한국컴퓨터정보학회 동계학술대회 논문집 제23권 제1호 (2015. 1)

범주별 고유 정보를 고려한 블로그 포스트의 자동 분류

김수아°, 오성탁*, 이지형*
°금오공과대학교 컴퓨터소프트웨어공학과
*성균관대학교 전자전기컴퓨터공학과
e-mail:sa4956@kumoh.ac.kr°, {ohsm014, john}@skku.edu*

Automatic Classification of Blog Posts Considering Category-specific Information

Suah Kim°, Sungtak Oh*, Jee-Hyong Lee*†
°Dept. of Computer Software Engineering, Kumoh National of Institute of Technology
*Dept. of Electrical and Computer Engineering, Sungkyunkwan University

● 요 약 ●

많은 블로그 제공 사이트는 블로그 포스트 작성자에게 미리 정의된 범주 (category)에 따라 포스트의 주제에 대하여 범주를 선택할 수 있는 환경을 제공한다. 그러나 블로거들은 작성한 포스트의 범주를 매번 수동으로 선택해야 하는 불편함이 있다. 이러한 불편함의 해결을 위해 블로그 포스트를 자동으로 분류해주는 기능을 제공한다면 블로그의 활용성이 증가할 것이다. 기존의 블로그 문서 분류의 연구는 각 범주의 고유 정보를 반영하는 것에 한계가 있었다. 이러한 문제를 해결하기 위해, 본 논문에서는 범주별 고유 정보를 반영한 어휘 가중치를 제안한다. 어휘 가중치의 분석을 위하여 범주별로 블로그 문서를 수집하고, 수집한 문서에서 어휘의 빈도와 문서의 빈도, 범주별 어휘빈도 등을 고려하여 새로운 지표인 CTF, CDF, IECDF를 개발하였다. 이러한 지표를 기반으로 기존의 Naive Bayes 알고리즘으로 학습하여, 블로그 포스트를 자동으로 분류하였다. 실험에서는 본 논문에서 제안한 가중치 방법인 TF-CTF-CDF-IECDF를 사용한 분류가 가장 높은 성능을 보였다.

키워드: 블로그(blog), 자동 분류(automatic classification), 어휘 가중치(term weighting), 가중치 결합(weighting combination)

I. 서 론

사용자의 참여를 중심으로 하는 인터넷 환경인 웹 2.0시대에서 1인 미디어로써 블로그가 떠오르고 있다. 개인적인 글이나 사회적 참여로 사용되던 블로그는 최근 여러 분야의 정보 공유 및 획득의 목적으로 사용되는 경우가 많아졌다 [1].

블로그 이용자들이 늘어나면서 블로그 서비스를 제공하는 여러 사이트에서는 주제와 정보에 맞게 블로그 포스트를 분류함으로써, 이용자가 정보 획득에 소요하는 시간과 비용을 줄일 수 있도록 하고 있다. 포스트의 분류를 위해 네이버 블로그는 포스트 작성 시 사용자가 내용에 맞는 분류를 직접 선택할 수 있도록 서비스를 제공한다. 하지만 블로그 작성자들은 분류선택의 애매함과 번거로움 때문에 분류를 선택하지 않는 경우가 많이 발생한다. 이러한 경우에는 해당 정보에 관심을 갖고 있는 사용자들이 그 정보를 찾아보기 어렵다는 문제가 발생하게 된다. 그렇기 때문에 블로그 사이트에서 블로거가 작성한 포스트를 자동으로 내용에 맞게 범주를 분류 해주는 기능을 제공한다면 사용자들의 블로그 활용성이 증가 할 수 있을 것이다.

지금까지 문서 내에서 중요한 단어를 뽑아내어 문서 분류를 하는 연구는 다소 진행되어 왔다. 그러나 기존의 블로그 문서 분류의 연구 [2]는 범주의 정보를 반영하는 것에 한계가 있었다. 이러한 문제를 해결하기 위해, 본 논문에서는 범주별로 분류된 네이버 블로그 포스트에서 특성을 추출하여, 범주의 고유 정보를 반영한 포스트의 자동 분류 방법을 제안한다. 제안방법은 기존의 TF-IDF를 확장해서, 범주의 고유 정보를 반영하는 CTF, CDF, IECDF 등 블로그 분류에 적합한 지표를 개발하였다. 그리고 기존의 여러 분류 알고리즘을 사용하여 블로그 포스트의 특성에 맞는 범주를 결정하였다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 관련 연구에 대하여 설명하고, 3장에서는 제안기법에 대해 설명한다. 4장에서는 실험 및 평가에 대해서 설명하고, 5장에서는 결론에 대해 설명한다.

II. 블로그 포스트의 자동분류

본 논문에서 제안하는 방법은 그림 1의 과정을 거쳐 진행된다. 먼저 블로그 사이트에서 범주별로 블로그 문서들을 수집하고, 형태소 분석기 [3]를 사용하여 명사를 추출한다. 그 후 추출된 명사 집합에

다양한 어휘 가중치를 적용하여 각 명사를 점수화 한다. 이를 통해 생성한 명사 점수표를 기반으로 학습 데이터에 기존의 분류 알고리즘을 사용하여 모델을 생성하고, 테스트 데이터로 모델을 평가한다.

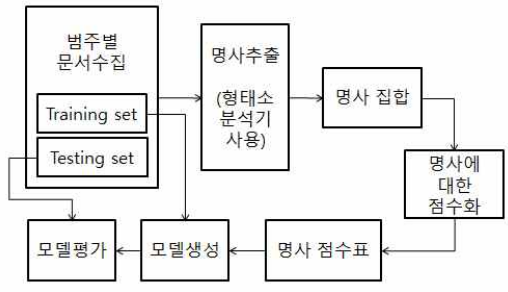

그림 1. 문서 특성 추출 및 분류모델 생성 과정
Fig 1. Flows of Document Feature Extraction and Classification Model Generation

1. 블로그 포스트 수집

국내에서 주로 이용하는 블로그 서비스 제공 사이트인 네이버, 티스토리, 다음에서 수집된 블로그 포스트를 기준으로 하여 분류의 적합성을 조사한 연구 [4]에서 네이버는 분류 일치도가 높은 결과로 나타났다. 본 논문에서는 주제 분류가 된 네이버 블로그 포스트를 학습 데이터로 사용한다.

네이버 블로그에서는 그림 2의 좌측에 보이는 30개의 범주가 있다. 이 중에서 일부는 부족한 양의 포스트로 학습데이터가 부족하거나 주제에 맞지 않는 광고 글이 많은 범주가 존재하여, 해당 범주를 제외하거나 병합하였다. 그 결과로 그림 2의 우측에 보이는 16개의 범주를 정하여 자동분류를 수행한다.

엔터테인먼트/예술	생활/노하우/쇼핑	취미/여가/여행	지식/동향
문학/책 영화 미술/디자인 공연/전시 음악 드라마 스타/연예인 만화/애니 방송	일상/생각 육아/결혼 애완/반려동물 좋은 글/이미지 패션/미용 인테리어/DIY 요리/레시피 상품리뷰	게임 스포츠 사진 자동차 취미 국내여행 세계여행 맛집	IT/컴퓨터 사회/정치 건강/의학 비즈니스/경제 어학/외국어 교육/학문

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

그림 2. 기존 네이버 블로그의 범주를 단순화 한 새로운 블로그 범주
Fig 2. Naver Blog Categories and Our Simplified Blog Categories

2. 단어별 주제 분별 점수 계산

문서 특성 추출을 위해 한국어 형태소 분석기를 사용하여 문서의 제목과 본문에서 발생하는 명사를 추출한 뒤, 각 단어를 분석한다. 본 논문에서는 문서의 특성을 파악하기 위해서 TF와 IDF 방법을 확장하여, 범주 내에서의 단어 빈도나 문서의 빈도를 반영하였다. 그 결과로 범주를 고려한 가중치인 CTF, CDF, IECDF를 제안한다. 아래에서는 각각의 가중치에 대하여 설명한다.

2.1 CTF

주제의 특성을 잘 나타내는 단어는 특정 분류, 즉 대표 분류에서 빈번하게 발생할 것으로 기대될 수 있다. 이를 분류 내 단어 빈도인 CTF(Category Term Frequency)로 수치화하여 나타낼 수 있다. 본 논문에서는 식 (1)과 식 (2)로 CTF를 구한다. 식 (1)은 특정 단어의 누적 빈도가 가장 높은 분류를 대표 분류로 보아, 단어 w_i의 대표 분류인 $Max\,C_{w_i}$를 구한다. 여기서 얻은 대표 분류에서 단어 w_i의 누적빈도인 $CTF(w_i)$를 식 (2)로 얻는다. 본 시스템에서는 각 분류별로 문서수를 동일하게 수집하여, CTF의 정규화가 이루어지도록 한다.

$$Max\,C_{w_i} = \arg\max_{C} \sum_{D \in C} freq(w_i, D) \qquad (1)$$

$$CTF(w_i) = \log \sum_{D \in Max\,C_{w_i}} freq(w_i, D) \qquad (2)$$

2.2. CDF

주제의 특성을 잘 나타내는 단어는 특정 분류에서 많은 문서에 나타날 것으로 기대할 수 있다. 이는 대표 분류에서의 문서 빈도 CDF(Category Document Frequency)로 수치화하여 나타낼 수 있다. 식 (3)는 식 (1)에서 구한 단어 w_i의 대표 분류인 $Max\,C_{w_i}$에서의 문서빈도인 $CDF(w_i)$를 구한다. 수식에서는 대표 분류에서 단어 w_i가 발생한 문서 수 $|D_{w_i, Max\,C_{w_i}}|$를 대표 분류의 전체 문서 개수인 $|D_{all, Max\,C_{w_i}}|$로 나눠, 단어 w_i가 분류의 대표성을 얼마나 나타내는지를 수치화한다.

$$CDF(w_i) = \frac{|D_{w_i, Max\,C_{w_i}}|}{|D_{all, Max\,C_{w_i}}|} \qquad (3)$$

2.3 IECDF

어떤 단어가 대표 분류를 제외한 분류의 문서에서 덜 발생할수록 주제의 특성을 잘 나타낸다고 기대할 수 있다. 이는 대표 분류를 제외한 분류에서의 IDF인 IECDF(Inversed, Excepted Category's, Document Frequency)로 수치화할 수 있다. 식 (4)은 IECDF를 구한다. 수식에서 $\overline{Max\,C_{w_i}}$는 전체 분류에서 $Max\,C_{w_i}$를 제외한 모든 분류를 나타낸다. 대표 분류를 제외한 분류에 있는 모든 문서 수를 단어 w_i가 발생한 문서수를 나누어, 대표 분류 이외의 문서에서 덜 나타날수록 더 높은 단어 가중치를 얻는다.

$$IECDF(w_i) = \log \frac{|D_{all, \overline{Max\, C_{w_i}}}|}{|D_{w_i, \overline{Max\, C_{w_i}}}|} \quad (4)$$

3. 주제 분별력 점수 결합

본 논문에서는 위에서 제안한 가중치를 결합하여 단어의 주제 분별 점수를 구한다. 결합 방법은 여덟 가지 방식을 사용하며, 각 결합 방식은 3.2에서 제안한 어휘가중치를 서로 곱하여 결합한다.

첫 번째 방식으로 일반적인 기존의 가중치 기법인 TF-IDF를 사용한다. CTF와 CDF의 유용성을 확인하기 위하여 두 번째 방식으로 TF-CTF를, 세 번째 방식으로는 TF-CDF를 사용한다. 네 번째 방식으로는 CTF와 CDF의 곱인 TF-CTF-CDF를 사용한다. 다섯 번째 방식으로 TF-CTF-IECDF를 사용한다. TF와 CTF, IECDF를 곱하여 대표 분류에서의 단어 빈도와 이외 분류에서의 IDF를 반영하여 가중치를 계산한다. 여섯 번째 방식으로는 TF-CDF-IDF를 사용하여, CDF를 사용해 해당 단어가 대표 분류에서 폭넓게 사용될수록, IDF를 통해 모든 문서에서 해당 단어가 희소성이 높을수록 높은 가중치를 얻도록 한다. 일곱 번째 방식으로는 TF-CDF-IECDF를 사용한다. 여섯 번째 방식에서 IDF 대신 IECDF를 사용하여 대표 분류를 제외한 나머지 분류에서의 희소성이 높을수록 높은 점수를 얻도록 한다. 마지막 방법으로는 위의 가중치를 모두 반영한 TF-CTF-IDF-IECDF 방법을 사용한다.

4. 분류 모델 생성

문서 별로 단어와 그 단어에 대한 주제 분별 점수가 구해지면 이를 이용해 분류 모델을 생성한다. 본 논문에서는 기존의 소프트웨어 WEKA 3.6.10 [10]에 구현된 Complement Naive Bayes와 Naive Bayes Multinomial 알고리즘을 사용하여 분류 모델을 생성하였다.

각 분류 알고리즘을 이용하여 문서에서 발생한 용어에 대한 주제 분별 점수를 입력으로 하고, 결과 범주를 출력으로 하여 분류 모델을 생성하였다. 분류 모델의 검증은 검증용 문서 집합을 이용하여 분류 모델의 정확도를 검증하였다.

III. 실험 및 평가

본 논문에서 제안하는 가중치 결합 방식과 2가지의 분류 방법을 통한 분류 정확률을 평가하였다. 실험에서는 16개의 분류에서 각각 임의로 600개씩 문서를 수집하여, 총 9,600개의 문서를 학습 데이터로 이용하였다. 테스트 데이터는 분류 별로 200개씩 수집하여, 총 3,200개의 문서를 이용하였다. 평가 척도로서 Precision을 이용하여 분류 결과를 제시한다. Precision은 분류 모델이 주제 범주에 따라 분류한 문서 중 정확하게 분류한 문서의 비율을 나타낸다.

표1. 각 분류 알고리즘 별 Precision
Table 1. Precision of Each Classifier

분류알고리즘 / 주제분별지표	Complement Naive Bayes	Naive Bayes Multinomial
TF-IDF	44.06%	43.31%
TF-CTF	68.02%	69.48%
TF-CDF	67.70%	63.60%
TF-CTF-CDF	66.64%	63.76%
TF-CTF-IECDF	70.31%	72.12%
TF-CDF-IDF	70.01%	70.29%
TF-CDF-IECDF	72.40%	72.61%
TF-CTF-CDF-IECDF	75.23%	73.57%

표 1은 결과를 보인다. 기존의 방식에서 이용한 TF- IDF를 결합한 방식은 50%가 되지 않는 정확률을 보였다. TF-IDF는 범주의 고유 정보를 반영하지 않기 때문에 낮은 정확률을 보이는 것으로 분석되었다. 이에 비하여 범주 정보가 반영된 TF-CTF와 TF-CDF는 이보다 높은 65% 내외의 정확률을 보여 범주에 기반을 둔 정보사용의 효과를 확인할 수 있었다.

결과에서 Complement Naive Bayes로 학습을 한 TF-CTF-CDF-IECDF가 75.23%로 가장 높은 정확률을 보였다. 전체적인 실험 결과로 역문서 빈도인 IDF를 사용한 방식보다 대표 분류를 제외한 분류에서의 역문서 빈도인 IECDF가 좋은 성능으로 나타났다. 또한, 대표 분류에서 어떠한 단어가 포함된 문서의 수와 그 단어가 발생한 누적 빈도 둘 다 유용하다는 것을 알 수 있었다. 그리하여 본 논문에서 제안한 가중치를 모두 고려했을 때 가장 높은 성능을 보이는 것으로 나타났다.

오류 분석으로는 수집한 문서에서 블로그 포스트 의 특성상 여러 분류 중에 하나의 분류를 명확히 선택하기 어려운 포스트가 있었다. 그리고 블로그 문서들은 최근 트렌드에 따라 특정한 단어가 많이 발생하는 경우가 있어 해당 분류와는 다르게 분류가 되는 경우가 있었다.

IV. 결 론

본 논문에서는 블로그 포스트를 자동으로 분류하는 것을 목적으로 TF-IDF를 확장하여 단어 주제 분별 점수를 계산하기 위한 다양한 가중치를 제안하였다. TF에 범주 정보를 반영하여 확장시킨 가중치 실험에서는, 단순 단어 빈도 TF만 사용했을 때보다 범주로 확장한 단어 빈도나 문서빈도인 CTF와 CDF를 같이 사용했을 때 더 정확한 결과를 보였다. IDF에 범주 정보를 반영하여 확장한 가중치 실험에서는, IDF보다 범주 정보를 반영한 IECDF가 블로그 포스트의 분류 정확률이 높게 나왔다.

블로그 포스트를 자동 분류하는 것은 한계가 있을 수 있다. 정형적인 텍스트가 아니고 오타나 신조어 등에 민감하다. 이러한 점은 고유명사 사전 등을 구축하거나 형태소 분석기의 성능이 향상되면 해결할 수 있을 것으로 기대한다. 추후 연구로는 문서의 자동 분류에서

문서 필터링으로 확장시킬 수 있을 것이다.

Acknowledgement

본 연구는 정부(미래창조과학부) 및 한국산업기술평가관리원의 SW컴퓨팅산업융합원천기술개발사업의 일환으로 수행된 연구임 (2014-044-024-002). 또한, 본 연구는 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단-차세대정보 컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(No. NRF-2014M3C4A7030503).

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] Hong Qu, Andrea La Pietra, and Sarah Poon “Automated blog classification: challenges and pitfalls,” AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp. 184-186, 2006.

[3] *http://cafe.naver.com/korlucene*, 2014.

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

[5] *http://www.cs.waikato.ac.nz/ml/weka/*, Accessed July 25, 2014.

[6] Hee-Sun Jho, Su-Ah Kim, and Hyun Ah Lee, “Automatic classification of blog posts,” 25th Annual Conference on Human and Cognitive Language Technology, 2013.