 부적합레코드를 훑어봐야 하는 탐색자의 부담을 덜기 위하여 적합성에 의한 레코드 순위부여 알고리즘을 도입한다.

(5) 불논리 검색에서는 검색된 레코드의 순위부여를 위해 탐색어의 출현빈도에 의한 통계적 기법, 핵심문헌과의 유사도 측정 기법, 용어간 어의적 인접성을 이용한 지식베이스 기반 기법 등을 활용할 수 있다.

(6) 검색의 정확성을 높이기 위하여 복합명사나 구의 탐색이 가능하도록 탐색어의 인접연산 기능을 도입한다.

(7) 통제언어 색인을 원할 경우에는 완전한 형태의 시소러스 대신 동의어 관계만을 표현한 시소러스를 개발하여 자동색인이 가능하도록 한다.

(8) 자동색인시스템은 현재의 단순한 형태소 분석에서 발전하여 부분적인 구문 분석 기능

을 도입하고, 더 나아가 지식베이스를 활용한 지식기반시스템으로 발전시킨다.

(9) 수작업 색인시스템에서는 색인전문가에 의해 색인하되 시소러스와 같은 통제어휘에 기반한 통제언어 색인이 바람직하다. 시소러스를 새로 개발해야 하는 경우 최소한의 수준은 동의어 관계만을 표현하는 것이고, 가능하면 탐색을 염두에 둔 계층관계를 포함하여 탐색확장이 가능하도록 한다.

6. 결 론

정보검색 환경이 발전하고 정보에 대한 일반의 수요가 증가함에 따라 데이터베이스에 대한 양적 및 질적 수요 또한 증가하고 있다. 특히 최근 들어 문장 형태의 정보를 담고 있는 텍스트 데이터베이스의 수는 급증하고 있으며, 이러한 데이터베이스의 효과적인 탐색을 위하여 텍스트의 내용을 대표할 수 있는 적절한 색인기법에 대한 요구가 점차 커지고 있다. 외국에서는 이미 지난 수십 년간 색인언어의 검색 성능에 관한 연구가 수행되어 왔으며, 대체로 자연언어 색인과 통제언어 색인은 검색 성능에 있어서 상호보완적이라는 결과가 나와 있다.

본 연구에서는 국내 텍스트 데이터베이스의 색인 실태를 먼저 파악한 후 기존의 검색 실험 연구들과는 다른 각도에서 색인언어의 검색 성능에 관한 실험을 수행하였다. 실험 결과 전반적으로 자연언어 색인이 통제언어 색인에 비해 검색 성능이 우수함을 발견하였고, 또한 성공적인 검색을 위하여 자연언어의 경우에는 탐색어의 적절한 선정이 무엇보다도 중요하며, 통제언어의 경우에는 색인어 선정의 적절성과 망라성이 중요함을 발견하였다. 따라서 국내 텍스트 데이터베이스들의 대부분이 자연언어 색인을 채택하고 있는 현실에 비추어 현재로는 자연언어 색인이 통제언어 색인보다 효율적이라는 결론에 도달하였다. 또한 자연언어 색인은 수작업에 의한 것보다는 자동색인 방법이 바람직하며, 시스템과의 대화방식의 후통제를 도입하는 것은 현재 수준의 자동색인시템에서는 색인의 정확성을 위해 필요할 것으로 보인다. 자연언어가 갖고 있는 단점은 적절한 탐색전략에 의해 극복될 수 있으며, 특히 탐색시 탐색용 시소러스나 동의어 사전 화일을 활용한 탐색어의 자동 확장은 자연언어의 검색 성능을 현저히 개선할 것으로 예상된다.

참 고 문 헌

데이터베이스 총람. 서울 : 산업기술정보원, 1993.

알기쉬운 한국의 데이터베이스 편람. 서울 : 한국데이터베이스진흥센터 / 한국데이터베이스산업진흥회, 1994.

정영미. 신문 시소러스 개발의 이론과 실제. 한국문헌정보학회지, 25 : 51-82, 1993.

한상길. 시소러스를 이용한 신문기사 데이타베이스 색인시스템에 관한 연구. 정보관리학회지, 11(1) : 125-144, 1994.

Aitchison, J. and Gilchrist, A. Thesaurus Construction. 2nd ed. London : Aslib, 1987.

Blair, D.C. and Maron, M.E. An evaluation of retrieval effectiveness for a full text document retrieval system. Comm. of the ACM, 28 : 289-299, 1985.

Calkins, M.L. Free text or controlled vocabulary : a case history step-by-step analysis ··· plus other aspects of search strategy, Database, 3(2) : 53-67, 1980.

Dubois, C.P.R. Free text versus controlled vocabulary : a reassessment. Online Review, 11(4) : 243-253, 1987.

Fidel, R. Searchers' selection of search keys : II. Controlled vocabulary or free-text searching. JASIS, 42(7) : 501-514, 1991.

Gale Directory of Databases, Volume 1 : Online databases and Volume 2 : CD-ROM, Diskette, Magnetic Tape, Handheld, and Batch Access Database Products. Detroit : Gale Research, Inc., July 1995.

Henzler, R.G. Free or controlled vocabularies. International Classification, 5 : 21-26, 1978.

Lancaster, F.W. Indexing and Abstracting : Theory and Practice. London : Library Association, 1991.

Markey, K. et al. An analysis of controlled vocabulary and free text search statements in online searches. Online Review, 4(3) : 225-236, 1980.

Ro, J.S. An evaluation of the applicability of ranking algorithms to improve the effectiveness of full-text retrieval. I. On the effectiveness of full-text retrieval. JASIS, 39(2) : 73-78, 1988.

Rowley, J. The controlled versus natural indexing languages debate revisited : a perspective on information retrieval practice and research. J. of Information Science, 20(2) : 108-119, 1994.

Tenopir, C. Full text database retrieval performance. Online Review, 9(2) : 149-164, 1985.

Williams, Martha E. The state of databases today : 1995. Foreword in Gale Directory of Databases. 2 vols. January 1995.