# 자동분류 알고리즘을 이용한 지능형 정보검색시스템 구축에 관한 연구

## A Study of Designing the Intelligent Information Retrieval System by Automatic Classification Algorithm

서 휘(Whee Seo)*

〈 목 차 〉



### 초 록

본 연구의 목적은 이용자의 탐색 행태, 시스템의 정보 구축 행태를 기반으로 초기 질의어의 범주에 해당하는 연관 용어들(해당 용어의 지식구조와 관련된 연관 용어들)을 학습기능을 통해 자동으로 제시해 줄 수 있는 지능형 검색 시스템을 구현하는 것이다. 이를 위해 학습을 통해 전문가 수준의 색인어를 추출할 수 있는 지능형자동색인 알고리즘, 자동분류에 관련한 클러스터링 알고리즘과 문서 범주화 알고리즘 그리고 범주 표현 알고리즘에 대한 이론적 연구를 수행하였으며, 이들 이론적 연구를 근거로 비용과 시간적인 측면에서 그리고 재현율과 정도율이란 측면에서 우수한 성능을 발휘할 수 있는 지능형검색시스템을 구현하였다.

키워드: 자동색인 알고리즘, 자동분류 알고리즘, 범주 표현 알고리즘, 자동정보검색 알고리즘, 지능형 정보검색 시스템

### ABSTRACT

This is to develop Intelligent Retrieval System which can automatically present early query's category terms(association terms connected with knowledge structure of relevant terminology) through learning function and it changes searching form automatically and runs it with association terms. For the reason, this theoretical study of Intelligent Automatic Indexing System abstracts expert's index term through learning and clustering algorism about automatic classification, text mining(categorization), and document category representation. It also demonstrates a good capacity in the aspects of expense, time, recall ratio, and precision ratio.

Keywords: Automatic Indexing Algorithm, Automatic Categorization Algorithm, Category Representation Algorithm, Automatic Information Algorithm, Intelligent Information Retrieval Algorithm

* 창원전문대학 문헌정보과 조교수(drs733m@changwon-c.ac.kr)
• 접수일: 2008년 11월 20일 • 최초심사일: 2008년 11월 25일 • 최종심사일: 2008년 12월 22일

# I. 서 론

## 1. 연구의 필요성과 목적

현재를 살아가는 우리는 인터넷에 수록된 수많은 정보를 활용하여 일상생활에 필요한 다양한 지적 요구를 만족시키고 있다. 우리는 취미활동에 필요한 정보에서부터 의사결정에 필요한 정보에 이르기까지 다양한 정보요구를 인터넷을 통해 해결하고 있다. 특히 인터넷은 과거의 정보독점적인 Know-how의 시대에서 정보개방형인 Know-where 시대로의 이행을 가속화하고 있다.

이상과 같은 정보환경의 변화에도 불구하고 전문적인 정보를 정확히(재현율이나 정도율이란 측면에서 만족할만한 수준으로) 해결하기 위해서는 상당히 많은 시간이나 노력 그리고 비용 등이 요구되는 것이 사실이다. 그 이유는 소수의 전문가들을 제외하고는 특정주제 분야에 대한 지식구조를 제대로 파악하고 있지 못하고 있기 때문이다. 이와 같은 문제점은 인터넷의 수많은 정보검색시스템에서 제공하는 이용자 지향적인 인터페이스가 존재함에도 불구하고 빈번하게 발생하고 있음은 주지의 사실이다. 따라서 본 연구에서는 이와 같은 인터넷의 정보탐색에 대한 문제점을 개선하는 방안을 제안하고자 한다.

그 방법은 이용자의 탐색 행태, 시스템의 정보 구축 행태를 활용한 지능형검색시스템을 구현하는 것이다. 이용자의 탐색 행태에 대한 주지하고 있는 사실은 인터넷을 이용하는 일반 이용자들이 정보탐색을 위해서 최소한 해당 주제 분야에 대한 한 개의 용어(두개의 용어)를 활용하고 있음이다. 또한 이용자는 그 용어가 시스템 입장에서 통제어인지 자연어인지를 알지 못한 채 사용하고 있다는 점이다. 따라서 본 연구에서는 이와 같은 이용자의 탐색 행태에 착안하여 사전에 용어들의 연관성과 계층을 정의하지 않은(non predefined) 방법으로, 한 개의 질의어(용어)를 이용하여 그 용어의 범주에 해당하는 형태가 다른 연관 색인어들(해당 용어의 지식구조와 관련된 연관 색인어들)을 자동으로 제시해 줄 수 있는 시스템을 구축할 것이다. 또한 제시된 연관 색인어들을 이용하여 질의어의 축소와 확장이 가능하며, 이를 통해 자동으로 탐색식을 구축하고 검색 작업을 자동으로 수행할 수 있는 지능형 검색 시스템을 구현할 것이다.

이와 같은 지능형 검색시스템을 구축하기 위해선 귀납학습 방법을 통한 자동색인 알고리즘, 자동분류 알고리즘, 범주 표현 알고리즘, 자동정보검색 알고리즘들이 결합되어야 가능할 것이다. 따라서 본 연구에서는 이들 각 알고리즘에 대한 선행 이론들을 조사하고 본 연구의 목적에 부합되는 알고리즘을 선택해 이를 적용한 지능형 검색시스템을 구축할 것이다.

### 2. 연구의 내용과 방법

본 연구는 다음과 같은 내용과 방법으로 수행하였다.

첫째, 자동분류 알고리즘을 이용한 지능형검색시스템을 구현하기 위하여 문헌 클러스터링과 문헌범주화 알고리즘에 관련된 선행 연구를 분석하였다.

둘째, 성능이 우수한 지능형검색시스템을 구현하기 위하여, 한글의 특성을 주입시킨 한글자동색인 알고리즘, 범주 표현 알고리즘, 정보검색 알고리즘에 관련된 선행 연구를 분석하였다.

셋째, 구축된 지능형검색시스템의 성능을 실험하기 위하여 대한기계학회 논문집의 '열 및 열 유체' 분야의 190편의 기사와 인지과학논문집에서 '인지, 의미, 언어'와 관련된 27편의 기사로 이루어진 실험데이터를 구성하였다.

넷째, 실험데이터는 표제명, 부표제명, 초록을 대상으로 자동색인시스템을 이용하여 색인어를 추출하였으며, 문헌 클러스터링 알고리즘과 범주표현 알고리즘을 적용해 입력데이터에 대한 범주화와 범주의 대표어를 자동으로 구성하고 선정하도록 하였다.

다섯째, 주제 분야가 상이한 실험데이터를 입력해서 본 연구에서 적용한 알고리즘에 의해서 범주화가 분리되어 구성되는지를 실험하였다.

여섯째, 실험데이터를 대상으로 초기질의어와 형태가 다른 연관 색인어들이 자동으로 제시되고 이들 연관 색인어들을 선택함에 의해 초기질의어와 연관 색인어가 결합된 탐색식의 자동 구성과 탐색 수행이 가능한지를 실험하였다. 또한 탐색결과의 정확성 여부를 확인하기 위해 해당 데이터의 원문을 비교함에 의해 탐색식에 포함된 용어들이 수록되어 있는지를 조사하였다.

본 연구에서 제시한 지능형정보검색시스템의 구현 방법과 이에 대한 검증은 소규모의 데이터로 이루어진 실험환경에서 이루어졌기 때문에 대규모의 실험환경에서의 검증을 통해 본 연구의 결과가 더욱 일반화될 수 있도록 계속되어야 할 것이다.

## Ⅱ. 이론적 연구

### 1. 자동색인 알고리즘

정보검색시스템은 입력 질의어와 문서 내용에 대한 색인어의 형태적 일치를 검사함으로써 적합문서 여부를 결정한다.[1] 따라서 입력 질의어와 일치하는 형태의 색인어를 문헌(문서)에서 추출하는 일은 적합성의 여부를 결정짓는 매우 중요한 작업이다. 이 같은 색인 작업은 인간의 수작업 방법,

1) 강승식, "한글 문서의 색인어와 색인 기법," 정보과학회지, 제22권 제4호(2004. 12), pp.72-77.

컴퓨터에 의한 자동색인방법, 인간과 컴퓨터가 협력하여 작업을 수행하는 반자동색인방법이 있다.

그 중 자동색인이라 함은 컴퓨터를 통해 입력된 문헌을 대표하는 색인어를 자동으로 추출하는 방법을 의미한다. 컴퓨터에 의한 자동색인은 통제어 기반 색인법과 일반 색인 기법(단일어 색인 기법)으로 나뉘며, 일반 색인 기법은 색인어를 선정하는 기준에 따라 통계적 기법, 언어학적 기법, 문헌구조적 기법으로 나뉘어진다. 대부분의 한글 자동색인 방법은 언어학적 기법을 이용하여 색인의 대상이 되는 명사나 명사구를 식별하고, 통계적 기법을 이용하여 식별된 명사나 명사구를 색인어로 적용시키는 방법을 채택하고 있다.[2)3)4)] 그 이유는 강현규의 한국어 자연어 질의 문장 사용 행태에 대한 조사 결과와 같이 한국어의 질의어는 거의 명사 유형이기 때문이다.[5)]

언어학적 기법은 어휘적 단계, 구문적 단계, 어의적 단계로 구분하며, 어휘적 단계기법은 불용어 제거 기법을 의미한다. 또한 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.

## 2. 자동분류 알고리즘

자동분류라 함은 인간이 특정 문헌의 내용을 충분히 이해하지 않고도 해당 문헌에 대한 범주(주제)를 쉽게 알아내는 능력을 컴퓨터에 학습시키고 이를 통해 분류업무를 자동화할 수 있는 방법이다.[6)7)8)9)]

이와 같은 자동분류는 학습방법에 따라 문헌범주화(text categorization) 알고리즘에 의한 방법과 문헌 클러스터링(clustering) 알고리즘에 의한 방법으로 구분된다.[10)] 문헌범주화 알고리즘은

2) 서휘, "클러스터링을 이용한 시소러스 브라우저의 설계에 관한 이론적 연구," 한국도서관·정보학회지, 제30권 제3호(1999. 9), pp.427-456.
3) 서휘, "자동정보검색을 위한 한글 시소러스 브라우저 구축에 관한 연구," 한국도서관·정보학회지, 제31권 제2호(2000. 6), pp.279-302.
4) 서휘, "자연어를 이용한 자동정보검색시스템 구축에 관한 연구," 한국문헌정보학회지, 제35권 제4호(2001. 12), pp.141-160.
5) 강현규, "개념 검색어 확장을 통해 질의 형식화를 도와주는 개념 마법사의 설계 및 구현," 정보처리학회논문지, 제9-B권 제4호(2002. 12), pp.437-444.
6) F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, Vol.34, No.1 (2002. 3), pp.1-47.
7) 심경, "문헌범주화에서 학습문헌수 최적화에 관한 연구," 정보관리학회지, 제23권 제4호(2006. 12), pp.277-294.
8) Dumais, Susan et al. "Inductive learning algorithms and representations for text categorization," Proceedings of ACM-CIKM 1998. pp.148-155.(online). 〈http://research.microsoft.com/~sdumais/cikm98.doc〉 [cited 2008. 10. 29].
9) 정영미, 임혜영, "SVM분류기를 이용한 문서 범주화 연구," 정보관리학회지, 제17권 제4호(2000. 12), pp.229-248.
10) Peter, Jakson, & Isabelle Moulinier. Natural Language Processing for Online Applications : Text Retrieval, Extraction and Categorization. Amsterdam : John Benjamins Publishing Co.

학습데이터가 이미 분류되어 있으며 분류마다 범주표시(labels)가 결정되어 있다는 의미에서 지도학습(supervised learning)에 속하며, 문헌 클러스터링 알고리즘은 학습데이터의 분류(범주 표시)가 사전에 결정되어 있지 않다는 의미에서 자율학습(unsupervised learning)에 속한다.[11][12]

따라서 분류나 범주화 표시가 되어 있지 않은 웹 데이터에 대한 이용자 지향적인 효과적인 검색시스템을 구축하기 위해서는 시스템에서는 자율 학습에 해당하는 클러스터링 알고리즘을 이용해 자동분류를 수행해야 하며, 이용자의 정보탐색 작업을 위해서는 - 클러스터링 알고리즘에 의해 형성된 분류(범주) 표시를 근거로 한 지도학습 방법을 수행할 수 있도록 -문헌범주화 알고리즘을 이용한 자동분류를 수행해야 할 것이다. 물론 웹 데이터가 사전에 분류(범주화)가 되어 있다면 해당 데이터에 대한 분류 방법으로 문헌범주화 알고리즘의 적용이 가능하겠으나 이용자와 시스템에서 사용하는 용어의 차이로 인하여 검색효율이 낮아질 수 있는 문제점이 발생할 것이다. 그 이유는 문헌범주화 알고리즘은 사전에 정의된(predefined) 범주정보를 활용하기 때문이다.

### 가. 자동분류 알고리즘의 종류

(1) 문헌 범주화 알고리즘

문헌 범주화 알고리즘은 앞에서 거론했듯이 사전에 정의된(predefined) 분류(범주) 정보가 있을 경우에 타당한 방법이다. 그럼에도 불구하고 이 알고리즘을 적용해 자동분류를 수행할 경우에는 시스템 내의 색인어와 이용자의 탐색어가 일치하지 않는 경우가 발생하여 검색 효율이 저하되는 문제점이 발생할 수 있다.

문헌범주화에 적용이 가능한 알고리즘은 다중회귀모형(multi-variate regression models), K-최근접 이웃 분류기(K-nearest neighbor classifiers), 네이브 베이지언 모형(Naive Bayesian models),[13][14][15] 결정트리(decision trees), 신경망(neural networks), SVM(Support Vector Machines : 지지벡터기) 알고리즘[16] 등이 있다. 본 연구에서는 가장 널리 알려져 있으며, 클러스터링 알고리즘에 의해 형성된 클러스터를 계층화하는데 적합한 방법인 K-최근접 이웃 분류기에 대해서만 소개하기로 한다.

---

11) B. Liu, Y. Dai, X. Li, W. S. Lee & P. S. Yu, "Building text classifiers using positive and unlabeled examples," Proceedings of the Third IEEE International Conference on Data Mining(ICDM-03), pp.179-188.

12) 김판준, 이재윤, "문헌간 유사도를 이용한 자동분류에서 미분류 문헌의 활용에 관한 연구," 정보관리학회지, 제24권 제1호(2007. 3), pp.251-271.

13) 고영중, 서정연, "문서관리를 위한 자동문서범주화에 대한 이론 및 기법," 정보관리연구, 33권 제2호(2002. 6) pp.19-32.(online). 〈http://nlp.sogang.ac.kr/pub/domestic/d02-jb002.pdf〉 [cited 2008. 10. 29].

14) 국민상, 정영미, "자질 선정에 따른 Naive Bayesian 분류기의 성능비교," 한국정보관리학회 학술대회논문집, 제7권 (2000.8), pp.33-36.

15) Susan. Dumais, et al. *op. cit.*, pp.148-155.(online). 〈http://research.microsoft.com/~sdumais/cikm98.doc〉 [cited 2008. 10. 29].

16) 정영미, 임혜영, 전게논문, pp.229-248.

K-최근접 이웃 분류기는 예제 기반 범주화 방법 중 가장 대표적인 것이다. K-최근접 이웃 분류 방법은 입력 문헌(문서나 질의어)가 주어졌을 때 학습 문헌(전체 문헌 집단 또는 데이터베이스) 중에서 입력 문헌과의 유사도가 가장 높은 K개의 문헌을 추출하고 그들을 사용하여 각 후보 범주의 순위를 매기는 방법이다. 여기서 K개의 추출된 학습문헌은 미리 정해진 범주가 있으므로, 각 범주와 입력 문헌과의 유사도는 각 범주별로 추출된 K개의 문헌과 입력 문헌과의 유사도의 합으로 계산된다.[17]

Yang이 소개한 최근접 이웃 분류 알고리즘의 수식은 다음과 같다.[18]

$$\mathrm{rel}(C_k|D_x) = \sum_{D_j \in k\text{개의 상위문헌}} \mathrm{Sim}(D_x, D_j) \cdot \mathrm{P}(C_k|D_j)$$

$\mathrm{rel}(C_k|D_x)$ = 신규문헌의 특정범주에 대한 적합성 척도

$$\mathrm{P}(C_k|D_j) = \frac{\text{범주 } C_k\text{가 문헌 } D_j\text{에 할당된 빈도}}{\text{학습집단에서 문헌 } D_j\text{가 출현하는 빈도}}$$

$\mathrm{Sim}(D_x, D_j)$ = 입력 문헌 $D_x$와 학습문헌 $D_j$간의 유사도 값 (2개 문헌 벡터간의 코사인 유사도 값)

위의 수식을 상세히 설명하면 다음과 같다. 학습문헌들($D_j$)로부터 각 문헌의 특성을 표현하고 있는 색인어를 추출하고 학습 문헌 및 각 학습문헌에 부여된 범주($C_x$)들을 벡터로 표현한 후, 입력문헌($D_x$)과 가장 유사한 K개의 학습문헌(k개의 $D_j$)을 찾아 그 문헌들에 이미 할당된 범주정보($C_k|D_j$)를 이용하여 입력문헌의 범주를 결정한다.[19]

K-최근접 이웃 분류기는 위의 수식을 이용하여 입력 문헌과 각 범주에 포함된 문헌들을 비교하여 가장 유사도가 높은 범주에 입력 문헌을 할당하는 방법을 택하고 있다. 그 방법은 입력문헌과 가장 유사한 K개 문헌들의 유사도 및 범주빈도(각 범주 당 학습문헌의 분류빈도, 한 학습문헌이 해당범주에 분류된 경우에는 $\mathrm{P}(C_k|D_j) = 1$, 아닌 경우에는 $\mathrm{P}(C_k|D_j) = 0$)를 합산(총 분류빈도)하여 그 값이 높은 범주를 시스템이 지정한 수 만큼 차례대로 입력문헌의 범주로 할당하는 방법을 사용하고 있다.

(2) 문헌 클러스터링 알고리즘

클러스터링 알고리즘은 비계층적 알고리즘(nonhierachical Clustering Algorithm)과 계층적

17) 고영중, 서정연(2002), 전게논문, pp.19-32.(online). 〈http://nlp.sogang.ac.kr/pub/domestic/d02-jb002.pdf〉 [cited 2008. 10. 29].

18) Y. Yang, "Expert Network : Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval," SIGIR'94. 13-22.

19) 이영숙, 정영미, "KNN 분류기의 범주할당 방법 비교 실험," 한국정보관리학회 학술대회논문집, 제7권(2000.8), pp.37-40.

클러스터링 알고리즘(Hierarchical Clustering Algorithm)으로 구분된다.[20] 비계층적 알고리즘은 문헌간의 계층을 형성하지 않으므로 자동분류방법으로는 적합하지 않아 설명을 생략하기로 한다. 계층적 클러스터링 알고리즘은 클러스터 대상물 간의 유사성(용어들의 동시 출현 빈도)을 측정하여 작성한 문헌-문헌 유사행렬을 이용하여 클러스터를 구성하는 방법이다. 계층적 클러스터링 알고리즘의 종류는 단일연결(single link), 완전연결(complete link), 그룹평균연결(group average), Ward 알고리즘 등이 있다. 본 연구에서는 클러스터의 센트로이드 표현과 클러스터들의 이웃 간 관계를 형성할 수 있는 그룹평균 연결 알고리즘과 Ward 방법만 소개하기로 한다.

(가) 그룹 평균 연결 알고리즘

그룹 평균 연결 알고리즘은 클러스터 대상 문헌 전체의 유사도의 평균 값을 근거로 클러스터를 형성하는 방법이다. 모든 객체들은 클러스터 간 유사성에 관련되어 있으므로 느슨한 형태의 단일 연결 클러스터와 견고한 형태의 완전연결 클러스터와 비교하여 중간 형태의 구조를 나타낸다. 따라서 이 알고리즘은 단일 연결이나 완전 연결 알고리즘에서 요구하는 것처럼 $O(N^2)$의 시간과 $O(N)$의 기억장소가 적용되지 않는다. 그 이유는 클러스터를 형성하는 센트로이드와 특정 문헌 간의 유사성이 전체 그룹평균값과 일치하기 때문이다. 이와 같은 이유에서 센트로이드는 모든 문헌 벡터들의 평균이므로, 중심 값은 $O(N)$의 기억장소만을 요구하며, 이 센트로이드가 클러스터들간의 유사성을 계산하는데 이용된다. 그러나 이 방법에서도 클러스터를 표현하는데 센트로이드를 이용하지 않고 출현한 모든 용어들로 표현하는 방법을 택한다.

(나) Ward 알고리즘

Ward 알고리즘은 최소분산방법으로 알려져 있는데, 그 이유는 각 단계에서 클러스터 쌍을 결합할 때, 문헌간의 거리를 유클리디안(euclidean)거리를 사용하여 - 비유사성 값을 사용하여 최소값을 갖는 것만을 연결하는 방법을 택하기 때문이다. 따라서 이 방법의 수학적 특성은 RNN(상호 근접 이웃 : reciprocal nearest neighbor)알고리즘의 적용이 가능하다. 따라서 이 방법은 어떤 클러스터나 근접 이웃(NN : nearest neighbor)이 존재하므로 소수의 객체쌍으로 이루어진 RNN까지도 구성할 수 있다.

이 방법은 대칭적 계층과 동질 클러스터를 만드는 경향이 있고, 클러스터의 무게 중심에 대한 정의는 클러스터를 새롭게 표현할 수 있는 가능성을 제시하고 있다. 이 방법은 클러스터 구조를 회복하는 데에는 좋으나 분리된 클러스터에 민감하고 늘어난 클러스터를 원상으로 회복하는 데는 부적합하다. RNN알고리즘 역시 $O(N^2)$의 시간과 $O(N)$의 기억장소에 대한 요구사항을 만족시킨다.

---

20) Gerald Salton, *Dynamic Information and Library Processing*(New-jersey : Prentice-Hall, 1975), p.329.

## 3. 범주 표현 알고리즘

범주의 표현 방법은 문헌범주화 알고리즘을 사용하던, 클러스터링 알고리즘을 사용하던 동일한 방법을 사용한다. 일반적으로 범주의 표현은 범주를 이루는 벡터 자질을 표현하는 것을 의미한다. 그 표현 방법은 범주 또는 클러스터에 대한 벡터 유사도를 제시하는 것이다. 범주를 벡터로 표현하기 위해서는 문헌-문헌 유사도 행렬을 이용해서 각 문헌에 포함되어 있는 색인어가 얼마나 함께 출현하는가를 벡터유사도 값으로 표현하는 것이다.

그러나 한 개의 범주를 표현하기 위해서 - 해당 범주에 포함된 모든 문헌들을 각기 특징지우는 색인어를 모두 이용해서 - 벡터 값으로 표현하는 것은 실용적인 측면에서- 속도나 비용적인 측면에서 많은 무리가 따르는 것이다. 그래서 범주벡터를 핵심적으로 표현하는 범주 센트로이드 벡터가 요구되는 것이다.[21][22]

범주의 대표어를 선정하는 방법은 단락빈도 이용방법, 단어빈도와 역단락빈도의 곱을 이용하는 방법, 클러스터 센트로이드를 이용하는 방법 등이 있다.[23] 단락빈도 이용방법은 특정 클러스터에 속한 색인어들 중에서 단락빈도가 가장 높은 색인어를 문헌 전체에서 주제적으로 의미가 있는 것으로 보고 해당 클러스터의 대표어로 선정하는 방법이다. 단어빈도와 역단락빈도의 곱을 이용하는 방법은 특정 색인어가 여러 개의 클러스터에 고르게 출현하는 것보다 특정 클러스터에 집중적으로 출현하는 것이 해당 클러스터를 대표하는 색인어라는 가정 하에 이를 대표어로 선정하는 방법이다. 그 수식은 다음과 같다.

$$cw = \ln(cf) \times \frac{cf}{df}$$

클러스터 센트로이드를 이용하는 방법은 클러스터에 속한 문헌들간의 유사도를 계산하여 유사도의 기준이 되는 색인어들을 클러스터의 대표어로 선정하는 방법이다. 한승희와 정영미는 대표어를 선정하는 실험에서 위에 제시한 세 가지 방법 중 단락빈도를 이용하는 방법이 재현율이나 정도율이란 측면에서 가장 우수하다고 주장하고 있다.

서휘는 Yu(1974)의 "이용자들이 일정 주제의 정보를 요구할 때 그 주제에서 출현빈도가 높은 용어를 이용해 정보를 검색할 것이다."란 가설[24]을 이용하여 다음과 같은 단락빈도를 이용한 범주

21) 서휘(1999), 전게논문.

22) 이재윤, "문헌간 유사도를 이용한 SVM 분류기의 문헌분류성능 향상에 관한 연구," 정보관리학회지, 제22권 제3호(2005. 9), pp.261-287.

23) 한승희, 정영미, "클러스터링 기법을 이용한 개별문서의 지식구조 자동 생성에 관한 연구," 정보관리학회지, 제21권 제3호(2004. 9), pp.251-267.

24) T. Yu Clement, "A Clustering Algorith Based on User Queries," *JASIS*, Vol.25, No.4(1974), pp.218-226.

별 센트로이드를 표현하는 알고리즘을 제안하였다.[25][26][27] 유의 알고리즘은 "이용자들이 일정 주제의 정보를 요구할 때 그 주제에서 출현빈도가 높은 용어를 이용해 정보를 검색할 것이다."란 가설에서 출발한다.[28]

## 4. 자동정보검색 알고리즘

자동정보검색이란 질의어 선정을 위한 사전탐색, 선정된 질의어들과 부울린 로직의 조합을 이용한 검색전략(검색식) 구축, 검색된 결과에 대한 평가를 근거한 피드백 탐색(사후 탐색) 등의 모든 과정이 결합된 검색방법을 의미한다.

정보탐색방법의 종류는 크게 부울린 탐색(Boolean search)과 매칭 함수(Matching functions)에 의한 탐색 방법으로 나누어진다. 부울린 탐색은 and, or, not 등의 연산자를 근거로 정보를 검색하는 방법이며, 매칭 함수(matching functions)를 이용한 탐색은 질의를 문헌이나 클러스터와의 연관성을 근거로 - Dice 계수, cosine 계수나 Tanimoto 계수 등의 연관성 측정법(association measure)을 적용해 일정 기준치를 통과하는 문헌만을 적합정보라고 판단하여 검색하는 방법이다. 매칭함수에 의한 탐색은 순차탐색(serial search), 클러스터탐색(clustered based search)이 존재한다.[29]

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. 적합성피드백에 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.[30][31]

---

25) 서휘(2000), 전게논문, pp.279-302.
26) 서휘(2001), 전게논문, pp.141-160.
27) 서휘(1999), 전게논문.
28) Gerald Salton, *Dynamic Information and Library Processing*(New-Jersey : Prentice-Hall, 1975), pp.353-357.
29) Van Rijisbergen, C. J. The Hyper-Textbook of the C. J. Van Rijsgergen's textbook on Information Retrieval. 1998. 〈http://www.dei.unipd.it/~melo/bible/〉 [cited 2008. 11. 10].
30) Helen J. Peat, and Peter. Willett, 1991. "The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems," *JASIS*, Vol.42 No.5(1991), pp.378-383.
31) 노정순, 1999, "탐색결과에 근거한 자연어질의 자동확장 및 응용에 관한 연구 고찰," 정보관리학회지, Vol.16 No.2 (1999. 6), pp.49-80.

이상과 같은 다양한 정보검색방법이 존재함에도 불구하고 정보검색은 어려운 일이다. 그 이유는 동의어나 복수 의미어에 의한 정확한 의미 파악이 어려우며, 질의어의 불확실성, 문헌 표현이 정확하지 않고 불충분하며 적합성 판단이 어렵기 때문이다. 그러므로 지능형 검색이 요구되는 것이다. 지능형 검색이라 함은 사용된 단어의 의미, 질의어의 순서 그리고 정보원의 신뢰성을 고려한 검색방법을 의미한다.

지능형 검색 알고리즘을 적용한 학습기반 지능형 검색시스템은 다음과 같은 기능을 갖고 있는 시스템을 의미한다.

첫째, 이용자의 직접 혹은 간접적 피드백을 통해 이용자의 요구를 정확히 파악하고 이용자 요구에 충실한 정보를 검색하도록 향상된 시스템 기능을 제공한다.

둘째, 정보검색을 쉽게 수행하도록 하기 위하여 기계학습 알고리즘을 이용해 문헌집합을 재구성하는 기능을 갖고 있는 시스템을 의미한다.

셋째, 자연언어 처리를 이용하여 문헌을 정확히 표현하여 검색결과를 향상시키거나 정보를 추출한다.

본 연구에서는 앞에 설명한 자동색인 알고리즘, 문헌 클러스터링 알고리즘, 문헌 범주화 알고리즘, 범주 표현 알고리즘을 이용해 이용자의 초기 질의어를 근거로 관련 색인어들을 제시해주고 이들 색인어끼리 부울린 로직의 AND 연산자를 이용해 자동으로 탐색식을 구성하고 원하는 정보를 검색할 수 있는 이용자 지향적인 정보검색시스템을 구현하였다.

## Ⅲ. 지능형 정보검색시스템의 구현

본 시스템의 실험 환경은 PC이며, 개발 Tool은 Paradox 7.0 DBMS를, 개발 언어는 델파이(Delphi 4.0 - PASCAL)를 사용하였다. 입력 데이터는 대한기계학회 논문집의 '열 및 열 유체' 분야의 190편의 기사와 인지과학논문집에서 '인지, 의미, 언어'와 관련된 27편의 기사에 수록된 표제, 부표제, 초록 등을 대상으로 하였다.

실험 데이터의 주제가 상이한 데이터를 사용한 이유는 다양한 주제를 수록하고 있는 웹에서도 본 시스템에서 사용하고 있는 알고리즘이 적용이 가능한지의 여부를 분석하기 위함이었다. 또한 주제가 상이함에 따라 동일한 형태의 용어가 다른 의미로 사용될 수 있음으로 연관 용어를 제시해 줌에 의해 해당 주제의 문헌을 탐색할 수 있는 가능성을 분석하기 위함이었다.

지능형정보검색시스템의 구현은 앞의 이론적 배경에서 설명한 자동색인 알고리즘, 문헌 클러스터링 알고리즘, 문헌 범주화 알고리즘, 범주 표현 알고리즘, 자동검색 알고리즘 등을 결합하여 구축하였다. 본 장에서는 지능형 정보검색시스템의 구축 과정을 실제로 작업한 순서대로 용어사전 데이

터베이스 구축, 색인어 추출, 클러스터(범주) 형성, 클러스터(범주) 대표어 선정, 지능형정보검색 등의 순으로 소개한다.

## 1. 용어사전 데이터베이스 구축

용어사전 데이터베이스는 〈그림 1〉과 같이 자동색인 작업을 위하여 한글어 특성별로 구축된 데이터베이스를 의미한다. 본 연구에서는 자동색인 작업을 시작하기 전에 조사리스트, 불용어리스트, 조사 겸 어미리스트, 어미리스트, 조사/어미형 명사리스트 등의 형태별 데이터베이스 구축의 초기 작업을 남영신의 우리말 분류사전을[32] 참고하였다.

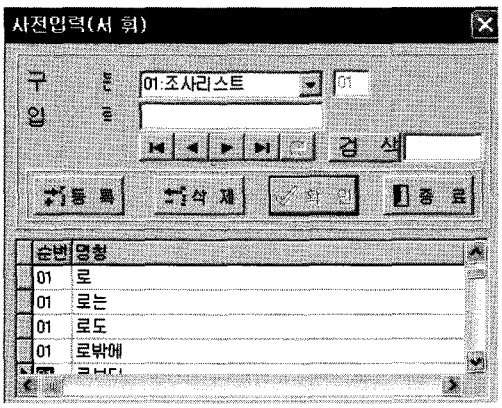

〈그림 1〉 자동색인용 형태별 용어사전 데이터베이스

조사리스트는 ~로부터, ~와, ~하고 등의 용어들로 이루어진 데이터베이스이며, 조사 겸 어미리스트는 ~는다, ~기도, ~하여 등의 용어들로 이루어진 데이터베이스이다. 어미리스트는 ~는, ~다, ~라도 등의 용어들로 이루어진 데이터베이스이며, 조사/어미형 명사리스트는 ~기계, ~기법, ~단계 등의 용어들로 구성된 데이터베이스이다. 또한 불용어 리스트는 $, # 등의 기호, 영어 대소문자, 숫자, *나라, *아래, 설사, 세계* 등과 같이 색인어의 역할을 수행할 수 없는 용어로 구성된 데이터베이스이다. 이들 중 조사리스트 데이터베이스에 수록된 사례는 〈그림 2〉와 같다.

본 연구에서는 자동색인의 과정에서 발생하는 한글 특성에 따른 새로운 용어들을 즉각적으로 각 용어사전 데이터베이스에 수록할 수 있도록 하기 위하여 〈그림 2〉와 같이 입력 창에 새로운 용어를 입력하고 등록, 삭제, 확인 등의 버튼을 이용하여 추가할 수 있는 기능을 제공하고 있다. 또한 입력된 각 용어들이 수록되었는지의 확인을 위하여 검색창과 검색버튼의 기능을 제시하였다.

32) 남영신 편, 우리말 분류사전(3) : 꾸밈씨 기타 편(서울 : 한강문화사, 1992).

이를 통해 실수에 의해 잘못 입력된 용어의 삭제나 새로운 용어의 추가가 수시로 가능하도록 구성하였다. 본 연구에서는 자동색인 시스템을 이용하여 88개의 용어로 구성된 조사리스트, 86개의 용어로 구성된 조사 겸 어미리스트, 264개의 용어로 구성된 어미리스트, 762개의 용어로 구성된 불용어리스트를 새롭게 구축하였다.



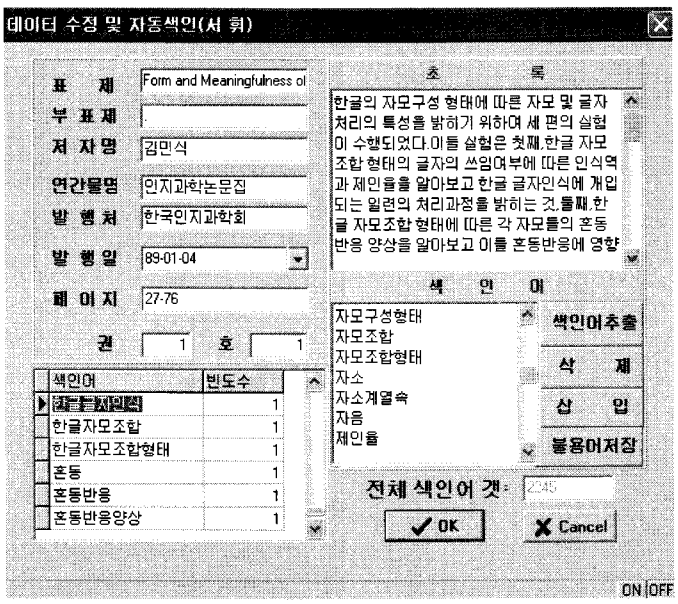

〈그림 2〉 자동색인용 조사리스트 데이터베이스

그리고 용어리스트는 〈그림 1〉과 같이 기계과학과 인지과학과 관련된 전공주제의 용어들로 구성된 데이터베이스인데, 기계과학 관련 용어들은 원문에 주어진 저자가 부여한 색인어를 근거로 구축하였으며, 인지과학 관련 용어들은 다양한 주제에 관련된 용어들이므로 사전에 관련 용어들을 입력하지 않았다. 본 시스템에서는 자동색인 작업에 의해 추출된 용어들을 출현용어리스트에 자동으로 수록되도록 하였으며, 이들 용어들은 용어리스트에도 자동으로 추가되도록 구성하였다(출현용어리스트에 등록된 용어만이 색인어이므로 그 이전까지의 작업에서 다루는 용어는 '용어'로 표현하겠음). 본 연구 과정 중 추출되어 출현용어리스트에 수록된 총 색인어 수는 〈그림 3〉에 나타난 바와 같이 2,245개이다.

본 연구에서는 이와 같이 각종의 용어사전데이터베이스를 이용하여 입력된 원문에 대한 자동색인 작업을 수행하였으며, 이 과정에서 새롭게 출현한 특성별 용어들을 수작업으로 확인하고 이를 해당 데이터베이스에 추가하는 과정을 통해 실시간으로 갱신되도록 하였다. 이와 같은 반복적인 작업을 통하여 새롭게 입력되는 원문에서 출현하는 동일한 형태의 각종 용어들은 이들 용어사전데이터베이스에 의해 자동으로 처리되도록 구성하였다.

## 2. 색인어의 자동 추출

색인어는 자동색인 알고리즘을 적용해서 구축한 자동색인시스템을 이용하여 추출하였다. 본 연구에서의 색인어 데이터베이스는 출현용어리스트 데이터베이스에 해당된다. 입력 문헌의 데이터베이스는 〈그림 3〉과 같이 표제, 부표제, 저자명, 연간물명, 발행처, 페이지, 초록, 색인어 등의 필드로 구성하였다. 자동색인 작업은 〈그림 3〉의 '색인어 추출' 버튼을 클릭함에 의해 시작되며, 입력 데이터 중 표제, 부표제 그리고 초록내의 문장들을 기 구축된 한글 기능에 의해 구분된 형태별 용어사전 데이터베이스에 수록된 용어들과 비교하는 과정을 통해 자동으로 색인어가 추출되도록 구성하였다.




〈그림 3〉 데이터 수정 및 자동색인 작업

〈그림 3〉에서의 색인어를 자동으로 추출하는 과정은 색인어추출 버튼을 클릭함에 의해 시작되며 추출된 용어 중에서 불용어에 해당하는 용어는 선택한 후 불용어 저장 버튼을 클릭하면 자동으로 불용어 리스트에 수록되도록 구성하였다. 또한 초록의 원문과 비교하여 색인어로 추출되지 못한 용어는 색인어 창에 직접 입력하고 삽입 버튼을 클릭함에 의해 출현용어리스트와 용어리스트에 수록되도록 해 새롭게 입력되는 원문에서는 자동으로 해당 용어들을 색인어로 추출할 수 있도록 구성하였다. 그리고 OK와 Cancel 버튼을 이용해 해당 데이터의 색인어가 잘 못 입력되는 오류를 줄일 수 있도록 구성하였다.

본 연구에서 구축한 자동색인 시스템에서 입력 데이터들이 〈그림 2〉와 〈그림 3〉의 과정을 거쳐 〈그림 1〉에 제시된 다양한 용어 사전 데이터베이스에 새로운 용어들을 추가하는 과정을 순서대로 소개하면 다음과 같다.

첫째, 공백기호에 의해 용어구를 추출한다.

둘째, 용어구에 어절 분리기호와 불용어리스트를 비교하여 새로운 용어구를 추출한다.

셋째, 용어구와 출현용어리스트(색인데이터베이스)와 비교하여 색인어를 추출한다.

넷째, 출현용어리스트와 일치하지 않는 용어구(특히 4음절 이상의 용어)는 명사 형태로 이루어진 용어리스트와 비교해 일치할 경우에는 용어리스트 중에서 복합명사 리스트와 비교해 색인어를 추출한다.

다섯째, 복합명사 리스트와 일치하지 않는 경우에는 용어리스트와 비교해 용어의 일부분과 일치하는지의 여부를 판단해 단일명사 형태의 색인어를 분리한다.

여섯째, 앞의 넷째 과정에서 용어리스트와 일치하지 않는 용어는 조사, 어미, 접미사 등이 추출된 명사 형태의 용어이므로 복합명사 리스트와 비교하고 단일명사와 비교하는 과정을 거친다.

일곱째, 여섯째 과정에서 새롭게 출현한 비명사 형태의 용어들은 그 성격에 맞게 조사, 어미, 조사·어미형 명사 리스트 등에 새롭게 등록한다.

여덟째, 앞의 여섯째 과정에서 새롭게 출현한 명사와 복합명사는 용어리스트와 출현용어리스트에 새롭게 등록한다.

아홉째, 출현용어리스트에 등록된 색인어들에 해당 문헌번호와 함께 문헌번호의 합계를 계산하여 추가시킨다.

앞의 과정 중에서 복합명사 여부를 대조하는 작업은 복합명사리스트의 등록과 함께 단일명사로 분리하기 위함이다. 복합명사를 단일명사로 분리하는 이유는 분할된 단일 명사들은 복합 명사에 비해 특정성이 떨어질 수 있지만 검색 결과의 재현율을 높일 수 있기 때문이다.

본 연구에서 구축한 자동색인시스템은 출현용어리스트와 용어리스트를 작업 과정의 앞에 위치하도록 하였다. 이렇게 위치토록 한 이유는 인간의 수작업색인 방법이 이와 유사하기 때문이다. 인간의 수작업 색인 추출 방법은 수년간의 학습과정에 의해서 다양한 명사 형태의 용어를 인지하는 지능을 확대시키고 이를 이용해 문헌 내에서 문장 분석을 통해 명사 형태의 색인어를 추출하는 방법을 채택하고 있다. 이와 같은 인간의 지능 확대 방법을 - 지식습득방법을 자동색인 알고리즘에 적용시키기 위하여 출현용어리스트와 용어리스트를 작업 과정의 앞 부분에 위치하도록 구성하였다.

또한 본 시스템에서 구현한 자동색인시스템에 의해 추출된 색인어는 출현용어리스트(색인어데이터베이스)에 새롭게 등록되며 색인어와 문헌번호 그리고 전체 문헌에서의 출현빈도의 합을 누적시키도록 하였다. 이와 같이 출현용어리스트를 구성한 이유는 부울린 로직(boolean logic)을 이용한 정보탐색과 클러스터(범주) 표현 알고리즘에 활용하기 위함이다. 본 시스템에서는 이와 같은

자동색인 작업에 의해 2,245개의 기계학 및 인지과학에 관련된 용어들이 수록된 출현용어(색인어) 데이터베이스를 구축할 수 있었다.

### 3. 클러스터(범주) 형성 및 대표어 선정

본 연구의 목적은 초기 질의어를 근거로 복수의 연관 색인어들을 제시하고 이를 근거로 질의어를 확대하거나 축소하여 자동으로 탐색식을 구성하고 탐색을 수행할 수 있는 시스템을 구현하는 것이다. 이를 위해 본 연구에서는 서휘가 제시한 알고리즘을 채택하였다.[33] 서휘 알고리즘의 가설은 색인어의 동시출현을 근거로 군집화된 문헌들은 계층화된 클러스터이며, 이 계층화된 클러스터를 가상의 새로운 문헌이라고 할 때, 이 가상의 새로운 문헌은 복수의 동시출현 색인어들로 표현될 것이며, 이들 복수의 동시출현 색인어들을 대표하는 대표 색인어가 결국은 계층화된 각각의 클러스터를 대표하는 핵심 색인어가 될 것이란 가설을 근거로 한다. 또한 색인어의 계층화 형성에 대한 알고리즘은 형성된 클러스터가 계층화되어 있으므로 이를 대표하는 색인어들을 순서대로 나열하면 색인어들의 계층화가 가능할 것이란 가설에서 출발한다. 이와 같은 가설을 적용했을 때 색인어들의 계층을 사전에 정의하지 않고도(non-predefined) 계층화된 연관 용어 구조를 구성할 수 있다.

계층별로 형성된 클러스터를 단일 색인어 또는 소수의 색인어로 표현토록 하는 알고리즘을 상세히 설명하면 다음과 같다.

첫째, 클러스터 내에서 출현 빈도수가 제일 높은 색인어를 최상위 계층의 대표어로 선정한다.

둘째, 2번째로 출현빈도가 높은 색인어 중 연결된 문헌의 전부가 최상위 계층에 포함되는 문헌의 일부와 완전히 일치되는 것을 차 순위 계층의 대표어로 선정한다.

셋째, 3번째로 빈도가 높은 색인어에 해당되는 문헌 전부가 앞의 2번째 클러스터와 전부 일치하면, 2번째 색인어와 연결된 하위 클러스터의 대표어로 선정한다. 만약 일치하지 않으면 최상위 대표어와 비교하는 작업을 수행하여, 일치하면 최상위 클러스터에 연결되는 클러스터로 인식하고 해당 색인어를 클러스터의 대표어로 선정한다.

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

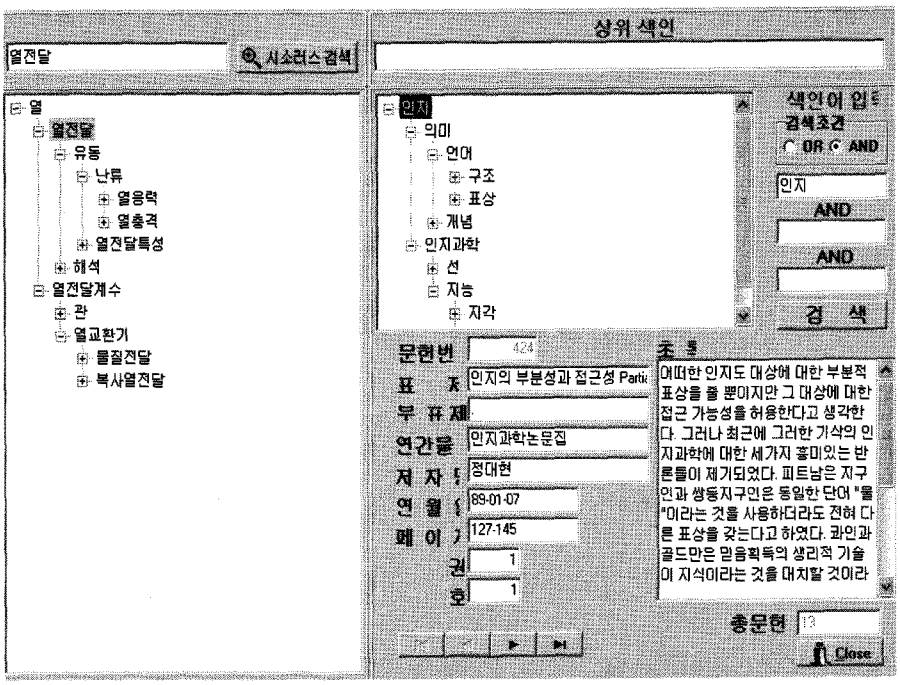
다섯째, 이 과정은 빈도수가 1회인 색인어까지 반복작업을 수행한다. 만약 빈도수가 1회인 색인어에 연결된 문헌이 동일문헌인 경우에는 앞의 과정처럼 대표어를 복수로 표시한다.

여섯째, 완전히 연결되지 않는 색인어들을 각 클러스터와 비교해 유사계수가 가장 높은 클러스터의 최상위 계층에 임의의 클러스터를 구성한다. 임의의 클러스터에 대한 대표어 표현은 각각의

33) 서휘(1999), 전게논문.

대표어를 OR로 묶어 표기한다. 이 작업은 모든 문헌이 전부 연결될 때까지 반복 작업을 한다.

본 연구에서는 이와 같은 알고리즘을 근거로 서로 상이한 주제(기계과학과 인지과학)의 입력데이터를 대상으로 클러스터(범주)화가 가능한가를 실험해 보았고 그 결과 〈그림 4〉와 같이 질의어를 근거로 자동으로 범주화된 관련 색인어들의 제시가 가능함을 알 수 있었다.




〈그림 4〉 형성된 클러스터 및 대표어

〈그림 4〉의 결과 화면에서 왼쪽의 화면에 출현하는 색인어들은 입력된 데이터 중에서 최상위 출현빈도 색인어인 '열'을 근거한 관련 색인어들이며, 오른쪽 화면에 출현한 색인어들은 색인어 입력 창에 용어를 입력하면 입력 용어를 근거로 실시간으로 새롭게 구성된 관련 색인어들이다. 오른쪽 화면의 색인어들은 '인지'란 용어를 근거로 구성된 관련 색인어들이다.

본 시스템에서 자동으로 구성해서 제시한 관련 색인어들은 입력한 질의어와 관련한 색인어들에 대한 유사도를 계산함에 의해 연관 색인어들을 추출하고 질의어와 연관 색인어간의 유사도 수치를 근거로 색인어들을 연결시키고 해당 색인어(대표어)에 범주(클러스터 또는 문헌)를 묶는 방법으로 구성된다.

본 시스템에서는 이용자의 초기 질의어와 관련된 색인어(문헌 내에 반드시 동일한 형태로 수록되어 있는 의미 있는 색인어)를 내장되어 있는 자체 알고리즘에 의해 자동으로 제시해 줄 수 있으

므로 이들 색인어들을 근거로 이용자들은 질의어의 확장이나 축소가 용이하며 확실한 정보 검색 결과를 보장받을 수 있다.

## 4. 지능형 정보검색시스템 구현

본 연구는 지능형 정보검색시스템을 구현하려는 목적에서 시도되었다. 지능형 정보검색시스템이란 귀납학습적 질의어 확장시스템의 기능을 수행할 수 있는 검색시스템을 의미한다. 귀납 학습적 질의어 확장 방법을 적용한 시스템은 이미 알려진 질의어와 각 질의어에 대한 적합한 문헌을 사전에 학습시킴으로써 특정한 질의어에 대한 가장 중요한 색인어가 어떤 것이지를 구별할 수 있게 할 수 있는 시스템이다. 또한 귀납학습 적용시스템은 이러한 각 질의어에 대한 적합한 색인어를 하나만 추출하는 것이 아니라 다양한 색인어들을 그 중요도와 함께 순위를 매길 수 있게 제공할 수 있어야 한다.[34]

앞의 장에서 살펴본 바와 같이 본 시스템은 어떠한 질의어도 시스템 내부에 수록된 한 개의 색인어와 일치하면 그 색인어와 연관된 관련 색인어를 자동으로 제시해줄 수 있는 기능을 제공하고 있다. 이와 같이 연관 색인어를 자동으로 제시해 줄 수 있는 기능을 수행하기 위해서는 앞에서 제시한 자동분류 알고리즘과 범주의 표현 알고리즘을 결합한 지능형정보검색시스템을 구현했을 때 가능하다. 또한 이들 연관 용어를 제시해 줄 수 있는 기능을 갖춘 지능형정보검색시스템은 이들 연관 용어를 이용하여 자동으로 재현율을 향상시키거나 정도율을 향상시킬 수 있는 방향으로 탐색식을 확대하고 축소할 수 있으며, 이를 통해 자동으로 정보탐색식을 구성하고 이를 이용한 정보탐색 기능의 수행이 가능하다.

다음의 장에서는 본 정보검색시스템을 이용하여 단일 주제(인지과학)의 용어에 대한 검색방법과 복수 주제 용어(인지과학과 기계과학에 동시에 출현한 동일형태의 용어)에 대한 검색방법을 설명하기로 한다.

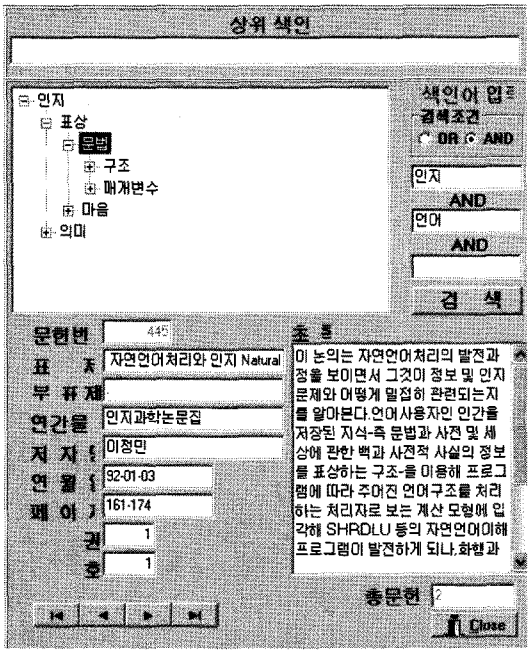
### 가. 단일 주제 용어의 검색

본 연구에서 구현한 지능형정보검색시스템의 질의어 확장과 질문식 자동구성 기능을 〈그림 4〉에서 제시된 사례를 근거로 설명하면 다음과 같다. 탐색자가 초기질의어를 인지과학 분야의 용어인 '인지'를 입력하고 검색버튼을 클릭하면 '인지'란 색인어가 수록된 13개의 결과를 제시하며 그 결과를 화살표시 버튼을 이용하여 해당 문헌의 내용들을 확인할 수 있도록 구성하였다. 또한 즉각적으로 '인지'란 색인어와 관련이 있는 형태가 다른 연관 색인어들을 자동으로 제시해주고 그들 연관

34) 김성희, "WWW상의 지능형 정보검색을 위한 기계학습 알고리즘 구현에 관한 연구," 정보관리학회지, 제17권 제2호(2000.6), pp.189-203.

색인어들 중에서 '언어'를 선택한다면 새로운 탐색식인(인지 AND 언어)가 구성되어 4개의 결과를 제시하며 그 결과 또한 확인할 수 있도록 구성하였다.

만약 초기 질의어를 복수의 용어들이 결합된 탐색식인(인지 AND 언어)로 입력하면 이 두 개 용어와 동시에 관련이 되는 새로운 형태의 연관용어들을 〈그림 5〉와 같이 제시하며, 그중 '문법'을 선택하면 질의어가 확장된 새로운 형태의 탐색식인(인지 AND 언어 AND 문법)이 구성되어 탐색을 수행하고 그 결과 2개의 탐색문헌을 확인할 수 있도록 구성하였다.



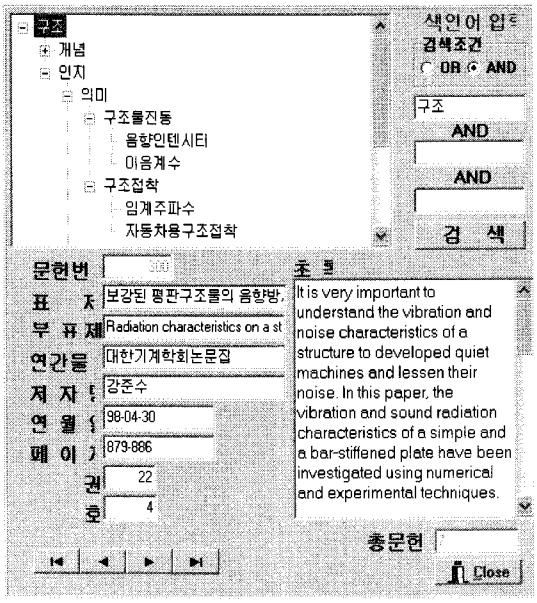

〈그림 5〉 AND를 이용한 복수의 초기 질의어를 이용한 검색 결과

본 시스템에서는 내장된 알고리즘에 의해 〈그림 5〉와 같이 자동으로 형성된 계층별 대표어 및 클러스터(범주 - 복수의 문헌)는 상위어와 하위어 간에 자동으로 AND 연산자가 채택됨에 의해 첫 번째 질의어의 의미를 한정시킬 수 있는 색인어와 연결된 탐색식을 자동으로 구성하고 이를 통한 입력데이터에 대한 탐색을 수행하는 기능을 제공한다.

### 나. 복합 주제 용어의 검색

정보검색에서는 동일한 형태의 탐색어가 다른 의미로 사용되거나 주제에 따라 관련된 용어들의 형태가 달라질 수 있다. 본 연구의 실험대상인 기계과학회 논문집-열 및 열처리 분야와 인지과학회 논문집에서 공통적으로 출현한 색인어는 〈그림 6〉에 제시된 바와 같이 '구조'이며, 이에 대한 검색

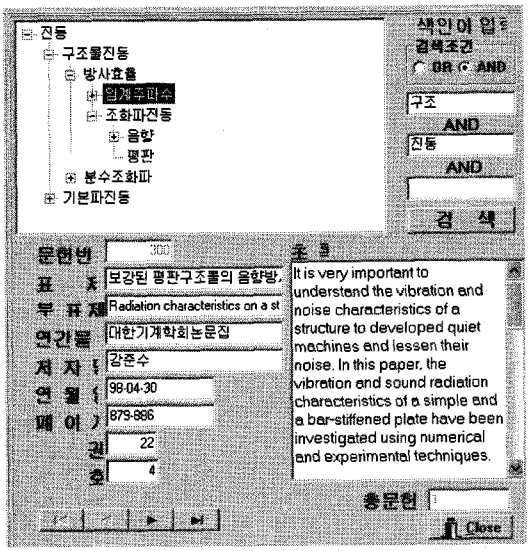
결과로 총 7개의 문헌을 확인할 수 있었다. 검색된 문헌은 인지과학에서 4개, 기계과학에서 3개가 검색되었다. '구조'와 연관되는 색인어들은 〈그림 6〉과 같이 제시되었으며 출현한 연관 색인어들은 27개로서 기계과학 분야와 인지과학 분야의 색인어들이 혼합되어 나타났다.




〈그림 6〉 각 주제의 공통 출현 용어인 '구조'와 관련된 검색결과 및 관련 색인어

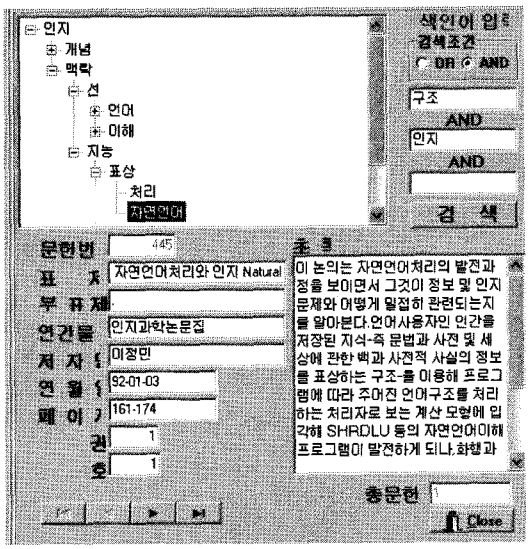
본 시스템에서는 이들 제시된 색인어들 중 특정 색인어를 클릭하면, 예를 들어 '이음계수'를 클릭하면(구조 and 이음계수)의 탐색식으로 전환하여 즉각적으로 1개의 검색결과를 제시할 수 있도록 하였다. 또한 인지를 클릭하면(구조 and 인지)의 탐색식으로 전환하여 3개의 검색결과를 제시하였다.

또한 '구조'와 연관된 용어 중 기계과학 분야의 '진동'이란 용어와 결합하면 '구조'와 '진동'의 두 개 색인어와 공통적으로 관련이 있는 새로운 형태의 연관 색인어 구조를 갖고있는 19개의 관련 색인어들을 제시해주며 이들 색인어들은 〈그림 7〉과 같이 기계과학 분야의 색인어들로만 구성되어 제시되었다. 또한 〈그림 7〉에 제시된 관련 색인어들 중 '임계주파수'를 클릭하면(구조 and 진동 and 임계주파수)에 해당하는 탐색식으로 전환해 1개의 검색결과를 제시해주었다.




〈그림 7〉 '구조'와 기계과학 용어인 '진동'이 결합된 탐색용 관련 용어

그리고 '구조'와 연관된 용어 중 인지과학 분야의 '인지'란 용어와 결합하면 19개의 관련 색인어들을 제시해주며 이들 색인어들은 〈그림 8〉과 같이 인지과학 분야의 용어들로만 구성되어 제시하였다. 또한 〈그림 8〉에 제시된 관련 용어 중 '자연언어'를 클릭하면(구조 and 인지 and 자연언어)에 해당하는 탐색식으로 전환해 1개의 검색결과를 제시해주었다.




〈그림 8〉 '구조'와 인지과학 용어인 '인지'가 결합된 탐색용 관련 용어

본 시스템에서는 이상과 같이 용어들의 계층과 관계를 사전에 정의하지 않고도(non predefined) 입력된 질의어를 근거로 연관 색인어들과 그 관계를 자동으로 제시할 수 있었으며, 이를 통하여 탐색식과 정보검색을 자동으로 수행할 수 있었다. 또한 색인어들은 반드시 원문에 출현한 자연어들을 근거로 구성하였기 때문에 이용자들이 일상적인 용어를 이용하여 정보검색 작업을 수행할 수 있도록 하였다.

정보검색시스템에서 탐색의 성공이란 문헌-색인어 형태로 축적된 문헌 데이터베이스를 근거로 탐색어가 색인어와 일치하는지의 여부를 의미한다. 따라서 탐색자가 색인어와 일치하지 않는 탐색어를 선택한다면 탐색에 성공할 수 없다. 반면에 탐색자의 탐색어가 용어의 특정성이란 측면에서 적절하다면 바람직한 탐색 결과를 제공받을 수 있는 것이다. 그러나 인터넷을 사용하는 일반 이용자들은 자신들이 선택한 첫 번째 질의어를 이용한 탐색 결과가 재현율이나 정도율이란 측면에서 매우 수준이 낮은 결과를 제공받고 있음은 누구나 경험하고 있는 주지의 사실이다. 또한 탐색자가 복수의 용어들을 입력하거나 문장 형태의 질문식을 입력하여도 동일한 결과가 발생하고 있었다. 그 이유는 일반 이용자들이 복수의 용어들을 사용할 때 대부분 그 용어들은 복합명사에 해당하기 때문이다.

그러나 본 시스템에서는 〈그림 8〉과 같이 지능형정보검색시스템을 이용해서 초기 질의어와 형태가 다른 연관용어들을 - 언어, 이해, 자연언어 등의 용어들을 - 제시해 줄 수 있기 때문에 앞에 제기한 정보검색시스템의 문제점은 축소될 수 있을 것이다. 또한 제시된 연관 색인어들을 이용하여 AND, OR 연산자를 적용해서 질의어의 확장과 탐색식의 수정이 즉각적으로 가능하기 때문에 이용자가 만족할만한 탐색결과를 보장받을 수 있을 것이다. 본 연구에서는 이와 같은 지능형검색시스템을 구현하였다.

## Ⅳ. 결 론

자동분류 알고리즘을 이용한 지능형정보검색시스템의 구축에 대한 본 연구의 결과를 요약하면 다음과 같다.

첫째, 지능형정보검색시스템을 구현하기 위해서는 자동색인 시스템이 필수적이다. 본 연구에서는 한글 자동색인에 적합한 알고리즘을 채택하여 학습을 통해 기능이 향상되는 자동색인시스템을 구축하였다.

둘째, 기존의 정보검색시스템은 초기 질의어를 입력하여 질의어를 확대하고자 할 경우, 초기 질의어의 형태가 포함된 복합명사 형태의 연관 용어만 제시해주는 기능을 제공하는 수준이다. 그러나 본 연구에서는 입력된 질의어를 근거로 질의어와 형태가 다른 연관 색인어들을 자동으로 제시하고

이를 근거로 초기질의어와 연관 색인어를 연결시킨 탐색식을 자동으로 구성하여 데이터베이스 내에서 해당 정보를 탐색해내는 기능을 수행하는 지능형정보검색시스템을 구현하였다.

셋째, 자동분류라 함은 주제가 상이한 입력데이터를 인간의 작업을 거치지 않고 범주화 작업을 수행할 수 있는 시스템을 의미한다. 본 연구에서는 주제가 상이한 기계 분야와 인지과학 분야의 데이터를 입력하고 자동범주화의 가능성을 실험한 결과 지능형검색시스템에서 주제가 자동으로 분리된 범주화된 결과를 확인할 수 있었다.

이상과 같은 결과를 근거로 할 때, 본 연구에서 개발한 알고리즘과 시스템은 자동색인, 자동분류, 자동정보검색 그리고 지식의 구조화를 통한 학습방법에도 널리 적용될 수 있을 것으로 기대된다.

〈참고문헌은 각주로 대신함〉