광고성 웹 게시글 판단 프로그램

배지선*, 오예림**, 김채원*, 박지원***, 홍진근***, 윤형기****
*백석대학교 컴퓨터공학부 정보보호학전공
**백석대학교 미디어선교학전공
***백석대학교 스마트IT공학부 빅데이터전공
****(주) 임팩트라인
bjs1138@bu.ac.kr, jem1610@bu.ac.kr , yescall29@bu.ac.kr, 20212076@bu.ac.kr,
jkhong@bu.ac.kr, hkyoon23@naver.com

Decision Program for Advertisement Web Posts

Ji-Seon Bae*, Ye-Rim Oh**, Chae-won Kim*, Ji-Won Park***,
Jin-Keun Hong***, Hyung-Ki Yoon****
*Dept. of Information Security, Division of Computer Engingeering, Baek-seok University
**Dept. of Mission & Media, Baek-seok University
***Dept. of Big Data, Division of Smart IT, Baek-seok University
****Impactline, Inc

요 약

흔히, 웹 플랫폼에서 검색했을 때, 게시글 마지막부분에 광고인지 여부를 판단 할 수 있는 관련 글들이 나타난다. 이 글들은 사용자의 판단력을 흐리게 할 수 있다고 판단되며 개선의 필요성이 제기된다. 따라서 본 논문에서는 사용자들에게 웹 게시글에서 나타나는 광고성 여부에 대해 신속한 판단이 가능하도록 하는 환경에 대한 연구를 하고자 한다. 본 논문에서는 게시글에 포함된 광고 관련 문구를 찾아 페이지 상단에 해당 정보를 제공하는 프로그램을 제작 게시함으로써, 광고여부를 판단할 수 있도록 하였다.

1. 서론

최근 유튜브 등 정보를 제공하는 플랫폼에서 '뒷광고' 논란이 화제가 되고 있다. 뒷광고는 업체 측으로 광고 요청을 받았지만 이를 알리지 않은 채 광고 효과를 제공하는 것이다. 영상 플랫폼과 함께 많은 검색은 웹 플랫폼에서 이루어지고 있다. 웹 플랫폼에서 정보에 대해 검색 시 이 글이 광고 글인지를 표시하여 사용자의 판단에 도움을 주는 프로그램에 대한 필요성이 대두된다.

사실 많은 검색이 웹 플랫폼에서 이루어지고 있다. 대부분 웹 게시글의 광고 여부는 글 마지막에 드러난다. 광고글은 객관적인 평가를 내리기 어렵고, 글을 끝까지 읽고 나서 광고 여부를 알게 되면 사용자는 시간 낭비라고 느끼게 된다. 만일 글을 끝까지 읽지 않는다면 해당 글이 광고인지 아닌지 알 수 없다. 따라서 본 논문에서는 사용자들에게 광고에 대한 신속한 판단이 가능하도록 하는 환경을 제공하는 것으로 목표로 연구되었다. 본 연구에서는 게시글에 포함된 광고 관련 문구를 찾아 페이지 상단에 해당 정보를 제공하는 프로그램으로 그 해결책을 제시하고자 한다[1-2].

본 논문의 구성은 다음과 같다. 2장에서 관련연구를, 3장에서 본 연구에서 제안하는 방법을, 그리고 4장에서 결론으로 맺고자 한다.

2. 관련연구

관련 연구에서 나철원 등은 최신 웹 크롤링 알고리즘 분석 및 선제적 크롤링 기업을 제안하였다[3]. 또한 이다예 등은 게시판 크롤링을 통한 선호도 기반 게시물 푸시 서비스를 구현하고 있다[4]. 최수렴 등은 OCR을 이용한 안드로이드 기반 텍스트 추출 및 검색 시스템을 구현하였다[5]. 그런데 이와 같은 연구들은 모두 게시글에 있는 광고성 문구를 찾아 사용자에게 전달하는 것, 이미지로 된 광고성 문구를 사용자에게 전달하는 것, 광고성 문구를 웹 페이지

상단에 띄우는 것과 관련되는 것이다. 그러므로 글과 이미지로 된 광고성 문구를 찾아 전달하는 것은 각각 웹 스크래핑 및 크롤링, 파싱으로 OCR, 이미지 제시의 방법으로 구현하는 것이다.

그런데 본 논문의 초점은 광고성 문구를 웹 페이지 상단에 띄우는 것으로 사용자의 접근성이 높은 크롬 확장 프로그램으로 구현한 것이 특징이다.[6]

3. 제안방법

본 논문에서는 다음 표1에서와 같은 개발 환경과 기술이 주로 적용되었다.

구분		개발환경	적용 기술
S/W 개발 환경	웹 스크래핑 및 파싱	JavaScript	게시글의 글을 크롤링 및 데이터 스크래핑
			스크래핑한 데이터를 명사 단위로 자연어 처리하고 의미 있는 문자열을 얻음
			이미지를 가져와 프로그램에 띄움.
	크롬 확장 프로그램	HTML	확장형프로그램 창 설정
		JavaScript	게시글에 포함된 광고성 문구와 이미지를 표시하는 프로그램을 제작
		CSS	사용자의 편의를 위한 프로그램 디자인

<표 1> 개발 환경 및 주요 적용 기술

3.1 게시글의 광고성 문구 획득 방법

본 논문에서 웹 게시글의 광고성 문구를 획득하는 방법은 크롤링 및 웹 스크래핑과 파싱 단계로 나뉜다. 크롤링 및 웹 스크래핑이란 웹 상에 게시된 정보를 데이터로 활용하기 위해 웹 페이지에 있는 내용을 가져오는 것을 말한다. 스크래핑은 데이터를 단순히 가져오는 것이기 때문에 양이 많고 정리되지 않은 문자열의 형태이다. 따라서 이것을 의미있는 정보의 형태로 얻는 파싱의 단계가 필요하다.

크롤링은 Python, JavaScript 등의 언어로 구현할 수 있다. 먼저 Python은 라이브러리 requests를 사용한다. 웹페이지의 본문 부분만을 가져오기 위해 태그언어 값을 타겟으로 주어 스크래핑한다. 파싱에는 BeautifulSoup을 사용한다.[7] JavaScript는 node.js를 활용하는 방법이 있다. axios와 cheerio를 모듈을 사용한다. axios를 사용해 url을 받아오고 cheerio에 내용이 load 되어서 파싱할 수 있게 된다. 또는 기본JavaScript와 크롬 확장 프로그램 개발과 접목시켜 bodyText로 스크래핑을, match로 파싱을 구현할 수 있다.[8]

3.2 이미지로 된 광고성 문구 획득 방법

또한 본 논문에서 이미지로 된 광고 문구를 얻는 방법은 이미지의 문자를 인식해 글자로 뽑아내는 방법이다. 이것을 OCR(Optical character recognition, 광학 문자 인식)이라고 한다. OCR을 구현하기 위해 광학 문자 인식 엔진인 Tesseract를 사용한다. Tesseract를 Python환경에서 구현했다. 먼저, Python의 Image 모듈을 사용해 url에서 이미지를 추출해 새로 저장한다. Tesseract를 실행한 후 저장한 이미지에서 pytesseract를 이용해 글을 추출한다.

4. 연구결과

본 논문에서는 사용자의 서비스 최적화를 위해 크롬 확장 프로그램을 제작했다. 웹 페이지 오른쪽 상단에 팝업창으로 구현했다. 광고 판단이 필요하지 않은 웹 사용 시 무의미한 프로세서 사용을 방지하기 위해 버튼을 누르면 기능이 작동하도록 하였다.

글로 된 광고성 문구를 획득하기 위해 JavaScript로 스크래핑과 파싱을 구현했다. 광고성 문구에 주로 쓰이는 단어를 배열에 저장해 게시글과 비교해 광고 문구 존재 여부를 판단한다. 게시글에 해당 단어가 존재한다면 그 단어를 저장해 html에 붙여서 사용자에게 시각적으로 확인할 수 있도록 한다.

이미지로 된 광고성 문구는 크롬 확장 프로그램과의 개발 환경 차이로 인해 OCR을 구현하지 않았다. 대신 이미지를 크롤링해 창에 띄워서 사용자에게 제시해 판단하도록 하는 방법을 사용했다.

(그림 1) 구현 결과 사진

사전 단어	빈도수	사전 단어	빈도수
원고료	20	업체로부터	10
무상	4	지급	6
소정의	10	무료	3
수수료		증정	2
지원금		제공받	5

<표2> 테스트 산출 빈도수

사전에 광고성 글들을 통해 모아본 특정 단어들을 사전 단어로 설정하였다. 위 표는 '샴푸'를 주제로 25개의 사이트에서 테스트하고 발견된 단어들의 빈도수를 나타낸 표이다.

5. 결론

본 논문에서는 사용자가 알고 싶어 하는 정보 글이 광고성 글인지 먼저 검사해 광고성 문구를 제시하였다. 사용자가 글을 읽기 전 판단에 필요한 정보를 상단에 제공함으로써 광고 여부인지를 판단할 때 효과적으로 할 수 있는 기능을 구현하였다. 이를 통해 일반 사용자들에게 스크롤을 내리지 않아도 광고성 유무를 판단할 수 있도록 시간 단축의 편리를 제공할 수 있도록 하였다. 사용자들은 유해한 광고로부터 간접적으로 보호받을 수 있다.

웹 상에서 기존에 제공되는 광고 차단 서비스는 광고 팝업 등 자체적인 광고를 사전에 제거하는 형태이다. 이러한 프로그램은 임의의 사용자가 작성한 글의 광고는 차단 불가하다. 본 연구의 차별점은 검색 설정으로 쉽게 걸러낼 수 있는 텍스트뿐만 아니라 이미지 형태의 광고성 문구도 제공하는 것이다. 또한 크롬 확장 프로그램이기 때문에 1회 설치만으로 별도 실행 없이 계속 사용이 가능하기 때문에 사용자의 편의와 접근성 면에서 높은 효율을 보인다.

※ 본 논문은 과학기술정보통신부
정보통신창의인재양성사업의 지원을 통해 수행한
ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] 이은순."“내돈내산”의 배신: 유튜브 뒷광고 논란으로 보는 뉴미디어윤리."한국방송학회 학술대회 논문집.(2020):55-57.

[2] 안휘훈, 박병준 (2014). "의견 마이닝을 위한 유사 광고성 상품평 추출." 대한전자공학회 학술대회, 1593-1596.

[3] 나철원, 온병원."최신 웹 크롤링 알고리즘 분석 및 선제적인 크롤링 기법 제안."인터넷정보학회논문지20.3(2019):43-59.

[4] 이다예, 박준홍, 이동욱, 전수빈."게시판 크롤링을 통한 선호도 기반 게시물 푸시 서비스 구현."한국정보과학회 학술발표논문집.(2020):1706-1708.

[5] 최수렴, 고은비, 하유진, 박영호 (2010). "OCR 을 이용한 안드로이드 기반 텍스트 추출 및 검색 시스템 구현." 한국멀티미디어학회 학술발표논문집, 469-473

[6] 조규철, 하진욱, 류성민. "Chrome Web Browser의 Chrome Extension을 활용한 웹 시스템 개발." 한국컴퓨터정보학회 학술발표논문집 25.2 (2017): 246-247.

[7] 승리, 윤수진, 우영운."파이썬을 이용한 다양한 형식의 웹 데이터 크롤링 기법."한국정보통신학회 종합학술대회 논문집23.2(2019):343-346.

[8] 이혜규, 김성희, 우영운."대규모 문서 분석을 위한 분석 라이브러리 연동 웹 서비스."한국정보통신학회 종합학술대회 논문집24.2(2020):305-307.