자동정보검색을 위한 한글 시소러스 브라우저 구축에 관한 연구

A Study of Designing the Han-Guel Thesaurus Browser for Automatic Information Retrieval

서 휘 (Whee Seo)*

〈목 차〉

초 록

본 연구는 질의어의 표현, 생성, 확장, 탐색식의 구성, 피드백 탐색 등 정보 탐색의 전과정을 자동으로 수행할 수 있는 한글 시소러스 브라우저 기반 자동정보검색 시스템을 구현하기 위해 시도되었다. 구현 시스템은 Delphi 4.0(PASCAL)으로 프로그래밍 되었으며, 자동색인, 클러스터링 기법, 시소러스의 구축과 표현, 자동정보검색이 가능하도록 구성되었다. 구현된 시스템의 평가결과는 새로운 알고리즘에 의해 구축된 시소러스 브라우저가 정보검색에 있어서 시소러스의 구축의 용이성, 이용의 편리성, 검색 속도, 검색의 적합성 수준에서 우수함을 입증하고 있다.

Abstract

This study is to develop a new automatic system for the Korean thesaurus browser by which we can automatically control all the processes of searching queries such as, representation, generation, extension and construction of searching strategy and feedback searching. The system in this study is programmed by Delphi 4.0(PASCAL) and consists of database system, automatic indexing, clustering technique, establishing and expressing thesaurus, and automatic information retrieval technique.
The results proved by this system are as follows: 1)By using the new automatic thesaurus browser developed by the new algorithm, we can perform information retrieval, automatic indexing, clustering technique, establishing and expressing thesaurus, information retrieval technique, and retrieval feedback. Thus it turns out that even the beginner user can easily access special terms about the field of a specific subject. 2) The thesaurus browser in this paper has such merits as the easiness of establishing, the convenience of using, and the good results of information retrieval in terms of the rate of speed, degree, and regeneration. Thus, it turns out very pragmatic.

* 창원전문대학 문헌정보과 조교수

Ⅰ. 서 론

1. 연구의 필요성 및 목적

인터넷을 통한 정보검색은 전세계적으로 급격히 확산되고 있고 접근할 수 있는 자원의 형태 또한 서지 데이터에서부터 멀티미디어에 이르기까지 매우 다양해지고 있다. 특히 전문검색 시스템의 개발은 데이터베이스에 축적된 문헌의 전문(full-text)을 대상으로 필요 정보를 검색하고, 필요한 정보를 담고 있는 전문의 전부 또는 일부분을 볼 수 있다는 점에서 이용자에게 많은 도움을 주고 있다. 그러나 현재 대부분의 정보검색시스템은 데이터베이스 내에 이용자 요구에 적합한 문헌들이 축적되어 있음에도 불구하고 색인어와 탐색어와의 불일치로 인해 적합문헌이 검색되지 못하는 현상과 요구에 부적합한 문헌들이 검색되는 현상이 발생하고 있다. 이러한 문제는 색인어가 자연어이던 통제어이던 간에 탐색자로 하여금 탐색 초기에 질의어의 표현, 생성, 확장 등을 온라인 상에서 조정할 수 있도록 시스템을 설계하여, 색인어와 탐색어의 일치확률을 최대한 증가시킴으로써 개선할 수 있을 것이다. 그러나 이러한 목적으로 개발된 기존의 시소러스 브라우저는 정보탐색자의 탐색전략이나 기법에 대한 인식 부족을 해결하는 기능보다는 단지 질의어에 대한 개념의 정의·표현·생성·확장에 대한 용어제시 수준에 머물고 있다.

더욱이 국내의 경우에는 각 주제에 대한 시소러스가 일부만 구성되어 있으며, 일부 구성된 한글 시소러스도 외국에서 제작된 시소러스에 대한 번역 중심으로 구축되었거나, 어휘(문헌보증)에 대한 객관적 검증이 없이 시소러스 작성자의 판단에 의한 어의 분석에 의존해서 작성되었기 때문에 질의어에 대한 표현·생성·확장의 기능이 실제로 검색효율 향상에 큰 영향을 주지 못하고 있다. 특히 대부분의 전문데이터베이스가 문헌정보의 표현을 시소러스내의 통제어휘로 표현하는 경우가 거의 없기 때문에 이같은 문제점은 더 심화되고 있다.

이상과 같은 문제점을 해결하기 위한 전제조건은 문헌보증과 이용자 보증[1]에 대한 객관적 검증이 보장된 어휘의 선정, 질의어에 대한 표현·생성·확장 기능의 자동화 및 탐색전략이나 기법의 자동 구성 기능을 갖춘 정보검색시스템의 구축에 있다.[2]

따라서 본 연구는 검색언어의 적절한 통제 및 조정기능을 갖춘 검색시스템을 구축하기 위하여 클러스터링 알고리즘을 이용한 시소러스 브라우저를 개발하는데 그 목적이 있다.

---

1) 서 휘, "클러스터링을 이용한 시소러스 브라우저의 설계에 대한 연구", ≪한국도서관·정보학회지≫ 제 32권, 제3호(1999.9), p. 435.

2) F. W. Lancaster. *Vocabulary Control for Information Retrieval.* 2nd ed. Virginia : Information Resources Press, 1986. pp.23-28.

2. 연구의 내용과 방법

본 연구는 다음과 같은 내용과 방법으로 수행하였다.

첫째, 시소러스 브라우저의 기능 확장 가능성을 실제로 구현하기 위하여 시소러스의 자동 구축에 관련된 선행 연구를 분석하였다.

둘째, 선행논문[3]에서 단계적으로 검증된 알고리즘을 근거로 검색시스템의 성능을 향상시키기 위한 새로운 시소러스 브라우저의 자동 구축과 이를 응용한 검색방법을 모색하였다.

셋째, 연구의 결과로 나타난 시소러스 브라우저 자동구축 방법의 타당성을 검증하기 위하여 대한기계학회 논문지의 1998년도에 간행된 194편의 기사를 대상으로 표제명과 영문초록 등으로 이루어진 실험용 데이터베이스를 구축하고, 표제명과 저자가 부여한 색인어들을 대상으로 디스크립터를 추출하고, 클러스터링 알고리즘을 이용하여 탐색용 시소러스 브라우저를 구축하였다.

넷째, 구축된 탐색용 시소러스 브라우저의 성능을 평가하기 위하여 기계공학 분야 교수들과 대학원생, 기계공학 전공 대학생과 문헌정보학과 대학생, 도서관 직원 등 113명을 대상으로 '열교환기 관련 문헌 검색' 등 다양한 질문식을 제시하고, 실험용 데이터베이스에 대한 시소러스의 자동 구축과정과 구축된 시소러스를 이용한 정보 검색 및 피드백 검색을 경험토록 하였다. 그리고 표현된 시소러스 계층의 적합성 여부와 시소러스를 이용한 정보검색의 효율성 평가 등 20개 항목으로 이루어진 설문조사를 수행하여 연구결과를 통해 구축된 본 시스템의 성능을 평가하였다.

본 연구에서 제시한 시소러스 브라우저의 자동구축 방법과 이에 대한 검증은 기계공학 분야의 제한된 실험적 환경에서 이루어 졌기 때문에 앞으로 대규모 실험환경에서의 검증을 통해 좀더 일반화될 수 있는 연구가 계속되어야 할 것이다.

Ⅱ. 선행 연구

본 연구와 관련된 선행연구의 분석은 시소러스의 계층 구축에 직접 관련된 것으로 한정하였다. 그 이유는 시소러스 브라우저의 자동 구축을 위해서는 자동색인, 클러스터링 기법, 자동정보검색 기법, 피드백 탐색 등 정보검색의 전 분야에 대한 선행조사를 해야 하지만, 각 단계

3) 서 휘, "클러스터링을 이용한 시소러스 브라우저의 설계에 대한 연구", ≪한국도서관·정보학회지≫ 제32권, 제3호(1999.9), pp. 427-456.

중에서 가장 핵심이 되는 부분은 용어들간의 계층 관계를 자동으로 형성하는 것이기 때문이다.

1. 해외 연구

G. Salton[4]은 1971년 SMART 프로젝트에서 용어들간의 계층관계를 형성하는 공식을 발표하였다. 그는 한 문헌 내에서 용어들의 출현 빈도를 근거로 계층관계를 형성하는 알고리즘을 개발하였다. 또한 W. B. Turski[5]는 시소러스를 기반으로 한 정보검색 모델을 최초로 연구하였다. 이 연구를 위한 시소러스는 용어들간의 동의 및 일반화관계(상위어와 하위어를 연결시키는 관계)를 이용하여 구축되었다.

D. Soergel[6]은 1974년 용어간의 연관성을 근거로 시소러스의 구축을 제안하였다. 그는 문헌을 대표하는 색인어들을 근거로 용어들이 동시에 출현하는 빈도를 통계적으로 측정해 용어들간의 관계를 형성할 것을 제안하였다.

T. Radecki[7]는 1976년 퍼지 시소러스를 이용해 정보검색 시스템의 수학적 모델을 개발하였다. 그는 용어들간의 유사도 관계를 퍼지 관계로 공식화하였으며, 이를 이용해 시소러스를 구축하였다. P. Shoval[8]은 1985년 이용자들이 정확한 검색어휘를 선정할 수 있도록 용어들간의 동의·계층·관련어 등의 관계를 의미망의 형태로 제공하는 전문가 시스템 형태의 지식기반 정보검색 시스템을 구축하였다.

Miyamoto[9]는 1990년 퍼 지시소러스 관련어 생성공식을 이용하여 관련어를 생성하였다. 그는 용어의 동시출현빈도와 퍼지집합 연산을 이용해 관련있는 키워드들의 집합을 구성하는 퍼지시소러스를 제안하였다. 이 연구에서 관련어 관계는 Jaccard 계수를, 계층관계는 Salton의 자동생성 공식을 적용시켰다.

Kimoto와 Iwadera[10]는 이용자의 특별한 관심사를 반영하는 동적 시소러스를 구축하였다.

---

4) Gerald Salton, *The SMART Retrieval System : Experiments in automatic Document Processing.* N.J. : Prentice-Hall, 1971, pp.133-141.

5) W. B. Turski, "On a model of information retrieval system based on thesaurus", *Information storage and Retrieval,* Vol. 7, No. 2(1971), pp. 89-94.

6) D. Soergel, "Automatic and Semi-Automatic Methods as an Aid in the Construction of Indexing Languages and Thesauri", *Intern. Classif,* Vol. 1, No. 1(1974) pp. 34-39

7) T. Radecki, "Mathematical model of information retrieval system based on the concept of fuzzy thesaurus", *Information Processing and Management,* Vol. 12, No. 5(1976), pp. 313-318.

8) P. Shoval, "Principles, procedures and rules in an expert system for information retrieval", *Information Processing and Management,* Vol. 21, No. 6(1985), pp. 475-487.

9) S. Miyamoto, "Information Retrieval Based on Fuzzy Associations", *Fuzzy Sets and Syst ems,* Vol. 38(1990), pp.191-205.

이 연구에서는 이용자가 자신의 정보요구에 적합한 문헌을 선택하도록 한 다음, 이 표본 적합 문헌으로부터 용어정보를 추출하여 개인별로 동적 시소러스를 구축하였다. 동적 시소러스의 구축시에 용어간의 계층관계는 기존의 수작업으로 편집된 시소러스를 이용하였다.

앞에 제시한 알고리즘들이 본 연구와 동일한 방법으로 시소러스를 구축하였다면, 정보검색의 효율성을 증진시키기 위해서 언어의 형태에 관계없이 모든 주제별 시소러스를 구축하였을 것이다. 그러나 아직 도서관 및 정보검색 관련 소프트웨어에 시소러스의 구축을 자동화하는 알고리즘이 적용된 사례가 없다. 따라서 아직 국내외적으로 시소러스를 구축하는데 많은 수작업의 과정이 요구되고 있다.

본 연구에서 구현한 시소러스는 용어들의 계층을 미리 정의하거나 사전에 결정하는 방법에 의해 구축한 것이 아니고, 연구 진행 과정 중 발견된 '용어들의 계층을 자동으로 형성하는 알고리즘'에 의해 구축하였다. 본 연구결과가 색인어가 부여된 문헌 데이터베이스에 적용된다면, 부여된 색인어의 언어 형태와 관계없이 시소러스의 자동구성이 가능할 것이다.

따라서 본 연구에서 수행한 방법과 동일한 방법으로 연구를 수행한 선행 논문은 발견할 수 없었다.

2. 국내 연구

국내에서 클러스터링을 이용한 시소러스의 자동 구축에 대한 최초의 연구는 1989년 색인어들간의 상관관계를 이용하여 정보검색의 효율화 방안을 제안한 김영환[11]의 계층적 시소러스 구축에 대한 연구이다. 이 연구에서는 미국의 Association for Computing Machinery에 의해 구축된 계층적 시소러스인 CRCS(computing reviews classification structure)를 이용하여(용어의 계층을 미리 정의하여) 한글이 아닌 영문 형태의 컴퓨터 관련 용어들의 계층을 구성하였다.

1994년 이재윤[12]은 92건의 증권분야 신문기사를 대상으로 동적 시소러스를 구축하였다. 이 연구에서는 자동색인과 용어의 가중치 방법, 용어간의 유사도 계산을 이용해 실험용 시소러스를 구축하였다. 이 시소러스에서 용어들간의 계층관계 등은 기존에 구축된 정적 시소러스의 계층을 근거로 구축되었다.

1994년 박영몽[13]은 자연어를 이용하여 정보요구를 표현하고 정보검색 시스템이 이를 색인

10) H. Kimoto and T. Iwadera. "Construction of a Dynamic Thesaurus and its Use for Associated Information Retrieval", *In Proceedings of the 13th International conference on Research and Development in Information Retrieval.* New York : The Association for Computing Machinery, 1990, pp. 227-240.

11) 김영환, 『계층적 개념 그래프를 이용한 지식기반 정보검색 모델』. 박사학위논문, 한국과학기술원, 1989.

12) 이재윤, 『동적 시소러스의 구축에 관한 실험적 연구』. 석사학위논문, 연세대학교 대학원, 1994.

어로 전환시켜 정보 검색의 효율을 높이기 위한 방안으로 인식론적 측면에서의 시소러스 구축을 제안하였다. 이 연구에서도 시소러스의 구축을 사전에 정의하는 방법으로 구축하였다. 1998년 정영미[14] 등은 용어 간의 관계를 통계적인 분석을 통해 자동으로 추출하기 위해 다양한 연관계수 공식의 성능을 측정하였다. 이 연구는 추후 정치, 경제, 사회 분야 신문기사에 수록되어 있는 100만 어절을 대상으로 실험을 할 계획이다.

이상과 같이 시소러스 구축에 관한 원리와 방법론, 그리고 문제점에 대한 많은 연구들이 있다. 그러나 일부만이 시소러스의 자동 구축에 이용되었을 뿐, 국내외 대부분의 시소러스 구축은 수작업 과정에 의존하고 있다. 수작업에 의한 시소러스의 생성은 상당히 개념적이고 지식-집적도가 높은 작업이며 노동-집약적인 특성을 지니고 있다. 따라서 자동적인 수단에 의해 구축되는 시소러스는 정보검색 시스템의 성능 확대에 큰 영향을 미칠 것이다.

Ⅲ. 시소러스 브라우저 설계단계의 설정

시소러스 브라우저의 목적은 최종 이용자의 주제탐색활동을 지원하는 것이므로, 시스템 구축시 이용자 중심 인터페이스, 화면 구성의 원칙, 정보검색시스템으로의 원칙, 주제탐색과 관련된 시소러스의 역할 및 기능적 측면을 고려해야 한다. 시소러스 브라우저가 이상과 같은 역할과 기능을 수행하도록 하기 위해서 다음과 같은 방향으로 시소러스 브라우저를 설계하고자 한다.

첫째, 시소러스 브라우저에 출현한 용어의 계층은 사전에 그 계층을 조정하지 않고(전통적인 수작업 시소러스에 의존하지 않고) 시스템에 내장된 알고리즘에 의해서만 계층이 형성될 수 있도록 한다.

둘째, 디스크립터는 이용자가 질의어를 구성할 때, 질의어가 문헌 내에서 저자가 사용하는 용어와의 일관성을 보장하는 문헌 보증의 역할을 해야 하므로, 반드시 문헌 내에 출현한 용어만으로 한정한다.

셋째, 디스크립터의 추출 작업은 시간과 비용을 절약하기 위해 자동색인 방법에 의하며, 디스크립터가 특정성과 망라성을 겸비할 수 있도록 표제 및 초록 내에 출현하는 용어와 저자가

13) 박영몽, 『지적 정보 검색을 위한 인식론적인 시소러스 시스템의 설계 및 구현』. 석사학위논문, 아주대학교 대학원, 1994.

14) 정영미. 이재윤, "한국어 텍스트 내 용어 연관성 분석을 위한 기초 연구", 《한국정보관리학회 학술대회 논문집》, 제5권(1998), pp. 243-246.

부여한 색인어들을 대상으로 통계적기법, 언어학적 기법, 문헌구조적 기법이 조합되어 의미있는 용어들을 디스크립터로 선정하도록 구성한다.

넷째, 시소러스 브라우저를 구축하기 위한 용어 클러스터링 알고리즘은 용어들의 계층화된 표현이 가능하도록 클러스터간의 계층관계를 최소 유사도만으로도 식별이 가능한 완전링크 클러스터링 알고리즘을 적용하며, 용어간의 계층 관계를 소수의 용어로 식별이 가능하도록 하기 위해 센트로이드 알고리즘을 병행해 사용한다.

다섯째, 시소러스 브라우저의 화면 및 용어 구성은 앞의 클러스터링 과정에서 형성된 용어간의 계층들을 윈도우 화면에서 제공하는 접기(fold in) 방식으로 제공하며, 이를 통해 검색어휘의 선정 및 검색식의 자동구축이 가능하도록 제공한다.

여섯째, 검색방법은 초기의 질의어를 핵심으로 하고, 질의어의 확장은 시소러스 브라우저의 인터페이스 기능을 이용해 관련된 질의어들을 제시하고, 이용자가 직접 탐색을 수행할 질의어들을 선택하면 이를 이용해 검색을 수행하는 방법을 택한다.

일곱째, 피드백탐색은 재현율과 정도율을 선택적으로 수용할 수 있도록 용어 간 AND와 OR의 기능을 이용하여 초기 시소러스 브라우저에 수록되어 있지 않은 특정 용어에 대한 검색이 가능하도록 한다.

1. 시스템적 접근

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

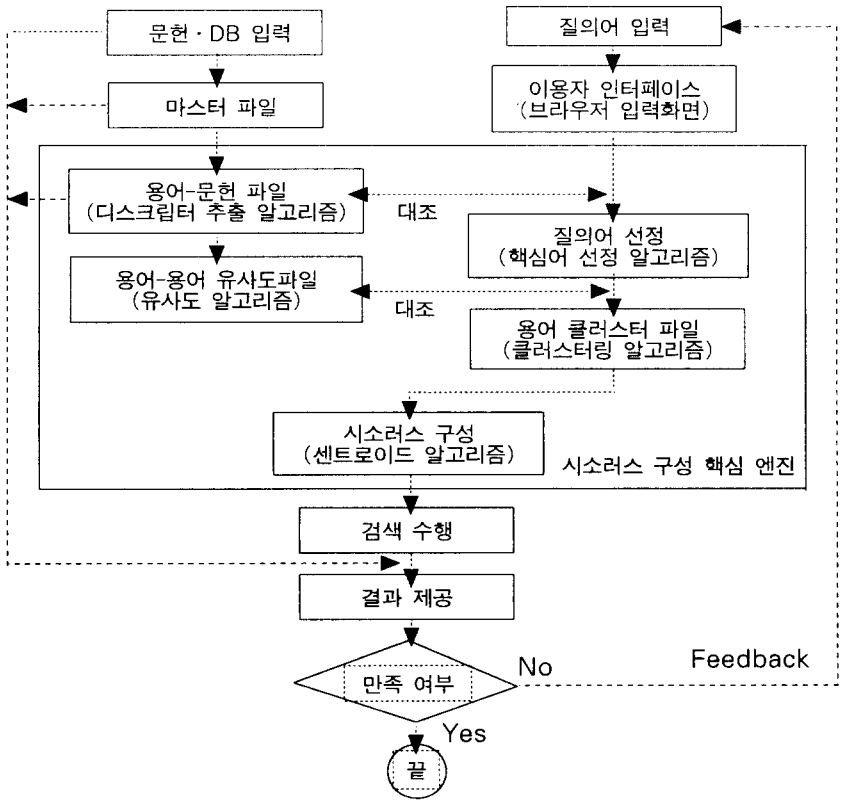
그러므로 정보검색의 효율을 높이기 위해서는 정보 분석과 축적, 질의어 확장 및 탐색의 수행 등 정보검색의 전 과정을 자동화할 수 있는 검색시스템이 필요하다. 즉 원문 내에 출현하는 용어를 자동으로 분석하는 자동색인의 기능, 추출된 색인어들의 계층을 자동으로 추출·표현할 수 있는 자동 클러스터링의 기능, 질의어를 분석해 핵심 색인어로 유도할 수 있는 기능, 선택된 핵심 색인어를 이용한 검색식의 자동 구축·탐색·피드백 탐색을 수행할 수 있는 검색시스템을 필요로 하고 있다.

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 자

동으로 디스크립터들을 추출할 수 있도록 하였으며, 추출된 디스크립터들의 출현 빈도를 근거로 클러스터링 알고리즘을 이용해 색인어들의 계층을 자동으로 형성 표현하도록 설계하였다.

또한 정보탐색시 앞에서 구축한 시소러스 브라우저를 근거로 이용자 불편의 축소와 검색의 효율을 높이기 위해서 입력된 질의어를 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(디스크립터 추출 알고리즘), 용어-용어 유사도 파일(유사도 알고리즘), 용어 클러스터 파일(클러스터링 알고리즘), 시소러스 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, 문헌 내에 출현한 용어들만을 대상으로 디스크립터를 추출하므로 검색시 시소러스 내의 통제어로 변환시켜야 하는 이용자의 불편을 줄일 수 있으며, 수시로 시소러스의 갱신이 가능함에 따라 신조어를 시소러스에 의해 검색하지 못하는 불편함을 해소할 수 있다.



〈그림 1〉 시소러스 브라우저 시스템 구성

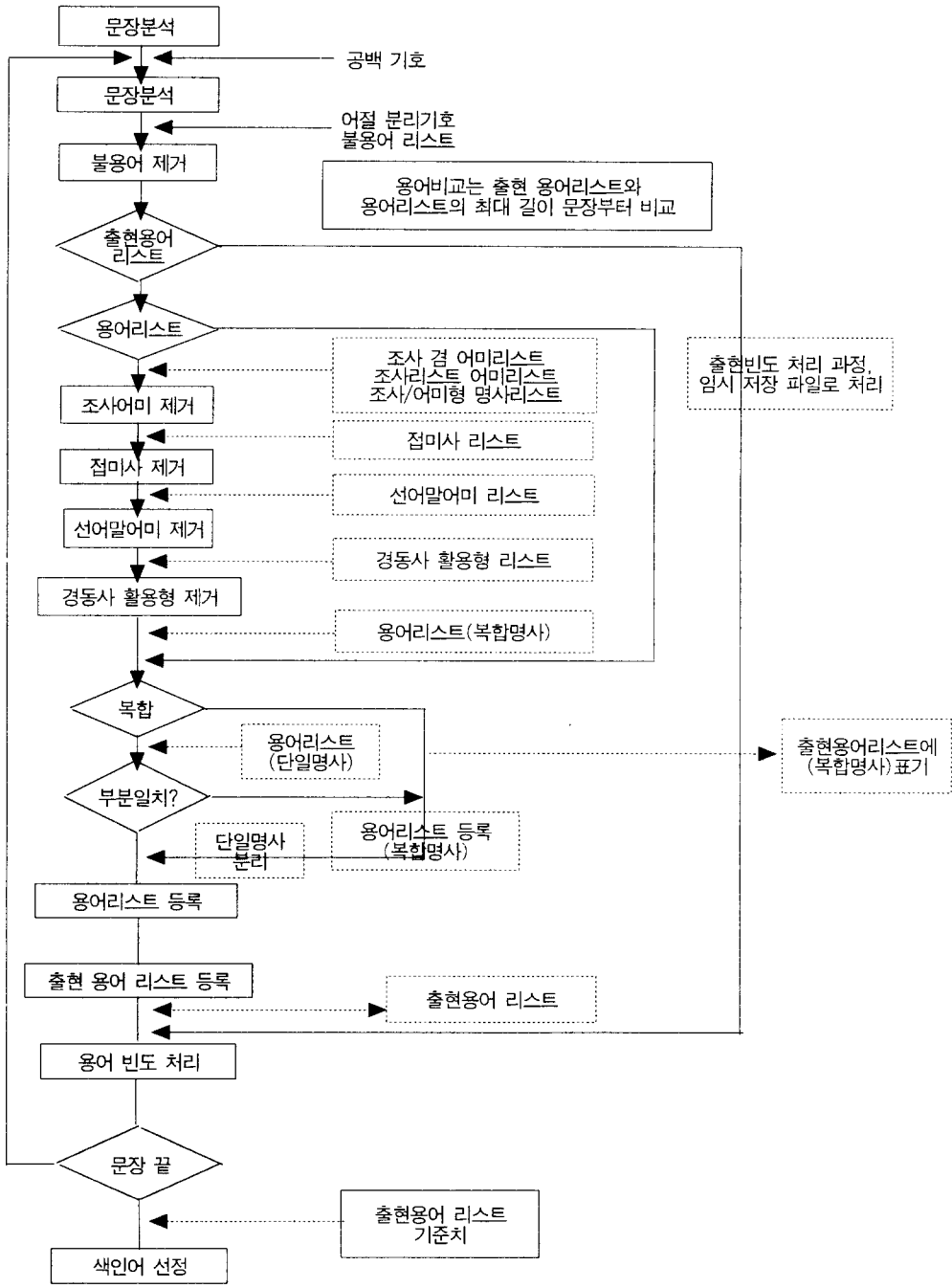
2. 디스크립터 추출 알고리즘

본 시스템에서 디스크립터의 자동 추출을 위해 각 단계에 적용한 단일어 자동색인법의 알고리즘은 다음과 같다. 먼저 문헌구조적 기법을 적용하여 표제, 초록 내에 출현하는 용어를 대상으로 분석하였으며, 언어학적 기법 중 어휘적 단계 기법을 적용하여 불용어 리스트 구축과 불용어를 제거하였다. 또한 구문적 단계 기법을 적용하여 용어 리스트의 구축과 갱신을 수행하였다.

그리고 구문분석과 어휘 분석을 통하여 색인의 대상이 되는 의미있는 어휘인 명사나 명사구를 식별하도록 했으며, 통계적 기법을 적용하여 특정문헌의 명사나 명사구 중 2회 이상의 출현빈도를 나타내는 것만을 색인어로 선정하였다. 2회 이상의 출현빈도를 갖는 명사를 채택하는 방법을 취한 이유는 초록 내에서 반복 출현하는 용어가 해당 문헌의 특성을 표현하는 중요한 용어라고 가정하였기 때문이다. 단, 2회 이상 출현한 용어가 거의 없는 경우를 예상하여 표제나 초록 내에 출현한 용어가 용어 리스트나 기 구축된 색인어 리스트와 일치하는 경우에는 1회만 출현한 용어도 디스크립터로 선정하였다.

또한 선택된 명사들이 디스크립터로 적합한 지의 최종 결정을 위해서 채택된 디스크립터의 대상어들을 즉각적으로 제시함으로서, 색인 담당자가 디스크립터로서의 활용 가능성에 대한 판단을 하도록 하고, 제시된 창에서 디스크립터의 추가와 삭제를 할 수 있도록 하였다. 이 같은 과정은 선정된 디스크립터에 대한 신뢰성을 보장하기 위하여 채택하였다.

본 시스템에서 적용한 2회 이상의 출현 빈도를 나타낸 용어만을 색인어로 선정하는 알고리즘은 데이터베이스에 띄어쓰기나 잘못 입력된 용어들을 색인어로 선정되지 못하도록 하는 장점을 갖는다. 이상과 같은 디스크립터 자동추출 알고리즘은 <그림 2>와 같이 축약된다.

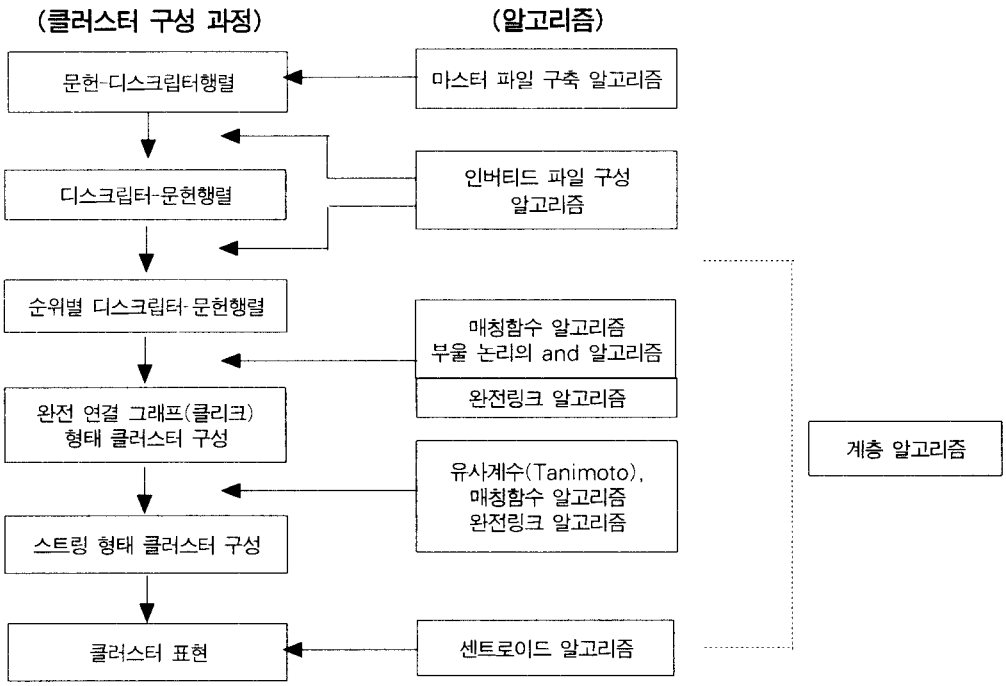


〈그림 2〉 디스크립터 자동 추출 플로우 차트

3. 시소러스 구성 알고리즘

1) 클러스터링 알고리즘

클러스터를 구성하기 위한 과정 중 각 단계에서 적용한 알고리즘은 다음과 같다. 먼저 마스터 파일을 이용해서 문헌-디스크립터 행렬과 디스크립터-문헌 행렬을 구성한다. 디스크립터-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. 연결 여부에 대한 분석은 동일 문헌 포함여부를 근거한 매칭함수 알고리즘을 적용한다.



〈그림 3〉 클러스터 구성을 위한 단계별 알고리즘

클러스터의 구성은 문헌을 가장 많이 포함하고 있는 최상위 디스크립터에 연결된 문헌들을 최정점의 클러스터로 그룹화하고, 이를 근거로 차 순위 디스크립터에 연결된 문헌들을 하부 클러스터로 구성하는 순으로 진행하여, 더 이상 분리되지 않는 최하위 계층 클러스터인 하나

의 문헌까지 순차적으로 비교한다. 이 과정은 인버티드 알고리즘과 매칭함수 알고리즘을 적용한다.

단, 클러스터간의 연결은 매칭함수 알고리즘을 이용하되, 가장 인접하는 하위 클러스터는 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 3>과 같다.

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 발생시킨다. 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.

이 가설을 적용해 앞의 '클러스터 구성작업'에서 형성한 계층 클러스터의 각 계층을 대표하는 디스크립터들을 해당 클러스터의 센트로이드로 표현한다. 계층별로 형성된 클러스터를 단일어 또는 소수의 용어로 표현토록 하는 센트로이드를 추출하는 새로운 알고리즘은 다음과 같다. 센트로이드 추출은 클러스터 구성 작업 순서와 동일하게 한다.

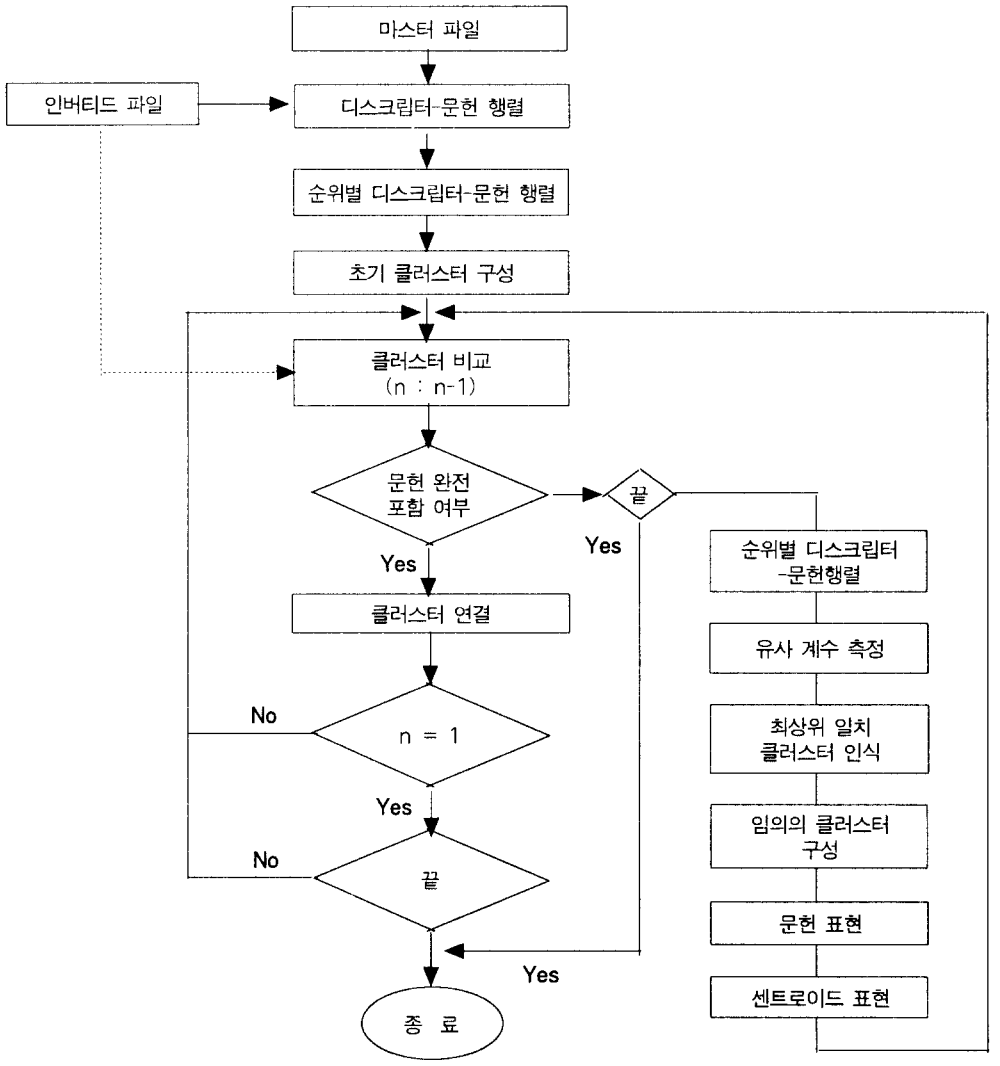
첫째, 출현 빈도수가 제일 높은 디스크립터를 최상위 계층의 센트로이드로 한다.

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

셋째, 3번째로 빈도가 높은 디스크립터의 문헌 전부가 앞의 2번째 클러스터와 전부 일치하면, 2번째 디스크립터와 연결된 하위 클러스터의 센트로이드로 한다. 만약 일치하지 않으면 최상위 센트로이드와 비교하는 작업을 수행하여, 일치하면 최상위 클러스터에 연결되는 클러

스터로 인식하고 해당 디스크립터를 클러스터의 센트로이드로 한다.

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



n = 디스크립터에 연결된 문헌 수

〈그림 4〉 클러스터 구성 플로우 차트

다섯 번째, 이 과정은 빈도수가 1회인 디스크립터까지 반복작업을 수행한다. 만약 빈도수가 1회인 디스크립터에 연결된 문헌이 동일문헌인 경우에는 앞의 과정처럼 센트로이드를 복수로 표시한다.

여섯 번째, 완전히 연결되지 않는 디스크립터들을 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 과정을 플로우 차트로 작성하면 <그림 4>와 같다.

4. 브라우저 화면 구성

일반적으로 정보를 탐색하는 대부분의 초기 탐색자들은 자신의 정보 욕구를 정확히 표현하지 못하는 경향이 많고, 정보탐색을 대행하는 정보검색 전문가는 특정 주제 분야의 전문 용어를 잘 알지 못하므로 검색효율이 저하되는 문제가 발생한다. 따라서 이 같은 문제점을 해결하기 위해서 앞에서 구축한 용어간의 계층 구조를 화면에 제시해줄 수 있는 시소러스 브라우저가 필요하다.

화면 구성은 앞의 과정에 의해 구성된 시소러스에 배열된 계층화된 디스크립터들을 이용해, 이용자가 입력한 질의어와 가장 일치하는 디스크립터의 계층을 화면에 제시해주는 방법을 채택한다. 화면 표현 방법은 접기(fold in) 방법을 이용한다.

5. 검색 알고리즘

1) 질의어 확장

클러스터링을 이용해 구축한 시소러스 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 정

보검색시스템의 구축과정에서 사전에(정보탐색 수행 이전에) 매칭기법과 클러스터링 기법에 의해 개념간의 매핑이 이루어졌기 때문이다. 또한 센트로이드에 포함된 용어들은 용어간의 계층관계, 동위관계, 등가관계 등 개념간의 매핑이 조화를 이룬 분류기호와 같은 성격의 용어들이기 때문이다.

질의어를 확장하는 과정은 질의 내의 출현 명사를 계층화된 시소러스 브라우저의 노드 값에 해당하는 디스크립터들과 대조하는 것으로 시작한다. 시소러스 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 시소러스 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.

(1) 동일 계층 검색

동일 계층의 질의어들인 경우에는 최하위어를 이용해 검색을 수행하면 되나, 특정 어휘 하나만을 입력할 경우에 잡음이 섞일 우려가 있으므로 동일 계층의 상위어들을 and로 조합해 검색식을 자동으로 구성해 검색을 수행한다.

(2) 분리 계층 검색

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.

3) 피드백 탐색

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. 클러스터링을 이용한 시소러스 브라우저에서의 피드백 탐색은 다음과 같다.

(1) 동일계층

동일 계층의 질의어들인 경우 최하위어를 중심으로 검색하면 검색결과가 너무 적게 나오므로 이에 대한 확대 과정이 필요하다. 확대방법은 상위어휘를 중심으로 검색식을 변경하는 방법을 채택한다.

(2) 분리계층

분리계층에 대한 검색시 대부분의 검색결과가 너무 적게 나올 것이므로 이에 대한 확대 과정이 필요하다. 그 과정은 앞의 동일계층의 방법과 비슷하게 분리 계층 중 하위계층의 용어를 확대하는 방법에 의해서 가능하지만, 만약 이용자가 만족해하지 않을 경우에는 계속 동일한 작업이 반복되어져야 한다. 따라서 반복작업을 줄이기 위해서 2용어를 OR로 묶어 처리한다.

Ⅳ. 시소러스 브라우저 자동구축 시스템의 구현 및 평가

1. 시스템 환경

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.

2) 실험 데이터

본 시스템에서 개발한 시소러스의 주제는 기계공학의 '열 및 열유체' 분야로 제한하였다. 기계과학을 실험 대상으로 선정한 이유는 대한기계과학회 논문집 데이터베이스(http://society.kordic.re.kr/~ksme/db_search/direct ory/)가 세부 주제를 다루고 있어 기계공학의 열 및 열유체 분야에서 전문 용어들의 계층이 뚜렷할 것이라고 판단했기 때문이다. 또한 논문집의 표제가 특정 주제를 의미하는 내용으로 이루어져 있으며, 저자가 부여한 한글 색인어가 제시되어 있어서 디스크립터의 선정이 용이하므로 본 연구에 가장 적합하다고 판단했기 때문이다. 수록한 내용은 1998년도에 간행된 22권 1호부터 22권 7호까지의 학술기사 194건을 대상으로 하였다. 디스크립터는 표제명, 부표제명, 저자가 부여한 색인어구를 대상으로 디스크립터를 자동으로 추출하는 방법에 의해 선정하였다. 입력 방식은 인터넷을 통해 각 필드의 내용을 복사해서 입력하는 방식으로 하였다.

2. 시스템의 구현

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

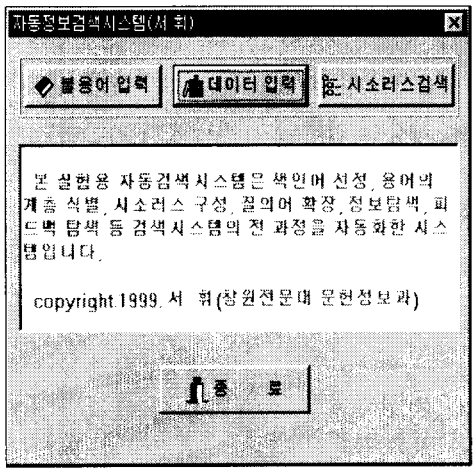
이 처럼 이용자와 시스템 간의 상호작용으로 업무의 수행이 이루어지는 본 시스템은 데이터베이스의 구축을 위해 서지사항과 초록을 입력하는 등록 시스템, 입력된 표제명, 부표제명, 초록 내의 용어를 분석하여 디스크립터를 추출하는 디스크립터 자동 추출 시스템, 추출된 디스크립터를 근거로 용어의 계층을 형성하는 클러스터링 시스템, 입력된 질의어를 근거로 관련 어휘들의 계층을 제시하는 시소러스 브라우저 시스템, 제시된 시소러스를 근거로 검색 용어를 선택하면 이를 이용해 탐색을 수행하는 탐색시스템으로 구성하였다.

1) 초기화면 및 용어리스트 입력

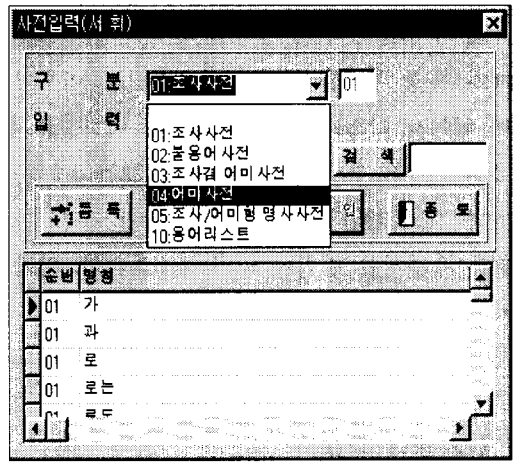
(1) 초기화면의 구성

실험용으로 구축한 본 시스템의 초기화면은 각종 용어리스트 입력, 데이터 입력, 시소러스 검색의 3개 메뉴로 구성되어 있다. 용어리스트 입력 메뉴는 자동으로 디스크립터를 추출하기 위해서 조사, 어미, 불용어 등의 용어들을 등록하거나 수정하는 시스템이며, 데이터 입력 메뉴

는 서지사항과 초록을 입력하고 수정하며, 디스크립터를 자동으로 추출하는 시스템이며, 시소러스 검색 메뉴는 검색을 원하는 질의어를 입력하면, 해당 용어와 관련된 디스크립터를 계층화해 시소러스 형태로 제시해주는 시스템이다. 초기화면은 다음의 <그림 5>와 같다.



(그림 5) 시소러스 자동구축시스템



(그림 6) 용어리스트 입력 및 수정 화면의 초기 화면

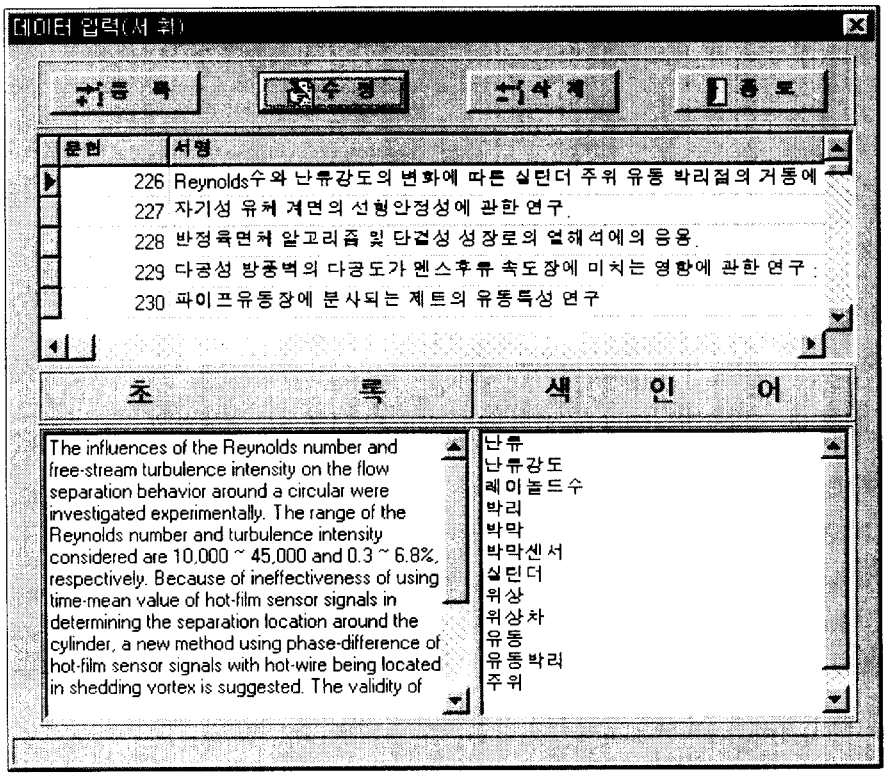
(2) 용어리스트 입력 화면

용어리스트(불용어) 입력화면은 조사, 불용어, 조사 겸 어미, 어미, 조사/어미형 명사, 전문용어 등을 입력하는 화면으로서, 입력한 각 항목들은 조사리스트, 불용어 리스트, 조사 겸 어미 리스트, 어미 리스트, 조사/어미형 명사 리스트, 용어 리스트에 데이터베이스 형태로 구축해 다음의 데이터 입력시 표제명, 부표제명, 초록 내에 출현한 용어(저자가 부여한 색인어)들과 비교하여 자동으로 디스크립터를 추출하는 작업을 담당한다. 용어리스트 입력화면은 앞의 <그림 6>과 같다.

2) 데이터 입력 및 디스크립터 자동 추출

(1) 데이터 입력 화면

데이터 입력의 초기 화면은 등록, 수정, 삭제, 종료 등의 메뉴로 구성하였다. 등록 메뉴는 새로운 데이터를 입력하는 기능을 담당하며, 수정 메뉴는 입력된 데이터를 수정하거나, 자동으로 디스크립터를 추출해 데이터베이스에 등록시키는 기능을 한다. 삭제는 특정 레코드를 제거시키는 역할을 하며, 종료는 초기화면으로 돌아가는 역할을 담당한다. 데이터 입력의 초기 화면은 다음의 <그림 7>과 같다.

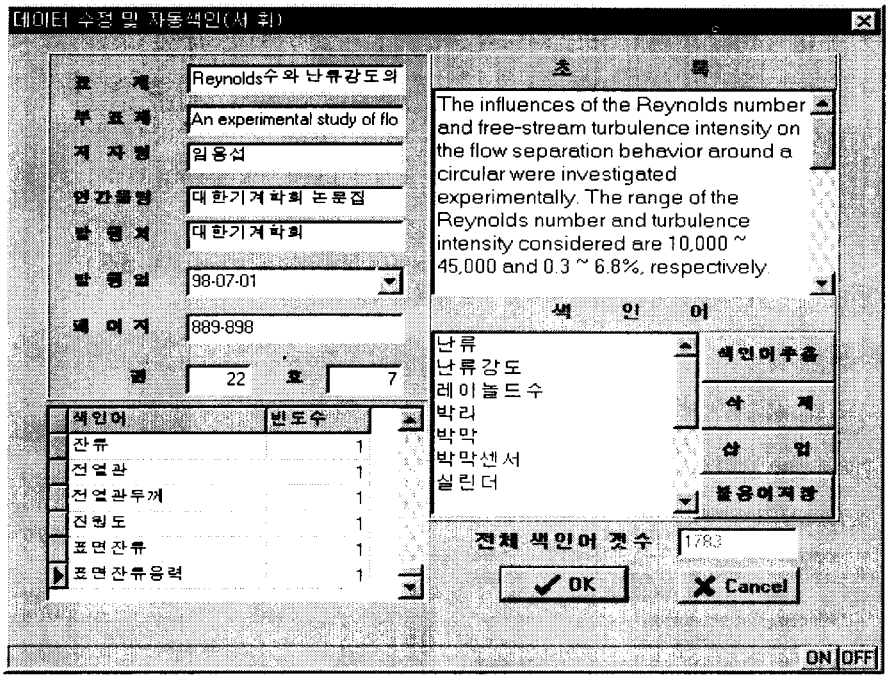


(그림 7) 데이터 입력 초기화면

(2) 데이터 등록, 수정 및 자동색인 추출 화면

데이터 등록 화면은 표제명(기사명), 부표제명(부표제), 저자명, 연속간행물명, 발행사항, 초록, 디스크립터 등의 필드를 입력할 수 있도록 구성되어 있다. 입력방법은 원문을 대상으로 직접 입력하는 방식과 인터넷을 통한 복사 입력방식의 병행이 가능하도록 구축하였다. 또한 디스크립터는 앞에서 입력한 용어리스트들을 근거로 표제, 부표제명, 초록(영문초록과 저자가 부여한 한글 색인어구)을 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 디스크립터 중 잘못 선정된 용어의 삭제 및 추가의 기능을 수행할 수 있도록 구축하였다. 데이터 수정 화면은 다음의 (그림 8)과 같다.



(그림 8) 데이터 수정 및 자동색인 작업 화면

3) 시소러스 브라우저의 구현

(1) 시소러스 브라우저

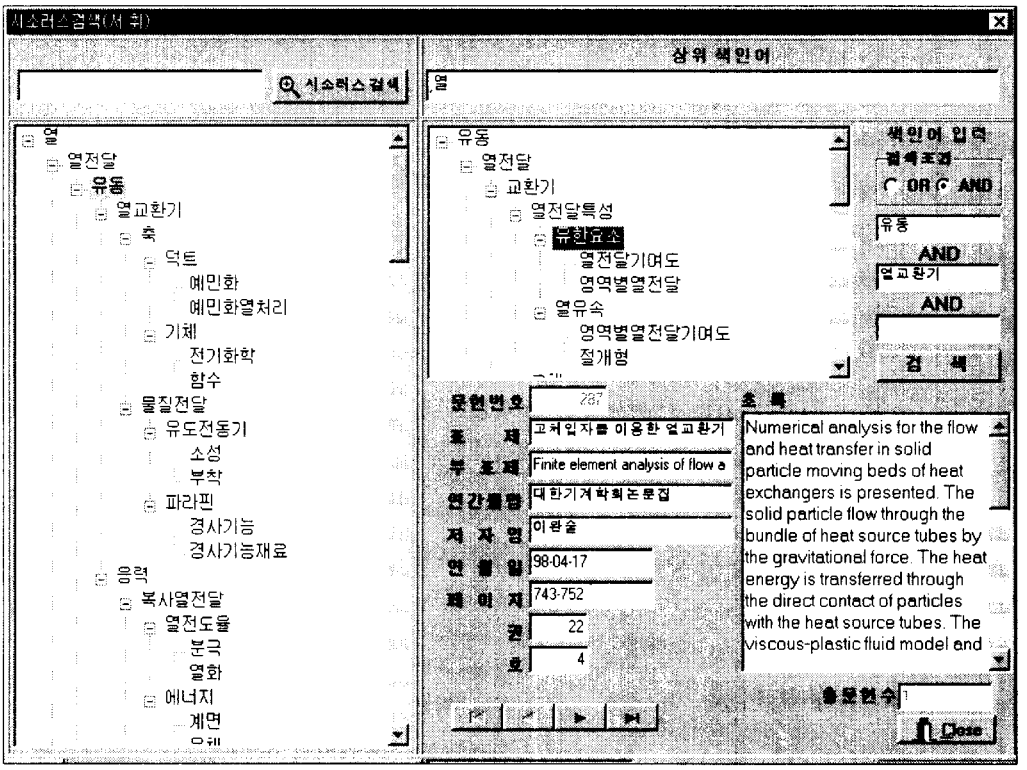
시소러스 브라우저는 (그림 9)와 같이 초기 시소러스 브라우저 화면(좌측 화면), 색인어 입력창과 입력 색인어를 중심으로 한 시소러스 브라우저 화면(우측 화면), 검색 문헌의 출력화면으로 구성하였다.

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

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

(2) 검색결과 출력화면

출력화면 중 초기 화면은 (그림 9)의 왼쪽 창에 제시한 바와 같이 가장 상위어를 중심으로 시소러스의 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.



(그림 9) 시소러스 브라우저 화면

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 디스크립터를 계층화해 제공해 주며, OR의 경우에는 검색어들 중 가장 상위의 용어에 해당하는 디스크립터를 중심으로 관련 디스크립터를 계층화해 시소러스 브라우저 화면에 제시하도록 구성하였다. 그리고 이용자가 제시된 디스크립터 중 핵심 검색어를 선택하면, 이에 해당하는 문헌들을 검색하여 출력하도록 구축하였다.

3. 시스템 평가

1) 평가방법

본 연구에 의해 구축된 시소러스 브라우저 자동구축시스템의 성능에 대한 평가를 수행하였다. 평가방법은 설문지 조사를 택하였고, 평가 항목은 새로운 시소러스가 기계공학 분야 전공용어로서의 적합성, 표현된 계층의 적합성, 이용자 지향적 인터페이스와 검색결과의 정확성 등의 내용으로 구성하여 시소러스 브라우저 자동구축 시스템의 전반적 성능을 검증하였다.
설문 조사는 이용자가 본 시스템을 직접 작동하면서 20개 항목으로 이루어진 설문에 응답하도록 하였다. 평가 결과에 대한 통계처리는 SPSS for windows를 이용해 단순 변인 처리방법으로 수행하였다.

시소러스 자동구축시스템의 성능 평가에 참여한 응답자들은 총 113명이며, 창원전문대학에 소속된 교수 12명(산업공학, 자동차공학, 기계설계 등 기계계열 소속 전공교수), 대학생 70명(산업공학, 자동차공학, 기계설계 등 기계계열 소속 29명, 문헌정보학 전공 41명), 도서관 직원 18명(창원대 소속 6명, 경남대 소속 12명), 창원대학교 기계공학과 석사과정 대학원생 13명 등이다.

설문 응답자 중 전공별 성향은 기계계열 전공자는 54명(46.9%)이며, 문헌정보학 등 기타 전공자는 59명(53.1%)이다. 설문 응답자의 전공을 기계계열과 문헌정보학으로 국한한 이유는 전공용어와 용어들의 계층 평가에는 기계계열 전공자들의 평가가 우수하다고 판단하였고, 검색방법과 시소러스의 개념 파악에는 문헌정보학 전공자들의 평가가 우수하다고 판단하였기 때문이다.

2) 설문조사 결과 분석

구축한 시스템의 성능 평가에 대한 설문 결과에 대한 종합적인 분석결과는 전체를 대상으로 '상당히 그렇다 ' 이상과 '상당히 아니다' 이하만 항목별로 기술한다.(괄호의 앞 수치는 '상당히 그렇다' 이상이며 뒤의 수치는 '상당히 아니다' 이하의 수치임)

먼저 시소러스에 수록된 용어들에 대한 적합성 평가 문항 - 전문용어(89%, 0%), 주제어 기능(92%, 0.9%), 보편성(90%, 0%), 비편협성(78%, 5.4%), 복합명사(87%, 0%) -을 근거로 하였을 때, 비편협성 문항이 상대적으로 저조하나 본 시스템에 의해 구축한 시소러스에 수록된 용어들이 기계공학 분야의 보편적인 전문 용어이며, 복합명사를 처리한 주제어로서의 기능을 갖고 있다고 평가하고 있다. 따라서 용어들의 적합성 평가 항목은 전반적으로 우수하다고 평가되었다.

시소러스의 기능 적합성에 대한 평가 항목 - 용어간의 관련성(95%, 0.9%), 계층성(83%,

3.6%), 계층 단계(77%, 5.3%), 재현율(92%, 0.9%), 정확률(88%, 0.9%) - 을 근거로 하였을 때, 본 시스템에서 구축한 시소러스가 시소러스로서의 관련성, 계층성, 재현율, 정도율의 기능을 충분히 발휘하고 있다고 평가하고 있다. 따라서 본 시스템이 구축한 시소러스 브라우저의 기능이 전반적으로 우수하다고 평가되었다. 단, 계층의 단계를 현재의 6단계에서 더 확장하는 것이 바람직하다는 의견도 있었다.

시소러스의 정보검색 기능에 대한 평가 문항 - 검색어 기능(92%, 0%), 탐색전략 지원(97%, 0%), 비전공자 지원(92%, 3.6%), 속도(78%, 6.2%) -을 근거로 하였을 때, 본 시스템의 정보검색 기능이 속도 문항을 제외하고 탐색용 시소러스의 기능을 충분히 수행하는 우수한 시스템이라고 평가되었다.

시스템의 이용자 지향적 성능에 대한 평가 문항 - 검색의 용이성(89%, 2.7%), 계층 제시 방법(84%, 3.6%), 사용방법의 편리성(94%, 0.9%), 화면 설계(90%, 0.9%) -을 근거로 하였을 때, 본 시스템은 이용자 지향적인 측면을 충분히 고려해 설계되었다고 평가되었다. 또한 시스템의 확장 가능성에 대한 평가 문항 - 온라인 적용성(95%, 0%), 대규모 실험 적용 가능성(92%, 0.9%) -을 근거로 하였을 때, 본 시스템은 대규모의 온라인 탐색엔진에 장착되어 검색의 효율성을 향상시킬 수 있는 시스템이라고 평가되었다.

이상과 같은 평가 결과를 근거로 하였을 때, 본 시스템에 적용된 시소러스 브라우저 자동 구축 알고리즘이 용어간의 관련성과 계층성을 이용한 질의어 확장, 상위어의 재현율과 하위어의 정확률을 이용한 탐색전략의 지원, 이용시 사용방법과 검색의 용이성이란 시소러스 브라우저의 기능을 효과적으로 수행할 수 있으며, 대규모의 온라인 탐색엔진에 장착되어 검색의 효율을 향상시킬 수 있는 알고리즘이라고 평가할 수 있다.

V. 결론 및 제언

클러스터링을 이용한 시소러스 자동 구축에 대한 본 연구의 결과를 요약하면 다음과 같다.

첫째, 전통적인 시소러스의 자동 구축 방법은 수작업으로 구축한 시소러스를 이용하여 사전에 결정된 용어의 계층을 적용시키는 방법으로 구성되나, 본 연구에서 구축한 시소러스는 용어들간의 계층을 미리 정의하지 않는 방법으로 구성하였다.

둘째, 시소러스의 자동 구축을 위해선 클러스터링에 의해 형성된 각 계층을 단일어로 표현하는 방법이 필요하다. 그러나 전통적인 센트로이드 표현 방법은 계층을 복수의 용어로 표현

하고 있어 자동으로 시소러스를 구축할 수 없었다. 본 연구에서는 클러스터의 각 계층을 소수의 용어로 표현하는 알고리즘을 개발하였다.

셋째, 이용자는 검색전략과 기법에 익숙하지 않으므로 데이터베이스 내에 수록되어 있는 정보를 쉽게 검색하지 못하고 있다. 이를 위해서는 질의어의 자동확장, 검색전략 및 검색식의 자동 구축, 탐색 및 피드백 탐색의 자동 수행 기능 등을 갖춘 검색시스템이 필요하다. 본 연구에서는 시소러스의 자동 구축 방법을 통해 이용자 지향적인 검색시스템의 기능을 갖춘 시소러스 브라우저를 구현하였다.

넷째, 시소러스 브라우저는 최신 용어들의 출현과 특정 용어들의 중요도 변화에 의해 검색결과의 차이가 발생할 수 있으므로 주기적으로 그 내용을 갱신해야 한다. 본 연구에서는 데이터가 입력되면, 별도의 노력이 없이도 그 즉시 전체 시소러스의 구조를 새롭게 변경할 수 있는 다이내믹 시소러스의 기능을 갖춘 시소러스 브라우저를 구축하였다.

이상과 같은 결과를 근거로 할 때, 본 연구에서 개발한 시스템의 특징은 용어간의 계층 관계를 자동으로 구축할 수 있도록 설계되었다는 점이다. 따라서 클러스터링을 이용한 용어 계층의 자동 구축 알고리즘에 의해 구축한 시소러스 브라우저는 주제별 용어간의 의미 네트워크를 통한 자동색인, 자동분류 및 자동검색 등에 있어서 필수적인 지식기반 데이터베이스로서의 기능을 발휘할 수 있을 것으로 기대된다.

〈참고문헌은 본문의 각주로 대신함〉