한국컴퓨터정보학회 하계학술대회 논문집 제26권 제2호 (2018. 7)

웹크롤링을 활용한 뉴스 어뷰징 추론 모델

정경록[O], 박구락[*], 정영석[*], 남기복[*]
[O*]공주대학교 컴퓨터공학과
e-mail: rokk1969@kongju.ac.kr[O], {ecgrpark, merope}@kongju.ac.kr[*], mtgood@naver.com[*]

News Abusing Inference Model Using Web Crawling

Kyoung-Rock Chung[O], Koo-Rack Park[*], Young-Suk Chung[*], Ki-Bok Nam[*]
[O*]Dept. of Computer Science & Engineering, Kongju National University

● 요 약 ●

기존 신문이나 티브이가 아닌 온라인과 모바일로 뉴스를 보는 사람이 더 많아지면서, 포털 사이트 뉴스난에 다른 언론사의 기사보다 더 많이 노출되기 위한 경쟁의 심화로 뉴스 어뷰징은 심각한 사회 문제로까지 대두되었다. 본 논문은 온라인상에서 생성, 유통되는 많은 뉴스 중에서 이용자의 시간을 낭비하고 양질의 정보를 찾기 힘들게 하는 뉴스 어뷰징을 판단하는 모델을 제안한다. 제안된 모델은 크롤링 기술을 사용하여 뉴스의 제목과 내용을 가져온 후 인공지능 기술을 이용한 유사도 검사로 기사의 어뷰징 여부를 판단하여 양질의 뉴스 정보를 사용자에게 제공될 수 있다.

키워드: 어뷰징(abusing), 웹크롤링(web crawling), 뉴스(NEWS), DOM(Document Object Model)

I. Introduction

매일매일 새로운 많은 뉴스가 온라인상에 유통되고 있는데 국내에는 약 6천여 개의 인터넷 신문사가 있고 그중에 100개 정도만이 언론으로서의 제 기능을 하는 것으로 알려져 있다.[1]

여러 신문사가 실시간으로 내용이 대동소이한 뉴스로 포털 사이트 뉴스난에 노출되기 위해 경쟁한다. 이런 많은 뉴스 중에는 광고료로 수익을 올리기 위해 클릭 수와 높은 트래픽 확보를 위해 다소 선정적인 머리기사로 내용과 일치하지 않는 어뷰징 기사도 많아 사용자 다수가 불편함을 느끼고 있으며 뉴스 본문이 광고 배너들로 도배되어 정작 뉴스는 알아보기 힘들다.

본 논문은 뉴스의 어뷰징을 판단하는 모델을 제안한다. 제안된 모델은 웹 크롤링을 활용하여 뉴스의 어뷰징 여부를 판단하며, 사용자에게 양질의 정보를 제공할 수 있다.

II. Related works

1. 뉴스 어뷰징(NEWS Abusing)

'남용·오용'이 사전적 의미인 어뷰징(Abusing)은 인터넷에서 기사의 수정과 업데이트가 쉽게 이루어지고, 기사 제목을 클릭하기 전에는 기사 내용을 파악할 수 없다는 점 때문에 일어난다. 서울중앙지방법원(2011)은 어뷰징을 "같은 내용의 뉴스 기사 반복 전송"이라고 전제하고 "뉴스 기사 검색 횟수를 정당하지 않은 방법으로 늘리기 위해 실질적으로 같은 뉴스를 작위적 제목 변경 또는 일부 내용 변경한 뉴스 기사 재송신 행위"로 정의했다.[2]

2. 웹 크롤링(Web Crawling)

웹 크롤링이란 웹 사이트에서 원하는 정보를 추출하는 것을 의미한다. 웹은 기본적으로 HTML 형태로 되어 있다. 소스는 개발자가 정형화된 형태로 관리하기 때문에 일정한 규칙이 있으며, 이러한 규칙을 분석해서 원하는 정보들만 뽑아오는 것이 웹 크롤링 작업이다.[3]

3. HTML DOM

DOM은 웹 페이지를 구성하는 모든 요소의 구조 정의다. 화면에 보이는 요소 내부 구조에 접근할 수 있도록 정의해 놓은 것이며 비동기식으로 처리한 데이터를 DOM을 사용해 동적으로 화면에 접근시킬 수 있어 비동기식 자바스크립트에 매우 중요하다. DOM을 사용하면 동적으로 변경할 수 있고 해당 페이지에 접근, 조작할 수 있다.[4] DOM은 HTML에 있는 스타일, 값 등을 객체로 통합하여 접근할 수 있게 하며 화면에 보이는 모든 내용과 해당 객체의 값을 조작할 수 있고 이러한 기능을 담고 있는 모델을 DOM TREE라고 한다. DOM NODE는 트리 구조 중 하나의 요소로 DOM TREE를 이루고 있는 가장 기초적인 단위를 나타낸다.[5]

III. The Proposed Scheme

본 논문은 크롤링 기술을 이용하여 뉴스의 제목과 내용을 취득하여,

어뷰징 뉴스를 판단하는 모델을 제안한다.

<Fig. 1.>은 논문에서 제안하는 어뷰징 뉴스를 판단하기 위한 크롤링 시스템 구성도이다.

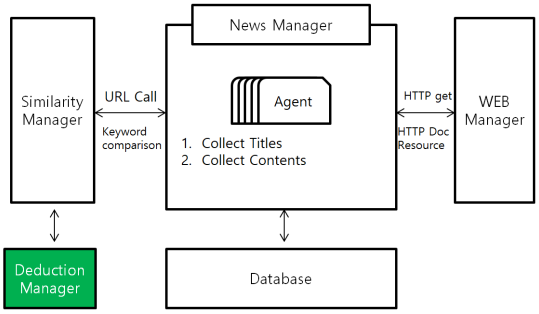

Fig. 1. Configuration

제안된 모델은 4개의 매니저로 구성하였다.

첫 번째, News Manager는 멀티로 크롤링을 처리하기 위해 뉴스의 제목과 내용만을 수집한다.

두 번째, WEB Manager는 크롤링하는 웹사이트를 방문해 HTML DOM Resource를 가져온다.

세 번째, Similarity Manager는 수집한 제목과 내용으로 유사도를 검사한다.

네 번째, Deduction Manager는 유사도 검사 결과로 뉴스 어뷰징 여부를 추론한다.

IV. Conclusions

포털 사이트에서 더 많이 노출되기 위한 경쟁 속에 뉴스에 다소 선정적인 머리기사가 생성되고 있다. 뉴스 어뷰징은 온라인 뉴스를 이용하는 사용자에게 불편을 주고 있다. 과도한 뉴스 어뷰징은 하는 언론사의 뉴스는 사람들이 점점 피하게 되어 결국 어뷰징을 하는 언론사의 손해로 되돌아올 것이다.

본 논문은 크롤링을 활용하여 뉴스를 수집하고, 수집된 정보의 유사도 검사로 어뷰징을 판단하는 모델을 제안했다. 향후 논문에서는 본 연구에서 제안한 모델을 구현하겠다.

REFERENCES

[1] ByoungHee Kim, "An Exploratory Study of the Affecting Factors for Preventing Abusing of News Articles When Developing Portal's Search Algorithm," Communication Theories, Vol. 11, No. 3, pp. 47~89, 2015.

[2] Moonki Hong and Byoung Hee Kim, "Institutional and Technological Approaches for Effective Damage Relief for Internet News Abusing," The Korean Journal of Advertising and Public Relations, Vol. 18, No. 2, pp. 112~147, 2016,

[3] https://m.blog.naver.com/potter777777/220605598446

[4] So-jin Nam, Do-Hoon Kim, Wan-Jung Kim, Yong-Hyuk Kim, "Dom-based Content Extraction for Improving Performance of Web Service", Korea Information Science Society, pp.218-222,Oct 2003.

[5] Kim Kyuheon, Park JungWook, Kim Byungchul, "Effective Method to Change Multimedia Scene Configuration Information Using DOM Update", Journal of Broadcast Engineering, Volume 18(1), pp. 43-58, 2013