연관규칙을 이용한 문헌정보학 전문용어 클러스터링 기법에 관한 연구*

A Clustering Technique Using Association Rules for The Library and Information Science Terminology

승 현 우(Hyon-Woo Seung)**
박 미 영(Mi-Young Park)***

목 차

1. 서 론
 1. 1 연구배경 및 필요성
 1. 2 논문의 구성
2. 관련연구
 2. 1 데이터마이닝의 정의 및 필요성과 중요성
 2. 2 데이터마이닝 지식발견 프로세스의 단계
 2. 3 텍스트마이닝을 이용한 클러스터링 기법
 2. 4 텍스트마이닝을 이용한 대표 색인어 추출기법
3. 전문용어 클러스터링 기법
 3. 1 전문용어 클러스터링 모델
 3. 2 전문용어 추출을 위한 전처리 과정
 3. 3 연관규칙 탐사를 이용한 전문용어 클러스터링
 3. 4 전문용어 클러스터링 모델의 데이터베이스 설계
4. 실험 결과 및 고찰
 4. 1 실험방법
5. 결론 및 향후 연구방향

초 록

본 논문에서는 대량의 웹 문서로부터 연관된 지식정보를 검색하기 위한 전문 검색엔진을 개발하기 위하여 텍스트에서 추출된 전문 용어를 효율적으로 클러스터링하기 위한 방법을 제안하고자 한다. 즉, 일반적인 용어들간의 무의미한 연관 규칙이 양산되는 것을 방지하기 위하여 전문 용어로 구성된 지식베이스 테이블을 이용하여 의미 있는 용어들간의 연관 규칙을 생성한다. 연관 규칙은 하나의 논문에서 사용된 전문 용어들의 집합을 트랜잭션 단위로 구성하여 Apriori 알고리즘을 적용하여 생성된다. 하나의 용어로부터 생성된 연관 규칙 집합은 해당 전문 용어와 관련된 클러스터로 구성된다.

ABSTRACTS

In this paper, an effective method for clustering terminologies extracted from text is proposed, in order to develope a search engine to extract relevant information from large web documents. To prevent frequency of the meaningless association rules among general terminologies, only useful association rules among terminologies are produced using database tables which consist of domain-specific terminologies. Such association rules are produced by applying the Apriori algorithm after forming transaction units from groups of association rules in a document. A group of association rules produced from a terminology forms in a cluster.

키워드: 데이터마이닝, 연관규칙, 클러스터링, TF*IDF, Apriori 알고리즘
Data Mining, Association Rules, Clustering, TF*IDF, Apriori Algorithm

---

* 본 논문은 서울여자대학교 2003년 교내특별연구과제지원에 의해 수행된 것임.
** 서울여자대학교 정보통신공학부 컴퓨터공학전공 교수(hwseung@swu.ac.kr)
*** 서울여자대학교 정보영상학부 문헌정보학전공 강사(ollive@swu.ac.kr)
논문접수일자 2003년 5월 13일
게재확정일자 2003년 6월 15일

1. 서 론

1. 1 연구배경 및 필요성

정보처리시스템과 통신기술의 발전으로 인하여 지식기반 사회가 도래함에 따라 인터넷에서 제공되는 웹 문서가 급속도로 증가하고 이에 이용자가 적합한 지식 정보를 검색하는 것은 점점 더 어려워지고 있다. 따라서 대량의 웹 문서로부터 의미 있는 유용한 지식정보를 추출하기 위한 지식탐사 기법이 필요하며 이러한 지식 탐사 기법은 관련 문서를 체계적으로 분류하여 원하는 정보를 쉽게 탐색하거나 관련용어간의 유사도 정도에 따라 클러스터링하는 방법을 주로 사용한다.

하지만 인터넷에서 정보를 검색하기 위해 사용하는 일반 검색엔진들은 주로 단일 키워드를 중심으로 웹 문서에 대한 인덱스를 생성하기 때문에 의미적으로 연관된 전문 내용을 포함하는 문서를 검색하는 것은 거의 불가능하다.

본 논문에서는 대량의 웹 문서로부터 연관된 지식정보를 검색하기 위한 전문 검색엔진을 개발하기 위하여 텍스트에서 추출된 전문용어를 효율적으로 클러스터링하기 위한 방법을 제안하고자 한다. 즉, 일반적인 용어들간의 무의미한 연관 규칙이 양산되는 것을 방지하기 위하여 전문 용어로 구성된 지식베이스 테이블을 이용하여 의미 있는 용어들간의 연관 규칙을 생성한다. 생성된 연관 규칙으로 형성되는 하나의 용어에 의해 발견되는 연관 규칙 집합을 이용하여 의미적으로 관련된 전문 용어들끼리 클러스터링을 한다.

실험대상으로는 한국문헌정보학회, 한국정보관리학회, 한국도서관정보학회에 게재된 학술 논문에서 문헌정보학 관련 전문 용어를 기초로 초기 지식 베이스 테이블을 생성한다. 지식베이스 테이블 구축의 전처리과정은 형태소 분석기 HAM2001(http://nlp.kookmin.ac.kr/HAM/Kor/download.html)을 통하여 추출된 용어 중에서 일반적인 단어는 배제시키고 문헌정보학 분야와 관련된 전문용어를 추출한다. 추출된 전문용어에 TF*IDF 알고리즘을 적용하여 가중치를 부여하는 과정으로 이루어진다.

연관 규칙은 하나의 논문에서 사용된 전문용어들의 집합을 트랜잭션 단위로 구성하여 Apriori 알고리즘을 적용하여 생성한다. 하나의 용어에 의해 생성된 연관 규칙 집합은 해당 전문 용어와 관련된 클러스터로 정의하고, 문헌정보학 학술논문을 대상으로 전문용어를 추출하여 관련된 용어들끼리 클러스터를 구성하는 실험을 통하여 제안된 방법의 효율성을 증명하고자 한다.

1. 2 논문의 구성

본 논문의 구성은 다음과 같다.

2장에서는 본 시스템의 중점 기법인 데이터 마이닝 기술에 대해 설명하고 관련 연구와 기존 연구의 문제점을 분석한 뒤 3장에서는 본 논문에서 제안하는 전문용어 클러스터링 모델에 대한 고찰과 데이터베이스를 설계한다. 4장에서는 전문 용어 클러스터링 모델의 구현 환경과 실험을 거쳐 얻어진 실험결과를 통하여 본 논문에서 제안한 기법의 우수성을 평가한다. 마지막으로 5장에서는 결론을 기술하며

현재의 논문에서 보완해야 할 향후 연구방향을 제시한다.

2. 관련연구

2. 1 데이터마이닝의 정의 및 필요성과 중요성

1960년대 이래로, 데이터베이스와 정보 기술은 원시적인 파일 처리 시스템(file processing system)으로부터 정교하고 강력한 데이터베이스 시스템으로 체계적으로 진화하여 왔다.

1970년대 이후, 데이터베이스 시스템에 대한 연구와 개발은 초기의 계층형 데이터베이스 시스템(hierarchical database system)과 네트워크형 데이터베이스 시스템(network database system)으로부터 데이터를 관계 테이블 구조에 저장하는 관계 데이터베이스 시스템(relational database system), 데이터 모델링 도구 그리고 인덱싱(indexing) 및 데이터 조직 기법의 발달로 진보하였다.

그밖에도 사용자는 질의어, 사용자 인터페이스, 최적화된 질의처리(query processing) 그리고 트랜잭션 관리를 통하여 편리하고 융통성 있게 데이터에 접근할 수 있게 되었다. 질의를 하나의 읽기전용 트랜잭션(read-only transaction)으로 보는 OLTP(on-line transaction processing)를 위한 효율적인 방법들의 개발에 힘입어 관계 데이터베이스 기술은 비약적인 발전을 하였으며, 대량의 데이터를 효율적으로 저장, 검색, 관리하기 위한 주요 도구로서 널리 받아들여지게 되었다.

1980년대 중반 이후의 데이터베이스 기술은 관계 데이터베이스 기술의 발전과 데이터베이스 시스템에 관한 연구와 개발의 급증으로 특징지을 수 있다. 확장 관계(extended relational), 객체 지향(object-oriented), 객체관계(object-relational) 그리고 연역적(deductive) 모델 등과 같은 진보된 데이터 모델들이 연구 개발되었다. 공간(spatial), 시간(temporal), 멀티미디어(multimedia), 능동적(active) 그리고 과학(scientific) 데이터베이스, 지식베이스(knowledge base) 및 사무정보 베이스(office information base) 등을 포함한 다양한 응용 지향적 데이터베이스 시스템들이 발전하였다. 데이터의 분산, 다양화 그리고 공유와 관련된 문제들이 광범위하게 연구되었다. 또한, 이질(heterogeneous) 데이터베이스 시스템과 WWW와 같은 인터넷 기반의 범세계적 정보 시스템 등이 등장하였고 정보 산업의 중추적 역할을 맡고 있다(승현우 2002).

최근, 정보 산업 분야에서 데이터 마이닝이 주목을 받고 있는 주된 이유는 데이터의 양적 팽창과 그러한 데이터를 유용한 정보와 지식으로 바꿔야하는 시급한 필요성에 기인한다. 획득된 정보와 지식은 기업경영, 생산관리 그리고 시장분석에서부터 공학설계와 과학탐구에 이르기까지 광범위한 응용분야에 이용될 수 있다.

데이터마이닝은 정보기술 분야에 있어서 자연스런 진화의 결과로 볼 수 있는데, 데이터베이스 산업 분야에서는 데이터 수집 및 데이터베이스 구축, 데이터 저장과 검색 그리고 데이터베이스 트랜잭션 처리를 포함하는 데이터

관리, 그리고 데이터 웨어하우스, 그리고 데이터 마이닝과 관련한 데이터 분석과 이해와 같은 기능적 진화 경로를 찾아 볼 수 있다.

예를 들어, 초기의 데이터 수집 및 데이터베이스 구축 기법의 발전은 나중의 데이터 저장과 검색 그리고 질의와 트랜잭션 처리 등을 위한 효과적인 기법의 발전을 위한 선행요건으로 작용하였다. 질의와 트랜잭션 처리를 통상적으로 제공하는 수많은 데이터베이스 시스템들과 더불어 데이터 분석과 이해는 자연스럽게 다음 목표가 되었다.

데이터마닝의 기법은 기존에 알려진 정보뿐만 아니라 쉽게 드러나지 않는 숨은 정보까지 데이터베이스로부터 찾아내고자 하는 정보 추출 방법론의 하나이다. 이러한 데이터 마이닝의 기법은 질의 도구(query tools), 통계적 기법(statistical technique), 가시화(visualization), 온라인 분석처리(OLAP, OnLine Analytical Processing), 사례기반학습(case based learning), 의사결정 트리(decisition tree), 연관규칙(association rule), 신경망(neural network), 유전자 알고리즘(genetic algorithm)등과 같은 다양한 접근방법에 의해 연구가 진행되고 있다(승현우 2002).

2. 2 데이터마이닝 지식발견 프로세스의 단계

데이터 마이닝은 데이터베이스나 데이터 웨어하우스 또는 그 밖의 다른 정보 저장소들에 저장되어 있는 대량의 데이터로부터 흥미로운 지식을 발견하는 과정이다. 데이터베이스에서의 지식 발견(Knowledge Discovery in Database : KDD)과 동의어로 취급하며 아래와 같은 지식 발견 절차를 나타내고 단계들의 반복적 연속으로 이루어져 있다(Jiawei Han 2001, 이란주 2001).

1. **데이터 정제** (잡음과 불일치 데이터의 제거)
2. **데이터 통합** (다수의 데이터 소스들의 결합)
3. **데이터 선택** (분석 작업과 관련된 데이터들이 데이터베이스로부터 검색된다)
4. **데이터 변환** (요약이나 집계 등과 같은 연산을 수행함으로써, 마이닝을 위해 적합한 형태로 데이터를 변환하거나 합병 정리한다)
5. **데이터 마이닝** (데이터 패턴을 추출하기 위하여 지능적 방법들이 적용되는 필수적 과정)
6. **패턴 평가** (몇 가지 흥미 척도들을 기초로, 지식을 나타내는 흥미로운 패턴들을 구별한다)
7. **지식 표현** (사용자에게 채굴된 지식을 보여주기 위하여 시각화와 지식 표현 기법들이 사용된다)

2. 3 텍스트마이닝을 이용한 클러스터링 기법

텍스트마이닝을 이용한 클러스터링 기법은 대상 집합에 따라 클러스터링의 기준이 되는 유사도를 다르게 정의한다. 먼저, 용어 클러스터링 기법에서는 주로 용어에 대한 총 빈도수와 한 문서에서 동시에 출현한 빈도수의 비율을 유사도로 사용한다. 문서 클러스터링 기법은 두 문서에서 추출된 용어의 총 개수와 동

시에 출현한 용어 개수의 비율을 유사도로 정의하여 클러스터링 한다(승현우 2002).

김호성(1992)등은 논문 제목에서 출현한 단어의 빈도와 단어간의 연관성에 의해서 용어를 분류하여 각 논문이 속하는 주제 분야를 분류할 수 있는 용어 클러스터링 시스템을 구현하였다. 하지만 주제와 관련 없는 용어가 제목에서 출현하거나 또 단순히 출현 빈도에 의해 클러스터를 구성하는 관계로 출현 빈도가 적은 용어는 클러스터에 포함되지 않은 문제점이 있다.

신진섭(2000)은 웹 상의 문서를 사용자 프로파일에 맞춰 분류하는 클러스터링 모델과 단어 연관성 모델을 제시하였다. 문서에 대한 대표 색인어를 찾기 위해 단어의 밀집성을 이용하였고, 두 단어가 같은 주제를 대표할 가능성에 대한 확률에 의해 연관성을 정의하였다.

Han(1997)등은 연관 규칙을 이용한 용어 클러스터링을 제시하였다. 연관 규칙에서의 최소 지지도를 만족하는 모든 빈발 항목집합을 대상으로 연관 규칙 하이퍼그래프(association rule hypergraph)를 생성하여, 생성된 하이퍼그래프를 신뢰도에 근거한 유사성 척도를 사용하여 분할하는 용어 클러스터링 알고리즘을 제시하였다.

서성보(2000)등은 트랜잭션에 대한 클러스터의 유사성을 측정하기 위해 주요 항목으로 구분하고, 각 트랜잭션의 최소 비용 계산을 통해 자동화된 문서 클러스터링 기법을 제안하였다. 하지만 주요 항목 집합의 기준으로 빈도수만 고려하여 무의미한 연관 규칙이 대량으로 발견될 수 있다.

이문기(2000)등은 클러스터링 시스템의 성능 향상을 위하여 같은 단어가 많이 나타나는 문서는 그만큼 서로 유사하다는 가정에 바탕을 두고 같은 범주 내의 문서들을 차별화 할 수 있는 문서 유사도식을 제안하여 웹 디렉토리 서비스를 위한 문서 클러스터링 기법을 제안하였다.

이정화(2000)는 전문검색엔진을 개발하기 위하여 텍스트에서 추출된 전문용어를 효율적으로 클러스터링 하기 위한 방법을 제안했지만 단일 명사로 구성된 용어만을 처리하였고 복합용어는 처리되지 못했다.

2. 4 텍스트마이닝을 위한 대표 색인어 추출 기법

비정형화된 대량의 문서를 대상으로 텍스트마이닝 기법을 적용하기 위한 핵심적인 기술 중의 하나는 문서를 대표할 수 있는 색인어를 효과적으로 추출하기 위한 방법이다. 대부분의 방법에서 추출된 단어의 순서보다는 문서 내에서 단어의 출현 여부나 빈도 수를 고려한 통계적인 정보를 기반으로 추출하고 있다.

문서에서 대표 색인어를 추출하기 위한 연구로는 일단 색인 대상 단어의 수를 줄이기 위해 어근 추출 알고리즘, 불용어 제거 알고리즘, 동의어 사전을 이용하는 방법과 문서에서 출현하는 각 용어에 대한 중요도를 확률적으로 계산하여 가중치를 조정하는 TF*IDF 알고리즘이나 키워드의 밀접성을 이용한 연구가 진행되었다. 또한, 각 주요 단어의 문서 길이에 따른 영향력 불균형을 해결해 주는 벡터 길이 정규화(vector length normalization) 알고리즘이 있다(Thorsten 1996).

전문 분야에 대한 지식 정보를 검색하기 위해 각 분야에서 사용되는 전문 용어를 효과적으로 추출하면, 관련 문서간의 유사성을 쉽게 판별할 수 있다. 이에 따라 문서에서 전문 용어를 효과적으로 추출하기 위한 연구도 진행되고 있다.

김호성(2000)등은 도서관 분류 체계에서 새로운 학문 분야를 반영하기 위하여 새로운 전문 용어의 개념을 자동으로 습득하여 용어 클러스터링 하는 방법을 제안하였다.

박정오(2000)등은 컴퓨터를 이용하여 전문 용어를 자동적으로 추출하기 위한 전문 용어 추출 시스템을 개발하였다. 이 시스템에서는 기존 문서에서 사용되는 특정어구를 이용하여 전문 용어를 추출하고, 후보 전문 용어에서 단어의 위치 정보를 이용하여 전문 용어를 추출하는 방법을 제안하였다.

3. 전문 용어 클러스터링 기법

3. 1 전문 용어 클러스터링 모델

본 논문에서 제안하는 클러스터링 모델은 크게 문서에서 전문 용어를 추출하기 위한 전처리 과정과 전문 용어간의 연관 규칙을 탐사하여 클러스터링 하는 과정으로 이루어진다. 다음 그림은 본 논문에서 제안하는 지식 베이스 테이블을 구축하기 위한 전문 용어 클러스터링 모델의 전체 구성도이다.

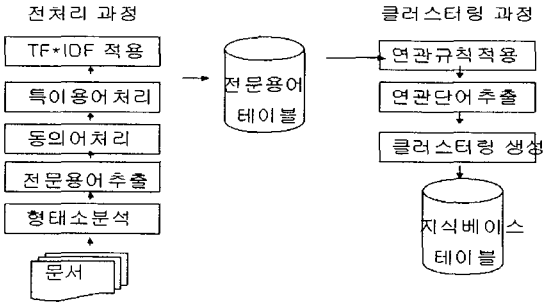
3. 2 전문 용어 추출을 위한 전처리 과정

본 논문에서 전문 용어를 추출하기 위한 실험 대상을 문헌정보학 분야의 논문으로 선택하였다. 따라서 본 과정은 문헌정보학 분야의 논문에서 문헌정보학에 관한 전문 용어만을 추출하기 위한 과정으로 전문 용어 추출 과정과 가중치 적용 과정으로 이루어진다.

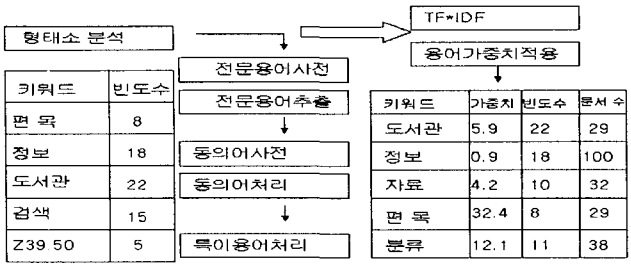
전문 용어 추출 과정은 형태소 분석 후 전문 용어를 추출하기 위한 과정, 동의어 처리 과정, 특이용어 제거 과정으로 이루어진다. 〈그림 2〉는 전처리 과정에 대한 전체적인 구성도이다.

3. 2. 1 전문 용어 추출

실험 대상 문서에 대한 형태소 분석을 통하여 문서에서 출현하는 모든 용어를 추출하였다. 형태소 분석기는 국민대학교 강승식 교수



〈그림 1〉 전문용어 클러스터링 모델 전체 구성도



〈그림 2〉 전처리 과정의 전체 구성도

팀이 개발한 공개용 형태소 분석기인 HAM 2001(http://nlp.kookmin.ac.kr/HAM/Kor/download.html)을 사용하였다.

일반 용어에 의해 생성되는 무의미한 연관 규칙을 배제하기 위해 형태소 분석기를 통해 추출된 실험 대상 문서의 모든 용어에 대해 숙명여자대학교 사공철 교수의 문헌정보학 용어 사전에 수록된 문헌정보학 분야의 전문 용어를 기준으로 전문 용어만을 추출하였다.

그리고 학술 논문에서 사용하는 전문 용어 중에서 같은 의미를 가진 용어이지만 저자에 따라 영어와 한국어를 혼용하거나 영문 용어를 한글화하는 과정에서 차이가 있는 관계로 동의어 사전을 구성하여 전문 용어를 표준화하였다. 예를 들어, '데이타베이스', '데이터베이스', 'database', '디지털', '디지탈' 'digital' 등과 같은 용어들에 대하여 하나의 대표 용어를 정의하여 표준화하였다.

또 전체 문서에서 출현하는 절대 빈도수가 매우 적어서 연관 규칙 탐사 대상이 되지 않는 용어와 용어들의 분포도가 매우 큰 관계로 무의미한 연관 규칙을 발생시킬 수 있는 용어를 특이 용어로 취급하여 연관 규칙 탐사과정에서 배제시켜 무의미한 연관 규칙의 양산을 방지하여 대표 색인어를 효율적으로 추출할 수 있도록 하였다. 이러한 용어는 연산시간을 낭비하고 최소지지도를 만족하지 못하여 연관 규칙으로 발견되지 않기 때문이다.

3. 2. 2 단어 빈도 가중치 조정

일반적으로 임의의 문서에서 그 문서를 대표할 수 있는 특징을 추출하기 위해서 단어의 빈도수(term frequency)를 많이 이용하고 있다. 그러나 한 문서에서 출현한 단어의 빈도수가 높다고 해서 그 문서를 정확히 대표하는 단어가 된다고 확신하기는 어렵다.

예를 들어 '정보'라는 용어는 문헌정보학 용어이고 빈도수는 높지만 대부분의 문헌정보학 관련 논문에서 공통적으로 출현하는 관계로 특정 문서를 대표하는 용어로 판정하기는 어렵다. 이러한 단어 빈도수의 문제점을 해결하기 위하여 여러 가지 가중치 공식들이 제안되었다.

본 논문에서는 TF*IDF 알고리즘을 적용하여 공통적으로 출현하는 단어에 대한 가중치를 조정하였다. TF*IDF 알고리즘은 역문서 빈도수(inverse document frequency)를 단어의 빈도수와 같이 적용함으로써 그 문서를 대표하는 단어들을 효율적으로 찾을 수 있는 알고리즘이다. 문서의 빈도 df_i는 문서들 중에서 단어 t_i가 존재한 문서의 개수를 의미

하며, 단어의 빈도 tf_{ij}는 문서 d_i에서 단어 t_i가 나타난 수를 의미한다. 이때 $\log(N/df_i)$는 역 문서 빈도수를 의미하며, 역 문서 빈도수와 단어 빈도수를 곱한 값을 문서 d_i에서 단어 t_i의 중요도 또는 영향력(weight)이라고 말하며 이를 w_{ij}라 한다(Thorsten 1996).

$$w_{ij} = tf_{ij} \times \log(N/df_i)$$

3. 2. 3 문서 길이 정규화

문서에서 추출된 단어들은 문서 길이에 따라 영향력을 달리 하기 때문에 문서 길이에 대한 정규화 과정이 필요하다. 일반적으로 벡터 길이 정규화(vector length normalization) 알고리즘을 이용하여 단어의 문서 길이에 따른 영향력 불균형을 해결하고 있다. 하지만 본 논문의 실험 대상 문서는 학술 발표 논문으로 문서 길이가 일정하므로 문서 길이 정규화는 고려하지 않았다.

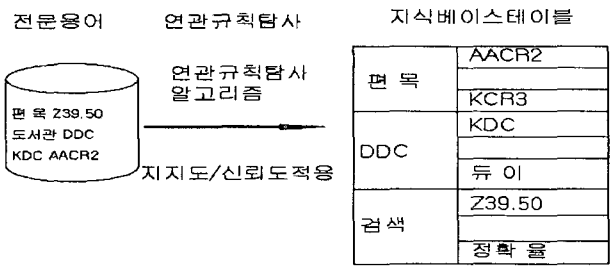
3. 3 연관 규칙 탐사를 이용한 전문용어 클러스터링

본 논문에서는 클러스터를 구성하기 위하여 데이터마이닝 기법의 장바구니 분석 과정에서 주로 사용하는 연관 규칙(association rule) 탐사 알고리즘 중 Apriori 알고리즘을 이용하여 전체 문서에서 추출된 전문 용어들간의 연관성을 분석하였다. 즉, 하나의 전문 용어와 연관된 용어는 최소 지지도와 신뢰도를 만족하는 연관 규칙의 결과로 구성된다. 각 전문 용어별로 구성된 클러스터는 지식베이스 테이블에 저장하여 지식 정보 검색엔진의 핵심적인 구성 요소로 사용하게 된다. 다음 그림은 연관 규칙 탐사 알고리즘을 이용하여 클러스터를 구성하는 과정에 대한 개념도이다.

3. 3. 1 연관 규칙의 정의

연관 규칙이란 데이터마이닝으로 얻어지는 여러 가지 지식 패턴 중의 하나로 1993년에 처음 소개되었다. 수많은 데이터마이닝 기법들 중에서 크게 주목받고 있는 중요한 지식 패턴의 하나로서 '어떤 사건이 발생하면 다른 사건이 일어난다'와 같은 연관성을 의미한다. 연관 규칙 탐사 알고리즘에서 하나의 장바구니에 담긴 상품 집합이나 단위 시간에 발생한 사건들의 묶음을 트랜잭션이라 정의한다.

연관 규칙 탐사란 이러한 트랜잭션 집합에서 미리 결정된 최소 지지도를 만족하는 빈발 항목집합들을 찾아내어 연관 규칙을 생성하는



〈그림 3〉 연관 규칙을 이용한 전문 용어 클러스터링

과정으로 이루어진다.

다음은 연관 규칙에 대한 정의이다. 먼저, $I=\{1, 2, 3, \cdots, m\}$을 항목들의 집합, D를 트랜잭션들의 집합이라 하면 각 트랜잭션 T는 $T \subseteq I$ 인 항목들의 집합이다. X가 한 트랜잭션에 포함된 항목들의 빈도를 고려하지 않은 항목들의 집합일 때 $X \subseteq T$이면 트랜잭션 T는 X를 포함한다. 이때 연관 규칙은 $R: X \rightarrow Y$ 형식의 함축이고, 이때 X와 Y는 서로 같은 원소를 갖지 않는 항목집합이다. 만일 한 트랜잭션이 X를 지지한다면, 어떤 확률에 의해 Y도 지지할 것이라고 예측하는 것이 연관 규칙이다. 이런 확률을 이 규칙에 대한 신뢰도(conf(R))라 한다. R의 신뢰도는 아래의 식처럼 X를 지지하는 T에 대하여 Y또한 지지할 조건부 확률로 정의된다.

$$conf(R) = p(Y \subseteq T \mid X \subseteq T)$$

$$= \frac{p(Y \subseteq T \mid X \subseteq T)}{p(X \subseteq T)} = \frac{supp(X \cup Y)}{supp(X)}$$

위에 있는 규칙 R에 대한 지지도는 $supp(X \cup Y)$로 정의한다. 지지도는 얼마나 자주 적용할 수 있는지를 나타내는 반면 신뢰도는 그 규칙이 얼마나 믿을만한지를 의미한다. 규칙이 데이터베이스에서 적절해지려면 충분한 지지도와 신뢰도를 가져야 한다. 그러므로 어떤 주어진 최소 신뢰도 C_{min}과 최소지지도 S_{min}에 대하여 $conf(R) \geq C_{min}$이고 $supp(R) \geq S_{min}$하면 규칙 R은 D에 대하여 성립한다고 할 수 있다. 규칙이 성립되기 위해 필요한 조건으로서 규칙의 조건부와 결과부는 모두 빈발해야 한다(R, Agrawal 1994).

3. 3. 2 연관 규칙 탐사 알고리즘

연관 규칙 탐사 알고리즘에서 연관 규칙을 탐사하는 과정은 다음과 같은 2단계로 이루어진다(박건호 1999).

1단계: 먼저, 빈발 항목집합(large item sets) l을 찾아낸다. 항목들의 전체집합 I의 부분집합이면서 몇 개의 항목들로 구성된 것을 항목집합이라 한다. 여기서 최소 지지도 S_{min} 이상의 트랜잭션 지지도를 가지는 항목집합들의 모든 집합들을 빈발 항목집합이라 한다.

2단계: 모든 빈발 항목집합 l에 대하여 l의 모든 공집합이 아닌 부분집합들을 찾는다. 이러한 부분집합 a에 대하여 supp(a)에 대한 supp(l)의 비율이 적어도 최소 신뢰도 C_{min}이상, 즉

$\frac{supp(l)}{supp(a)} \geq C_{min}$, $a \Rightarrow (l-a)$ 의 형태의 규칙을 연관 규칙으로 출력한다.

3. 3. 3 연관 규칙을 이용한 클러스터링

본 논문에서는 의미적으로 서로 연관된 전문 용어들끼리 클러스터로 구성하기 위하여 연관 규칙 탐사 알고리즘을 사용하였다. 하나의 논문에서 추출한 전문 용어들의 집합을 트랜잭션(transaction)이라 정의하고, 각 논문에서 추출된 전문 용어를 항목(item)으로 정의하였다. 일반적으로 지지도와 신뢰도 값의 증가에 따라 생성되는 연관 규칙의 수는 반비례 관계가 있다. 즉, 지지도와 신뢰도 값이 높을수록 조건을 만족하는 연관 규칙의 수는 줄어든다.

하나의 전문 용어에 대해 구성되는 클러스

터의 크기와 클러스터에 포함되는 용어를 결정하기 위한 명확한 기준을 정하기는 어렵다. 본 연구에서는 다양한 구간에 대한 지지도와 신뢰도를 만족하는 연관 규칙에 대하여 전문용어간의 연관성을 판정하여 최적의 클러스터로 정의하였다.

3. 4 전문 용어 클러스터링 모델의 데이터베이스 설계

데이터베이스 설계 과정은 개념적 데이터베이스 설계 단계와 논리적 데이터베이스 설계 단계로 이루어진다. 설계된 데이터베이스는 MS SQL 2000 SERVER 데이터베이스를 기반으로 스키마를 생성하였다.

3. 4. 1 개념적 데이터베이스 설계 및 ER 다이어그램

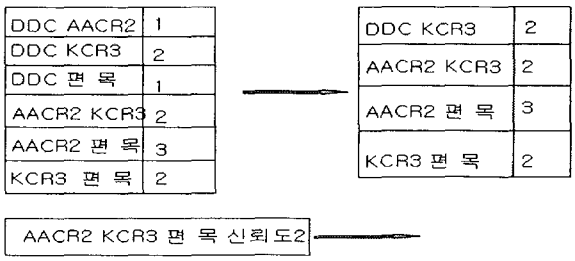
다음 그림은 본 논문의 전문 용어 클러스터링 모델을 위한 개념적 데이터베이스 설계도인 ER 다이어그램이다.

〈그림 6〉은 논문(Thesis), 색인어(Keyword), 전문용어(Major_Term), 동의어 사전(Same_Term), 연관규칙(Know-ledge)을 개체로 갖는 ER 다이어그램이다. 논문과 색인어 1:n, 전문용어와 색인어 1:n, 전문용어와 지식베이스 1:n, 전문용어와 동의어 1:n의 관계를 갖는다.

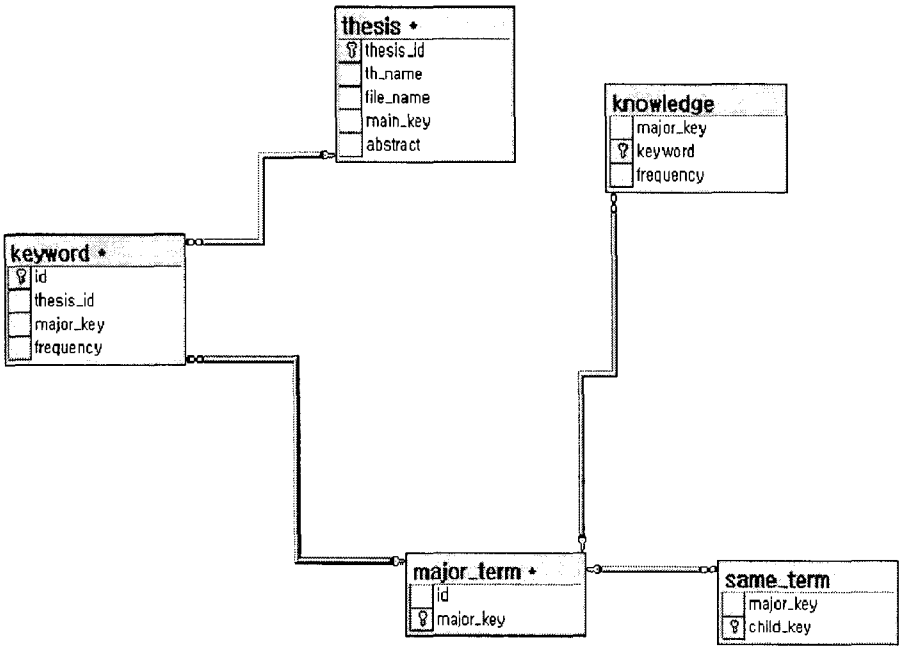
3. 4. 2 논리적 데이터베이스 스키마 설계

위의 ER 다이어그램을 토대로 다음과 같이 5개의 논리적 스키마를 구성하였다.

〈그림 4〉 연관규칙 탐사 알고리즘 1



〈그림 5〉 연관규칙 탐사 알고리즘 2



〈그림 6〉 전문 용어 클러스터링 모델의 ER 다이어그램

전문 용어 클러스터링 모델을 구성하는 테이블은 크게 실험 대상 논문에 대한 정보를 저장하기 위한 논문 테이블, 하나의 논문에서 추출된 용어들의 집합인 각 트랜잭션에서 발견된 키워드를 저장하기 위한 색인어 테이블, 모든 색인어 중에서 일반 용어는 배제시키고 문헌정보학 전문 용어만을 저장하는 전문용어 테이블, 영어나 한국어를 혼용하거나 영문 용어를 한글화하는 과정에서 발생하는 용어 표현의 차이를 해결하고 용어를 표준화하기 위해 필요한 동의어 사전 테이블, 연관 규칙을 가지는 키워드를 저장하기 위한 연관 규칙 테이블로 구성된다.

1) 논문 테이블(Thesis Table)

본 논문에서는 문헌정보학 관련 학회의 발표 논문을 실험 대상으로 하였다. 논문 테이블은 논문번호, 논문제목, 논문이 저장된 디렉토

〈표 1〉 논문 테이블 구조

Column Name	Data Type	Key Type	Null/Unique	Comment	Sample Data
Thesis_id	number(5)	Primary Key	NN	논문번호	1
Th_name	varchar2(100)			논문제목	디지털도서관...
File_name	varchar2(200)			저장디렉토리	paper1/1.txt
Main_key	varchar2(50)			전문용어	
abstract	varchar2(400)			초록	

리 경로 등과 같은 논문에 관한 세부적인 정보를 저장하기 위한 테이블이다.

2) 색인어 테이블(Keyword Table)

하나의 논문에서 추출된 용어들의 집합인 각 트랜잭션에서 발견된 키워드를 저장하기 위한 테이블이다. 본 논문에서는 실험의 대상인 논문에 대해 용어 분석을 위하여 국민대학교 강승식 교수팀이 개발한 공개 형태소 분석기인 HAM 2001을 이용하여 키워드를 추출하였다. 색인어 ID와 추출된 논문 번호, 색인어(키워드)와 발생 빈도수에 해당하는 항목이 저장된다.

3) 전문 용어 테이블(Major_term Table)

전문 용어 사전 테이블은 형태소 분석을 통하여 HAM 2001에서 추출된 모든 색인어 중에서 일반 용어는 배제시키고 문헌정보학 전문 용어만을 저장하는 테이블이다. 전문용어를 입력할 때 정렬된 값을 가지기 위하여 전문용어ID, 해당 전문용어가 저장된다.

4) 동의어 사전 테이블(Same_term Table)

영어나 한국어를 혼용하거나 영문 용어를 한글화하는 과정에서 발생하는 용어 표현의 차이를 해결하고 용어를 표준화하기 위해 필요한 동의어 사전 테이블이다. 대표용어, 동의

〈표 2〉 색인어 테이블 구조

Column Name	Data Type	Key Type	Null/Unique	Comment	Sample Data
ID	number(15)	Primary Key	NN	키워드 ID	9
Thesis_id	varchar(4)			논문번호	1
Major_key	varchar2(50)			키워드	편목
Frequency	number(4)			빈도수	10

〈표 3〉 전문용어 테이블 구조

Column Name	Data Type	Key Type	Null/Unique	Comment	Sample Data
ID	number(10)			단어번호	11
Major_Key	varchar2(50)	Primary Key	NN	용어이름	편목

〈표 4〉 동의어 테이블 구조

Column Name	Data Type	Key Type	Null/Unique	Comment	Sample Data
Major_key	varchar2(30)		NN	키워드	편목
Child_key	varchar2(30)	Primary Key	NN	키워드	AACR2

〈표 5〉 연관 규칙 테이블 구조

Column Name	Data Type	Key Type	Null/Unique	Comment	Sample Data
Major_Key	varchar2(30)			키워드1	1
Keyword	varchar2(30)	Primary Key	NN	키워드2	13
Frequency	number(5)			발생빈도수	3

어로 구성된다.

5) 연관 규칙 테이블(Knowledge Table)

연관 규칙을 가지는 키워드를 저장하기 위한 테이블이다. 의미적으로 서로 연관된 전문 용어 쌍과 용어간의 동시 발생 빈도수를 저장한다.

4. 실험 결과 및 고찰

4. 1 실험 방법

전문 용어 클러스터링 구현 환경은 Microsoft Windows 2000 SERVER 운영체제에서 MS SQL DBMS 2000을 기반으로 하였다.

4. 1. 1 전문 용어 추출

전처리 과정의 형태소 분석을 통해 추출된 용어에 대하여 문헌정보학사전에 수록된 문헌정보학용어를 기준으로 추출하였다. 그리고 동의어 사전을 구성하여 동일한 전문 용어에 대하여 다르게 표현된 용어를 표준화하였다. 또한 전체 문서에서 출현하는 절대 빈도수가 매우 적거나 분포도가 매우 큰 전문 용어들은 특이 용어로 처리하여 제외시켰다. 이러한 특이 용어들은 최소 지지도를 만족하지 않는 관계로 연관 규칙 탐사 대상에서 배제되거나 무의미한 연관 규칙을 양산할 수 있기 때문이다. 전처리 과정의 마지막 단계로 단어 빈도수에 의한 불균형 문제를 해결하기 위하여 TF*IDF 알고리즘을 이용하여 각 논문에서 추출된 전문 용어에 대한 가중치를 조정하였다.

문헌정보학 관련분야 논문 100편을 대상으로 실험한 결과, 전체 논문에서 추출된 전문 용어는 약 10,498개 정도이며 편당 평균 104개이다. 동의어 처리를 통해 용어를 표준화한 결과 전체 용어 수는 9,221개, 평균 92개로 줄어들었다. 그리고 전체 출현 빈도수가 2이하인 용어는 약 202개이고, 전체 문서 수에 대한 특정 용어의 출현 문서 수에 대한 표준 편차가 8이하로 분포도가 큰 전문 용어들의 수는 162개이다. 이러한 특이 용어를 제외한 최종적인 전문 용어의 수는 598개이다.

4. 1. 2 단어 빈도 가중치 조정

본 실험에서는 TF*IDF 알고리즘을 이용하여 한 문서에서 추출된 전문 용어간의 가중치를 계산하여 영향력이 현저하게 떨어지는 용어를 연관 규칙 적용 대상에서 제외시켰다. 영향력이 낮은 용어에 대한 기준은 가중치가 1이하인 용어로 정의하였다. TF*IDF 알고리즘에서 가중치 값이 1이하인 용어는 한 문서에서 3번 출현한 용어가 전체 논문의 50%에서 출현한 것을 의미하므로 일반 용어와 비슷한 의미를 가지게 된다. 따라서 이러한 용어는 문헌정보학 분야의 전문 용어이지만 모든 문서에 고르게 분포되어 무의미한 연관 규칙을 양산할 수 있다.

다음 표는 문헌정보학 정보조직 분야의 한 논문에서 추출된 전문 용어에 대한 가중치를 계산하여 상하위 10%값을 비교한 결과이다. 〈표 6〉에서 처럼 상위 가중치에 대한 용어들은 문헌정보학 정리조직 분야 논문에서 자주 사용되는 'AACR2', '편목', 'KCR3' 등과 같은 전문 용어로 구성되어 전체적인 출현 빈도에

〈표 6〉 문헌정보학 정보조직분야의 논문에서 추출된 전문 용어에 대한 가중치 비교

가중치상위 10%			가중치하위 10%		
용어	빈도수	TF*IDF	용어	빈도수	TF*IDF
AACR2	32	35.147	조직	3	0.239
KCR3	24	22.248	문헌	4	0.698
편목	21	18.351	자료	2	0.457
KDC	18	15.654	관계	1	0.239
DDC	24	21.321	분류	4	0.785

비해 문헌정보학 정리조직 분야의 출현 빈도가 더 높은 것을 알 수 있다. 그러나 하위 값을 가지는 '자료', '문헌', '분류' 등과 같은 용어는 비록 문헌정보학 분야의 전문 용어이긴 하지만 특정 분야와 상관없이 모든 분야에서 사용되는 관계로 특정 용어와 연관된 클러스터에 포함하기 어려운 용어임을 알 수 있다.

4. 1. 3 연관 규칙을 이용한 전문 용어 클러스터링

전처리 과정에서 추출된 전문 용어에 대하여 의미적으로 연관된 용어끼리 클러스터로 구성하기 위하여 연관 규칙 알고리즘을 적용하였다. 하나의 전문 용어에 대하여 발견되는 연관 규칙은 최소 지지도와 최소 신뢰도에 따라 다양한 크기로 출력된다. 즉, 지지도와 신뢰도 값이 높을수록 발견되는 연관 규칙의 수는 줄어든다. 다음 〈표 7〉은 문헌정보학 정보조직 분야의 대표적인 전문 용어인 "편목"에 대하여 지지도/신뢰도의 변화에 따라 발견된 연관 규칙의 수를 나타낸 결과이다.

〈표 7〉에서 처럼 전문 용어별로 발견된 연관 규칙은 최소 지지도와 신뢰도의 변화에 따라 다양하게 출력된다. 여기서 지지도는 전체 문서에서 연관 규칙을 이루는 전문 용어 쌍이 동시에 출현한 문서 수를 의미한다.

지지도가 너무 낮을 경우에는 연관성이 높지 않은 단어에 대해서도 연관 규칙은 만족하므로 지나치게 많은 수의 클러스터를 형성한다. 본 실험에서는 대상 문서 100편의 20% 정도인 지지도 20을 최소지지도로 설정하였다. 그리고 신뢰도는 연관 규칙 a => b에서

〈표 7〉 전문 용어 "편목"에 대한 지지도/신뢰도별 연관 규칙의 수

신뢰도 \ 지지도	5 이상	10 이상	15 이상	20 이상	25 이상	30 이상	35 이상	…
5% 이상	428	224	132	87	55	40	31	…
15% 이상	128	128	128	87	55	40	31	…
25% 이상	51	51	51	51	51	40	31	…
35% 이상	30	30	30	30	30	30	30	…
45% 이상	13	13	13	13	13	13	13	…
55% 이상	10	10	10	10	10	10	10	…
…	…	…	…	…	…	…	…	…

a 용어를 기준으로 a와 b가 동시에 출현하는 비율을 의미한다. 하지만 본 실험에서는 문헌정보학 분야의 모든 전문 용어에 대하여 관련된 용어를 클러스터로 구성하기 때문에 신뢰도는 큰 의미가 없다. 즉, 연관규칙 a=>b에서 신뢰도를 높이면 b의 출현 빈도에 따라 연관규칙의 수는 줄어들게 된다.

다음 〈표 8〉은 문헌정보학 분야에서 주로 사용되는 대표적인 용어에 대해 최소 지지도가 20일 때, 연관 규칙으로 발견된 용어의 클러스터를 구성한 예이다.

〈표 8〉의 결과에서처럼 본 논문에서 제안한 방법에 의해 전문 용어와 의미적으로 관련된 용어끼리 효과적으로 클러스터를 구성할 수 있었다. 이 결과를 지능형 전문 검색엔진이나 지식탐사시스템에 적용할 경우 단순히 키워드가 들어가는 문서를 찾아주는 기존 일반 검색엔진보다 키워드와 의미적으로 연관된 용어가 포함된 지식 문서를 검색할 수 있음으로써 좀 더 유용한 정보를 찾을 수 있을 것이다.

5. 결론 및 향후 연구 방향

최근 대량의 텍스트 문서로부터 의미 있는 패턴이나 연관 규칙을 발견하기 위한 텍스트 마이닝 기법에 대한 연구가 활발히 전개되고 있다. 하지만 비정형의 구조를 가진 대량의 텍스트 문서로부터 추출된 용어의 수는 불규칙적이고 의미 없는 일반적인 용어가 많이 추출되는 관계로 일반적인 연관 규칙 탐사 방법을 사용하게 되면 무의미한 연관 규칙이 대량으로 생성되어 사용자에게 필요한 유용한 지식 정보를 효과적으로 검색하기 어렵다.

본 논문에서는 대량의 문서로부터 적합한 지식 정보 검색을 제공하는 지능형 검색엔진을 개발하기 위한 전문 용어 클러스터링 방법을 제안하였다. 클러스터링 과정은 논문에서 전문 용어만을 추출하기 위한 전처리 과정과 추출된 전문 용어에 연관 규칙 탐사 알고리즘을 적용하여 하나의 전문 용어에 대한 발견되는 연관 규칙 집합을 클러스터로 구성하는 과정으로 이루어졌다.

제안한 클러스터링 기법의 효율성을 검증하기 위하여 문헌정보학 관련 학회에서 발표된 100편의 학술 논문에서 추출한 문헌정보학 용어를 대상으로 실험을 하였다. 전처리 과정에서 일반 용어는 제외하고 문헌정보학과 관련된 전문 용어만을 추출하여 연관 규칙 탐사 알고리즘을 적용하였다. 따라서 무의미한 연관 규칙의 양산을 방지하고 관련 용어들간에 발견된 연관 규칙에 의해 전문 용어와 의미적으로 관련된 전문용어를 클러스터링할 수 있었다.

〈표 8〉 최소지지도 20, 최소신뢰도 55%일 때 연관 규칙 생성 결과

전문 용어	관련 용어 클러스터
DDC	KDC, 클러스터링, 듀이
편목	AACR2, KOMARC, OPAC, MARC
시소러스	z39.50, 정확율, 재현율

본 연구 결과를 특정 분야에 대한 지능형 전문 검색엔진에 적용할 경우 현재 단일 키워드로만 검색하는 기존의 검색엔진과는 달리 관련된 전문 용어에 대한 클러스터링 정보와 지식 정보를 효과적으로 검색할 수 있는 지능적인 전문 검색엔진을 개발할 수 있을 것이다. 또한 일반 문서에 적용하여 관련 있는 문서끼리 지식 정보의 연관성에 따라 자동으로 분류하여 효율적인 지식 탐사 시스템 개발에 적용할 수 있을 것이다.

향후 연구에서는 문헌구조에 따른 정보량에 차이가 있다는 사실에 근거하여 단어가 출현한 위치에 따른 빈도수와 각 문헌요소에 따른 가중치를 복합적으로 고려하여 가중빈도를 산출하는 것을 제안하고자 한다. 문헌요소에 대한 가중빈도 산출방식은 다음과 같다.[20]

① 표제(국문, 영문) : 단순빈도 * 30
② 요약(국문, 영문) : 단순빈도 * 10
③ 서론 : 단순빈도 * 5
④ 본론 : 단순빈도 * 1
⑤ 결론 : 단순빈도 * 5

이와 같은 가중치의 부여형태는 학술논문의 문헌구조와 같은 문헌형태에 한정된 것으로 앞으로 다양한 형태의 문헌집단에 적용하기 위해서는 문헌내부의 각 문헌요소에 출현한 단어간의 단순출현빈도에 의한 상대적 가중치를 부여하는 방향으로 연구가 진행되어야 할 것이다.

참 고 문 헌

강승식, HAM : 한국어 형태소 분석 라이브러리. 〈http://nlp.kookmin.ac.kr/HAM/Kor/download.html〉

김호성, 고희정. 1992. 용어 빈도수를 이용한 영문 문헌정보의 점진적인 개념적 집단화. 『한국정보과학회논문지』, 19(1): 12-23.

박건호. 1999. 『Apriori 알고리즘 연관규칙 마이닝기법을 이용한 정보검색』. 석사학위논문, 고려대학교 대학원, 컴퓨터학과.

박정오, 황도삼. 2000. 전문 용어 추출 시스템. 『한국정보과학회』, 27(1): 316-318.

박종수, 유원경, 홍기형. 1998. 연관 규칙 탐사와 그 응용. 『한국정보과학회 SIGDB 춘계튜토리얼』.

서성보, 김선철, 이준욱, 류근호. 2000. 주요 항목 집합을 이용한 문서 클러스터링 및 연관 규칙 탐사 기법. 『한국정보과학회』, 27(1): 169-171.

승현우, 박미영, 조용한, 강미나. 2002. 상용 DataMining Tool비교 분석. 『제3회 한국정보처리학회 정보기술워크숍』

승현우, 박미영, 황정민. 2002. 데이터마이닝 기법과 연구동향분석. 『제4회 한국정보처리학회 정보기술워크숍』

승현우, 박미영, 황정민. 2002. 데이터마이닝 기법을 이용한 문헌정보학 전문용어

클러스터링 데이터베이스 설계에 관한 연구. 『서울여자대학교 자연과학연구 소논문집』

신진섭. 2000. 『웹 문서 분류를 위한 단어의 연관성 모델과 클러스터링 모델』. 박사학위논문, 건국대학교 대학원, 컴퓨터정보통신공학과.

이란주. 데이터마이닝에 관한 연구. 〈http://203.241.185.12/asd/main.cgi?board=Data&ryal=&view=2&back=&search=%C0%CC%B6%F5%C1%D6&where=1&how=1〉

이문기, 권오욱, 이종혁. 2000. 웹 디렉토리 서비스를 위한 문서 클러스터링. 『한국정보과학회』. 27(1): 351-353.

이정원 외 12인. 2000. 『데이터마이닝 알고리즘 분석』. 이화여자대학교 과학기술대학원 컴퓨터학과 EIST Research Report Series.

이정화. 2000. 『데이터마이닝 기법을 이용한 전문용어 클러스터링』. 석사학위논문, 창원대학교 대학원, 컴퓨터학과.

Eui-Hong Han, Vipin Kumar, George Karypis. 1997. 『Hypergraph Based Clustering In A HighDimensional Data Sets : A Summary of Results』. IEEE.

Han, Jiawei. 2002. *Data Mining: Concepts and Techniques*. Canada: Morgan Kaufmann publishers.

R. Agrawal, Chu Xu. 1994. 『Fast Algorithms for Mining Association Rules』, In Proc. of VLDB.

Thorsten Joachims. 1996. 『A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization』. CMU-CS-96-18.

William B. Frakes, Ricardo Baeza-Yates. 1992. 『Information Retrieval : Data Structures & Algorithms』. Prentice Hall.