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

Numerical Research of Accident using SNS Crawling

Jiin Min[1], Minsu Jung[1O], Do Geun Yoo[2], Seunghyeok Hong[1*]
[1]Division of Data Science, The University of Suwon, South Korea
[2]Department of Civil Engineering, The University of Suwon, South Korea

요 약

인천시 적수 피해 사건과 더불어 물 공급 사고를 조기에 파악하기 위하여, SNS 웹 크롤링의 수치화 연구를 진행하였다. 연구 기법으로 snscrape, TfidfVectorizer, cosine similarity 등이 사용되었다. 소셜 네트워크 서비스를 통한 문제 인식의 확산이 사건 전후에 검색량 지수의 확연한 증가로 확인되었다. 이러한 수치적 정보를 데이터베이스화 함으로써, 수질사고 뿐만 아니라 더 많은 분야에 관련된 다양한 사건 양상을 조기 파악하고, 추가적인 동향 분석을 할 수 있을 것으로 기대된다.

1. 서 론

2019 년 5 월 30 일 인천시 서구 검암·백석·당하동에서 수돗물 대신 붉은 물이 나온다는 민원이 접수됐다. 4 일 뒤엔 중구 영종도, 15 일 후엔 강화도에서 적수 피해 신고가 잇따랐다[1]. 그로부터 다음 해인 2020 년 7 월 13 일 붉은 수돗물 사태가 벌어졌던 인천 서구 일대에서 이번에는 유충이 보인다는 민원이 제기됐다[2]. 이와 같이 2019 년부터 지속된 수질사고로 인해 국민들은 제대로 된 물공급을 받지 못하게 되는 피해를 입었다. 이렇게 지속적인 수질사고 발생은 소셜 네트워크 서비스(SNS)를 통해 빠르게 확산되며 물공급 과정에 대한 부정적인 인식 증가와 신뢰도 저하를 초래한다. 따라서, 물공급 과정에서 발생하는 수질사고를 빠르게 인지하는 방법론의 적용을 통해 피해 최소화를 위한 노력이 필요로 한다.

본 연구는 물공급 과정 내 수질사고인지를 위한 SNS 별 웹 크롤링을 제안하고, 적용 결과를 분석한다. 웹 크롤링이란 웹페이지를 방문해 자료를 수집하는 것을 말한다. 웹은 기본적으로 HTML 형식으로 되어있다. HTML 형태로 어떻게 보이는지는 소스를 통해 볼 수 있는데 이런 소스들은 개발자들이 정형화된 형태로 관리하고 있다. 그렇기 때문에 일정한 규칙이 생기게 되고 이런 규칙을 분석해 원하는 정보만 추출해오는 것을 웹 크롤링 작업이라고 한다[3].

위와 같이 웹 크롤링을 수행하여, 관련 게시물의 수와 그 의미를 분석할 수 있다. 이때 핵심 키워드(수돗물, 적수) 조합을 활용한다. 그 검색량 정도를 지표(metrics)로 표현하여 유사도와 파급력 정도를 판단하였다.

2. 연구 방법

2.1 연구 대상

본 연구는 방법론 구현에 앞서, 연구에 필요한 대상에 대한 각종 SNS 별 웹 크롤링 가능 여부와 정보 획득 기간을 확인하여 물공급 과정에서 발생하는 수질사고를 연구 대상으로 선정하였다. 과거 유사 수질사고 발생 시 영향력과 관련 게시글이 크게 나타난 네이버와 트위터를 중심으로 웹 크롤링 절차를 진행한다. 네이버와 트위터에 검색할 키워드는 행정구역 명과 핵심 키워드인 '수돗물, 적수'를 조합해 함께 검색한다.

2.2 연구 기법

2.2.1 snscrape

소셜 네트워킹 서비스(SNS)용 스크레이퍼이다. 사용자 프로필, 해시태그, 검색 관련 게시물을 내보내준다. 트위터에서 게시글 작성자의 id 와 날짜, 게시글의 내용을 크롤링 해올 때 사용되었다. 검색 키워드는 지역명과 핵심 키워드를 함께

조합하여 검색한다. 본 연구에서는 검색 키워드를 '인천 수돗물 적수'로 설정하고 기간은 처음 수질 사고 민원이 접수된 5 월 30 일을 기점으로 사고 전인 5 월 10 부터 10 일 단위로 총 100 일 동안의 분석을 진행하였다. 이때 필터를 통해 기간과 검색 키워드, tweet 개수를 변경할 수 있도록 했다 (본 연구에서는 10 개 이하).

2.2.2 requests

HTTP, HTTPS 웹 사이트에 요청하기 위해 자주 사용되는 모듈 중 하나로 크롤링 과정에서 웹 사이트의 소스코드를 가져온다[4]. 네이버에서 크롤링을 진행할 때, url 에 쓰이는 검색어 부분은 미리 설정해둔 키워드 '인천 수돗물 적수'로 넣어주고 기간은 트위터와 동일하게 사고 전인 5 월 10 부터 10 일 단위로 총 100 일 동안의 분석을 진행하였다. 이때 검색어에 들어가는 날짜에는 '.'과 같은 숫자 이외의 문자가 포함되면 안 되기 때문에 전처리를 해주었다. 설정한 url 을 requests.get 을 이용해 가져와 변수에 지정하고 변수의 text 를 html 로 지정해 주었다.

2.2.3 Beautiful Soup

HTML 및 XML 파일에서 원하는 데이터를 쉽게 가져오기 위한 Python 라이브러리이다. 즐겨 찾는 파서와 함께 작동하여 구문 분석 트리를 탐색, 검색 및 수정하는 관용적 방법을 제공해 준다[5]. 앞에서 지정한 html 을 가져오고 'html.parser'를 통해 BeautifulSoup 의 인자 값을 지정해 주었다. 네이버에서 크롤링해 오는 것은 제목, 신문사, 기사 본문, 링크이다. 기사 검색 방식은 '관련도 순, 최신 순, 오래된 순' 중 '관련도 순'으로 설정하였다. 본문은 정제화 작업이 필요하기 때문에 정제화 함수를 활용했다. 검색 결과가 트위터에 비해 상대적 결과가 많은 네이버에서는 보다 더 정확한 분석을 위해 100 페이지까지 크롤링하였다.

2.2.4 TfidfVectorizer

TF-IDF 값은 특정 단어의 상대적인 빈도수를 나타내 주는 값이다. 값이 클수록 현재 문서에서는 자주 언급되면서 다른 문서에서는 잘 언급되지 않음을 뜻하고, 값이 작을수록 다른 문서에는 자주 언급되면서 현재 문서와 관련성이 낮음을 의미한다[6]. 트위터와 네이버에서 크롤링 한 결과를 수치화해 지표로 표현하였다. 이때 결과와 키워드를 비교하기 위해 키워드가 쓰여있는 행을 마지막에 추가하였다.

2.2.5 cosine similarity

두 벡터가 얼마나 유사한지 수치로 나타낸 것이다. 벡터 방향이 비슷할수록 두 벡터는 서로 유사하며, 벡터 방향이 90 도 일 때는 두 벡터 간의 유사성이 없음을 의미한다. 또한 벡터 방향이 반대가 될수록 두 벡터는 반대 관계를 보인다. 트위터와 네이버의 결과를 수치화한 지표를 마지막 행인 '인천 수돗물 적수'와 코사인 유사도를 도출한 후 평균하여, 백분율로 표기하였다. 소셜 네트워크의 검색기능은 포괄적으로 정보를 제공해야 하므로, 실제 검색어와 유사도가 높은 글만 제공되지는 않기 때문에 지수화 과정이 필요하였다.

3. 연구 결과

본 논문에서는 소셜 네트워크 서비스인 트위터, 네이버의 크롤링 방법을 바탕으로 검색량과 유사도를 도출해 분석하였다. 검색 결과를 도출한 뒤 결과로 나온 코사인 유사도의 평균을 구해서 해당 웹페이지의 검색 결과를 백분율 화하는 과정을 거쳤다.

표 1 에서 확인할 수 있듯이, 트위터의 경우 수질 사고 발생 전인 '2019-05-10~2019-05-19'과 '2019-05-20~2019-05-29'을 대상으로 크롤링을 진행한 결과, 관련 트윗 글이 없음을 확인할 수 있었다.

적수 사고 직후인 '2019-05-30'일부터 10 일간 크롤링 해본 결과로 게시글이 3 개, '2019-06-09'일부터 10 일간 4 개, '2019-06-19' 부터 10 일간 5 개의 게시글이 수집된 것을 확인하였다.

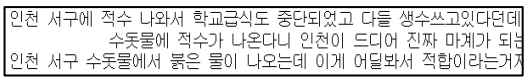

그림 1. 트위터 '2019-05-30'부터 10 일간 크롤링 결과

'2019-06-29'과 '2019-07-09'의 각 10 일간의 결과는 다시 0 개임을 보았고 '2019-07-19' 부터 10 일간의 결과는 1 개임을 확인하였다. 마지막으로 '2019-07-29'과 '2019-08-08' 부터 10 일간은 0 개임을 확인하였다.

네이버 메인 검색도 수질 사고 발생 전인 5 월 10 일부터 100 일 동안 10 일 단위로 크롤링한 결과를 도출하였다.

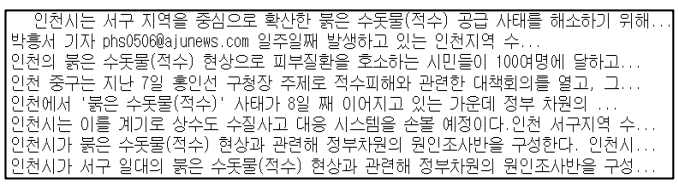

그림 2. 네이버 '2019-05-30'부터 10 일간 크롤링 결과

표 1 과 같이, 수질 사고 발생 전에는 관련 트윗 글과 네이버 기사가 0 개임을 볼 수 있고 사고 이후부터 글의 개수가 급증했다가 감소한 것을 볼 수 있다. 해당하는 키워드와의 코사인 유사도를 도출하였다.

표 1. 트위터, 네이버 기간별 글 수 및 코사인 유사도 표

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

트위터와 네이버 모두 수질 사고 발생 전에는 관련 검색 결과가 존재하지 않았기 때문에 유사도 결과가 0 인 것을 확인할 수 있다. 트위터의 그래프에서 알 수 있듯이 사고 전과 사고 후를 비교해 보았을 때, 기존 0%에서 직후에는 0.97%부터 4.17%까지 상승한 것을 볼 수 있고 이후 다시 3.19%에서 0%으로 감소하는 것을 확인하였다.

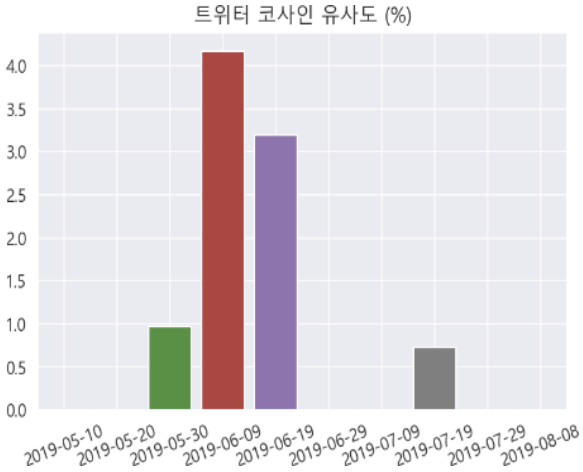

그림 3. 트위터 기간별 코사인 유사도

네이버의 경우도 수질 사고 발생 전과 후를 비교해 보았을 때, 0%에서 직후에는 7.5%까지 상승한 것을 볼 수 있고 다시 5.99%로 감소하는 듯하다가 그 후에도 꾸준히 높은 유사도를 유지하고 있는 것을 확인하였다.

우리는 본 연구를 통해 소셜네트워크 서비스를 통한 확산과 그 파급력을 확인할 수 있었다. 앞으로 우리는 이러한 정보를 분석하면서 다양한 방법론을 적용해 수질사고 정보의 전파 및 확산 방식이 '유충' 발생등과 같이 다른 수질 사고에서 동작할 수 있음을 확인하였다.

또한 영등포 문래동에서 일어난 '적수 사태'나 강원도 춘천에서 일어난 '수돗물 대란'과 같은 여러 유사 사례에 적용해 계속해서 연구를 진행할 수 있을 것이다. 더 나아가 수질사고 뿐만 아니라 더 많은 분야에 관련된 다양한 문제를 인지하고 그 확산 방식과 파급력을 확인해 추가적으로 분석할 수 있을 것으로 기대된다.

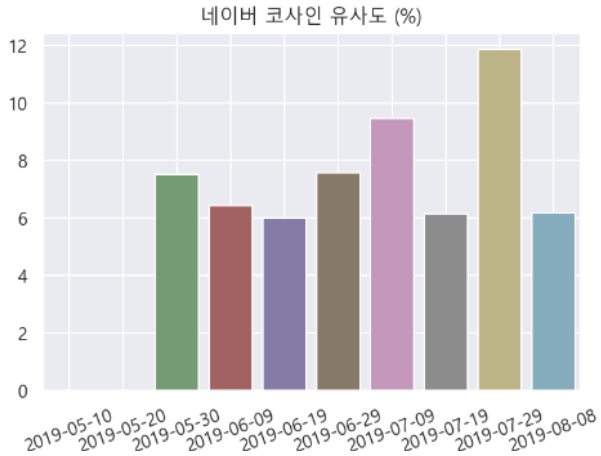

그림 4. 네이버 기간별 코사인 유사도

4. Acknowledgements

본 연구는 중소벤처기업부의 기술개발사업[S3245343]과 한국수자원공사(K-water)의 개방형혁신 R&D 사업(21-BT-001)의 일환으로 수행되었습니다.

5. 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] "Web Scraping(웹 크롤링) 이란?." Junee01, 2016 년 1 월 22 일 수정, 2022 년 4 월 10 일 접속, https://m.blog.naver.com/potter777777/220605598446

[4] "Python requests 모듈(module) 사용법." me2nuk. 2021 년 04 월 21 일 수정, 2022 년 4 월 10 일 접속, https://me2nuk.com/Python-requests-module-example/#requests-

[5] "Beautiful Soup Documentation", Leonard Richardson, 2022 년 4 월 10 일 접속, https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[6] "자연어 처리(NLP) - TF-IDF, TfidfVectorizer(), SGDClassifier()," 답을 찾아가는 과정, 2020 년 2 월 19 일 수정, 2022 년 4 월 10 일 접속, https://blog.naver.com/handuelly/221814212246