색인파일 기반의 질의어 확장용 지식베이스 구축에 관한 연구

A Study of Designing the Knowledge Base System for the Query Extension by Index File

서 휘(Whee Seo)*

〈 목 차 〉

초 록

본 연구의 목적은 이용자 지향적인 정보검색을 수행하기 위한 질의확장용 지식베이스를 구축하는 것이다. 이를 위해 개념기반 정보검색방법과 통계적 기반 정보검색방법을 이용한 지식베이스 구축에 관련된 다양한 이론 연구를 수행하였다. 이들 지식베이스 구축방법에 있어서 공통된 가설은 연관용어의 출현은 문헌집합내의 동시출현 빈도임을 재확인하였고, 이 가설을 근거로 색인파일 알고리즘과 부울 논리의 And 연산자를 이용하여 질의확장용 지식베이스를 구축하였다. 본 지식베이스의 실험 주제는 교육학이며, 교육학개론이란 단행본을 이용하여 색인어들의 연관용어를 자동으로 제시해줄 수 있는 실험용 지식베이스를 구축하였다. 실험용 지식베이스는 자연어색인방법과 통제어색인방법을 이용하여 두 개의 지식베이스를 구축해 각 지식베이스 시스템의 질의확장 성능에 대한 평가 작업을 수행하였다.

키워드: 지식베이스, 질의 확장, 개념 기반 정보검색, 자동정보검색 알고리즘, 지능형 정보검색 시스템

ABSTRACT

This study is to develop knowledge base system for query extension to the user oriented information retrieval. This study has survey the theories of the concept-based information retrieval method and statistic based information retrieval method. In the construction method of knowledge base, the common hypothesis is that the emergence of related term is the frequency of simultaneous emergence of a set of documents. Using the subject index file algorithms and the 'and' operator of boolean logic based on this hypothesis, this study builds the knowledge base. In this research experiment, a subject of knowledge base is education. Using the book of the Introduction to Education, two experimental knowledge base systems is constructed by the different indexing method. One system has constructed by controlled language indexing method, and another system has constructed by natural language indexing method. The performance of two knowledge base system is evaluated.

Keywords: Knowledge Base, Query Extension, Concept-based Information Retrieval, Automatic Information Retrieval Algorithm, Intelligent Information Retrieval Algorithm

* 창원전문대학 문헌정보과 조교수(drs733m@changwon-c.ac.kr)
• 접수일: 2009년 5월 18일 • 최초심사일: 2009년 5월 28일 • 최종심사일: 2009년 6월 22일

Ⅰ. 서 론

현재 인터넷을 이용한 온라인 정보검색시스템은 이용자 편의적인 시스템으로 발전하고 있다. 특히 문장 형태의 질문을 입력하여 원하는 정보를 탐색하는 방법이나 초기 질의어와 의미적으로 관련이 되는 소수의 연관 용어(association terms)들을 자동으로 제시하는 방법들은 이용자 편의적인 정보검색시스템의 대표적 사례라고 할 수 있다. 그러나 아직까지도 다양한 형태의 관련어들을 제시해주고 이를 활용하여 정보 탐색을 수행하는 온라인 정보검색시스템은 거의 존재하지 않고 있다. 이와 같은 이유 때문에 온라인 정보탐색 작업 시 불필요한 정보가 추가되거나 필요한 정보가 누락되는 현상이 발생하고 있는 것이다.

이와 같은 문제점을 축소하기 위해서 상업용 데이터뱅크인 DIALOG와 같은 정보검색시스템에서는 Expand 명령어 기능을 이용하여 초기 질의어에 관련된 용어들을 확장하는 방법을 제공하고 있다. 물론 이와 같은 DIALOG의 용어 확장 방법은 특정 개념어에 대한 확대와 축소가 가능한 시소러스 기반의 지식베이스가 존재하기 때문에 가능하다.

또한 일부 정보탐색 전문가들은 정보 탐색 전략 중, 핵심 참고문헌 활용 확장 검색 기법(pearl citation growing method)을 이용하여 이와 같은 문제점을 해결하고 있다. 핵심 참고문헌 활용 확장 검색 기법은 초기 질의어에 의해 탐색된 문헌들 중에서 적합문헌의 색인어 필드(descriptor field, identifier field)에서 관련어들을 수작업으로 확인하고 이를 피드백 탐색에 활용하는 방안이다. 그러나 이 방법은 정보탐색 전략에 익숙하지 않은 일반인에게는 적용하기 어려운 방법이며, 전문가에게도 수작업의 과정에 의해 색인어 필드를 확인해야만 하는 번거로운 과정이 수반된다. 또한 인터넷에서 유통되고 있는 정보들에는 메타 데이터(meta data)들이 거의 없어 - 별도의 색인어들이 존재하지 않으므로 - 이 방법은 적용하기 불가능하다.

현재 시스템 측면에서 인터넷을 이용해 정확한 정보 검색을 수행하기 위해서 채택할 수 있는 초기 질의어 자동 확장 방법은 다음의 두 가지 방법일 것이다. 첫 번째 방법은 초기 질의어에 의해 탐색된 결과 중, 적합문헌을 선택해서 문헌 내의 텍스트들을 자동색인 알고리즘으로 분석해서 관련어들을 제시해주는 방법이다. 두 번째 방법은 학문별 주제별 지식베이스를 구축해서 이를 이용해 초기 질의어와 관계하는 관련어들을 제시해주는 방법이다.

본 연구에서는 두 번째 방법인 학문별 주제별 지식베이스를 구축해서 초기 질의어를 확장해서 정보검색의 효율성을 보장할 수 있는 방안을 제시해보고자 한다. 따라서 본 연구에서는 초기 질의어의 확장방법에 대한 선행 이론을 소개할 것이며, 이를 근거로 지식베이스를 구축할 것이다. 본 연구에서 구축할 실험용 지식베이스의 주제는 교육학이며, 교육학 지식베이스의 구축방법은 교육학 개론 단행본을 근거로 장, 절, 항목, 내용 단락별로 나누어 원문을 입력하고 색인어를 추출하여 구축할 것이다. 교육학 지식베이스는 색인 추출방법을 달리하여 자연어 색인을 근거한 교육학 지식

베이스와 통제어 색인을 근거한 교육학 지식베이스를 구축할 것이다.

구축된 두 개의 교육학 지식베이스에 대하여 문헌정보학과와 유아교육학과 학생들을 대상으로 색인어의 전문성, 일반성, 특정성, 망라성을 근거로 지식베이스의 연관용어 특성에 대한 평가를 수행할 것이다. 또한 탐색식 작성의 편리성, 탐색 결과의 재현율과 정도율을 근거로 지식베이스의 정보검색 지원 가능성을 평가할 것이다. 그리고 지식베이스를 통해 제시된 연관용어들의 관계에 대한 개념적 밀접성과 이들 관계를 통해 교수나 학습의 지원 가능성에 대해 평가하고 그 결과를 제시할 것이다.

Ⅱ. 질의어 확장 시스템에 대한 이론적 연구

1. 질의어 확장의 정의

온라인 정보탐색 작업은 이용자가 자신의 정보 요구(needs)를 분석해서 부울린 로직을 적용시킨 정보 탐색식으로 바꾸어 정보검색시스템에 입력함에 의해 시작된다. 앞의 문장에서 정보 탐색식은 초기 질의어이다. 정보검색 시스템에 있어서 초기 질의어의 입력 방식은 단일 명사나 복합 명사 또는 복수의 명사들을 직접 입력하는 방식과 자연어 문장 형식으로 입력하는 두 가지 방식이 있다. 자연어 문장 형식의 질의어는 자동색인 알고리즘(파싱과 스테밍 알고리즘)을 이용해서 불용어를 제외한 개념어들을 추출해서 이를 초기 질의로 선정해 입력하는 방식을 취한다.

일반적으로 초기 질의어를 입력한 후 탐색 결과가 만족스러우면 탐색 작업이 종료되나 그렇지 못할 경우 피드백 탐색이 요구된다. 보통 피드백 탐색 시 초기 질의에 대한 변경이나 수정이 요구된다. 이와 같은 초기 질의어에 대한 변경이나 수정을 질의어 확장(Query Expansion) 작업이라고 한다. 질의어 확장방법은 지식 기반 확장방법과 탐색 결과 기반 확장방법이 있다. 지식 기반 확장 방법은 시소러스나 의미 네트워크 등과 같은 지식베이스를 이용하여 확장하는 방법이며, 탐색 기반 확장방법은 초기 질의어를 근거로 검색된 문헌 중 적합 문헌에 출현한 용어들을 사용해서 확장하는 방법이다.[1)2)3)4)] 본 연구에서는 지식 기반 확장 방법에 대해서만 기술하고자 한다.

---

1) 노정순, "탐색 결과에 근거한 자연어질의 자동확장 및 응용에 관한 연구 고찰," 정보관리학회지, Vol.16, No.2 (1999. 3), pp.49-80.

2) 김성희, "www상의 지능형 정보검색을 위한 기계학습 알고리즘 구현에 관한 연구," 정보관리학회지, Vol.17, No.2 (2000. 6), pp.189-203.

3) L. A. Paris & H. R. Tibbo, "Free style vs. Boolean : a comparison of partial and exact match retrieval systems," *Information Processing & Management*, vol.34(1998), pp.175-190.

4) P. Borlund & P. Ingwersen, "The development of a method of the evaluation of interactive information retrieval systems," *Journal of Documentations*, Vol.53, No.3(September 1997), pp.225-250.

2. 지식 기반 질의어 확장방법의 종류

질의어 확장 시스템은 초기 질의어와 의미적으로 관련이 있는 용어들을 수작업이 아닌 자동화된 방법으로 이용자들에게 제시해주어 정보 검색의 효율을 향상시킬 수 있는 검색시스템이다. 이와 같은 기능을 하는 지식 기반 질의어 확장시스템은 연관용어를 선정하는 알고리즘에 따라 개념 기반 구축방법과 통계적 유사도 기반 구축방법이 있다.

가. 개념 기반 구축방법

개념 기반 구축방법은 연관용어의 선정을 각 용어의 개념 관계에 의해 선정하는 방법이다. 즉 기존의 시소러스를 이용한 확장방법이 이 방법이다. 개념 기반 구축 방법을 이용한 질의어 확장시스템은 이용자의 초기 질의어를 수록하고 있는 문헌뿐만 아니라 질의어와 연관된 색인어가 속해있는 문헌까지 검색할 수 있는 기능을 수행한다. 이와 같은 기능을 수행하기 위해서 지식베이스의 구축과 함께 개념 확장 알고리즘이 적용되어야 한다. 일반적으로 지식베이스는 해당 검색시스템에 입력되는 문헌 데이터베이스를 이용하여 자동으로 구축한다.

개념 기반 구축방법을 적용한 정보검색시스템은 구축된 지식베이스를 근거로 초기 질의어에 대한 개념 확장을 수행한 후(관련어들을 추가한 후) 입력된 문헌 데이터베이스를 대상으로 관련 정보를 검색한다. 지식베이스를 기반으로 개념 확장(용어 확장)을 수행하는 알고리즘은 bnb 알고리즘(branch-and bound expansion activation algorithm)과 홉필드 넷 알고리즘(Hopfield net algorithm)이 사용되고 있다.[5)6)]

개념 기반 구축방법에 있어서의 문제점은 용어간의 개념관계를 설정하는 방법이 어렵다는 점이다. 특정 주제의 모든 용어에 대하여 상위개념과 하위개념 또는 동의어 개념을 설정하는 것은 매우 복잡한 작업과 함께 시간과 노력 그리고 비용이란 점에서 많은 투자가 요구되는 작업이다.

나. 통계적 유사도 기반 구축방법

통계적 유사도 기반 구축방법은 용어-문헌 매트릭스, 용어-용어 매트릭스를 근거로 통계적 방법을 적용하여 연관용어를 선정하는 방법이다. 연관 용어에 대한 통계적 유사도 기반 구축방법은 텍스트 마이닝 알고리즘(text mining algorithm)과 용어 클러스터링 알고리즘(term clustering algorithm)이 있다. 텍스트 마이닝 알고리즘은 구조화된 데이터를 대상으로 하는 데이터 마이닝 기법인 연관 규칙 마이닝을 비구조화된 형태인 텍스트 처리에 적용시켜 연관 용어를 형성하는 알

5) 노영희, "개념기반 검색을 위한 시소러스 관계의 효과적 활용방안에 관한 연구," 정보관리학회지, Vol.17, No.4 (2000. 12), pp.47-65.

6) 노영희, 정영미, "의미망 지식베이스를 이용한 개념기반 정보검색기법의 성능 연구," 정보관리학회지, Vol.17, No.3(2000. 9), pp.45-70.

고리즘이다. 텍스트 마이닝 알고리즘은 연관성 척도인 지지도(support), 신뢰도(confidence), 향상도(lift)를 근거로 문헌 텍스트로부터 연관 용어 집합을 생성한다. 용어 클러스터링 알고리즘은 용어 간 연관성 척도인 다양한 유사계수 측정 공식을 적용하여 연관 용어 집합을 생성하는 알고리즘이다.[7)]

이들 모든 알고리즘의 공통된 기본적 가설은 문헌 집합에서 각 문헌에 특정한 두 개의 용어가 동시에 출현하는 빈도가 높을 경우, 이 두 개의 용어들은 서로 연관성이 있는 용어라는 점이다. 이와 같은 통계적 유사도 기반 구축방법에 있어서의 문제점은 용어 간의 유사도관계는 문헌-용어 매트릭스와 용어-용어 매트릭스 간의 유사도 연산 계산 값과 기준치 값의 변화를 통해 형성되기 때문에 신규문헌과 신규 용어가 입력될 때마다 많은 처리 시간과 저장 용량이 요구된다는 점이다.

3. 색인 파일 기반 질의어 확장 시스템 구축 방안

색인 파일의 구축 작업은 문헌-용어 매트릭스, 용어-용어 매트릭스를 근거로 용어 간의 연관성을 측정하는 통계적 유사도 기반 질의어 확장 시스템의 사전 작업에 해당한다. 따라서 주제어 색인 파일만을 이용해 초기 질의어에 의미적으로 관련성이 있는 연관용어를 제시해줄 수 있다면 데이터베이스의 용량이나 시간 그리고 비용이란 측면에서 효율적일 것이다.

주제어 색인 파일을 이용한 질의어 확장 시스템의 핵심 이론은 앞의 통계적 유사도 기반 구축방법의 가설과 동일하다. 즉 각 문헌에 특정한 두 개의 용어가 동시에 출현하는 빈도가 높을 경우 이 두 개의 용어들은 서로 연관성이 있는 용어라는 점이다.

가. 주제어 색인 파일의 구조

주제어 색인 파일은 일반적으로 〈그림 1〉과 같이 사전 파일(dictionary file)과 도치파일(inverted file)로 구성되어 있다. 사전 파일에는 색인어와 저장 주소(address) 시작번지와 색인어 수록 레코드의 개수(색인어 + 시작 주소(91번지) + 50개)로 구성되어 있다. 도치 파일은 주소 번지수와 레코드 번호(91번지 + 50번 레코드)로 구성된다.[8)9)]

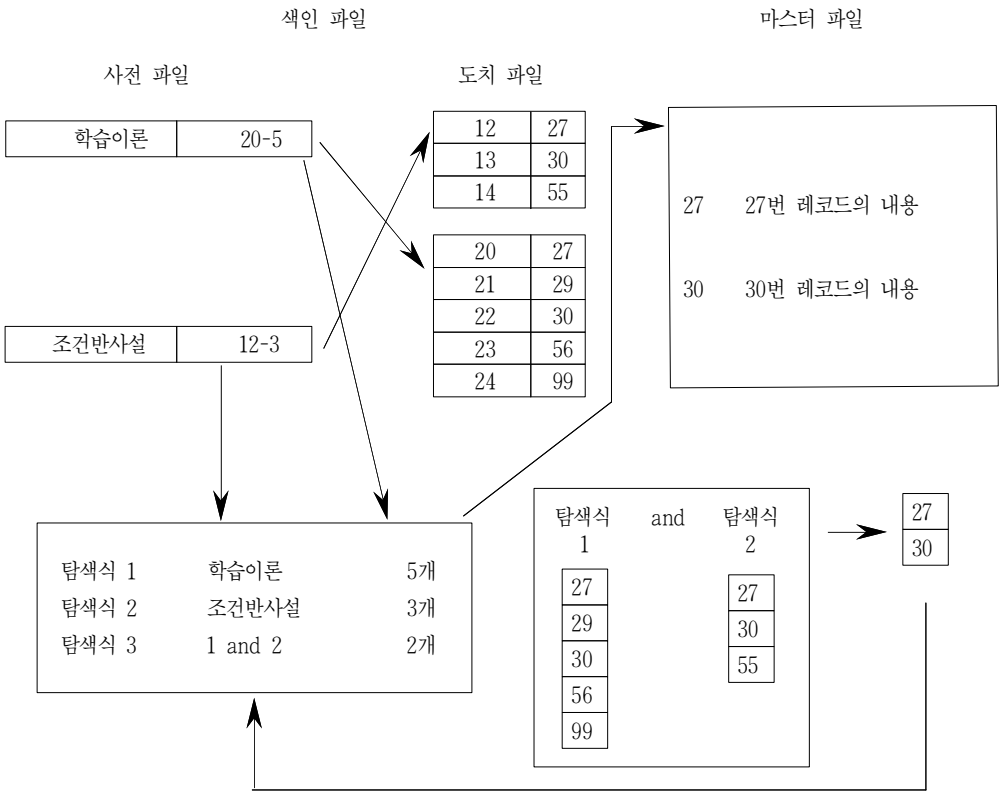
〈그림 1〉과 같은 구조를 갖는 데이터베이스에서 색인 파일은 사전 파일과 도치 파일을 이용하여 이용자의 질문식에 관계하는 적합정보를 탐색하는 핵심적 역할을 수행한다.

---

7) 김수연, 정영미, "텍스트 마이닝 기법을 이용한 연관 용어 선정에 관한 실험적 연구," 정보관리학회지, Vol.23, No.3(2006. 9), pp.147-165.

8) 津田良成編. 圖書館·情報學槪論(東京 : 勁草書房, 1983). pp.121-122.

9) William B. Frakes & Ricardo Baeza-Yates ; 김진호 & 류근호 공역, 정보검색(서울 : 시그마프레스, 1995). pp.43-63.



〈그림 1〉 데이터베이스의 파일 구조

나. 주제어 색인 파일 기반 용어 관계 설정

색인 파일을 이용한 용어 관계의 설정은 부울린 로직의 AND 연산자를 이용하면 가능하다. 〈그림 1〉에서 학습이론과 조건반사설이란 색인어는 각기 다른 문헌 집합(27번 문헌, 30번 문헌)에서 공통으로 출현한 용어이다. 따라서 27번 문헌과 30번 문헌의 원문 내용을 개념적으로 대표하는 용어는 학습이론과 조건반사설이다.

앞의 통계적 유사도 기반 질의어 확장시스템에서의 구축방법의 가설과 같이 동일한 문헌집합에서 각기 다른 형태의 용어가 동시에 출현한다면 이들 용어들은 분명 개념적으로 연관용어라고 할 수 있다. 더구나 두 개의 문헌집합이 아니고 수십 개 이상의 문헌집합에서 이들 두 개의 용어가 동시에 출현한다면 이 두 개의 용어는 개념적으로 매우 밀접한 연관관계를 갖는다는 가설이 가능하다. 이와 같은 문헌 집합 내에서 동시출현 용어를 선정하는 방법은 앞에서 기술한 연관용어 선정 방법인 개념 기반 구축방법과 통계적 유사도 기반 구축방법에서 사용한 다양한 알고리즘을 적용하지 않고도 색인파일의 구조를 이용하면 구현이 가능하다.

다. 주제어 색인 파일 기반 질의어 확장 방법

색인 파일을 이용한 질의어 확장 방법은 - 초기 질의어에 대한 연관용어 제시 방법은 데이터베이스의 각 문헌 레코드에서 주제 개념을 갖는 필드(서명 필드, 초록 필드, 원문 필드)들을 분석해서 색인어 필드(디스크립터 필드, 식별자 필드)를 구성하고 부울린 로직의 AND 연산자를 적용하면 가능하다.

데이터베이스에서 각 문헌 레코드의 주제어 색인 필드에 수록되어 있는 각기 다른 형태의 색인어들은 특정 레코드의 내용을 대표하는 개념을 대표하는 색인어들이다. 따라서 특정 레코드에 수록되어 있는 색인어들은 상호간 개념적으로 밀접한 관계를 갖는 용어들이 될 수밖에 없다. 그러므로 색인 필드에 수록되어 있는 색인어들을 이용하면 초기 질의어와 관련된 연관 용어들을 제시해줄 수 있다. 주제어 색인 파일을 이용한 질의어 확장 방법은 다음과 같은 과정을 거쳐 제시된다.

첫째, 초기 질의어를 근거로 색인 파일 내에서 동일한 형태의 색인어를 탐색한다.

둘째, 초기 질의어와 동일한 형태의 색인어를 수록하고 있는 레코드를 찾는다.

셋째, 초기 질의어와 동일한 형태의 색인어를 수록하고 있는 레코드를 적합레코드라고 가정하고 해당 레코드의 색인 필드에 수록되어 있는 색인어들을 추출한다.

넷째, 추출된 복수의 색인어들을 초기 질의어와 AND 연산자를 적용시켜 탐색 작업을 수행한다.

다섯째, AND 연산자에 의해 문헌 집합에서 동일 문헌 내에 초기 질의어와 동시에 출현하는 용어들을 빈도가 높은 용어부터 순서대로 제시한다.

이상과 같은 과정에 의해 구성된 연관 용어들의 순서는 연관 용어 간의 개념 밀접도 순위와 유사할 것이다. 따라서 상위의 위치에 배열된 용어는 초기 질의어와 개념적으로 매우 밀접한 관계를 갖는 용어일 것이며, 하위의 위치에 배열된 용어는 개념적으로 소원한 관계를 갖는 용어일 것이다.

이상과 같이 색인 파일을 이용한 연관용어의 자동 설정 방법은 초기 질의어와 관련된 연관 용어를 제시해줄 수 있으므로 탐색 효율 향상을 위한 용어 확장 방법의 이상적 알고리즘이 될 수 있을 것이다.

Ⅲ. 색인 파일 기반 질의어 확장 시스템의 구현

1. 시스템환경

본 연구에서 구축한 질의어 확장 시스템은 앞의 색인 파일 기반 질의어 확장시스템의 구축방법을 근거로 Paradox 7.0 DBMS와 Delphi 4.0(PASCAL)을 이용해 개발하였다. 질의어 확장 시스템의 실험용 데이터의 주제는 교육학으로 하였다. 실험용 데이터는 단행본으로 하였으며, 입력 단

행본은 내용 분량이 305페이지인 교육학개론[10]의 원문을 대상으로 하였다. 이와 같은 실험데이터를 근거로 질의어 확장을 위한 지식베이스를 구축하였다.

실험데이터를 교육학개론이란 단행본으로 선정한 이유는 개론서에는 특정 학문에 대한 모든 전공용어들이 수록되며 또한 전공용어에 대한 상세한 설명이 기술되어 있어, 이들을 이용해 특정 색인어와 관련 색인어들의 추출이 용이할 것이라고 판단했기 때문이다.

2. 실험데이터 입력방법

실험데이터의 입력방법은 1개 레코드에 표제어와 페이지 수 그리고 해당 페이지의 원문과 색인어 필드로 구성하였다. 표제어는 장과 절 또는 항목의 제목으로 하였다. 페이지 수는 입력 내용이 교육학개론에 위치하는 페이지의 숫자를 입력하였다. 페이지 원문은 1개 레코드에 약 2.5페이지의 분량을 입력하였다. 이 같은 레코드 별 입력 분량은 원문을 읽어보고 단락의 내용변화를 근거로 결정하였다. 레코드 별 원문의 입력 분량을 단행본 내 항목의 제목 별로 입력하지 않고 세분한 이유는 입력대상이 개론서이기 때문에 소항목내에도 내용들이 세분되어 있기 때문이었다.

색인어 입력방법은 자연어색인 입력방법과 통제어색인 입력방법으로 하였다. 본 연구에서는 색인어의 입력방법에 따라 2개의 지식베이스 - 질의어 확장시스템을 구현하였다.

가. 자연어색인 입력

자연어색인의 추출 및 입력 방법은 1개 레코드를 입력할 때마다 자동색인 방식과 수작업 방식을 병행하여 처리하였다. 자동색인 방식과 수작업 색인 방식을 병행해서 색인어를 선택한 이유는 교육학 관련 색인어가 학문의 특성상 용어들에 역사, 철학, 심리학, 사회과학 등 광범위한 주제를 갖고 있으며 전공용어에 일반 용어의 성격이 강한 용어들이 많이 포함되므로 색인어 선택 과정이 자연과학의 문헌보다 훨씬 복잡한 작업이었기 때문이었다.

색인어의 추출은 레코드 1개마다 다음과 같은 작업 과정을 거쳐 선택하였다.[11]

첫째, 1개 레코드를 대상으로 기 구축한 조사 리스트, 조사 겸 어미 리스트, 조사 어미형 명사 리스트, 불용어 리스트 등을 근거로 자동색인 방식으로 처리하였다. 추출된 색인어는 수작업으로 대조하기 편리하도록 가나다 순서로 배열하여 제시하였다.

둘째, 특정 레코드의 원문을 수작업으로 분석해 색인어로 필요한 단어들이 누락이 되는 경우에는 새롭게 색인어로 추가하였다. 수작업으로 색인어를 추가할 때, 해당 레코드의 자동색인 작업에

10) 손유식, 교육학개론(마산 : 도서출판 경남, 2006). pp.1-305.
11) 서휘, "자동분류 알고리즘을 이용한 지능형 정보검색시스템 구축에 관한 연구," 한국도서관·정보학회지, Vol39, No.4(2008. 12), pp.283-304.

서 추출한 색인어를 중복으로 입력할 수도 있으므로 입력과 동시에 자동으로 대조해 동일 형태의 명사는 1개의 색인어만 입력할 수 있도록 구성하였다.

셋째, 색인어로 필요한 단어들이 자동색인 작업에서 누락이 되는 이유는 조사 겸 어미 리스트나 조사 어미형 명사 리스트 또는 불용어 리스트에 그 단어가 수록되어 있기 때문이다. 따라서 1개 레코드의 색인 작업이 종료된 이후 해당 리스트에서 검색 기능을 이용해서 특정 단어를 삭제시키는 작업을 해 추후 자동색인에 의해 색인어로 추출되도록 하였다.

넷째, 자동으로 추출된 색인어 중에서 색인어로 해결이 되지 않아야 할 용어들은 추후 색인어로 추출되지 못하도록 불용어 사전에 등록하였다.

다섯째, 복합명사를 자동으로 단일 명사로 분리하는 경우, 동형이의어 형태의 색인어가 추출되는 경우가 많으므로 수작업으로 대조하여 필요 없는 용어는 레코드별 색인 작업시 즉각 색인어에서 누락시켰다.

여섯째, 이와 같은 과정에 의해 자동색인시스템의 성능이 강화되어 정련된 색인어를 추출할 수 있지만, 앞의 다섯 번째 과정에서 발생하는 동형이의어와 같은 불필요한 색인어가 추출될 수 있으므로 입력 작업이 종료될 때까지 수작업의 과정을 병행해서 정확한 색인어를 선택하였다.

그 결과 141개의 레코드와 3,787개의 색인어로 형성된 자연어 색인 기반 질의어 확장시스템을 구축하였다. 아래의 그림은 자연어 색인을 이용한 1개 레코드의 입력 사례이다.

표제어 : 1. 교과중심 교육과정 저자 : 손유식 서명 : 교육학개론
페이지 : 215-218 원문 : 생략
색인어 : 가치, 객관적, 경험, 경험해석, 계발, 교과, 교과서, 교과영역, 교과중심, 교과중심교육과정, 교사, 교사중심교육과정, 교수학습방법, 교육, 교육과정, 교육과정개정, 교육내용, 교육성과, 교육자, 교재, 그리스, 기하, 논리적, 논리학, 명료, 문법, 문제해결, 문화유산, 문화유산가치, 문화유산중심교육과정, 분류, 사고방식, 사회생활, 산술, 설명력, 설명위주교수법, 수사학, 습득, 역사적, 연계성, 음악, 의미, 이성, 이성계발, 이성주의, 이해, 인류, 인류문화유산, 일반적, 입학요건, 전달, 조직, 주지주의, 중앙집권적통제, 지도, 지식, 지식습득, 지식체계, 천문학, 체계, 체계적, 측정평가, 통제, 특징, 평가, 학교, 학문, 학습, 학습결과, 학습자, 학예, 학자, 한정교과영역

〈그림 2〉 자연어 색인 입력사례

나. 통제어색인 입력

색인어의 추출 방법은 1개 레코드를 입력할 때마다 원문을 읽고 내용분석을 함에 의해 선정하였다. 색인어의 추출은 레코드 1개마다 다음과 같은 작업과정을 거쳐 선택하였다. 통제 색인어는 디스크립터 프로파일링 방법을 적용시켜 추출하였다.[12] 단행본의 경우에는 학술 기사와 같이 문헌에

12) 김판준 & 이재윤, "연구 영역 분석을 위한 디스크립터 프로파일링에 관한 연구," 정보관리학회지, Vol.24, No.4 (2007. 12), pp.285-303.

부여된 디스크립터가 존재하지 않으므로, 권말색인과 장과 절 그리고 항목의 제목에 출현한 복합명사들을 우선 처리하는 디스크립터 프로파일링 방법을 적용하였다.[13]

레코드 별 통제 색인어 추출과정은 다음과 같다.

첫째, 항목의 제목을 색인어로 선정하였다.

둘째, 표제어의 내용이 항목인 경우에는 상위항목인 장과 절의 제목을 색인어로 추가하였다.

셋째, 원문에 출현하는 용어 중, 표제어와 밀접한 관계를 갖는 복합명사 형태의 용어만을 색인어로 추출하였다.

그 결과 141개의 레코드와 1,294개의 색인어로 형성된 통제어 탐색 확장시스템을 구축하였다. 아래의 그림은 자연어 색인을 이용한 1개 레코드의 입력 사례이다.

표제어 : 1. 교과중심 교육과정	저자 : 손유식	서명 : 교육학개론
페이지 : 215-218	원문 : 생략	
색인어 : 교과중심교육과정, 교사중심교육과정, 교실학습제일주의, 교육, 교육과정, 교육과정유형, 문화유산중심교육과정, 설명위주교수법, 전통적교육과정, 중앙집권적통제용이성, 한정교과영역-교수학습		

〈그림 3〉 통제어 색인 입력사례

3. 색인 파일 기반 질의어 확장시스템의 구현

교육학 지식베이스 또는 교육학 분야 질의어 확장시스템은 색인어의 입력방법에 따라 두 개의 시스템을 구현하였다. 한 개 시스템은 자연어 색인을 이용한 질의어 확장시스템이며, 다른 한 개 시스템은 통제어 색인을 이용한 질의어 확장시스템이다.

가. 자연어색인 기반 지식베이스

자연어색인을 이용한 지식베이스는 교육학 관련 자연어 형태의 질의어에 대한 연관용어를 제시하기 위해 구축한 질의 확장 시스템이다. 교육학과 관련한 자연어를 질의어 입력창에 입력하면 지식베이스에 형성된 연관용어 관계를 이용해 입력된 질의어와 밀접한 관계를 갖고 있는 연관용어들을 제시해 이용자가 특정 연관용어를 선택함에 의해 질의 확장 기능을 할 수 있도록 구성되어 있다. 〈표 1〉은 초기 질의어 '학습이론'과 의미적으로 밀접한 관계를 갖는 연관용어들의 리스트이다.

자연어색인을 이용한 지식베이스는 〈표 1〉과 같이 자동으로 제시되는 연관용어들을 대상으로 인간의 두뇌 작업을 통해 특정 용어를 선택함에 의해 초기 질의를 확장하거나 축소할 수 있다. 단,

13) 서휘, "자동정보검색을 위한 한글 시소러스 브라우저 구축에 관한 연구," 한국도서관·정보학회지, Vol31, No.2 (2000. 6), pp.279-302.

자연어 색인을 이용한 지식베이스는 입력된 색인어가 일반성이 강해 초기 질의어인 '학습이론'과 밀접한 관계를 갖는 연관용어(조건반사설, 통찰설)와 일반적 용어(자극, 체계화, 학생, 학습, 형성)들이 함께 제시되고 있어 용어확장에 불편함이 있음을 알 수 있다.

〈표 1〉 자연어 색인 기반 지식베이스의 질의어 확장

(초기 질의어(학습이론)의 질의 확장)

동시 출현 빈도	색인어	동시 출현 빈도	색인어
8	학습이론	7	학습
5	행동	4	형성
3	학생		
2	강화 경험 기술적 동물 반응 시행착오 이해 자극 조건반사설 조건반응 체계화 통찰설 파블로프 학습상황 학습현상 현상 효과적		
1	계획 과정 교재 기술 목표 변화 신념 의미 인간 조건화 조작 지식 학습자 형태 환경		

그러나 자연어색인을 이용한 지식베이스의 장점은 일반 용어를 이용하여 특정 전공용어를 확인할 수 있도록 하는 기능을 갖고 있다는 점이다. 〈표 2〉의 연관용어 리스트는 초기질의어(학습 AND 행동)를 입력하여 이 두 개의 용어와 동시에 출현하는 관련 용어를 제시한 것이다. 〈표 2〉의 연관용어 리스트를 분석해보면 초기질의어를 복수의 일반용어를 입력함에 따라 학습이론이나 교수학습과정과 같은 특정 전공용어를 확인할 수 있다.

〈표 2〉 자연어 색인 기반 지식베이스의 복수 질의어 확장

(초기 질의어(학습 AND 행동)의 질의 확장)

동시 출현 빈도	색인어	동시 출현 빈도	색인어
26	학습 행동	13	학생
11	인간	8	수업 형성
7	지식 학교		
5	교육 반응 변화 이해 태도 학습이론 학습자 형태 환경		
4	강화 교수학습과정 성취 수업목표 신경 연구 의미 자아 자연 정의 출발점 통합		
3	교사 교수 목적 심리 아동 자극 정신 조건 특성		

나. 통제어색인 기반 지식베이스

통제어색인을 이용한 지식베이스는 교육학 관련 통제어 형태의 질의어에 대한 연관용어를 제시하기 위해 구축한 질의어 확장 시스템이다. 교육학과 관련한 통제어를 질의어 입력창에 입력하면 지식베이스에 형성된 연관용어 관계를 이용해 입력된 질의어와 밀접한 관계를 갖고 있는 연관용어들을 자동으로 제시해 이용자가 특정 연관용어를 선택함에 의해 질의 확장 기능을 할 수 있도록

구성되어 있다. 〈표 3〉은 초기 질의어 '학습이론'과 의미적으로 밀접한 관계를 갖는 연관용어들의 리스트이다.

통제어색인을 이용한 지식베이스는 〈표 3〉과 같이 제시되는 연관용어들을 대상으로 인간의 두뇌 작업을 통해 특정 용어를 선택함에 의해 초기 질의를 확장하거나 축소할 수 있다. 〈표 3〉에 제시된 연관용어를 분석해보면 대부분의 연관용어들이 초기 질의어 '학습이론'과 개념적으로 밀접한 관계를 갖는 교육학 분야의 전공용어임을 알 수 있다. 따라서 통제어색인을 이용한 지식베이스를 구축하면 초기질의어와 밀접한 관계를 갖는 성능이 좋은 연관용어 관계를 형성할 수 있다. 그러나 통제어색인을 이용한 지식베이스의 단점은 초기질의어를 자연어 형태의 일상적인 용어를 입력하면 결과를 확인할 수 없다는 점이다. 그러나 특정 주제에 대해 정보탐색을 할 경우 이용자가 사용하는 초기질의어는 대부분 전공용어임을 감안한다면 결과를 확인할 수 없는 경우는 거의 없을 것이다.

〈표 3〉 통제어 색인 기반 지식베이스의 질의어 확장

(초기 질의어(학습이론)의 질의 확장)

동시 출현 빈도	색인어	동시 출현 빈도	색인어
8	학습이론	5	교수학습과정
4	조건반응	3	조건반사설 파블로프
2	고전적조건반사설 교육 동물실험 보상자극 스키너 스키너강화이론 스키너상자 시행착오학습 쏜다이크 연합설 왓슨 자극반응연합설 조건반사 조건형성 조작적조건형성 조작적조건형성이론 조작행동 프로그램학습 행동주의이론		
1	강화자극 반응행동 발생빈도 수정기술 시행착오설 외부자극 의도적행동 자극 정신질환 착오반응 프로그램 학습 행동반응		

통제어색인을 이용한 지식베이스는 복수의 용어를 질의어로 선택했을 경우, 〈표 3〉과 같이 동시출현빈도는 낮지만 보다 의미적으로 밀접한 관계를 갖는 연관용어를 제시해줄 수 있다.

〈표 4〉 통제어 색인 기반 지식베이스의 질의어 확장

(초기 질의어(학습이론 AND 교수학습과정)의 질의 확장)

동시 출현 빈도	색인어	동시출현빈도	색인어
5	교수학습과정 학습이론	2	조건반응
1	강화자극 고전적조건반사설 교수 교수법 교실사태-교수연구 다의적-학습 문제해결학습 반응행동 보상자극 스키너 스키너강화이론 스키너상자 시행착오학습 쏜다이크 연합설 왓슨 의도-무의도적-학습 의도적-교수 일의적-교수 자극반응연합설 조건반사 조건반사설 조건형성 조작적조건형성 조작적조건형성이론 조작행동 지각이론 퀼러 통찰설 파블로프 프로그램학습 학습 학습지도 행동주의이론 형태주의자 형태주의자		

Ⅳ. 색인 파일 기반 질의어 확장시스템의 평가

본 연구에 의해 구축된 두 개의 질의어 확장시스템 - 자연어 색인 기반 지식베이스와 통제어 색인 기반 지식베이스의 성능에 대한 평가를 수행하였다. 평가 항목은 교육학 분야 전공용어로서의 적합성, 초기질의어와 연관용어간의 의미 연관성, 제시된 연관용어를 이용한 인터넷 정보검색의 적합성, 교육학 분야 학습의 지원 가능성 등이다. 설문조사는 이용자가 본 시스템을 직접 작동하면서 〈부록 1〉에 나타난 바와 같이 8개 항목으로 이루어진 설문에 응답하도록 하였다.

이들 평가 항목을 이용하여 본 시스템을 구현할 때 규정한 다음과 같은 다섯 가지 가설을 입증하는 작업을 수행하였다.

첫째, 지식베이스 구현을 위해 사용한 색인어는 수록된 내용을 정확히 표현한 색인어이다.(주제색인어의 질적 수준)

둘째, 지식베이스 구현을 위한 색인어 추출방법은 통제어 색인방법이 자연어 색인방법보다 우수할 것이다.(색인어 추출방법에 따른 성능 차이)

셋째, 주제색인 파일 기반 질의어 확장시스템의 구축 방법에 대한 가설은 동일 문헌 내에 동시에 출현하는 색인어들은 의미적으로 밀접한 관계를 갖는 용어들이다. 따라서 이 가설을 근거로 구축한 질의어 확장시스템에서 제시하는 용어들은 상호간 개념적으로 밀접한 관계를 갖게 될 것이다.(연관용어의 개념적 연관성)

넷째, 주제색인 파일 기반 질의어 확장시스템을 이용하면 효율적인 정보검색이 가능할 것이다.(효율적인 정보검색 지원 시스템의 가능성)

다섯째, 주제색인 파일 기반 질의어 확장시스템은 특정 학문의 교수학습 지원 시스템의 기능을 발휘할 것이다.(학습지원 시스템의 가능성)

이와 같은 다섯 가지 가설을 입증하기 위해 본 연구 과정 중 구축한 2개의 시스템을 시연한 후 창원전문대학에 재학 중인 문헌정보학과 2학년 학생 22명과 유아교육과 3학년 학생 27명을 대상으로 설문조사를 실시하였다. 교육학 전공 교수와 문헌정보학 전공 교수들의 응답은 통계 분석에 의미 있는 사례 수에 미치지 못하여 분석에서 제외하기로 한다.

평가 설문에 대한 통계적 처리는 통계 프로그램인 SPSS 12.0을 이용하였다. 색인어 추출방법의 성능 차이에 대한 가설의 효과성 검증은 t-test pair로 검증하였으며, 기타의 4가지 가설에 대한 평가는 카이 제곱(X^2) 검증 방법을 실시하여 입증하였다.

1. 주제 색인어의 질적 수준

주제 색인어의 질적 수준은 첫 번째 가설인 “지식베이스 구현을 위해 사용한 색인어는 수록된

내용을 정확히 표현한 색인어이다."란 가설을 입증하는 방법으로 처리하였다. 설문 문항에서 색인용어의 전문성, 특정성, 망라성을 근거로 분석하였다. 구현한 색인 파일 기반 질의어 확장시스템은 주제색인어의 전문성(전공용어 성격), 특정성, 망라성을 근거로 색인어의 질적 수준을 평가하였다. 그 결과는 〈표 5〉, 〈표 6〉, 〈표 7〉과 같았다.

가. 색인어의 전문성

"본 연구에서 구현한 시스템에 수록되어 있는 색인어는 교육학 분야의 전공용어이다."란 가설을 입증하기 위한 방법은 카이 제곱(X^2) 검증 방법을 사용하였다. 검증의 결과는 〈표 5〉와 같다.

〈표 5〉 색인용어의 전문성 - X^2검증

( )안은 %

색인어	전적 비동의	비동의	보통	동의	전적 동의	합계	X^2
자연어	0(0)	3(6.1)	10(20.4)	18(36.7)	18(36.7)	49(100)	12.80***
통제어	1(2.0)	1(2.0)	14(28.6)	13(26.5)	20(40.8)	49(100)	29.27***
전 체	1(1.0)	4(4.1)	24(24.5)	31(31.6)	38(38.8)	98(100)	54.96***

*** $p < .001$

자연어의 전문용어 검증은 $X^2 = 12.80$, 통제어의 전문용어 검증은 $X^2 = 29.27$로 나타났으며, 자연어는 73%, 통제어는 67%가 전문용어라고 답하였다. 자연어와 통제어를 구분하지 않고 이 시스템에서 사용한 색인어의 전문용어 성격에 대한 검증은 $X^2 = 54.96$으로 전체 응답 중 약 70% 이상이 색인어의 전문성에 대해 긍정적인 반응을 보였으며, 부정적인 반응은 5%로 나타났다($p < .001$). 따라서 본 시스템에서 적용한 색인어는 전문성의 성격이 강한 교육학 분야의 전공용어임이 증명되었다. 표 내에서 *** 표시는 표 아래의 *** $p < .001$의 유의미한 조건을 만족하는 항목이다.

나. 색인어의 특정성

"본 연구에서 구현한 시스템에 수록되어 있는 색인어는 원문의 세부적인 내용까지도 표현한 특정성이 높은 색인어이다."란 가설을 입증하기 위한 방법은 카이 제곱(X^2) 검증 방법을 사용하였다. 검증의 결과는 〈표 6〉과 같다.

색인어의 특정성에 대한 평가에서 자연어와 통제어 모두 70% 이상이 긍정적인 응답을 하였으며, 부정적인 응답은 8%로 나타나 본 시스템에서 사용한 색인어는 원문의 세부적인 내용까지 정확히 표현한 용어임이 입증되었다($p < .001$).

〈표 6〉 색인어의 특정성 - X^2 검증

( )안은 %

색인어	전적 비동의	비동의	보 통	동 의	전적 동의	합 계	X^2
자연어	1(2.0)	2(4.1)	11(22.4)	16(32.7)	19(38.8)	49(100)	26.82***
통제어	2(4.0)	3(6.1)	13(26.5)	12(24.5)	19(38.8)	49(100)	34.88***
전 체	3(3.0)	5(5.1)	24(24.5)	28(28.6)	38(38.8)	98(100)	75.51***

*** $p < .001$

다. 색인어의 망라성

"본 연구에서 구현한 시스템에 수록되어 있는 색인어는 원문의 내용을 적절하게 표현한 망라성이 높은 색인어이다."란 가설을 입증하기 위한 방법은 카이 제곱(X^2) 검증 방법을 사용하였다. 검증의 결과는 〈표 7〉과 같다.

〈표 7〉 색인어의 망라성 - X^2 검증

( )안은 %

색인어	전적 비동의	비동의	보 통	동 의	전적 동의	합 계	X^2
자연어	0(0)	4(8.2)	12(24.5)	19(38.8)	14(28.6)	49(100)	9.53*
통제어	1(2.0)	2(4.1)	16(32.7)	22(44.9)	8(16.3)	49(100)	33.56***
전 체	1(1.0)	6(6.1)	28(28.6)	41(41.8)	22(22.4)	98(100)	54.35***

* $p < .05$, *** $p < .001$

색인어의 망라성에 대한 긍정적인 응답은 64.2%이며, 부정적인 응답은 7.1%로 나타났다($p<.001$). 따라서 본 시스템에서 사용한 색인어는 자연어와 통제어 모두 수록되어 있는 내용에 대해 적절한 수준으로 색인작업을 수행한 망라성이 높은 색인어인 것으로 증명되었다.

이상과 같은 색인 용어의 전문성, 특정성, 망라성에 대한 검증 결과 본 시스템에서 사용한 색인어들은 특정성과 망라성을 보장할 수 있는 교육학 분야의 전문적인 용어이다.

2. 색인어 추출방법의 성능 차이

색인어 추출방법의 차이가 지식베이스의 성능에 영향을 줄 것이란 판단으로 본 연구에서는 두 개의 각기 다른 시스템을 구현하였다. 두 개 시스템의 성능 평가는 "지식베이스 구현을 위한 색인어 추출방법은 통제어색인방법이 자연어 색인방법보다 우수할 것이다."란 가설에 대한 검증 결과와 동일하다. 가설에 대한 검증방법은 t검증 방법을 사용했으며, 검증에 대한 분석 결과는 〈표 8〉의 내용과 같다.

〈표 8〉 색인어 추출방법에 따른 성능 차이 검증 - t 검증

색인어	평 균	표준편차	t	학 과	평 균	표준편차	t
자연어	32.78	4.37	.29	문헌정보과	33.90	4.01	1.69
				유아교육과	31.85	4.50	
통제어	32.57	4.58		문헌정보과	33.23	4.91	.91
				유아교육과	32.04	4.31	

자연어 색인방법에 대한 반응점수는 32.78점, 통제어 색인방법은 32.57점으로 색인어 추출방법에 따른 성능 차이에 대한 응답점수에는 차별성이 없었으며, 통계적으로도 유의미한 차이가 없었다(t=.29, p〉.05). 또한 학과간의 차이검증에서도 자연어 색인방법과 통제어 색인방법의 차이에 따라 반응점수에 약 1~2점의 차이가 나타났으나, 통계적으로 유의미한 차이를 보이지는 않았다(p〉.05). 따라서 색인방법의 차이에 따른 질의어 확장시스템의 성능에는 의미 있는 차별성이 없었다.

일반적으로 이론적 측면에서 통제어색인이 자연어색인보다 정보검색의 효율성이란 측면에서 우수한 것으로 알려져 있었는데, 본 지식베이스에서는 이에 대한 차이를 입증할 수 없었다. 그 이유는 저자가 자연어 색인의 추출 작업시 임의적으로 통제어형태의 색인어를 추가했기 때문인 것 같다. 또 다른 이유는 설문 응답자가 자연어색인과 통제어색인의 차이를 명확히 인식하지 못했기 때문인 것으로 짐작된다. 따라서 통제어 색인 기반 질의어 확장시스템이 자연어 기반 질의어 확장 시스템보다 성능이 우수할 것이란 가설은 본 연구에서는 기각되었다.

3. 연관용어의 개념적 연관성

본 연구에서 구현한 주제색인 파일을 근거한 연관용어 선정의 정확성에 대한 평가는 "동일 문헌내에 동시에 출현하는 색인어들은 의미적으로 밀접한 관계를 갖는 용어들일 것이다. 따라서 주제색인 파일과 and 연산자를 이용해 자동으로 처리된 연관용어 관계는 개념적으로 밀접한 관계일 것이다"란 가설을 X^2검증방법을 이용하여 입증하는 방법으로 처리하였다. 이에 대한 검증 결과는 〈표 9〉와 같다.

〈표 9〉 연관용어의 개념적 연관성 - 카이 제곱(X^2) 검증

( )안은 %

색인어	전적 비동의	비동의	보통	동의	전적 동의	합계	X^2
자연어	0(0)	3(6.1)	15(30.6)	11(22.4)	20(40.8)	49(100)	12.80**
통제어	0(0)	1(2.0)	9(18.4)	12(24.5)	27(55.1)	49(100)	28.95***
전체	0(0)	4(4.1)	23(23.5)	24(24.5)	47(48.0)	98(100)	37.92***

* p〈.05 ** p〈.01 *** p〈.001

자연어 색인 기반 연관용어의 개념적 밀접성에 대해 63.2%의 학생이 우수하다는 반응을 보여 자연어 색인 기반 연관관계 선정방법은 그 가설이 적합하였다($X^2=12.80$, $p<.01$).

통제어 색인 기반의 연관용어들의 개념적 밀접성에 대한 평가는 매우 높게 나타났으며, 자연어에 의한 적합성 응답비율보다 높았다($X^2=28.95$, $p<.001$).

자연어와 통제어의 구분 없이 지식베이스에 의한 선정된 연관용어들의 개념적 밀접성은 72.5%의 반응율로 적합하다는 반응을 보였으며, 적합하지 않다는 반응(4.1%)은 매우 낮았다($X^2=37.92$, $p<.001$). 따라서 주제어 색인 파일과 AND 연산자를 이용한 연관용어 선정방법은 타당한 방법임이 입증되었다.

4. 정보검색 지원 시스템의 가능성

주제색인 파일 기반 질의어 확장시스템의 효율적인 정보검색 지원 가능성은 "주제색인 파일 기반 질의어 확장시스템을 이용하면 효율적인 정보검색이 가능할 것이다."란 가설을 정보검색의 편이성과 정보검색의 효율성(정도율, 재현율) 관련 항목으로 평가하였다. 가설 검증 방법은 카이 제곱(X^2) 검증을 사용하였으며 이들 항목에 대한 검증 결과는 〈표 10〉, 〈표 11〉, 〈표 12〉와 같다.

가. 정보검색의 편리성

주제색인 파일 기반 질의어 확장시스템을 인터넷 정보검색에 이용하면 편리할 것이란 질의에 대한 응답은 〈표 10〉과 같이 자연어 색인 방법이 약 77%의 학생이 편리하다고 응답하였다($X^2=44.78$, $p<.001$). 또한 통제어 색인 방법도 자연어 색인 방법과 비슷한 결과를 보였다($X^2=19.98$, $p<.001$). 학과와 색인어 추출방법의 구분 없이 본 연구에서 구축한 시스템이 정보검색에 편리한 기능을 제공한다는 검증 결과가 나왔다($X^2=80.98$, $p<.001$). 따라서 주제 색인 파일 기반 질의어 확장 시스템은 인터넷 정보검색시 이용자의 편의성을 보장하는 방법임이 입증되었다.

〈표 10〉 인터넷 정보검색의 편리성 - 카이 제곱(X^2) 검증

( )안은 %

색인어	전적 비동의	비동의	보통	동의	전적 동의	합계	X^2
자연어	2(4.08)	1(2.04)	8(16.3)	11(22.5)	27(55.1)	49(100)	44.78***
통제어	2(4.1)	0	11(22.4)	12(24.5)	24(49.0)	49(100)	19.98***
전체	4(4.1)	1(1.0)	19(19.4)	23(23.5)	51(52.0)	98(100)	80.98***

*** $p<.001$

나. 정보검색의 정확성 보장

주제색인 파일 기반 질의어 확장시스템을 인터넷 정보검색에 이용하면 정확한 정보를 찾는데 도움이 될 것이란 질의에 대한 응답은 〈표 11〉의 결과와 같다. 색인 방법의 구분 없이 전체적으로 정확하였다는 응답이 약 80% 내외의 반응을 보였으며, 정확하지 않다는 대답은 4% 미만으로 답하였다(자연어 - $X^2=19.82$, 통제어 $X^2=20.96$, $p<.001$). 주제색인 파일 기반 질의어 확장시스템을 이용하니 필요한 정보를 정확하게 찾을 수 있다는 응답이 매우 높았으며, 통계적으로 매우 유의미하였다($X^2=38.24$, $p<.001$) 따라서 주제색인 파일 기반 질의어 확장시스템은 인터넷 정보검색 시 이용자에게 정확한 정보를 찾을 수 있는 방법임이 입증되었다.

〈표 11〉 정보검색의 정확성 보장 - 카이 제곱(X^2)검증

( )안은 %

색인어	전적 비동의	비동의	보통	동의	전적 동의	합계	X^2
자연어	0(0)	1(2.0)	9(18.4)	20(40.8)	19(38.8)	49(100)	19.82***
통제어	0(0)	2(4.1)	9(18.4)	14(28.6)	24(49.0)	49(100)	20.96***
전체	0(0)	3(3.1)	18(18.4)	34(34.7)	43(43.9)	98(100)	38.24***

*** $p<.001$

다. 정보검색의 재현성 보장

주제색인 파일 기반 질의어 확장시스템을 인터넷 정보검색에 이용하면 누락된 정보가 없이 재현율이 높은 정보검색을 보장할 수 있을 것이란 가설에 대한 검증 결과는 〈표 12〉와 같았다. 재현율 보장에 대한 검증 결과는 자연어 색인방법과 통제어 색인 방법 모두 75% 이상이 동의했으며, 8% 정도가 동의하지 않는 것으로 나타났다. 또한 정보검색의 재현성 보장에 대한 가설 검증 결과는 타당한 것으로 검증되었다($X^2=68.74$, $p<.001$). 따라서 주제색인 파일 기반 질의어 확장시스템을 인터넷 정보검색에 이용하면 정보검색에 있어서 재현율이 높은 결과를 보장할 수 있는 것으로 입증되었다.

〈표 12〉 인터넷 정보검색의 재현성 - 카이 제곱(X^2)검증

( )안은 %

색인어	전적 비동의	비동의	보통	동의	전적 동의	합계	X^2
자연어		4(14.8)	7(14.3)	16(32.7)	22(44.9)	49(100)	16.71***
통제어	1(2.0)	3(6.1)	8(16.3)	12(24.5)	25(51.0)	49(100)	37.02***
전체	1(1.0)	7(7.1)	15(15.3)	28(28.6)	47(48.0)	98(100)	68.74***

*** $p<.001$

5. 학습지원 시스템의 가능성

"본 연구에서 구현한 주제색인 파일 기반 질의어 확장시스템은 특정 학문의 교수학습 지원 시스템의 기능을 발휘할 것이다"란 가설에 대한 X^2 검증의 분석 결과는 〈표 13〉과 같다. 학습지원 시스템의 가능성에 대한 평가에서 색인 방법에 따른 성능 차이는 유의미한 결과가 나타나지 않았다. 그러나 약 70~80% 학생은 이 시스템이 학습이나 개념 파악에 도움이 될 것이란 긍정적인 반응을 하였다($X^2=88.43$, $p<.001$).

〈표 13〉 학습지원 시스템의 가능성 - 카이 제곱(X^2)검증

( )안은 %

색인어	전적 비동의	비동의	보통	동의	전적 동의	합계	X^2
자연어	1(2.0)	1(2.0)	5(18.5)	15(30.6)	27(55.1)	49(100)	51.10***
통제어	2(4.1)	3(6.1)	7(14.3)	11(22.4)	26(53.1)	49(100)	38.65***
전체	3(3.1	4(4.1	12(12.2)	26(26.5)	53(54.1)	98(100)	88.43***

*** $p<.001$

앞의 다섯 가지 평가 항목 - 주제색인어의 질적 수준, 색인 작성방법의 성능 차이, 연관용어의 개념적 연관성, 정보검색 지원 가능성, 학습지원 시스템의 가능성 등에 대한 평가 결과를 종합하면 다음과 같다.

본 연구에서 구현한 주제색인 파일 기반 용어 확장 시스템은 수록되어 있는 색인어가 교육학 분야의 전공 용어 성격이 강한 것으로 나타났다. 이와 같은 성격의 전공용어들을 이용해 자연어 색인 기반 용어 확장 시스템과 통제어 기반 용어 확장 시스템에 대한 성능 평가를 한 결과 연관용어 관계, 정보검색의 효율성, 학습지원 가능성 등에서 유의미한 차이는 없는 것으로 나타났다.

색인 파일 기반 지식베이스를 이용한 초기질의어와 연관용어들의 개념상의 관계는 매우 밀접한 것으로 나타났다. 또한 이 지식베이스를 이용한 질의어 확장 시스템은 정보검색시스템에 적용시킨다면 매우 효율성이 높은 정보 검색 결과를 보장받을 수 있는 것으로 나타났다. 그리고 이 지식베이스는 교수학습 과정에 있어서 유용한 학습지원 시스템으로서 역할을 수행할 수 있는 것으로 나타났다.

이상과 같은 평가를 종합하면 색인 파일 기반 질의어 확장 방법이 정보검색 지원 시스템과 학습지원 시스템으로서 유용한 지식베이스 시스템의 구축 방법으로 타당한 방법임을 알 수 있었다.

V. 결 론

본 연구에서는 정보검색의 효율성을 보장해줄 수 있는 용어 확장 방법에 대한 선행 이론을 조사하였다. 용어 확장 방법인 지식 기반 확장방법과 탐색 결과 기반 확장방법에 대한 간략한 비교와 함께 지식기반 용어확장 방법인 개념 기반 구축방법과 통계적 유사도 기반 구축방법의 문제점을 제시하였다.

본 연구에서는 기존의 지식기반 용어확장 방법이 아닌 새로운 알고리즘으로 지식기반 용어확장 방법을 제시하였다. 그 방법은 색인 파일을 이용한 용어 확장방법이다. 색인 파일 기반 용어 확장 방법은 앞의 개념 기반 구축방법과 통계적 유사도 기반 구축방법의 복잡한 알고리즘을 적용함에 따른 시간, 노력, 비용, 컴퓨터 용량 등의 문제점을 해소할 수 있으며 유사한 수준의 성능을 보장할 수 있는 방법이다.

본 연구에서는 교육학개론이란 단행본을 근거로 색인파일을 이용한 지식베이스를 구축하였고 이를 이용해 교육학과 관련된 연관용어 관계를 형성할 수 있었다. 지식베이스는 색인어의 망라성-특정성 수준에 따른 연관용어 관계 형성 수준을 분석하기 위해서 자연어 색인을 이용한 지식베이스와 통제어 색인을 이용한 두 개의 지식베이스를 구축하였다.

구축한 두 개의 지식베이스는 성능 측정을 위해 교육학 전공 교수와 유아교육학 전공 학부 학생들을 대상으로 주제 색인어의 질적 수준을 평가하였다. 또한 이들 주제색인어들로 구성된 색인파일을 이용해 구축한 지식베이스에 대하여 용어 연관성, 정보검색 지원 가능성, 교수학습 지원 가능성 등에 대해 평가한 결과 우수한 성능을 갖고 있는 것으로 입증되었다.

따라서 본 연구에서 수행한 주제색인 파일 기반 질의 확장 시스템의 구축방법은 기존의 방법보다 비용이나 시간적인 측면에서 효율적인 용어확장시스템을 구현할 수 있다고 판단된다. 또한 본 연구에서 수행한 색인파일을 근거로 구축한 용어 확장 알고리즘은 정보검색시스템과 다양한 학습 인지시스템에서 유용한 알고리즘이 될 것으로 판단된다.

다만 본 연구에서 자연어 색인방법과 통제어색인 방법에 따른 용어확장 시스템들의 성능이 현격한 차이가 없다는 결과에 대한 정확한 원인 분석은 추후의 연구과제로 할 예정이다.

〈참고문헌은 각주로 대신함〉

〈부록 1〉 질의 확장용 시스템 평가용 설문 문항

(시스템 평가용 설문)

소속 : 직위 : (교수, 학생)

전공 : 학년 :

내 용	용어	전적 동의				전적 비동의
1. 탐색어 확장시스템에 출현한 용어는 교육학 분야의 전공용어들입니까?(색인 용어의 전문성)	자연어	5	4	3	2	1
	통제어	5	4	3	2	1
2. 초기 입력어(탐색어)와 관련해서 제시된 용어들은 교육학분야의 세부적인 내용을 잘 표현하고 있습니까?(색인어의 특정성)	자연어	5	4	3	2	1
	통제어	5	4	3	2	1
3. 초기 입력어(탐색어)와 관련해서 제시된 용어들은 원문을 근거로 분석했을 때, 주제를 표현하는 적절한 수준의 색인어라고 생각합니까?(색인어의 망라성)	자연어	5	4	3	2	1
	통제어	5	4	3	2	1
4. 초기 입력어(탐색어)를 근거로 제시된 용어들은 초기 입력어(탐색어)와 밀접한 의미 관계가 있는 용어들입니까?(연관 용어의 개념적 연관성)	자연어	5	4	3	2	1
	통제어	5	4	3	2	1
5. 본 시스템이 인터넷에 장착되면 정보검색의 초보자부터 전문가까지 편리하게 이용할 수 있다고 생각하십니까?(인터넷 정보 검색의 편리성)	자연어	5	4	3	2	1
	통제어	5	4	3	2	1
6. 초기 입력어(탐색어)와 제시된 관련 용어들은 인터넷 정보검색시 정확한 정보를 찾는데 도움이 되겠습니까?(인터넷 정보검색의 적합성 - 정확성)	자연어	5	4	3	2	1
	통제어	5	4	3	2	1
7. 초기 입력어(탐색어)와 제시된 관련 용어들은 인터넷 정보검색시 관련 정보를 많이 찾는데 도움이 되겠습니까?(인터넷 정보검색의 적합성 - 재현성)	자연어	5	4	3	2	1
	통제어	5	4	3	2	1
8. 초기 입력어(탐색어)와 제시된 관련 용어들은 교육학의 전공 학습이나 개념 파악에 도움이 되겠습니까?(교육학 분야 학습 지원 가능성)	자연어	5	4	3	2	1
	통제어	5	4	3	2	1