체감 안전도 향상을 위한 뉴스 빅데이터 분석 연구: 범죄율과 보도량의 왜곡현상을 중심으로

정준호*, 이병도**, 장종욱*, 손윤식*
*동국대학교 컴퓨터공학과
**동국대학교 법공학연구센터
e-mail : sonbug@dongguk.edu

Big Data Analysis For Improving Citizens Safety Perception: Focusing on Distortion Phenomenon Between Crime Rates and Numbers of News Articles

Junho Jeong*, Byung Do Lee**, Jong.wook Jang*, Yunsik Son*
*Dept of Computer Science and Engineering, Dongguk University
**justice Engineering Convergence Research Center, Dongguk University

요 약

오늘날 안전한 사회를 위한 시민 체감 안전도의 향상이 매우 중요한 사회적 이슈가 되고 있다. 그중에서도 많은 시민들이 온라인 뉴스를 통해 오늘날 주요 사회 범죄와 그 위험도를 체감하고 있다. 하지만 이 온라인 뉴스를 통해 보도되는 범죄들은 국내에서 발생하는 범죄의 통계적 비율과는 다소 많은 차이가 있다. 따라서 실제 범죄 위험도와는 상이한 시민들의 체감 안전도가 형성된다는 문제가 있다. 본 논문에서는 체감 안전도 향상을 위한 온라인 뉴스로부터 데이터를 수집하고 정제하여 범죄 발생율과 보도량이 왜곡되는 현상을 빅데이터 분석을 통해 연구하기 위한 모델을 제시하고자 한다. 또한 기초적인 빅데이터 수집 및 분석의 결과를 바탕으로 실제 경찰청에서 제공되는 범죄율과 비교하여 시민들의 체감 안전도의 왜곡현상을 살펴보고자 한다.

1. 서론

오늘날 안전한 사회를 위한 시민 체감 안전도의 향상이 매우 중요한 사회적 이슈가 되고 있다. 실질적으로 많은 시민들이 온라인 뉴스를 통해 주요 사회 범죄와 그 위험도를 체감하고 있다. 하지만 국내에서 발생하는 범죄의 실제 통계적 비율과 온라인에서 나타나는 범죄 뉴스의 보도량에는 많은 차이가 있다.

예를 들어 온라인 뉴스에서는 인기를 위해서 일반적인 단순 범죄보다는 자극적인 기사를 자주 보도할 수 있는 특수 범죄의 비율이 높은 편이다. 그에 따라 경찰청 범죄 통계에서 발생건수가 적은 '살인'이나 '강도'와 같은 범죄에 대해서 더 자세하고 정제되지 않은 많은 기사를 전파하여 오히려 시민들의 체감 안전도를 낮추는 문제가 있다[10].

그에 반해서 '절도'에 대해서는 전체적인 보도량이 매우 낮으며 오히려 최근에는 한국이 얼마나 절도에 안전한 곳인지를 확인하기 위한 다양한 실험들이 SNS를 통해 알려짐에 따라 많은 시민들이 '절도'에 대한 체감 안전도가 상대적으로 높게 인지되고 있다고 할 수 있다[11].

이와 같이 언론으로부터 보도되는 범죄별 보도량은 시민들의 체감 안전도에 큰 영향을 주고 있으며 실제 통계자료와 다를 수 있기 때문에 시민들의 체감 안전도가 왜곡되는 문제가 있다. 본 논문에서는 체감 안전도 향상을 위해 온라인 뉴스의 범죄 보도들을 크롤러를 통해 수집하고 필터링을 통해 정제된 데이터에 대해서 분석한 범죄 보도율과 실제 범죄 발생률 통계를 비교 분석한 결과를 대시민 서비스로 제공하기 위한 분석 모델을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서 대시민서비스를 제공하기 위한 분석 모델을 제시하고, 3장에서는 수집된 데이터의 데이터 분석 결과를 소개하며 마지막으로 4장에서 결론을 맺는다.

2. 체감 안전도 향상을 위한 범죄 뉴스 빅데이터 분석을 이용한 대시민서비스 모델

본 논문에서 제안하고자 하는 대시민서비스 모델을 구축하기 위한 데이터의 분석 프로세스는 그림 1과 같다.

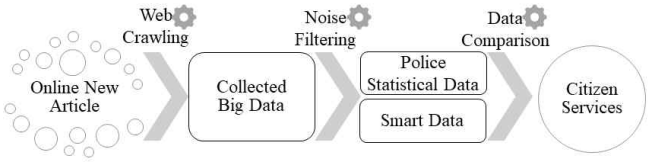

그림 1. 체감 안전도 향상을 위한 데이터 분석 프로세스

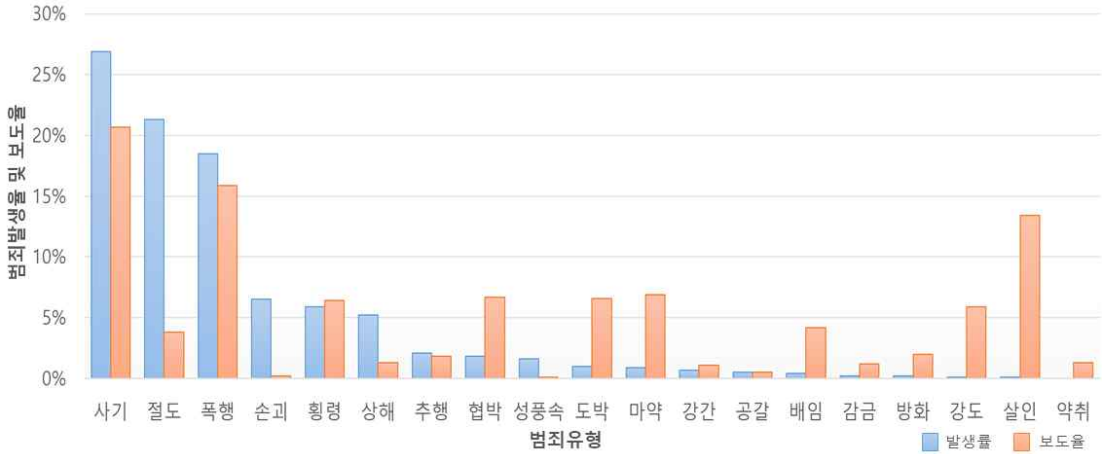

그림 2. 범죄유형에 따른 범죄발생률과 보도율

그림 1과 같이 온라인인 뉴스 기사를 웹크롤링을 이용하여 수집하여 빅데이터를 구축하고, 빅데이터에 존재하는 노이즈들을 필터링하여 스마트 데이터로 생성하는 일반적인 웹 기반의 빅데이터 분석 모델[12]의 결과를 경찰청에서 발행하는 공식적인 통계결과와 비교하여 그 영향을 분석하고자 한다. 이 모델에서 중요한 것은 웹 크롤링을 통한 수집된 빅데이터의 정확도와 이를 노이즈 필터링을 통해 정제된 스마트데이터의 유효함이다.

제안 연구에서는 특정 키워드가 포함된 기사를 웹 크롤러를 통해 수집하였다. 특정 키워드는 범죄유형단어(살인, 강도 등)로 이 범죄유형단어들은 경찰청에서 제공하는 2017년도 경찰통계연보의 주요 죄종별 발생·검거 통계자료에서 선정하였고 살인, 강도, 강간, 추행, 방화, 절도, 상해, 폭행, 감금, 협박, 약취, 공갈, 손괴, 사기, 횡령, 배임, 도박, 성풍속, 마약 이렇게 19가지 단어들이 기준이다.

그리고 제안 연구의 유용성을 분석하기 위해서 국내에서 가장 많은 사용자가 있는 네이버 포털을 대상으로 2017년에 보도된 기사에 대해서 주요 범죄유형단어들을 이용하여 키워드 별로 수집한 결과 약 120 만 건의 빅데이터를 수집하였으며 제목, 내용, 작성일 등의 패턴에 따라 기초 필터링이 완료된 스마트 데이터 약 70만 건을 수집하여 2017년도 경찰청통계연보와 비교 분석하였다.

3. 범죄 위험도 왜곡현상 분석

그림 2는 범죄유형에 따른 경찰청에서 제공하는 2017년도 주요 죄종별 발생 건수와 온라인에서 수집된 범죄 보도 데이터에서 범죄유형에 따른 수를 비율로 표현한 것이다.

경찰청 통계를 기반으로 실질적으로 가장 높은 범죄의 비율은 사기, 절도, 폭행 순이었고, 반대로 강도, 살인, 약취 순으로 사건 발생률이 낮았다는 점을 알 수 있다. 그러나 범죄 보도율은 실제 통계와는 다르게 살인, 협박, 도박, 마약과 같은 자극적인 범죄에 대한 비중이 월등이 높은 것을 확인 할 수 있었다.

그리고 절도의 경우에는 통계적으로는 두 번째로 높은 발생 비율을 가지고 있는 범죄이나 기사로 언급되는 것은 19 개 범죄 중 10 번 째 정도밖에 되지 않았다. 이와 같은 현상은 생명에 위협이 되거나 자극적인 요소가 중복 확대 재생산 되는 기사들의 특징으로 발생하는 것으로 판단되며 시민들의 체감 안전도를 낮추는 주된 요인으로 판단된다.

4. 결론 및 향후 연구

본 연구는 범죄 뉴스의 빅데이터 분석을 통한 시민들의 체감 안전도를 높이기위한 대시민서비스 모델을 제안하였다. 또한 기초적인 필터링을 통해 얻어진 스마트 데이터를 이용하여 통계자료와 비교 분석하여 실질적인 범죄 위험도와의 왜곡이 발생하여 시민의 체감 안전도가 낮아지는 근본적인 원인을 분석하였다. 하지만 수집된 데이터의 필터링에서 기사의 내용을 고려하지 않아 긍정오류가 다수 존재하는 것을 확인 할 수 있었다.

향후 연구로서 기사의 내용을 고려한 필터링을 통해 스마트데이터의 정확도를 높여, 전체적인 분석 모델의 정확도를 높이기 위한 연구를 수행하고, 이를 통해 뉴스의 트렌드가 실제 범죄 발생과는 다를 수 있다는 점을 안내하여 대시민 체감 안전도를 향상시키기 위한 서비스를 구축하고 제공하고자 한다.

감사의 글

이 성과는 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행되는 연구임(No.2018R1A5A7023490).
이 성과는 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행되는 연구임(NRF-2017R1D1A3B03029906).

참고문헌

[10] 박형민, 이민아, “강력사건 및 자살에 대한 언론보도의 실태와 문제점,” 제 2009 권, 12호 pp. 13-137, 2009.
[11] KTV국민방송, “ ‘한국은 얼마나 안전할까?’ 실험하는 외국인 영상,“ https://www.youtube.com/watch?v=vT4u7bGlH0c”, 2017.
[12] D. García-Gil, J. Luengo, S. García, and F. Herrera, “Enabling smart data: noise filtering in big data classification,” Information Sciences, Vol. 479, pp. 135-152, Apr. 2019.