자연어를 이용한 자동정보검색시스템 구축에 관한 연구

A Study of Designing the Automatic Information Retrieval System based on Natural Language

서 휘(Whee Seo)*

목 차

초 록

본 연구에서는 자연어를 이용하여 자동으로 정보검색을 수행하는 시스템을 구축하였다. 구현 시스템은 Delphi 4.0(PASCAL)으로 프로그래밍 하였으며, 자동색인, 클러스터링 기법, 자연어 계층관계의 구축과 표현, 자동정보탐색이 가능하도록 구성했다. 이 시스템을 이용하여 질의어의 표현, 생성, 확장, 탐색식의 구성, 피드백 탐색 등 정보탐색의 전과정을 자동으로 수행할 수 있었다.

ABSTRACTS

This study is to develop a new system for conducting the information retrieval automatically. The system in this study is programmed by Delphi 4.0(PASCAL) and consists of automatic indexing, clustering technique, establishing and expressing term hierarchic relation, and automatic information retrieval technique. Thus this browser system can automatically control all the processes of information searching such as representation, generation and extension of queries and construction of searching strategy and feedback searching.

키워드 : 자동정보검색, 클러스터링 알고리즘, 시소러스, 용어계층 구축 알고리즘
automatic information retrieval, clustering algorithm, thesaurus, term hierarchy construction algorithm

---

* 창원전문대학 문헌정보과 조교수(drs733m@changwon-c.ac.kr)
논문접수일자 2001년 10월 21일
게재확정일자 2001년 12월 7일

1. 서 론

1. 1 연구의 필요성과 목적

컴퓨터와 네트워크 기술의 급속한 발전으로 인해 전통적인 서지정보는 물론이고 다양한 전문(fulltext)정보에 대한 정보검색이 가능하다. 특히 인터넷상에 수록되어 있는 다양한 주제, 다양한 형태의 전문정보들로 인해 이용자들은 연령이나 학력수준에 제약을 받지 않고 필요한 정보를 적시에 검색해 이용할 수 있다. 그럼에도 불구하고 전문 데이터베이스의 일반 이용자들은 정보검색시 자연어 검색효율이 저하되는 현상으로 인해 - 필요한 정보가 누락되거나 불필요한 정보를 대량으로 포함시켜 제공하고 있는 검색엔진들로 인해 - 제시된 결과들에 대한 적합성 여부를 재확인해야 하는 불편함을 감수하고 있다.

이 같은 결과는 이용자들의 정보검색기법(탐색전략, 질의어 표현·생성·확장)에 대한 미숙함이나 해당사이트가 제공하는 검색 알고리즘의 문제점에 기인하기도 하지만, 동일한 의미에 대해 정보제공자와 탐색자가 동일한 형태의 언어를 사용하지 않고 있기 때문이다. 이를 해결하기 위해서는 검색엔진들이 원문내의 용어들에 대한 정확한 용어통제 기능을 적용하거나, 동의어(異形同義語)나 동형이의어(同形異意語)들을 정확히 안내할 수 있는 시소러스의 기능을 도입해야 할 것이다.

물론, 현재 유통되고 있는 대부분의 국내외 검색엔진들은 시소러스의 기능을 제공하고 있기는 하다. 그러나 현재 제공되고 있는 검색엔진들의 시소러스 기능 중 상하위어의 관계를 분석해 보면 해당 사이트의 디렉토리 구조를 그대로 옮겨놓고 있는 실정일 뿐이다. 또한 일부 시스템에서는 수많은 비용과 전문인력을 동원해 수작업으로 구축한 시소러스를 그대로 적용시켜 정보검색에 활용하고 있기는 하나, 이 또한 원문내의 용어형태와 일치하지 않고 있으므로 사용에 불편할 뿐이다.

이상과 같은 문제점을 해결하기 위해서는 원문에 표현(출현)되어 있는 자연어를 근거로 구축된 용어 관계의 틀을 제공해, 이용자에게 미숙한 정보검색기법(탐색전략, 질의어 표현·생성·확장)을 대신할 수 있는 이용자 지향적인 검색시스템이 구현되어야 할 것이다. 따라서 본 연구에서는 전문정보 검색시 검색효율을 상향시킬 수 있도록, 사전에 정의한(predefined 혹은 predetermined) 용어의 계층관계를 적용하지 않고 자연어를 이용한 자동 정보검색 시스템을 구축하고자 한다.

1. 2 연구의 내용과 방법

본 연구는 다음과 같은 내용과 방법으로 수행하였다.

첫째, 자동정보검색 시스템 구현을 위한 선행조건으로 원문 내에 출현한 자연어들간의 계층관계를 자동으로 형성할 수 있는 이론적 배경을 확보하기 위해 용어들간 계층관계 형성에 밀접한 관계를 갖는 시소러스 자동구축과 시소러스 브라우저와 관련된 선행 연구를 분석하였다.

둘째, 원문 내에 출현한 자연어들간의 계층관계를 자동으로 구현할 수 있는 센트로이드(centroid) 표현 방법에 관련된 알고리즘을 제

시하였으며, 정보검색시 형성된 자연어들간의 계층관계를 이용하여 질의어 확인, 질의어 확장, 검색식 구축 등에 관계하는 자동검색 알고리즘을 제시하였다(서휘 1999).

셋째, 연구의 결과로 나타난 자연어를 이용한 자동정보검색 알고리즘의 타당성을 검증하기 위하여 한국인지과학회의 1989년 1월(Vol. 1. No. 1)부터 1997년 3월(Vol. 8. No. 1)까지의 122편의 학술기사를 대상으로 표제명과 초록 등으로 이루어진 실험용 데이터베이스를 구축하였다.

본 연구에서 제시한 자연어 자동검색용 용어관계 형성과 이를 적용한 자동검색시스템은 소규모의 실험적 환경에서 이루어졌으므로 앞으로 대규모 실험환경에 적용시킴을 통해 좀 더 일반화될 수 있는 연구가 계속되어야 할 것이다.

2. 자연어 자동정보검색 시스템의 필요성

자연어 자동검색 시스템이란 이용자의 정보탐색 행태, 자동색인 작성법, 클러스터링 기법, 시소러스 구축방법, 시소러스 구현방법, 정보탐색 기법 등 정보검색의 모든 분야를 종합한 알고리즘에 의해 구현될 수 있다. 본 장에서는 이용자 정보탐색 행태와 정보검색 기법이란 이용자 관점에서 자동정보검색 시스템에 대한 이용자 요구사항을 기술할 것이다. 자동정보검색 시스템에 대한 이용자 요구사항은 초기 질의어에 관련된 용어들을 자동으로 제시할 수 있는 질의어 확장기능, 질의어와 관련된 용어들을 결합해 적합한 문헌들을 찾아낼 수 있도록 자동으로 탐색식을 구성하는 탐색전략 자동 구축기능이다.

2. 1 질의어 확장용 시스템

현재 인터넷 상에 유통되고 있는 정보는 서지사항이나 초록 등의 2차 정보서비스에서 원정보인 전문 전체를 수록하고 있는 전문데이터베이스 서비스로 급속하게 변화하고 있다. 그러나 대부분의 전문데이터베이스가 사전에 색인작업을 수행하지 않기 때문에 탐색자에게 큰 부담으로 작용한다. 그 이유는 탐색식 구성시 용어의 조합을 적합문헌에 출현하는 정확한 용어를 예측해 조합해야 하지만, 특정주제에 대한 포괄적인 검색을 하는데 필요한 동의어, 계층어, 관련어를 예측하는 것은 저자의 저작유형이 다양하여 탐색자가 이러한 용어를 생각해내기는 무리이기 때문이다(Lancaster 1985, 313).

특히 일반 이용자는 단일어(single word)만을 이용해 탐색을 수행하는 경향이 많기 때문에, 질의어에 포함된 용어가 적합문헌에 출현할 것이란 예측을 통해 검색작업을 수행하지만, 그 용어가 부적합문헌에도 출현할 가능성을 배제할 수 없기 때문에 예측했던 부적합문헌의 수보다 많은 부적합문헌이 검색될 수 있다(Moid 1991, 369).

이상과 같은 문제점을 해결하기 위해 질문확장 또는 질의어 확장(query expansion)이 요구되는 것이다. 질의어 확장이란 이용자의 초기 질문이 주제와 관련이 있는 소수의 용어로만 구성되기 때문에 검색효율이 떨어질 수 있

으므로 초기 탐색어 집합에 이형동의어, 동형이의어, 관련어 등을 추가하거나 제거하는 과정을 의미한다. 그 방법은 시소러스나 의미 네트웍 등과 같은 정보원을 사용해 초기질의어를 확장하는 지식기반확장 방법과 초기질의벡터의 탐색으로 검색된 문헌 중 적합문헌에 출현한 용어들을 사용해 초기질의어를 확장하는 탐색결과 기반 확장 방법 등이 있다(Helen and Peter 1991, 378). 현재 대부분의 검색엔진은 시소러스를 브라우저의 형태로 제공하여 초기 질의어에 대한 동의어, 상위어, 하위어 등의 관련 용어들을 제시해 주는 시스템을 채택하고 있다.

그러나 이같은 질의어 확장방법이 오히려 검색의 효율성을 저하시키는 원인이 되고 있다. 왜냐하면 전통적인 시소러스는 수작업 색인(서지 데이터베이스 검색용 색인)용으로 작성되었기 때문에 자연언어색인을 기본으로 하고 있는 (색인작업을 수행하지 않는) 전문데이터베이스의 검색에는 부적합하기 때문이다. 따라서 정보검색의 효율을 향상시키기 위해서는 색인어와 탐색어가 문헌 보증(literary warrant)과 이용자 보증(user warrant)원칙(Lancaster 1986, 23-28)에 따라 원문 내에 출현하는 자연어와 동일한 형태로 선정되어야 한다.

초기 질의어에 대하여 문헌보증과 이용자 보증이 가능한 색인어와 탐색어로 확장함을 통하여 검색효율을 향상시키기 위해서는 원문 내에 출현한 자연어들을 대상으로 계층관계를 구성해야 한다. 원문 내에 출현한 용어들만을 대상으로 형성된 계층관계는 질의어에 대한 동의어, 계층관계어, 관련어 들을 즉각적으로 탐색에 적용시킬 수 있도록 유도하므로 적합문헌을 검색할 수 있는 가능성을 높일 수 있다. 따라서 자동정보검색을 위한 질의어 확장 시스템은 통제어를 대상으로 구축된 기존의 시소러스 브라우저가 아닌 원문 내에 출현하는 자연어들을 대상으로 계층관계를 제시할 수 있는 브라우저가 필요하다.

2. 2 탐색전략 구축용 시스템

대부분의 전문데이터베이스는 시간적인 측면이나 비용적인 측면 때문에 별도의 색인작업(시소러스 등을 근거한 디스크립터 추출작업)을 수행하지 않으므로 검색된 정보에 불필요한 정보가 포함되거나 필요한 정보가 누락되는 현상이 발생한다. 이같은 결과는 색인어와 탐색어가 동일한 형태로 표현될 수 있도록 하는 도구인 시소러스를 적용하지 않기 때문이기도 하지만, 이용자가 적용한 탐색용어의 부정확성, 용어 조합의 오류, 탐색 전략의 부적합성 때문이다.(Mckinin 외 1991, 303) 또한 Borgman(1996)은 이같은 현상에 대해 최종 이용자가 특정 데이터베이스나 시스템에서 사용하는 시소러스나 주제명표목, 용어사전화일, 시스템 언어 등에 대한 이해 부족은 물론, 부울(boolean) 로직을 이용한 검색결과의 축소 및 확대에 필요한 탐색전략(strategy)과 기법(tactic)등에 대한 이해가 부족한데서 기인한다고 주장하고 있다(Borgman 1996, 493-503).

일반적으로 정보검색의 과정은 사전탐색(presearching), 데이터베이스 선택(DB selection), 탐색전략 구축(searching strategy construction), 온라인 탐색(online searching), 사후탐색(postsearching)의 단계를 거친다. 앞의

질의어 확장 시스템은 사전탐색과 사후탐색에 해당되어 초기질의어와 초기 검색결과를 이용해 계층관계를 갖는 용어와 동의어 등의 관련 어휘를 추출하는 과정이다. 그러나 이 과정에 의해 추출된 확장된 어휘를 이용하더라도, 일반 이용자는 탐색전략 구축방법(and, or, not과 같은 부울 로직을 적용시키는 방법)에 익숙하지 못하기 때문에 정확한 탐색식을 구축하지 못하고 있다.

따라서 탐색전략의 구축과 적용을 최종 이용자를 대신해서 수행할 수 있는 시스템이 필요하다. 이같은 시스템은 Rowley(1994), B. H. Weinberg(1995) 그리고 Susan Jones (1995)의 시소러스의 기능 확장에 대한 선행 연구결과에 나타난 바와 같이 현재 일부 검색시스템에 적용되고 있다. 그 방법은 시소러스가 GUI 또는 윈도우 환경에서 탐색자에게 통제어휘 리스트를 제공하고, 이용자가 이 리스트 중 한 항목을 선택하면 해당 통제어에 대한 레코드의 검색을 수행함을 통하여, 최종 이용자를 대신해 탐색전략의 구축과 적용이 가능한 이용자 인터페이스로서의 기능을 수행하는 방법으로 채택되고 있다. 다른 하나의 방법은 시소러스가 지능 인터페이스를 제공하는 지식베이스로서의 기능을 제공함을 통해, 통제어를 자연어로 변경시키거나 자연어를 통제어로 변경시키는 과정을 통해 질의어를 자연어로 확장하는 탐색으로도 이용이 가능하다. 이와 같이 시소러스는 단어들 사이의 관계를 정의하고, 확장하거나 축소하는 등의 다양한 방법을 통하여 자동적으로 이용자 탐색을 유도할 수 있는 이용자 인터페이스의 기능을 수행할 수 있다.

하지만 이와 같은 방법 또한 시소러스를 구축하는데 필요한 막대한 시간과 노력이 요구되며, 원문 내의 자연어와 시소러스를 연결하는데 필요한 알고리즘을 구축하는 데에도 동일한 시간과 노력이 요구될 것이다. 또한 앞에서 기술한 바와 같이 전통적인 시소러스가 갖는 동일한 문제점을 갖고 있기 때문에 전문데이터베이스에 대한 자동정보검색 방법으로 채택하기에는 많은 문제점이 있다.

이같은 문제점을 해결하기 위해서는 원문을 근거한 자연어 계층관계 브라우저가 검색엔진에 도입되어야 한다. 원문을 근거한 계층관계 브라우저는 용어간 계층을 이용해 부울 로직을 적용시켜 정확한 정보검색이 가능하며, 계층관계를 축소하거나 확대함을 통해 검색결과의 수량을 조절할 수 있다. 또한 자연어에 대한 클러스터링을 이용한 계층관계 브라우저는 부울 로직과 비부울 로직(매칭함수에 의한 확률 검색방법)이 결합된 검색방법이므로 최종 이용자가 탐색전략과 기법 등에 대한 이해 부족에서 발생하는 검색성능의 저하를 해결할 수 있다. 따라서 앞에서 기술한 질의어 확장기능을 갖는 자연어 계층관계 브라우저는 시소러스 브라우저가 갖고 있는 정보검색의 기능을 동일하게 수행할 수 있으며, 계층 내에 표현된 용어들이 전문데이터베이스에 표현된 자연어와 동일한 형태를 갖고 있으므로 정확한 정보검색이 가능하다고 판단된다.

3. 자연어 자동검색의 이론적 배경

정보검색의 전과정을 자동으로 수행하기 위해서는 원문을 근거로 검색의 실마리가 되는

색인어를 자동으로 추출하는 자동색인 알고리즘, 색인어의 계층 구성이나 문헌에 대한 분류를 자동으로 형성토록 하는 클러스터링 알고리즘, 이용자의 탐색 질의어를 확장하고 이 질의어를 이용해 탐색전략과 기법 등 탐색전략을 자동으로 구성토록 하는 자동탐색 알고리즘 등이 종합되어 시스템으로 구성되어야 한다.

3. 1 자동색인 알고리즘

자동색인은 컴퓨터에 입력된 문헌을 대상으로 분석한 후 문헌의 내용을 나타낼 수 있는 단어나 단어구를 추출하는 과정이며 색인 과정에서 분석대상이 되는 부분은 문헌의 전문이나 초록이 된다. 컴퓨터에 의한 자동 색인은 시소러스 이용여부에 따라 시소러스 기반 색인법과 일반 색인 기법(단일어 색인 기법)으로 나뉘며, 일반 색인 기법은 색인어를 선정하는 기준에 따라 통계적 기법, 언어학적 기법, 문헌구조적 기법의 3가지로 나뉘어진다.

시소러스 기반 색인 기법은 연구자들에 따라 그 성능 평가가 다르다. 특히 국내의 시소러스는 문헌 내의 용어 출현 여부보다는 전문가들의 합의에 의한 방법과 해외 시소러스를 단순 번역해 구축한 것들이 대부분이므로 색인 용어의 불철저성(non-exhaustivity)이란 문제점을 안고 있으므로 이를 이용해 색인어를 추출하는 과정은 비합리적이다. 즉 잘못된 시작을 근거로 잘못된 결과를 발생케하는 문제점을 야기시킬 수 있으므로 본 논문에서는 색인작성법을 일반색인 기법(단일어 색인 기법)으로 한정한다.

통계적 기법은 단어의 출현 빈도가 높을수록 그 단어가 문헌의 주제를 대표할 확률이 높다는 가설을 근거한 것으로서, 색인어 선정 방법은 단어의 출현 빈도를 근거로 주제어로서의 중요도를 측정해 색인어를 선정한다.

언어학적 기법은 어휘적 단계, 구문적 단계, 어의적 단계로 나뉘며, 어휘적 단계 기법은 불용어 제거 기법을 의미하며, 구문적 단계 기법은 단어의 구문적 범주 결정을 위해 단어 사전을 사용하는 방법이 포함된다. 이 방법은 단서어 기법과 구문분석 기법이 해당되는데 그 중에서 구문분석 기법이 주류를 이루고 있으며 대부분의 구문분석 기법은 어의분석까지 포함하고 있다.

문헌구조적 기법은 문헌 속에 단어가 나타난 위치에 의해 색인어를 선정하는 기법으로서 서론, 본론, 요약 등의 제목을 갖는 특정한 부분에 나타난 주제들을 색인어로 선택하는 방법과 각 문단의 첫 문장과 마지막 문장과 같은 주제적 문장을 선택하여 이 문장 속에 나타난 주제어를 색인어로 선택하는 방법이 있다.

대부분의 한글 자동색인법은 언어학적 기법을 이용하여 색인의 대상이 되는 명사나 명사구를 식별하고, 통계적 기법을 이용하여 식별된 명사나 명사구를 색인어로 적용시키는 방법을 채택하고 있다(남영준 1994).

3. 2 클러스터링 알고리즘

클러스터링 알고리즘은 비계층 클러스터링 알고리즘(nonhierachical Clustering Algorithm)과 계층 클러스터링 알고리즘(Hierarchical Clustering Algorithm)으로 구분된다.

비계층 클러스터링 알고리즘은 용어간의 계층을 형성하지 않으므로 자연어 계층관계 브라우저 구축방법으로는 적합하지 않아 설명을 생략하기로 한다. 계층 클러스터링 알고리즘(이하부터 계층 알고리즘)은 클러스터 대상물간의 유사성을 측정하여 작성한 문헌-문헌 유사행렬을 이용하여 클러스터를 구성하는 방법이다. 계층 알고리즘은 각 문헌이나 클러스터들이 모두 연결될 때까지 중복을 허용하는 방법으로 링크를 계속하는 방법을 택하므로, 비계층 알고리즘에 비해 공간(space)과 시간은 많이 요구되나 대상 문헌과 클러스터들이 계층을 형성케 되므로 문헌정보 검색에 더 적합한 알고리즘이다.

클러스터를 구성하는 일반적인 과정은 다음과 같다. 먼저 문헌-색인어 행렬을 대상으로 유사치(Similarity) 측정 공식(Jardine and Van Rijisbergen 1971, 225-226 ; Salton 1975, 329)을 적용시켜 각 문헌쌍 {Di, Dj} (i = 1, 2, ···, k / j = 1, 2, ···, k) 들간의 유사계수를 계산해 문헌-문헌 유사계수 행렬을 구성하는 과정에서부터 시작한다. 여기에 일정 기준치(Trash-hold Value) 'T'를 부여해 Sim(Di, Dj) ≥T 인 경우에는 '1'로, Sim(Di, Dj) <T 인 경우는 '0' 으로 하여 '1' 값을 갖는 문헌 쌍들에 소속된 문헌들을 하나의 클러스터에 소속토록 해, 서로 다른 문헌들을 동일한 문헌으로 간주하는 방법이다. 이 알고리즘에 의해 형성된 클러스터들은 유사도 순위에 따라 이원 나무 구조(binary tree struction)로 조직된다.

계층적 알고리즘은 클러스터를 형성하는 방법에 따라 응집적 방법(agglomerative)과 분열적(divisive) 방법이 있다. 응집적 방법은 클러스터가 이루어지지 않은 n개의 문헌 아이템에서 시작하여 n-1번의 결합이 이루어지며, 분열적 방법은 특정 클러스터에 소속된 모든 문헌 아이템들을 대상으로 n-1번의 결합을 통해 더 작은 클러스터들을 형성하는 과정을 거친다. 분열적인 방법은 거의 이용되지 않고 있어 유용한 알고리즘도 거의 존재하지 않고 있다. 현재 주로 이용되고 있는 응집적 방법은 단일연결(single link), 완전연결(complete link), 그룹평균연결(group average), Ward방법(Ward's method) 등이며, 기타 중앙값(Median)방법과 중심값(Centroid)방법이 있다.

3. 3 자동탐색 알고리즘

정보검색의 효율성을 보장하기 위한 정보탐색과정의 핵심은 질의어 선정을 위한 사전탐색, 선정된 질의어들과 부울린 로직의 조합을 이용한 검색전략(검색식) 구축, 검색된 결과에 대한 평가를 근거한 피드백 탐색(사후 탐색)이다. 그러나 최종 이용자들은 특정 데이터베이스나 시스템에서 사용되는 시소러스나 주제명표목, 용어사전화일, 시스템 언어 등에 대한 이해 부족은 물론, 부울린 로직을 이용한 검색결과의 축소 및 확대에 필요한 탐색전략(strategy)과 기법(tactic)등에 대한 이해가 부족하다. 따라서 이용자를 대신해 검색 질의어 확장, 탐색전략 구축 및 탐색, 피드백 탐색 등의 과정을 수행하는 시스템이 필요하다.

질의어 확장은 효과적인 검색을 위해서 이용자의 초기 질의어에 포함된 탐색어에 이형동의어, 동의어, 관련어 등을 추가하는 과정으로써 자동탐색전략의 전단계에 해당된다. 질

의어 확장방법에는 시소러스나 의미네트웍 등과 같은 정보원을 사용해 확장되는 지식기반 확장 방법(Peat and Willett 1991, 378)과 초기 질의벡터의 탐색으로 검색된 문헌 중 적합문헌에 출현한 용어들을 사용해 확장되는 탐색결과 기반 확장 방법(노정순 1999, 69-71) 등이 있다.

탐색방법의 종류는 크게 부울린 탐색(Boolean search)과 매칭 함수(Matching functions)에 의한 탐색 방법으로 나뉘어진다. 부울린 탐색은 and, or, not 등의 연산자를 근거로 정보를 검색하는 방법이며, 매칭 함수(matching fuctions)를 이용한 탐색은 질의를 문헌이나 클러스터와의 연관성을 근거로, 즉 Dice 계수, cosine 계수나 Tanimoto 계수 등의 연관성 측정법(association measure)을 적용해 일정 기준치를 통과하는 문헌만을 원하는 정보라고 판단하여 검색하는 방법이다. 매칭함수에 의한 탐색은 순차탐색(serial search), 클러스터 탐색(clustered based search)이 존재한다.

순차탐색은 모든 문헌의 용어들과 질의 용어를 매칭함수로 비교해 정보를 검색하는 것으로 검색결과는 특정 기준치(threshold value)에 의해 제공되거나, 매칭함수에 의한 문헌의 서열을 근거한 절단서열위치(cutoff rank position)에 의해 제공된다. 클러스터 탐색은 클러스터 알고리즘에 의해 구성된 클러스터의 표현(representatives)을 매칭함수로 비교해 정보를 검색한다. 그 방법은 하향식(top-down) 탐색법과 상향식(bottom-up) 방법이 있다(Van Rijisbergen 98).

피드백 탐색은 검색결과가 만족스럽지 못할 경우 새롭게 탐색을 수행하는 과정을 의미하는 것으로 탐색결과에 근거한 질의확장(query expansion search based on search results)이라고도 한다. 탐색결과를 근거로 자동으로 질의를 확장하여 정보를 탐색하는 방법은 매칭함수를 이용해 구성이 가능하다. 자동 피드백 탐색 방법은 먼저 초기 질의어에 의해 검색된 문헌들에 대해 매칭함수를 이용한 적합성 서열화가 이루어져야 하며, 이를 근거로 가장 적합하다고 판단되는 10% 이내의 문헌에 출현하는 용어를 근거로 초기질의를 확장 또는 수정하여 탐색을 수정하는 과정을 거친다. 적합성 피드백에 의한 초기질의 수정 또는 확장 방법은 질의어 자동 수정 방법(Automatic Query Modification), 질의어 자동 확장 방법(Automatic Query Extension - All), 질의어 자동 선별 확장 방법(Automatic Query Extension - Sellect), 탐색자 개입 질의어 확장 방법(Interactive Query Expansion) 등이 있다.

4. 자연어 자동검색 시스템의 구현

본 시스템은 색인어 자동 추출 기능, 자연어 계층관계 자동 구축의 기능, 질의어 확장 기능, 탐색전략 자동 구축 등을 통해 인간의 간섭을 최소화하면서도 검색의 성능을 극대화할 수 있도록 설계한 자연어 기반 자동검색 시스템이다. 이 시스템은 화면에 제시된 순서대로 이용자가 작업을 수행하면, 장착된 알고리즘에 의해 데이터베이스를 구축하고, 검색시 자연어 계층관계 브라우저를 통해 질의어 확장 및 검색식의 구축, 탐색의 수행 및 피드백 탐색이 가능하도록 설계하였다.

4. 1 시스템 환경

4. 1. 1 개발 환경

본 시스템의 개발은 개인용 컴퓨터를 이용하였으며, 개발 tool은 Paradox 7.0 DBMS를, 개발 언어는 델파이(Delphi 4.0 - PASCAL)를 사용했으며, 한글처리는 KSC5601(행망용 한글코드)을 사용하였다. Paradox DBMS를 사용한 이유는 데이터베이스를 구축하기 쉽고, 상업용 DBMS가 제공하는 다양한 기능을 이용할 수 있기 때문이다.

4. 1. 2 실험 데이터

본 시스템에서 개발한 자연어 계층관계 구성의 주제는 인지과학을 대상으로 하였다. 인지과학을 실험 대상으로 선정한 이유는 한국인지과학회의 인지과학논문집 데이터베이스(http://hawk.kordic.re.kr/~kjcs/db_search/index.html)가 심리학, 언어학, 전산학, 철학, 신경과학, 영문학 등 다양한 주제를 다루고 있어 유사한 주제어에 대해서 용어 간의 통일이 이루어지지 않아 통제어에 의한 정보검색보다 자연어에 의한 정보검색이 더 유리하며, 발표된 학술기사내의 용어들이 학문간 용어통일이 이루어지지 않아 사용된 용어 간의 계층관계를 형성하기에 어려운 주제라고 판단했기 때문이다.

수록한 내용은 1989년 1월(Vol 1. No. 1)부터 1997년 3월(Vol 8. No. 1)까지의 학술기사 122건을 대상으로 하였다. 색인어는 표제명, 부표제명을 대상으로 색인어를 자동으로 추출하는 방법에 의해 선정하였다. 입력 방식은 인터넷을 통해 각 필드의 내용을 복사해서 입력하는 방식으로 하였다.

4. 2 자연어 정보검색시스템 구축 알고리즘

자연어를 이용한 자동검색 시스템은 원문내에 출현한 용어들의 계층관계를 수작업으로 처리하는 과정이나 수작업으로 구축한 시소러스의 계층관계를 도입하지 않고 원문에 출현한 용어들간의 문헌 내 동시출현 빈도를 근거로 계층관계를 자동으로 구성할 수 있도록 설계되었다. 이 시스템의 핵심은 용어들의 계층관계를 자동으로 형성하는 알고리즘과 질의어에 근거한 자연어 자동검색 알고리즘이다. 따라서 본 장에서 기술할 내용은 종합적 알고리즘과 계층관계 형성 알고리즘 그리고 질의어 확장에 관련된 알고리즘으로 한정한다. 자동색인 알고리즘을 제외한 이유는 본 연구에 적용된 알고리즘이 선행 연구와 큰 차이가 없기 때문이다.

4. 2. 1 시스템적 접근

최종 이용자는 정보검색을 할 때 자신이 인지하고 있는 간단한 용어로 원하는 정보를 탐색하는 경향이 있으며, 질의어 확장 방법이나 탐색식 구축에 익숙하지 못하다. 따라서 데이터베이스 내에 관련정보가 수록되어 있음에도 불구하고, 정보가 누락되거나 필요없는 정보가 검색되는 현상이 발생한다. 또한 시스템 입장에서도 엄청나게 쏟아져 나오는 정보를 모두 수작업으로 색인 작업을 하기는 매우 어려운 작업이며, 더욱이 통제어휘로 변환시키는 작업은 불가능하다.

그러므로 정보검색의 효율을 높이기 위해서는 정보 분석과 축적, 질의어 확장 및 탐색의

수행 등 정보검색의 전 과정을 자동화할 수 있는 검색시스템이 필요하다. 즉 원문 내에 출현하는 용어를 자동으로 분석하는 자동색인의 기능, 추출된 색인어들의 계층을 자동으로 추출·표현할 수 있는 자동 클러스터링의 기능, 질의어를 분석해 핵심 색인어로 유도할 수 있는 기능, 선택된 핵심 색인어를 이용한 검색식의 자동 구축 등을 수행할 수 있는 검색시스템을 필요로 하고 있다.

본 시스템은 이와 같은 자동화된 검색시스템의 가능성을 실제로 구현하기 위해 〈그림 1〉과 같이 설계하였다. 먼저 문헌의 서지사항과 초록은 직접 입력하는 방법과 온라인으로 입력하는 방법을 병행해서 사용할 수 있도록 하였으며, 입력된 레코드는 자동색인 알고리즘을 근거로 색인어들을 자동으로 추출할 수 있도록 하였으며, 추출된 색인어들의 출현 빈도를 근거로 클러스터링 알고리즘을 이용해 색인어들의 계층을 자동으로 형성 표현하도록 설계하였다.

또한 정보탐색시 앞에서 구축한 자연어 계층관계 브라우저를 근거로 입력된 질의어 및 관련 용어들을 계층화해 제시해주고, 이를 통하여 탐색을 자동으로 수행할 수 있도록 설계하였다. 이상과 같은 기능을 갖는 시스템을 효과적으로 운용하기 위해 형성되는 데이터베이스와 적용 알고리즘은 마스터 파일, 용어-문헌 파일(자동색인 알고리즘), 용어-용어 유사도 파일(유사도 알고리즘), 용어 클러스터 파일(클러스터링 알고리즘), 계층관계 파일(센트로이드 알고리즘) 등이다.

이상과 같은 방법에 의해 구성한 본 시스템은 기존의 시소러스와는 달리 용어들을 사전에 정의된 계층에 의존하지 않고도 용어들의 계층 구조를 자동으로 구성할 수 있으며, 문헌내에 출현한 용어들만을 대상으로 색인어를 추출하므로 검색시 시소러스 내의 통제어로 변환시켜야 하는 이용자의 불편을 줄일 수 있으며, 수시로 용어 계층관계의 갱신이 가능함에 따라 신조어를 검색하지 못하는 불편함을 해소할 수 있다.

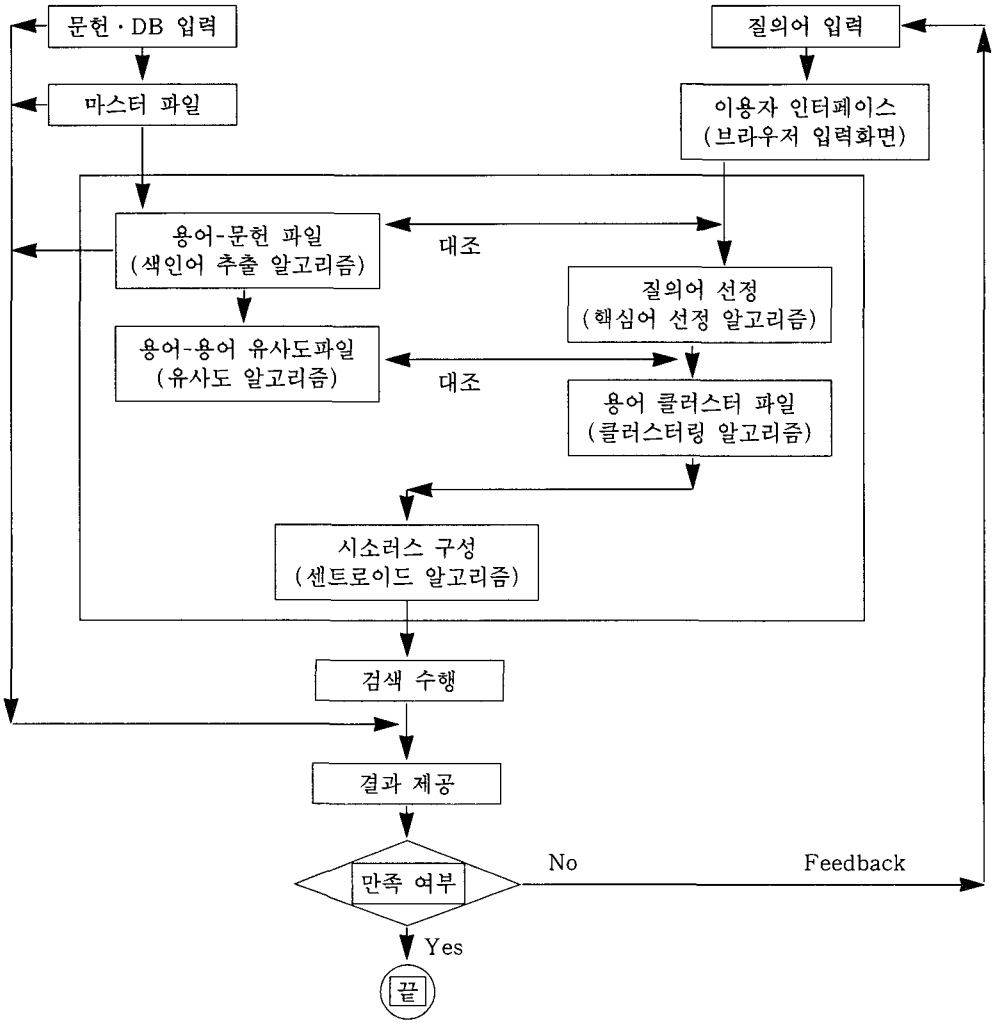
4. 2. 2 자연어 계층관계 형성 알고리즘

(1) 클러스터링 알고리즘

문헌 클러스터를 구성하기 위한 과정 중 각 단계에서 적용한 알고리즘은 다음과 같다. 먼저 마스터 파일을 이용해서 문헌-색인어 행렬과 색인어-문헌 행렬을 구성한다. 색인어-문헌 행렬의 구성은 인버티드 파일 구성 알고리즘을 적용해 색인어의 문헌 출현빈도를 계산한다. 그리고 이를 이용해서 색인어들을 최상위 빈도에서 최하위 빈도순으로 비교해 색인어 간의 완전 연결(포함)여부를 분석한다. 연결여부에 대한 분석은 동일 문헌 포함여부를 근거한 매칭함수 알고리즘을 적용한다.

클러스터의 구성은 문헌을 가장 많이 포함하고 있는 최상위 색인어에 연결된 문헌들을 최정점의 클러스터로 그룹화하고, 이를 근거로 차 순위 색인어에 연결된 문헌들을 하부 클러스터로 구성하는 순으로 진행하여, 더 이상 분리되지 않는 최하위 계층 클러스터인 하나의 문헌까지 순차적으로 비교한다. 이 과정은 인버티드 알고리즘과 매칭함수 알고리즘을 적용한다.

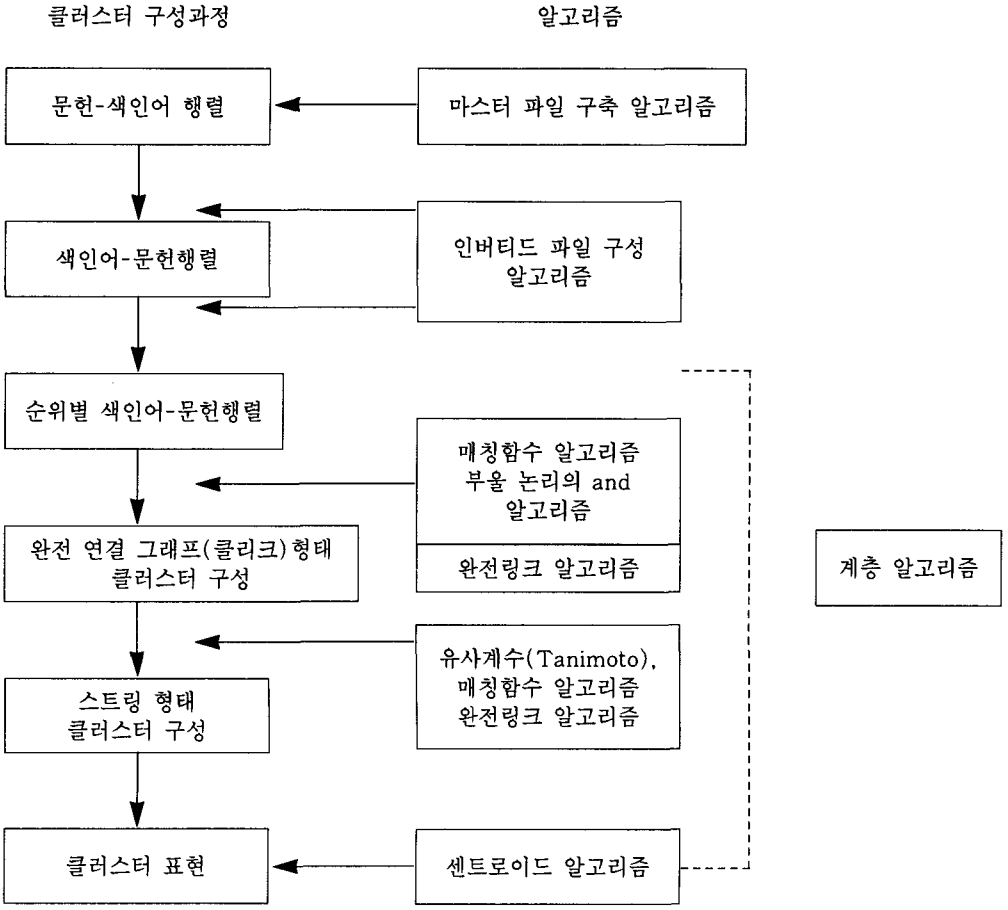
단, 클러스터간의 연결은 매칭함수 알고리즘을 이용하되, 가장 인접하는 하위 클러스터



〈그림 1〉 자연어 자동검색 시스템

는 소속된 문헌들이 상위 클러스터에 소속된 문헌들의 일부와 완전히 일치하는 것으로 한정한다. 이 과정은 부울 논리의 AND 알고리즘을 적용한다.

완전 매칭이 이루어지지 않는 색인어에 포함된 클러스터(문헌)들을 계층화된 클러스터에 연결시키기 위해 Tanimoto공식을 적용해 유사도 측정을 한다. 측정한 결과를 근거로 가장 유사하다고 판단되는 클러스터의 위에 임의의 클러스터를 만들고, 그 표시는 센트로이드 알고리즘을 이용해 각 클러스터의 센트로이드(색인어들)를 동시에 표현한다. 새로이 구성된 센트로이드에 해당 색인어에 포함된 클러스터(문헌)를 연결한다. 이때 적용되는 알고리즘은 매칭함수 알고리즘이다. 각 단계에서 적용되는 알고리즘은 〈그림 2〉와 같다.



〈그림 2〉 클러스터 구성을 위한 단계별 알고리즘

(2) 센트로이드 표현 알고리즘

기존의 센트로이드 표현 방법은 앞 단계의 클러스터 구성을 근거로 공통 용어를 나열하는 방식이었다. 이 같은 표현 방식은 센트로이드의 표현이 길어지게 하므로 핵심 역할보다는 주변 역할을 통해 정보를 검색하도록 해주며, 용어들간의 계층을 식별하기 어려운 문제를 발생한다. 따라서 클러스터를 대표하는 센트로이드는 센트로이드가 의미하는 바와 같이 소수의 핵심어(중심어)로 표현되어야 한다.

센트로이드를 시소러스 구성에 활용할 수 있도록 계층별로 소수의 핵심어로 표현하는 방법의 가설은 '센트로이드는 클러스터를 식별토록 하는 요인이므로 해당 계층에서 동시에 출현하는 용어 중 전체 문헌에서 가장 출현빈도가 낮은 용어가 해당 클러스터를 대표할 수 있다'는 것이다.[1]

이 가설을 적용해 앞의 '클러스터 구성작업'에서 형성한 계층 클러스터의 각 계층을 대표하는 색인어들을 해당 클러스터의 센트로

이드로 표현한다. 계층별로 형성된 클러스터를 단일어 또는 소수의 용어로 표현토록 하는 센트로이드를 추출하는 새로운 알고리즘은 다음과 같다. 센트로이드 추출은 클러스터 구성 작업 순서와 동일하게 한다.

첫째, 출현 빈도수가 제일 높은 색인어를 최상위 계층의 센트로이드로 한다.

둘째, 2번째로 출현빈도가 높은 색인어 중 연결된 문헌의 전부가 최상위 계층에 포함되는 문헌의 일부와 완전히 일치되는 것을 차순위 계층의 센트로이드로 한다.

셋째, 3번째로 빈도가 높은 색인어의 문헌 전부가 앞의 2번째 클러스터와 전부 일치하면, 2번째 색인어와 연결된 하위 클러스터의 센트로이드로 한다. 만약 일치하지 않으면 최상위 센트로이드와 비교하는 작업을 수행하여, 일치하면 최상위 클러스터에 연결되는 클러스터로 인식하고 해당 색인어를 클러스터의 센트로이드로 한다.

넷째, 4번째 순위 색인어를 역순으로 비교해 전부 일치하는 색인어에 연결시키고, 해당 색인어를 센트로이드로 표현한다. 단, 동일빈도의 색인어가 동일 문헌을 포함하는 경우에는 미리 형성된 클러스터와 동일한 것으로 간주하고, 앞에서 형성한 센트로이드 옆에 해당 색인어를 괄호로 묶어 같이 표기한다.

다섯째, 이 과정은 빈도수가 1회인 색인어까지 반복작업을 수행한다. 만약 빈도수가 1회인 색인어에 연결된 문헌이 동일문헌인 경우에는 앞의 과정처럼 센트로이드를 복수로 표시한다.

여섯째, 완전히 연결되지 않는 색인어들을 각 클러스터와 비교해 유사계수가 가장 높은 클러스터의 최상위 계층에 임의의 클러스터를 구성한다. 임의의 클러스터에 대한 센트로이드 표현은 각각의 센트로이드를 OR로 묶어 표기한다. 이 작업은 모든 문헌이 전부 연결될 때까지 반복 작업을 한다.

이상과 같은 클러스터 구성 과정은 다음의 〈그림 3〉과 같다.

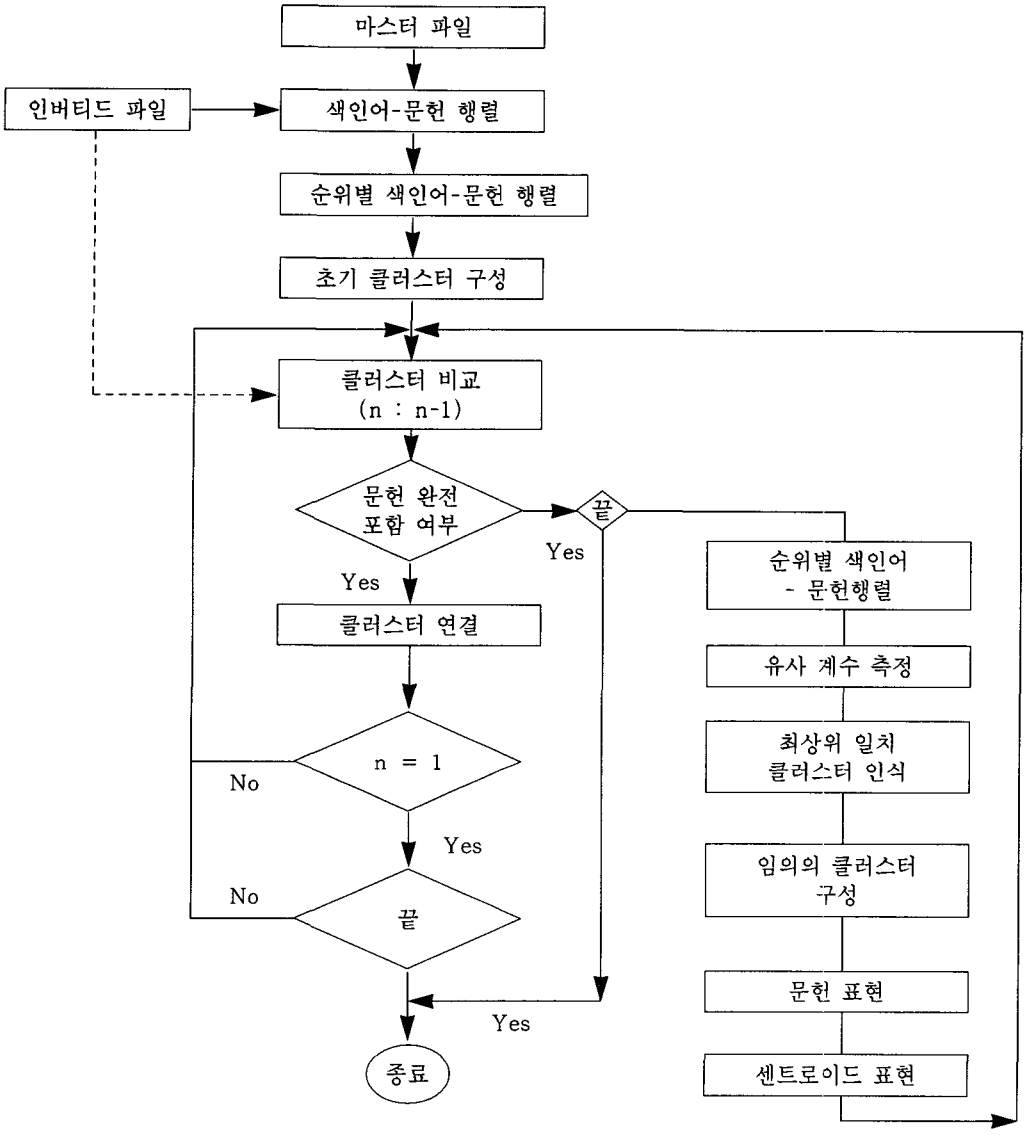
4. 2. 3 자연어 자동검색 알고리즘

(1) 질의어 확장

클러스터링을 이용해 구축한 자연어 계층관계 브라우저에서의 질의어 확장 방법은 문헌에서 실제로 빈번히 사용되고 있는 용어로 질의어를 확장하는 것이다. 앞의 클러스터링 과정에서 생성한 용어 집합은 노드 값에 따라 분리되는데, 이 노드들은 부모 노드와 자식 노드의 관계를 형성한다. 즉 상위 노드의 센트로이드 값은 하위 노드의 센트로이드 값보다 상위개념이 되며, 하위 노드의 센트로이드 값은 하위개념이 된다. 이는 시소러스의 광위어와 협의어의 관계로 설명된다.

정보요구에 대한 질의어 확장 중 초기 질의어 확장에 관심을 갖는 것은 지식기반데이터베이스(시소러스 브라우저)를 이용한 정보검색에서는 문헌탐색과 관련된 용어들이 자동화된 정보검색시스템의 구축과정에서 사전에(정보탐색 수행 이전에) 매칭기법과 클러스터링 기법에 의해 개념간의 매핑이 이루어졌기 때문이다. 또한 센트로이드에 포함된 용어들은

1) 서휘. 2000. 6. 자동정보검색을 위한 한글 시소러스 브라우저 구축에 관한 연구. 『한국도서관·정보학회지』, 31(2): 290



n = 색인어에 연결된 문헌 수

〈그림 3〉 센트로이드 표현 알고리즘

용어간의 계층관계, 동위관계, 등가관계 등 개념간의 매핑이 조화를 이룬 분류기호와 같은 성격의 용어들이기 때문이다.

질의어를 확장하는 과정은 질의 내의 출현 명사를 계층화된 자연어 계층관계 브라우저의 노드 값에 해당하는 색인어들과 대조하는 것으로 시작한다. 자연어 계층관계 브라우저는 부울 논리의 AND기능을 통해 입력된 초기 질의어들이 전부 수록되어 있는 센트로이드들을 매칭함수에 의해 서열화해 제시한다. 물론

이 과정에서 이용자의 간섭없이 자동으로 질의어를 확장할 수 있으나, 검색의 결과는 이용자가 스스로 선택한 질의어에 의해 더 만족될 수 있기 때문에 관련 질의어들을 서열화해 제시하도록 한다. 만약 이용자가 검색어에 대한 간섭을 하지 않을 경우에는 최하위의 용어를 검색의 핵심어로 자동 선정한다.

(2) 정보 탐색

클러스터링을 이용해 구축한 자연어 계층관계 브라우저에서의 검색방법은 최종 이용자가 겪는 탐색의 어려움 중에서 부울 논리를 이용한 검색결과의 축소 및 확대에 필요한 탐색전략과 기법 등에서 발생하는 어려움을 해결한다. 이 과정에서 필요한 알고리즘은 부울 논리, 매칭함수, 기준치를 근거한 탐색방법 등이다.

매칭함수는 질의어 확장에 적용하여 시소러스 브라우저의 개념 노드에 해당하는 센트로이드가 포함하는 용어들을 제시하기 위해 사용한다. 매칭함수는 초기 질의어에 수록된 용어들의 일부가 센트로이드에 포함되어 있지 않더라도 가장 유사한 센트로이드를 검색하도록 해주며, 검색된 확장 질의어 대상들의 서열을 제시하기 위해 적용한다. 부울 논리는 확장된 질의어를 실제 문헌 내에 수록되어 있는 용어들과 대조해 검색하는 과정에서 주로 적용한다. 이 과정에서 질의어들은 모두 AND로 결합해 탐색을 수행한다.

탐색식은 질의어가 동일 계층에 속한 용어인 경우에는 최하위어를 이용해 검색을 수행하면 되나, 특정 어휘 하나만을 입력할 경우에 잡음이 섞일 우려가 있으므로 동일 계층의 상위어들을 and로 조합해 검색식을 자동으로 구성한다. 반면에 질의어가 동일 계층에 속하지 않고, 분리되어 표현될 경우에는 분리된 각 계층의 용어들을 같이 표기하여 AND로 묶어서 검색식을 제시한다. 단, 검색결과가 너무 적어 불만족스러울 경우에는 피드백 탐색을 통해 검색식을 확장한다.

피드백 탐색은 앞의 검색방법을 적용한 결과가 너무 광범위한 내용이거나 협소한 내용이기 때문에 이용자가 만족하지 못할 때 적용한다. 너무 광범위할 경우에는 하위 계층의 용어로 변환해 검색을 수행하며, 너무 협소할 경우에는 상위 계층의 용어로 변환해 검색을 수행한다.

4. 3 시스템의 구현

이용자와 시스템 간의 상호작용으로 자연어에 대한 정보검색이 자동으로 수행되는 본 시스템은 데이터베이스의 구축을 위해 서지사항과 초록을 입력하는 등록 시스템, 입력된 표제명, 부표제명, 초록 내의 용어를 분석하여 색인어를 추출하는 색인어 자동 추출 시스템, 추출된 색인어를 근거로 용어의 계층을 형성하는 클러스터링 시스템, 입력된 질의어를 근거로 관련 어휘들의 계층을 제시하는 자연어 계층관계 브라우저 시스템, 제시된 계층관계를 근거로 탐색 용어를 선택하면 이를 이용해 탐색을 수행하는 탐색시스템으로 구성하였다.

4. 3. 1 데이터 입력 및 색인어 자동 추출

(1) 데이터 입력 화면

데이터 입력의 초기 화면은 등록, 수정, 삭제, 종료 등의 메뉴로 구성하였다. 등록 메뉴

는 새로운 데이터를 입력하는 기능을 담당하며, 수정 메뉴는 입력된 데이터를 수정하거나, 자동으로 색인어를 추출해 데이터베이스에 등록시키는 기능을 한다. 삭제는 특정 레코드를 제거시키는 역할을 하며, 종료는 초기화면으로 돌아가는 역할을 담당한다. 데이터 입력의 초기화면은 다음의 〈그림 4〉와 같다.

(2) 데이터 등록, 수정 및 자동색인 추출 화면

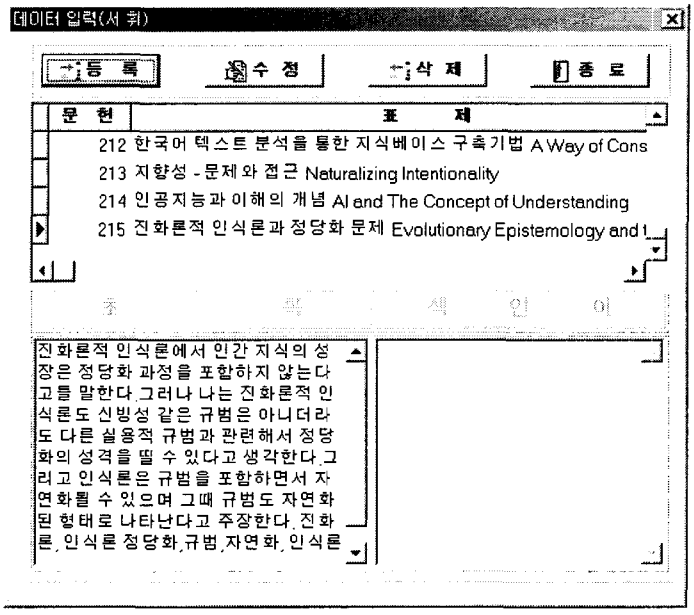
데이터 등록 화면은 표제명(기사명), 부표제명(부표제), 저자명, 연속간행물명, 발행사항, 초록, 색인어 등의 필드를 입력할 수 있도록 구성되어 있다. 입력방법은 원문을 대상으로 직접 입력하는 방식과 인터넷을 통한 복사 입력방식의 병행이 가능하도록 구축하였다. 또한 색인어는 앞에서 입력한 용어(조사, 어미, 조사 겸 어미 용어 등 불용어)리스트들을 근거로 표제, 부표제명, 초록을 대상으로 자동으로 추출할 수 있도록 하였다.

데이터 수정 화면은 앞의 '데이터 입력의 초기 화면'에서 수정 메뉴를 선택하면, 입력된 각 필드의 내용을 수정할 수 있도록 구성하였다. 또한 자동으로 추출된 색인어 중 잘못 선정된 용어의 삭제 및 추가의 기능을 수행할 수 있도록 구축하였다. 데이터 수정 화면은 다음의 〈그림 5〉와 같다.

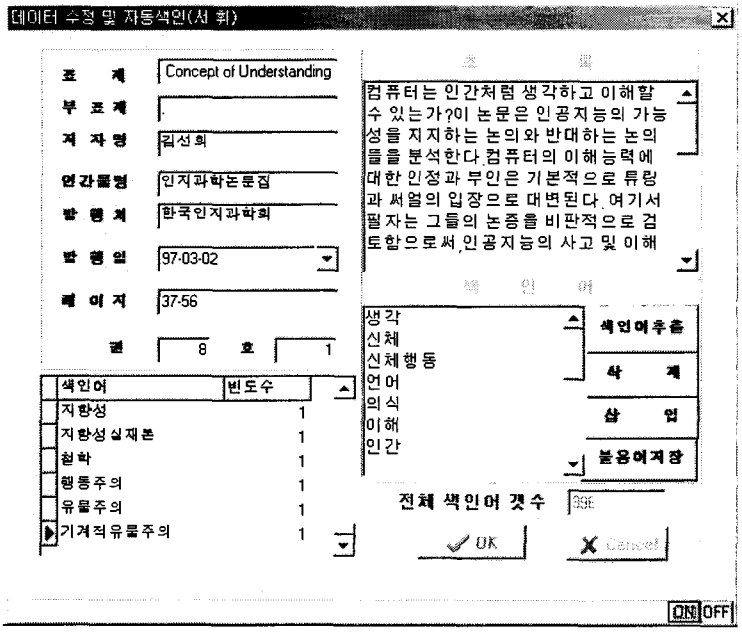
4. 3. 2 자연어 계층관계 브라우저의 구현 및 검색

(1) 자연어 계층관계 브라우저

자연어 계층관계 브라우저는 〈그림 6〉과 같이 초기 브라우저 화면(좌측 화면), 질의어 입력창과 입력 질의어를 중심으로 한 자연어 계층관계 브라우저 화면(우측 화면), 검색 문헌



〈그림 4〉 데이터 입력 초기화면



〈그림 5〉 데이터 수정 및 자동색인 작업 화면

의 출력화면으로 구성하였다.

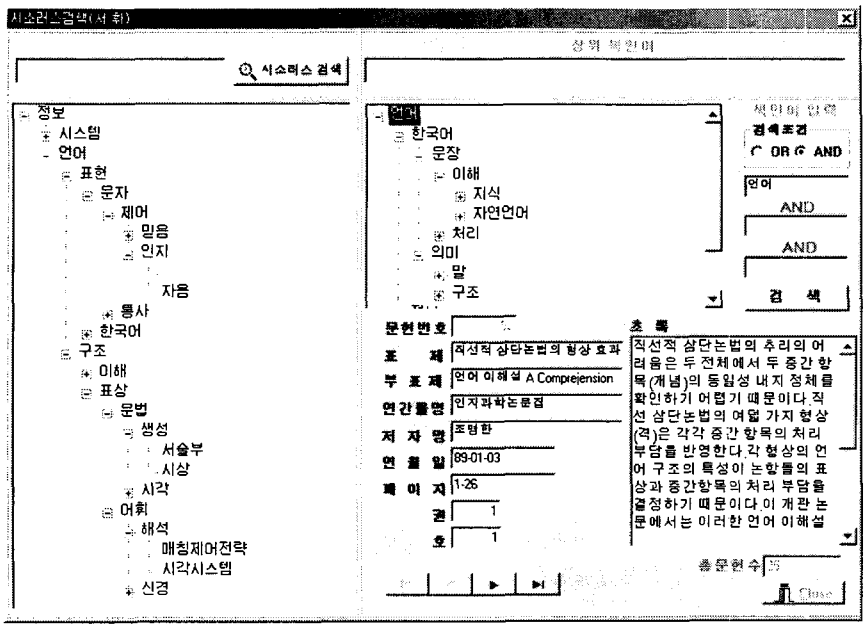
좌측 창의 초기 자연어 계층관계 브라우저는 색인어 중 최상위어를 정점으로 하며, 우측 창의 자연어 계층관계 브라우저는 입력한 질의어를 정점으로 한 색인어의 계층을 제공한다. 각 계층에 연결되는 하위어들은 접기(fold-in)방식을 이용해 제시하며, 화면에 제시된 용어를 선택하면, 해당 용어를 중심으로 상위 색인어 중 한 개 이상을 AND로 결합해 서지사항을 출력하도록 구축하였다.

질의어 입력은 자동색인처럼 구문형식으로 입력해 핵심 질의어를 단일어 형식으로 추출하는 방법도 가능하나 본 시스템에서는 단일어 단위로 입력하는 방식을 채택하였다.

(2) 검색결과 출력화면

출력화면 중 초기 화면은 〈그림 6〉의 왼쪽 창에 제시한 바와 같이 가장 상위어를 중심으로 자연어 계층관계 브라우저의 계층을 나타내 주며, 이용자가 계층에 출현한 용어를 선택하면 이에 해당하는 문헌 레코드를 표제명(기사명), 부표제명(부기사명), 저자명, 연속간행물명, 발행사항, 초록 등의 항목으로 출력하며, 화살표 키를 이용해 다음 문헌의 레코드를 출력하도록 구현하였다.

또한 출력화면을 근거로 이용자의 만족도를 조사해 필요하면 오른쪽 창에서 피드백 탐색을 수행하며, 피드백의 탐색은 3개의 검색어까지 AND나 OR로 검색을 수행하도록 구현하였다. AND의 경우에는 입력한 검색어들을 동시에 갖고 있는 문헌들이 수록하고 있는 색인어를 계층화해 제공해 주며, OR의 경우에는 검색어들 중 가장 상위의 용어에 해당하는 색인어를 중심으로 관련 색인어를 계층화해 시



〈그림 6〉 시소러스 브라우저 화면

소러스 브라우저 화면에 제시하도록 구성하였다. 그리고 이용자가 제시된 색인어 중 핵심 검색어를 선택하면, 이에 해당하는 문헌들을 검색하여 출력하도록 구축하였다.

5. 결 론

본 연구는 인터넷 상의 전문 데이터베이스가 갖고 있는 문제점을 해결하기 위해 수행되었으며, 그 결과로서 자연어를 근거로 한 자동 정보검색 시스템을 구현하였다. 자연어 자동 검색 시스템의 구현에 대한 본 연구의 결과는 다음과 같다.

첫째, 용어들의 계층관계 형성에 관계하는 전통적인 시소러스는 인간의 두뇌에 의한 수작업 과정에 의해 구축되었으나, 본 연구에서 수행한 자연어 계층관계는 용어들의 계층관계를 미리 정의하지 않는 방법으로 구성하였다.

둘째, 클러스터의 각 계층을 대표하는 센트로이드의 표현방법은 복수의 용어로 표현되는 방법으로 정의되었으나, 본 연구에서는 단일(또는 소수) 용어로 표현하는 방법을 개발하였다.

셋째, 본 연구에서는 자동정보검색을 수행하기 위한 전제조건인 질의어 자동 확장과 탐색전략 자동 구축 방법을 지원할 수 있는 자연어 계층관계 브라우저를 자동으로 구축할 수 있는 시스템을 구현하였다.

넷째, 본 시스템은 전문 데이터베이스 내에 신조어가 출현하면, 이를 자동으로 자연어 계층관계 내에 포함시킬 수 있는 다이내믹 시소러스의 기능을 갖추었으며, 질의어의 조합에 따라 다이내믹한 구조로 용어계층을 변형시켜

제공할 수 있도록 설계되었다.

다섯째, 자연어 계층관계가 수작업의 과정을 거치지 않고, 통계적인 방법에 의해 구성되었기 때문에 한글 뿐만 아니라 외국어에도 자연어 계층관계를 자동으로 구축할 수 있도록 설계되었다.

이상과 같은 결과를 근거로 할 때, 본 연구에서 개발한 시스템의 특징은 자연어간의 계층관계를 자동으로 구축할 수 있도록 설계되었다는 점이다. 따라서 본 시스템은 통제색인어의 처리가 이루어질 수 없는 전문데이터베이스 시스템에서 효과적인 정보탐색을 지원할 수 있을 것이다.

참 고 문 헌

남영준. 1994. 『색인어형태분석에 의한 한국어 자동색인기법 연구』. 박사학위논문, 중앙대학교 대학원, 문헌정보학과.

노정순. 1999. 탐색결과에 근거한 자연어질의 자동확장 및 응용에 관한 연구 고찰. 『정보관리학회지』, 16(2): 49-80.

서 휘. 1986. 『정보검색을 위한 인버티드화일과 클러스터화일의 비교분석』. 석사학위논문, 중앙대학교 대학원, 문헌정보학과.

서 휘. 1999. 클러스터링을 이용한 시소러스 브라우저의 설계에 대한 연구. 『한국도서관정보학회지』, 32(3): 427-456.

서 휘. 1999. 『클러스터링을 이용한 시소러스 자동구축에 대한 연구』, 박사학위논문, 중앙대학교 대학원, 문헌정보학과.

Borgman, C. L. 1996. "Why are online catalogs still hard to use ?." JASIS, 47(7): 493-503.

Jardine, N. and Van Rijisbergen, C. J. 1971. "The use of hierachic clustering in information retrieval." Information Storage and Retrieval, 7: 217-226.

Jones, Susan et al. 1995. "Interactive thesaurus navigation: intelligence rules OK?." JASIS, 46(1): 52-59.

Lancaster, F. W. 1985. 『정보검색시스템』. 윤구호, 김태승 공역. 서울: 구미무역.

Lancaster, F. W. 1986. 『Vocabulary Control for Information Retrieval』. 2nd ed. Virginia: Information Resources Press.

Mckinin, E. J. et al. 1991. "The Medline/full-text Tesearch Project." JASIS, 42(4): 217-307.

Peat, Helen J. and Willett, Peter 1991. "The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems." JASIS, 42(5): 378-383.

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

Salton, Gerald. 1975. 『Dynamic Information and Library Processing』. New-jersey : Prentice-Hall, 1975.

Siddiqui, Moid A. 1991. "Full-Text Database." Online Review, 15(6): 367-372.

Van Rijisbergen, C. J. 1998. 『The Hyper-Textbook of the C.J. Van Rijsgergen's textbo ok on Information Retrieval』. [online] 〈http://www.dei.unipd.it/~melo/bible/docu ments〉.

Weinberg, B. H. 1995. "Library classification and information retrieval thesauri: compa rison and contrast." Cataloging and Classification Quarterly, 19(3): 23-44.