SNS 크롤링 기반 사건 발생의 지수화 연구

민지인[1], 정민수[1O], 유도근[2], 홍승혁[1*]

[1]수원대학교 데이터과학부

[2]수원대학교 토목공학과

qpflwldls@naver.com, alstnwjd0424@naver.com,
dgyoo411@suwon.ac.kr, *shongdr@gmail.com

서론

2019년 5월 30일 인천시 일대 수돗물에서 적수, 유충이 보인다는 민원이 제기됐다. 위 사태는 소셜 네트워크 서비스(SNS)를 통해 물 공급 과정에 대한 부정적 인식을 초래했다. 따라서 수질사고를 빠르게 인지하는 방법론을 통해 피해를 최소화 하는 노력이 필요로 한다.

본 연구는 물 공급과정 사고 분석을 위한 SNS 웹 크롤링을 제안, 분석한다. 웹 크롤링이란 웹 페이지에서 자료를 수집하는 것이고 정형화된 규칙 소스를 분석해 원하는 정보만을 추출하게 된다. 이때 핵심 키워드(수돗물, 적수) 조합을 활용하여 검색량 정도를 지표로 표현하여 유사도와 파급력 정도를 파악하였다.

연구 방법

연구 대상

본 연구는 필요 대상에 대한 SNS별 웹 크롤링 가능 여부와 정보 획득 기간을 확인하여 물 공급 과정에서 발생하는 수질 사고를 연구 대상으로 선정하였다. 과거 사건 발생의 관련도가 높은 네이버와 트위터를 중심으로 웹 크롤링을 진행한다. 네이버, 트위터에 검색할 키워드는 행정구역 명과 핵심 키워드(수돗물, 적수)를 함께 검색한다.

연구 기법

Beautiful Soup

HTML, XML 파일에서 원하는 데이터를 쉽게 가져오기 위한 Python 라이브러리이다. 지정한 html을 가져오고 'html parser' 를 통해 Beautiful Soup의 인자 값을 지정해 주었다. 네이버에서 제목, 신문사, 기사 본문, 링크(관련도 순)를 가져왔다. 본문 정제화를 위해 관련 함수를 활용했고 상대적으로 검색 결과가 많은 네이버에서는 100페이지 까지 크롤링 하였다.

TfidVectorizer

TF-IDF 값은 특정 단어의 상대적인 빈도수를 나타내는 값으로써 클수록 현재 문서에는 자주 언급되면서 다른 문서에는 잘 언급되지 않음을 뜻한다. 각 매체의 크롤링 결과를 수치화 하여 지표로 표현하였다.

Cosine similarity

두 벡터의 유사도를 수치로 나타낸 것이다. 트위터와 네이버의 결과를 코사인 유사도로 도출 후 평균하여 백분율로 표기하였다. 검색 기능의 정확도를 높이기 위해 지수화 과정을 포함하였다.

연구 결과

인천시는 서구 지역을 중심으로 확산한 붉은 수돗물(적수) 공급 사태를 해소하기 위해...
박흥서 기자 phs0506@ajunews.com 일주일째 발생하고 있는 인천지역 수...
인천의 붉은 수돗물(적수) 현상으로 피부질환을 호소하는 시민들이 100여명에 달하고...
인천 중구는 지난 7일 홍인선 구청장 주제로 적수피해와 관련한 대책회의를 열고, 그...
인천에서 '붉은 수돗물(적수)' 사태가 8일 째 이어지고 있는 가운데 정부 차원의 ...
인천시는 이를 계기로 상수도 수질사고 대응 시스템을 손볼 예정이다.인천 서구지역 수...
인천시가 붉은 수돗물(적수) 현상과 관련해 정부차원의 원인조사반을 구성한다. 인천시...
인천시가 서구 일대의 붉은 수돗물(적수) 현상과 관련해 정부차원의 원인조사반을 구성...

그림 1. 네이버 '2019-05-30'부터 10일간 크롤링 사례

그림1과 같이 웹 크롤링 결과로부터, 코사인 유사도를 도출하였다.

표1과 같이 사건 발생 전에는 관련 트윗 글이 없음을 알 수 있었다. 적수 사고 직후 10일간 크롤링 해본 결과 게시글이 3개, 4개, 5개 수집된 것을 확인 하였다.

수질 사고 발생 전에는 관련 네이버 기사가 없다가 사고 이후 600여개로 급증 하였다.

트위터는 키워드와 유사도가 사건 전에 0 이지만 사건 직후는 4.17% 까지 상승하였다. 네이버 역시 7.5%로 상승 한 것을 볼 수 있다. 네이버의 경우, 트윗과 달리 지속적으로 과거 기사가 영향을 미침을 알 수 있다.

향후, 수질 사고 뿐만 아니라 특정 사건의 SNS 확산 방식을 빅데이터화 할 수 있으며, 유사시에는 적재적소에 대응할 수 있도록 기여할 수 있을 것이다.

표 1. SNS 기간별 글 수 및 코사인 유사도

기간	Tweet	유사도 (%)	Naver	유사도 (%)
05.10~05.19	0	0	0	0
05.20~05.29	0	0	0	0
05.30~06.08	3	0.97	421	7.50
06.09~06.18	4	4.17	857	6.41
06.19~06.28	5	3.19	679	5.99
06.29~07.08	0	0	354	7.55
07.09~07.18	0	0	271	9.46
07.19~07.28	1	0.72	199	6.12
07.29~08.07	0	0	236	11.83
08.08~08.17	0	0	187	6.15

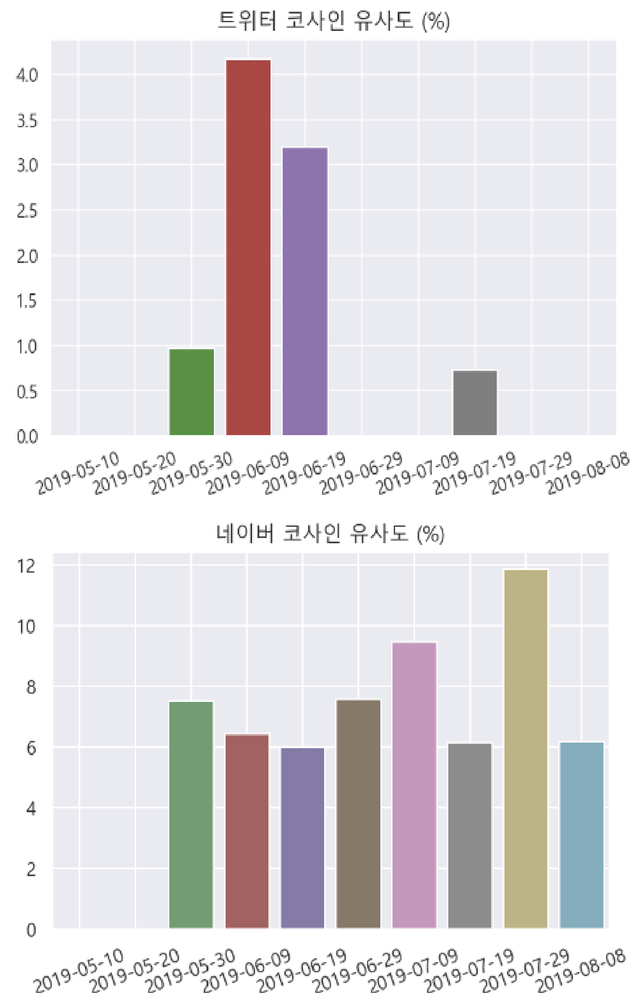

그림 2. 트위터, 네이버 키워드 유사도

Acknowledgements

본 연구는 한국수자원공사(K-water)의 개방형혁신 R&D 사업(21-BT-001)의 일환으로 수행되었습니다.

참고 문헌

[1] 심석용, "시장 고개 숙인 인천 붉은 수돗물 사태 1 년, 어디까지 왔나", 중앙일보, 2020.06.13.
[2] 김주영, "붉은 수돗물 터진 인천, 이번엔 유충 보인다", 세계일보, 2020.07.13.
[3] "Beautiful Soup Documentation", Leonard Richardson, 2022 년 4 월 10 일 접속, https://www.crummy.com/software/BeautifulSoup/bs4/doc/