효율적인 문서 자동 분류를 위한 대표 색인어 추출 기법

김 지 숙* · 김 영 지* · 문 현 정* · 우 용 태*

A Feature Selection Technique for an Efficient Document Automatic Classification

Ji-Suk Kim*, Young-Ji Kim*, Hyeon-Jeong Mun*, Yong-Tae Woo*

Abstract

Recently there are many researches of text mining to find interesting patterns or association rules from mass textual documents. However, the words extracted from informal documents are tend to be irregular and there are too many general words, so if we use pre-exist method, we would have difficulty in retrieving knowledge information effectively. In this paper, we propose a new feature extraction method to classify mass documents using association rule based on unsupervised learning technique. In experiment, we show the efficiency of suggested method by extracting features and classifying of documents.

※ 본 논문은 한국과학재단의 2001년 목적기초연구(2001-1-30300-015-1) 지원으로 수행되었음.
* 창원대학교 컴퓨터공학과

1. 서 론

인터넷의 대중화에 따라 웹 문서가 기하급수적으로 늘어나고 있지만 검색엔진에서 인덱스하는 웹 문서 비율은 오히려 감소하여 사용자들이 인터넷에서 원하는 정보를 효율적으로 찾기가 어려워지고 있다. 이에 따라 검색엔진들은 검색 효율을 개선하기 위해 지능형 검색 기법을 개발하고 있다. 또한 기업이나 정부기관들이 보유한 전문 지식 문서가 급격하게 증가하고 있으며, 대량의 지식 문서를 체계적으로 관리하여 기업 경영에 활용하기 위한 지식관리시스템 구축이 필수적으로 요구되고 있다.

최근에 지능형 검색엔진이나 지식관리시스템 개발을 위해 데이타베이스에 저장된 대량의 지식 문서로부터 연관된 패턴이나 규칙을 발견하여 사용자가 원하는 지식 정보를 정확하게 검색하기 위한 지식탐사시스템(KDD, Knowledge Discovery in Databases)에 대한 연구가 활발하게 진행되고 있다. 이러한 지식탐사시스템에서 가장 핵심적인 요소 기술의 하나는 대량의 지식 문서를 자동적으로 분류하기 위한 문서분류시스템이다.

문서분류시스템에서 문서의 내용에 따라 적절한 카테고리를 지정하여 문서를 자동적으로 분류하기 위해서는 각 카테고리를 대표하는 색인어가 먼저 구성되어야 한다. 이러한 카테고리별 대표 색인어를 추출하기 위한 연구 방법은 학습용 문서의 사전 분류 여부에 따라 감독학습 기법과 비감독학습 기법으로 나뉘어진다. 감독학습 기법은 비교적 정확한 대표 색인어가 추출되지만, 사전에 각 카테고리별로 학습용 문서를 분류하기 위한 비용과 카테고리를 재구성할 경우에 대표 색인어를 다시 추출해야하는 어려움이 있다. 이에 반해 비감독학습 기법은 사전에 분류된 학습 문서를 사용하지 않고 분류대상 문서에서 직접 대표 색인어를 추출하여 분류하기 때문에 비용이 적게 들고, 동적으로 카테고리를 재구성할 수 있다.

본 논문에서는 비감독학습 기법에 의해 대표 색인어를 추출하는 방법을 제안하였다. 먼저, 각 카테고리를 대표하는 Seed Keyword 집합을 선정한다. 여기서 Seed Keyword 집합은 각 카테고리를 대표하는 특징단어 집합이다. 그리고 연관 규칙 탐사 알고리즘을 적용하여 각 Seed Keyword와 관련된 용어집합을 추출한다. 추출된 용어집합에서 상위 N개의 용어들로 1차 연관 용어 집합 $K_{si} = \{T_1, T_2, T_3, \cdots, T_n\}$을 구성한다. 두 번째 단계에서는 연관 용어 집합 K_{si}의 각 원소 $T_i(1 < i < n)$에 대해 연관 용어 집합을 추출하여 1차 연관 용어 집합과 그룹화한다. 각 그룹에 대해 최소 지지도 n% 이상의 상위 N개의 연관 용어들을 재그룹하여 2차 연관 용어 집합을 추출하여 대표 색인어로 구성한다.

제안된 기법의 효율성을 검증하기 위하여 컴퓨터 관련 논문을 대상으로 분류 실험을 하였다. 대표 색인어 추출을 위한 문서는 컴퓨터 관련 논문 240편을 사용하였고, 분류 실험에서는 대표 색인어 추출과정에 사용하지 않은 분야별 30여편의 논문을 사용하였다. 기존의 감독학습 기법 중에서 대표 색인어 추출 성능이 우수하다고 알려진 χ^2(Chi Square) 기법, DF(Document Frequency) 기법[진훈 외 1인 2001][홍진혁 외 2인 2001][Yang & Pdedersen 1997]과의 비교 실험을 통하여 제안된 기법에 대한 성능을 평가하였다.

2. 문서 자동 분류 기법

2.1 문서 자동 분류의 정의

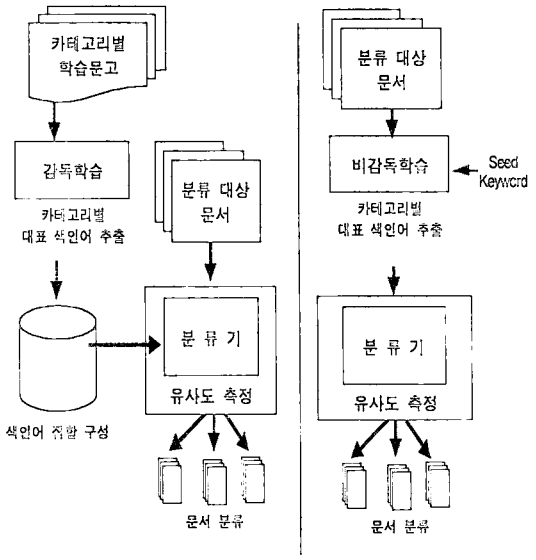
문서 자동 분류란 컴퓨터를 이용하여 문서를 대표하는 특징으로 구성된 색인어 집합과 유사한 문서들을 같은 그룹으로 분류하는 기법이다. 컴퓨터를 이용하여 문서를 자동으로 분류하기 위한 시도는 70년대 말 Salton에 의해 체계화되기 시작하였다[정영미 1993]. 문서 자동 분류 기법은 지식관리시스템의 가장 핵심적인 요소로서 사용자가 원하는 지식정보를 효과적으로 검색하기 위해 대량의 문서를 자동 분류하는 기술이다. 문서 분류 과정은 전체 문서 집합에서 분류에 영향을 주지 않는 의미없는 단어를 제거하기 위한 전처리 과정, 문서 내용에서 중심이 되는 특징을 구성하기 위한 대표 색인어 추출 과정, 그리고 추출된 대표 색인어를 이용하여 문서간의 유사도에 따라 문서를 분류하는 클러스터링 과정으로 구성된다. 이러한 문서 자동 분류 과정에서 문서 분류의 성능 개선을 위하여 가장 중요한 과정은 대표 색인어를 추출하는 과정이다.

2.2 기존의 대표 색인어 추출 기법

대표 색인어 추출을 위한 기존의 방법은 사전에 분류된 학습 문서의 사용 여부에 따라 감독학습 기법과 비감독학습 기법으로 나뉘어진다. 다음 (그림 1)은 감독학습 기법과 비감독학습 기법을 비교한 그림이다.

2.2.1 감독학습 기반의 대표 색인어 추출 기법

감독학습 기반의 대표 색인어 추출 기법은 카테고리별로 사전에 분류된 문서를 대상으로 대표 색인어를 추출하는 기법이다. 이 방법은 사전에 분류된 학습용 문서를 이용하기 때문에 비



a. 감독학습 기법 b. 비감독학습 기법

(그림 1) 감독학습 기법과 비감독학습 기법의 비교

감독학습 기법보다 비교적 정확하게 대표 색인어 추출이 가능하다. 하지만 사전에 각 카테고리별로 학습용 문서를 분류하기 위한 비용과 카테고리를 재구성할 경우에 대표 색인어를 재구성해야 하는 어려움이 있다. 감독학습 기법에 의한 대표 색인어 추출 기법에는 DF(Document Frequency), IG(Information Gain), MI(Mutual Information), χ^2(Chi Square), TS(Term Strength), LSI(Latent Semantic Indexing) 등이 있다[Yang & Pdedersen 1997].

먼저, DF 기법은 단어가 출현한 문서의 절대 빈도수만을 고려하는 기법으로 단순하고, 분류 성능도 비교적 우수하지만 출현 빈도만으로 문서의 카테고리를 결정할 수 없는 경우도 많이 발생할 수 있다. IG 기법은 정보검색에서 주로 사용되는 색인어 추출 기법으로, 카테고리별 단어의 평균 빈도수를 고려하는 방법이다. MI 기법은 텍스트 분류에서 많이 사용되는 색인어 추출 기법으로 단어와 카테고리간의 독립성을 고려하여 카테고리별 대표 단어를 추출한다. 하지

만 빈도수가 극히 낮은 단어를 제거하기 어렵다. χ^2 기법은 우연성 테이블을 이용하여 단어와 카테고리의 독립성을 고려하는 기법으로 텍스트 분류에서 비교적 높은 정확도를 나타낸다. 하지만, 저빈도 단어들에 대해서는 고려하기 어려운 단점을 가지고 있다. TS 기법은 코사인 계수와 같은 식에 의한 유사도 계산을 통해 문서들을 사전에 클러스터링 한 후, 유사한 문서 쌍 내에서 출현 확률이 높은 단어만을 대표 색인어로 추출하는 기법이다. 하지만 문서 쌍간의 유사도 비교에 따른 클러스터링에 소요되는 비용이 클 뿐만 아니라, 분류 성능도 다른 기법에 비해 낮다[Yang & Pdedersen 1997]. LSI 기법은 정보 검색에서 동의어, 다의어 문제를 해결하기 위해 SVD(Singular Value Decomposition)를 이용한 기법이다. 하지만 문서와 용어의 수가 많아질수록 계산량이 $O(N^2*K^3)$로 증가하여 대량의 문서에는 적용하기 어렵다[Hull 1994].

2.2.2 비감독학습 기반의 대표 색인어 추출 기법

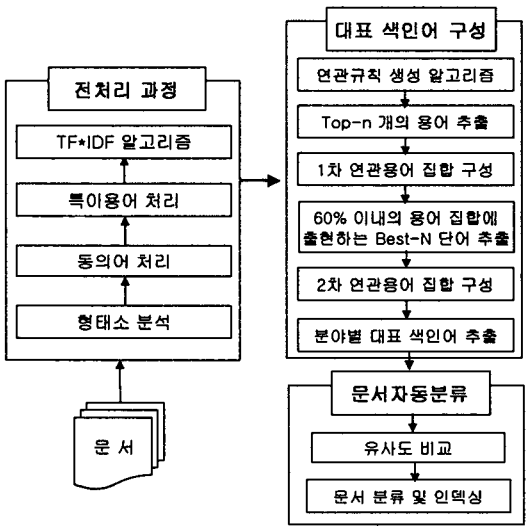
비감독학습 기반의 대표 색인어 추출 기법은 기분류된 학습 문서 집합을 사용하지 않고 분류하고자 하는 문서에서 직접 대표 색인어를 추출하는 방법이다. 이 방법은 학습용 문서를 분류하는 비용이 불필요하며 카테고리를 동적으로 재구성할 수 있는 장점이 있다. 그러나 비감독학습 기법에 의해 문서에 대한 관심 영역 즉, 카테고리를 예측하여 분류하는 방법에 대한 연구가 필요하다[백혜정 1997].

비감독학습 기법에 기반한 대표 색인어 추출 기법에는 TF*IDF 알고리즘이 있다. TF*IDF 알고리즘은 임의의 문서에 대한 TF*IDF 값을 구하여 대표 색인어를 추출하는 방법이다. 따라서 TF*IDF 알고리즘을 이용한 대표 색인어 추출은 사전에 분류된 학습용 문서없이 분류대상 문서에서 직접 대표 색인어를 추출함으로써 저비용으로 분류가 가능하다. 그러나 감독학습 기법과는 다르게 관심 영역에 대한 특징을 직접 추출하기 어렵고 단지 그 문서에 대한 중요한 단어만 추출할 수 있다. 이러한 TF*IDF 기법은 분야별 대표 색인어 추출보다는 전처리 단계에서 특징벡터의 수를 줄이는데 주로 사용되고 있다. 본 논문에서도 TF*IDF를 전처리 단계에서 단어의 빈도에 따른 가중치를 조정하여 단어의 수를 줄이는데 사용하였다.

3. 효율적인 대표 색인어 추출 기법

3.1 대표 색인어 추출 기법

본 논문에서는 데이타마이닝 기법 중 하나인 연관 규칙 탐사 알고리즘을 사용하여 비감독학습 기법에 의한 대표 색인어 추출 기법을 제시하였다. 즉, 사전에 분류되지 않은 대량의 문서로부터 직접 대표 색인어를 추출하기 위한 방법이다. 제안한 대표 색인어 추출 기법은 분류대상 문서로부터 전문용어를 추출하기 위한 전처



(그림 2) 비감독학습 기법에 의한 대표 색인어 추출 기법에 의한 문서 자동 분류 모델

리 과정, 전문용어 집합을 대상으로 연관 규칙 탐사 알고리즘을 적용하여 1차 연관 용어 집합을 생성하는 단계, 1차 연관 용어 집합 중에서 일정한 빈도수 이상으로 나타나는 용어들을 포함하는 2차 연관 용어 집합을 생성하는 단계로 이루어진다. 다음 (그림 2)는 본 논문에서 제안한 비감독학습 기법에 의한 대표 색인어 추출 기법을 사용한 분류 모델이다.

3.2 전처리 과정

3.2.1 전문용어 추출

먼저, 전체 문서 집합을 대상으로 형태소 분석을 통하여 문서에서 출현하는 모든 단어를 추출하였다. 형태소 분석기는 공개용 형태소 분석기인 HAM4.0a[강승식 1999]를 사용하였다. 형태소 분석 과정에서 추출된 단어 집합에 대하여 컴퓨터용어 사전에 수록된 전문용어만을 별도로 추출하였다.

3.2.2 동의어 처리

전문용어 중에서 같은 의미를 가진 용어이지만 저자에 따라 영어나 한국어 용어를 혼용하고 있다. 특히 영어로 된 전문용어를 한글로 표기하는 경우에 자주 발생한다. 이러한 동의어는 별도의 동의어 사전을 구성하여 표준화하였다. 예를 들어 '데이타베이스', '데이터베이스', 'database', 'databases', 'db' 등과 같은 용어는 하나의 용어로 표준화하였다.

3.2.3 특이용어 처리

전체 문서에서 출현하는 절대 빈도수가 매우 적은 용어는 연산 시간만 낭비하고 최소지지도를 만족하지 못하기 때문에 연관 규칙으로 발견되지 않는다. 그리고 전문용어이지만 모든 분야에서 공통적으로 사용되는 전문용어는 특정 분야를 대표하는 용어로 보기 어렵다. 본 논문에서는 이러한 용어를 특이용어로 처리하여 연관 규칙 탐사 과정에서 제외시킴으로서 무의미한 연관 규칙의 양산을 방지하여 대표 색인어를 효율적으로 추출할 수 있도록 하였다.

3.2.4 단어 빈도 가중치 조정

문서에서 동의어나 특이용어를 처리하더라도 여전히 많은 단어를 포함하고 있다. 따라서 분류과정의 비용을 줄이고 분류 효율을 향상시키기 위하여 분류과정에 영향을 적게 미치는 단어들을 효과적으로 제거하는 방법이 필요하다. 일반적으로 문서 분류에 영향을 미치지 않는 단어를 제거하기 위하여 단어 빈도수(Term Frequency)를 가장 많이 고려한다. 하지만 한 문서에서 출현하는 단어의 빈도수가 높다고 그 문서를 대표하는 단어가 된다고 확신하기 어렵다. 예를 들어 '시스템' 이라는 용어는 컴퓨터 용어이지만 대부분의 컴퓨터 관련 논문에서 공통적으로 출현하며 빈도수도 높기 때문에 특정 분야를 대표하는 용어로 판정하기 어렵다. 이러한 단어 빈도수에 의한 문제점을 해결하기 위하여 여러 가지 형태의 가중치 공식들이 제안되었다[이재윤 2000].

본 논문에서는 TF*IDF 알고리즘을 적용하여 모든 문서에서 공통적으로 출현하는 단어에 대한 가중치를 조정하였다. TF*IDF 알고리즘은 하나의 문서에서 출현하는 단어의 빈도수에 역문서 빈도수(Inverse Document Frequency)를 가중치로 적용하여 문서를 대표하는 단어들을 효과적으로 선별할 수 있는 알고리즘이다[Salton 1991].

3.3 대표 색인어 추출

3.3.1 1차 연관 용어 집합 생성

전체 문서에서 연관 규칙 탐사 알고리즘을 적

용하여 전문용어들간의 연관성을 분석하고 1차 연관 용어 집합을 구성하였다. 그리고 Seed Keyword 별로 연관성이 높은 단어들을 하나의 집합으로 구성하였다. 여기서 Seed Keyword 집합은 각 카테고리를 대표하는 특징단어 집합이다. 연관 규칙을 발견하기 위한 트랜잭션 단위는 하나의 문서에서 추출된 전문용어 집합이다. 전문용어 집합은 전처리 과정에서 형태소 분석을 통하여 추출된 모든 용어에 대하여 전문용어 사전에 수록된 용어를 추출하여 구성하였다. 그리고 같은 의미를 가지는 동의어를 표준화하고 불필요한 연산이나 연관 규칙을 양산할 수 있는 특이용어도 제거하였다. 다음 <표 1>은 약 240편의 컴퓨터 분야 논문에 대하여 연관 규칙 알고리즘을 적용하여 생성된 1차 연관 용어 집합의 예이다.

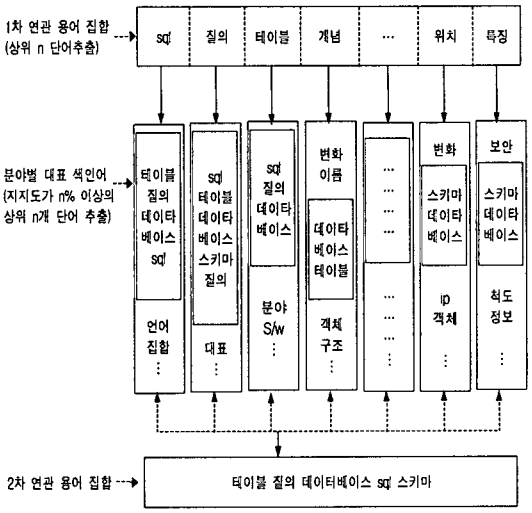
〈표 1〉 1차 연관 용어 집합

연관 용어	연관용어집합	연관 용어	연관용어집합
데이타 베이스	SQL, 질의, 인스턴스, 객체, 테이블, …	SQL	질의, 테이블 데이타베이스 인스턴스, 백업, …
테이블	SQL, 질의, 데이타베이스, 인스턴스, 위치, …	질의	데이타베이스, SQL, 인스턴스, 관계형, …
인스 턴스	SQL, 관계형, 데이타베이스, 스키마, 특징, …	네트 워크	IP, 프로토콜, 전송, 통신, 데이타베이스, …

3.3.2 2차 연관 용어 집합 생성

1차 연관 용어 집합 중에서 일정한 빈도수 이상 공통으로 나타나는 용어들을 포함시켜 2차 연관 용어 집합을 생성하여 카테고리별 대표 색인어로 구성하였다. 다음 (그림 3)은 1차 연관 용어 집합 중에서 일정한 빈도수 이상 공통으로 포함된 2차 연관 용어 집합을 구성하는 과정을 나타낸다.

위 (그림 3)에서처럼 1차 연관 규칙 탐사의 결과에서 추출된 '개념', '위치', '특징' 등과 같은 단어들은 지지도가 임계치 이하이기 때문에 제거되었다. 그리고 '스키마'는 1차 연관 규칙 탐사 결과에서는 추출되지 않았지만 주어진 Seed Keyword인 '데이타베이스'와 밀접한 관련이 있는 단어로 추출되어 대표 색인어에 포함되었다.



(그림 3) 연관 용어 집합을 이용한 대표 색인어 추출

3.3.3 추출된 대표 색인어를 이용한 문서 분류

문서 분류 과정에서는 대표 색인어와 문서간의 유사도를 계산하여 문서에 대한 분류 실험을 하였다. 모든 분야를 대상으로 계산한 유사도 중에서 최대값을 가지는 분야가 해당 문서가 속하는 분야이다. 대표 색인어와 문서간의 유사도 계산을 위하여 코사인 계수를 사용하였다. 코사인 계수는 비교하고자 하는 두 대상에 대한 특징간의 일치 정도를 측정하는 기법으로 문서 분류에서 주로 사용되는 유사도 계수이다[Goldszmidt & Sahami 1998]. 다음 식 (1)은 대표 색

인어와 문서간의 유사도를 계산하기 위한 코사인 계수식이다. 여기서 X는 분류하고자 하는 문서에 대한 단어 벡터이고, Y는 추출된 분야별 대표 색인어를 나타낸다.

$$\cos\theta(X, Y) = \frac{\sum_{i=1}^{n} X_i \cdot Y_i}{\sqrt{\sum_{i=1}^{n}(X_i)^2 \cdot \sum_{i=1}^{n}(Y_i)^2}} \quad (1)$$

3.3.4 제안된 대표 색인어 기법의 성능 평가

문서분류 결과에 대한 성능 평가를 위한 척도로 Recall, Precision, F-measure 값을 주로 사용한다[Michael 1994]. Recall 값은 카테고리내의 전체 문서 중에서 정확하게 분류된 문서의 분류 비율을 의미한다. 다음 식 (2)는 Recall의 정의식이다.

$$Recall = \frac{\text{정확하게 분류된 문서수}}{\text{해당 카테고리의 문서수}} \times 100 \quad (2)$$

높은 Recall 값은 카테고리내의 대부분 문서가 정확하게 분류되었다는 것을 의미한다. 하지만 다른 카테고리의 문서가 해당 카테고리로 오분류된 문서에 대해서는 고려하기 어렵다. 예를 들면, 해당 카테고리에 속한 실제 문서수는 100건이지만 분류기를 통해 다른 카테고리의 문서 100건도 같이 분류되더라도 Recall 값은 여전히 100%가 된다. 따라서 분류의 정확성 측면에서 신뢰성이 다소 떨어진다.

Precision 값은 분류된 문서중에서 정확하게 분류된 문서의 비율을 의미한다. 다음 식 (3)은 Precision의 정의식이다.

$$Precision = \frac{\text{정확하게 분류된 문서수}}{\text{분류된 전체 문서수}} \times 100 \quad (3)$$

높은 Precision 값은 해당 카테고리에 분류된 거의 모든 문서들이 정확하게 분류되었다는 것을 나타낸다. 하지만 분류 자체의 오류에 대해서는 고려하지 못하는 단점이 있다. 극단적인 경우의 예를 들면, 실제 해당 카테고리의 문서수가 100건이지만 분류된 문서수가 1건이더라도 해당 문서가 정확하게 분류되면 Precision 값은 100%가 된다.

이러한 Recall과 Precision 값은 서로 반비례 관계에 있으므로 적절한 조정 과정이 필요하다. Lewis 등은 Recall과 Precision의 문제점을 보완하기 위하여 Recall과 Precision을 결합한 F-Measure 개념을 제안하였다[Lewis & Gale 1994].

다음 식 (4)는 F-measure에 대한 정의식이다.

$$F_\beta = \frac{(\beta^2+1) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall} \quad (4)$$

β는 Recall 값과 Precision 값의 중요도에 따른 가중치를 나타낸다. 즉 $\beta = 0$ 이면 F 값은 Precision 값과 동일하고, $\beta = \infty$ 이면 F 값은 Recall 값과 동일하다. $\beta = 1$ 이면 Recall 값과 Precision 값에 동일한 가중치를 적용하여 F 값을 계산한다. 그리고 $\beta = 0.5$이면 Recall 값에 0.5배의 가중치를 적용하여 Precision 값에 대한 중요도를 높여서 계산한다. $\beta = 2$ 이면 Recall 값에 2배의 가중치를 적용하여 Recall 값에 대한 중요도를 높여서 계산한다. 그러므로 Recall 값과 Precision 값의 중요도에 따라 β의 가중치를 선택적으로 조정할 수 있다.

본 논문에서는 제안된 대표 색인어 추출 기법의 정확성을 검증하기 위하여 $\beta = 1$ 즉, Recall 값과 Precision 값에 동일한 가중치를 적용하여 분류 성능을 평가하였다.

4. 실험 결과 및 고찰

본 논문에서 제안한 비감독학습 기법에 의한

대표 색인어 추출기법에 대한 효율성을 검증하기 위하여 컴퓨터 관련 논문을 대상으로 분류 실험을 하였다. 제안한 기법의 성능을 평가하기 위하여 감독학습 기법에 의한 대표 색인어 추출 기법 중에서 비교적 우수한 성능을 지닌 χ^2 기법, DF 기법과 비교 실험을 하였다.

4.1 실험 환경

대표 색인어 추출 기법에 대한 실험 환경은 Sun Solaris 2.6에서 오라클 DBMS 8.0.5를 기반으로 구현하였다.

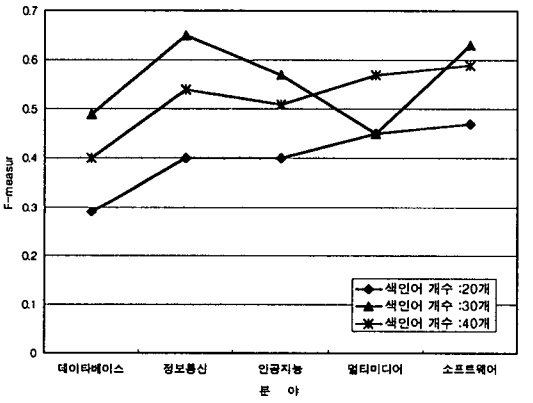
분류 실험은 컴퓨터 관련 분야 240편의 논문을 대상으로 실험하였다. 전체 논문에서 추출된 전문용어는 24,986개이며 편당 평균 104개의 전문용어가 추출되었다. 동의어 처리를 통해 용어를 표준화한 결과 전체 용어수는 22,838개, 평균 95개로 줄어들었다. 그리고 전체 출현 빈도수가 2이하인 용어는 약 242개이고, 전체 문서수에 대한 특정 용어의 출현 문서수의 표준 편차가 8 이하로 분포도가 큰 전문용어는 148개이다. 이러한 특이용어를 제외한 최종적인 전문용어의 수는 1,499개이다.

4.2 실험 결과

먼저, 각 카테고리를 대표하는 Seed Keyword를 선정하였다. 그리고 연관 규칙 탐사 알고리즘을 적용하여 각 Seed Keyword와 관련된 용어 집합을 추출하였다. 추출된 용어 집합에서 상위 50개의 용어로 1차 연관 용어 집합 $K_{si} = \{T_1, T_2, T_3, \cdots, T_n\}$을 구성하였다. 두 번째 단계에서는 연관 용어 집합 K_{si}의 각 원소 $T_i(1 < i < n)$에 대해 연관 용어 집합을 추출하여 1차 연관 용어 집합과 그룹화 하였다. 각 그룹에 대해 최소 지지도 60% 이상의 상위 30개의 연관 용어을 추출하여 2차 연관 용어 집합 즉, 대표 색인어로 구성하였다. 이때 상위 50개와 지지도 60% 이상의 상위 30개는 다양한 경우에 대한 분류 실험을 통하여 최적의 값으로 선정하였다. 다음 <표 2>와 (그림 4)에서처럼 대표 색인어 수를 30개로 구성하였을 때 모든 분야에서 가장 좋은 분류 성능을 보였다.

〈표 2〉 F-measure 값에 의한 대표 색인어 개수별 문서 분류 결과

색인어 개수 \ 분야	데이타베이스	정보통신	인공지능	멀티미디어	소프트웨어	평균
20	0.29	0.4	0.4	0.45	0.47	0.40
30	0.49	0.65	0.57	0.45	0.63	0.56
40	0.4	0.54	0.51	0.57	0.59	0.54



(그림 4) 대표 색인어 개수별 성능 평가 비교

실험결과에 따라 최적의 대표 색인어 개수는 분야별로 상위 30개의 용어집합으로 구성하였다. 제안된 방법을 통하여 전처리 과정을 거친 1,499개의 전문용어 중에서 71%의 용어를 제거한 총 210개의 단어로 대표 색인어를 구성하여 분류를 위한 계산 비용을 최소화하였다. 다음 <표 3>은 제안 기법에 의해 추출된 분야별 대표 색인어이다.

제안된 방법과 χ^2 기법, DF 기법에 의해 추출한 대표 색인어를 이용한 비교 실험을 하였

다. 분류대상 문서에 대하여 코사인 계수를 사용하여 분야별 대표 색인어와 분류대상 문서간의 유사도를 계산하여 해당 문서를 가장 유사한 카테고리에 분류하였다.

〈표 3〉 제안 기법으로 추출한 대표 색인어 집합

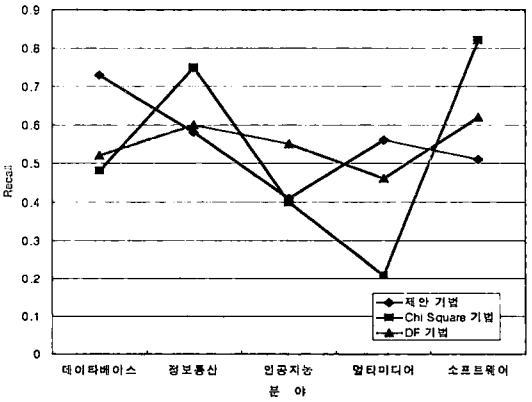
분 야	대 표 색 인 어
데이터베이스	데이타베이스, 디자인, 접근, 객체, 관계, 질의, 연결, 표현, 테이블, 검색, 컨텐츠, sql, 선택, 인터페이스, 특징, 영역, 전달, 통신, 추출, 개념, 위치, 분류, 컨트롤, 알고리즘, 변경, 변환, 전송, 네트워크, 이름, 공간
정 보 통 신	통신, 디자인, 연결, 접근, 프로토콜, 관계, 패킷, 호스트, 가상, 컨텐츠, 데이타베이스, 표현, 객체, 검색, tcp, 선택, 인터페이스, 특징, 영역, 전달, 위치, ip, 분류, 컨트롤, 개념, 변경, 알고리즘, 네트워크, 변환, 전송
소프트웨 어	디자인, 관계, 객체, 표현, 연결, 접근, 검색, 데이타베이스, 인터페이스, 컨텐츠, 선택, 특징, 요구사항, 이벤트, 재사용, 초점, 캡슐, 캡슐화, 컴포넌트, 영역, 개념, 추출, 구성요소, 분류, 개발자, 스펙, 소프트웨어, 식별자, 자원, 참조

다음 <표 4>와 (그림 5)는 Recall 값을 이용한 제안 기법, χ^2 기법, DF 기법을 비교 실험한 결과이다.

실험결과, 제안 기법은 데이터베이스 분야, 멀티미디어 분야에서 높은 Recall 값을 가지고, Recall 값의 평균은 0.56으로 가장 높았다. χ^2 기법은 정보통신 분야, 소프트웨어 분야에서 높은 Recall 값을 가진다. χ^2 기법의 Recall 값의 평균은 0.53이다. DF 기법은 인공지능 분야에서만 높은 Recall 값을 가지며, 인공지능 분야에서만 다른 기법에 비해 정확하게 분류되었음을 알 수 있다. DF 기법의 분야별 Recall 값의 평균은 0.55이다. 제안 기법은 분야별로 Recall 값의 차이가 있지만 평균적으로 다른 기법에 비해 가장 우수한 성능을 보였다.

〈표 4〉 제안 기법과 다른 기법과의 Recall 값 비교

Recall	데이타베이스	정보통신	인공지능	멀티미디어	소프트웨어	평 균
제안 기법	0.73	0.58	0.41	0.56	0.51	0.56
Chi Square 기법	0.48	0.75	0.40	0.21	0.82	0.53
DF 기법	0.52	0.60	0.55	0.46	0.62	0.55



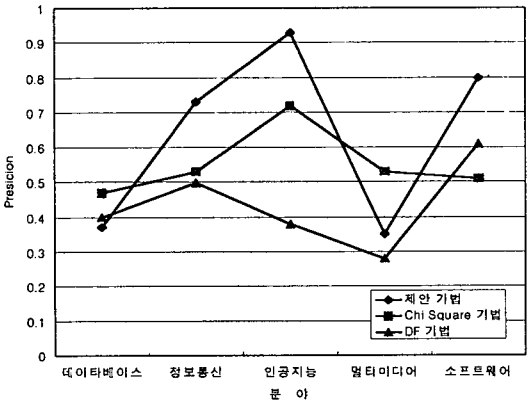
(그림 5) 제안 기법과 다른 기법과의 Recall 값 비교

다음 <표 5>와 (그림 6)은 Precision 값을 이용한 제안 기법, χ^2 기법, DF 기법을 비교 실험한 결과이다.

실험결과, 제안 기법은 정보통신 분야, 인공지능 분야, 소프트웨어 분야에서 높은 Precision 값을 가지고, 분야별 Precision 값의 평균은 0.64으로 가장 높았다. χ^2 기법은 데이터베이스 분야, 멀티미디어 분야에서 높은 Precision 값을 가진다. χ^2 기법의 분야별 Precision 값의 평균은 0.55이다. DF 기법은 모든 분야에서 가장 낮은 Precision 값을 가지며, 분야별 평균 Precision 값은 0.43이다. 제안 기법의 분야별 평균 Precision 값은 0.64로 다른 기법에 비해 가장 우수한 성능을 보였다.

〈표 5〉 제안 기법과 다른 기법과의 Precision 값 비교

Precision	데이타베이스	정보통신	인공지능	멀티미디어	소프트웨어	평 균
제안 기법	0.37	0.73	0.93	0.35	0.8	0.64
Chi Square 기법	0.47	0.53	0.72	0.53	0.51	0.55
DF 기법	0.40	0.50	0.38	0.28	0.61	0.43



(그림 6) 제안 기법과 다른 기법과의 Precision 값 비교

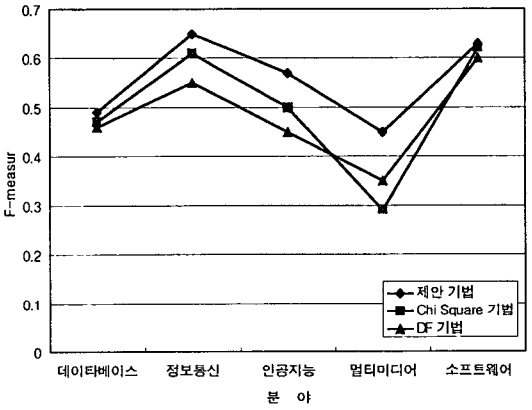
그러나 Recall 값은 오분류된 것에 대해서는 고려되지 않기 때문에 Recall 값이 높다그 해서 정확하게 분류되었다는 것을 의미하지는 않는다. 또한 Precision 값은 분류된 문서 중에서만 정확도를 계산하기 때문에 Precision 값이 높다고 해서 정확하게 분류되었다는 것을 의미하는 것은 아니다. 그러므로 신뢰성 있는 분류 성능 측정을 위해서는 Recall 값과 Precision 값을 결합한 F-measure 값에 대한 비교가 필요하다.

다음 <표 6>과 (그림 7)은 F-measure 값을 이용한 제안 기법, χ^2 기법, DF 기법을 비교 실험한 결과이다.

실험 결과, 제안 기법은 모든 분야에서 높은 F-measure 값을 가지며, 분야별 평균 F-measure 값은 0.56이다. χ^2 기법은 제안 기법보다는 낮지만 DF 분야보다는 높은 F-measure 값을 가지며 분야별 평균 F-measure 값은 0.50이다. DF 기법은 제안 기법보다는 낮지만 멀티미디어 분야, 소프트웨어 분야에서 χ^2 기법보다 높은 F-measure 값을 가지며 분야별 평균 F-measure 값은 0.48이다. 제안 기법은 분야별 평균 F-meausre 값이 0.56으로 다른 기법에 비해 가장 우수한 성능을 보였다.

〈표 6〉 제안 기법과 다른 기법과의 F-measure 값 비교

F-measure	데이타베이스	정보통신	인공지능	멀티미디어	소프트웨어	평 균
제안 기법	0.49	0.65	0.57	0.45	0.63	0.56
Chi Square 기법	0.47	0.61	0.50	0.29	0.62	0.50
DF 기법	0.46	0.55	0.45	0.35	0.60	0.48



(그림 7) 제안 기법과 다른 기법과의 F-measure 값 비교

5. 결 론

본 논문에서는 대량의 문서를 자동으로 분류하기 위하여 비감독학습 기법에 의해 카테고리별 대표 색인어를 구성하기 위한 방법을 제안하였다. 제안된 방법에서는 사전에 문서를 분류하지 않고 대표 색인어를 추출하기 위하여 데이타마이닝 기법 중의 하나인 연관 규칙 탐사 알고리즘을 이용하였다.

먼저, 각 카테고리를 대표하는 Seed Keyword를 선정한다. 그리고 연관 규칙 탐사 알고리즘을 적용하여 각 Seed Keyword와 관련된

용어 집합을 추출한다. 추출된 용어 집합에서 상위 50개의 용어들로 1차 연관 용어 집합 $K_{si} = \{T_1, T_2, T_3, \cdots, T_n\}$을 구성한다. 두 번째 단계에서는 연관 용어 집합 K_{si}의 각 원소 $T_i(1 < i < n)$에 대해 연관 용어 집합을 추출하여 1차 연관 용어 집합과 그룹화한다. 각 그룹에 대해 최소 지지도 60% 이상의 상위 30개의 연관 용어들을 재그룹하여 2차 연관 용어 집합을 추출한다.

제안된 기법의 성능을 검증하기 위하여 컴퓨터 관련 논문을 대상으로 분류 실험을 하였다. 대표 색인어 추출을 위한 학습용 문서는 컴퓨터 관련 학회에서 발표한 논문을 240편을 사용하였고, 분류 실험에서는 대표 색인어 추출 과정에서 사용하지 않은 분야별 30여편의 논문을 사용하였다. 대표적인 감독학습 기법인 χ^2 기법, DF 기법과의 비교 실험을 통하여 제안 기법의 성능을 평가하였다.

실험 결과 감독학습 기법의 대표 색인어 추출 기법 중에서 우수하다고 알려진 χ^2 기법과 DF 기법보다 우수한 분류 성능을 보였다.

참 고 문 헌

[1] 정영미, 정보검색론, 구미무역(주) 출판부, 1993

[2] 강승식, "HAM:한국어 형태소 분석 라이브러리", http://ham.hansung.ac.kr/ham/ham-intr.html

[3] 백혜정, 박영택, 윤석환, "사용자 관심도를 이용한 웹 에이전트", 정보처리학회지, 제4권 제5호, pp.88-100, 1997

[4] 이재윤, "문헌 자동분류에서 용어 가중치 기법에 대한 연구", 한국정보관리학회 학술대회 논문집, pp.41-44, 2000

[5] 진 훈, 김인철, "문서 분류를 위한 특징 선택", 2001 봄 학술발표논문집, 한국정보과학회, 제28권 1호, pp.262-264, 2001

[6] 홍진혁, 류중원, 조성배, "실세계의 FAQ 메일 자동분류를 위한 문서 특징추출 방법의 성능 비교", 2001 봄 학술발표논문집, 한국정보과학회, 제28권, 제1호, pp.271-273, 2001

[7] Hull, D., "Improving Text Retrieval for the Routing Problem using Latent Semantic Indexing," Proceeding of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.282-291, 1994

[8] Lewis, D. and Gale, W.A., "A Sequential Algorithm for Training Text Classifiers," In Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, London, Springer-Verlag, pp.3-12, 1994

[9] Michael Buckland, "The Trade-off between Recall and Precision," Journal of the American Society for Information Science, pp.12-19, 1994

[10] M. Goldszmidt, M. Sahami, "A Probabilistic Approach to Full-Text Document Clustering," Technical Report ITAD-433 MS-98-044, SRI International, pp.434-444, 1998

[11] Salton G., "Developments in Automatic Text Retrieval," Science, Vol.253, pp.974-979, 1991

[12] Yang Y., J.O. Pdedersen, "A Comparative Study on Feature Selection in Text Categorization," Proceedings of the 14th International Conference on Machine Learning ICML-97, pp.412-420, 1997

■ 저자소개

김 지 숙

경남대학교 전자계산학과에서 공학사, 창원대학교 대학원 전자계산학과에서 이학석사를 취득하였으며, 주요관심분야는 텍스트마이닝, 데이터마이닝 등이다.

김 영 지

창원대학교 전자계산학과에서 이학사, 창원대학교 대학원 전자계산학과에서 이학석사를 취득하였으며, 현재 창원대학교 대학원 전자계산학과 박사과정 중이다. 주요관심분야는 데이터마이닝, 검색엔진, 추천 알고리즘 등이다.

문 현 정

한국방송통신대학 전자계산학과에서 이학사, 창원대학교 대학원 전자계산학과에서 이학석사를 취득하였으며, 창원대학교 대학원 전자계산학과 박사과정 수료하였으며, 주요관심분야는 텍스트마이닝, 데이터마이닝, CRM 등이다.

우 용 태

경북대학교 전자공학과에서 공학사, 경북대학교 대학원 전자공학과에서 공학석사, 경북대학교 대학원 전자공학과에서 공학박사를 취득하였으며, 현재 창원대학교 컴퓨터공학과 교수로 재직중이다. 주요관심분야는 데이터마이닝, CRM, 추천 알고리즘 등이다.