기계학습을 활용한 소셜 텍스트의 주요 정보 추출 기법

김소현*, 김한준*
*서울시립대학교 전자전기컴퓨터공학부
e-mail : sohyeon24@gmail.com

Extracting Significant Information from Social Text using Machine Learning

So-Hyeon Kim*, Han-joon Kim *
*School of Electrical and Computer Engineering, University of Seoul

요 약

빅데이터 시대를 맞이하여 텍스트마이닝과 오피니언마이닝의 활용도가 커지고 있는 시점에서 소셜 네트워크 데이터로부터 유용한 데이터를 추출하는 작업은 매우 중요하다. 이에 본 논문은 블로그 HTML 문서에서 추출한 태그 특징에 로지스틱 회귀 및 앙상블 기법을 적용하여 본문을 포함하는 태그를 분류하는 모델을 구성한 뒤 태그의 깊이 특징을 이용하여 주요 본문을 찾는 방법을 제안한다. 직접 수집한 데이터를 이용한 실험에서 태그 분류 정확도가 0.990, 본문을 찾아낸 문서의 비율이 80.5%로 나왔다.

1. 서론

최근 다양한 분야에서 소셜 네트워크 서비스(SNS, Social Networking Service)에서 얻은 데이터에 텍스트마이닝(Text Mining)과 오피니언마이닝(Opinion Mining)을 적용하고자 하는 시도가 많아지고 있다.[1,2] 텍스트마이닝이란 비정형 텍스트 데이터에서 의미 있는 정보를 찾아내는 기술이며, 이것의 세부 분야인 오피니언마이닝은 소셜 데이터를 분석하여 극성 및 감성 분석을 하는 기술을 말한다. 위에서 언급한 기술을 적용하였을 때 유용한 정보를 얻을 수 있는 데이터 출처가 바로 소셜 네트워크 서비스이다. 소셜 네트워크 서비스에서는 제품에 대한 사용자들의 의견이나 사회적 이슈에 대한 네티즌의 의견 등 다양한 분야에서 중요하게 사용될 수 있는 정보를 얻을 수 있다. 이러한 소셜 네트워크 서비스 중에서도 블로그는 '1 인 미디어'라고 불리는 만큼 주관적인 글을 얻을 수 있기 때문에 오피니언마이닝을 위한 유용한 데이터를 얻을 수 있다.

그러나 블로그 웹문서에서 핵심 정보를 담고 있는 본문의 텍스트를 추출하는 과정은 고려해야 할 요소가 많다. 대개 블로그 웹페이지는 주요 내용을 담고 있는 본문이외에 광고, 메뉴, 댓글 등과 같은 불필요한 텍스트 영역이 많이 포함되어 있기 때문에 이러한 영역들을 정밀하게 분별하여 본문을 추출할 수 있어야 한다. 또한 블로그는 개인이 운영하는 만큼 HTML 형식이 매우 자유롭고 시간이 흐름에 따라 변동이 크기 때문에 정해진 형식에 얽매이지 않고 본문을 추출할 수 있어야 한다. 본 논문은 이와 같은 블로그 데이터에 대한 본문 추출 문제를 풀기 위해 로지스틱 회귀(Logistic Regression) 및 앙상블(Ensemble) 기법을 활용한 주요 본문의 추출 기법을 제안한다

2. 블로그 HTML 태그을 이용한 본문 추출

본 논문은 HTML 로 작성된 블로그 문서를 트리 구조로 표현하였을 때 각 태그가 가지는 구조적 특징, 텍스트 밀도 특징과 본문 제목과의 연관성 특징을 기계학습 기법에 적용하여 본문 태그를 분류하는 방법을 제안한다. 지금부터 본 논문에서 언급되는 '본문 태그'는 본문을 포함하고 본문을 제외한 광고, 메뉴, 댓글 등의 영역을 최소로 가지고 있는 태그로 정의한다.

2.1 HTML 문서의 태그 특징

HTML 문서는 각각의 HTML 태그를 노드로 가지는 DOM(Document Object Model) 트리 형식으로 나타낼 수 있고[3,4,5] 그 예는 그림 1 과 같다. 이러한 HTML 트리 구조의 특징을 통해, 본문 태그를 분류하는 모델을 학습할 때의 입력 특징(feature)으로 각 태그의 부모 태그의 개수(깊이, depth)와 자식 태그의 개수를 사용할 수 있다. 예를 들면, 그림 1 에서 <div(article)> 태그의 부모 태그 개수는 1 (<body>)이고 자식 태그 개수는 4 (<div(content)>, <a>, <a>, <a>, <p>)이다. 본 연구는 블로그 HTML 에 자주 사용되는 22 개의 태그인 , , <a>, <b>,
, <font>, <h1>, <h2>, <h3>,

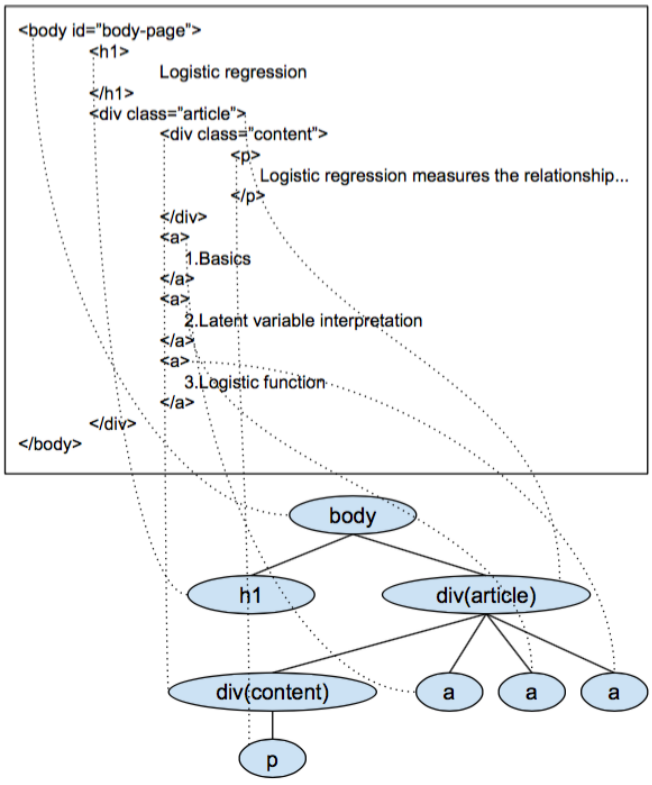
<h4>, <h5>, <h6>, <img>, <li>, <ul>, <ol>, <p>, <pre>, <q>, <table>, <tr>, <td>에 속하는 것만 자식 태그의 개수에 포함시켰다.

본문은 대개 광고, 메뉴, 댓글 영역보다 많은 텍스트를 가지고 있기 때문에 각 태그가 포함하는 텍스트의 길이를 입력 특징으로 고려해볼 수 있다. 따라서 HTML 문서에 포함된 전체 글자 수를 $T(total)$, 각 태그가 포함하는 글자 수를 $T(node)$로 표기하고, 텍스트 밀도 $D(node)$를 식(1)과 같이 정의한다.

$$D(node) = \frac{T(node)}{T(total)} \quad (1)$$

또한 블로그의 본문 제목과 연관성이 높은 태그가 본문 태그일 가능성이 높기 때문에 각 태그가 본문 제목의 단어를 얼마나 포함하는지를 고려해볼 수 있다. 한 태그가 가지는 전체 단어의 수를 N, 해당 태그가 가지는 단어를 Wi라 하고 본문 제목의 단어 집합을 $S(title)$이라고 하여 식 (2)에 따라서 본문 제목과의 연관성을 계산하여 태그의 특징으로 사용하였다.

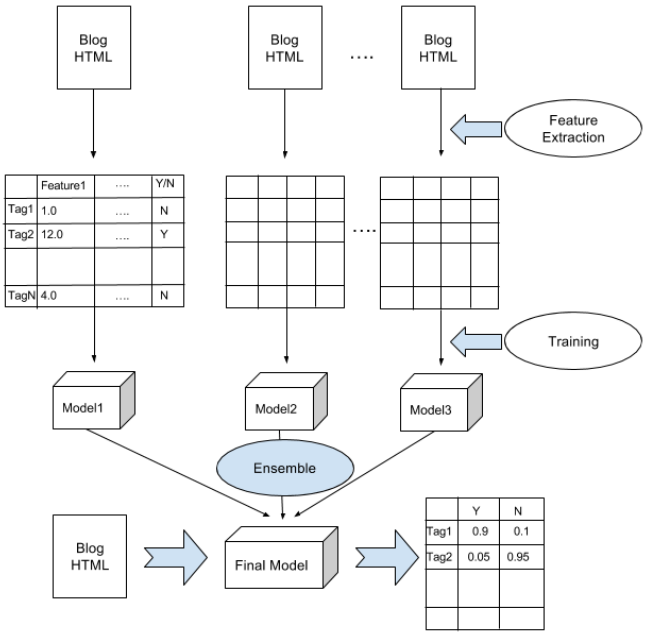
$$\frac{\sum_{i=1}^{N} ti}{N},\ ti = \begin{cases} 1 & when\ Wi \in S(title) \\ 0 & when\ Wi \in \neg S(title) \end{cases} \quad (2)$$



(그림 1) DOM 트리 구조 예시

2.2 본문 태그 추출 알고리즘

크롤러(Crawler)를 이용하거나 사용자가 직접 수집하여 얻은 블로그 HTML 에서 본문 태그가 될 수 있는 태그를 표시하고 학습에 불필요한 부분을 제거하는 전처리를 한다. 그 다음으로 한 HTML 문서를 하나의 트리로 보고 각 태그에 대해 부모 태그의 개수, 22 개의 특정 태그에 포함되는 자식 태그의 개수, 텍스트 밀도 특징 그리고 본문 제목과의 연관성 특징의 총 25 개 특징을 추출하여 각 태그마다 특징 벡터를 구성한다. 추출한 특징 벡터를 입력으로 로지스틱 회귀 기법을 이용해 한 블로그마다 하나의 본문 태그 분류 모델을 만든 뒤 모든 모델을 앙상블하여 최종 모델을 만든다. 이렇게 만들어진 분류 모델을 이용하여 본문 추출을 위해 주어진 블로그 HTML 문서로부터 본문 태그일 확률이 가장 높은 태그를 k 개 색출하고, 그 중 깊이가 가장 큰 태그를 본문 태그로 식별한다.



(그림 2) 본문 태그 추출 모델 개발 과정

3. 실험 및 평가

3 장에서는 실험 데이터, HTML 문서의 태그에 본문 태그를 표시한 방법, 전처리 과정, 본문 태그 분류 모델의 개발 과정과 성능 평가를 설명하였다.

3.1 HTML 문서 전처리와 특징 추출

본 연구에서는 여행 후기, 맛집, 아이폰 7, 삼성 노트 7, 데이터마이닝, mongodb, hiv, spark, hadoop 와 같은 일반적인 주제의 224 개의 블로그 HTML 을 직접 수집하여 실험 데이터로 사용하였다. 웹문서 개발 과정에서 본문 영역은 다양한 이유로 한 개 이상의 태그로 감싸지기 때문에 한 블로그 당 본문 태그를 1~4 개 사이로 정하였다. HTML 태그는 (키=값) 으로 구성된 속성을 가질 수 있으며, 그 예가 그림 1 의 id="body-page", class="article", class="content"이다. 이와 같은 HTML 태그의 특성을 이용하여 직접 수집한 140 개의 블로그 문서의 본문 태그들에 this="main_content" 속성을 추가하였다. 또한 본문 태그를 표시한 HTML 문서에서 학습에 불필요하다고 판단되는 <script>태그와 <!-- -->의 형식으로 쓰이는 주석 부분을 제거한다.

이렇게 정제된 HTML 문서에서 <body> 태그에 속하는 태그만을 검사하여 부모 태그의 개수, 22 개의

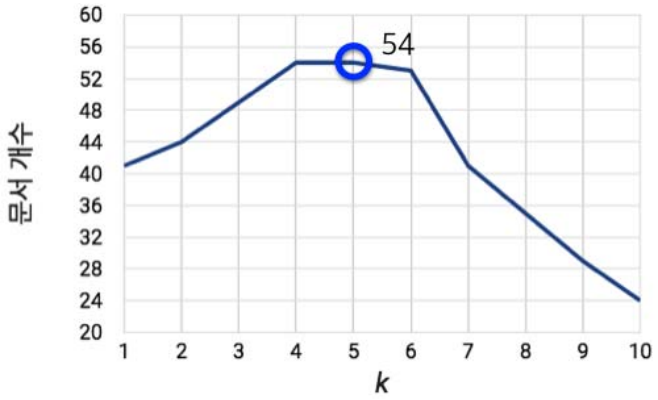
특정 태그에 포함되는 자식 태그의 개수, 텍스트 밀도 특징 그리고 본문 제목과의 연관성 특징의 총 25 개 특징을 추출하여 각 태그의 특징 벡터를 만든다.

3.2 블로그 HTML 의 본문 태그 분류 모델 개발

전처리한 224 개 블로그 HTML 문서를 7:3 의 비율로 나누어 각기 학습 데이터와 테스트 데이터로 사용하였으며 이를 표 1 에 나타내었다. 학습 데이터 문서의 태그 특징 벡터를 로지스틱 회귀 학습의 입력으로 사용하여 한 문서당 하나의 모델을 구성시킨다. 이렇게 학습된 모델을 앙상블을 통해 하나의 모델로 통합하고 이를 이용하여 각각의 테스트 문서의 태그들 중에서 본문 태그일 확률이 가장 높은 태그를 k 개 추출한다. 이 k 개의 태그 중에서 깊이가 가장 깊은 태그를 본문 태그로 추출한다. 이때 적절한 k 값은 실험적으로 테스트 데이터에서 본문 태그를 찾아낸 문서의 개수가 가장 큰 5 로 정하였고 이를 그림 3 에 나타내었다.

<표 1> 실험 데이터 명세

블로그 문서	문서 개수	태그 개수	본문 태그 개수	그 외 태그 개수
학습 데이터	157	112,796	259	112,537
테스트 데이터	67	42,864	108	42,756



(그림 3) k 값에 따른 본문 태그를 식별한 문서 개수

3.3 모델 성능 평가 및 분석

학습한 본문 태그 분류 모델의 성능을 평가하기 위해 67 개의 테스트 문서에서 각 블로그의 주제(여행 후기, 맛집, 아이폰 7, 삼성 노트 7, 데이터마이닝, mongodb, hiv, spark, hadoop)별로 태그 분류의 정확도(accuracy)를 계산했고 이를 표 2 에 나타내었다. 정확도는 분류(예측)한 태그들 중에서 올바르게 분류된 태그들의 비율을 의미한다. 또한 이 실험에서는 한 문서에서 본문 태그를 정확하게 찾아냈는지 여부가 중요하기 때문에 모든 테스트 문서에서 본문 태그를 찾아낸 문서의 비율을 계산하였다.

실험 결과 전체 테스트 문서에서 태그 분류 정확도는 평균 0.990 이고 본문 태그를 정확하게 찾아낸 문서의 비율은 80.5%으로 측정되어 제안한 추출 기법이 복잡하고 변동이 심한 블로그 문서에 대하여 매우 효과적임을 확인하였다.

또한 직접 수집한 데이터를 이용하여 L3C 연구소가 제안한 본문 추출 방법[6,7]에서 착안한 태그 특징을 모델의 입력 특징으로 사용하여 성능 비교 실험을 해보았다. 비교 실험에서는 각 태그의 단어 밀도($D_{WORD}(node)$)와 링크 밀도($D_{LINK}(node)$)를 각각 식 (3), (4) 와 같이 정의하여 태그의 특징으로 사용하였다.

$$D_{WORD}(node) = \frac{Word(node)}{Sentence(node)} \quad (3)$$

$$D_{LINK}(node) = \frac{LinkedWord(node)}{Word(node)} \quad (4)$$

식 (3), (4)에서 $Word(node)$는 각 태그가 포함하는 단어의 개수, $Sentence(node)$는 각 태그가 포함하는 문장의 개수, $LinkedWord(node)$는 각 태그가 포함하는 <a>태그의 단어 개수를 의미한다. [7]에서는 실험적으로 80 개의 단어까지를 하나의 문장으로 정의하였지만 본 논문에서는 개행 문자(\n)로 끝나는 문자를 하나의 문장으로 정의하였다.

이렇게 얻어진 태그 특징을 본 논문에서 소개한 본문 태그 분류 모델의 입력 특징으로 사용하여 실험한 결과 전체 테스트 문서에서 태그 분류 정확도가 0.976 로 나왔다. 결과적으로 본 논문에서 제안한 모델의 정확도가 약 0.024 높게 나와서 더 세밀하게 본문 태그를 분류한 것을 확인하였다.

<표 2> 블로그 주제에 따른 태그 분류 정확도

	제안 모델	기존 모델
여행 후기	0.989	0.982
맛집	0.991	0.981
아이폰 7	0.988	0.967
삼성노트 7	0.987	0.973
데이터마이닝	0.990	0.961
mongodb	0.992	0.945
hive	0.987	0.939
spark	0.990	0.968
hadoop	0.991	0.955
전체 블로그	0.990	0.976

4. 결론

텍스트마이닝과 오피니언마이닝이 여러 분야에서 활발하게 쓰여지고 있는 상황에서 소셜 네트워크 서비스에서 유의미한 데이터를 정확하게 추출하는 것이 중요해졌다. 이에 본 논문은 소셜 네트워크 데이터 중에서도 태그 구조가 복잡하고 변동이 심한 블로그 문서로부터 본문을 추출하는 방법을 제안하였다. 주목할 점은 블로그 HTML 의 구조적 태그 특징과 함께 텍스트 밀도, 본문 제목과의 연관성과 같은 특징을

기계학습에 적용한 뒤 깊이 특징을 이용하여 본문 영역을 정확히 추출할 수 있다는 것이다.

5. 감사의 글

본 연구는 국토교통부 도시건축연구사업의 연구비 지원(16AUDP-B100356-02)에 의해 수행되었습니다.

참고문헌

[1] 배정환, 손지은, 송민. "텍스트 마이닝을 이용한 2012 년 한국대선 관련 트위터 분석." 한국지능정보시스템학회, 지능정보연구 19.3 (2013): 141-156.

[2] 이윤주, 서지훈, 최진탁. "SNS 텍스트 콘텐츠를 활용한 오피니언마이닝 기반의 패션 트랜드 마케팅 예측 분석." 한국정보기술학회논문지 12.12 (2014): 163-170.

[3] Narawade, Shubhada Maruti, et al. "A Web Based Data Extraction Using Hierarchical (DOM) Tree Approach." International Journal for Innovative Research in Science and Technology 2.11 (2016): 255-257.

[4] Geng, Hua, Qiang Gao, and Jingui Pan. "Extracting content for news web pages based on DOM." IJCSNS International Journal of Computer Science and Network Security 7.2 (2007): 124-129.

[5] Kadam, Vinayak B., and Ganesh K. Pakle. "DEUDS: Data Extraction Using DOM Tree and Selectors." International Journal of Computer Science and Information Technologies 5.2 (2014): 1403-1410.

[6] Kohlschütter, Christian, Peter Fankhauser, and Wolfgang Nejdl. "Boilerplate detection using shallow text features." Proceedings of the third ACM international conference on Web search and data mining. ACM (2010): 441-450.

[7] Tomaz K, Evaluating Text Extraction Algorithms. [Online]. Available: http://tomazkovacic.com/blog/