우리말 정보자료를 처리하는 지능형 정보검색시스템의 설계

(Design of a Korean Intelligent Information Retrieval System)

정 영 미*

□ 목 차 □

초 록

본 연구에서는 지능형 정보검색의 개념을 정립하고, 지능형 정보검색시스템의 모형을 제시하였다. 실제로 우리말 문헌을 검색하는 지능형 시스템을 구현하였으며, 이 시스템은 자연언어 인터페이스, 이용자 모형화, 격관계를 이용한 자동색인, 복수의 검색기법 등을 수용한다.

키워드: 지능형 정보검색시스템, 정보검색시스템, 전문가시스템, 이용자모형화, 자연언어처리, 지능형정보검색, 지식베이스, 자동색인, 격분법, 격관계색인, 탐색확장, 매칭함수검색

ABSTRACT

A prototype model of intelligent information retrieval system is presented with the definition of intelligent information retrieval. An intelligent information retrieval system for Korean documents was designed, and the system was implemented with Turbo Prolog 2.0 and Turbo Pascal 5.5. The characteristics of the system include natural language interface, user modeling, automatic indexing by case relationship, and multiple retrieval techniques.

1. 서 론

1.1 연구개요

정보검색이란 정보자료의 내용을 분석, 가공하여 구축한 정보화일로부터 이용자의 정보요구에 적합한 정보를 탐색하여 찾아내는 일련의 과정을 의미한다. 컴퓨터를 정보검색에 응용한 이래 정보검색에 있어서의 주된 연구과제는 문헌의 내용을 가능한 한 정확히 표현할 수 있는 자동색인기법과, 높은 검색효율을 가져올 수 있는 효과적인 검색기법의 개발이었다.

정보검색이란 용어 자체는 제 1 세대 컴퓨터

이 논문은 1990년도 문교부 지원 한국학술진흥재단의 자유공모과제 학술연구조성비에 의하여 연구되었음.

* 연세대학교 문헌정보학과 교수

가 출현한 1950년대초에 미국에서 처음 사용되었으나, 실제로는 유럽을 중심으로 하여 1940년대에 특히 활발하게 전개되었던 다큐멘테이션 개념을 대신한 것이다. 정보검색의 발달과정을 보면, 1940년 이전에는 인쇄물을 이용한 수작업 시스템을 통해 색인과 검색작업이 이루어졌으나 1940-1950년대에는 반기계화된 정보검색시스템들이 나타났으며, 1960년대에는 컴퓨터를 이용한 시스템들이 개발되었다. 1970년대 이후는 네트웍을 통해 정보검색시스템에 접근할 수 있게 됨으로써 전세계적인 정보유통시대가 시작되었으며, 1980년대 이후로는 인공지능을 응용한 전문가시스템과 지능형 정보검색시스템에 관한 연구가 특히 활발하게 진행되고 있다.

본 연구의 목적은 차세대 정보검색시스템은 다양한 지식베이스와 추론기능을 이용하는 지능적인 지식기반시스템이 되어야 할 것이라는 가설하에, 지능형 정보검색시스템의 모형을 제시하고, 특히 우리말 문헌을 처리하는 시스템을 구현하는 데 있다. 시스템 구현은 연구기간의 제약으로 인하여 이용자-시스템 인터페이스 모듈, 자동색인모듈, 검색모듈에 국한하였으며, 특히 자연언어 처리기능을 강조하였다. 실험대상 문헌으로는 「정보관리학회지」 제1권-제7권(1984년-1990년)에 수록된 논문 63편을 선정하였으며, 프로그래밍 언어로는 Turbo Prolog 2.0과 Turbo Pascal 5.5를 사용하였고, IBM PC/AT 호환기종상에서 시스템을 구현하였다.

1.2 지능형 정보검색시스템에 관한 요구

정보검색에 컴퓨터가 본격적으로 이용되기 시작한 1960년대와 네트웍을 통해 정보검색시스템에 온라인으로 접근하게 된 1970년대까지만 하더라도, 데이타베이스의 탐색은 거의가 도서관의 사서나 기타 정보관련기관의 정보전문가에 의하여 수행되었다. 그러나 1980년대에 들어서 온라인 정보검색시스템의 단말기로 사용할 수 있는 개인용 컴퓨터의 보급이 확대되고, 또한 일반인이 관심을 갖는 생활정보를 제공하는 데이타베이스 및 정보검색시스템이 증가하면서, 데이타베이스의 탐색이 최종이용자에 의하여 수행되는 경향이 나타나기 시작하였다. 구체적인 지표로는 온라인 정보검색시스템이나 데이타뱅크를 이용하는 개인가입자의 비율을 살펴볼 수 있는데, 그 이유는 가입자가 기관이 아니고 개인인 경우는 데이타베이스의 탐색자는 사서나 정보전문가가 아닌 최종이용자라고 가정할 수 있기 때문이다. 국내에서도 1991년 7월 현재 한국데이타통신의 통신망을 통해 미국의 데이타뱅크인 DIALOG을 이용하는 가입자수는 758기관(명)이며, 이 가운데 개인가입자가 180명으로 전체의 약 24%를 차지하고 있었다. 주로 생활정보를 제공하는 국내 데이타뱅크인 천리안 Ⅱ의 경우는, 개인가입자의 비율이 이보다 훨씬 높아서 전체가입자 9035기관(명) 가운데 80%가 개인가입자인 것으로 조사되었다. 앞에서 언급한 요인 이외에 CD-ROM 형태의 데이타베이스의 증가는 온라인 정보검색시스템에 비해 값이 싸고 편리한 CD-ROM 시스템의 이용을 국내외적으로 증대시키고 있으며, 이것 또한 '최종이용자에 의한 탐색'의 보편화 현상에 크게 기여하고 있다.

이와같이 최종이용자가 바로 정보탐색자가 되는 현상은 데이타베이스 탐색기법에 관한 전문적인 지식이 없이도 효과적으로 데이타베이

스를 탐색할 수 있도록 도와 줄 시스템에 대한 요구를 발생시켰으며, 이러한 요구는 문헌정보학 영역에서 초기에 개발된 전문가시스템들이 거의가 대규모 데이타뱅크의 온라인 탐색을 보조하는 탐색중개 전문가시스템이었던 점에 잘 반영되어 있다[1-4]. 이러한 온라인 탐색중개 전문가시스템들은 최종이용자와의 대화기능, 자연언어 형태의 질문 처리기능, 정형의 탐색문 형성과 탐색전략의 수립기능, 데이타베이스 선택기능 등의 제공을 목표로 하고 있으나 실제 시스템에 따라 기능상에 차이가 있으며, 또한 모체가 되는 정보검색시스템이 요구하는 정형의 탐색문 형성을 주목적으로 하기 때문에 이용자 개개인의 특성을 고려한 탐색전략의 수립, 이용자 피드백에 의한 탐색확장, 보다 완벽한 자연언어 인터페이스 등의 바람직한 기능은 제공하지 못하고 있다. 따라서 단순한 탐색중개인의 역할을 벗어나 효과적이고 지능적인 탐색이 가능하도록 하는 시스템이 요청되었으며, 이러한 요구가 지능형 정보검색시스템에 관한 연구를 촉진시킨 한 요인이 되었다고 볼 수 있다. 실제로 전형적인 지능형 정보검색시스템으로 꼽고 있는 I^3R(Intelligent Intermediary for Information Retrieval)은 시스템 명칭이 말해 주고 있는 것과 같이 지능적인 '전문가 중개(expert intermediary)' 시스템으로 개발된 것이라는 점에 유의할 필요가 있다.

이와같이 최종이용자를 위해 자연언어 인터페이스와 이용자 특성에 맞는 다양한 탐색전략의 수립 등의 기능을 갖는 '이용자 편의시스템'으로서의 정보검색시스템의 개발이 지능형 정보검색시스템을 등장시킨 중요한 요인이었으나, 이외에 데이타베이스의 수량적 증가도 이에 못지 않는 요인으로 생각할 수 있을 것이다. 최근 미국에서 출판된 데이타베이스 디렉토리(The Cuadra Directory of Databases on Disc)에는 일반인이 접근할 수 있는 온라인 데이타베이스가 4700개 이상 수록되어 있으며, CD-ROM이나 디스켓 등에 소장된 데이타베이스도 1500개 이상이 수록되어 있는 것을 볼 수 있다[5]. 현재의 온라인 데이타베이스의 수는 1980년의 600여개, 1984년의 2400여개에 비해 볼 때[6] 크게 증가하였으며, 이로부터 앞으로의 증가추세를 가늠할 수 있을 것이다. 데이타베이스의 수량적 증가는 많은 양의 정보를 보다 효율적이고 정확하게 색인할 수 있는 자동색인기법에 대한 필요성과 높은 검색효율을 얻을 수 있는 효과적인 검색기법에 대한 필요성을 더욱 증대시키게 되는 것이다. 결국 효과적인 이용자-시스템 인터페이스와 자동색인 및 검색은 모두 다양한 지식에 기반한 기능적인 방법이 되어야 할 것이라는 가설하에 지능형 정보검색시스템에 관한 연구가 수행되고 있다고 볼 수 있을 것이다.

1.3 전문가시스템과 지능형 정보검색시스템

인공지능을 응용한 정보시스템에 관한 연구는 1980년대에 들어서 본격화되었으며, 초기에는 온라인 탐색, 참고업무, 분류, 색인, 편목 등의 특정한 업무를 수행하는 전문가시스템에 관한 연구가 주류를 이루었다. 그러나 1980년대 후반에 이르러 '지능형 정보검색시스템'이라고 부르는 정보검색시스템들이 실험적으로 개발되기 시작하면서 '지능형 정보검색'은 정보검색분야의 새로운 연구주제로 관심을 끌고 있다.

MYCIN이나 DENDRAL 등 대표적인 전문

가시스템들이 의사나 화학자 등 특정한 주제분야의 전문가를 보조할 목적으로 개발되었던 것과 마찬가지로 문헌정보학 영역의 전문가시스템들은 참고사서, 색인전문가, 분류전문가, 편목전문가, 온라인 탐색전문가 등 특정한 정보처리업무를 수행하는 전문인을 대신하거나 보조할 목적으로 개발되었음이 주지의 사실이다. 이러한 전문가시스템들의 특징은 전문적인 지식을 담고 있는 지식베이스와 추론기능을 제공하는 지식기반시스템이라는 것이다. 지능형 정보검색시스템도 기본적으로 지식기반시스템이며 제공되는 정보가 문제해결에 사용된다는 점에서는 전문가시스템과 유사하지만, 시스템의 목적이 전문인을 대신하는 것이 아니라 정보검색과 관련된 모든 기능을 지능적으로 처리하여 시스템의 성능을 높이는 데 있다는 점에서 전문가시스템과 구별된다. 브룩스(Brooks)도 전문가시스템과 지능형 정보검색과의 관계를 다룬 논문에서, 정보검색이 전문가시스템의 응용영역이 되기 어려운 점으로서 처리할 기능의 다양성과 폭넓은 주제영역 및 필요로 하는 지식의 다양성을 우선적으로 꼽고 있음을 볼 수 있다[7].

정보검색시스템이 사람처럼 '지능적'이기 위해서는 단순한 데이타나 정보 이외에 체계화된 지식을 소장하고 이용할 수 있어야 하고, 또한 자연언어 이해능력과 문제해결을 위한 추론능력을 가져야 한다. 따라서 지능형 정보검색시스템은 첫째, 자연언어 형태의 질문을 이해할 수 있어야 하며, 둘째, 자연언어로 된 문헌 텍스트를 처리하되 언어의 의미론적 지식과 주제영역 지식을 이용하여 문헌의 내용을 정확히 표현함으로써 문헌에 관한 색인지식을 구축할 수 있으며, 셋째, 이용자에 관한 지식을 이용하여 다양한 탐색전략을 실현할 수 있는 정보검색시스템이 되어야 할 것이다.

'지능형 정보검색(intelligent information retrieval)'이란 개념의 정립은 1983년 스파크 존스(Sparck Jones)에 의해 시도된 것으로 보이는데, 그녀는 지능형 정보검색시스템을 '정보요구와 문헌간의 관계를 결정하기 위한 추론능력과 지식베이스를 갖는 시스템'이라고 정의하였다[8]. 브룩스(Brooks)의 정의는 좀더 구체적이며 특히 이용자에 관한 지식의 활용을 강조하고 있다[7]. 즉, 지능형 정보검색시스템은 주제지식, 문헌에 관한 지식(색인), 이용자에 관한 지식을 소장하며, 개별적인 이용자에 관한 정보와 이용자가 해결하고자 하는 정보가 입력되면 위의 지식을 이용하여 문제해결에 필요한 문헌들을 추론하여 검색하는 컴퓨터시스템이라고 보고 있다. 이러한 시스템은 이용자 개개인의 특성을 고려해야 하며, 이용자의 문제기술 수준에서 문제를 처리할 수 있는 능력이 있어야 한다는 것이다.

실질적으로 지능형 정보검색에 관한 연구는 기존의 통계적인 정보검색기법이 안고 있는 검색효율상의 문제점에 대한 인식에서 시작된 것으로 보이며, 크로프트(Croft)가 지적한대로 이 분야의 연구는 (1)정보검색을 응용대상으로 한 인공지능의 각 연구영역(자연언어처리, 지식표현, 추론 등)에서의 기본적 연구와, (2)전통적인 정보검색기법과 인공지능 분야의 기법을 접목한 정보검색시스템에 관한 연구로 나뉘어 수행되고 있다[9].

자연언어 처리, 지식표현, 추론 등이 인공지능분야의 연구대상이 되어왔음을 고려할 때, 앞에서 언급한 바와 같은 특성과 정의를 갖는 지

능형 정보검색시스템은 결국 '인공지능을 이용하여 정보의 축적과 검색을 수행하는 시스템'이라고 말할 수 있다. 구체적으로 지능형 정보검색시스템은 다양한 지식베이스 및 데이타베이스의 구축, 지식기반 색인, 자연언어 질문의 처리, 추론을 통한 적합정보의 검색, 지식베이스를 이용한 탐색확장 등의 기능을 수행하도록 설계되어야 할 것이다. 또한 궁극적으로는 다양한 형태의 정보매체에 소장된 정보를 검색할 수 있는 다중매체(multimedia)시스템과 하이퍼텍스트(hypertext)시스템으로서의 기능도 제공하여야 할 것이다.

2. 지능형 정보검색시스템 실례

지능형 정보검색시스템의 범주에 속하는 시스템으로 I³R[10], IOTA[11], CODER[12], RESEARCHER[13], 비스와스(Biswas) 등이 개발한 지식기반 문헌검색시스템[14] 등을 꼽을 수 있다. 또한 온라인 정보검색시스템에 대한 인터페이스시스템으로 개발된 IR-NLI Ⅱ 시스템은 지능형 정보검색시스템이 요구하는 이상적인 인터페이스 기능을 제공하고 있다는 점에서 주목할 만하다[15]. 이러한 시스템들은 실제로 시스템 구성과 기능면에서 다양성을 보이고 있기 때문에 지능형 정보검색시스템의 전형적인 모형을 제시하지 못하고 있다.

이 가운데 CODER 시스템은 전문가시스템 기법과 지식베이스를 이용한 시스템으로 설계되었으나 일부만 구현된 상태에서 발표되었으므로 간단히 살펴보면 다음과 같다. 이 시스템은 전자우편을 통해 전송되는 다양한 유형의 메시지를 검색하는 시스템으로서, 시스템 구성요소로는 문헌분석과 검색에 각각 사용되는 2개의 블랙보드와 이에 관련된 여러개의 전문가 모듈, 이용자인터페이스 관리기, 질문분석기, 이용자모형 구축기, 그리고 이용자모형 지식베이스, 사전, 문헌 데이타베이스 등이 있다. CODER 이외의 시스템에 관해서는 시스템의 구성과 특성을 중심으로 하여 시스템별로 살펴보고자 한다.

2.1 I³R

이 시스템은 정보검색의 여러 단계에서 이용자를 보조하는 기능을 갖고 있으며, 인터페이스 관리기와 스케쥴러 외에 다음과 같이 일련의 전문기능(expert) 모듈들로 구성되어 있다.

(1) 이용자모형 구축기능: 시스템 사용경험, 탐색의 목표(높은 재현율이나 높은 정확률 등), 관심주제영역 등 이용자에 관한 정보를 수집하여 정보요구모형 구축과 탐색전략 선택에 사용한다.

(2) 정보요구모형 구축기능: 이용자의 질문으로부터 정보요구모형을 구축하며, 검색된 문헌에 대한 적합성 정보를 이용하여 정보요구모형을 수정한다. 이 기능은 질문으로부터 색인어와 가중치를 추출하며 다양한 형태의 질문(자연언어 질문, 불리안 논리식, 가중치가 부여된 단어나 단어구)을 처리한다.

(3) 주제영역지식 이용기능: 이용자모형과 지식베이스내 주제지식을 이용하여 원질문에 포함된 개념과 관련된 다른 개념을 추론한다.

(4) 탐색제어 기능: 두가지 검색기법(확률적 검색과 클러스터화일 검색)가운데 적절한 검색기법을 선택하여 수행한다.

(5) 브라우징 기능: 지식베이스를 브라우즈

함으로써 적합문헌을 찾아내는 비공식적인 검색방법을 제공한다. 브라우징과정은 특정한 문헌, 저자, 또는 색인어로부터 시작할 수 있으며 링크를 따라 지식베이스내의 다른 항목으로 이동한다.

(6) 설명기능: 일반적인 규칙기반시스템에서와 같이 이용자의 요청에 따라 시스템의 행위를 설명한다.

I³R시스템의 지식베이스는 문헌에 관한 지식, 색인지식, 이용자에 관한 지식, 주제영역지식으로 구성되는데, 문헌에 관한 지식베이스는 서지 데이타베이스와 유사하며 색인지식은 도치색인화일에 해당된다. 주제영역지식은 개념과 개념간의 관계, 개념과 색인어간의 관계를 포함하고 있으며 시소러스가 제공하는 지식과 매우 유사하다.

이 시스템의 특징으로는 첫째, 검색전에는 주제지식을 담고 있는 지식베이스와 이용자에 관한 지식을 이용하여 탐색문을 수정할 수 있으며, 검색후에는 이용자피드백에 의해 탐색문을 수정하여 재탐색을 할 수 있다는 것, 둘째, 이용자가 지식베이스를 브라우즈할 수 있는 기능을 제공한다는 것, 셋째, 이용자모형에 따라 적절한 탐색전략을 허용한다는 것이다. 제한점으로는 자연언어 형태의 질문을 허용하고 있기는 하나 언어학적으로 처리하는 대신 단순한 통계적 기법에 의해 탐색어를 추출한다는 것과, 색인작업시에도 통계적 기법을 사용하기 때문에 문헌의 내용을 정확히 표현하기 힘들다는 점을 들 수 있다.

2.2 IOTA

지능형 정보검색시스템의 프로토타입으로 개발된 프랑스의 IOTA 시스템의 특징은 전체시스템 구조에 있어서 전문가시스템과 정보검색시스템을 별개의 구성요소로 유지하되, 전문가시스템을 두 시스템의 협력과정에 있어서의 스케쥴러로 이용하고 있다는 것이다.

전문가시스템 부분에서는 지식베이스, 단기 데이타베이스, 절차베이스를 유지하며, 정보검색시스템 부분에서는 문헌 텍스트 데이타베이스 이외에 단어사전, 색인화일, 시소러스 등을 유지한다. 전문가시스템의 지식베이스내의 지식은 생성규칙 형태로 표현되는데, 생성규칙들의 조건부는 변수에 부여되는 조건(예: IF task=query-analysis)을 기술하며 실행부는 조건이 만족될 때 처리할 일련의 작업(예: THEN call parser)을 지시한다. 단기 데이타베이스는 전문가시스템의 현재 상태에 관한 모든 정보를 소장하며, 절차베이스는 지식베이스를 구성하는 생성규칙의 실행부에 올 수 있는 모든 절차들을 소장한다. 반면 정보검색시스템의 단어사전과 시소러스는 문헌텍스트의 자동색인, 질문의 분석과 이해, 탐색문의 재형성 등의 작업에 사용된다.

정보검색시스템 부분이 수행하는 주된 기능은 질문의 처리와 검색이며 질문의 처리는 질문분석모듈과 구문패턴 대조모듈에서 수행된다. 질문분석모듈은 불어 자연언어로 표현된 질문의 구문을 분석하여 명사(구)를 추출한 다음 불리안 논리를 사용한 일차 탐색문을 생성한다. 이 모듈은 질문의 분석결과를 출력시킴으로써 질문의 분석이 잘못된 경우에는 이용자가 수정을 하도록 한다. 구문패턴 대조모듈은 질문분석모듈에서 생성된 일차탐색문의 개념들을 구문패턴 대조 알고리즘을 통해 시스템의 색인어로

변환시켜 최종탐색문을 작성한다. 검색모듈은 불리안 논리에 의한 검색을 수행하며, 검색결과는 전문가시스템의 평가와 관련된 규칙에 의해 평가된 다음 평가결과가 만족스럽지 않을 때는 시소러스를 이용하여 탐색어를 확장하거나 탐색문의 논리구조를 변경하여 탐색문을 재형성한다.

이 시스템의 특징은 자연언어 질문을 분석하여 불리안 논리 탐색문을 생성한 다음 이용자의 질문을 구성한 원래의 개념을 시스템의 색인어로 대체함으로써 검색효율을 높이고 있다는 것이다. 또 하나의 특징은 이용자에 관한 지식을 이용하여 검색결과를 평가하고 탐색문을 재형성하도록 한 점이다. 실제로는 이용자의 유형(초보자, 중간경력자, 전문가)에 관한 정보와 검색결과로부터 얻은 몇가지 파라미터(적합성 함수에 의해 평가된 검색문헌의 적합성, 검색문헌의 수, 평균 적합성 수준 등)를 조건으로 하는 규칙에 의해 검색결과를 평가한다. 예를 들어 이용자유형이 전문가이고 검색문헌수가 20개 이상이면 검색결과는 나쁜 것으로 판정되어 탐색문을 재형성하게 되는데, 이것은 이용자 피드백에 의한 평가에 비해 지나치게 시스템중심적인 평가로 보인다. 이 시스템의 제한점은 질문의 분석과정을 강조하고 있는 반면 검색과정이 상대적으로 소홀히 취급되고 있다는 것이다. 이 시스템은 단순한 불리안 논리 검색만을 허용하고 있으므로 이용자의 특성이나 요구수준을 충분히 고려한 탐색전략의 수립이 어려운 것으로 보인다.

2.3 비스와스(Biswas) 등의 지식기반 문헌검색시스템

비스와스 등이 개발한 지식기반 문헌검색시스템은 지식베이스를 이용하여 자연언어로 된 질문을 분석하고, 추론을 통해 적합문헌을 검색한다는 점에서 지능형 정보검색시스템이라고 볼 수 있을 것이다. 이 시스템은 자연언어 인터페이스, 추론기제, 시스템 제어기, 지식베이스, 문헌 데이타베이스로 구성되며, 각 구성요소에 관한 설명은 다음과 같다.

(1) 자연언어 인터페이스: 자연언어 처리기에 의해 자연언어 질문을 처리하여, 주제개념, 출판년, 검색문헌 수의 세가지 탐색요소로 구성되는 내부형식으로 변환한다.

(2) 추론기제: 뎀스터-세이퍼(Dempster-Shafer) 증거이론을 이용하여 적합문헌을 검색하는 검색기제와 검색된 문헌에 순위를 부여하는 순위부여기제로 구성된다.

(3) 시스템 제어기: 자연언어처리기와 추론기제를 연결해 주며, 효율적인 탐색을 위한 탐색전략을 제공한다.

(4) 지식베이스: 주제영역의 중요한 개념들을 나타내는 키워드(구)로 구성되며 개념간의 관계도 포함한다.

(5) 문헌 데이타베이스: 색인전문가에 의해 각 문헌은 색인어와 가중치 벡터로 표현된다.

이 시스템의 가장 큰 특징은 자연언어 질문을 ATN 문법과 도식대조방식에 의해 분석하되 퍼지개념을 포함한 질문을 처리할 수 있다는 것이다. 예를 들어 "Retrieve a few very recent and important survey articles about multivalued logic"과 같은 질문에서 a few, very recent, important와 같은 한정어들은 퍼지함수에 의해 적절히 해석된다. 이 시스템의 다른 특징은 검색시 지식베이스를 이용하여 탐

색어를 확장할 수 있으며, 뎀스터-세이퍼 이론에 기초한 수학적 검색모형을 사용하여 질문과 문헌간의 관련성을 추론한다는 것이다. 그러나 이 시스템이 전형적인 지능형 정보검색시스템이 되기 위해 해결해야 할 과제로는 문헌의 내용을 정확히 표현할 수 있는 적절한 자동색인 기법의 도입, 이용자에 관한 지식의 수집과 이용, 검색모형의 실용화 및 검색전략의 다양화 등이 있다.

2.4 RESEARCHER

RESEARCHER는 다른 정보검색시스템들이 문헌의 내용을 단순한 단어(구)로 표현하여 색인을 작성하는 것과는 달리 자연언어 텍스트를 처리한 결과를 일정한 형식으로 표현하여 지식베이스를 구성하고, 이로부터 해답을 찾아내는 일종의 질문응답시스템으로 설계되었다. 이 시스템은 특허자료의 초록을 분석하여 지식베이스를 작성하며, 특히 디스크 드라이브장치와 같이 부품들간의 계층적 관계를 설정할 수 있는 물체에 관한 지식을 표현하고 있다. 이 시스템은 텍스트를 이해하고, 새로 입력되는 유사한 지식을 통합하여 일반화된 지식을 추론하며, 지식베이스로부터 질문에 대한 응답을 찾아내는 기능을 갖고 있다. RESEARCHER는 궁극적으로 (1)입력된 텍스트로부터 일반화된 지식을 유추해냄으로써 새로운 지식을 지식베이스에 추가하는 기능(학습기능), (2)단어에 관한 의미지식 이외에 이미 입력된 텍스트에 관한 지식을 이용하여 텍스트를 처리하는 기능, (3)이용자의 수준에 맞는 응답을 제공하는 기능 등을 갖는 지능형 정보검색시스템을 지향하고 있다.

2.5 IR-NLI Ⅱ(Information Retrieval-Natural Language Interface Ⅱ)

일반이용자로 하여금 직접 온라인 정보검색시스템에 접근하도록 하는 전문가 인터페이스 시스템으로 개발된 IR-NLI Ⅱ 시스템은 이용자 모형화 기능을 갖는 대표적인 시스템이며, 정보검색 전문가 하부시스템과 이용자 모형화 하부시스템으로 구성된다. 정보검색 전문가 하부시스템이 수행하는 주된 기능은 (1)자연언어 대화의 처리(이해 및 대화모듈), (2)이용자 정보요구의 파악과 탐색문 형성(추론모듈), (3)탐색전략의 수립과 정보검색시스템으로의 전달(형식화 모듈) 등이다. 정보검색 전문가 하부시스템의 지식베이스로는 탐색중개인의 기술과 지식을 모형화한 '전문가 지식베이스'와 시소러스와 사전 등의 '주제영역 지식베이스'가 있다.

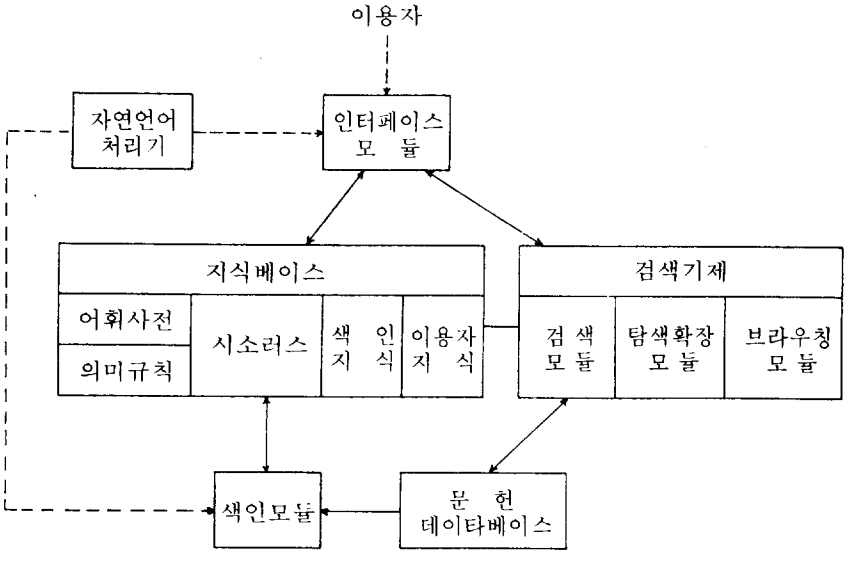
이용자 모형화 하부시스템은 탐색세션과 기타 여러 세션을 통해 이용자 모형화작업을 수행하며, 기본적으로 (1)이용자와 시스템간의 대화로부터 이용자 모형화에 관련된 정보의 도출과 (2)이용자모형의 구축과 갱신의 기능을 갖는다. 시스템의 현재 이용자에 관한 정보는 정보검색 전문가 하부시스템으로부터 이용자 모형화 하부시스템으로 전달되어 이용자모형을 구축하며, 역으로 이용자 모형화 하부시스템에서 구축된 이용자 모형은 정보검색 전문가 하부시스템에서 탐색전략 수립에 이용된다. 이용자 모형화 하부시스템은 앞의 두가지 기능을 수행하는 모형구축기와 히스토리 관리기의 두 모듈로 구성되며, 히스토리 관리기는 이용자 개개인이 수행한 각 탐색세션의 요약기록을 장기 데이타베이스에 소장한다. 구체적인 이용자 모형 구축과정에 관해서는 「4.1 이용자-시스템 인터

페이스 모듈」에서 상술할 것이다.

3. 지능형 정보검색시스템의 모형

2장의 시스템 실례에서 살펴본 바와 같이 지능형 정보검색시스템은 다양한 지식베이스를 활용하는 지식기반시스템이어야 하며, 동시에 소장된 지식을 이용하여 적절한 추론을 행할 수 있는 시스템이어야 한다. 추론기능은 다음과 같이 여러 단계에서 제공될 수 있다. 첫째, 색인지식을 이용한 정보요구와 문헌과의 관련성 추론, 둘째, 이용자지식을 이용한 적절한 탐색전략의 추론, 셋째, 주제지식을 이용한 새로운 탐색어의 추론, 넷째, 기존지식을 이용한 새로운 지식의 추론이 가능해야 할 것이다. 지식베이스에 소장되는 지식의 유형과 표현방법도 추론기능에 따라 다양해지며 주제영역지식, 이용자에 관한 지식, 탐색전략 선택에 관한 지식, 자연언어 형태의 질문과 문헌 텍스트 처리를 위한 언어학적 지식 등이 이에 관련된다.

그림-1은 지능형 정보검색시스템의 기본적인 모형으로서 시스템 구성요소로는 인터페이스 모듈, 색인모듈, 검색모듈, 탐색확장 모듈, 브라우징 모듈과 문헌 데이타베이스 및 다양한 지식베이스를 포함한다. 이 모형에서 색인모듈과 인터페이스모듈은 각각 문헌 텍스트의 분석과 자연언어 질문의 처리에 독립된 자연언어 처리기를 호출하여 사용하도록 하였다. 지식베이스에는 주제관련 지식베이스, 언어학적 지식베이스, 이용자에 관한 지식베이스가 있으며, 주제영역지식을 담고 있는 시소러스와 문헌내용을 표현한 지식을 담고 있는 색인지식베이스가 주제지식베이스를 구성하며, 어휘사전과 의미해석규칙 등이 자연언어 처리를 위한 언어학적 지식베이스를 구성한다. 이용자에 관한 지식은 단기적인 지식과 장기적인 지식으로 구성된다.



(그림 1) 지능형 정보검색시스템 구성도

4. 우리말 문헌을 처리하는 지능형 정보검색시스템

4.1 이용자-시스템 인터페이스모듈

브래즈닉(Brajnik) 등은 정보검색시스템에 대한 인터페이스시스템의 설계시 다음의 두가지 문제를 고려해야 한다고지적하였다[15].

-이용자와 시스템간의 언어적 차이를 어떻게 극복할 것인가?

-이용자의 정보요구의 분석, 적절한 탐색전략의 형성, 검색결과의 평가 등에 있어서 각 개념적 단계에서 이용자를 어떻게 지원할 것인가?

실제로 시스템을 설계하는 데 있어서, 첫번째 문제는 자연언어 형태의 질문을 완벽하게 처리할 수 있는 기능을 통해 해결할 수 있을 것이며, 두번째 문제는 이용자 개개인의 특성을 반영하는 이용자 모형화를 통해 해결하려는 시도가 계속되고 있다.

4.1.1 이용자 모형화

이용자 모형화(user modeling)란 일반적으로 사람-기계간의 상호작용을 증진시키기 위해 시스템이 이용자에 관해 갖고 있는 정보를 활용하는 것을 의미한다[16]. 다니엘스(Daniels)는 정보검색에 관련된 인지모형에 관한 리뷰논문에서 현재 정보검색분야 연구의 주목표는 이용자가 직접 데이타베이스를 탐색하도록 하는 것이라고 보고, 정보검색시스템의 성능을 향상시키기 위해서는 이용자 개개인에 관한 이용자모형을 활용하는 것이 필요함을 역설하고 있다[17]. 그는 이용자모형을 '시스템이 이용자에 관해 갖고 있는 인지모형'이라고 정의하고, 특히 분석적 인지모형으로서의 이용자모형은 수행중인 작업에 관한 이용자의 지식과 이용자의 목표, 배경, 계획, 선호하는 학습/상호작용 스타일, 시스템 사용경험 등에 관한 지식을 포함하는 것이라고 기술하고 있다.

효과적인 이용자-시스템 인터페이스를 통해 시스템의 성능을 높이기 위해서 시스템 이용자에 관한 모형의 구축이 필요하다는 것은 지난 수년간 강조되어 왔지만, 실제 지능형 정보검색시스템에서 이용자에 관한 지식이 충분히 활용되기 위해서는 더 많은 연구가 필요할 것으로 보인다. 2장에서 기술한 시스템들 가운데 IOTA 시스템에서는 검색결과를 평가하기 위해 이용자 유형에 관한 지식을 이용하였고, I^3R 시스템에서는 이용자 유형 외에 이용자의 탐색목표(높은 재현율 등), 관심주제 등의 지식을 수집하여 이용자-시스템 상호작용 유형의 결정과 탐색전략의 선택 등에 사용하였다.

가장 전형적인 이용자 모형화는 IR-NLI Ⅱ 시스템에서 시도되었으며, 이 시스템의 이용자 모형 구축과정은 다음과 같다.

(1) 시스템은 대화를 통해 이용자에 관한 기본적인 정보(교육, 직업배경 등)를 수집한다.

(2) 수집된 정보를 이용하여 스테레오타입 지식베이스내의 스테레오타입의 활성화 방법을 검사한다. 이때 만족되는 모든 스테레오타입이 활성화된다.

(3) 활성화된 스테레오타입들 가운데 이용자 모형 구축의 기초가 될 가장 적절한 스테레오타입을 판별해 낸다.

(4) 선택된 스테레오타입에 기초하여 정보의 수집과 확인과정을 통해 이용자모형을 반복적으로 정련해간다.

(5) 탐색세션 종료시 완성된 이용자모형이 이용자모형 지식베이스에 소장된다.

이용자 모형화에서 사용되는 '스테레오타입(stereotype)'이란 개념은 공통적인 특성을 공유하는 이용자 집단에 대한 기술을 의미하며, 새로운 이용자에게는 먼저 적절한 스테레오타입이 이용자모형으로 제공된다. 스테레오타입은 전형적인 이용자모형 구축과정에서 보편적으로 사용되고 있다[10, 18]. IR-NIL Ⅱ 시스템의 이용자모형은 〈이용자 이름〉 〈모형 히스토리〉 〈이용자 프로화일〉 〈이용자의 지식〉의 네 요소로 구성되며 프레임 형태로 표현된다. 〈이용자 프로화일〉은 교육배경, 직업배경, 정보검색 배경, 개인적 특성, 일반적 탐색 요구사항(탐색주제영역, 탐색목표, 출력형식 등)을 포함한다. 〈이용자의 지식〉은 IR-NIL Ⅱ 시스템이 운영되는 환경에 대한 이용자의 지식(주제영역, 데이타베이스, 정보검색시스템 등에 관한 지식)을 포함한다.

앞에서 언급한 시스템들 이외에도 제한적이긴 하지만 이용자 모형화를 시도한 예로는 THOMAS[19]와 ASK[20] 시스템이 있으며, 보다 본격적인 시도는 이용자 모형화기법을 실험할 목적으로 도서관 이용자에게 소설을 추천하는 전문가시스템으로 개발된 Grundy[18]와 원예분야의 참고정보를 제공하는 전문가시스템인 PLEXUS[21]에서 찾아볼 수 있다.

브래즈닉(Brajnik) 등은 지능형 정보검색시스템에 적합한 이용자모형의 특성을 다음과 같이 지적하고 있다[15].

- 모형유형은 이용자 개개인에 관한 지식을 표현하는 분석적 인지모형일 것.
- 모형구축방법은 이용자가 모형에 포함될 지식을 직접 제공하는 것이 아니라 시스템에 의해 필요한 정보가 수집되어 모형이 구축되는 암시적 모형화(implicit modeling) 방법일 것.
- 모형에 소장되는 지식의 특성에 있어서는 단기적 지식(특정한 정보요구, 이용자 계획과 의도, 탐색의 목적 등)과 장기적 지식(이용자 교육수준, 주제에 관한 지식, 시스템 사용경험 등)을 다 포함하기 위해 단기적 모형화와 장기적 모형화 방법을 모두 사용할 것.

4.1.2 자연언어 처리

자연언어를 처리하는 컴퓨터시스템에 관한 연구는 1950년대에 시작되어 지금까지 계속되고 있으며 시기별로 응용분야의 변화를 보이고 있다[22]. 1950년대와 1960년대 초기까지의 연구대상은 주로 기계번역이었으며, 1960년대 중기부터 1970년대 중기까지는 질문응답시스템(LUNAR, LSP, PROTOSYNTHEX 등)에 관한 연구가 활발하였다. 1970년대 후기에 들어서는 실용화된 데이타베이스시스템에 자연언어로 접근할 수 있는 전단시스템들(REQUEST, INTELLECT, PLANES 등)이 개발되었다. 이외에 자연언어 텍스트를 언어학적 기법에 의해 처리하는 자동색인에 관한 연구가 1970년대 이후 지속되고 있으며, 최근에는 지식베이스를 이용하는 지식기반 색인시스템에 관한 연구가 수행되었다[23]. 1980년대에 들어서는 특히 자연언어 인터페이스를 제공하는 전문가시스템과 지능형 정보검색시스템에 관한 연구가 활발히 수행되고 있다. 국내에서도 한국어 처리에 관한 연구가 꾸준히 진행되고 있으며, 특히 질문응답

시스템과 데이타베이스 전단시스템, 자동색인 시스템 등에 자연언어 처리기술을 응용한 예를 많이 찾아볼 수 있다[24-32].

자연언어 처리시스템들은 초기에는 주로 특정한 단어(구)의 도식대조방식에 의해 문장을 분석하는 초보적인 시스템들이었으나, 후에 ATN이나 변형문법 등을 이용하면서 본격적인 구문분석을 시도하였으며, 구문분석과 의미분석을 병행한 시스템들도 나타났다. 실제로 자연언어 처리시스템을 설계할 때 기본적으로 고려할 점으로는 구문분석의 수준(완전한 분석, 부분적 분석)과 사용할 문법의 선택이 있다. 또한 의미분석의 수준도 고려할 수 있는데, 불완전한 구문분석을 보조하는 수단으로서 사용하는 경우와; 완전한 구문분석을 수행하되 의미자질을 이용하여 보다 완전한 처리를 목적으로 하는 경우에 각각 다른 수준의 의미분석을 수행하게 된다. 실제로 완전한 구문분석을 수행하는 시스템이라도 자연언어의 모든 현상을 완벽하게 처리하기는 힘들다. 따라서 처리가능한 구문의 범위가 시스템에 따라 다르게 설정되며, 실제로 개발된 시스템들은 제한된 형식의 문장만을 처리하는 시스템에서부터 내포문의 처리, 복잡한 명사구의 처리, 관계대명사나 생략문의 처리 등이 가능한 시스템에 이르기까지 다양하다. 또한 대부분의 시스템이 소규모의 어휘사전과 의미사전이나 의미해석규칙 등을 이용하여 특정한 주제영역만을 처리대상으로 하고 있기 때문에 시스템의 범용성 내지는 이식성이 떨어지는 문제가 제기된다. 현재로는 자연언어 처리에 있어서 분석대상이 되는 주제영역의 폭과 언어학적 분석의 깊이는 상호 조정관계에 있는 것으로 보인다. 다시 말해 주제영역의 폭이 넓으면 심층적이고 완벽한 언어학적 분석은 기대하기 어렵다는 것이다.

빅커리(Vickery) 등이 제안한 자연언어 인터페이스시스템이 갖춰야 할 기능 가운데 특히 지능형 정보검색시스템에서 고려해야 할 기능들을 열거하면 다음과 같다[21].

- 생략문, 숙어, 문법적 오류, 단편적 문장 등의 처리가 가능한 융통성있는 구문분석 기능
- 대화의 정확한 전달과 해석이 가능한 강력한 커뮤니케이션 기능
- 시스템이 '무엇을' '왜' 할 수 있는지, 했는지, 하고자 하는지에 관한 설명기능
- 설명을 통해 대상물을 식별할 수 있는 기능
- 이용자의 수준을 진단하고 이용자모형을 구축하는 기능
- 이용자가 입력오류를 수정할 수 있는 기능
- 이용자 편의시스템으로서의 기능
- 이용자를 위한 지도 기능
- 적절한 수준의 시스템 응답시간

4.1.3 인터페이스 모듈의 기능

지능형 정보검색시스템의 인터페이스 모듈은 기본적으로 이용자 모형화와 자연언어 질문의 처리가 가능해야 하며, 다음과 같은 기능을 수행하여야 할 것이다.

(1)이용자-시스템간의 대화형 인터페이스를 제공한다. 구체적으로 이용자에 관한 지식의 수집, 이용자 질문의 입력, 검색결과의 출력, 검색결과에 대한 설명, 브라우징 화면의 제공, 이용자 피드백의 입력 등을 수행한다.

(2) 자연언어 형태의 질문을 분석하여 정형의 탐색문을 생성한다.

(3) 이용자 모형을 구축한다.

(4) 이용자 모형을 통해 이용자에게 적합한 인터페이스 방식을 제공하며, 적절한 탐색전략을 선택한다.

(5) 이용자가 입력오류를 수정할 수 있으며, 질문의 분석결과를 출력하여 잘못분석된 경우 수정을 가하도록 한다.

위의 기능 가운데 본연구에서 구현한 기능은 이용자에 관한 지식의 수집과 단기적인 이용자 모형의 구축, 이용자모형을 이용한 탐색전략의 수립, 자연언어 질문의 분석과 탐색문의 생성 등이며, 입력오류나 질문 분석결과의 수정기능은 구현하지 않았다. 이용자모형은 이용자-시스템간의 대화를 통해 구축하였으며, 시스템 사용경험과 현행 탐색목표에 관한 정보를 수집하여 단기적인 지식으로 사용하였다. 이용자 개인에 관한 지식(초보자/유경험자; 높은 재현율/높은 정확률)을 담은 이용자모형은 블랙보드에 소장하여 인터페이스 모듈과 검색모듈에 의해 접근하도록 하였다. 이용자모형이 구축된 다음 이용자가 '초보자'인 경우에는 시스템에 관한 설명화면을 출력하고, '유경험자'인 경우에는 설명화면을 생략한다. 또한 이용자가 '높은 재현율'을 원하는 경우에는 검색모듈에서 매칭함수에 의한 검색을 수행하며, '높은 정확률'을 원하는 경우에는 격관계 그래프에 의한 검색을 수행하게 된다.

이용자모형의 구축과정이 완료되면 이용자의 정보요구를 자연언어 형태로 입력하도록 하는데 시스템의 효율성을 고려하여 형식상의 제약을 가하였다. 실제로 사용될 수 있는 문장형태의 질문은 '..에 관한 문헌을 검색하라', '..에 대한 논문을 찾아주시오', '..에 관한 문헌을 원한다' 등 다양한 형식이 될 수 있으나 밑줄친 부분은 잉여정보이므로 생략하는 것이 이용자나 시스템을 위하여 효율적이다. 마찬가지로 '에 관한 논문'이나 '을 다룬 논문' 등도 생략할 수 있으나 이용자가 정보요구를 기술할 때 주제만을 표현하기 보다는 '..에 관한 문헌'과 같은 형식으로 기술하는 것이 자연스럽다는 점을 고려하여 이 부분은 생략하지 않도록 하였다. 또한 정보요구를 표현할 때 주제개념 이외에 문헌의 출판시기에 관한 용어, 출판언어에 관한 용어, 검색문헌수를 지시하는 용어 등이 포함될 수 있으나 본연구에서는 출판시기에 관한 용어만을 포함하도록 하였다. 이 연구에서 구현한 시스템은 우리말 문헌을 처리하는 시스템이므로 언어는 당연히 한국어가 될 것이며, 자료의 형태는 논문에 국한하였으므로 이와 관련된 용어는 처리할 필요가 없었다. 그러나 데이타베이스의 규모가 커질 때에는 학위논문, 단행본, 논문 등으로 구분하여 원하는 자료만을 검색하도록 하는 것이 바람직할 것이다. 검색문헌수는 질문을 입력하기 전에 이용자가 선택한 탐색목표에 관한 응답이 검색문헌수에 대한 이용자의 요구수준을 대략 표현하기 때문에 이에 관련된 용어도 처리하지 않았다. 따라서 이 시스템이 허용하는 질문은 '한글문헌의 자동색인에 관한 최신 문헌'이나 '한글문헌의 자동색인에 관한 1985년 이후 출판된 논문' 등과 같이 주제개념을 나타내는 용어와 출판시기를 나타내는 용어로 구성된다.

이 시스템에서 질문을 처리하는 과정은 다음과 같다.

(1) 입력된 질문에서 '에 관한(대한) 문헌(논문/자료/글/기사/논고)', '에 관해(대해) 쓴

(쓰여진) 문헌', '을(를) 다룬(연구한) 문헌' 등의 패턴을 인식하여 제거한다. '최신문헌', '최근에 출판된 문헌', '1985년 이후 출판된 문헌', '1990년 출판된 문헌' 등과 같이 출판시기와 관련된 수식어가 포함된 경우에는 '최신'이나 연도 등 해당되는 용어를 인식하여 탐색어로 변환하도록 한다. '최신'은 탐색년을 기준으로 하여 지난 3년간으로 변환되며(예: 탐색년이 1991년인 경우 1991, 1990, 1989년이 해당됨), '-년 이후'는 지정한 연도에서 탐색년까지를 탐색어로 사용한다.

(2) 주제개념으로만 구성된 명사구를 파싱한다. 파싱과정은 색인모듈에서와 같으며, '명사+명사' 형태로 띄여서 입력된 복합어는 사전과 대조하여 붙여쓴다.(예: 한글 문헌 --> 한글문헌)

(3) 색인모듈에서 사용한 두가지 색인기법을 사용하여 격관계 그래프에 의한 탐색문과 키워드리스트로 구성된 탐색문을 생성한다. 예를 들어 '한글문헌의 자동색인에 관한 문헌'은 다음과 같이 두가지 탐색문을 생성한다.

(한글문헌 -- 대상격 --> 자동색인)

(한글문헌, 자동색인)

4.2 자동색인 모듈

4.2.1 색인대상 텍스트의 선정

자동색인에서는 일반적으로 색인어의 선정을 위해 표제, 초록, 문헌전문을 분석대상 텍스트로 이용하고 있다. 분석대상 텍스트로 무엇을 선택하는가는 문헌의 유형과 자동색인기법에 따라 결정된다. 학술논문을 색인함에 있어서 통계적 기법을 사용하는 경우에는 초록과 전문을 선택하는 것이 바람직하며, 불용어 제거기법만을 사용하는 경우에는 대개 표제와 초록을 대상으로 하여 색인어를 선정한다. 반면 신문기사나 판결문과 같이 전문의 길이가 짧고 전문 데이타베이스가 구축되는 경우에는 어떠한 색인기법을 사용하든지 분석대상 텍스트는 전문이 된다.

본 연구에서는「정보관리학회지」제1권-제7권에 수록된 논문 63편을 선택하여 실험대상 문헌집단을 구성하였다. 색인어 선정을 위해 분석할 텍스트를 표제와 초록으로 할 것인가, 표제만으로 할 것인가를 결정하기 위하여 시스템의 '효율'과 '효과'의 두 측면을 고려하였다. 시스템의 효율을 높이기 위해서는 분석대상 텍스트가 짧을수록 좋으며, 반면에 일반적으로 재현율과 정확률로 측정되는 시스템의 '효과'를 높이기 위해서는 표제뿐만 아니라 초록을 분석대상으로 삼는 것이 바람직하다. 따라서 표제만을 분석대상으로 함으로써 효율을 높이고 동시에 적절한 수준의 효과(검색효율)를 유지할 수 있는 방법을 모색하기 위하여 63편의 논문의 초록을 분석하여 보았다.

「정보관리학회지」의 논문초록은 저자가 작성한 초록이므로 초록의 성격에 있어서 지시적 초록과 통보적 초록이 혼합되어 있었다. 63편 가운데 지시적 초록이 45편, 통보적 초록이 18편이었으며, 초록의 길이는 컴퓨터에 의해 한번의 처리대상이 되는 토큰을 한 단어로 세었을 때 평균 56단어였다. 초록을 구성하는 단어 가운데 주제어의 수는 평균 5개 정도였으며, 표제의 주제어와 초록의 주제어와의 일치율은 논문에 따라 차이가 컸으나 대략 25%-30% 정도인 것으로 나타났다. 표제어와 일치하지 않는 초록의 주제어중에는 색인어로 부적합한 색인어가 상

당수 있었으며, 색인어로 적합한 주제어는 대체로 표제어와 어의적인 관계를 갖고 있었다. 이것은 잘 구성된 시소러스를 이용하여 표제어를 확장하면 초록에 포함된 색인어를 대신할 수 있으리라는 것을 의미한다. 예를 들어 '논문의 동시인용을 통한 지적 구조의 규명에 관한 연구'에서 초록으로부터 추출한 주제어는 '동시인용', '계량서지학적 기법', '지적 구조', '사회학 분야', '다차원 공간적 측면', '저자동시인용', '문헌동시인용'의 일곱 단어였는데, 이 가운데 '다차원 공간적 측면'은 색인어로는 부적합하였으며, 표제어와 일치하지 않는 나머지 네 단어 중 '사회학 분야'를 제외한 세 단어는 모두 다음과 같이 표제어와 계층적 관계를 갖고 있었다.

동시인용 NT 저자동시인용
문헌동시인용
BT 계량서지학적 기법

위와 같이 표본초록들을 분석한 결과를 참고하여 본 연구에서는 표제만을 분석대상 텍스트로 삼아 색인어를 선정하고, 초록에서 추출될 수 있는 색인어는 탐색시 시소러스를 이용하여 탐색문에 추가하는 방법을 채택하였다.

4.2.2 자동색인 기법

이상적인 색인은 정확한 색인어의 선정뿐만 아니라 색인개념간의 구문적인 관계와 어의적인 관계를 표현해 줄 수 있는 것이어야 한다. 컴퓨터를 이용한 자동색인시스템에서 색인개념간의 어의적 관계는 대개 시소러스나 이와 유사한 주제지식베이스에 정의되며 이 관계를 색인이나 탐색시 활용하고 있다. 구문적 관계는 색인어 선정시 텍스트 분석과정에서 표출되나, 대부분의 자동색인 시스템은 구문분석 결과 특정한 구문단위를 색인어로 선정하는데 그치며 색인어간의 관계는 표현하지 못하고 있다.

색인개념간의 구문적 관계를 표현해 준 초기의 컴퓨터시스템으로는 질문응답시스템으로 개발된 Protosynthex Ⅰ이 있다[33]. 이 시스템은 의존문법(dependency grammar)에 기초한 의존구조를 이용하여 문장을 구성하는 단어들의 구문관계를 그래프로 표현하였다. 한 문장을 의존구조에 의해 표현할 때 수식어가 되는 단어는 수식되는 단어 아래 오며 화살표에 의해 서로 연결된다. Protosynthex Ⅰ에서는 다음 예에서처럼 주동사가 주어에 의존하도록 정의되었다. 그림-2는 "Worms eat grass."와 "Horses with worms eat grain."을 각각 의존구조로 표현한 것이다.

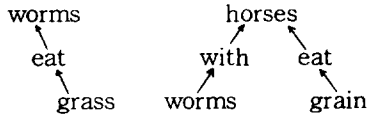


그림-2. Protosynthex Ⅰ에서의 의존구조

최근에는 독일에서 개발된 정보검색시스템에서 Protosynthex Ⅰ에서와 같은 의존구조를 이용하여 문헌텍스트를 구성하는 모든 구들을 색인한 것이 발표되었다[34]. 이와 같이 의존구조를 이용하여 색인하는 것은 색인개념간의 의존관계는 표현할 수 있으나 개념간에 구체적으로 어떠한 구문적 관계가 있는지는 표현할 수 없다. 국내에서는 장재경[35]이 격문법을 이용하여 색인개념간의 격관계를 설정함으로써 구문

적인 관계를 표현하였으나, 패러데인(Farradane)의 관계색인 등의 용어열색인에서와 마찬가지로 격관계를 사람이 부여하도록 하였고 자동색인시스템으로는 구현하지 못하였다.

본 연구에서는 색인개념간의 구문적 관계를 표현해 주어야만 문헌의 내용을 정확히 표현할 수 있다고 보고, 자연언어 처리결과 표제를 구성하는 주제어간의 구문적 관계를 격관계로 표현하여 색인표목을 구성하였다. 자동색인 모듈에서는 격관계 색인 이외에 키워드색인도 작성하였으며, 각 색인은 검색모듈에서 이용자의 탐색목표에 적합한 검색기법과 함께 사용되도록 하였다.

자동색인 모듈에서 수행되는 색인절차는 다음과 같다.

(1) 간단한 자유문맥문법(context-free grammar: CFG)을 사용하여 표제를 파싱한다. 파싱한 결과 각 단어에는 품사와 의미자질이 부여된다.

(2) 파싱한 결과를 격문법에 의해 분석하여 색인어를 선정하고 색인어간의 격관계를 부여한다. 이때 '..에 관한 연구', '..에 관한 고찰' 등 표제속에서 전형적으로 사용되는 부분은 미리 제거한다. 단 '사적 연구', '계량서지학적 연구' 등 연구방법론을 의미하는 용어가 포함된 경우에는 제거하지 않는다.

(3) 색인결과 {색인어--격관계-->색인어}를 기본단위로 한 격관계 그래프를 생성한다.

(4) 격관계 그래프에서 격관계를 제외한 색인어만을 추출한 뒤 단독으로는 주제어가 될 수 없는 색인어(예: 이론, 실제, 현황, 문제점 등)를 불용어사전에 의해 제거하여 키워드리스트를 생성한다.

다음은 '컴퓨터와 연관된 지적소유권보호책의 현황'이란 표제를 격관계색인법에 의해 색인한 결과 생산된 격관계 색인그래프이다.

(컴퓨터--관계격-->지적소유권보호책)--대상격-->현황

위의 격관계 그래프로부터 생산되는 키워드리스트는 '컴퓨터, 지적소유권보호책'이다.

4.2.3 표제의 형태소/구문분석

학술논문의 표제는 대체로 명사구 형식이며 일정한 구문적 패턴을 갖는다. 실제로 본 연구의 표본문헌집단의 표제는 모두 명사구였으므로 다음과 같이 CFG의 명사구 다시쓰기 규칙을 사용하여 표제를 파싱하였다. 파싱을 위하여 단어의 품사와 의미자질을 수록한 어휘사전을 사용하였다.

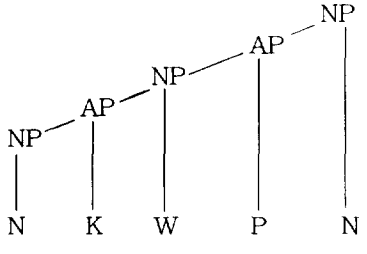
〈명사구〉::=〈관형어〉〈명사구〉
〈명사구〉::={명사}*
〈명사구〉::=〈관형구〉〈명사구〉
〈명사구〉::=〈관형구〉〈명사구1〉
〈명사구〉::=〈관형절〉〈명사구〉
〈명사구1〉::=〈명사구〉{〈접속조사〉〈명사구〉}*
〈관형절〉::=〈명사구〉〈격조사〉〈동사〉
〈관형구〉::=〈명사구〉〈격조사〉
〈관형구〉::=〈명사구〉〈후치사〉
〈관형구〉::=〈명사구1〉〈격조사〉
〈관형구〉::=〈명사구1〉〈후치사〉

위의 규칙 가운데 관형구는 명사구에 격조사나 후치사가 결합되어 구성되는 것으로 정의되

어 있다. 후치사라는 임의의 카테고리를 설정한 이유는 대부분의 표제에서 〈조사+관형어〉의 형식이 명사에 붙어 한 단위로 사용되며 격조사와 동일한 역할을 수행하기 때문이다. 후치사로 처리된 단어는 다음과 같다.

-에 관한 -에 따른 -에 대한
-을(를) 위한 -에 의한 -에 미치는
-와(과) 연관된 -에 기초한
-에 있어서 -에 있어서의 -을(를) 통한
-을(를) 중심으로 한

간단한 예로 '용어의 자동분류에 관한 연구'를 위의 다시쓰기 규칙에 의해 파싱한 결과를 나무구조로 표현하면 다음과 같다.(NP: 명사구, AP: 관형구, N: 명사, K: 격조사, P: 후치사)



용어 의 자동분류 에 관한 연구

표제의 파싱결과 각 단어에는 품사와 의미자질이 부여되며, 이것은 격문법에 의한 의미분석에 이용된다. 어휘사전에 수록된 의미자질은 올바른 격관계 설정을 위해 필요한 것으로서 해당되는 의미자질이 없는 경우에는 'null'로 표시하였다. 사용된 의미자질로는 기관, 장소, 사람, 시스템, 의존성 주제어 등이 있다.

4.2.4 격문법에 의한 의미분석

필모어(Fillmore)에 의해 발전된 격문법(case grammar)은[36] 구문분석과 의미분석을 통합시키고자 한 초기시도 가운데 하나로 인식되고 있다. 필모어에 의해 설정된 원래의 격범주는 행위자격(agentive), 도구격(instrumental), 여격(dative), 결과격(factitive), 처소격(locative), 대상격(objective)의 여섯개였으며, 이후 국내외적으로 다양한 격분류가 시도되었다. 현재로는 격범주의 설정에 있어서의 통일이 이루어지지 않은 상태이며, 그 원인은 기본적으로는 격에 대한 정의의 차이에서 찾을 수 있으나 실질적으로는 자연언어처리시스템의 응용분야와 주제영역에 따라 다른 격분류가 사용되고 있는 것으로 보인다.

국어문법에서 격에 대한 정의는 크게 세가지 유형으로 나뉜다[37]. 첫째 유형은 문장성분의 종류와는 상관없이 체언이 문장속의 다른 말에 대하여 가지는 일정한 자격이 격이라고 보는 견해다. 이 정의에서는 조사가 부착된 체언이 관계를 맺는 말은 다른 체언일 수도 있고 용언일 수도 있으며, 두 말 사이의 관계는 통사론적인 것이 아니라 의미적인 것이다. 둘째 유형은 문장성분으로서의 자격을 격이라고 보는 견해이다. 이때의 격은 서술어에 이끌리는 관계에 있는 것만을 일컫는 것은 아니다. 셋째 유형은 체언이 일정한 문장성분으로서 문장안에서 차지하는 자리를 격으로 보되, 서술어에 의해 이끌리는 관계에 있는 것이어야 한다는 것이다. 이것이 필모어의 격문법에서 말하는 격에 해당되는 것이며, 이때의 격은 서술어에 이끌리는 관계에 있는 문장성분의 기능이 되며, 격조사는 체언으로 하여금 이러한 기능을 가지게 하는 말이라고 정의된다.

격문법의 기본구조는 다음과 같다.

$S \dashrightarrow P+M$

$P \dashrightarrow V+C_1+\cdots+C_n$

$C_i \dashrightarrow NP+K$

S:문장 P:명제 M:양태
V:술어 C:격
NP:명사구 K:조사

위의 구조에 의하면 모든 문장은 명제와 양태로 이루어지며, 명제는 하나의 술어와 여러개의 격으로 구성된다. 격은 명사구와 조사에 의해 표현되며 명사구와 술어와의 기본관계를 나타낸다. 본연구에서는 분석대상이 되는 표제가 대체로 술어가 포함되지 않는 명사구의 형태이므로 위의 기본구조를 그대로 적용하지 않고 격에 관한 규칙만을 적용하였다. 이 규칙에서 C_i는 명사(구)와 조사 또는 후치사로 구성되는 관형구를 의미한다. 또한 격범주가 명사구와 술어와의 관계인 반면 격관계는 표제를 구성하는 개념간의 구문적 관계를 나타내는 것이므로, 격관계는 조사나 후치사 전후에 오는 명사(구)간의 관계를 파악함으로써 결정되도록 하였다. 이 방법은 앞에서 기술한 격에 관한 세가지 정의 가운데 첫번째 정의를 응용한 것이다. 예를 들어 '용어의 자동분류에 관한 연구'에서 '용어'와 '자동분류' 사이에는 '의'라는 격조사가 있으며, 이때의 '의'는 격관계 부여규칙에 의해 대상격관계를 나타내는 조사로 식별된다.

본연구에서 설정한 격관계와 대응되는 조사 및 후치사는 다음과 같다.

관계격 : -와/과 연관된
관형격 : -의
도구격 : -에 의한, -을/를 활용한, -을/를 이용한, -을/를 통한
대상격 : -에 관한, -에 대한, -의, -에서의, -에 있어서(의)
처소격 : -에서의, -에 있어서(의), -의, -내에서의
행위자격 : -의, -이/가
여격 : -에
도달격 : -을/를 위한, -을/를 전제로 한.
비교격 : -와/과(의), -와/과 -의
원천격 : -에 기초한, -을/를 중심으로 한
원인격 : -으로/로 인한
접속격 : -와/과, -이(나), 및
재료격 : -을/를 사용한, -으(로) 된, 격조사 없이 명사만 나열된 경우
동격 : -의, 격조사 없이 동격인 명사가 나열된 경우

위에서 보는 바와 같이 '의', '-에서의', '-에 있어서(의)'는 여러개의 다른 격관계를 나타낼 수 있으므로 올바른 격관계를 선택하기 위한 규칙이 필요하다. 이외에도 접속격조사인 '와/과'의 처리를 위한 규칙과 특정한 규칙이 나란히 나타날 때 서로 직접적인 관계를 갖는 개념끼리 묶는 규칙 등 올바른 격관계를 부여하기 위한 다수의 의미해석 규칙이 요구된다. 격관계 색인그래프를 생성하는 데 있어서의 기본원칙은 첫째, 격관계가 명사구 전체에 걸리는 경우에는 해당되는 명사구전체를 묶어주는 것이며, 둘째, '와/과', '및' 등 접속격조사가 '의'와 함께 나타나는 경우에는 접속격조사를 우선적으로 처리하는 것이다. 격관계부여를 위해 사용한

조사 및 후치사의 격관계 테이블과 관련된 규칙들은 부록-1에 수록하였다.

다음은 격관계 부여가 단순하지 않은 여러 경우와 각 경우의 해결방안을 기술한 것이다.

(1) '의'는 행위자격, 대상격, 관형격, 처소격, 동격 관계를 나타낸다. 올바른 격관계를 부여하기 위해서는 문맥정보와 어휘사전의 의미자질을 이용한다. 각 격관계를 결정하는 규칙은 다음과 같다.

행위자격 : '의'앞의 명사(구)가 갖는 의미자질이 사람이나 기관이고 다음에 오는 명사에 접사 '-하다'를 붙이면 동사가 되는 경우

예 : CAB의 기능과 활동에 관한 연구
물리학자들의 학술정보 이용과 전달에 관한 연구

대상격 : a. '의' 앞의 명사의 의미자질이 사람, 기관, 장소가 아니고 다음의 명사가 접사 '-하다'가 붙어 동사가 될 수 있는 경우

예 : 데이타베이스의 이용이 참고업무에 미치는 영향에 관한 연구

b. '의' 다음의 명사가 단독적으로는 주제어가 될 수 없는 것으로 의존성 주제어란 의미자질을 갖는 경우(예 : 이론, 실태, 전망, 개념, 회고 등)

예 : 연관색인법의 이론과 실제

처소격 : '의'앞의 명사의 의미자질이 장소나 기관인 경우

동격 : '의' 전후에 오는 명사의 자질이 같은 경우(조사가 생략된 경우 '의'로 간주한다.)

예 : ETLARS(의) 한글정보검색시스템 개발에 관한 연구

관형격 : 위의 네가지 경우가 아니면 관형격으로 처리한다.

예 : 우리나라의 학술잡지의 발달과정에 관한 연구

(2) '-에 있어서(의)', '-에서의' 앞에 오는 명사의 의미자질이 장소나 기관인 경우에는 처소격을 부여하고 그외에는 대상격을 부여한다.

예 : 중소기업에서의 기술정보유통에 관한 연구(처소격)
연구활동에 있어서의 비공식 커뮤니케이션(대상격)

(3) 하나의 복합명사가 아닌 것으로 명사 사이에 조사가 없는 것은 사이에 '의'를 삽입하여 처리한다. '의'의 격부여는 (1)의 규칙에 따른다.

예 : 전문도서관(의) 수서업무(의) 전산화에 관한 연구

(4) '-의, -과'와 같이 접속격조사 '와(과)'에 조사 '의'가 선행하는 경우, '와(과)' 전후에 오는 명사가 같은 의미자질을 갖는 경우 접속격조사 전후의 명사는 '의' 앞의 명사와 동일한 격관계를 갖는다.

예 : 연관색인법의 이론과 실제

(5) '-과, -의'와 같이 접속격조사 '와/과' 다음에 조사 '의'가 오는 경우, '와/과' 전후에 오는 명사가 같은 의미자질을 갖는 경우 '의' 다음의 명사와 동일한 격관계를 갖는다.

예 : KWIC색인과 Descriptor색인의 검색효율성:
KWIC색인--관형격-->검색효율성<--관형격--Descriptor색인

(6) <도구격+대상격>의 형식으로 분석되는

경우 도구격관계는 바로 다음에 오는 명사와의 관계가 아니라 대상격관계 다음에 오는 명사('-하다'가 붙을 수 있는 명사)와의 관계가 되도록 연결한다.

예 : 이용자피이드백에 의한 검색질문의 자동수정:

이용자피이드백--도구격-->자동수정<--대상격--검색질문

(7) 〈도달격+대상격〉의 형식으로 분석되는 경우, 도달격관계를 나타내는 후치사 '-을 위한' 앞까지의 명사구 전체가 후치사 다음에 오는 명사와 도달격관계를 갖도록 연결한다.

예 : 우리말신문기사 검색을 위한 질문응답시스템 구현에 관한 연구:

(우리말신문기사-대상격->검색)-도달격->질문응답시스템)-대상격->구현

(8) '-이, -에 미치는 영향', '-에 있어서, -의 역할', '-과, -의 비교분석(비교연구)' 등과 같이 표제속에 자주 나타나는 구문형식은 도식대조방식에 의해 일정한 구문관계로 분석한다.

a. 데이타베이스 이용이 참고업무에 미치는 영향:

(데이타베이스-대상격->이용)-영향격->참고업무

b. 정보검색을 위한 인버티드화일과 클러스터화일의 비교분석:

정보검색-도달격->(인버티드화일-비교격->클러스터화일)

c. 연구활동과 과학지식생산성에 있어서 학술연구전산망의 역할:

연구활동-역할격->학술연구전산망<-역할격-과학지식생산성

4.3 검색/ 탐색확장 모듈

지능형 정보검색에서는 이용자의 수준이나 탐색목표에 따라 가장 적합한 검색기법을 선택하거나 탐색전략을 수립할 수 있어야 하며, 또한 검색된 정보는 이용자의 정보요구에 적합한 것이어야 한다. 질문과 문헌간의 유사성을 평가하는 방법으로서 지금까지 다양한 검색기법이 개발되어 있다. 검색기법은 기본적으로 완전일치기법과 부분일치기법으로 분류되며, 완전일치기법인 불리안 논리검색이 대부분의 대규모 정보검색시스템에서 채택되고 있다. 부분일치기법으로는 가중치에 의한 검색, 확률검색, 매칭함수에 의한 검색, 클러스터화일 검색, 퍼지집합 검색 등이 있다. 불리안 논리검색이 대부분의 상용시스템에서 사용되고 있지만, 이 기법의 문제점으로 첫째, 부분적으로 일치하는 문헌은 검색할 수 없으며 검색된 문헌에 적합성에 따른 순위를 부여할 수 없다는 것과, 둘째, 탐색개념간의 관계를 표현할 수 없다는 점 등이 지적되고 있다. 첫번째 문제점을 해결할 수 있는 기법으로 퍼지집합 검색[38]과 확장된 불리안 논리검색[39]이 논의되고 있으나 아직은 실험적인 단계에 머물고 있다. 두번째 문제점을 해결하기 위해서는 개념간의 구문적 관계를 표현할 수 있는 자동색인기법의 개발과 탐색문 생성에의 적용이 요구된다.

최근에는 텍스트를 논리적이거나 그래프 등의 형식으로 표현하고 추론에 의해 적합정보를 검색하는 검색기법이 제시되고 있다[20, 34, 40]. 특히 1986년 리즈버겐(van Rijsbergen)이 제시한 비전통적인 논리에 기초한 검색이론은 이 분야 연구자들에게 큰 영향을 미치고 있다[41]. 리즈버겐은 통계적 검색기법들은 이제 이

론적 한계에 도달했다고 지적하고, 차세대 정보검색시스템을 위해서는 새로운 검색이론이 필요함을 강조하였다.

본 연구에서는 이용자 모형을 통해 높은 재현율을 원하는 이용자에게는 부분일치기법의 하나인 매칭함수에 의한 검색을 선택하고, 높은 정확률을 원하는 이용자에게는 격관계 색인그래프와 대조하는 완전일치기법을 선택하도록 하였다. 탐색시 주제지식베이스인 시소러스를 이용하여 유사어 및 아랫개념어를 탐색어로 추가함으로써 자동적으로 탐색확장이 되도록 하였다.

4.3.1 매칭함수에 의한 검색

매칭함수에 의한 검색은 이용자가 높은 재현율을 원하는 경우에 선택되며 검색기준치는 탐색확장시 이용자요구에 따라 조절될 수 있다. 본연구에서는 이용자 모형에 '높은 재현율'과 '높은 정확률'의 두가지 정보만을 포함하였기 때문에 검색기준치를 일단 0.6으로 정하였다. 매칭함수에 의한 검색은 질문과 문헌의 유사도, 구체적으로는 문헌에 부여된 색인어와 질문을 구성하는 탐색어의 일치도를 유사계수공식에 의해 산출하여, 유사계수값이 검색기준치를 넘는 문헌을 검색하는 것이다. 본연구에서는 다음과 같은 다이스계수(Dice's coefficient) 공식을 사용하였다.

$$\text{유사계수} = \frac{2\,|D \cap Q|}{|D| + |Q|}$$

D:문헌에 부여된 색인어 집합
Q:질문에 부여된 탐색어 집합

매칭함수에 의한 검색을 수행하게 되면 검색모듈은 유사계수가 0.6 이상인 문헌을 검색하여 유사계수 값의 순으로 출력한다. 탐색시 질문속에 포함된 탐색개념을 자동으로 확장하도록 하였는데, 탐색어와 동의어 관계에 있는 용어와 계층적 관계(아랫개념어)에 있는 용어를 탐색어로 추가하여 새로운 탐색어리스트를 생성한다. 탐색문의 자동확장시 윗개념어를 추가하지 않은 이유는 윗개념어를 탐색어로 사용하는 경우 질문의 내용보다 포괄적인 탐색이 수행됨으로써 지나치게 광범위한 주제의 문헌이 검색될 수가 있기 때문이다. 그러나 검색되는 문헌의 수가 적을 때는 이용자피드백에 의해 윗개념어를 추가하는 탐색확장을 수행하도록 하였다.

예를 들어 질문이 '우리말문헌의 자동색인에 관한 문헌'인 경우 생성되는 탐색어리스트는 〈우리말문헌, 자동색인〉이며, 지식베이스를 이용한 결과 〈한글문헌, 자동색인〉, 〈한국어문헌, 자동색인〉, 〈한글, 자동색인〉, 〈우리말, 자동색인〉, 〈한국어, 자동색인〉 등이 추가로 생성된다. 이러한 탐색어리스트에 의해 검색되는 문헌은 검색기준치를 0.6으로 했을 때 다음과 같다.

D1 : 통계적 기법에 의한 한글 자동색인의 연구 (유사계수=0.8)
D2 : 한글문헌의 자동색인에 관한 실험적 연구 (유사계수=0.8)
D3 : 언어학적 기법에 의한 한글문헌 자동색인 (유사계수=0.8)

같은 주제의 질문인 '우리말로 된 문헌의 자동색인에 관한 문헌'을 처리할 때는 '우리말'과 '문헌'의 관계를 재료격관계로 규정한 뒤, 두 단

어를 붙여쓴 형태가 어휘사전에 있는가 확인하여 나와있는 경우에는 하나의 복합어로 처리하게 된다.

이용자가 탐색결과에 만족하지 않을 때는 검색기준치를 높임으로써 정확률을 향상시키는 방법과 윗개념어를 추가하거나 검색기준치를 낮춤으로써 재현율을 향상시키는 방법을 도입할 수 있다. 예를 들면 질문이 '전문가시스템에서의 생성규칙의 이용에 관한 문헌'인 경우 본 연구의 실험용 데이타베이스로부터 검색되는 문헌은 없다. 그러나 탐색확장을 원하는 경우 지식베이스로부터 '전문가시스템'의 윗개념어인 '지식기반시스템'이 추가되고, '생성규칙'의 윗개념어인 '지식표현'이 추가됨으로써 데이타베이스에 소장되어 있는 '문헌정보학영역 지식기반시스템에서의 지식표현'이란 표제의 문헌이 검색된다. 격관계 그래프에 의한 검색에서도 이용자피드백에 의해 탐색확장을 수행하는 경우 같은 검색결과를 얻게 된다.

4.3.2 격관계 그래프에 의한 검색

이용자가 높은 정확률을 원하는 경우에는 격관계 그래프에 의한 검색을 수행한다. 입력된 질문은 인터페이스 모듈에서 분석되어 격관계 탐색그래프를 생성하며, 탐색그래프와 일치하거나 탐색그래프를 하부그래프로 하는 색인그래프가 있는 경우 해당되는 문헌을 검색한다. 앞에서 예로 든 질문과 검색문헌(D3)의 격관계 그래프는 아래와 같으며, 이때 탐색그래프는 색인그래프의 하부그래프가 되어 이 문헌은 검색되는 것이다.

색인그래프 : 언어학적 기법--도구격-->자동색인<--대상격--한글문헌

탐색그래프 : 한글문헌--대상격-->자동색인

일차로 생성된 격관계 탐색그래프를 구성하는 탐색개념들은 주제지식베이스에 의해 확장되어 여러개의 탐색그래프를 추가로 생성한다. 격관계 그래프 검색에서는 높은 정확률이 탐색의 목표이므로 일차 탐색확장시에는 동의어 관계만을 이용하며, 검색된 문헌이 없는 경우 자동으로 아랫개념어와 윗개념어를 추가하여 탐색확장을 하도록 한다. 일차 검색결과 검색된 문헌이 있는 경우에도 이용자가 더많은 문헌을 원할 때에는 아랫개념어와 윗개념어를 추가하여 이차적인 탐색그래프를 생성한다.

격관계 그래프를 이용한 검색은 개념간의 구문적 관계를 나타냄으로써 질문과 문헌의 내용을 정확히 표현하고, 따라서 질문에 적합한 문헌만을 검색할 수 있는 장점이 있는 반면, 같은 내용을 표현하는 데 전혀 다른 구문구조를 사용한다든가 단독으로는 주제어가 될 수 없는 용어가 그래프에 포함되는 경우에는 검색시 문제가 발생하게 된다. 예를 들어 '한글문헌의 자동색인에서의 언어학적 기법에 관한 연구'는 '언어학적 기법에 의한 한글문헌의 자동색인'과는 구조가 다른 격관계 그래프를 생성하게 된다.

(한글문헌--대상격-->자동색인)--대상격-->언어학적 기법

언어학적 기법--도구격-->자동색인<--대상격--한글문헌

결과적으로 두 문헌은 '한글문헌의 자동색인에 관한 문헌'이라는 질문에 대해서는 모두 적

합문헌으로 검색되지만, 질문이 '언어학적 기법에 의한 자동색인에 관한 문헌'인 경우에는 두번째 문헌만이 검색된다. 이와 같은 문제를 해결하기 위해서는 용어의 의미자질을 이용한 의미분석을 더욱 강조할 필요가 있으며, 위의 경우 '언어학적 기법'은 방법론이라는 의미자질을 갖게 하여 '-하다'가 붙을 수 있는 명사(핵심개념어)와는 항상 도구격관계를 갖는 것으로 분석하는 규칙을 이용함으로써 같은 격관계를 부여할 수 있다.

위에서 언급한 또하나의 문제점으로 인해 '전문가시스템에서의 생성규칙의 이용에 관한 문헌'과 같은 질문의 처리시 '이용'이 탐색그래프에 포함됨으로써 '전문가시스템에서의 생성규칙'이나 탐색확장결과 검색대상이 되는 '지식기반시스템에서의 지식표현' 등 '이용'이 포함되어 있지 않은 표제는 적합문헌이면서도 검색되지 않는다. 이러한 문제는 탐색그래프를 구성할 때 '이용'과 같은 '의존성 주제어'를 제거함으로써 해결할 수 있다. 결과적으로 중요한 탐색개념을 포함하는 격관계 색인그래프는 모두 검색하게 된다.

격관계 그래프에 의한 검색기법을 도입한 목적은 불리안 논리검색이 가져오는 정확률의 저하문제를 해결하는 데 있다. 그러나 본연구의 실험문헌집단은 표제를 구성하는 개념의 수가 적고 또한 개념간의 관계가 다양하지 않으므로 인해서 격관계 그래프에 의한 검색결과는 불리안 논리검색 결과와 큰 차이가 없었다. 다음과 같은 경우는 동일한 색인개념을 포함하나 구문관계는 다른 표제를 격관계 색인법에 의해 색인함으로써 검색결과에 차이를 가져올 수 있는 예이다. 즉, 불리안 논리검색에서는 '데이타베이스 이용이 참고업무에 미치는 영향에 관한 문헌'이라는 질문에 대해 아래의 두 문헌이 모두 검색되는 반면 격관계 그래프에 의한 검색에서는 첫번째 문헌만이 검색된다.

D4 : 데이타베이스 이용이 참고업무에 미치는 영향:
(데이타베이스--대상격-->이용)--영향격-->참고업무

D5 : 참고업무가 데이타베이스 이용에 미치는 영향:
참고업무--영향격-->(데이타베이스--대상격-->이용)

구문적 관계가 결정적인 역할을 하는 경우는 특히 두개 이상의 복합주제가 접속격조사에 의해 연결된 경우가 될 것이다. 격관계 그래프에 의한 검색에서는 개념간의 구문적 관계가 표현됨으로써 불리안 논리검색에서 개념의 잘못조합으로 인해 발생하는 정확률의 저하를 방지할 수 있다. 예를 들어 '최근 미국의 정보전문가 교육의 동향과 한국 사서교육과정 개정의 기본방향'이란 표제의 문헌은 키워드색인을 통해 불리안 논리검색을 하는 경우에 '한국의 정보전문가 교육의 동향'이란 질문에 대해서도 검색될 것이다.

5. 결 론

현재 운용되고 있는 대부분의 정보검색시스템은 다음과 같은 문제점을 안고 있다.

(1) 이용자가 자신의 정보요구를 자연언어 형태로 표현하지 못하고 정형의 탐색문을 작성

하여 입력하여야 한다.

(2) 정형의 탐색문은 탐색어간의 구문적이고 어의적인 관계를 반영하지 못하고 있다.

(3) 이용자의 특성이나 탐색에 관련된 요구를 수용하지 못하고 있다.

(4) 문헌의 내용을 정확히 표현할 수 있는 효과적이며 효율적인 자동색인기법이 개발되어 있지 않다.

(5) 현재 사용되고 있는 검색기법들은 통계적인 기법으로서 검색효율을 현재수준 이상으로 끌어올리는 데 한계가 있다.

(6) 시스템과 이용자간에 충분한 인터페이스가 제공되지 못하고 있다.

이러한 문제점을 해결하기 위해 1980년대에 들어서부터 온라인 탐색중개 전문가시스템과 실험적인 지능형 정보검색시스템들이 개발되고 있으나 아직 이상적인 정보검색시스템의 모형이 제시되지 못한 상태이다. 본연구에서는 현재의 정보검색시스템이 안고 있는 문제점을 해결하기 위하여는 무엇보다도 자연언어 처리와 다양한 지식베이스의 활용이 가능한 지능형 정보검색시스템이 필요하다고 보고, 이러한 지능형 정보검색시스템의 전형적인 모형을 제시하였다. 또한 이 기본적인 모형을 이용하여 우리말 문헌을 처리하는 시스템을 부분적으로 구현하였으며, 특히 질문의 분석과 자동색인을 위한 자연언어 처리과정을 집중적으로 연구하였다. 본연구를 통해 개발한 정보검색시스템의 특성과 제한점은 다음과 같다.

(1) 이용자에 관한 단기적인 지식을 이용하여 인터페이스 방식과 탐색전략을 선택하도록 하였다.

(2) 시스템의 효율을 고려하여 표제만을 분석대상으로 하여 색인하였으며, 색인개념간의 구문적 관계를 격관계로 표현하였다.

(3) 제한된 형식의 자연언어 질문을 처리하였으며, 탐색개념간의 구문적 관계를 격관계로 표현하였다.

(4) 매칭함수에 의한 검색과 격관계 그래프에 의한 검색이 가능하며, 이용자의 요구를 고려하여 적절한 검색기법을 선택하도록 하였다.

(5) 검색모듈과 탐색확장 모듈을 함께 구현함으로써 주제지식베이스를 이용한 탐색확장이 자동적으로 수행되도록 하였다.

(6) 검색결과 이용자피드백에 의한 탐색확장을 수행하도록 하였으나 적합성 판정에 의한 탐색확장은 구현하지 않았다.

(7) 자연언어 처리시 구문분석과 의미분석을 병행하였으나 처리가능한 구문형식이 명사구에 국한되었다.

앞으로 정보검색분야의 연구과제는 자동색인과 검색기법에 관한 지속적인 연구와 함께, 주제영역에 제한받지 않는 이식성이 높은 자연언어처리기의 개발, 실험용이 아닌 완벽한 어휘사전과 분야별 시소러스 등 지식베이스의 구축, 텍스트 이외에 그림, 소리, 그래픽 등의 정보를 처리하는 다중매체시스템과 정보를 비연속적으로 조직하는 하이퍼텍스트시스템으로서의 기능을 갖는 미래지향적인 정보검색시스템의 개발 등이 되어야 할 것이다.

참 고 문 헌

[1] G.L. Horowitz and H.L. Bleich, "Paperchase: a computer program to search the medical literature," New England J.

of Medicine, 305(16): 924-930, 1981.

[2] M.I. Crystal and G.E. Jacobson, "FRED, a front end for databases," Online, 27-30, 1982.

[3] D.E. Taliver, "OL'SAM: an intelligent front-end for bibliographic information retrieval," ITL, 1(4): 317-326, 1982.

[4] C.T. Meadow, et al., "A computer intermediary for interactive data base searching. 1. Design," JASIS, 33(5): 325-332, 1982.

[5] "Information Marketplace," Bulletin of the ASIS, 17(3): 27, 1991.

[6] M.L. Neufeld and M. Cornog, "Database history: from dinosaurs to compact discs," JASIS, 37(4): 183-190, 1986.

[7] H.M. Brooks, "Expert systems and intelligent information retrieval," IPM, 23(4): 367-382, 1987.

[8] K. Sparck Jones, "Intelligent retrieval," in K.P. Jones, ed., Intelligent Information Retrieval: Informatics 7. London: ALSIB, 1983.

[9] W.B. Croft, "Approaches to intelligent information retrieval," IPM, 23(4): 249-254, 1987.

[10] W.B. Croft and R.H. Thompson, "I^{3}R: a new approach to the design of document retrieval systems," JASIS, 38(6): 389-404, 1987.

[11] Y. Chiaramella and B. Defude, "A prototype of an intelligent system for information retrieval: IOTA," IPM, 23 (4): 285-303, 1987.

[12] E.A. Fox, "Development of the CODER system: a testbed for artificial intelligence methods in information retrieval," IPM, 23(4): 341-366, 1987.

[13] M. Lebowitz, "An experiment in intelligent information systems: RESEARCHER," in R. Davies, ed., Intelligent Information Systems: Progress and Prospects. Chichester: Ellis Horwood, 1986. pp.127-149.

[14] G. Biswas, et al., "Knowledge-assisted document retrieval: 1. The natural-language interface: 2. The retrieval process," JASIS, 38(2): 83-96: 97-110, 1987.

[15] G. Brajnik, et al., "User modeling in intelligent information retrieval," IPM, 23(4): 305-320, 1987.

[16] D. Sleeman, "User modelling panel," Proceedings of the 9th IJCAI, 1985, 1298-1302.

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

[18] E. Rich, "Users are individuals: individualizing user models," in R. Davies, ed., Intelligent Information Systems: Progress and Prospects. Chichester: Ellis Horwood, 1986, pp. 184-201.

[19] R.N. Oddy, "Information retrieval through man-machine dialogue," J. of

Doc., 33: 1－14, 1977.

[20] N.J. Belkin, et al., "ASK for information retrieval: 1. Background and theory," J. of Doc., 38: 61－71, 1982.

[21] A. Vickery, et al., "An expert system for referral: the PLEXUS project," in R. Davies, ed., Intelligent Information Systems: Progress and Prospects. Chichester: Ellis Horwood, 1986. pp. 154－183.

[22] T. Winograd, Language As a Cognitive Process. Reading, Massachusetts: Addison-Wesley, 1983.

[23] S.M. Humphrey, "MedIndEx system: medical indexing expert system," IPM, 25(1): 73－88, 1989.

[24] 정영미, "우리말 신문기사 검색을 위한 질문응답시스템 구현에 관한 연구," 정보관리학회지, 4(1): 3－23, 1987.

[25] 신성우, 주식투자영역에서의 한글 자연어 처리시스템의 설계 및 구현. 연세대학교 석사학위논문, 1990.

[26] 윤덕호, 한국어 질의응답시스템의 설계 및 구현에 관한 연구. 서울대학교 석사학위논문, 1987.

[27] 김성기, 자연 한글 질의어처리를 위한 인터페이스의 설계 및 구현. 서울대학교 석사학위논문, 1985.

[28] 박혁로 외, "한글문서를 위한 자동색인시스템 개발," 우리말 정보화잔치 '91 논문집, 1991.

[29] 최원태, 격문법을 이용한 자동색인 및 탐색확장에 관한 연구. 연세대학교 석사학위논문, 1986.

[30] 장순걸, 언어학적 기법에 의한 한글문헌 자동색인. 경북대학교 석사학위논문, 1987.

[31] 안현수, 한글문헌의 자동색인에 관한 실험적 연구. 연세대학교 석사학위논문, 1986.

[32] 이영주, "자동색인을 위한 한국어 형태소 분석 알고리즘," 1989년도 한글날 기념 학술대회 발표논문집, 1989.

[33] R.F. Simmons, et al., "Indexing and dependency logic for answering English questions," Amer. Doc., 15: 196－204, 1964.

[34] G. Ruge, et al., "Effectiveness and efficiency in natural language processing for large amounts of text," JASIS, 42 (6): 450－456, 1991.

[35] 장재경, 우리말 문헌정보 검색을 위한 지식베이스 설계에 관한 연구. 연세대학교 석사학위논문, 1986.

[36] C. Fillmore, "The case for case," in E. Bach and R.T. Harms, ed., Universals in Linguistic Theory. Chicago: Holt, Rinehart and Winston, 1968.

[37] 남기심, "국어 문법에서 격(자리)은 어떻게 정의되어 왔는가?" 애산학보 5: 57－71, 1987.

[38] A. Bookstein, "Probability and fuzzy-set applications to information retrieval," ARIST 20: 117－151, 1985.

[39] G. Salton, et al., "Extended Boolean information retrieval," Comm. of the ACM, 26(11): 1022－1036, 1983.

[40] R.F. Simmons, "A text knowledge base from the AI handbook," IPM 23(4): 255-266, 1987.

[41] C.J. van Rijsbergen, "A new theoretical framework for information retrieval," Proc. of the ACM Conference on Research and Development in Information Retrieval, 1986.

부록-1. 격관계테이블과 관련규칙

주1) '과', '와', '및'이 '의'와 함께 한 명사구에 나타나는 경우, '과', '와', '및'을 우선적으로 처리한다.

주2) 현재의 조사를 'k1'으로 두고 뒤에 나오는 조사들은 순차적으로 번호를 매겨 처리한다. 현재의 조사 앞에 있는 명사는 항상 'np1'이고, 현재의 조사 뒤에 있는 명사는 'np2'이며, 그 뒤에 나오는 명사들은 순차적으로 번호를 매겨 처리한다. ㉮는 의미자질을 나타낸다.

조사(후치사)	규 칙	가능한 격관계
의	규칙1	행위자격관계 대상격관계 관형격관계 처소격관계 동격관계
에 있어서의 에 있어서 에서의	규칙2	대상격관계 처소격관계
이(가)	규칙3	행위자격관계 영향격관계
에 의한 에 따른	규칙4	도구격관계
을(를) 위한	규칙5	도달격관계
와 연관된	규칙6	관계격관계
와(과) 및	규칙7	비교격관계 접속격관계
을(를) 중심으로 한 을(를) 기초한	규칙8	원천격
를(을) 이용한	규칙9	도구격
에 관한 에 대한	*	대상격
에 미치는	*	미리주어졌음

〈규칙1〉

If (np1=n and np1 ㉮ "‥"=np2)

then「np1--동격관계--np2」처리한 다음 이를 괄호로 묶어 준다.

else (np1=n and np1 ㉮=사람) and (k2='의') and (np3=nv)

then「np1--행위자격관계--np3」

else (np1=n and np1 ㉮=사람) and (np2=nv and np2 ㉮== 자동)

then「np1--행위자격관계--np2」

else {(np1=n and np1 ㉮ 〈〉 (사람, 기관, 장소)) and (np2=nv)} or

{(np1=n) and (np2=n and np2 ㉮ 의존)
then「np1--대상격관계--np2」
else (np1=nv) and (np2=nv)
then np1까지의 모두를 괄호로 묶어 준 다음,
「괄호로 묶은 전체--대상격관계--np2」
else (np1=n and np1 ㉮=(장소, 기관)) and (np2=n)
then「np1--처소격관계--np2」
else
then「np1--관형격관계--np2」처리한 다음 이를 괄호로 묶어 준다.

〈규칙2〉

If (np1=n and np1 ㉮=(기관, 장소)) and (np2=n)
then「np1--처소격관계--np2」
else
then「np1--대상격관계--np2」

〈규칙3〉

If (k2='에 미치는') and (np3='영향')
then「np1까지의 문자열 묶은 것--영향격관계--np2」
else
then「np1--행위자격관계--np2」

〈규칙4〉

If (k2='의') and (np3=nv)
then「np1--도구격관계--np3」

else
then「np1--도구격관계--np2」

〈규칙5〉

항상「np1까지의 문자열 묶은 것--도달격관계--np2」

〈규칙6〉

항상「np1--관계격관계--np2」로 처리하되 이를 괄호로 묶어 준다.

〈규칙7〉

If (k2='의') and (np3=('비교', '비교관계', '비교연구'))

then「np1--비교격관계--np2」

else

then np1과 np2의 ㉮를 비교하여 같은 값이면 이를 분리하여 각각에 대한 명사구 내 다른 단위들과의 관계를 구해낸 뒤, 구해 진 각각의 관계 전체 사이에 접속격 관계를 준다.

else

then k1 앞에 있는 모두와 k2 뒤에 있는 모두를 각기 괄호로 묶어 이들 괄호 사이에 접속격을 준다.

〈규칙8〉

항상「np1--원천격관계--np2」로 처리하되 이를 괄호로 묶어 준다.

〈규칙9〉

명사구 내의 마지막 명사가 동사화명사('-하다'가 붙을 수 있는 명사)인지를 확인한 뒤

If 마지막명사=동사화명사

then「np1--도구격관계--동사화명사 바로 앞 명사」

else「np1--도구격관계--np2」