데이타마이닝 기법을 이용한 효율적인 전문 용어 클러스터링

이정화* · 남상엽* · 문현정* · 우용태*

An Efficient Terminology Clustering Method Using Datamining Technique

Jung-Hwa Lee*, Sang-Yub Nam*, Hyeon-Jeong Mun*, Yong-Tae Woo*

요 약

최근 대량의 텍스트 문서로부터 의미 있는 패턴이나 연관 규칙을 발견하기 위한 텍스트마이닝 기법에 대한 연구가 활발히 전개되고 있다. 하지만 비정형 텍스트 문서로부터 추출된 용어의 수는 불규칙적이고 일반적인 용어가 많이 추출되는 관계로 일반적인 연관 규칙 탐사 방법을 사용하게 되면 무의미한 연관 규칙이 대량으로 생성되어 지식 정보를 효과적으로 검색하기 어렵다. 본 논문에서는 연관 규칙 탐사 기법을 이용하여 대량의 문서로부터 유용한 지식 정보를 찾기 위하여 의미적으로 연관된 전문 용어들끼리 클러스터링 하기 위한 방법을 제안하였다. 학술 논문을 대상으로 전문 용어를 추출하여 관련된 용어들끼리 클러스터를 구성하는 실험을 통하여 제안된 방법의 효율성을 보였다.

Key words : 데이타마이닝, 텍스트마이닝, 전문 용어 클러스터링, 연관 규칙 탐사

I. 서 론

지식기반 사회가 도래함에 따라 일반적인 문서나 인터넷에서 제공되는 웹 문서가 급속도로 증가하고 있지만 사용자에게 필요한 유용한 지식 정보를 찾는 과정은 점점 어려워지고 있다. 이에 따라 대량의 문서로부터 의미 있는 지식 정보를 추출하기 위한 지식 탐사 기법이 필요하게 되었다. 최근에는 대량의 데이타로부터 관련 있는 패턴이나 연관 규칙을 발견하기 위한 데이타마이닝 기법을 응용한 텍스트 마이닝 기법에 대한 연구가 활발하게 전개되고 있다.

연관 규칙 알고리즘을 이용한 텍스트마이닝 기법은 텍스트 문서를 대상으로 서로 연관된 용어 집합을 추출하고, 유용한 지식 정보를 효과적으로 검색하기 위한 기법이다. 하지만 대량의 비정형 텍스트 문서로부터 추출된 단어를 정형화된 수치로 표현하기 어렵기 때문에 일반적인 연관 규칙 알고리즘을 그대로 적용하기 어렵다.
또한 텍스트 문서에서 추출된 용어의 수가 불규칙적이고 일반적인 용어가 많이 추출되는 관계로 무의미한 연관 규칙이 대량으로 생성되어 지식 정보를 효과적으로 추출하기 어렵다. 특히 인터넷에서 정보를 찾기 위해 사용하는 검색엔진들은 주로 단일 키워드를 중심으로 웹 문서에 대한 인덱스를 생성하는 관계로 의미적으로 연관된 전문 내용을 포함하는 문서를 찾기 어렵다.

본 논문에서는 대량의 웹 문서로부터 유용한 지식 정보를 찾기 위한 전문 검색엔진을 개발하기 위하여 텍스트에서 추출된 전문 용어를 효율적으로 클러스터링하기 위한 방법을 제안하였다. 즉, 일반적인 용어들간의 무의미한 연관 규칙이 양산되는 것을 방지하기 위하여 전문 용어로 구성된 지식베이스 테이블을 이용하여 의미 있는 용어들간의 연관 규칙을 생성하였다. 실험 대상을 하나의 용어에 의해 발견되는 연관 규칙 집합을 이용하여 의미적으로 관련된 전문 용어들끼리 효율적으로 클러스터링 할 수 있게 하였다.

제안된 기법의 효율성을 검증하기 위하여 학술 논문에서 사용된 컴퓨터 관련 전문 용어를 기초로 초기 지식베이스 테이블을 생성하였다. 지식베이스 테이블은 형태소 분석기를 통해 추출된 용어 중에서 일반적인 단어는 배제시키고 컴퓨터 분야와 관련된 전문 용어를 중심으로 구성하였다.

연관 규칙은 하나의 논문에서 사용된 전문 용어들의 집합을 트랜잭션 단위로 구성하여 연관 규칙 탐사 알고리즘을 적용하여 생성하였다. 하나의 용어에 의해 생성된 연관 규칙 집합은 해당 전문 용어와 관련된 클러스터로 정의하였다. 또한 클러스터에 속하는 용어들간의 연관 정도를 수치로 표현하기 위하여 연관된 용어간의 출현 빈도 수를 가중치로 부여하였다. 학술논문을 대상으로 전문 용어를 추출하여 관련된 용어들끼리 클러스터를 구성하는 실험을 통하여 제안된 방법의 효율성을 보였다.

* 창원대학교 전자계산학과

II. 텍스트마이닝 기법

2.1 텍스트마이닝 기법의 정의

최근에 데이타마이닝 기법을 응용하여 대량의 문서로부터 숨겨진 패턴이나 연관 규칙을 발견하기 위한 텍스트마이닝 기법에 대한 연구가 활발하게 진행되고 있다. 텍스트마이닝 기법은 대량의 텍스트 문서로부터 유용한 지식 정보를 효과적으로 찾기 위한 기법이다. 하지만 비정형 텍스트 문서로부터 추출된 단어를 정형화하기 어렵기 때문에 일반적인 연관 규칙 알고리즘을 그대로 적용하기 어렵다. 또한 텍스트 문서에서 추출된 용어의 수가 불규칙적이고 일반적인 용어가 많이 추출되는 관계로 무의미한 연관 규칙이 대량으로 생성되어 지식 정보를 효과적으로 추출하기 어렵다.

텍스트마이닝의 일반적인 접근 방법은 텍스트 요약(Text Summarization), 텍스트 범주화(Text Categorization), 그리고 텍스트 클러스터링(Text Clustering) 방법으로 구분된다. 텍스트 요약이란 주어진 텍스트의 내용에서 대표가 되는 일부를 사용자에게 제시함으로써 텍스트의 내용이 쉽게 파악 되도록 하는 방법이다. 텍스트 범주화는 텍스트에 대해 사전에 범주를 설정하고 임의의 텍스트에 대한 범주를 자동으로 설정하기 위한 방법이다. 텍스트 클러스터링은 주어진 텍스트 집단을 내용의 유사성에 따라 소집단으로 분할하는 방법이다[1].

2.2 관련 연구

2.2.1 일반적인 클러스터링 기법

패턴 인식이나 문자 인식에서 사용되는 일반적인 클러스터링 방법은 분할 접근과 계층적 접근 방법으로 나눌 수 있다. 분할 접근 방법은 범주 함수를 최적화 시키는 K개의 분할영역을 결정해 나가는 방법으로 유클리디안 거리 측정법에 기반한다. 대표적인 알고리즘으로는 K-means 방법과 K-medoid 방법이 있다. 계층적 접근은 초기에 각각의 데이타 점을 하나의 클러스터로 설정한 후 이들 쌍간의 거리를 기반으로 하여 분할과 합병을 반복하는 방식이다. 쌍간의 거리 측정 방법에 따라 단일 연결, 완전 연결, 중심 연결 등이 있다. 대표적인 알고리즘으로는 CLARANS, BIRCH, DBSCAN, DBCLASD, CURE 등이 있다[2].

2.2.2 텍스트마이닝을 이용한 클러스터링 기법

텍스트마이닝을 이용한 클러스터링 기법은 대상 집합에 따라 크게 용어 클러스터링과 문서 클러스터링으로 나뉘어진다. 또한 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 동시에 출현한 용어 개수의 비율을 유사도로 정의하여 클러스터링 한다.

김호성 등[3]은 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 출현 빈도에 의해 클러스터를 구성하는 관계로 출현 빈도가 적은 용어는 클러스터에 포함되지 않은 문제점이 있다. 신진섭[4]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 등[5]은 연관 규칙을 이용하여 용어에 대한 하이퍼그래프를 생성하여 신뢰도에 근거한 유사성 척도를 사용하여 분할하는 용어 클러스터링 알고리즘을 제시하였다. 서성보 등[6]은 트랜잭션에 대한 클러스터의 유사성을 측정하기 위해 주요 항목과 비 주요 항목으로 구분하고, 각 트랜잭션의 최소 비용 계산을 통해 자동화된 문서 클러스터링 기법을 제안하였다. 하지만 주요 항목 집합의 기준을 빈도수만 고려하여 무의미한 연관 규칙이 대량으로 발견될 수 있다.

2.2.3 텍스트마이닝을 위한 대표 색인어 추출 기법

비정형화된 대량의 문서를 대상으로 텍스트마이닝 기법을 적용하기 위한 핵심적인 기술중의 하나는 문서를 대표할 수 있는 색인어를 효과적으로 추출하기 위한 방법이다. 대부분의 방법에서는 추출된 단어의 순서보다는 문서 내에서 단어의 출현 여부나 빈도 수를 고려한 통계적인 정보를 기반으로 추출하고 있다.
문서에서 대표 색인어를 추출하기 위한 연구는 문서에서 출현하는 각 용어에 대한 중요도를 확률적으로 계산하여 가중치를 조정하는 TF*IDF 알고리즘[7]이나 키워드의 밀집성[4]을 이용한 연구가 진행되었다.

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 등[3]은 도서관 분류 체계에서 새로운 학문 분야를 반영하기 위하여 새로운 전문 용어의 개념을 자동으로 습득하여 용어 클러스터링 하는 방법을 제안하였다.
박정오 등[8]은 컴퓨터를 이용하여 전문 용어를 자동적으로 추출하기 위한 전문 용어 추출 시스템을 개발하였다. 이 시스템에서는 기존 전문 용어에서 사용되는 특정어구를 이용하여 전문 용어를 추출하고, 후보 전문 용어에서 단어의 위치 정보를 이용하여 전문 용어를 추출하는 방법을 제안하였다.

III. 효율적인 전문 용어 클러스터링 기법

3.1 전문 용어 클러스터링 모델

본 논문에서는 대량의 웹 문서로부터 유용한 지식 정보를 찾기 위한 지능형 검색엔진을 개발하기 위하여 의미적으로 연관된 전문 용어들끼리 효율적으로 클러스터링하기 위한 방법을 제안하였다. 제안한 클러스터링 모델은 크게 문서에서 전문 용어를 추출하기 위한 전처리 과정과 전문 용어간의 연관 규칙을 탐사하여 클러스터로 구성하는 과정으로 이루어진다. 다음 그림 1은 본 논문에서 제안한 전문 용어 클러스터링 모델의 전체적인 구성도이다.

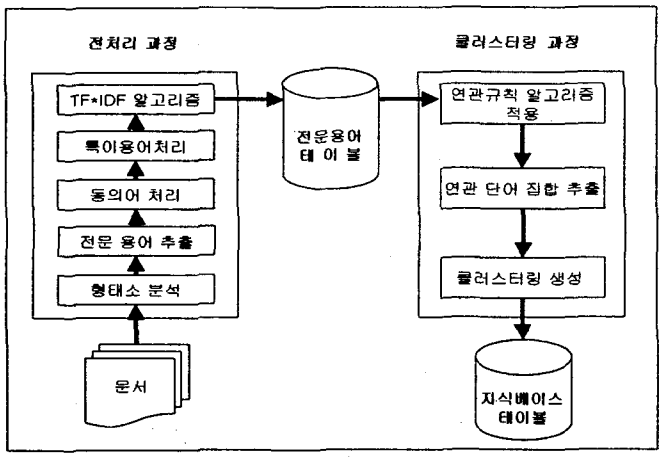

그림 1. 전문 용어 클러스터링 모델의 전체 구성도

3.2 전문 용어 추출을 위한 전처리 과정

전문 지식을 담고 있는 학술 논문에서도 전문 용어보다는 일반적인 용어들이 많이 출현한다. 따라서 전문 용어가 아닌 일반 용어를 배제하고 전문 용어만을 추출하기 위한 전처리 과정이 필요하다. 본 논문에서 전문 용어를 추출하기 위한 실험 대상을 컴퓨터 분야의 논문으로 선택하였다. 따라서 본 과정은 컴퓨터 분야의 논문에서 컴퓨터에 관한 전문 용어만을 추출하기 위한 과정으로 전문 용어 추출 과정과 가중치 적용 과정으로 이루어진다. 그림 2는 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 한성대학교 강승식 교수팀이 개발한 공개용 형태소 분석기인 HAM4.0a[9]를 사용하였다. 그리고 신성대학 김현숙 교수의 컴퓨터 용어 사전에 수록된 컴퓨터 분야의 전문 용어를 기준으로 전문 용어를 추출하였다. 그리고 학술 논문에서 사용하는 전문 용어는 저자에 따라 영어나 한국어를 혼용하거나 영문 용어를 한글화하는 과정에서 차이가 있는 관계로 동의어 사전을 구성하여 전문 용어를 통일하였다. 그리고 전체 문서에서 출현하는 절대 빈도수가 매우 적어서 연관 규칙 탐사 대상이 되지 않는 용어와 용어들의 분포도가 매우 큰 관계로 무의미한 연관 규칙을 발생할 수 있는 용어를 특이 용어로 취급하여 연관 규칙 탐사과정에서 배제시켰다.

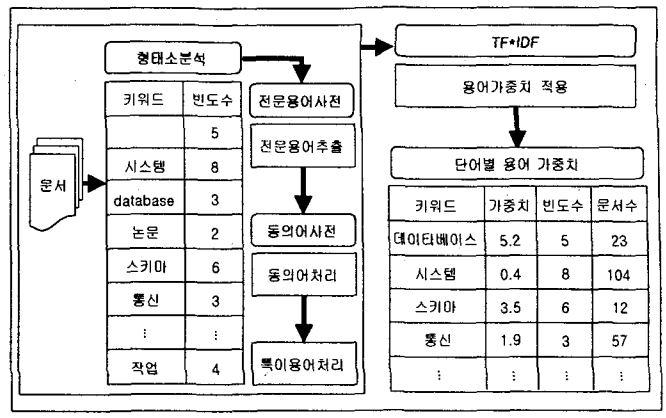

그림 2. 전처리 과정의 전체 구성도

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 '시스템'이라는 용어는 컴퓨터 분야에서 많이 사용되는 용어이고 빈도수는 높지만 대부분의 컴퓨터 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 역 문서 빈도수(Inverse Document Frequency)를 단어의 빈도수와 같이 적용함으로써 그 문서를 대표하는 단어들을 효율적으로 찾을 수 있는 알고리즘이다[7]. 문서의 빈도 df_i는 N개의 문서들 중에서 단어 t_j가 존재한 문서의 개수를 의미하며, 단어의 빈도 tf_{ij}는 문서 d_i에서 단어 t_j가 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_j의 중요도 또는 영향력(Weight) w_{ij}라고 정의한다.

$$w_{ij} = tf_{ij} \log(N/df_i) \quad (1)$$

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 논문에서 실험 대상 문서는 학술 발표 논문으로 문서 길이가 일정하므로 문서 길이 정규화는 고려하지 않았다.

3.3 연관 규칙을 이용한 전문 용어 클러스터링

본 논문에서는 데이타마이닝 기법을 이용하여 서로 연관된 전문 용어들끼리 클러스터를 구성하였다. 클러스터를 구성하기 위하여 데이타마이닝 기법 중에서 장바구니 분석 과정에서 주로 사용하는 연관 규칙 탐사 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 그림 3은 연관 규칙 탐사 알고리즘을 이용하여 클러스터를 구성하는 과정에 대한 개념도이다.

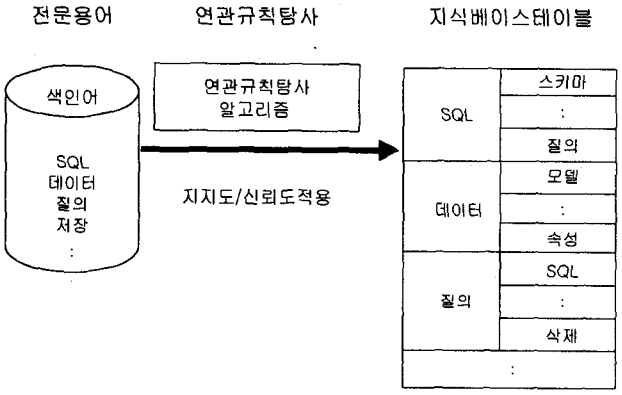

그림 3. 연관 규칙을 이용한 전문 용어 클러스터링

3.3.1 연관 규칙의 정의

연관 규칙이란 '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 최소 지지도와 신뢰도를 만족하는 의미 있는 연관 규칙을 발견하는 과정을 의미한다.

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. X를 한 트랜잭션에 포함된 항목들의 빈도를 고려하지 않은 항목들의 집합일 때 $X \subseteq T$ 이면 트랜잭션 T는 X를 포함한다. 이때 연관 규칙(Association Rule)은 R: $X \rightarrow Y$ 형식의 함축이고, 이때 X와 Y는 서로 같은 원소를 갖지 않는 항목집합이다. 만일 한 트랜잭션이 X를 지지한다면, 또한 어떤 확률에 의해 Y도 지지할 것이라는 예측하는 것이 연관 규칙이다. 이 확률을 이 규칙에 대한 신뢰도(conf(R))라 한다. R의 신뢰도는 다음 식(2)처럼 X를 지지하는 T에 대하여 Y 또한 지지할 조건부 확률로 정의된다.

$$conf(R) = p(Y \subseteq T \mid X \subseteq T) = \frac{p(Y \subseteq T \bigwedge X \subseteq T)}{p(X \subseteq T)} = \frac{supp(X \bigcup Y)}{supp(X)} \quad (2)$$

D에 있는 규칙 R에 대한 지지도는 $supp(X \cup Y)$로 정의한다. 지지도는 얼마나 자주 적용할 수 있는 지를 나타내는 반면 신뢰도는 그 규칙이 얼마나 믿을만한지를 의미한다. 규칙이 데이타베이스에서 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 조건으로 규칙의 조건부와 결과부는 모두 빈발해야 한다[10].

3.3.2 연관 규칙 탐사 알고리즘

연관 규칙 탐사 알고리즘에서 연관 규칙을 탐사하는 과정은 다음과 같은 2 단계로 이루어진다.

1 단계 : 먼저, 빈발 항목 집합(Large Itemsets) 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 항목 집합이라 한다.

2 단계 : 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 연구에서는 의미적으로 서로 연관된 전문 용어들끼리 클러스터로 구성하기 위하여 연관 규칙 탐사 알고리즘을 사용하였다. 하나의 논문에서 추출한 전문 용어들의 집합을 트랜잭션(Transaction), 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 클러스터의 크기와 클러스터에 포함되는 용어를 결정하기 위한 명확한 기준을 정하기는 어렵다. 본 연구에서는 다양한 구간에 대한 지지도와 신뢰도에 대하여 연관 규칙을 찾는 실험을 반복하였다. 그리고 일정한 지지도와 신뢰도를 만족하는 연관 규칙에 대하여 전문 용어간의 연관성을 판정하여 최적의 클러스터로 정의하였다.

IV. 실험 결과 및 고찰

4.1 실험 방법

본 논문에서 제안한 클러스터링 기법의 효율성을 검증하기 위하여 컴퓨터 관련 학회에서 발표된 240편의 학술 논문에서 추출한 컴퓨터 용어를 대상으로 실험을 하였다.

4.1.1 전문 용어 추출

전처리 과정을 통하여 형태소 분석을 통해 추출된 용어에 대하여 신성대학 김현숙 교수의 컴퓨터 용어 사전에 수록된 컴퓨터 전문 용어로 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.

컴퓨터 관련분야 논문 240편을 대상으로 실험한 결과, 전체 논문에서 추출된 전문 용어는 약 24,986개 정도이며 편당 평균 104개이다. 동의어 처리를 통해 용어를 표준화한 결과 전체 용어 수는 22,838개, 평균 95개로 줄어들었다. 그리고 전체 출현 빈도수가 2이하인 용어는 약 242개이고, 전체 문서 수에 대한 특정 용어의 출현 문서 수에 대한 표준편차가 8이하로 분포도가 큰 전문 용어들의 수는 148개이다. 이러한 특이 용어를 제외한 최종적인 전문 용어의 수는 1,499개이다.

4.1.2 단어 빈도 가중치 조정

한 문서 내에서 추출된 전문 용어들간에도 그 문서에서 차지하는 영향력이 차이가 난다. 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 가지게된다. 따라서 이러한 용어는 컴퓨터 분야의 전문 용어이지만 모든 문서에 고르게 분포되어 무의미한 연관 규칙을 양산할 수 있다.

다음 표 1은 정보 통신 분야의 한 논문에서 추출된 전문 용어에 대한 가중치를 계산하여 상하위 10%값을 비교한 결과이다. 표 1에서처럼 상위 가중치에 대한 용어들은 정보 통신 논문에서 자주 사용되는 'tcp', '패킷' 등과 같은 전문 용어로 구성되어 전체적인 출현 빈도에 비해 정보 통신 분야의 출현 빈도가 더 높은 것을 알 수 있다. 그러나 하위 값을 가지는 '접근', '컨트롤' 등과 같은 용어는 비록 컴퓨터 분야의 전문 용어이긴 하지만 특정 분야와 상관없이 모든 분야에서 사용되는 관계로 특정 용어와 연관된 클러스터에 포함하기 어려운 용어임을 알 수 있다.

표 1. 데이타베이스 분야의 논문에서 추출된 전문 용어에 대한 가중치 비교

가중치 상위 10%			가중치 하위 10%		
용어	빈도수	TF*IDF	용어	빈도수	TF*IDF
ack	24	40.349	정도	1	0.609
tcp	24	20.680	컨트롤	1	0.461
송신자	10	15.351	접근	1	0.350
패킷	16	12.450	관계	1	0.323

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 2는 데이타베이스 분야의 대표적인 전문 용어인 "데이타베이스"에 대하여 지지도/신뢰도의 변화에 따라 발견된 연관 규칙 결과이다.

표 2. 전문 용어 "데이타베이스"에 대한 지지도/신뢰도별 연관 규칙의 변화

…	…	…	…	…	…	…	…	…
55%이상	8	8	8	8	8	8	8	…
45%이상	14	14	14	14	14	14	14	…
35%이상	31	31	31	31	31	31	31	…
25%이상	51	51	51	51	51	41	33	…
15%이상	128	128	128	89	57	41	33	…
5%이상	437	243	141	89	57	41	33	…
신뢰도 / 지지도	5 이상	10 이상	15 이상	20 이상	25 이상	30 이상	35 이상	…

표 2에서처럼 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. 본 실험에서는 대상 문서 240편의 10% 정도인 20을 최소지지도로 설정하였다. 그리고 신뢰도는 연관 규칙 a => b에서 a 용어를 기준으로 b가 동시에 출현하는 비율을 의미한다. 본 실험에서는 컴퓨터 분야의 모든 전문 용어에 대하여 관련된 용어를 클러스터로 구성하기 때문에 신뢰도는 큰 의미가 없다. 즉, 연관 규칙 a => b에서 신뢰도를 높이면 b의 출현 빈도에 따라 연관 규칙의 수는 줄어들게 된다.

다음 표는 정보 통신 분야에서 주로 사용되는 대표적인 5개의 용어에 대해 최소 지지도가 20일 때, 연관 규칙으로 발견된 용어의 상위 빈도 10개를 클러스터로 구성한 예이다.

표 3. 지지도 20이상, 신뢰도 55%이상일 때 연관 규칙 생성 결과

전문 용어	관련 용어 클러스터
프로토콜	전송, 연결, 네트워크, 통신, 전달, 디자인, 컨트롤, 접근, 메세지, 인터페이스
패킷	전송, 프로토콜, 네트워크, 연결, ip, 통신, 전달, 호스트, 디자인, 컨트롤
교환	디자인, 전송, 전달, 연결, 통신, 접근, 네트워크, 메시지, 객체, 컨텐츠
수신	전송, 네트워크, 연결, 프로토콜, 디자인, 통신, 메시지, 접근, 전달, 컨트롤
네트워크	연결, 전송, 통신, 프로토콜, 디자인, 접근, 전달, 컨트롤, 선택, 특징

표 3의 결과에서처럼 본 논문에서 제안한 방법에 의해 전문 용어와 의미적으로 관련된 용어끼리 효과적으로 클러스터를 구성할 수 있었다. 이 결과를 검색엔진이나 지식경영시스템에 적용할 경우 단순히 키워드가 들어가는 문서를 찾아주는 기존 검색엔진보다 키워드와 의미적으로 연관된 용어가 포함된 지식 문서를 검색할 수 있을 것이다.

V. 결 론

본 논문에서는 대량의 웹 문서로부터 유용한 지식 정보를 찾기 위한 지능형 검색엔진을 개발하기 위하여 텍스트에서 추출된 용어 중에서 의미적으로 연관된 전문 용어들끼리 효율적으로 클러스터링하기 위한 방법을 제안하였다. 클러스터링 과정은 논문에서 전문 용어만을 추출하기 위한 전처리 과정과 전문 용어간의 연관 규칙을 탐사하여 클러스터로 구성하는 과정으로 이루어진다. 제안한 클러스터링 기법의 효율성을 검증하기 위하여 컴퓨터 관련 학회에서 발표된 240편의 학술 논문에서 추출한 컴퓨터 용어를 대상으로 실험을 하였다. 관련 용어들간에 발견된 연관 규칙에 의해 전문 용어와 의미적으로 관련된 용어끼리 효과적으로 클러스터를 구성할 수 있었다. 본 연구 결과를 특정 분야에 대한 전문 검색엔진이나 관련된 전문 용어에 대한 클러스터링 정보를 이용하여 지식 정보를 효과적으로 검색할 수 있는 지능적인 전문 검색엔진을 개발할 수 있을 것이다. 또한 일반 문서에 적용하여 관련 문서끼리 지식 정보의 연관성에 따라 분류하여 효율적인 지식 탐사 시스템 개발도 가능하다. 본 논문에서는 컴퓨터 분야의 전문 용어 클러스터링에 대한 한정된 연구를 하였으나, 향후로는 다른 전문 분야에 대한 연구도 점진적으로 진행할 계획이다.

참 고 문 헌

[1] 조태호, "텍스트 마이닝에 대한 소개와 기능," 한국정보처리학회, '98추계학술발표논문집, pp.27-29, 1998

[2] 이정원 외 12인, "데이터 마이닝 알고리즘 분석," 이화여자대학교 과학기술대학원 컴퓨터학과, EIST Research Report Series, 2000

[3] 김호성, 고희정, "용어 빈도수를 이용한 영문 문헌정보의 점진적인 개념적 집단화," 한국정보과학회 논문지, 제19권 1호, pp.12-23, 1992

[4] 신진섭, "웹 문서 분류를 위한 단어의 연관성 모델과 클러스터링 모델," 건국대학교 대학원 컴퓨터정보통신공학과 박사학위논문, 2000

[5] Eui-Hong Han, Vipin Kumar, George Karypis, "Hypergraph Based Clustering In A High-Dimensional Data Sets : A Summary of Results," IEEE, 1997

[6] 서성보, 김선철, 이준욱, 류근호, "주요 항목 집합을 이용한 문서 클러스터링 및 연관 규칙 탐사 기법," 2000 봄 학술발표논문집, 한국정보과학회, 제27권 1호, pp.169-171, 2000

[7] Thorsten Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," CMU-CS-96-18, March 1996

[8] 박정오, 황도삼, "전문 용어 추출 시스템," 2000 봄 학술발표논문집(B), 한국정보과학회, 제27권 1호, pp.316-318, 2000

[9] 강승식, "HAM: 한국어 형태소 분석 라이브러리," http://ham.hansung.ac.kr

[10] 박종수, 유원경, 홍기형, "연관 규칙 탐사와 그 응용", 한국정보과학회 SIGDB 춘계튜토리얼, 1998