JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 25, No. 12: 1709~1715, Dec. 2021

태그 경로 및 텍스트 출현 빈도를 이용한 HTML 본문 추출

김진환[1] · 김은경[2*]

HTML Text Extraction Using Tag Path and Text Appearance Frequency

Jin-Hwan Kim[1] · Eun-Gyung Kim[2*]

[1]Graduate Student, Department of Computer Science & Engineering, Korea University of Technology and Education, Cheonan, 31253 Korea
[2*]Professor, School of Computer Science & Engineering, Korea University of Technology and Education, Cheonan, 31253 Korea

요 약

웹 페이지에서 필요한 텍스트를 정확하게 추출하기 위해 본문이 존재하는 곳의 태그와 스타일 속성을 웹 크롤러에 명시하는 방법은 웹 페이지 구성이 변경될 때마다 본문을 추출하는 로직을 수정해야 하는 문제가 있다. 이러한 문제점을 해결하기 위해 이전 연구에서 제안한 텍스트의 출현 빈도를 분석하여 본문을 추출하는 방법은 웹 페이지의 수집 채널에 따라 성능 편차가 크다는 한계점이 있었다. 따라서 본 논문에서는 텍스트의 출현 빈도뿐만 아니라 웹 페이지의 DOM 트리로부터 추출된 텍스트 노드의 부모 태그 경로를 분석하여 다양한 수집 채널에서 높은 정확도로 본문을 추출하는 방법을 제안하였다.

ABSTRACT

In order to accurately extract the necessary text from the web page, the method of specifying the tag and style attributes where the main contents exist to the web crawler has a problem in that the logic for extracting the main contents. This method needs to be modified whenever the web page configuration is changed. In order to solve this problem, the method of extracting the text by analyzing the frequency of appearance of the text proposed in the previous study had a limitation in that the performance deviation was large depending on the collection channel of the web page. Therefore, in this paper, we proposed a method of extracting texts with high accuracy from various collection channels by analyzing not only the frequency of appearance of text but also parent tag paths of text nodes extracted from the DOM tree of web pages.

키워드 : 웹 크롤링, 웹 스크래이핑, 빅데이터 수집, 텍스트 빈도 분석, 태그 경로 분석

Keywords : Web crawling, Web scrapping, Big data collection, Text frequency analysis, Tag path analysis

Received 9 September 2021, **Revised** 14 September 2021, **Accepted** 27 September 2021

*** Corresponding Author** Eun-Gyung Kim(E-mail: egkim@koreatech.ac.kr Tel:+82-41-560-1350)
Professor, School of Computer Science & Engineering, Korea University of Technology and Education, Choenan, 31253 Korea

Open Access http://doi.org/10.6109/jkiice.2021.25.12.1709 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

최근 사회적 이슈에 대한 대중의 반응을 분석하여 사회적 인식 변화를 탐색하기 위한 목적으로 SNS와 뉴스, 블로그 등 다양한 수집 채널에서 수집한 텍스트를 이용한 빅데이터 분석이 활발하게 이루어지고 있다[1-3]. 브라우저 상에 나타나는 무수히 많은 텍스트 중에서 빅데이터 분석 목적에 부합하는 본문 텍스트를 추출해야 하지만, 수많은 텍스트와 태그의 조합으로 구성되어 있는 웹 페이지에서 본문 텍스트만을 추출하기란 쉽지 않다.

웹 페이지에서 빅데이터 분석에 필요한 본문 텍스트만을 정확하게 추출하기 위해 본문이 존재하는 곳의 태그와 스타일 속성을 웹 크롤러에 명시하는 방법이 지금까지 주로 이용되고 있다[4]. 하지만 이러한 방법은 사람이 직접 태그와 스타일 속성을 찾기 위해 웹 페이지의 구조를 분석해야 하므로 상당히 번거롭고 오랜 시간이 소요될 뿐만 아니라, 웹 페이지 구성이 변경될 때마다 본문을 추출하는 로직을 수정해야 하는 단점이 있다.

이러한 단점을 보완하기 위해 웹 페이지에 나타나는 텍스트나 태그 정보를 학습하여 본문을 추출하는 다양한 방법들이 제안되었다[5-13]. 하지만 이런 방법들은 웹 페이지의 구조가 수정되어도 본문을 추출하는 로직을 변경하지 않아도 되는 장점이 있지만, 수집 채널에 따라 본문 추출 성능의 편차가 커서 일반성이 떨어진다는 단점이 있다.

본 논문에서는 이전 연구[13]에서 제안한 텍스트 출현 빈도 분석 방법뿐만 아니라 웹 페이지의 DOM (Document Object Model) 트리로부터 추출한 텍스트 노드의 부모 태그 경로를 이용하여 다양한 수집 채널에서 높은 정확도로 본문을 추출하는 방법을 제안하였다.

본 논문의 2장에서는 웹 페이지에서의 본문 추출과 관련된 최근 연구에 대해 소개하고, 3장에서는 본 논문에서 제안하는 본문 추출 방법에 대해 자세히 설명하였다. 4장에서는 다양한 수집 채널에서의 본문 추출 성능을 측정하여 제안한 방법의 우수성을 입증하였다.

Ⅱ. 관련 연구

[5]에서는 웹 페이지에서 텍스트 노드를 추출한 후, 텍스트 노드의 글자 수와 텍스트 노드와 연결된 링크 수로 계산한 단어/링크 밀도와 인접한 태그 정보를 학습하여 텍스트 노드의 본문 여부를 구분하였다. [6]에서는 본문 추출에 참고하는 인접한 태그 정보를 확장하여 [5]의 성능을 향상시켰다. [7]에서는 웹 페이지에서 불필요한 DOM 요소들을 제거하고 텍스트가 긴 텍스트 노드를 군집하여 본문을 추출하는 방법으로 [5]의 성능을 향상시켰다. [8]에서는 웹 페이지의 DOM 트리에서 본문에 해당하는 텍스트 노드들의 공통된 조상 노드(ancestor node)의 위치를 나타내는 특징을 포함한 총 11개의 특징을 이용하여 본문을 추출하였으며, [9]에서는 DOM 트리 상에 나타나는 노드의 위치와 폰트, 텍스트, 태그, 링크 등의 특징과 본문 텍스트 노드가 연속적으로 나타난다는 특징을 이용하여 본문을 추출하였다. [10]에서는 각 태그의 부모 태그 및 특정 태그에 포함되는 자식 태그의 개수, 텍스트 밀도, 본문 제목과의 연관성 등 총 25개의 특징을 이용하여 본문을 추출하였다. [11]에서는 정규표현식을 만족하는 텍스트의 포함여부를 비롯하여 128개의 규칙에 따라 추출한 특징을 학습하여 시퀀스 레이블링(Sequence Labeling)으로 본문을 추출하였다. 또한 [12]에서는 추출해야 하는 텍스트 노드의 특징이 많다는 [11]의 단점을 보완하기 위해, 특징의 종류를 줄이고 특징 추출 자동화를 위해 텍스트 노드의 부모 태그 및 단어 분포로만 구성된 입력 데이터를 이용한 학습 방법을 제안하였다. 본 연구의 이전 연구에서는 본문 텍스트가 비본문 텍스트에 비해 웹 페이지에서 출현 빈도가 낮다는 특징을 이용하여 기존의 단어/링크 밀도 [5-7]나 시퀀스 레이블링[12]를 이용한 방법보다 우수한 성능으로 본문을 분류하였다[13].

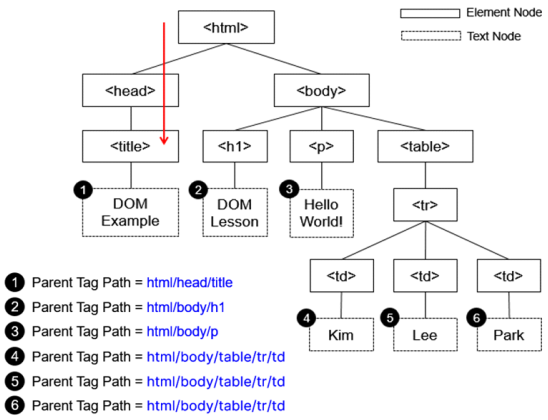

Fig. 1 Parent tag paths for 6 text nodes

그러나 이런 방법들은 수집 채널에 따라 성능 차이가 발생한다는 단점이 있다. 따라서 본 논문에서는 이전 연구에서 제안한 텍스트 출현 빈도뿐만 아니라 텍스트 노드의 부모 태그 경로를 이용하여 다양한 수집 채널에서 높은 정확도로 본문을 추출하는 방법을 제안하였다.

Ⅲ. 태그 경로와 텍스트 빈도 분석을 이용한 HTML 본문 추출

3.1. 부모 태그 경로 분석

DOM 트리는 웹 페이지를 트리 구조로 나타내기 때문에 최상위 태그부터 깊이 우선(depth-first)으로 탐색하여 그림 1과 같이 텍스트 노드의 부모 태그 경로(Parent Tag Path: PTP)를 분석할 수 있다. 이렇게 부모 태그 경로 패턴을 이용하면 본문과 비본문 텍스트 노드를 구분할 수 있다. 그리고 이때 본문 및 비본문 텍스트 노드 모두에 중복되는 부모 태그 경로 패턴이 존재할 수도 있다. 만약 부모 태그 경로 패턴이 중복되지 않는다면 부모 태그 경로 패턴을 이용하여 웹 페이지로부터 본문을 정확히 추출할 수 있을 것이다.

그림 2는 트위터와 인스타그램에서 수집한 웹 페이지의 부모 태그 경로 패턴이 표본의 개수에 따라 어떻게 변화하는지를 나타낸 것이다. 트위터의 경우 그림 2의 상단 왼쪽 그래프에서 알 수 있듯이 전체 부모 태그 경로(All PTP)와 본문에 해당하는 부모 태그 경로(Contents PTP)의 패턴 수는 표본 수에 비례하여 증가하다가 표본 수가 1,000개 정도 되었을 때부터 큰 변화를 보이지 않았다. 반면 비본문에 해당하는 부모 태그 경로(Non-contents PTP) 패턴의 수는 표본 수의 증가와는 무관하게 20개 정도임을 알 수 있다. 인스타그램의 경우 그림 2의 하단 그래프에서 알 수 있듯이 표본의 수가 증가해도 전체적인 부모 태그 경로 패턴의 수는 변화가 없었다.

한편, 그림 2의 오른쪽 그래프에서 알 수 있듯이 트위터와 인스타그램 모두 본문 및 비본문의 부모 태그 경로 가운데 중복되는 패턴(Duplicated PTP)은 없었다. 이처럼 중복되는 부모 태그 경로 패턴이 존재하지 않으면 부모 태그 경로를 분석해서 웹 페이지로부터 본문과 비본문 텍스트 노드를 정확히 구분하는 것이 가능하다. 하지만 만약 트위터나 인스타그램과는 달리 부모 태그 경로

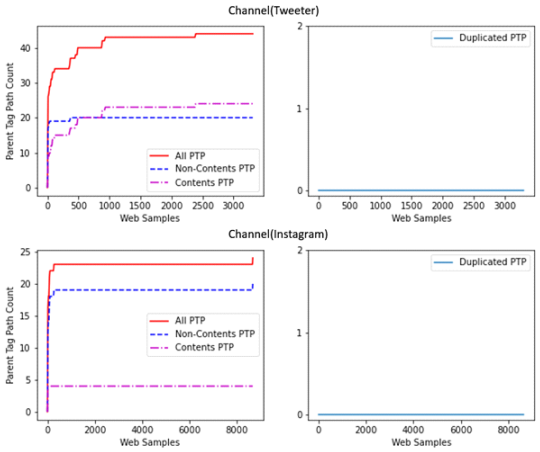

Fig. 2 Change in the number of PTP according to the number of web samples (Twitter, Instagram)

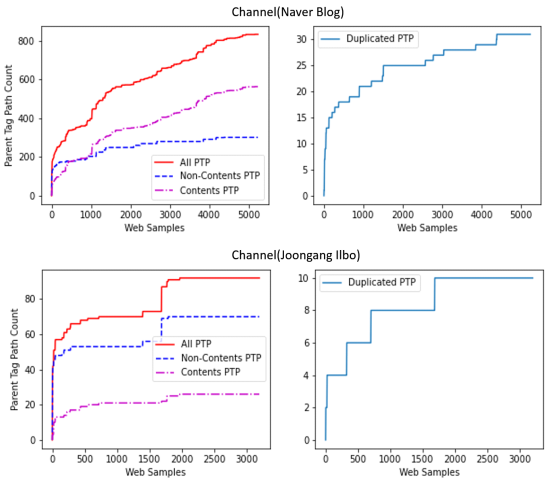

Fig. 3 Change in the number of PTP according to the number of web samples (Naver Blog, JoongAng Ilbo)

패턴이 중복되는 경우에는 부모 태그 경로만을 이용하여 본문을 정확히 구분하기 어렵다. 이를 확인하기 위해 본 연구에서는 본문을 다양한 템플릿으로 보여주는 국내 대표 블로그인 네이버 블로그와 뉴스 사이트 중 하나인 중앙일보에서 웹 페이지를 수집하였다.

그림 3은 네이버 블로그와 중앙일보에서 수집한 웹 페이지 표본 수에 따른 부모 태그 경로 패턴 수의 변화를 나타낸 것이다. 그림 3에서는 그림 2와 달리 표본의 수가 증가함에 따라 부모 태그 경로 패턴의 수도 증가함을 알 수 있다. 또한 중복되는 패턴 역시 표본의 수가 증가함에 따라 꾸준히 증가함을 확인할 수 있다. 이런 중

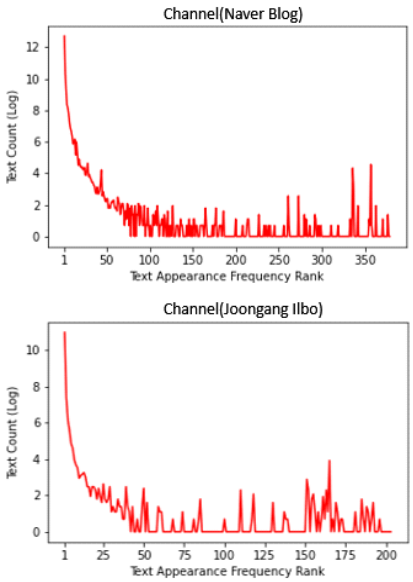

Fig. 4 Change in the number of texts according to text appearance frequency ranks

복 패턴으로 인해 블로그의 경우 중복되는 부모 태그 경로 패턴을 갖는 비본문 텍스트 노드의 비중은 전체의 6.7%, 중앙일보는 55.8%로 나타났다. 따라서 이 같은 경우 부모 태그 경로 패턴만을 이용한 본문 추출은 정확도가 떨어지는 한계가 있다.

3.2. 텍스트 출현 빈도 분석

네이버 블로그나 중앙일보처럼 중복되는 부모 태그 경로 패턴이 존재하는 경우의 문제점을 보완하기 위해 본 연구에서는 텍스트 출현 빈도(Text Appearance Frequency: TAF) 분석을 추가로 활용하였다.

텍스트는 내용과 표현, 목적에 따라 출현 빈도가 다르게 나타난다. 안내나 지시문구, 광고와 같은 비본문 텍스트는 서비스 이용의 편의성을 도모하거나 홍보 목적을 달성하기 위해 최대한 많은 이용자들에게 자주 노출되어야 하므로, 여러 웹 페이지에서 빈번하게 나타난다. 반면 본문에 해당하는 텍스트는 여러 서비스 이용자들에 의해 작성되기 때문에 내용과 표현이 매우 다양하므로 동일한 텍스트가 나타나는 경우는 매우 드물다.

그림 4는 네이버 블로그와 중앙일보의 웹 페이지 표본에서 분석한 텍스트 출현 빈도별 텍스트의 개수를 나타낸 것이다. 그림 4에서 x축은 출현 빈도 순위를, y축은 로그 스케일로 나타낸 텍스트의 개수를 의미한다. 이때 출현 빈도 순위는 출현 빈도가 낮을수록 높은 순위가 부여된다. 예를 들어, 출현 빈도가 1, 100, 200이라면 출현 빈도 순위는 차례대로 1, 2, 3이 된다. 그림 4에서는 출현 빈도 순위가 낮은 구간에 상당히 많은 텍스트가 분포하고 있음을 알 수 있다.

그림 5는 네이버 블로그에서 텍스트 출현 빈도의 구간별 텍스트 개수의 변화를 나타낸 그래프와 출현 빈도별 텍스트의 예시를 표현한 것이다. 출현 빈도가 낮은 구간에 분포된 텍스트는 대체적으로 문장의 길이가 길고 내용이 어떤 주제와 연관되어 있는 본문에 해당된다는 것을 알 수 있으며, 출현 빈도가 높은 구간에 분포된 텍스트는 어떤 기능에 대한 안내나 지시 문구 등에 해당하는 비본문임을 알 수 있다. 이처럼 출현 빈도가 낮은 구간에서 본문 텍스트 노드의 비중이 높다는 특징을 이용하면, 동일한 부모 태그 경로 패턴을 갖는 텍스트 노

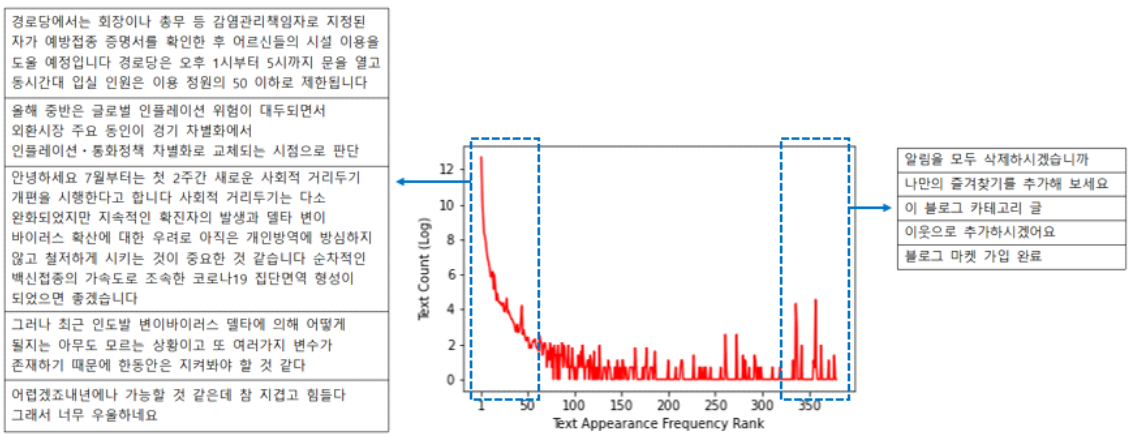

Fig. 5 Text details according to section of text appearance frequency

드의 본문 구분 성능을 크게 향상시킬 수 있다. 본 연구에서는 실제로 네이버 블로그의 5,233개 웹 페이지와 중앙일보의 3,201개 기사의 웹 페이지를 분석하여 동일한 부모 태그 경로 패턴을 갖는 텍스트 노드 중에서 상위 5순위까지의 출현 빈도를 갖는 본문 텍스트 노드의 비율을 확인하였다. 그 결과 본문 텍스트 노드의 비율이 네이버 블로그는 92.6%, 중앙일보는 86.2%로 나타났으며, 결과적으로 본문 텍스트 노드는 출현 빈도가 낮은 구간에 집중적으로 분포하고 있음을 확인하였다. 따라서 보다 정확한 본문 분류를 위해 어떤 출현 빈도 순위까지를 본문 텍스트로 분류할 지 결정할 필요가 있다.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \quad (1)$$

본 연구에서는 본문 텍스트로 분류할 출현 빈도 순위를 결정하기 위해 식 (1)과 같이 평균제곱오차(Mean Squared Error: MSE)를 이용하였다. N은 웹 페이지에서 추출된 텍스트 노드의 개수, Y는 텍스트 노드의 실제 본문 여부, $\hat{Y}$은 출현 빈도 순위에 따라 결정된 본문 여부를 나타낸다. 텍스트 노드의 본문 여부는 본문일 때 1, 비본문일 때는 0으로 표현하였다.

앞서 언급한 5,233개의 네이버 블로그 웹 페이지에서는 112 이하인 출현 빈도 순위를 갖는 텍스트 노드를 본문으로 분류하였을 때 오차가 0.004로 가장 낮았고, 정확도는 99.6%로 가장 높게 나타났다. 3,201개의 중앙일보 웹 페이지의 경우 149 이하인 출현 빈도 순위를 갖는 텍스트 노드를 본문으로 분류하였을 때 오차가 0.045로 가장 낮았고, 정확도는 93.6%로 가장 높게 나타났다. 특히 단지 낮은 출현 빈도를 갖는 텍스트 노드만을 본문으로 분류하였을 때와는 달리, 해시태그나 자주 사용되는 단어나 관용 표현까지 본문으로 분류할 수 있기 때문에 보다 세밀한 본문 추출이 가능하다는 것을 확인할 수 있었다.

Ⅳ. 성능 분석

4.1. 실험 데이터

실험을 위해 표 1과 같이 네이버 블로그와 중앙일보, 트위터, 인스타그램에서 2020년 3월 1일부터 2021년 6월 30일 기간 중에 작성된 '코로나 백신'과 관련된 웹 페이지 가운데 일부 웹 페이지에서 텍스트 노드를 추출하여 사용하였다. 전체 텍스트 노드 가운데 80%는 부모 태그 경로 패턴과 텍스트 출현 빈도의 임계값을 분석하기 위한 용도로 이용하였으며, 나머지는 성능 검증에 활용하였다.

Table. 1 Dataset per collection channels

Channel	Web pages	Total text node	Train text node	Test text node
Naver Blog	5,233	1,380,526	1,104,421	276,105
Joongang Ilbo	3,201	967,197	773,757	193,440
Tweeter	3,313	277,194	221,756	55,438
Instagram	8,664	406,702	325,362	81,340

4.2. 성능 평가 방법

본 논문에서 제안한 방법과 텍스트 출현 빈도만을 이용한 이전 연구[13]의 성능 비교를 위해, 텍스트 노드를 본문/비본문으로 분류하고 표 2와 같이 TP, FN, FP, TN의 네 가지 유형으로 분류하였다.

Table. 2 F-measure confusion matrix

		Actual Class	
		Contents	Non Contents
Predict Class	Content	TP (True Positive)	FP (False Positive)
	Non Contents	FN (False Negative)	TN (True Negative)

TP는 분류 모듈이 본문으로 분류한 텍스트 노드가 실제 본문인 경우의 수, FN은 분류 모듈이 비본문으로 분류한 텍스트 노드가 실제 본문인 경우의 수를 의미한다. 또한 FP는 분류 모듈이 본문으로 분류한 텍스트 노드가 실제 비본문인 경우의 수, TN은 분류 모듈이 비본문으로 분류한 텍스트 노드가 실제 비본문인 경우의 수를 의미한다.

본 연구에서는 재현율(Recall)과 정밀도(Precision), $F_1 - Score$ 등의 평가지표를 이용하여 성능을 측정하였다[14]. 해당 지표들은 데이터셋의 특정 클래스의 비중이 높더라도 성능을 정확하게 측정할 수 있다는 장점이 있기 때문에 분류 문제의 평가에서 광범위하게 이용되고 있다[6]. 특히, $F_1 - Score$는 식 (2)과 같이 재현율과 정밀도의 조화평균으로 계산하기 때문에 실제 본문의

수집률과 분류 모듈의 본문 수집 정확도를 반영하여 성능을 측정할 수 있다. 이때 재현율과 정밀도는 각각 식 (3)과 식 (4)과 같이 구할 수 있다.

$$F-Measure = \frac{2\times \mathrm{Precision}\times Recall}{\mathrm{Precision}\times Recall} \quad (2)$$

$$\text{재현율}(Recall) = \frac{TP}{TP+FN} \quad (3)$$

$$\text{정밀도}(\mathrm{Precision}) = \frac{TP}{TP+FP} \quad (4)$$

4.3. 성능 비교

본 논문에서 제안한 부모 태그 경로와 텍스트 출현 빈도를 모두 활용한 본문 추출 방법(PTP+TAF)의 성능을 평가하기 위해, 부모 태그 경로(PTP)와 텍스트 출현 빈도(TAF)를 각각 이용한 본문 추출 성능을 측정하여 비교하였으며, 그 결과는 표 3과 같다.

Table. 3 Performance comparison

Channel	Method	Recall (%)	Precision (%)	F1-Score (%)
Naver Blog	PTP	100.0	77.9	87.5
	TAF	82.9	71.1	76.6
	PTP+TAF	99.7	98.8	99.2
Jongang Ilbo	PTP	100.0	50.7	67.3
	TAF	94.1	71.2	81.1
	PTP+TAF	100.0	88.9	94.1
Tweeter	PTP	100.0	100.0	100.0
	TAF	97.9	84.5	90.7
	PTP+TAF	100.0	100.0	100.0
Insta-gram	PTP	100.0	100.0	100.0
	TAF	91.5	80.8	85.8
	PTP+TAF	100.0	100.0	100.0

네이버 블로그의 경우 PTP+TAF는 PTP보다 $F_1-Score$가 11.7%, TAF보다 22.6% 향상되었으며, 중앙일보의 경우 TAF보다 PTP+TAF를 이용하였을 때 13% 성능이 향상되었다.

본문 및 비본문의 부모 태그 경로 가운데 중복된 부모 태그 경로 패턴이 존재하지 않는 트위터와 인스타그램의 경우, PTP만으로도 PTP+TAF와 마찬가지로 정확하게 본문 텍스트 노드를 분류하였고, TAF보다 각각 9.3%, 14.2% 성능이 향상되었다.

결과적으로 본 논문에서 제안한 PTP+TAF의 성능이 모든 채널에서 우수하게 측정되었을 뿐만 아니라, 수집 채널별 성능 편차 역시 다른 방법에 비해 큰 차이가 없는 것으로 나타났다.

Ⅴ. 결 론

본 논문에서는 부모 태그 경로와 텍스트 출현 빈도를 이용하여 다양한 수집 채널에 적용할 수 있는 본문 추출 방법을 제안하였으며, 다양한 채널에서 수집된 웹 페이지를 대상으로 성능을 비교하였다. 그 결과 트위터와 인스타그램처럼 본문과 비본문의 부모 태그 경로 패턴이 중복되지 않는 수집 채널에서는 부모 태그 경로 분석만을 이용해도 본문 추출 성능이 우수하였고, 네이버와 중앙일보처럼 본문과 비본문의 부모 태그 경로 패턴이 중복되는 수집 채널에서는 부모 태그 경로 패턴과 텍스트 출현 빈도를 동시에 이용했을 때 본문 추출 성능을 크게 향상되었음을 확인하였다. 따라서 부모 태그 경로 패턴과 텍스트 출현 빈도를 동시에 이용하면 다양한 수집 채널에서 높은 성능으로 웹 페이지의 본문 추출이 가능할 것으로 기대된다. 향후 본 논문에서 제안한 방법을 적용하여 빅데이터 수집 시스템을 구현할 계획이다.

References

[1] Y. J. Kim, H. S. Kim, and H. S. Kim, “Understanding the Effects of COVID-19 on the Starbucks Perception through Big Data Analytics: A Comparative Study,” *Culinary Science & Hospitality Research*, vol. 27, no. 6, pp. 276-279, 2021.

[2] Y. R. Suh, K. P. Koh, and J. W. Lee, “An analysis of the change in media’s reports and attitudes about face masks during the COVID-19 pandemic in South Korea: a study using Big Data latent dirichlet allocation (LDA) topic modelling,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 25, no. 5, pp. 731-740, 2021.

[3] C. H. Lee, K. H. Kang, Y. H. Kim, H. N. Lim, J. H. Ku, and K. H. Kim, “A Study on the Factors of Well-aging through

Big Data Analysis: Focusing on Newspaper Articles," *Journal of the Korea Academia-Industrial cooperation Society*, vol. 22, no. 5 pp. 354-360, 2021.

[4] J. H. Lee, "Building an SNS Crawling System Using Python," *Journal of the Korea Industrial Information Systems Research*, vol. 23, no. 5, pp. 61-76, 2018.

[5] C. Kohlschuer, P. Fankhauser, and W. Nejdl, "Boilerplate detection using shallow text features," in *Proceedings of the third ACM international conference on Web Search and Data Mining (WSDM)*, New York: NY, pp. 441-450, 2010.

[6] W. M. Song and M. G. Kim, "Contents Extraction from HTML Documents using Text Block Context," *Journal of KISS : Software and Applications*, vol. 40, no. 3, pp. 155-163, 2013.

[7] H. G. Jeon and C. Koh, "Text Extraction Algorithm using the HTML Logical Structure Analysis," *Journal of Digital Contents Society*, vol. 16, no. 3, pp. 445-455, 2015.

[8] J. H. Mo and J. M. Yum "Korean Web Content Extraction using Tag Rank Position and Gradient Boosting," *Journal of KIISE*, vol. 44, no. 6, pp. 581-586, 2017.

[9] S. Wu, J. Liu, and J. Fan, "Automatic Web Content Extraction by Combination of Learning and Grouping," in *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*, pp. 1264-1274, 2015.

[10] S. H. Kim and H. J. Kim, "Logistic Regression Ensemble Method for Extracting Significant Information from Social Texts," *KIPS Transactions on Software and Data Engineering*, vol. 6, no. 5, pp. 279-284, 2017.

[11] T. Vogels, O. E. Ganea, and C. Eickhoff, "Web2text: Deep structured boilerplate removal," in *Proceedings of the 40th European Conference on Information Retrieval*, pp. 167-179, 2018.

[12] J. Leonhardt, A. Anand, and M. Khosla, "Boilerplate Removal using a Neural Sequence Labeling Model," in *Companion Proceedings of the Web Conference 2020 (WWW '20)*, New York: NY, pp. 226-229, 2020.

[13] J. H. Kim and E. G. Kim, "HTML Text Extraction Using Frequency Analysis," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 25, no. 9, 2021.

[14] Tharwat, A, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17 no. 1, pp. 168-192, 2021.

김은경(Eun-Gyung Kim)

1983년 2월 : 숙명여자대학교 물리학과 졸업
1986년 2월 : 중앙대학교 전자계산학과 석사
1991년 2월 : 중앙대학교 컴퓨터공학과 박사
1992년 3월~ 현재 : 한국기술교육대학교 컴퓨터공학부 교수
※관심분야 : 빅데이터 분석, 딥러닝, 트리즈 등

김진환(Jin-Hwan Kim)

2016년 2월 : 한국기술교육대학교 컴퓨터공학부 졸업(학사)
2019년~현재 : 한국기술교육대학교 컴퓨터공학과 석사과정
※관심분야 : 빅데이터, 텍스트마이닝, 웹 크롤링, 기계학습