시소러스를 이용한 신문기사 데이타베이스 색인시스템에 관한 연구*

A Study of Indexing System Based on Thesaurus for Newspaper Database

한 상 길 (Sang-Gil Han)**

□ 목 차 □

초 록

신문기사 색인을 위한 시소러스에 대한 논의는 오래전부터 있어 왔다. 특히 CTS에 의한 신문제작 전산화 이후 대부분의 신문사가 신문기사 DB구축과 자동색인에 많은 관심을 기울이고 있으나 아직 국내에서는 이렇다할 성과가 없었다.

본 논문은 중앙일보사가 국내 최초로 구축한 JOINS시소러스에 대해서 살펴보고 시소러스를 이용한 신문기사 데이타베이스 온라인 자동색인 및 후통제 수작업색인의 효율성을 비교하고 바람직한 색인방안을 모색해 본다.

ABSTRACT

The Matter of vocabulary control for newspaper database has been studied for a long time. These efforts hadn't made any good achievements until JOINS Thesaurus system developed.

The purpose of this paper is to introduce JOINS Thesaurus which the Joong-ang Daily News has developed for the first time in Korea. In addition to that, this study is compares the efficiency of Auto-Indexing system with post-controlled indexing system for newspaper database on thesaurus.

* 이 논문은 정보통신분야 제조업 경쟁력강화를 위한 체신부 기술개발지원에 의하여 이루어 졌음.

** 중앙일보 조사부 기자

I. 서 론

과학기술의 발달에 따른 정보량의 증가로 필요한 정보를 신속 정확하게 검색하는 것이 정보를 수집하는 일보다 더욱 큰 과제로 등장하게 됐다.

정보를 검색하기 위해서는 색인의 과정을 거쳐야 하며 검색의 효율성을 높이기 위해서는 색인단계에서 적절한 용어통제가 필요하다. 시소러스란 일종의 통제어휘집으로 용어의 동등관계를 디스크립터라고 하는 대표용어로 규정하고 이들 사이의 계층관계와 관련관계를 체계화 해 놓은 것을 말한다.

정보검색 분야에 시소러스가 우리나라에 처음 알려지게 된 것은 약 20년전인 1970년대 초였다. 1972년 趙敏元씨가 최초로 시소러스에 관한 연구를 시작했으며, 司空哲씨에 의해 많은 연구가 이루어 졌다.

특히 1980년대 이후 대학도서관을 비롯한 대규모 도서관이나 기술정보실과 같은 정보처리기관 등에서 정보량이 늘어나고 컴퓨터가 도입되는 등 정보처리환경이 바뀜에 따라 시소러스의 필요성이 본격적으로 인식되기 시작했다. 그러나 시소러스 구축작업은 짧은 시간에 단순한 단어의 나열로만 그칠 수 없는 많은 시간, 노력, 비용이 수반되는 대규모 작업이므로, 시소러스의 필요성을 강조한 개별 주제에 관한 이론적 연구나 수작업 색인을 위한 참조용 시소러스가 실험적으로 이용되었을 뿐 진정한 의미의 시소러스는 지금까지 전무했다. 또한 정보처리 기술의 발달에 따른 자동색인 기법의 발달에 따라 일부에서는 시소러스의 無用論이 대두되기도 했다.

그러나 자연어 색인 역시 한글의 특성이나 대상문헌의 특수성 등으로 인해 정도율이 떨어지거나 검색잡음이 많은 등 검색효율이 떨어져 자연어나 통제어 한가지 유형을 고수하기보다는 자연어와 통제어를 병용하는 시스템으로 발전하고 있는 추세이다.

특히 주제의 폭이 넓고 빈번한 용어의 변경, 축약언어, 띄어쓰기의 미비 등많은 어의적 문제를 가지고 있어 자연어색인으로는 검색의 효율을 보장할 수 없는 신문기사색인시스템에 있어서는 시소러스에 의한 어휘통제가 필수적인 것으로 인식되어 왔다.

중앙일보사에서는 신문기사색인에서의 어의적 문제를 해결하고 JOINS 기사DB의 검색효율을 높이기 위해 1989년부터 시소러스 개발에 착수 약 6년여에 걸쳐 시소러스 사전구축을 완료했다. 또한 컴퓨터 신문제작시스템인 CTS시스템과와 기사DB를 직접 연계 온라인화하고, 시소러스를 통한 자동색인 시스템을 구현하였다.

이 색인시스템은 자연어 추출 이외에 빈도,위치분석 등과 함께 시소러스사전을 이용한 가중치 부여로 자동생성된 보다 정확한 색인어에 수작업색인을 첨가한 완벽한 색인시스템이다.

이 연구는 중앙일보 JOINS시소러스와 온라인 환경에서 실행되는 시소러스를 이용한 신문기사데이타베이스 자동색인에 대해서 살펴본다. 아울러 신문기사의 자동색인과 수작업 색인의 효율성을 비교하고 바람직한 기사데이타베이스 색인 방안에 대해서 고찰하고자 한다.

II. 이론적 배경

2.1 시소러스의 정의

시소러스란 事件表, 確率에 관한表라는 뜻을 가진 희랍어 「Onσαυρos」에서 유래한 말(Saracevic, 1970)로 시소러스가 지식의 寶庫, 辭典, 百科事典의 의미로 알려지기 시작한 것은 1736년 「Shorter Oxford English Dictionary」이다(Gilchrist, 1971).

시소러스라는 용어가 저술에서 처음으로 쓰인 것은 1852년 로제(P.M. Roget)가 작성한 「Thesaurus of English Word and Phrase」였는데 여기에서의 의미는 "어떤 개념에 대하여 그것을 가장 적절히 표현 할 수 있는 표목을 선택하기 위하여 만들어진 어구의 집대성"으로 나타났다. 이말을 바꾸어 말하면 시소러스란 용어를 알고 내용을 찾는 일반사전과는 반대로, 뜻을 알고 있으나 그 개념에 해당하는 적당한 용어를 찾는데 사용되는 용어집이라고 할 수 있다.

정보검색분야에 시소러스라는 용어가 처음 도입된 것은 1957년 5월 영국 도킹(Docking)에서 개최된 「제1회 정보검색용어분류법에 관한 국제회의」에서 였다. 이 회의에서 브론슨(H. Brownson, 1957)은 「관련된 의미네트웍에서 비롯된 시소러스」라는 주제강연을 통해 "한 문헌에 표현되는 여러 개념과 그것들의 관계를 보다 정비된 용어로 바꾸는 것이 정보검색의 과제이며, 그 해결방법으로, 문장을 구성하고 있는 구문들을 단문화하고, 동의어를 통제하여 의미를 상호관련시키는 방법을 적용하는 것이다"고 하였다. 한편 ISO(국제표준기구)에서 출판된 「Documenatation of Monolingual thesauri」에서는 시소러스를 기능적인 면과 구조적인 면으로 구분하여 "기능적인 면에서는 문헌, 색인작성자, 이용자의 자연어를 통제가 가해진 시스템언어로 변환시킬때 사용되는 용어통제표이며 구조적인 면에서는 지식의 어떤 특수한 영역을 포함한 일반적이며, 상위개념의 용어와 하위개념의 용어를 의미론으로 밝힌 동적인 어휘집"(F.I.D., 1970)이라고 정의하고 있다. 이외에도 룬(H.P.Luhn, 1957), 길럼(T.L.Gillum, 1964), 소에르겔(D.soergel, 1974)등의 정의가 있다.

이상의 여러 정의들을 종합 정의하면,시소러스란 문헌정보의 축적과 검색에서 색인작성자와 검색자가 사용하는 용어를 표준화된 어휘로 통일한 어휘집으로, 용어간의 개념관계를 동의어, 계층관계, 관련성 등의 측면으로 결합하여 체계적으로 배열해 놓은 용어통제표라 할수 있다.

이상에서 살펴 본 것처럼 시소러스의 정의는 전통적인 도서관에서 주제명목록카드의 작성과 배열을 위해 편찬한 「주제명표목표」의 개념에서 시작하여, 컴퓨터시스템의 문헌정보 검색 도입 시점부터는 컴퓨터 정보검색을 위한 정보생산자, 입력자, 이용자 사이에 용어통일의 역활을 수행하는 용어집으로 그 개념이 바뀌고 있음을 알 수 있다.

2.2 시소러스의 발달과 사용실태

최초의 시소러스라 할 수 있는 1952년 로제가 출판한 「Thesaurus of English Word and Phrases」는 원래 영어로 문장을 구성할때 적절한 용어를 선택할 수 있도록 용어를 개념에 따

라 배열한 동의어 사전으로, 문장에서 사용하고자 하는 적절한 용어를 안내해주는 사전에 지나지 않았다. 따라서 지금의 정보검색분야에서 적용되는 시소러스와는 커다란 차이가 있다.

정보검색을 위한 최초의 시소러스는 1959년 듀폰(Du Pont)의 시소러스가 효시를 이루며, 이어서 ASTIA(1960), 미국화학공업연구소(1961)에서 시소러스를 발행하다.

이후 미국에서는 미국국립의학도서관의 「MESH」(1963), 미국 공학관계학회연합회와 미국국방문헌센터가 합동으로 편찬한 「EJC thesaurus」, 「TEST(Thesaurus of Engineering and Scientific Terms)」(1967), 「NASA thesaurus」(1967), 「INIS thesaurus」(1970)등 많은 시소러스가 간행되었다.

또한 대표적 주제명표목표인 미국의회도서관의 「Library of Congress Subject Headings」를 1986년 온라인 포맷으로 바꾸어 상호참조기호를 시소러스 기호로 변경하여 쓰고 있다.

한편 일본에서는 일본과학기술정보센터의 「JICST科學技術シソ-ラス」(1975)가 대표적이며, 「中日ニュ-ス シソ-ラス」(中日新聞社 1974), 「NK-MEDIA シソ-ラス」(日刊工業新聞社 1984), 「日經シソ-ラス」(日本經濟新聞社 1988)등이 있다.

이외에도 영국규격협회의 「ROOT Thesaurus」(1981)와 교육분야 시소러스인 ERIC(1966), 과학기술분야 시소러스인 INSPEC(1973)등이 있으며, 오늘날 자연어검색기법이 발달하고 있음에도 불구하고 특정분야 문헌검색을 위한 시소러스는 계속 간행되고 있는 추세다.

국내의 시소러스는 아직 개념정립단계를 벗어나지 못하고 있다. 우리나라에서는 1980년 중앙일보사가 「중앙 IR-thesaurus」를 간행한 적이 있으며, 한국교육개발원의 「KEDI교육시소러스」(1981), 한국농촌개발연구원의 「농업경제문헌검색어집」(1985), SERI의 「과학기술용어시소러스」, 한국언론연구원의 「신문기사 종합시소러스」(1993)등이 간행되었다. 그러나 이들은 모두 책자형 시소러스로 색인이나 검색과정에서 자동으로 용어를 통제하기 불가능한 참조용 시소러스에 머물고 있다. 이외에 최근 국내 여러 기관에서 온라인 시소러스를 구축하고 있거나 계획중인 곳이 몇곳 있다.

2.3 정보검색 시스템에서의 시소러스

정보검색이란 생성된 정보를 정보전문가가 분석 가공하여 축적해둔 축적매체에서 이용자의 요구에 적합한 정보만을 탐색해내는 일련의 과정(이영자, 이경호, 1987)을 말하는데, 이를 위해 수반되는 시스템이 정보검색시스템이다.

정보검색과정은 크게 축적과 검색으로 구분되며 축적을 위해서는 주제분석과 색인작업이 필요하다.

색인이란 정보이용자가 쉽게 접근할 수 있도록 정보원에 포함된 정보내용을 쉽게 탐지할수 있는 소재지시기호를 달아 일정한 순서로 배열한 것을 말하는 것으로, 이용자가 목적에 따라 다양하고 특정한 정보를 필요로하는 관점에서 접근할 수 있도록 한 것이다.

색인은 관점에 따라 여러가지 형태로 분류할 수 있는데 언어통제 有無에따라 통제언어색인과 자연언어색인으로, 방법에따라 수동색인과 자동색인, 이들을 혼용해서 쓰는 반자동색인으로 구분할 수 있다.

자연어색인의 언어적문제를 해결하고 보다 효과적인 탐색을 하기 위해서는 이용자의 요구표현과 문헌표현을 일치시키고, 의미상 관련 있는 용어를 한자리에 모으는 언어통제가 필요하다. 이러한 통제된 언어에는 전통적인 도서관에서 관련문헌을 한 곳에 모으기 위해 주제명목록카드의 작성과 배열을 고려하여 편찬한 주제명표목표와 컴퓨터검색에서 관련문헌을 한꺼번에 탐색하기 위해 각 용어간의 상하관계를 체계적으로 구조화한 시소러스가 있다.

한편 색인 작성 방법에서 볼 때 훈련된 사서나 주제전문가가 수행하던 수동색인이 문헌량이 늘어남에 따라 컴퓨터가 처리하는 자동색인으로 바꾸어 나가지 않으면 안되게 되었다. 자동색인은 "컴퓨터에 입력한 문헌의 본문을 컴퓨터가 특수한 분석기법으로 분석한후 문헌의 내용을 나타낼 수 있는 단어나 단어구를 추출하여 작성한 색인(정영미, 1987)"으로, 과학기술문헌에 대한 보다 신속하고 완벽한 탐색 및 정보의 폭발현상에 대한 효율적 통제의 필요성과, 非수치정보를 처리할 수 있는 컴퓨터의 기능향상, 컴퓨터를 이용해서 언어의 구조와 의미를 분석하는 전산언어학의 발달, 그리고 인공지능의 발전 등으로 인해 등장하게 된 기법이다(사공철 외, 1990).

색인자의 매개없이 컴퓨터를 이용하여 색인어를 추출하는 자동색인방법에는 문헌속에서 중요개념을 추출하는 방법과 문헌의 모든 형태소를 추출하는 방법이 있는데, 전자가 자연어검색의 관점이라면 후자는 통제어검색 관점이라고도 할 수 있겠다. 그러므로 자동색인시스템인 경우 주제색인법에 있어서 색인자의 배경지식이 지식베이스구조에 표현되어야 하며 문헌의 모든 형태소를 추출하는 방법론과 복합적으로 연결되어야 하는데(최원태, 1986) 문헌의 중요개념을 추출하기 위해서는 시소러스가 중요한 도구가 된다.

자동색인은 색인어를 판정하는 기준에 따라 통계적 기법, 언어학적 기법, 문헌구조적 기법의 3가지로 나눌 수 있는데(정영미, 1987) 통계적 기법은 단어의 출현빈도에 따라 주제어로서의 중요도를 측정한 후 색인하는 방법을 말하며, 문헌구조적기법은 문헌내에서 단어가 나타난 위치에 따라 색인하는 것이다. 언어학적 기법은 불용어기법과 구문분석기법이 있는데, 불용어기법은 기능어를 제외한 다른 용어로 색인어를 선별해 내는 방법이고, 구문분석적 기법은 특정한 구문적 기능을 수행하는 단어나 단어구가 문헌의 내용을 나타낸다고 보고 구문을 단위로 분석하는 방법이다.

이상적인 자동색인은 대상문헌에서 색인어를 추출하고, 필요하다면 본문에서 사용되지 않은 단어를 추출해야 하며, 그 단어를 통제어휘 형태로 변환시킬 수 있어야 한다(Jones, 1976).

그러므로 자동색인에서 시소러스의 역할은 비디스크립터 용어일 경우 이를 통제어휘인 디스크립터용어로 바꾸어 색인어를 일치시키고, 여러가지 색인기법을 사용하여 주제어를 선정할 때 그 단어가 시소러스용어인가 아닌가를 확인, 시소러스 디스크립터용어일 때 가중치를 많이 부여해 검색효율을 높이는 것이다.

또한 시소러스는 검색시에도 색인어 개념에 대한 용어 안내, 용어 사이의 상호참조, 적절한 질의어 선정에 쓰인다. 동의어나 동음이의어 등과 같이 서로 달리 표현된 용어를 디스크립터로 통합해 검색할 수 있고 시소러스 용어간 관

계를 이용해 질의어와 관련된 모든 단어를 추가하거나 개념을 확장해서 검색 할 수 있게 하는 확장검색기법 등으로 재현율을 높일 수 있게 한다.

그러므로 시소러스는 정보의 급증, 정보매체와 정보요구의 다양화로 특징지울 수 있는 현대정보사회에서 정보검색시스템에 필수적인 요소로, 자동색인에서 시소러스 디스크립터용어로 색인어를 추출하고 검색시에도 다양한 이용자의 질문어를 디스크립터용어로 바꾸어 요구 정보를 검색하므로 재현율과 정도율을 높이는 역할을 수행한다.

2.4 신문기사데이타베이스의 특성

사회에서 일어나는 제현상을 전달해 주는 정보전달 매체인 신문은 "특정한 개인이나 기관이 뉴스를 수집 처리하여 新聞紙라는 대중매체를 통하여 독자들에게 제공하여 그들의 정신적 욕구를 만족시켜 주고 그 대가로 이윤을 추구하는 경제적인 동시에 문화적인 커뮤니케이션 행위라고 말할 수 있다"(박유봉 등, 1983).

신문에 담긴 내용인 신문기사는 사회 모든 분야에서 일어난 일들에 대한 객관적이고 충실한 기록인 동시에 사실의 기록이며, 속보성, 망라성을 지닌다(神尾達夫, 1983). 그러므로 "신문기사는 단순히 정치, 경제, 사회, 평론 등이 있는 거시적인 정보원일 뿐만 아니라 신문의 종류를 적절히 선택하면 과학기술, 기업, 업계, 상품활동 등 미시적인 동향을 알 수 있는 유력한 정보원이 될 수 있다"(竹谷一郎, 1978).

정보로서 가치를 지닌 신문기사를 데이타베이스로 만들기 위해서는 신문이 갖는 특성과 우리말 특성을 고려해야 한다.

먼저 신문이 갖는 특성으로는 신문은 일반적인 관심사에서 전문분야에 이르기까지 主題의 폭과 깊이가 다양하고, 복합 주제의 기사가 많으며, 표제와 기사내용이 일치하지 않는 경우가 많고, 일정한 간격을 두고 반복되는 주제의 기사가 많으며, 수명이 짧으며, 어휘가 어렵고 쉽게 변한다(한상길, 1991). 또한 신문기사는 일반문헌에 비해 문장이 짧으며,중요한 사항이나 결론적인 내용을 문두에 기술하는 특징이 있으며(神尾達夫, 1989), 일시적인 유행어, 기술용어, 관공 서식의 딱딱한 표현,완곡한 어법, 주의를 끄는 문구 등 다양한 문체와 관점을 지녀 색인작업을 어렵게 한다(Rothman, 1968).

다음으로는 신문기사가 갖는 언어적 문제인데, 국어는 구미언어와 달리 어근을 중심으로 어미를 덧붙여서 단어를 만드는 첨가어이며, 주어+목적어+동사의 문장구조로 모든 문법적 형태소는 반드시 어근이나 어간뒤에 오며, 주어가 생략되는 경우가 많고, 용언의 활용어미가 다양한 의미를 갖고 연결어미가 다양한 등의 특징을 지닌다(남기심, 고영근, 1991). 그러므로 어절에서 조사나 어미변화를 분리해 내고 색인할 수 있는 정형화된 규칙으로된 지식베이스를 만드는 것은 거의 불가능에 가깝다. 또한 한국어는 띄어쓰기가 자유롭다. 특히 신문기사는 맞춤법 규정이나 외래어표기편람 등이 있지만 지면의 제약 등으로 인해 무시되는 경우가 많기 때문에 이를 통제할 수 있는 시소러스가 필요하다.

III. 중앙일보 JOINS시소러스 개요와 색인시스템

3.1 시소러스의 개요

중앙일보사에서는 이미 15년 전인 1978년 신문기사 데이타베이스 구축을 목적으로 시소러스 구축에 착수 한 바 있다. 그 결과 초보적이기는 하지만 「중앙IR시소러스」라는 이름으로 국내 최초로 순 우리말로된 시소러스를 낸 바가 있었다. 그러나 언론통폐합과 정보산업에 대한 사회전반의 인식부족 등으로 무려 10년여 동안이나 중단되었다가 89년말부터 다시 구축을 시작해 약 6년간에 걸쳐 시소러스를 완료하고 CTS완성과 함께 본격 가동에 들어갔다.

JOINS시소러스는 1차적으로 신문기사 데이타베이스 관리를 위한 시소러스이다. 따라서 특정 주제에 대한 전문 시소러스가 아니라 정치, 경제,산업, 사회 등 모든 주제분야를 색인, 검색하기 위한 종합시소러스이다.

3.1.1 시소러스의 구성과 구조

본 시소러스는 '시소러스 파일'과 '시소러스 지원파일'로 구성되어 있으며 시소러스파일에는 일반용어에 대한 상하위 계층관계와 관련관계 등을 정의하는 사전파일,상하관계파일,관련관계파일로 구성되어 있다.

시소러스 지원파일은 회사명,기관·단체명,인명 등의 고유명사파일과 시소러스에 속하지는 않지만 정보검색시 필요하다고 판단되는 일반

	SEQ	상하관계용어	관련용어	
1단계상위어→	001	..... 컴퓨터	01 PC 통신	←관련어
검색어 →	002	..... [illegible]	02 MSX	
1단계하위어→	003	..... 8비트컴퓨터		
	004	..... 16비트컴퓨터		
	005	..... 32비트컴퓨터		
	006	..... 슈퍼퍼스컴		
	007	..... 다기능컴퓨터		
	008	..... 휴대용컴퓨터		
2단계하위어→	009	..... 랩탑		
	010	..... 팜톱		
	011	..... 노트북컴퓨터		
	012	..... 노트패드컴퓨터		
	013	..... 교육용컴퓨터		

〈그림 1〉 시소러스의 구조

용어를 계층·관련 관계 없이 나열한 보조키워드파일, 이상의 각 파일에 수록된 용어에 대한 동의어파일, 보조키워드 또는 자연어와 시소러스의 특정 관계를 정의한 특정어파일로 구성된다. 시소러스 구조는 〈그림 1〉과 같다.

3.1.2 시소러스의 규모

시소러스의 총 규모는 〈표 1〉에서 볼 수 있는 것처럼 시소러스파일은 디스크립터 17,487어, 비디스크립터 38,618어로 전체 56,105어이며 보조키워드, 기관단체명 40,902, 회사명 68,260, 인명 72,000을 포함한 전체 규모는 237,267어에 이른다.

시소러스에 채용되는 디스크립터 수는 시소러스의 전체 내용을 파악하고 또 각 용어간의 관계, 체계 모순을 배제하기 위해 가능한 억제할 필요가 있으며 비디스크립터의 경우는 색인의 일관성 또는 표준화, 검색의 편의(용어 선택의 용이성과 검색누수 방지)를 고려해 필요한 용어를 수록하게 되는데, 일반적으로 비디스크립터는 많을 수록 좋은 것으로 알려져 있다(岡

〈표 1〉 시소러스 용어의 전체 규모

파 일 명	디스크립터수	非디스크립터수	총 계
시소러스파일	17,487	38,618	56,105
보조KW, 기관·단체명	14,075	26,827	40,902
회사명	21,046	47,214	68,260
인 명	12,000	60,000	72,000
총 계	64,608	172,659	237,267

〈표 2〉 시소러스파일의 주제별 용어수및 관련어·어군수

주 제	디스크립터	非디스크립터	관련어	어 군
정 치	1,781	3,516	637	321
사 회	3,283	6,688	1,100	584
국 제	592	1,296	123	128
문화·예술	1,409	2,485	353	293
경 제	2,672	6,165	1,531	361
산 업	4,934	11,131	1,676	548
스 포 츠	1,159	5,092	179	133
학 문	1,657	2,245	188	151
총 계	17,487	38,618	5,787	2,519

野弘行, 1989).

본 시소러스는 현재 '디스크립터 : 비디스크립터'의 비율이 17,487 : 38,618어로 약 1 : 2 의 비율이다. 이는 본 시소러스가 완전 온라인 형태로 개발중이기 때문에 기존의 책자형 출판을 목적으로 한 시소러스와는 달리 각 디스크립터에 대한 비디스크립터의 갯수를 엄격히 제한하고 있지 않은 결과이다.

시소러스파일의 주제별 디스크립터 분포는 〈표 2〉에서 볼수 있는 것처럼 산업, 사회, 경제 순이며 비디스크립터의 분포도 같은 양상을 보인다, 관련어 분포에서는 산업, 경제, 사회 순으로 많으며, 어군별 분포는 사회가 584개의 어군으로 가장 많고 다음이 548개인 산업, 그 다음이 경제, 정치 순이다.

한편 디스크립터 계층별 용어수는 〈표 3〉에 나타난 것처럼 제2계층이 48.6%로 압도적으로 많고 다음이 제3계층, 제1계층 순이며 제5,6계층에서는 용어가 극히 적다. 이는 신문기사는 특정사안을 깊이 있게 다루기보다 광범위한 주제를 포괄적으로 다루기 때문이며, 산업과 경제부문 용어가 많은 것은 경제기사를 고려했기 때문이다.

3.1.3 시소러스의 기능

랑카스터(Lancaster)는 시소러스의 기능을 동의어, 동의적 표현 및 동음이의어를 통제함으로 색인자와 검색자가 일관되게 주제를 표현할 수 있게 하고 서로 관련 있는 단어를 연결시키므로 특정한 주제를 포괄적으로 검색할 수 있도록 하는데 있다(Lancaster, 1986)고 했다. 일반적인 책자형 시소러스가 색인이나 검색시 용어를 통일시키거나 확인하기 위해 참고로 삼는 보조수단정도로만 쓰이고 있는데 비해 완전히 온라인화된 중앙일보시소러스는 그 자체가 하나의 지식베이스이며 색인과 검색시스템의 일부로 작용한다.

〈표 3〉 디스크립터의 계층별 용어수

주 제 계 층	제1계층	제2계층	제3계층	제4계층	제5계층	제6계층
정 치	321	939	427	82	7	5
사 회	584	1,590	821	244	39	5
국 제	128	139	269	49	7	0
문화·예술	293	787	279	42	8	0
경 제	361	1,231	698	282	70	30
산 업	548	2,181	1,610	513	71	11
스 포 츠	133	891	123	11	1	0
학 문	151	757	499	211	35	4
총 계	2,519	8,515	4,726	1,434	238	55

가) 색인시스템에서 기능

본 시소러스는 자연어색인시스템 안에서 형태소 분석을 통해 불용어를 제거하고 명사를 추출한다. 그리고 이들 형태소 분석 결과 얻어진 약어, 유행어, 신조어 등 다양하게 표기되는 용어를 통일해 비디스크립터 용어를 디스크립터로 변환시킨다. 또한 자동색인에서 일반적인 가중치 부여방법인 출현위치와 빈도에 시소러스용어를 추가하여 시소러스용어 특히,어군내에서 하위어 일수록 높은 가중치를 부여하므로 보다 정확한 색인어를 얻을수 있게 한다.

또한 데이타 내에 어떤 시소러스 용어가 많이 등장했는가에 따라 데이타가 자동 분류되며 해당 주제의 시소러스 용어에 가중치가 부여되기도 하며 특정어파일을 통하여 자연어와 통제어 사이의 특정 관계를 이용하여 특정어를 자동 생성함으로써 시소러스와 자연어 시스템 모두의 활용성을 높일 수 있다. 이는 기사내에 '노대통령' '김대통령' 같은 불완전한 용어가 있을때 '노태우' '김영삼' '대통령' 처럼 완전한 용어로 자동 변환시키거나, '현대자동차'처럼 업계 전반에서 보면 단편적인 용어에 대해 '자동차업계'라는 대표성 있는 용어를 자동 생성함으로써 정보의 질을 높일 수 있다. 또 '부동산가'라는 복합어에서 '부동산' '가격' '부동산경기' 등 세분된 개념 또는 관련 용어를 자동 생성해 자연어와 시소러스 양자를 보완하기도 한다.

나) 검색시스템에서 기능

본 시소러스는 검색 시스템 내에서 용어확인 기능과 하위어 포괄검색 기능이 있다.

용어 확인기능은 검색하고자 하는 용어의 상하위어, 관련어를 확인함으로정확하게 정보에 접근할 수 있도록 해준다. 예를들어, 퍼스컴을 검색할 경우 이용자는 실제로 랩탑 관련 기사를 검색하고 싶지만 정확한 용어가 떠오르지 않을 경우가 있다. 이때 이용자는 퍼스컴이나 컴퓨터 어군을 검색함으로써 자기가 필요로 하는 용어인 랩탑을 찾아볼 수 있다. 또 실제로는 PC통신 관련 기사가 필요할 때 단순히 PC로만 접근해도 PC통신에 대한 안내를 받을 수 있다.

하위어 포괄검색은 어군 참조를 확장해주는 기능에 해당한다. 즉 어군에 대한 단순한 참조를 넘어 특정 어군에 속하는 용어들에 대한 개별적인 탐색을 한꺼번에 수행하는 것이다.

가령 '퍼스컴'에 대한 포괄검색을 실시함으로써 '퍼스컴'의 하위어로 등록된 '8비트퍼스컴' '슈퍼퍼스컴' '휴대용컴퓨터' 등 하위어에 관련된 데이타 까지 한꺼번에 검색할 수 있다.

이 외에도 검색어 통제 기능이 있는데 이용자가 정확한 용어로 접근하지 못하는 경우라도 시스템 내에서 동의어파일 탐색 등을 통해 검색어를 디스크립터 형태로 자동 변환하여 이용자가 정확한 검색어를 다시 지정하지 않고도 정보를 검색할 수 있도록 한다.

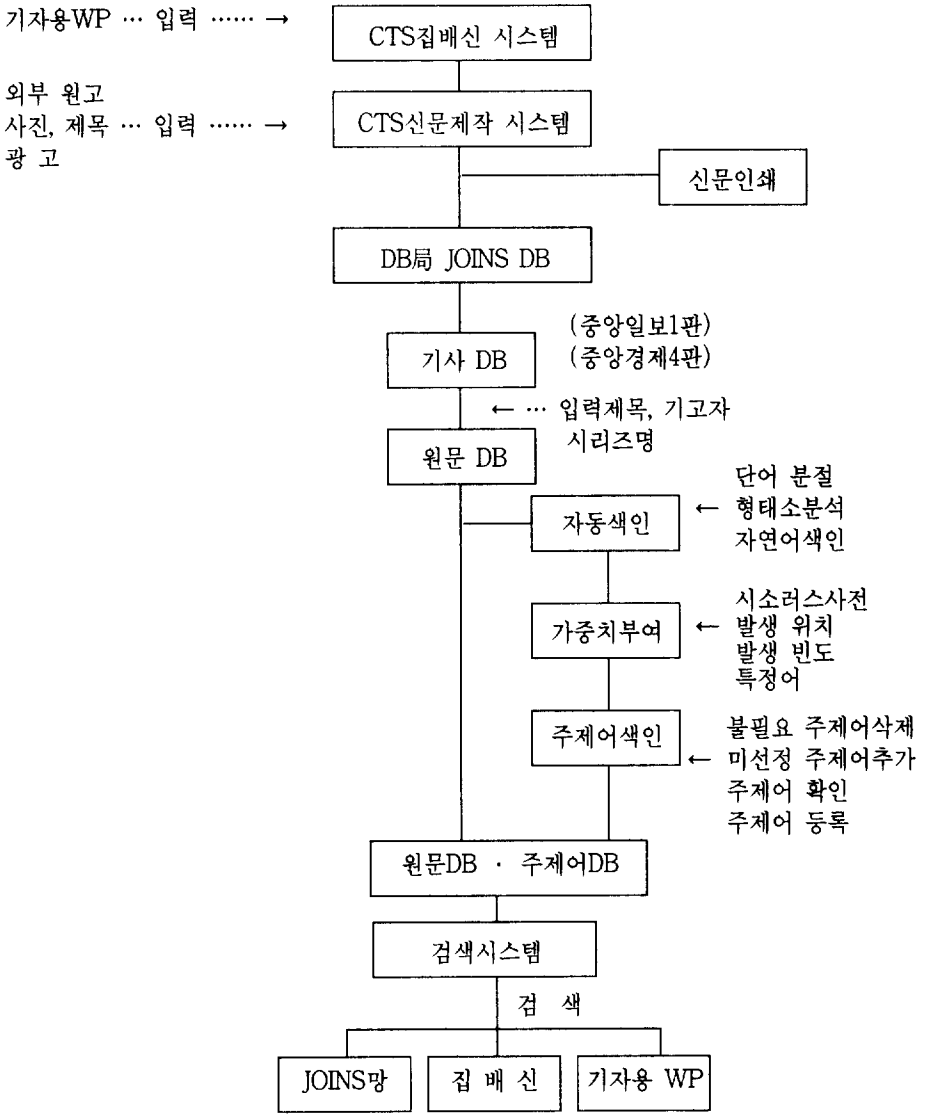
3.2 시스템 설계 및 구현

3.2.1 입력

JOINS 기사DB는 89년 11월부터 현재까지 약 30만건의 기사가 축적되어 있으며 하루 약 300건씩 늘어나고 있다. JOINS 기사DB의 전체흐름도는 〈그림 2〉와 같이 CTS신문제작시스템에

서 신문이 인쇄된뒤 全文이 전송되어 JOINS DB로 넘어오면 제목, 기고자, 시리즈, 기사유형, 사진유무등을 입력하여 원문 DB를 구축 한다. 명사사전, 불용어사전, 형태소사전을 이용하여 불용어처리와 형태소분석을 하고 시소러스사전의 지원을 받아 주제어를 추출하는 컴퓨터 자동색인 과정을 거쳐 색인자의 최종 확인을 거쳐 기사데이타베이스에 등록된다.

원문의 한자는 그대로 입력되며,제목은 대부분 그대로 입력하나 독자의 관심을 끌기 위해



〈 그림 2〉 중앙일보 색인시스템 전체 흐름도

주제와 관련이 적은 선정적표현으로된 제목은 가중치부여 효과를 높이기 위해 될수 있는데로 시소러스 용어가 포함된 제목으로 바꿔 입력한다. 또 기고자, 시리즈 등을 입력하고 국제기사를 구분하는데,이것은 기고자, 시리즈명의 색인어를 생성시켜 기사검색의 접근접으로 활용하기 위해서다.

입력 양식은 〈그림3〉과 같은데, 이 기사는 94년 3월 1일 중앙일보 1판 1면(종합 1면)기사로 내용은 국제기사이고 사진을 포함하고 있다. 이 기사를 쓴 기자는 이규연이고 시리즈 '경찰과 시민사회' 11번째 이다. 이 경우 '국제기사', '기고자이규연', '시리즈경찰과시민사회'라는 주제어가 자동 생성 된다. 이렇게 하므로 국제기사, 시리즈, 기고자명 만으로 검색하거나 배제해 검색할 수 있다.

미확인 : 30 ——— 〈 DB 축적용 기사교정 〉 ——— PAGE: 01/04

J940301 판 : 1 면 : 1 종합1면 구분 : F (F : 국제) 사진 : Y

AUTHOR : 이규연 COLUMN : 경찰과시민사회

표 제 : 〈경찰과시민사회〉 11. 일본은 하이테크 경찰시대-인공지능 110센터와 연결 범죄신고 즉시조회...용의자 색출

내용

001 (오오사카府 경찰본부내 범죄신고 번호인 110번 통신지령센터. 낮 12시55분)
002 (쯤 세쯔市 지역에서 살인사건이 발생했다는 제보전화가 걸려 왔다. 한 모니)
003 (터 요원이 받은 제보 내용을 중앙 컴퓨터에 입력함과 동시에 다른 요원은 발)
004 (생지점에 가깝게 있는 경찰서와 파출소 직원들을 무선으로 호출, 현장출동)
005 (을 명령했다. 경찰이 사건현장에 도착한 시각은 12시59분. 신고 전화가 접)
006 (수된지 불과 4분만에 출동한 것이다.)
007 (곧이어 "범인은 차량으로 도주했으며 예상 도주 방향은 **쪽"이라는 현장)
008 (경찰관의 보고가 접수 됐다. 중앙 컴퓨터의 인공 지능 시스템은 제보 내용)
009 (과 현장 경찰관의 보고를 토대로 도주 가능 범위와 검문 지점등을 예측, 대)
010 (형 상황판에 원모양으로 도주범위를 정해 빨간 전구로 검문지점을 표시)
011 (해 낸다.)
012 (컴퓨터의 분석 결과에 따라 현장에 배치된 경찰관들은 순찰차에 장착된 컴)
013 (퓨터단말기(이동 데이타 터미널) 를 통해 110센터와 교신하며 수배 차량과)
014 (동일수법 전과자를 즉석에서 조회 한다. 순찰차 1대가 움직이는 파출소 역할)
015 (을 하고 있는 것이다.)

PF:1-분리, 2-결합, 4-한칸, 5-1ST, 6-2ST, 7-BAK, 8-FOR, 9-삭제, 12-등록

SB 한글 전자

〈그림 3〉 입력 양식

3.2.2 자연어 추출

명사사전과 불용어 사전으로 불용어제거 후 단어를 분절하여 입력된 기사를 첫글자로부터 최장 일치되는 단어를 선정하고 일치된 단어의 다음 글자로부터 위의 방법으로 다시 최장 일치를 실행한다. 일치되는 단어가 없을 경우, 앞 용어의 뒷 글자를 탈락시킨 후 다시 최장 일치되는 용어를 찾는다. 앞의 용어가 없는 경우, 뒷 글자로부터 가장 길게 일치되는 어미나 조사를 삭제하고 나머지 앞 글자를 신규단어로 가정한다. 또한 분석하지 못한 용어 전체도 신규 단어로 가정한다. 마지막으로 분절된 단어를 합성하여 합성단어를 생성시키고 시소러스파일과 대조하여 동의어는 디스크립터로 변환시킨다.

자연어 추출시스템 흐름은 〈그림 4〉와 같다.

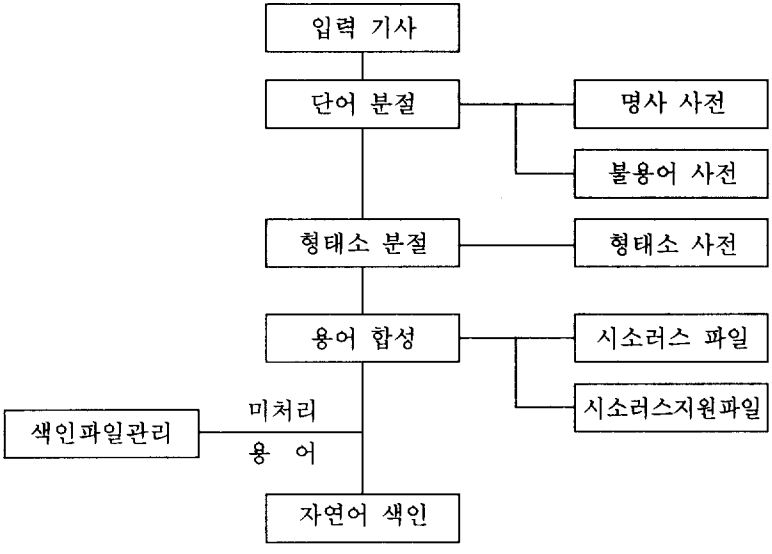
3.2.3 가중치 부여에 의한 주제어 추출

① 발생위치 : Free Term의 발생위치에 따른 중요도를 산정한다. 기사의 제목에 나타난 용어에 대한 가중치를 높이고, 기사내용을 5등분하여 각각의 가중치를 부여한다

② 발생 빈도 : 용어 발생빈도를 가중치에 고려한다. 이때 발생빈도가 지나치게 높은 일반명사 처리를 고려해야 한다.

③ 합성어의 처리 : 합성된 용어는 명사만으로 이루어진 합성어와 그렇치 않은 합성어로 나눌 수 있으며 참조파일에 없는 명사만의 합성어는 일반명사로 처리하며, 그외의 합성어는 주제어 선정에서 탈락시킨다.

④ 시소러스 용어 : 시소러스와 시소러스 지원파일 용어에 대한 가중치를 부여한다. 가중



〈그림 4〉 자연어추출시스템

치는 기사내의 시소러스 용어가 지닌 주제 분류를 분석하여, 빈출된 주제에 해당하는 용어에 가중치를 높인다.

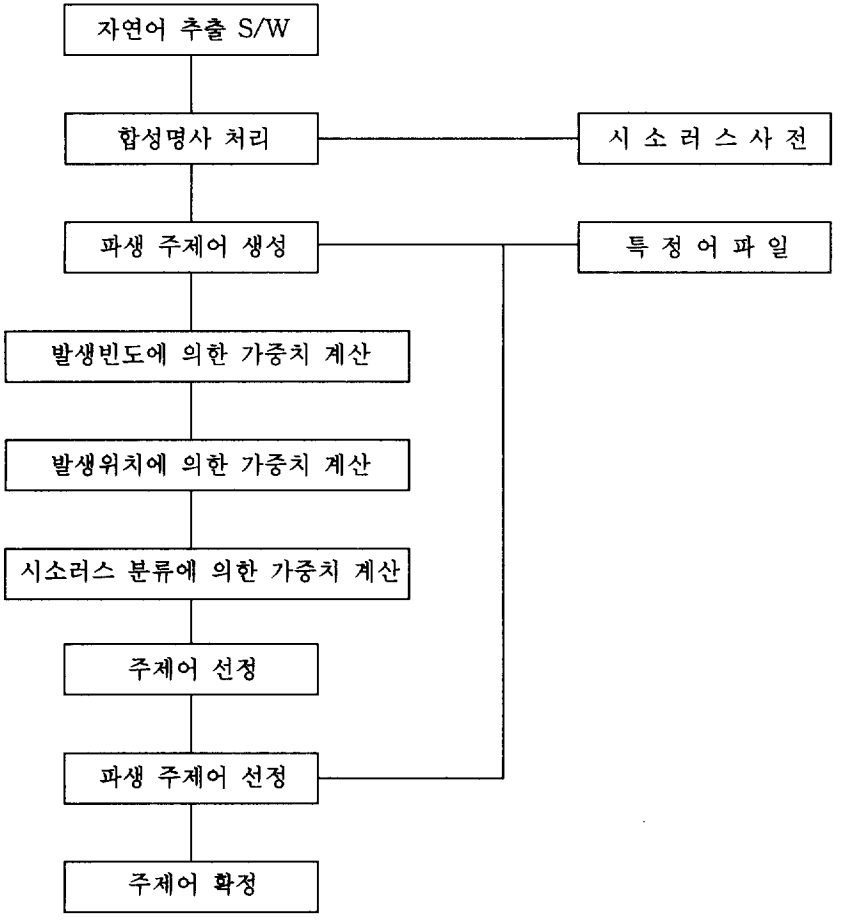
⑤ 특정어 추출 : 특정어를 색인할 때 특정어 파일에 정의된 추가 주제어를 함께 발생시키며 파생주제어는 특정주제어와 같은 가중치를 부여한다.

이외에도 병렬로 표현된 용어의 가중치를 줄이고 강조된 용어에 대한 가중치를 늘린다. 병렬표현어와 강조표현어는 형태소 분석을 포함한 구문분석 단계에서 처리할 수 있다.

주제어 추출시스템의 구성은 〈그림 5〉와 같다.

3.2.4 수작업 색인

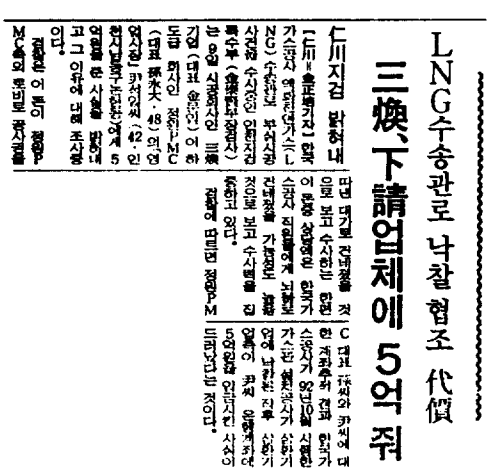
자연어색인과 통제어색인의 효율성에 대한 논란은 끊임없이 제기되고 있는데, 조성호(1989)는 "한글체제의 특성으로 볼 때 후통제처리를 반드시 필요로 한다"고 했고, 福岡 克



〈그림 5〉 주제어 추출시스템

(1992)도 "신문기사를 대상으로 하는 색인시스템에서는 단순한 자동추출에 의한 주제어로는 검색 만족을 줄 수 없기 때문에 상당부분의 주제어 추가가 필요하다고 했다. 신문기사의 경우 주제와 관련이 없거나 적은 용어가 많아 일반적인 자연어색인 가중치 부여방식으로는 검색할 때 많은 잡음이 나타나게 된다. 그러므로 이를 제거하고, 기사내에 나타나지 않았거나 나타났더라도 시소러스용어에 없는 신조어등을 입력하는 수작업색인을 병용할 필요가 있다.

수작업색인의 기준으로는 검색시스템에서 논리연산검색을 전제로하여 색인하며 본문에는



LNG수송관로 낙찰 협조 代價

三煥,下請업체에 5억 줘

仁川지검 밝혀내

〈 주 제 어 화 면 〉

NEWS KEY : J940309 핵심어 : 0017 분석어 : 0060 PAGE : 01/01

제목 : 三煥기업, LNG수송관로 낙찰협조 對價로 下請업체에 5억 줘
인천지검 밝혀내

001 기고자김정배
002 LNG
003 수사
004 부실공사
005 삼환기업
006 뇌물
D 007 하청업체
D 008 공사
009 한국가스공사
D 010 정원
D 011 한국
D 012 지방검찰청
013 인천
D 014 검찰
015 정원PMC
D 016 운수
D 017 수송관
*** 입찰부정
*** 하도급부조리

입력단어 ( )
선 택 ( )
PF2 : 저장 PF3 종료 PF5 수정취소 PF7 앞 PF8 뒤 PF9 색인취소 PF12 저장확인

SB 한글 전자

〈그림 6〉 수작업색인 화면

등장하지 않지만 전체내용을 잘 나타내주는 용어를 입력하고 띄어쓰기, 접속사로 인해 추출하지 못한 색인어를 추가한다.

〈그림 6〉 수작업색인 화면에서 볼 수 있듯이 시스템이 추출한 자연어 60어 중 가중치부여로 생성된 주제어는 입력단계에서 정형화된 색인어로 추출되는 '기고자김정배'를 포함 17개이다. 주제어순서는 가중치 점수순으로 배열되며, 이 과정에서 시스템은 '부실시공'과 같은 비디스크립터 용어는 '부실공사'라는 디스크립터 용어로 생성시킨다. 또한 주제와 관련 없는 '정원'이나 '한국'같은 시소러스용어도 생성시키게 된다. 시스템이 추출한 주제어 중에서 색인어로서의 가치가 없는 주제어를 제거시키고 시스템에서 찾아내지 못한 색인어이지만 이 기사와 같은 기사를 검색할 수 있는 일관된 주제어 즉,이 기사의 중심주제인 '입찰부정', '하도급비리'를 입력 시킨다. 이렇게 해서 시스템이 추출한 주제어 17개 중 8개를 제거하고, 2개를 추가시켜 11개의 주제어를 최종 등록 한다.

이 색인시스템은 컴퓨터로 자동적으로 시소러스사전을 탐색하여 동의어, 유사동의어들을 디스크립터로 통일시키고 여러가지 가중치 기법을 동원한 자연어색인으로 주제의 정확도를 높인뒤 수작업으로 최종적인 색인등록을 하므로 시간과 경비를 줄이고 정확한 색인어를 선정할 수 있게 한다.

그러나 수작업색인 역시도 일관된 색인을 위해서는 색인자를 위한 수작업 매뉴얼을 갖출 필요가 있으며 분류번호 검색 방식도 고려할 필요가 있을 것이다.

IV. 시스템 평가

본 연구의 실험대상은 중앙일보 94.3.1-3.31일까지 휴간일 2일을 제외한 29일의 기사중 DB로 처리할 수 없는 도표,만화,광고 등과 DB로 구축할 가치가 없다고 생각되는 소설,날씨,바둑기보, 독자투고 등을 제외한 4,101건의 기사를 대상으로 했다. 실험대상 기사건수를 1개월로 제한한 것은 이 연구가 신문기사의 주제분석이나 검색성능을 비교하기 위한 것이 아니라 색인시스템의 성능만을 측정하는 실험이므로 실험대상 기사수를 많이 잡을 필요가 없다고 보았기 때문이다.

분석 결과를 살펴보면 다음과 같다.

1) 주제어 추출 결과

3월 1개월 전체 분석 대상기사는 4,101건으로 일평균 141건에 이르렀고, 자연어처리시스템에서 시스템이 분석한 명사 수는 기사당 75개, 가중치를 부여한 뒤 시스템이 추출한 일정 점수 이상인 핵심주제어는 기사당 10.9개였다. 이 가

〈표 4〉 주제어 추출결과

구 분	총분석 기사	자연어처리	핵심주제어	등록된색인어
전 체	4,101	307,508	44,653	19,413
평 균	141/일	75/기사	10.9/기사	4.7/기사

운데 후통제(수작업색인)과정을 거쳐 최종 등록된 색인어는 기사마다 평균 4.7개 였다.

이것은 입력단계에서 정형 파일에 입력 정형화된 색인으로 추출되는 시리즈,기고자, 국제기사 등의 색인어는 제외시킨 숫자로, 이를 포함하면 기사당 평균 색인어는 6개 정도이다.

2) 위치별 색인어 분석

등록된 최종색인어를 분석해보면 〈표 5〉에서 볼 수 있듯이 제목에 등장한 용어가 가장 높은 32.6% 비율로 나타났으며, 다음으로 본문외 용어, 기사 첫문장, 본문 순이다. 이것은 신문기사의 구성방식이 역피라미드 즉, 중요한 내용을 위로 올리고 다음에 설명을 덧붙이는 방식이기 때문이며, 특히 제목에 있는 용어가 가장 높은 비율로 나타나는 것은 제목 입력때 가능하면 선정적 용어를 배제하고 시소러스 용어 중심으로 입력하기 때문이다.

본문외 색인어는 본문에 쓰인 용어가 시소러스의 비디스크립터 즉 동의어나 유사동의어일 경우 시스템 내에서 자동으로 디스크립터로 변환시키는 경우와 시스템이 추출 할 수 없는 색인어를 색인자가 수작업 과정에서 입력한 경우이다.

3) 색인어로 추출된 시소러스용어 분석

최종 등록된 색인어 가운데서 시소러스사전에 있는 용어가 차지하는 비율은 〈표 6〉에서 나타난 바와 같이 79.7%로, 시소러스 사전에 등록되지 않은 인명, 기관단체명, 기업명 등 고유명사와 신조어 등의 시소러스 이외의 용어보다 월등히 높다.

이밖에도 본문외 용어 5,693개 중에서 시소러스 용어와 일치된 비율이 1,387개로 비디스크립터를 디스크립터로 전환한 비율이 25%에 이르고, 전체 등록된 색인어 중에서 사람이 등록한 색인어와 시스템이 선정한 색인어의 일치율은 63.9%이며, 최종 등록된 색인어중 본문에 나타

〈표 5〉 위치별 색인어 비율

구 분	최종색인어	제 목	첫 문 장	본 문	본 문 외
색인어수	19,413	6,333	4,512	2,873	5,693
비 율		32.6	23.3	14.8	29.3

〈표 6〉 색인어로 추출된 시소러스 용어 비율

구 분	시소러스 용어	시소러스외 용어	계
용 어 수	15,482	3,931	19,413
비 율	79.7	20.3	

$$Np = \frac{\text{수작업 색인어} \cap \text{자동추출 색인어}}{\text{수작업 색인어}} \times 100$$

$$Ne = \frac{(\text{수작업색인어-본문외 색인자추가어}) \cap \text{자동추출 색인어})}{\text{수작업색인어 - 본문외 색인자추가어}} \times 100$$

$$Snp = \frac{\Sigma\ Np}{\text{총 색인 기사 건수}}$$

$$Sne = \frac{\Sigma\ Ne}{\text{총 색인 기사 건수}}$$

Np = 기사당 색인어 일치율
Ne = 기사당 색인어 보정 일치율
Snp = 색인어 일치율 평균
Sne = 색인어 보정일치율 평균
Np = 색인어 일치율의 합
Np = 색인어 보정일치율의 합

나지 않은 색인어를 빼고 난 순수한 자동색인 일치율은 80.99%에 이른다.

이상 결과에서 볼 수 있는 것은 최종 입력된 색인어는 제목에 가장 많이 포함되어 있기 때문에 기사데이타베이스 구축시 신문지면에 나타나는 본문과 관련없는 선정적이거나 축약어를 사용한 제목을 수정할 필요가 있다. 또 가중치 부여방법과 시소러스 사전을 사용함에도 불구하고 본문 이외의 용어가 많이 등록되는데 이것은 신문기사가 갖는 특성과 복합명사 사이의 띄어쓰기, 접속사 등이 있으므로 시스템에 의한 자동색인만으로는 주제어 선정에 미흡함이 있음을 알 수 있다. 이 문제를 해결하기 위해서는 단일문장내에 나타나는 명사들을 결합시켜 시소러스사전에 등록된 용어와 비교 일치될 경우 색인어로 추출하는 기법 등이 개발되어야 할 것으로 여겨진다.

V. 결 론

이상의 결과를 종합하면 다음과 같다.

1. 최종 등록된 색인어는 정형적으로 등록되는 기사유형, 기고자등에 따른 정형화된 색인어를 제외하면 기사당 평균 4.7개로 기사당 평균 75개정도로 추출되는 후통제를 하지 않은 순수 자연어색인에 비해 검색 노이즈를 크게 줄일 수 있다.
2. 최종 등록된 색인어 분포 비율은 제목에서 가장 많이 발생하고 있는데 이것은 제목 입

력 과정에서 시소러스 용어를 중심으로 제목을 수정할 필요가 있음을 알 수 있다.

3. 자연어 색인과정에서 동의어, 유사동의어 등 비디스크립터를 디스크립터로 바꾸고, 가중치를 부여할 때 중요한 역할을 하는 시소러스 사전은 신문기사색인시스템에서 긴요한 역할을 하고 있음을 알 수 있다.
4. 최종 등록된 전체 색인어중 제목이나 본문에 나타나지 않는 용어가 29.3%에 이르므로 자연어처리시스템이 아무리 우수한 성능을 지녔다고 해도 후통제 수작업색인이 필요하며, 복합명사에서 색인어를 추출하기 위해서는 단일 문장내에 등장하는 명사를 결합해 복합명사로 구성, 시소러스사전과 대조해 일치되면 색인어로 추가 등록하는 기법이 개발되어야 할 것이며, 또한 아무리 완벽한 시소러스를 구축했다고 해도 신조어, 새로 발생하는 사건 등에 대비 시소러스파일을 계속적으로 수정 보완할 필요가 있다.
5. 개별 주제에 관한 일관된 검색을 위해서는 분류번호 검색기법의 도입이나 색인된 용어를 시소러스파일과 비교, 분류번호를 생성하는 자동분류 기법을 도입하는 문제를 고려할 필요가 있다.

참고문헌

권정임, "우리말 신문기사를 위한 시소러스 개발에 관한 연구". 미간본석사학위논문, 연세대학교 대학원, 1991.

岡野弘行, "JICST 科學技術用語シソラス," 情報の科學と技術, v.39, n.12(1989), pp.558-566.

남기심, 고영근, 표준국어문법론, 서울 : 탑출판사, 1991.

ランカスタ, F. W., 松村多美子, 鈴木祐滋 譯, シソラスの構築と利用, 東京 : 情報科學技術協會, 1989.

朴有鳳 外, 新聞學理論, 서울 : 박영사, 1983.

福岡 克, "新聞DB 檢索語を點檢 : 標準化とシステチュックを望む," 新聞研究, N.142 (1992, 4), pp.111-115.

사공철, "정보검색에 있어서의 Theasurus 導入에 관한 基礎研究," 未刊本碩士學位論文, 연세대학교 산업대학원, 1972.

사공철 외, 최신정보검색론, 서울 : 구미무역 출판부, 1990.

神尾達夫, "新聞情報," ドクメンテーション研究, v.33, n.8(1983,8), pp521-528.

_____, "新聞記事データベスにおけるキーワード自動抽出," 情報管理, v.32, n.4 (1989, 6), pp. 283-293.

日本ドキュメンテーション協會, シソーラス, 동경 : 同協會, 1970.

이영자, 이경호, 정보학개론, 대구 : 경북대 출판부, 1987.

정영미, 정보검색론, 서울 : 정음사, 1987.

조민원, "文獻情報檢索을 위한 한글 Theasurus 作成研究," 未刊本碩士學位論文, 고려대학교 경영대학원, 1972.

조선희, "KINDS에 있어서 자연어와 동의어의 검색 효율성 비교 연구," 未刊本碩士學位論文, 숙명여자대학교 대학원, 1993.

조성호, "컴퓨터를 이용한 한글 自動索引시스템 구축에 관한 연구," 未刊本 碩士學位論

文, 연세대학교 산업대학원, 1989.

竹谷一郎, "新聞情報の加工分析," ドクメンテーション研究, v.28, n.11(1978, 6), pp.521-528.

최원태, 격문법을 이용한 자동색인 및 탐색확장에 관한 연구, 미간본석사학위논문, 연세대학교 대학원, 1986.

韓相吉, "記事資料標準分類表에 관안 分析的 考察," 未刊本碩士學位論文, 경북대학교 대학원, 1991.

홍승의, "신문기사의 자동색인시스템에 관한 연구," 미간본석사학위논문, 연세대학교 대학원, 1991.

Aitchison, Jean, Gilchrist, *Thesaurus Construction*, 2nd, London : Aslib, 1972.

Browson, H, *In Proceeding df the Informational Study on Conference on Classifition for Internation Retrieval*, London : Aslib, 1957.

Gilchrist, Alan, *The Thesaurus in Retrieval*, London : Aslib, 1771.

Gillum, T.L., "Compiling a Technical Thesaurus," *Journal of Chemical Documentation*, v.4, no.1(1974). pp.29-32.

International Organization for Standardization, ISO 2788 : *Guideline for the Establishment and Development of Monolingual Thesauri*, Geneva : ISO, 1974.

Jones, Kevin P., "Towards theory of Indexing", *Jounal of Documention*, v.36(1976), pp. 118-120

Lancaster, F.W., *Vocabulary control for information retrival*, 2nd, Alington : Information Resources Press, 1968.

Luhn,H.P., "A Statistical Approaches to Mechanized Encodig and Searching Library Information", *IBM Journal of Research and Development* v.1(1957). pp.309-317.

Rothman, John, "Automated Information Processing at the New York Times," *Proceedings of the 31st annual Meeting of ASIS*, v.5, (1968), p.85-87.

Saracevic, Tefko, *Introduction to Information Science*, New York: Borker, 1970.

Soergel, D, *Indexing Languages and Thesaurus Construction and Maintenance*, Los Angeles : Melvile, 1974.

Wall, Eugene, "Seymiotic Development of Thesauri and Intromation Systems : A Cse Study," *Journal of ASIS*, v.26(1975), pp. 71-72.