# 키워드 추출 기법에 관한 연구

신성윤*, 정경택**, 표성배***, 이양원*
*군산대학교 컴퓨터정보공학과
**군산대학교 전자정보공학부
***인덕대학 컴퓨터 소프트웨어과
e-mail:{s3397220, ywrhee}@kunsan.ac.kr, pyosb@induk.ac.kr

# A Study for Keyword Extraction Method

Shin Seong Yoon*, Jeong Kyong Taek**, Rhee Yang Won*
*Dept. of Computer Information Engineering, Kunsan National University
**School of Electrical and Information Engineering, Kunsan National University
*** Dept. of Computer Software, Induk Institute of Technology

본 논문에서는 대량의 문제를 자동으로 분류하기 위하여 비감독 학습 기법에 의해 카테고리별 키워드를 구성하기 위한 방법을 제안하였다. 제안된 방법에서는 사전에 문제를 분류하지 않고 키워드를 추출하기 위하여 데이터마이닝 기법 중의 하나인 연관 규칙 탐사 알고리즘을 이용하였다. 먼저, 각 카테고리를 대표하는 핵심 키워드를 선정하고, 연관 규칙 탐사 알고리즘을 적용하여 각 핵심 키워드와 관련된 용어 집합을 추출한다.

키워드 : 비감독 학습 기법, 키워드, 연관 규칙 탐사 알고리즘, 데이터마이닝

## I. 서론

문서분류시스템에서 각 카테고리를 대표하는 키워드가 먼저 구성되어야 한다. 이 키워드는 문서의 내용에 따라 적절한 카테고리를 지정하여 문서를 자동적으로 분류하기 위해서이다. 이렇게 카테고리별 키워드를 추출하기 위한 연구 방법은 여러 개가 존재한다. 그중에서 학습용 문서의 사전 분류 여부에 따라 감독 학습 기법과 비감독 학습 기법으로 나누어진다. 감독 학습 기법은 비교적 정확한 키워드가 추출된다. 하지만 사전에 각 카테고리별로 학습용 문서를 분류하기 위한 비용과 카테고리를 재구성할 경우에 키워드를 다시 추출해야 하는 어려움이 있다. 반면에 비감독 학습 기법은 사전에 분류된 학습 문서를 사용하지 않고 분류 대상 문서에서 직접 키워드를 추출하여 분류하기 때문에 적은 비용으로 동적 카테고리를 재구성할 수 있다.

본 논문에서는 비감독 학습 기법에 의해 키워드를 추출하는 방법을 제안한다. 먼저, 각 카테고리를 대표하는 핵심 키워드 집합을 선정하는데, 여기서 핵심 키워드 집합이란 각 카테고리를 대표하는 특정 단어 집합이다. 그리고 연관 규칙 탐사 알고리즘을 적용하여 각 핵심 키워드와 관련된 용어 집합을 추출한다. 추출된 용어 집합에서 상위 N개의 용어들로 연관 용어 집합 $K_{si}=\{T_1, T_2, T_3, \ldots, T_n\}$을 구성한다. 이렇게 추출된 연관 용어 집합을 대표 키워드로 구성한다.

제안된 기법의 효율성을 검증하기 위하여 컴퓨터 관련 논문을 대상으로 분류 실험을 하였다. 키워드 추출을 위한 문서는 컴퓨터 관련 문제 500개를 이용하였고, 분류 실험에서는 키워드 추출 과정에 사용되지 않은 분야별 100여개의 문제를 이용하였다. 기존의 감독 학습 기법 중에서 대표 키워드 추출 성능이 우수하다고 알려진 $X^2$ 기법, Document Frequency(DF) 기법[1][2][3]과의 비교 실험을 통하여 제안된 기법에 대한 성능 평가를 수행하였다.

## II. 관련 연구

문제 자동 분류란 컴퓨터를 이용하여 문제를 대표하는 특징으로 구성된 키워드 집합과 유사한 문제들을 같은 그룹으로 분류하는 기법이다. 컴퓨터를 이용하여 문제를 자동으로 분류하기 위한 시도는 70년대 말 Salton에 의해 체계화되기 시작하였다[4]. 문제 자동분류 기법은 지식관리시스템의 가장 핵심적인 요소로서 사용자가 원하는 지식정보를 효과적으로 검색하기 위해 대량의 문제를 자동 분류하는 기술이다. 문제 분류 과정은 전체 문제 집합에서 분류에 영향을 주지 않는

의미 없는 단어를 제거하기 위한 전처리 과정, 문제 내용에서 중심이 되는 특징을 구성하기 위한 키워드 추출 과정, 그리고 추출된 키워드를 이용하여 문제간의 우유사도에 따라 문제를 분류하는 클러스터링 과정으로 구성된다. 이러한 문제 자동 분류과정에서 문제 분류의 성능 개선을 위하여 가장 중요한 과정은 키워드를 추출하는 과정이다.

키워드 추출을 위한 기존의 방법은 사전에 분류된 학습 문제의 사용 여부에 따라 감속 학습 기법과 문제의 사용 여부에 따라 감독 학습 기법과 비감독 학습 기법으로 나누어진다. 다음 〈그림 1〉은 비감독 학습 기법을 나타낸다.

비감독 학습 기반의 키워드 추출 기법은 기분류된 학습 문제 집합을 사용하지 않고 분류 하고자 하는 문제에서 직접 키워드를 추출 하는 방법이다. 이 방법은 학습용 문제를 분류하는 비용이 불필요하며 카테고리를 동적으로 재구성할 수 있는 장점이 있다. 그러나 비감독 학습 기법에 의해 문제에 대한 관심 영역 즉, 카테고리를 예측하여 분류하는 방법에 대한 연구가 필요하다[5].

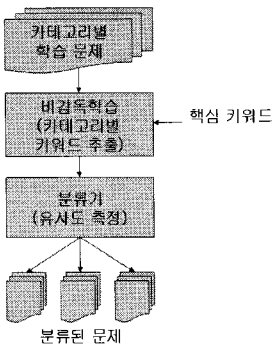


그림 1. 비감독 학습 기법

## III. 키워드 추출 기법

본 논문에서는 데이터 마이닝 기법 중 하나인 연관 규칙 탐사 알고리즘을 사용하여 비감독 학습 기법에 의한 키워드 추출 기법을 제안하였다. 즉, 사전에 분류되지 않은 대량의 문제로부터 직접 키워드를 추출하기 위한 방법이다. 제안한 키워드 추출 기법은 분류대상 문제로부터 전문 용어를 추출하기 위한 전처리 과정, 전문 용어 대상을 대상으로 연관 규칙 탐사 알고리즘을 적용하여 연관 용어 집합을 생성하는 단계로 이루어진다.

### 3.1 전처리 과정

첫째, 전체 문제 집합을 대상으로 형태소 분석을 통하여 문제에서 출현하는 모든 단어를 추출하였다.

둘째, 동의어는 별도의 동의어 사전을 구성하여 표준화하였다.

셋째, 전체 문제에서 출현하는 절대 빈도수가 매우 적은 용어는 연산 시간만 낭비하고 최소지지도를 만족하지 못하기 때문에 연관 규칙으로 발견되지 않는다.

넷째, 일반적으로 문제 분류에 영향을 미치지 않는 단어를 제거하기 위하여 단어 빈도수(Term frequency)를 가장 많이 고려한다. 하지만 한 문제에서 출현하는 단어의 빈도수가 높다고 그 문제를 대표하는 단어가 된다고 확신하기 어렵다. 예를 들어 '시스템' 이라는 용어는 컴퓨터 용어이지만 대부분의 컴퓨터 관련 논문에서 공통적으로 출현하는 용어로 판정하기 어렵다. 이러한 단어 빈도수에 의한 문제점을 해결하기 위하여 여러 가지 형태의 가중치 공식들이 제안되었다[6].

본 논문에서는 TF*IDF 알고리즘을 적용하여 모든 문제에서 공통적으로 출현하는 단어에 대한 가중치를 조정하였다 TF*IDF 알고리즘은 하나의 문제에서 출현하는 단어의 빈도수에 역문제 빈도수(Inverse Document Frequency)를 가중치로 적용하여 문제를 대표하는 단어들을 효과적으로 선별할 수 있는 알고리즘이다[7].

### 3.2 키워드 추출 절차

전체 문제에서 연관 규칙 탐사 알고리즘을 적용하여 전문 용어들 간의 연관성을 분석하고 연관 용어 집합을 구성하였다. 그리고 핵심 키워드 별로 연관성이 높은 단어들을 하나의 집합으로 구성하였다. 여기서 핵심 키워드 집합은 각 카테고리를 대표하는 특징단어 집합이다. 연관 규칙을 발견하기 위한 트랜잭션 단위는 하나의 문제에서 추출된 전문 용어 집합이다. 전문 용어 집합은 전처리 과정에서 형태소 분석 사전에 수록된 용어를 추출하여 구성하였다. 다음 〈표 1〉은 약 500개의 컴퓨터 분야 문제에 대하여 연관 규칙 알고리즘을 적용하여 생성된 연관 용어 집합의 예이다.

표 1. 연관 용어 집합의 예

| 연관용어 | 연관용어집합 |
|---|---|
| 운영체제 | 운영체제, 사용자인터페이스, 프로세스, 자원, 메모리, 보안, … |
| 네트워트 | 네트워크, 전송매체, 유형, 랜, 통신장치, 중계기, 허브, 스위치, … |
| 인터넷 | 인터넷, 프로토콜, 유알엘, 디엔에스, 포트, 웹페이지, 검색엔진, … |

키워드와 문제은행의 문제 사이의 유사도 계산을 위하여 코사인 계수를 사용하였다. 코사인 계수는 비교하고자 하는 두 대상에 대한 특징간의 일치 정도를 측정하는 기법으로 문제 분류에서 주로 사용되는 유사도 계수이다[8]. 다음 식(1)은 키워드와 문제간의 유사도를 계산하기 위한 코사인 계수 식이다. 여기서 X는 분류하고자 하는 문제에 대한 단어 벡터이고, Y는 추출된 분야별 키워드를 나타낸다.

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

3.3 제안된 키워드 기법의 성능 평가

문제 분류 결과에 대한 성능 평가를 위한 척도 Recall, Precision, F-measure 값을 주로 사용한다. Recall 값은 카테고리 내의 전체 문제(Tq) 중에서 정확하게 분류된 문제(Cq)의 분류 비율을 의미한다. 다음 식 (2)는 Recall의 정의식이다.

$$Recall=\frac{Cq}{Tq}\times 100 \quad \cdots\cdots (2)$$

높은 Recall 값은 카테고리 내의 대부분 문제가 정확하게 분류 되었다는 것을 의미한다. 하지만 다른 카테고리의 문제가 해당 카테고리로 잘못 분류된 문제에 대해서는 고려하기 어렵다. 예로서, 해당 카테고리에 속한 실제 문제수는 100개이지만 분류기를 통해 다른 카테고리의 문제 100건도 같이 분류되더라도 Recall 값은 여전히 100%가 된다. 따라서 분류의 정확성 측면에서 신뢰성이 어느 정도 떨어진다.

Precision 값은 분류된 문제(Ctq) 중에서 정확하게 분류된 문제(Cq)의 비율을 의미한다. 다음 식 (3)은 Precision을 정의한 것이다.

$$Precision=\frac{Cq}{Ctq}\times 100 \quad \cdots\cdots (3)$$

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

이러한 Recall과 Precision 값은 서로 반비례 관계에 있으므로 적절한 조정 과정이 필요하다. Lewis 등은 Recall과 Precision을 결합한 F- Measure 개념을 제안하였다.

다음 식 (4)는 F-measure에 대하여 정의한 식이다.

$$F_\beta=\frac{(\beta^2+1)\bullet <ision\bullet recall}{\beta^2\ \bullet <ision\bullet recall} \quad \cdots\cdots (4)$$

β는 Recall 값과 Precision 값의 중요도에 따른 가중치를 나타낸다. 즉 β=0 이면 F 값은 Recall 값과 동일하다. β=∞이면 F값은 Recall 값과 동일하다. β=1이면 Recall 값과 Precision 값에 동일한 가중치를 적용하여 Recall 값에 동일한 가중치를 적용하여 F 값을 계산한다. 그리고 β=0.5이면 Recall 값에 0.5배의 가중치를 적용하여 Precision 값에 대한 중요도를 높여서 계산한다. β=2이면 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 키워드 추출 기법의 정확성을 검증하기 위하여 β=1 즉, Recall 값과 Precision 값에 동일한 가중치를 적용하여 분류 성능을 평가하였다.

## IV. 성능평가 및 분석

분류 실험은 컴퓨터 관련 분야 500개의 문제를 대상으로 실험하였다. 전체 문제에서 추출된 전문 용어는 1,656개이며 문제당 평균 3.312개의 전문용어가 추출되었다. 동의어 처리를 통해 용어를 표준화한 결과 전체 용어수는 1,360개로서 문제당 평균 2.72개로 줄어들었다. 그리고 전체 출현 빈도수가 2이하인 용어는 156개이고, 전체 문제수에 대한 특수 용어의 출현 문제수의 표준 편차가 8이하로 분포도가 큰 전문용어는 107개이다. 또한, 특수용어를 제외한 최종적인 전문용어의 수는 995개이다.

제안된 방법과 $X^2$ 기법 DF 기법에 의해 추출한 키워드를 이용한 비교 실험을 하였다.

분류 대상 문제에 대하여 코사인 계수를 사용하여 분야별 키워드와 분류대상 문제간의 유사도를 계산하여 해당 문제를 가장 유사한 카테고리에 분류하였다.

표 2 키워드의 카테고리화

| 분야 | 키워드 |
|---|---|
| 운영체제 | 운영체제, 사용자인터페이스, 프로세스, 자원, 메모리, 보안, 드라이버, 부팅, 커널… |
| 네트워트 | 네트워크, 전송매체, 유형, 랜, 통신장치, 중계기, 허브, 스위치, 무선기술, 전송매체… |
| 인터넷 | 인터넷, 프로토콜, 유알엘, 디엔에스, 포트, 웹페이지, 검색엔진, 아이피주소… |

다음 〈표 3〉는 Recall 값을 이용한 제안 기법, $X^2$ 기법, DF 기법을 비교 실험한 결과이다.

실험결과, 제안 기법은 분야별로 Recall 값의 차이가 있지만 평균적으로 다른 기법보다 가장 우수한 성능을 보였다.

표 3. 제안 기법과 다른 기법과의 Recall 값 비교

| Recall | 운영체제 | 네트워크 | 인터넷 | 평균 |
|---|---|---|---|---|
| 제안기법 | 0.75 | 0.63 | 0.57 | 0.65 |
| $X^2$ 기법 | 0.51 | 0.71 | 0.46 | 0.56 |
| DF 기법 | 0.55 | 0.62 | 0.59 | 0.59 |

다음 〈표 4〉은 Precision 값을 이용한 제안 기법, $X^2$기법, DF 기법을 비교 실험한 결과이다. 실험결과, 제안 기법의 분야별 평균 Precision 값은 0.71로 다른 기법에 비해 가장 우수한 성능을 보였다.

표 4. 제안 기법과 다른 기법과의 Precision 값 비교

| Precision | 운영체제 | 네트워크 | 인터넷 | 평 균 |
|---|---|---|---|---|
| 제안 기법 | 0.45 | 0.76 | 0.91 | 0.71 |
| $X^2$ 기법 | 0.54 | 0.59 | 0.78 | 0.55 |
| DF 기법 | 0.46 | 0.53 | 0.43 | 0.47 |

그러나 Recall 값은 잘못 분류된 것에 대해서는 고려되지 않기 때문에 Recall 값이 높다고 해서 정확하게 분류되었다는 것을 의미하지는 않는다. 또한 Precision 값은 분류된 문제 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.

표 5. 제안 기법과 다른 기법과의 F-measure 값 비교

| F-measure | 운영체제 | 네트워크 | 인터넷 | 평 균 |
|---|---|---|---|---|
| 제안 기법 | 0.56 | 0.69 | 0.7 | 0.65 |
| $X^2$ 기법 | 0.52 | 0.64 | 0.58 | 0.58 |
| DF 기법 | 0.5 | 0.57 | 0.5 | 0.52 |

〈표 5〉의 실험 결과, 제안 기법은 모든 분야에서 높은 F-Measure 값을 가지며, 분야별 평균 F-measure 값은 0.65이다. $X^2$ 기법은 제안 기법보다는 낮지만 DF 분야보다는 높은 F-measure 값을 가지며, 분야별 F-measure 값은 0.58이다. DF 기법은 제안 기법과 $X^2$ 기법보다 전 분야에서 낮은 값을 가진다. 분야별 평균 F-measure 값은 0.52이다. 제안 기법은 분야별 평균 F-measure 값이 0.65로 다른 기법에 비해 가장 우수한 성능을 보였다.

## V. 결론

본 논문에서는 대량의 문제를 자동으로 분류하기 위하여 비감독 학습 기법에 의해 카테고리별 키워드를 구성하기 위한 방법을 제안하였다. 제안된 방법에서는 사전에 문제를 분류하지 않고 키워드를 추출하기 위하여 데이터마이닝 기법 중의 하나인 연관 규칙 탐사 알고리즘을 이용하였다. 먼저, 각 카테고리를 대표하는 핵심 키워드를 선정하고, 연관 규칙 탐사 알고리즘을 적용하여 각 핵심 키워드와 관련된 용어 집합을 추출한다. 추출된 용어 집합에서 상위 50개의 용어들로 연관 용어 집합 $K_{si}=\{T_1, T_2, T_3, \ldots, T_n\}$을 구성한다. 두 번째 단계에서는 연관 용어 집합 $K_{si}$의 각 원소 $T_i(1\langle i\langle n)$에 대해 최소 지지도 60% 이상의 상위 20개의 연관 용어를 추출하여 키워드로 구성하였다.

제안된 기법의 성능을 검증하기 위하여 컴퓨터 관련 분야를 대상으로 분류 실험을 하였다. 키워드 추출을 위한 학습용 문제는 컴퓨터 관련 서적에 발표한 문제를 500개를 사용하였고, 분류 실험에서는 키워드 추출 과정에서 사용하지 않은 분야별 30개의 문제를 사용하였다. 대표적인 감독 학습 기법인 $X^2$ 기법, DF 기법과의 비교 실험을 통하여 제안 기법의 성능을 평가하였다. 실험 결과 감독 학습 기법의 키워드 추출기법 중에서 우수하다고 알려진 $X^2$ 기법과 DF 기법보다 우수한 분류 성능을 보였다.

## 참고문헌


[1] 진 훈, 김인철, "문제 분류를 위한 특징 선택," 2001 봄 학술발표논문집, 한국정보과학회, 제28권, 제1호 pp. 20권 제1호, pp. 769-772, 1993.

[2] 홍진혁, 류중원, 조성배, "실세계의 FAQ 메일 자동분류를 위한 문제 특징 추출 방법의 성능 비교," 2001 봄 학술발표논문집, 한국정보과학회, 제28권, 제1호, pp. 271-273, 2001.

[3] 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.

[4] 정영미, 정보검색론, 구미무역(주) 출판부, 1993.

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

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

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

[8] 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.