# 블로그 포스트의 자동 분류 시스템

조희선O, 김수아, 이현아
금오공과대학교 컴퓨터소프트웨어공학과
shinhwa3528@naver.com, sa4956@nate.com, halee@kumoh.ac.kr

# Automatic Classification of Blog Posts

Hee-Sun JhoO, Su-Ah Kim, Hyun-Ah Lee
Kumoh National Institute of Technology, Department of Computer software Engineering

요 약

편리한 블로그 사용과 블로그에서의 정보 탐색을 위해서는 내용에 기반한 분류가 필요하다. 대부분의 블로그 사이트에서는 내용 기반 분류를 제공하고 있으나, 블로거들은 자신이 작성한 블로그에 대한 수동 분류를 입력하지 않는 경우가 많다. 본 논문에서는 분류가 제공되는 블로그 사이트에서 각 분류별 문서를 수집하고, 어휘빈도와 문서빈도, 분류별 빈도를 활용하여 문서 내 어휘의 자질 가중치를 부여하고, 다양한 학습기를 이용하여 분류 모델을 생성한 뒤 블로그의 특성에 적합한 자질 추출 알고리즘과 분류 알고리즘을 찾아낸다. 실험에서는 본 논문에서 고안한 CTF-IECDF와 나이브 베이즈 멀티노미얼로 조합한 분류 모델이 75.40%의 분류 정확률을 보였다.

주제어: 블로그, 자동 분류, 자질 추출

## 1. 서론

정보 개방과 사용자 참여를 가능하게 하는 웹 2.0의 시대에 블로그는 웹 2.0의 특성을 살린 1인 미디어로 부상하고 있다. 최근에는 많은 이용자들이 블로그를 개인적 기록이나 사회 참여의 도구로 사용할 뿐만 아니라, 취미나 관심 분야의 정보 획득 및 공유의 목적으로 많이 사용하고 있다[1]. 블로그 사용자가 늘어남에 따라서 블로그 사이트에서는 주제와 목적에 맞게 블로그를 분류하는 서비스를 제공하고 있다. 블로그 글이 주제에 맞게 분류가 되어있다면, 블로그를 통해 정보를 찾으려는 사용자는 많은 시간과 노력을 감소할 수 있다. 하지만, 다양한 분야(예를 들어 영화, IT, 주식)의 블로그 포스트는 하나의 분류로 대응시키기 적합하지 않을 수 있고, 블로그 포스트를 작성할 때 마다 포스트의 분류를 결정하게 하는 네이버 블로그의 경우 대부분의 블로거들이 기본 분류를 선택하는 양상을 보이고 있다.

본 논문에서는 네이버 블로그에서 주제별 분류가 등록된 포스트들을 수집하고, 이를 학습 데이터로 사용하여 자동으로 포스트의 주제별 분류를 추천하기 위한 시스템을 제안한다. 시스템에서는 분류의 특성을 추출하기 위하여 네가지의 방식을 제시한다. 분류에서는 다양한 학습방법을 적용해 본다. 각각의 결과에 대한 분석과 평가를 통해 블로그 글의 특성에 맞는 특성 추출 알고리즘과 분류 알고리즘으로 블로그 자동 분류 시스템을 구현한다.

## 2. 관련 연구

문서에 대한 자동 분류 연구에서는 문서에 포함된 단어의 특성을 주로 이용한다. 국내 연구 중에서 [2]는 웹 문서에 대한 텍스트 자동 분류를 위한 특성 추출 기법을 제안한다. 학습 문서 벡터는 웹 디렉터리 내의 문서로부터 추출된 단어 및 관련 문서를 기반으로 구성하였으며, 학습 문서 구성 후 SVM 학습기를 통해 모델을 구성하여 문서 분류를 수행하였다. 이 연구에서는 클래스간의 연관성이 높은 경우 낮은 성능을 보이는 문제점이 있었고 추가적인 학습 문서의 정제가 필요하였다.

[3]에서는 한글 웹 문서에 사용된 한글 형태소 및 키워드의 빈도에 기초하여 문서의 특성을 추출하는 방법을 제시하고, 이를 기초로 비구조적인 문서의 주제를 자동으로 분류하는 방법을 제시하였다. 의사 결정 트리, 신경망 모델 및 SVM 방법을 사용하였으며, 주제간 구분이 명확할수록 정확도는 증가함을 보였다.

## 3. 블로그 포스트의 자동 분류 시스템

### 3.1 블로그 글 수집

국내의 대표적인 블로그 사이트인 네이버, 다음, 티스토리에서 수집된 블로그 문서를 기준으로 적합한 분류를 조사한 [4]에서 네이버 블로그는 장르 분류 일치도에서도 높은 결과를 보였다. 본 연구에서는 네이버의 장르, 즉 분류가 부착된 블로그를 이용하여 학습에 사용한다.

네이버 블로그에는 총 30개의 분류가 존재한다. 하지만, 각 분류들은 글을 수집하여 분석하고 학습 데이터로 쓸 만큼 충분한 양의 글이 올라오지 않는 경우가 있었다. 또한, 분류의 주제에 맞지 않는 광고성 글이 존재하는 경우도 존재했다. 이러한 문제를 보완하기 위해, 네이버의 30개의 분류 중 일부를 제거하거나 병합하여 아래 [그림 1]과 같은 16개 분류를 얻고, 이를 이용하여 자동 분류를 수행한다.

| | | | |
|---|---|---|---|
| 문학-책 | 육아-결혼 | 여행 | 요리-레시피/맛집 |
| 영화 | 건강-의학 | 스포츠 | 미술-디자인/공연-전시 |
| 음악 | 패션-미용 | 자동차 | 사회-정치 |
| 인테리어 | IT-컴퓨터 | 게임 | 애완-반려동물 |

[그림 1] 블로그 분류를 위한 16개 분류

### 3.2 단어별 주제 분별 점수 계산

문서 분류를 위한 문서 특성은 제목과 본문에서 사용되는 명사에서 추출한다. 단어들을 정확히 추출하기 위해서 조사나 어미의 구분이 필요하다. 한국어 형태소 분석기를 이용하여 문서 내의 단어를 추출한 뒤, 각 단어의 빈도를 분석한다.

각 문서의 단어 점수는 문서에서 발생한 단어의 빈도 TF에 단어의 주제 분별 점수를 곱하여 얻는다. 이때 문서의 크기가 커지면 문서에서 발생한 단어의 절대 빈도도 커지므로 정규화가 필요하다. 문서 $D$에서 발생한 단어 $w_i$의 빈도수에 문서 $D$의 총 단어수로 나누어 정규화한 $TF_D(w_i)$를 식(1)로 구한다.

$$TF_D(w_i) = \frac{\text{문서 } D\text{에서 단어 } w_i\text{의 빈도}}{\text{문서 } D\text{의 총 단어 빈도}} \quad (1)$$

정규화한 단어 빈도 $TF_D(w_i)$와 아래에 설명할 IDF, CTF-IECDF, CDF-IDF, CDF-IECDF의 각 4가지 방식으로 구한 주제 분별 점수를 곱하여 문서 내 단어의 주제 분별력 점수를 생성한다. 아래에서는 각 주제별 점수 계산 방식을 설명한다.

#### 3.2.1 IDF

IDF는 문서 빈도를 이용하여 단어의 희소성이나 정보성을 표현하는 통계적 방법으로, 첫 번째 방식에서는 IDF로 단어 $w_i$의 주제 분별 점수를 구한다. $IDF(w_i)$는 전체 문서수에서 단어 $w_i$가 발생한 문서의 빈도를 나눈 값에 log를 취한 값으로서 식으로 표현하면 다음 식 (2)과 같다.

$$IDF(w_i) = \log \frac{\text{전체문서의수}}{\text{단어 } w_i\text{가존재하는문서의수}} \quad (2)$$

#### 3.2.2 CTF-IECDF

CTF-IECDF는 분류 $c$에서의 단어 누적 빈도와 $c$를 제외한 분류에서의 IDF값을 통해, 분류 $c$에서의 단어 중요도를 계산한다. CTF는 문서내 빈도인 TF를 분류 내 빈도로 확장시킨 값이며, IECDF는 특정 분류를 제외한 IDF를 의미한다. 이 방식에서는 단어 $w_i$의 누적빈도가 가장 큰 분류를 단어 $w_i$의 대표 분류로 보고, 해당당 분류 $C_{max}$를 수식 (3)으로 얻는다. $C_{max}$에서의 단어 $w_i$의 누적빈도 $CTF(w_i)$를 수식 (4)로 얻는다.

$$C_{\max} = \arg\max_c (c\text{에서의 단어} w_i\text{의 누적빈도}) \quad (3)$$

$$CTF(w_i) = C_{\max}\text{에서 단어} w_i\text{의 누적빈도} \quad (4)$$

수식 (5)는 IECDF를 구한다. 수식에서 $C_{others}$는 전체 분류에서 $C_{max}$를 제외한 모든 분류를 의미한다. 분류 $C_{others}$ 분류들에서의 단어 $w_i$의 IDF 값인 IECDF와 (4)의 CTF값을 곱한 후 루트를 취하여 단어 $w_i$의 주제 분별 점수를 구한다. CTF-IECDF의 값은 각각의 점수 차이가 크기 때문에 루트를 취하여 편차를 줄인다.

$$IECDF(w_i) = \log \frac{C_{others}\text{의 문서수}}{C_{others}\text{에서 } w_i\text{가 존재하는 문서수}} \quad (5)$$

$$CTF-IDCDF(w_i) = \sqrt{CTF(w_i) \times IECDF(w_i)} \quad (6)$$

#### 3.2.3 CDF-IDF

CDF-IDF에서 CDF는 단어의 $C_{max}$ 내 문서빈도를 의미한다. 수식(3)의 $C_{max}$를 이용하여, $C_{max}$에서의 문서빈도인 $CDF_i$를 수식 (7)으로 구한다.

$$CDF(w_i) = \frac{C_{\max}\text{에서 단어} w_i\text{가 발생한 문서수}}{\text{카테고리 } C_{\max}\text{의문서수}} \quad (7)$$

$CDF-IDF$에서는 식 (2)의 IDF와 식 (7)의 CDF를 곱하여 주제 분별 점수를 구한다. CDF를 통해 해당 단어가 대표 분류에서 폭넓게 사용될수록, IDF를 통해 해당 단어가 희소성이 높을수록 높은 점수를 얻는다.

$$CDF-IDF(w_i) = CDF(w_i) \times IDF(w_i) \quad (8)$$

#### 3.2.4 CDF-IECDF

CDF-IECDF는 앞의 방식에서 사용한 CDF값과 IECDF 값을 곱하여 주제 분별력을 계산한다.

CDF를 통해 해당 단어가 대표 분류에서 폭넓게 사용될수록 높은 점수를 얻을 수 있고, 특정 분류를 제외한 IDF인 IECDF를 사용하여 나머지 분류에서의 희소성이 높을수록 높은 점수를 얻는다.

$$CDF-IECDF(w_i) = CDF(w_i) \times IECDF(w_i) \quad (9)$$

### 3.3 분류 모델 생성

문서별로 주제 분별 점수가 구해지면 이를 이용하여 분류 모델을 생성한다. 먼저, 학습용 문서 집합을 이용하여 분류 모델을 생성하고 평가한다. 분류 모델 생성을 위하여 기존의 소프트웨어 WEKA 3.6.10에 구현된 컴플리먼트 나이브 베이즈, 나이브 베이즈 멀티노미얼, SVM 알고리즘을 사용하였다.

각 분류 알고리즘을 이용해 입력을 각 문서 별 단어에 대한 주제 분별 단어점수로 하고 출력을 주제 필드로 설정하여 각각의 분류 모델을 생성한다. 분류 모델의 검증은 생성과 마찬가지로 검증용 문서 집합을 이용하여 생성된 각 분류 모델의 정확도를 검증한다.

## 4. 실험 및 평가

본 논문에서 제안하는 네 가지의 특성 추출 방식과 3가지의 분류 학습기를 통한 분류 정확률 실험 및 평가한

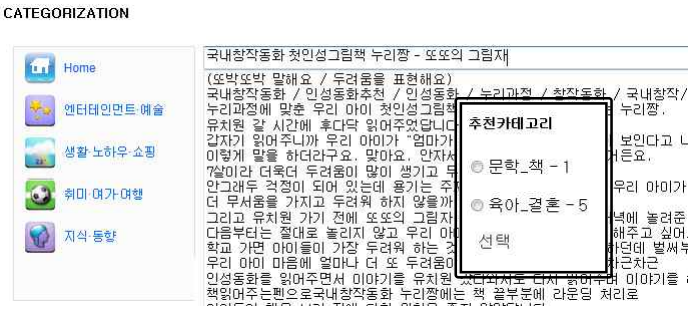
다. 학습 데이터와 테스트 데이터는 따로 수집하였다. 실험에서는 각 분류별 500개 총 8000개의 학습 데이터를 이용하였으며, 실험 데이터는 분류별 200개, 총 3200개를 사용한다. [표 1]은 결과를 보인다.

[표 1] 분류 학습기에 따른 분류 정확률

| Test data 200 | 컴플리먼트 나이브 베이즈 | 나이브 베이즈 멀티노미얼 | SVM |
|---|---|---|---|
| TF-IDF | 44.06% | 43.31% | 6.25% |
| CTF-IECDF | 73.26% | 75.40% | 69.06% |
| CDF-IDF | 70.01% | 58.64% | 25.96% |
| CDF-IECDF | 72.40% | 72.61% | 58.63% |

기존의 TF-IDF는 3가지 분류 학습기를 이용한 결과 모두에서 50%가 되지 않는 분류 정확률을 보였다. TF-IDF는 분류 정보가 반영되지 않고, 키워드의 단순 빈도와 단순한 문서에서의 IDF를 이용하여 문서를 분류하여 낮은 정확률을 보이는 것으로 분석되었다. 컴플리먼트 나이브 베이즈로 학습한 CTF-IECDF는 TF-IDF와 29.2%의 차이로 더 높은 분류 정확률을 보였다. CDF-IDF와 CDF-IECDF는 약 2% 차이로 전체적으로 비슷한 정확률을 보였다. 나이브 베이즈 멀티노말로 학습을 한 결과에서 CTF-IECDF가 3가지 방식의 실험 중 75.40%로 가장 높은 정확률을 보였고, 반대로 CDF-IDF는 가장 낮은 결과를 보였다. 전체적인 실험 결과는 CTF-IECDF가 다른 방식들보다 가장 높은 정확률을 보였다. 그리고 SVM으로 학습을 한 결과는 다른 분류 학습기를 사용 했을 때 보다 비교적 낮은 정확률을 보였다.

본 연구에서는 실험 결과에서 정확률이 높은 상위 6개를 이용하여 블로그 자동 분류 시스템을 구축하였다. [그림 2]는 실행 예를 보인다. 시스템에서 사용자가 포스트를 작성한 뒤 저장 버튼을 누르면 자동으로 추천 카테고리를 제시하는 방법으로 구동된다. 시스템에서는 6개의 분류 기법에서 얻어진 분류를 투표 방식(voting)을 적용하여, 가장 많이 추천된 분류부터 순서대로 사용자에게 제시된다. 실행 예에서 포스트는 육아와 관련한 책을 소개하는 글로서, [문학-책]과 [육아_결혼]의 분류를 추천하는 결과를 볼 수 있다.




[그림 2] 블로그 포스트를 시스템에 입력한 결과

## 5. 결론

문서 분류에 TF-IDF를 변형하여 새로운 알고리즘을 만드는 연구는 다양하게 존재한다. 하지만 기사와 같이 의도가 분명한 글에 관한 분류와 달리 다양한 개성을 가진 사용자들의 글이 존재하는 블로그 문서에 대한 연구는 드물다.

본 논문에서는 블로그 포스트를 자동으로 분류하기 위해 TF와 IDF를 분류로 확장시킨 특성 추출 알고리즘을 사용하여 정확률을 실험하였다. 제안된 방식에서 나이브 베이즈를 사용한 분류 모델이 비교적 높은 정확률을 보였고, IDF를 사용한 것 보다 IECDF를 사용했을 때 블로그 문서의 분류가 더 정확하게 수행되었다. TF보다는 분류로 확장한 CDF나 CTF를 사용하는 것이 더 정확한 결과를 보였다.

블로그 문서는 정형적인 텍스트가 아니기 때문에 오타나 신조어 등에 민감 할 수 있는데 이러한 점은 형태소 분석기의 성능이 향상되거나 고유 명사 사전 등을 구축하면 해결 할 수 있을 것이라고 기대한다. 그리고 문서 자동 분류에서 문서 필터링으로의 확장이 추후 연구가 될 수 있다.

## 참고문헌

[1] Young-Ju Kim, "A Study on the Blog as a Media : Focused on Media Functions and the Problems of the Blog", Korean journal of journalism & communication studies, Vol.50, No.2, 2006.

[2] 박단호, 최원식, 김홍조, 이석룡, "한글 형태소 및 키워드 분석에 기반한 웹 문서 분류", 정보처리학회 논문지(D), pp.263-270, 2012.

[3] 강윤희, 박용범, "SVM을 이용한 디렉토리 기반 기술 정보 문서 자동 분류시스템 설계", 전기전자학회 Vol.8.No.2, pp. 186-194, 2004.

[4] Hae Young Kim, "An Experimental Study On Semi-Supervised Classification of Blog Genres", MS Thesis, Yonsei University, 2009.