KIPS Tr. Software and Data Eng.
Vol.6, No.5 pp.279~284 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2017.6.5.279

Logistic Regression Ensemble Method for Extracting Significant Information from Social Texts

Kim So Hyeon† · Kim Han Joon††

ABSTRACT

Currenty, in the era of big data, text mining and opinion mining have been used in many domains, and one of their most important research issues is to extract significant information from social media. Thus in this paper, we propose a logistic regression ensemble method of finding the main body text from blog HTML. First, we extract structural features and text features from blog HTML tags. Then we construct a classification model with logistic regression and ensemble that can decide whether any given tags involve main body text or not. One of our important findings is that the main body text can be found through 'depth' features extracted from HTML tags. In our experiment using diverse topics of blog data collected from the web, our tag classification model achieved 99% in terms of accuracy, and it recalled 80.5% of documents that have tags involving the main body text.

Keywords : Machine Learning, Information Extraction, Ensemble, Logistic Regression, Social Media

소셜 텍스트의 주요 정보 추출을 위한 로지스틱 회귀 앙상블 기법

김 소 현† · 김 한 준††

요 약

빅데이터 시대를 맞이하여 텍스트마이닝과 오피니언마이닝의 활용도가 커지고 있는 시점에서 소셜 네트워크 서비스로부터 유용한 정보를 추출하는 작업은 매우 중요한 연구 주제 중 하나이다. 이에 본 논문은 블로그 HTML 문서에서 주요 본문을 찾는 로지스틱 회귀 앙상블 기법을 제안한다. 먼저, 블로그 HTML 태그에서 구조적 특징, 텍스트 특징을 추출한다. 그 다음, 블로그 HTML 문서에서 추출한 태그 특징에 로지스틱 회귀 및 앙상블 기법을 적용하여 본문을 포함하는 태그를 분류하는 모델을 구성한다. 본 연구의 중요한 발견 중 하나는 태그의 깊이 특징을 이용하여 주요 본문을 찾을 수 있다는 점이다. 다양한 주제의 국내 블로그 데이터를 이용한 실험에서 태그 분류 정확도가 99%, 본문을 찾아낸 문서의 비율이 80.5%로 평가되었다.

키워드 : 기계학습, 정보 추출, 앙상블, 로지스틱 회귀, 소셜 네트워크 서비스

1. 서 론

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 유용하게 사용될 수 있는 정보를 얻을 수 있다. 그에 따라 소셜 텍스트에서 유용한 정보를 추출하고자 하는 여러 연구가 진행되고 있다[3, 4]. 이러한 소셜 네트워크 서비스 중에서도 블로그는 '1인 미디어'라고 불리는 만큼 주관적인 글을 얻을 수 있기 때문에 오피니언마이닝을 위한 유용한 정보를 얻을 수 있다[5].

그러나 블로그 웹문서에서 주요 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 포함되

※ 본 연구는 국토교통부 도시건축연구사업의 연구비지원(17AUDP-B100356-03)에 의해 수행되었습니다.
※ 이 논문은 2016년도 한국정보처리학회 추계학술발표대회에서 '기계학습을 활용한 소셜 텍스트의 주요 정보 추출 기법'의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 서울시립대학교 전자전기컴퓨터공학과 석사과정
†† 정 회 원 : 서울시립대학교 전자전기컴퓨터공학과 정교수
Manuscript Received : December 16, 2016
First Revision : January 23, 2017
Accepted : January 27, 2017
* Corresponding Author : Kim Han Joon(khj@uos.ac.kr)

어 있기 때문에 이러한 영역들을 정밀하게 분별하여 본문을 추출할 수 있어야 한다. 블로그 웹 페이지 구조의 예시를 Fig. 1에 나타내었다.

기존의 몇 가지 연구는 웹 페이지의 구조적 반복성에 주목하여 본문을 추출하였다[6, 7]. 이 중에는 선험적 알고리즘(Apriori Algorithm)[6]과 같은 연관규칙(Association Rule)을 이용하여 웹 페이지간의 유사성을 토대로 본문을 추출하는 연구도 있다. 그러나 이러한 접근은 최근에 웹 애플리케이션의 발전과 개인 맞춤형 블로그 호스팅 서비스가 늘어남에 따라 심화된 블로그 구조의 개인화 특성을 고려하지 못하고 있다. 즉, 블로그의 HTML 형식이 매우 자유롭고 시간이 흐름에 따라 변동이 크기 때문에 정해진 형식에 얽매이지 않고 본문을 추출할 수 있어야 한다. 따라서 본 논문은 로지스틱 회귀(Logistic Regression) 및 앙상블(Ensemble) 기법을 활용하여 유동적인 블로그 웹페이지 구조에 대해 주요 본문을 추출하는 기법을 제안한다.

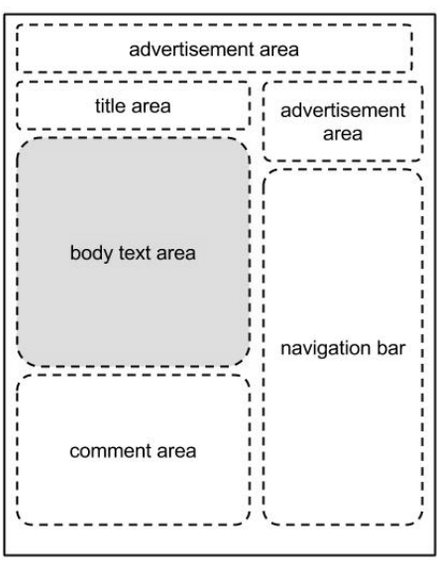

Fig. 1. Example of Blog Web Page's Structure

2. 관련 연구

기존의 다양한 연구가 웹 페이지의 본문을 추출하기 위해 기계학습을 사용한다[8-10]. 이러한 연구들에서 기계학습의 입력 특징(feature)을 정의하기 위해 보편적으로 이용하는 것이 웹 페이지의 HTML 태그의 글자 수와 하이퍼링크(hyperlink) 그리고 문서 객체 모델(DOM, Document Object Model)이다.

L3S 연구소[8]는 각 태그의 단어 수, 문장 수, 하이퍼링크 수에 따라 단어 밀도 특징과 링크 밀도 특징을 정의하여 기계학습을 하였다. 그러나 해당 기법은 본문의 주제와 같은 텍스트의 의미적인 성질을 이용하지 못하고 있다. 일반적으로 광고, 메뉴, 댓글과 같은 영역보다 본문이 주제와 관련된 단어를 포함하고 있을 가능성이 크다. 따라서 본문 주제를 고려한 특징을 정의한다면 학습 성능을 더 올릴 수 있을 것이다. 본 논문에서는 글자 수, 하이퍼링크 수에 따른 특징뿐만 아니라 각 태그가 본문 제목의 단어를 얼마나 포함하는지에 따른 본문 제목과의 연관성 특징도 정의하여 학습의 성능을 높이고자 했다.

HTML 문서는 각각의 HTML 태그를 노드로 가지는 문서 객체 모델(DOM, Document Object Model) 형식으로 나타낼 수 있고[11-13] 그 예는 Fig. 2와 같다. 문서 객체 모델이란 구조화된 모델을 표현하는 형식으로서 트리 구조로 나타낼 수 있다. 문서 객체 모델 트리 구조를 이용한 연구[10]에서는 [8]와는 다르게 태그 특징을 정의하였다. [8]은 각 태그가 가지는 단어 수, 문장 수, 하이퍼링크 수만을 고려해서 특징을 정의하였지만 [10]은 문서 객체 모델의 트리 구조를 이용하여 각 태그의 하위 태그들의 글자 수와 하이퍼링크 수를 모두 더해서 해당 태그의 텍스트 밀도 특징을 정의하였다. 그러나 [10]은 트리 구조에서 가장 일반적으로 생각할 수 있는 깊이 특징을 고려하지 않았다. 본 논문에서 실험한 결과, 태그의 깊이가 본문 태그를 찾는데 중요한 역할을 할 수 있음을 알아내었다. 이에 따라 본 논문은 기계학습의 결과로 본문 태그일 확률을 구한 뒤 깊이 특징을 이용한 후처리를 통해 더 정밀하게 본문을 추출하였다.

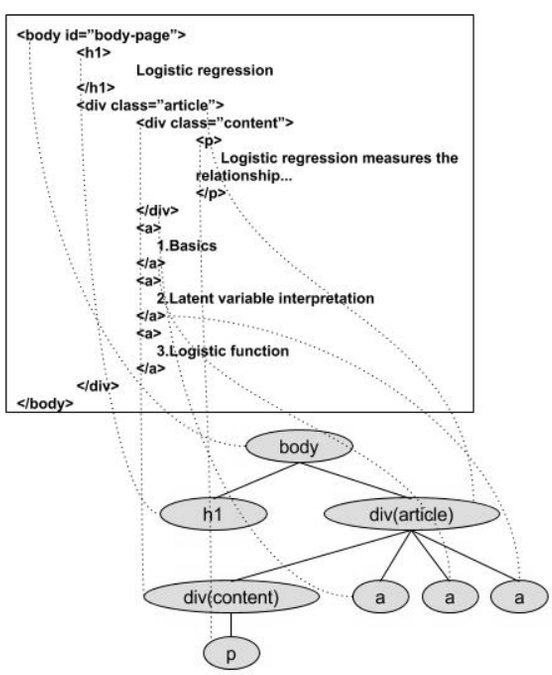

Fig. 2. Example of DOM Tree

3. 블로그 HTML 태그를 이용한 본문 추출

본 논문은 HTML로 작성된 블로그 문서를 트리 구조로 표현하였을 때 각 태그가 가지는 구조적 특징, 텍스트 밀도 특징과 본문 제목과의 연관성 특징을 정의하고 이러한 특징

들에 기계학습 기법을 적용하여 본문 태그를 분류하는 방법을 제안한다. 지금부터 본 논문에서 언급되는 '본문 태그'는 본문을 포함하면서 본문을 제외한 광고, 메뉴, 댓글 등의 영역을 최소로 가지는 태그로 정의한다.

3.1 HTML 문서의 태그 특징

문서 객체 모델 트리 구조의 계층적 성질을 이용하여, 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, Fig. 2에서 <div(article)> 태그의 부모 태그 개수는 1 (<body>)이고 자식 태그 개수는 4 (<div(content)>, <a>, <a>, <a>, <p>)이다. 본 연구는 블로그 HTML에 자주 사용되는 22개의 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, 댓글보다 많은 텍스트를 가지고 있기 때문에 각 태그가 포함하는 텍스트의 길이를 고려한 태그 특징을 생각해 볼 수 있다. 따라서 HTML문서에 포함된 전체 글자 수를 $L(total)$, 각 태그가 포함하는 글자 수를 $L(node)$로 표기하여 텍스트 밀도 $D(node)$를 Equation (1)과 같이 정의한다. 여기서 L는 텍스트 길이(text length), D는 텍스트 밀도(Density of text) 그리고 $node$는 각각의 태그(tag, node)를 나타낸다. 즉, 텍스트 밀도 $D(node)$는 각 태그가 포함하는 글자 수를 HTML문서에 포함된 전체 글자 수로 나눈 것이다.

$$D(node) = \frac{L(node)}{L(total)} \tag{1}$$

또한 블로그의 본문 제목과 연관성이 높은 태그가 본문 태그일 가능성이 높기 때문에 각 태그가 본문 제목의 단어를 얼마나 포함하는지를 고려한 태그 특징을 생각해 볼 수 있다. 따라서 한 태그가 가지는 전체 단어 개수를 N, 해당 태그가 가지는 각각의 단어를 $W_i(0 \le i \le N)$ 그리고 본문 제목의 단어 집합을 $S(title)$로 표기하여 본문 제목과의 연관성$R(node)$를 Equation (2)와 같이 정의한다. 여기서 S는 단어 집합(set of words), R은 연관성(relationship), $node$는 각각의 태그(tag, node) 그리고 $title$은 본문 제목을 나타낸다. 즉, 본문 제목과의 연관성 $R(node)$는 각 태그가 포함하는 본문 제목의 단어 집합에 속하는 단어 개수를 해당 태그가 가지는 전체 단어 개수로 나눈 것이다.

$$R(node) = \frac{\sum_{i=1}^{N} t_i}{N}, t_i = \begin{cases} 1 \ when \ W_i \in S(title) \\ 0 \ when \ W_i \in \neg S(title) \end{cases} \tag{2}$$

3.2 본문 태그 추출 알고리즘

크롤러(Crawler)를 이용하거나 사용자가 직접 수집하여 학습에 필요한 블로그 HTML 문서를 얻는다. 수집한 블로그 HTML에서 본문 태그가 될 수 있는 태그를 표시한 뒤 학습에 불필요한 부분을 제거하는 전처리를 한다. 그 다음으로 한 블로그 HTML문서를 하나의 트리 구조로 분석하여 각 태그에 대해 부모 태그의 개수, 22개의 특정 태그에 포함되는 자식 태그의 개수, 텍스트 밀도 특징 그리고 본문 제목과의 연관성 특징의 총 25개 특징을 추출하여 각 태그마다 특징 벡터를 구성한다. 추출한 특징 벡터를 입력으로 로지스틱 회귀 기법을 이용해 한 블로그마다 하나의 본문 태그 분류 모델을 학습한다. 학습한 모든 본문 태그 분류

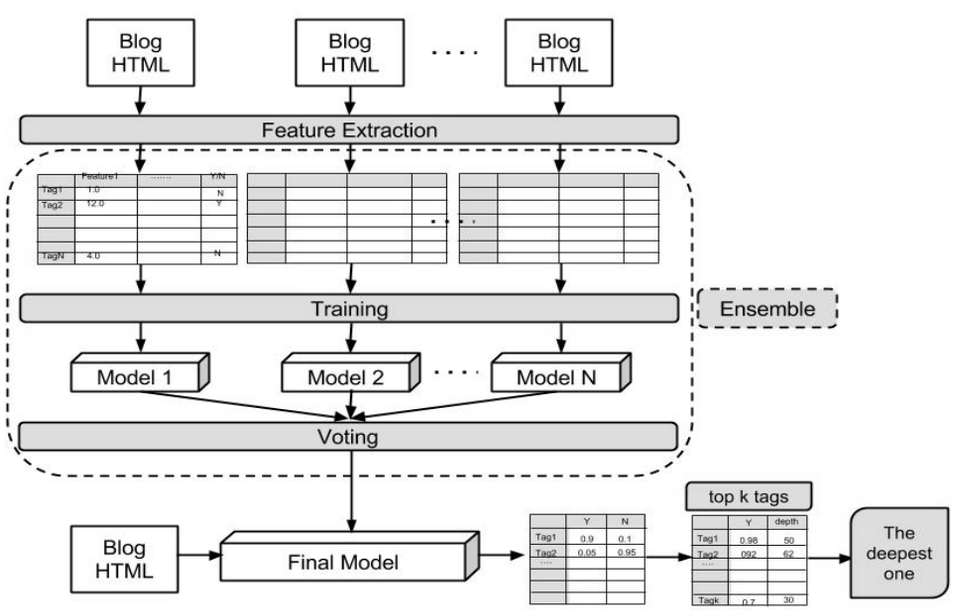

Fig. 3. Process of Extracting Tags that Contain the Main Body Text

모델에서 얻은 본문 태그일 확률의 평균값을 구하는 확률적 보팅을 하여 앙상블을 한다[14, 15]. 앙상블 기법을 사용하는 이유는 모든 문서에 대해 하나의 모델을 만드는 것보다 각 문서마다 하나의 모델을 만들어서 앙상블 하는 것이 실험적으로 성능이 더 좋게 나왔기 때문이다. 또한 본문 태그일 확률을 결과값으로 내는 이유는 다음 단계에서 이 확률을 이용해서 확률이 높은 k개의 태그에 깊이 특징을 적용하는 후처리를 하기 때문이다.

결과적으로 한 블로그 HTML 문서에서 최종 모델을 이용하여 본문 태그를 추출하는 과정은 다음과 같다. 본문을 추출하고자 하는 블로그 HTML 문서에서 25개의 특징을 추출하여 특징 벡터를 구성한다. 특징 벡터를 앙상블의 결과로 얻은 최종 본문 태그 분류 모델에 넣어 모든 태그들의 본문 태그일 확률과 본문 태그가 아닐 확률을 얻는다. 이때, 한 태그가 본문 태그일 확률이 아닐 확률보다 높으면 그 태그는 1차적으로 본문 태그 후보군에 들어간다. 본문 태그 후보군 중 본문 태그일 확률이 가장 높은 태그를 k개 색출하고 그 중 깊이 특징이 가장 큰 태그를 본문 태그로 식별한다. 본문 태그 추출 기법의 프로세스를 Fig. 3에 나타내었다.

4. 실험 및 평가

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

4.1 HTML 문서 전처리와 특징 추출

본 연구에서는 여행 후기, 맛집, 아이폰7, 삼성 노트7, 데이터마이닝, mongodb, hiv, spark, hadoop와 같은 일반적인 주제의 224개의 블로그 HTML을 직접 수집하여 실험 데이터로 사용하였다. 웹문서 개발 과정에서 본문 영역은 다양한 이유로 한 개 이상의 태그로 감싸지기 때문에 한 블로그당 본문 태그를 1~4개 사이로 정하였다. HTML 태그는 (키="값") 으로 구성된 속성을 가질 수 있으며, 그 예가 Fig 2의 id="body-page", class="article", class="content"이다. 이와 같은 HTML태그의 특성을 이용하여 직접 수집한 224개의 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 제거하였다.

대부분의 블로그 HTML 문서는 <body>태그에 본문이 있기 때문에 실험의 효율성을 위해 정제된 블로그 HTML 문서에서 <body> 태그에 속하는 태그만을 검사하여 부모 태그의 개수, 22개의 특정 태그에 포함되는 자식 태그의 개수, 텍스트 밀도 특징 그리고 본문 제목과의 연관성 특징의 총 25개 특징을 추출하여 각 태그의 특징 벡터를 만든다.

4.2 블로그 HTML의 본문 태그 분류 모델 개발

전처리한 224개 블로그 HTML 문서를 7:3의 비율로 나누어 각기 학습 데이터와 테스트 데이터로 사용하였다. 각 문서 개수, 태그 개수, 본문 태그 개수와 그 외 태그 개수를 Table 1에 나타내었다. 각각의 학습 데이터 문서의 태그 특징 벡터를 로지스틱 회귀 학습의 입력 데이터로 사용하여 한 문서 당 하나의 모델을 구성시킨다. 이렇게 학습된 모델들을 앙상블을 통해 하나의 모델로 통합한다. 앙상블한 모델에 테스트 문서들을 개별적으로 넣어 한 테스트 문서 당 본문 태그일 확률이 가장 높은 태그를 k개 추출한다. 마지막으로 이 k개의 태그 중에서 깊이가 가장 깊은 태그를 본문 태그로 추출한다. 이때 적절한 k값은 실험적으로 테스트 데이터에서 본문 태그를 찾아낸 문서의 개수가 가장 큰 5로 정하였고 이를 Fig. 4에 나타내었다.

Table 1. Specification of the Data

	Documents	Total Tags	Tags Involving Main Text	Other Tags
Training Data	157	112,796	259	112,537
Test Data	67	42,864	108	42,756

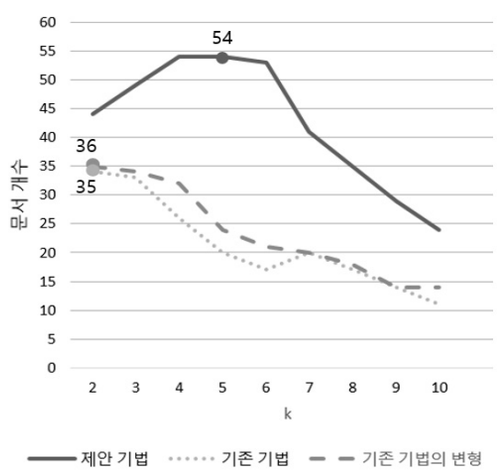

Fig. 4. Changes of the Number of Documents Containing the Main Body Text from Varying k Values

4.3 모델 성능 평가 및 분석

학습한 본문 태그 분류 모델의 성능을 평가하기 위해 67개의 테스트 문서에서 각 블로그의 주제(여행 후기, 맛집, 아이폰7, 삼성 노트7, 데이터마이닝, mongodb, hiv, spark, hadoop)별로 태그 분류의 정확도(accuracy)를 계산했고 이를 Table 2에 나타내었다. 이때 정확도는 앙상블한 모델이 본문 태그 후보군으로 분류(예측)한 태그들 중에서 올바르게 분류된 태그들의 비율을 의미한다. 또한 이 실험에서는 각각의 블로그 HTML 문서에서 본문 태그를 정확하게 찾아냈는지 여부가 중요하기 때문에 모든 테스트 문서에서 본문 태그를 찾아낸 문서의 비율을 계산하였다. 이때 본문 태그로 예측한 태그는 최종적으로 본문 태그 후보군에서 본문 태그일 확률이 높은 k개의 태그에서 깊이 특징이 가장 큰 태그이다.

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

또한 기존 기법과의 성능 비교를 위해 직접 수집한 데이터를 이용하여 비교 실험을 해보았다. 비교 실험에서는 L3S 연구소가 제안한 본문 추출 방법[8, 9]의 태그 특징을 본문을 포함하면서 광고, 메뉴, 댓글 등의 영역을 최소로 가지는 태그를 본문 태그로 정의한 본 논문의 목적에 맞게 약간 변형하여 사용하였다. 기존 기법에서 착안한 태그 특징은 다음과 같이 정의하였다. 각 태그가 포함하는 단어 개수를 $Word(node)$, 각 태그가 포함하는 문장 개수를 $Sentence(node)$ 그리고 각 태그가 포함하는 <a>태그의 단어 개수를 $LinkedWord(node)$로 표기하여 각 태그의 단어 밀도 특징 $D_{WORD}(node)$와 링크 밀도 특징 $D_{LINK}(node)$를 각각 Equations (3), (4)와 같이 정의한다. 즉, 단어 밀도는 각 태그가 포함하는 단어 개수를 해당 태그가 포함하는 문장 개수로 나눈 것이고 링크 밀도는 각 태그가 포함하는 <a>태그의 단어 개수를 해당 태그가 포함하는 단어의 개수로 나눈 것이다. [9]에서는 실험적으로 80개의 단어까지를 하나의 문장으로 정의하였지만 본 논문에서는 개행문자(\n)로 끝나는 부분까지를 하나의 문장으로 정의하였다.

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

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

이렇게 얻어진 태그 특징을 본 논문에서 소개한 본문 태그 분류 모델의 입력으로 모델을 학습한 결과 기존 기법이 전체 테스트 문서에서 태그 분류 정확도가 0.976로 나왔고 각 블로그 주제에 대한 정확도를 Table 2에 기존 기법으로 나타냈다. 또한, 테스트 문서에서 본문 태그를 찾아낸 비율이 k가 1일 때 52.2%로 가장 높게 나왔고 이를 Fig. 4에 나타내었다.

Table 2. Tag Classification Accuracy for the Topics of Blogs

	Proposed Method	Conventional Method	Variation of Conventional Method
Travel	0.989	0.982	0.984
Restaurant	0.991	0.981	0.983
iPhone7	0.988	0.967	0.966
Galaxy Note7	0.987	0.973	0.975
Data Mining	0.990	0.961	0.961
Mongodb	0.992	0.945	0.947
Hive	0.987	0.939	0.941
Spark	0.990	0.968	0.968
Hadoop	0.991	0.955	0.958
Total	0.990	0.976	0.977

또한, 제안 기법은 모델의 학습에서도 깊이 특징을 사용했기 때문에 이를 기본 기법과 비교하기 위해 기존 기법의 단어 밀도 특징, 링크 밀도과 함께 깊이 특징을 입력 특징으로 추가하여 모델 학습의 입력으로 넣어주는 변형된 기존 기법에 대해서도 실험하였다. 그 결과 변형된 기존 기법이 전체 테스트 문서에서 태그 분류 정확도가 0.977로 나왔고 각 블로그 주제에 대한 정확도를 Table 2에 기존 기법의 변형으로 나타냈다. 또한, 테스트 문서에서 본문 태그를 찾아낸 비율이 k가 1일 때 53.7%로 가장 높게 나왔고 이를 Fig. 4에 나타내었다.

결과적으로 본 논문에서 제안한 모델의 정확도가 기존 기법과 기존 기법의 변형과 비교하여 각각 0.024, 0.013 정도 높게 나와서 더 세밀하게 본문 태그를 분류한 것을 확인하였다. 또한 본문 태그를 찾아낸 비율은 기존 기법과 기존 기법의 변형과 비교하여 매우 높게 나타났기 때문에 제안 기법의 본문 추출 프로세스에 제안 기법에서 정의한 25개의 태그 특징이 매우 적절하게 적용된 것을 알 수 있었다.

5. 결 론

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 HTML의 구조적 태그 특징과 함께 텍스트 밀도, 본문 제목과의 연관성과 같은 특징을 기계학습에 적용한 뒤 깊이 특징을 이용하여 본문 영역을 정확히 추출할 수 있다는 것이다.

향후에는 다섯 가지 관점에서 본 연구를 더 발전시키고자 한다.

첫 번째로, 본문 태그를 더 엄격하게 정의하여 앙상블 모델의 성능을 높이고자 한다. 제안 기법의 모든 단계를 거치고 나서 최종적으로 본문 태그를 찾아낸 문서의 비율은 80.5%로 나오지만 중간 단계인 앙상블 모델의 분류 결과는 정밀도(precision)와 재현율(recall)이 낮게 나온다. 이를 높이기 위해 본문 태그를 문서 당 하나로 제한하는 것이 효과가 있을 것으로 생각된다. 본 논문에서는 본문 태그를 본문을 포함하면서 댓글, 메뉴, 광고 등 본문이 아닌 영역을 최소로 가진 태그로 정의하여 한 문서 당 1~4개의 태그를 본문 태그로 정하였다. 향후 연구에서는 다른 기준을 통해 문서 당 태그를 하나만 선별하여 본문 태그로 정의하고자 한다.

두 번째로, 본문 태그 분류 모델의 입력 특징에 대한 좀 더 세밀한 검증을 통해 태그 분류 정확도를 높인다. 본 논문에서 제시하는 본문 태그 분류 모델은 태그가 가지는 구조적 특징, 텍스트 밀도 특징과 본문 제목과의 연관성 특징 등 25개의 특징을 정의하여 해당 모델의 입력 특징으로 사용하였다. 그러나 입력 특징을 정의하는 과정에서 각 특징들 간의 관계와 중요도를 실험적으로 측정하여 특징 선택(feature selection)을 한다면 모델의 성능을 더 올릴 수 있을 것이다[16, 17].

세 번째로, 제안한 기법의 마지막 과정인 본문 태그일 확률이 가장 높은 k개의 태그들에서 깊이가 가장 큰 태그를

찾아내는 과정을 좀 더 발전시키고자 한다. 깊이 특징은 각 태그들의 구조적 연관성에서 나온 특징이기 때문에 단순히 깊이가 가장 큰 태그를 본문 태그로 예측하기 보다는 k개의 태그들을 다시 트리 구조로 표현하여 이를 통해서 태그 깊이에 대한 정밀한 탐색을 하는 것이 더 정확한 본문 태그를 찾을 것으로 예상된다.

네 번째로, 블로그 HTML 문서에서 본문 태그 모델을 학습할 때 블로그의 주제와 관련된 특징을 정의하여 활용할 수 있다. 각 주제마다 블로그 웹 페이지의 구조나 본문의 위치 등이 다를 수 있고 본문은 블로그 주제와 관련된 단어를 많이 포함할 수 있기 때문에 주제 관련 특징을 정의하는 것도 의미 있는 연구가 될 수 있다.

마지막으로, 블로그 문서의 본문뿐만 아니라 더 나아가서 블로그 댓글에서 주요 정보를 추출하는 것을 생각할 수 있다. 블로그 문서의 댓글에서는 본문에서 추출할 수 없었던 해당 블로그의 관리자가 아닌 네티즌들의 의견을 수집할 수 있기 때문에 오피니언마이닝에 더욱 유용하게 사용할 수 있을 것이다.

References

[1] Jung-hwan Bae, Ji-eun Son, and Min Song, “Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques,” *Journal of Intelligence and Information Systems,* Vol.19, No.3, pp.141-156, 2013.

[2] Yoon-Ju Lee, Ji-Joon Seo, and Jin-Tak Choi, “Fashion Trend Marketing Prediction Analysis Based on Opinion Mining Applying SNS Text Contents,” *Journal of Korean Institute of Information Technology (KIIT),* Vol.12, No.12, pp.163-170, 2014.

[3] Imran, Muhammad et al., “Extracting information nuggets from disaster-related messages in social media,” *Proc. of ISCRAM*, Baden-Baden, Germany, 2013.

[4] So-hyeon Kim and Han-joon Kim, “Extracting Significant Information from Social Text using Machine Learning,” *Korea Information Processing Society, The KIPS Fall Conference,* Vol.23, No.2, pp.742-745, 2016.

[5] Wang, Changzhi et al., “Opinion Mining Research on Chinese Micro-blog,” *First International Conference on Information Science and Electronic Technology*, 2015.

[6] Gulhane, Pankaj et al., “Exploiting content redundancy for web information extraction,” *Proceedings of the VLDB Endowment,* Vol.3, pp.578-587, 2010.

[7] Bronzi, Mirko et al., “Extraction and integration of partially overlapping web sources,” *Proceedings of the VLDB Endowment,* Vol.6, No.10, pp.805-816, 2013.

[8] 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*, pp.441-450, 2010.

[9] Tomaz K, Evaluating Text Extraction Algorithms [Internet], http://tomazkovacic.com/blog/.

[10] Sun, Fei, Dandan Song, and Lejian Liao, “Dom based content extraction via text density,” *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp.245-254, 2011.

[11] Narawade, Shubhada Maruti et al., “A Web Based Data Extraction Using Hierarchical (DOM) Tree Approach,” *International Journal for Innovative Research in Science and Technology*, Vol.2, No.11, pp.255-257, 2016.

[12] 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*, Vol.7, No.2, pp.124-129, 2007.

[13] Kadam, Vinayak B., and Ganesh K. Pakle, “DEUDS: Data Extraction Using DOM Tree and Selectors,” *International Journal of Computer Science and Information Technologies*, Vol.5, No.2, pp.1403-1410, 2014.

[14] Kuswanto, Heri et al., “Logistic Regression Ensemble for Predicting Customer Defection with Very Large Sample Size,” *Procedia Computer Science*, Vol.72, pp.86-93, 2015.

[15] Wang, Hong, Qingsong Xu, and Lifeng Zhou, “Large unbalanced credit scoring using Lasso-logistic regression ensemble,” *PloS one,* Vol.10, No.2, e0117844, 2015.

[16] Chandrashekar, Girish, and Ferat Sahin, “A survey on feature selection methods,” *Computers & Electrical Engineering,* Vol.40, No.1, pp.16-28, 2014.

[17] Jurado, Sergio et al., “Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques,” *Energy*, Vol.86, pp.276-291, 2015.

김 소 현

e-mail : sohyeon24@gmail.com

2016년 서울시립대학교 전자전기컴퓨터공학부(공학사)

2016년~현 재 서울시립대학교 전자전기컴퓨터공학과 석사과정

관심분야 : Data Mining, Machine Learning, Deep Learning, Information Retrieval, Data Warehouse

김 한 준

e-mail : khj@uos.ac.kr

1994년 서울대학교 계산통계학과(공학사)

1996년 서울대학교 전산과학과(공학석사)

2002년 서울대학교 컴퓨터공학부 (공학박사)

2002년~현 재 서울시립대학교 전자전기컴퓨터공학과 정교수

관심분야 : 텍스트마이닝, 기계학습, 온톨로지, 정보검색, 데이터베이스