# 범주 대표어의 가중치 계산 방식에 의한 자동 문서 분류 시스템

이경찬[0], 강승식
국민대학교 컴퓨터학부, 첨단정보기술연구센터
{ayin, sskang}@cs.kookmin.ac.kr

# Automatic Document Classification by Term-Weighting Method

Kyung-Chan Lee, Seung-Shik Kang
School of Computer Science, Kookmin University and Advanced Information Technology Research Center

**요 약**

자동 문서 분류는 범주 특성 벡터와 입력 문서 벡터의 유사도 비교에 의해 가장 유사한 범주를 선택하는 방법이다. 문서 분류 시스템을 구현하기 위하여 각 범주의 특성 벡터를 정보 검색 시스템의 역파일 형태로 구축하였으며, 용어 가중치를 계산하는 방법을 달리하여 문서 분류 시스템의 정확도를 실험하였다. 실험 문서는 일간지의 신문 기사들을 무작위로 추출한 문서 집합을 대상으로 하였으며, 정보 검색 모델에서 보편적으로 사용되는 TF-IDF 방식이 변형된 방식에 비해 더 나은 성능을 보였다.

## 1. 서론

정보 자료의 양이 많지 않을 때는 문서를 관리하는 작업을 수동으로 처리할 수 있었으나, 월드와이드 웹의 등장으로 인한 웹 문서의 폭발적인 증가와 더불어 대부분의 자료들이 디지털화 됨에 따라 문서를 보다 효율적으로 관리하기 위한 문서 분류의 필요성이 증가하고 있다[1,2,4]. 많은 문서들을 효율적으로 관리하여 사용자 요구에 적합한 정보를 검색하거나 정보 제공 서비스를 하기 위해서는 문서들을 유형에 따라 분류해야 한다. 그런데 대량의 문서들을 수동으로 분류하는 것은 시간적, 경제적으로 매우 비효율적이므로 자동 분류에 대한 연구가 활발하게 진행 중이다. 자동 문서 분류 시스템은 입력 문서에 한 개 또는 그 이상의 범주를 자동으로 할당해 준다.

문서 분류 시스템은 각 범주의 특징을 표현하는 범주 특성 벡터와 입력 문서의 특징을 표현하는 문서 벡터 사이의 유사도를 계산하여 유사도 값에 따라 입력 문서에 범주를 할당한다[5,6,8,12]. 자동 문서 분류 시스템은 문서에서 출현한 용어들에 대해 각 범주를 판별하는데 기여하는 정도에 따라 가중치를 부여한다. 입력 문서에 대해서도 문서에 출현한 용어들을 추출하고 해당 문서에 대한 가중치를 계산하여 문서 벡터를 구한다. 입력 문서 벡터와 각 범주들의 특성 벡터 사이의 유사도 계산에 의해 입력 문서의 범주를 결정한다.

문서 분류 시스템은 각 문서의 특징 벡터를 학습하는 방법과 입력 문서의 벡터를 계산하는 방법에 따라 그 성능에 차이가 있다. 문서 분류 방법은 확률적인 방법으로 범주 벡터를 학습하는 방법을 중심으로 연구되어 왔다. 범주 벡터의 학습 방법으로는 나이브 베이스 분류자(Naive Bayes's classifier), 결정 트리(decision tree), kNN(k-Nearest Neighbor), SVM(Support Vector Machine), 신경망 모델(Neural Network Model) 등이 있다[3,7,10]. 이러한 학습 알고리즘들은 그 성능에 약간의 차이가 있으며, 동일한 조건하에서 각 알고리즘의 성능을 평가했을 때 SVM의 성능이 가장 좋은 것으로 알려져 있다[13,14].

자동 문서 분류는 범주와 입력 문서간의 유사도의 계산에 의한 순서화로 정의할 수 있으며, 이는 정보 검색 모델과 매우 유사하다. 본 논문에서는 일반적인 정보검색 시스템을 구현할 때 사용되는 역파일 구조를 이용하여 문서 분류 시스템을 구현한다. 이 때, 범주에 대한 범주 대표어의 용어 가중치를 부여하는 방법을 변화시켜서 범주 판정의 정확도를 비교한다. 입력 문서와 범주 특성 벡터 사이의 유사도 계산은 벡터 모델에서 사용하는 코사인 계산식을 사용한다.

범주 대표어의 가중치를 계산하는 방식으로는 범주별 용어의 출현 확률을 이용하는 방법, 역 범주 빈도를 이용하는 방법, 그리고 용어의 출현 확률 대신에 정규화된 용어 빈도를 이용하는 방법을 적용하여 각 방법에 대한 문서 분류 시스템의 성능을 실험한다.

## 2. 문서 분류 시스템의 구조

### 2.1 문서 범주의 분류

문서 범주는 웹 검색 엔진의 디렉토리 서비스 메뉴를 참조하여 구분하였으며, 표 1과 같이 구분된 12개의 범주와 이를 세분화한 하위 범주들로 구성된다.

표 1. 최상위 범주의 분류

| | | |
|---|---|---|
| 가정, 여성 | 건강, 의학 | 게임 |
| 교육, 참고자료 | 레크리에이션 | 북한 |
| 비즈니스, 경제 | 쇼핑 | 스포츠 |
| 엔터테인먼트 | 컴퓨터, 인터넷 | 학문, 과학 |

상위 12개의 범주를 시작으로 이를 세분화하여 불필요하거나 중복된다고 판단되는 범주는 임의로 제거하였으며, 최종적으로 총 61개의 문서 범주로 분류하였다. 상위 범주로부터 세부 범주들로 세분화한 예는 다음과 같다.

예) 스포츠 => 축구, 야구
건강, 의학 => 의학, 전통의학, 증상, 질병
비즈니스, 경제 => 무역, 부동산, 투자, 금융, 재태크

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

2.2 범주 대표어 추출

각 범주의 특성을 기술하는 용어(범주 대표어)는 수작업으로 구축하거나 혹은 학습 알고리즘을 적용하여 학습한다. 본 논문에서는 문서 범주를 웹 검색 사이트의 디렉토리 분류 체계의 일부를 실험 대상으로 선정하였다. 각 디렉토리 범주들을 기술하는 문서 내용으로부터 범주를 기술하는 용어들을 추출하였다.

각 디렉토리 서비스의 해당 범주에 관한 설명문 및 범주를 기술하는 문자열들을 해당 범주의 대표 문자열들이라고 간주하여 이를 텍스트 문서로 저장하고, 형태소 분석기를 이용하여 명사들을 추출한다. 각 용어의 가중치는 해당 범주에 출현한 용어의 출현 빈도 및 범주 빈도에 따라 계산한다.

2.3 자동 문서 분류 시스템

① 역파일 생성

문서 분류 시스템의 구조는 정보검색 시스템의 구조와 유사하다. 범주별로 추출된 용어와 가중치 쌍을 용어별로 재구성하여 그림 1과 같이 역파일을 생성한다.

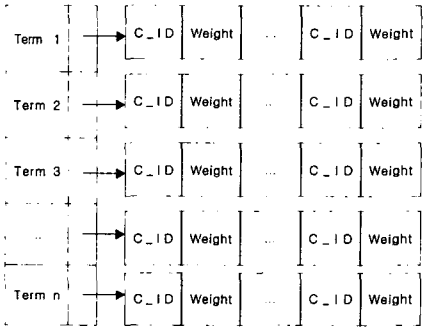


그림 1. 역파일 구조

그림 1의 역파일은 각 용어들에 대하여 이 용어가 범주 결정에 작용하는 기여도를 나타내는 가중치를 저장하고 있다. 각 범주를 기술하는 용어의 선정 및 가중치 계산의 신뢰도가 높다고 가정할 때, '박찬호'라는 용어는 '야구' 범주에 대한 가중치가 가장 높고, '스포츠' 범주에 대한 가중치가 두 번째로 높을 것이다.

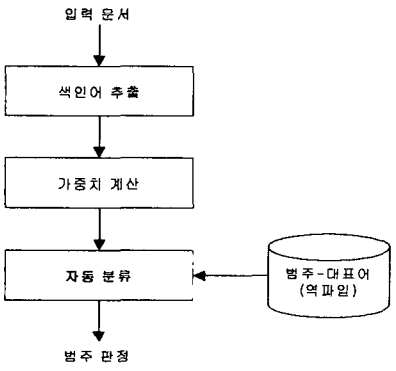


그림 2. 문서 분류 과정

② 범주 결정

범주를 기술하는데 사용된 용어와 그 가중치가 계산되어 역파일로 구성되었으면, 이를 바탕으로 임의의 입력 문서에 대한 범주를 판정할 수 있다. 예를 들어, 어떤 문서의 주제어가 '박찬호'라고 가정할 때 이 문서의 범주는 '야구'일 확률이 가장 높고, 다음으로 '스포츠'가 될 것이다. 그런데 입력 문서의 주제어가 무엇인지를 자동으로 판별하기는 어려우므로 입력 문서로부터 색인어를 추출하여 출현 빈도와 범주 빈도에 의해 <색인어, 가중치>쌍을 구하여 코사인 계수에 의해 문서와 범주간의 유사도를 계산한다. 임의의 문서에 대해 범주를 판정하는 과정은 그림 2와 같다.

## 3. 범주 대표어의 가중치 계산 방법

자동 문서 분류 시스템은 (1) 범주 대표어의 가중치 계산 방법, (2) 입력 문서의 용어 가중치 계산 방법, (3) 각 범주와 입력 문서 간의 유사도 계산 방법에 따라 그 성능이 달라질 수 있다. 본 논문에서는 유사도 계산은 코사인 계수를 이용하고, 범주 대표어 및 입력 문서의 용어 가중치 계산 방법을 다르게 하여 실험 결과를 비교하였다.

$$sim(d,c) = \frac{\sum_t (W_{t,d} \times W_{t,c})}{\sqrt{\sum_t w_{t,d}^2} \times \sqrt{\sum_t w_{t,c}^2}}$$

3.1 방법-1

범주 C에 대한 용어 t의 가중치 $W_{t,c}$는 용어 t가 범주 C를 대표하는 정도를 반영하기 위하여 용어 t의 빈도수를 범주 C를 기술하는 모든 용어들의 빈도수 합으로 나눈 값으로 계산한다.

$$W_{t,c} = \frac{freq(t)}{\sum_i freq(t_i)}$$

입력 문서에 출현한 용어의 가중치 $W_{t,d}$는 TF-IDF에 의해 아래와 같이 계산한다. $W_{t,d}$를 계산할 때 IDF에 log를 취하지 않은 이유는 문서 분류의 경우에 주제어와 비주제어를 명확히 구분하기 위한 목적으로 특정 범주에 대한 기여도가 높은 용어의 가중치를 높이기 위해서이다.

$$W_{t,d} = freq(t) \times \frac{N}{n_t}$$

$N$ : 총 문서 범주의 개수

$n_t$ : 용어 t가 출현한 문서 범주의 개수

3.2 방법-2

이 방법은 일반적으로 벡터 공간 모델의 정보 검색 시스템에서 사용되는 방법이다. 다만, 입력 문서의 용어 가중치를 질의어 가중치 계산식이 아니라 일반 문서에 대한 가중치로 계산한 것만 다르다. 각 범주 대표어 및 입력 문서에 출현한 용어의 가중치는 아래와 같이 동일한 방법으로 계산한다.

$$W_{t,c} = freq(t) \times \log \frac{N}{n_t}$$

$$W_{t,d} = freq(t) \times \log \frac{N}{n_t}$$

3.3 방법-3

이 방법은 용어의 출현 빈도를 각 문서에 출현한 최대 빈도수로 나누어 정규화하고 역문헌빈도를 이용하여 가중치를 계산한다. 각 범주 대표어와 입력 문서의 용어 가중치 계산은 아래와 같다.

$$W_{t,c} = \frac{freq(t)}{\max TF} \times \log \frac{N}{n_t}$$

$$W_{t,d} = \frac{freq(t)}{\max TF} \times \log \frac{N}{n_t}$$

## 4. 실험 및 평가

실험 대상의 문서집합은 일간 스포츠, 스포츠 서울, 스포츠 조선, 중앙일보, 매일경제, 동아일보 등 웹에서 서비스하고 있는 여러 신문사의 신문 기사 중에서 본 논문에서 구현된 범주에 속하는 기사들을 무작위로 추출하였다. 실험에 사용된 문서의 개수는 50개이다. 평가 방법은 결과 우선 순위 중 1~3 순위까지의 정확률을 계산하여 평가하였다. 세 가지의 방법을 비교하여 나온 결과는 표 2와 같다.

표 2. 문서 분류 실험

| 순위 / 방법 | 1순위 | 2순위 | 3순위 |
|---|---|---|---|
| 방법 1 | 71.0% | 76.3% | 78.9% |
| 방법 2 | 73.6% | 84.2% | 86.8% |
| 방법 3 | 73.6% | 81.5% | 81.5% |

실험 결과에 의하면 1순위의 정확률은 역문헌 빈도에 log를 취한 방법 2, 3이 용어의 총 빈도수를 이용한 계산 방법보다 1순위 판정 결과에서 좀 더 우수한 것으로 나타났다. 그리고 2, 3순위를 비교해 보면 관련된 범주의 유사도면에서 방법 2가 최대 빈도 용어로 정규화시킨 방법 3보다 관련 범주에 대한 판정 결과에 비하여 좀 더 나은 결과를 나타낸 것을 알 수 있다.

## 5. 결론

본 논문에서는 문서 분류의 범주 판정 과정을 정보 검색 시스템의 구조와 같이 역파일 형태로 구성하고, 문서와 범주의 용어 가중치를 계산하기 위하여 세 가지 방법에 대해 실험하였다. 역파일 형태로 대표어를 저장하는 방법은 속도면에서 일반 문서당 전체 범주의 비교 방식보다는 나아졌으며, 해당 용어에 대한 범주의 총 빈도수를 이용한 방법 1보다는 해당 용어에 대한 범주의 총 빈도수의 log(ICF)값을 이용한 방법 2,3이 1순위 판정에서 좋은 결과를 보였다. 이는 2순위 3순위로 갈수록 관련된 범주에 대한 판정이 더 나아짐을 볼 수 있다. 방법 2와 해당 용어에 대한 범주의 최대 빈도를 이용한 정규화 방법 3도 1순위에서는 차이를 보이고 있지 않지만, 2순위 3순위에서는 방법 2가 좀더 나은 결과를 보이고 있다.

향후 문서 분류 시스템의 성능 개선을 위하여 학습을 통한 용어별 가중치를 부여하여 성능을 개선하는 방법과 통계적인 구문 패턴을 이용하여 자연어 처리에 접목시킬 수 있는 다양한 연구 방법이 모색되어야 할 것이다.

## 참고문헌


[1] 강원석, 강현규, 김영섬, "가중치 부여 휴리스틱을 이용한 개념 기반 문서분류기 TAXON의 개선", 한국정보과학회 가을 학술발표 논문집, Vol.25, No.2, pp.153-155, 1998.

[2] 강원석, 황도삼, 최기선, "의미의 상하위 정보를 이용한 웹문서 분류 시스템", 제11회 한글 및 한국어 정보처리 학술발표 논문집, pp.36-39, 1999.

[3] 권오욱, 이종혁, 이근배, "Nearest Neighbor 방법을 이용한 문서 범주화에서 범주 자질의 평가", 제9회 한글 및 한국어 정보처리 학술발표 논문집, pp.7-14, 1997.

[4] 오효정, 임정묵, 이만호, 맹성현, "점진적으로 계산되는 분류 정보와 링크 정보를 이용한 하이퍼텍스트 문서 분류 모델", 제11회 한글 및 한국어 정보처리 학술발표 논문집, pp.89-96, 1999.

[5] 정성화, 이종혁, "문서 구조 정보에 기반한 웹 페이지 범주화 모델", 제10회 한글 및 한국어 정보처리 학술발표 논문집, pp.91-96, 1998.

[6] 조광제, 김준태, "역 카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동 분류", 한국정보과학회 봄 학술발표 논문집, Vol.26, No.1, pp.507-510, 1997.

[7] 조태호, "신경망 또는 k-NN에 의한 신문기사 분류와 그의 성능 비교", 한국정보과학회 가을 학술발표 논문집, Vol.25, No.2, pp.363-365, 1998.

[8] 최동시, 정경택, "카테고리와 키워드의 밀접성 정보에 의한 문서 자동 분류 시스템 설계 및 구현", 한국정보과학회 가을 학술발표 논문집, Vol.22, No.2, pp.639-642, 1995.

[9] 하얀, 최봉진, 김용성, 김순기, "2단계 필터링을 이용한 문서 선별 및 순위", 한국정보과학회 봄 학술발표 논문집, Vol.26, No.1, pp.315-317, 1999.

[10] 허준회, 고수정, 김태용, 최준혁, 이정현, "문서의 주제어별 가중치와 말뭉치를 이용한 한국어 문서의 자동 분류: 베이지안 분류자", 한국정보과학회 가을 학술발표 논문집, Vol.26, No.2, pp.154-156, 1999.

[11] R. Hoch, "Using IR Techniques for Text Classification in Document Analysis", SIGIR'94, 1994.

[12] Thorsten Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization", International Conference in Machine Learning, 1996.

[13] David D. Lewis, "Evaluating Text Categorization", Proceedings of Speech and Natural Language Workshop, 1991.

[14] Yiming Yang and Xin Liu, "A Re-examination of Text Categorization Methods", SIGIR'99, pp.42-49, 1999.