J. Korea Water Resour. Assoc. Vol. 55, No. S-1 (2022), pp. 1261-1270
doi: 10.3741/JKWRA.2022.55.S-1.1261
pISSN 2799-8746
eISSN 2799-8754

Water leakage accident analysis of water supply networks using big data analysis technique

Hong, Sung-Jin[a] · Yoo, Do-Guen[b]*

[a]Master of Engineering, Department of Civil Engineering, The University of Suwon, Suwon, Korea
[b]Assistant Professor, Department of Civil Engineering, The University of Suwon, Suwon, Korea

Paper number: 22-063
Received: 31 August 2022; Revised: 1 November 2022; Accepted: 4 November 2022

Abstract

The purpose of this study is to collect and analyze information related to water leaks that cannot be easily accessed, and utilized by using the news search results that people can easily access. We applied a web crawling technique for extracting big data news on water leakage accidents in the water supply system and presented an algorithm in a procedural way to obtain accurate leak accident news. In addition, a data analysis technique suitable for water leakage accident information analysis was developed so that additional information such as the date and time of occurrence, cause of occurrence, location of occurrence, damaged facilities, damage effect. The primary goal of value extraction through big data-based leak analysis proposed in this study is to extract a meaningful value through comparison with the existing waterworks statistical results. In addition, the proposed method can be used to effectively respond to consumers or determine the service level of water supply networks. In other words, the presentation of such analysis results suggests the need to inform the public of information such as accidents a little more, and can be used in conjunction to prepare a radio wave and response system that can quickly respond in case of an accident.

Keywords: Waster supply networks, Leakage accident, Web-crawling, Unstructured text data, Big data

R기반 빅데이터 분석기법을 활용한 상수도시스템 누수사고 분석

홍성진[a] · 유도근[b]*

[a]수원대학교 토목환경공학과 공학석사, [b]수원대학교 건설환경공학과 조교수

요 지

본 연구의 목적은 사람들이 쉽게 접할 수 있는 포털의 뉴스 검색 결과를 활용하여 쉽게 접근, 활용하지 못하는 상수도 누수 관련 정보를 모아 분석하는 것이다. 상수도 시스템의 누수사고 빅데이터 뉴스의 추출을 위한 웹크롤링 기법을 적용하고 정확한 누수사고 뉴스를 획득하고자 알고리즘을 절차화하여 제시하였다. 또한 추출된 누수사고 기사에서 발생일시, 피해영향, 발생지점, 피해원인, 피해시설 등과 같은 추가적인 정보의 획득이 가능하도록 상수도 누수사고 정보 분석에 적합한 데이터 분석 기법을 개발하였으며 그에 따른 적용결과를 제시하였다. 본 연구에서 제안한 빅데이터 기반 누수 분석을 통한 가치 추출은 기존의 상수도통계 결과와 비교를 통한 유의미한 가치를 추출하는 데 1차적 목표가 있으며, 이와 같은 분석 결과를 활용하여 향후 누수 사고 대응에 있어 소비자의 반응에 효과적으로 대응하거나 서비스 수준을 결정하는데 활용할 수 있다. 즉, 이와 같은 분석결과의 제시를 통해 사고와 같은 정보를 대중에 조금더 알려야하는 필요성을 제시하고, 사고 발생시 빠른 대처가 가능할 수 있는 전파 및 대응 체계를 마련하는데 연계활용할 수 있다.

핵심용어: 상수관망, 누수사고, 웹크롤링, 비정형 텍스트 자료, 빅데이터

***Corresponding Author.** Tel: +82-31-229-8676
E-mail: godqhr425@naver.com (Yoo, Do-Guen)

1. 서 론

2019년 기준 전국 상수도 보급률은 99.3%로 전년 대비 0.1% 증가했으며, 수돗물을 제공받는 인구는 5,274만 7천여명으로 전년대비 94만 명이 증가했다(ME, 2021). 그러나 1970에서 1980년대 산업화에 의한 상수도시설의 집중적 보급에 따라 수도시설의 21년 이상의 경년관 비율이 점차 증가되고 있으며, 이것은 수도시설의 내구연한에 따라 매년 일정하고 지속적인 상수관망 정비 및 개선이 필요함을 의미한다. 2019년 기준 전국 상수도의 누수율은 10.5%(약 7억 1백만 m^3)로, 수돗물 누수로 인한 손실액을 경제적 손실로 환산하면 생산원가 물 1 m^3당 914.3원 기준 약 6,592억 원이다(ME, 2019). 따라서 누수에 의한 피해를 저감하기 위한 노후관 파악 및 정비, 과학적 누수 탐지, 그리고 누수 발생 관로의 빠른 복구 등이 필수적이라 할 수 있다.

최근 국내의 상수도 공급과정에서의 유지, 관리 패러다임은 최근의 물관리 일원화와 최근연속적으로 발생한 중대형 누수·수질 사고 등을 계기로 수량 중심에서 수량-수질 통합관리와 대응 위주로 변화하고 있다. 특히, 상수도 보급률의 증가로 소비자는 1차적 목표인 물공급 뿐만 아니라, 고품질의 안전한 물 공급을 요구하는 등 높은 서비스수준을 요구하고 있다. 따라서, 소비자의 관점에선 생명에 직결되는 먹는 물 안전을 위한 상수도 관련 정보의 투명한 공개와 사고 발생 시 대응을 위한 관련 정보의 빠른 제공이 무엇보다 중요하다 할 수 있다. 스마트폰의 보급, 소셜네트워크서비스(Social Network Service, SNS) 활성화 등에 따라 수많은 데이터가 수집되고 있고, 데이터의 중요성에 따라 저장되어야 하는 데이터양은 기하급수적으로 증가하고 있다. 각종 정보가 실시간으로 생성되고 다양한 정보기술의 발전에 수없이 많은 데이터를 빅데이터 분석기법을 활용하여 분석하고 새로운 가치를 찾아내는 것이 중요하다. 그러나, 현재까지 국민이 접근할 수 있는 국내의 공식적인 상수도 관련 정보는 상수도의 일반적인 시설, 운영정보를 담아 매년 1월 발간되는 상수도통계와 일부 지자체에서 운영하는 실시간 운영관리 시스템의 주요 물공급 지점의 유량, 수압, 수질 자료에 한정된다. 상수도통계는 검·인증 과정의 시간소요로 인해 실시간 정보의 접근성이 떨어진다는 문제가 존재하며, 실시간 계측자료의 경우 누수 및 수실 사고 등과 같은 위기상황에 대한 인지, 대응, 복구 등과 관련한 정보는 아직 제공하지 않는다. 이와 같은 제한된 정보의 제공은 디지털 전환으로 대변되는 4차 산업의 가속화에 따라 보다 빠르고 실시간적인 정보의 제공을 요구받게 될 것이다.

대규모 뉴스 기사를 분석하기 위한 방법으로는 텍스트 마이닝, 오피니언 마이닝, 웹 마이닝 등이 있다. 텍스트 마이닝은 비정형 텍스트에서 자연어 처리와 형태소 분석기술에 기반을 두어 유용한 단어를 추출해 빈도수를 분석하고, 맥락수준의 의미를 찾아내는 분석방법이다. 오피니언 마이닝은 소셜미디어 및 웹사이트 등에서 여론 및 다양한 의견을 분석하여 이를 유용한 정보로 재가공하는 기법이다. 웹 마이닝은 인터넷 상에서 수집된 정보를 데이터 마이닝 방법으로 분석하여 통합하는 기법으로, 다양한 인터넷 매체를 활용하여 정보를 추출하는 기술을 말한다. 이 중 뉴스 기사와 같은 비정형 데이터를 처리하는 기법은 데이터마이닝 기법이다. 일정한 기준이 적용된 상식적인 범위에서 정제된 데이터베이스를 기반으로 부분적인 데이터를 다루는 정형 데이터마이닝의 한계를 뛰어넘는 기법들을 의미한다. Park *et al*. (2016)은 호텔 연구 동향 파악을 위해서 국제 주요 저널 4개(1994-2016)를 분석대상으로 하였으며, 텍스트마이닝과 소셜 네트워크 분석 방법을 적용하였다. 더불어, 중심성 분석, 토픽 모델링, 빈도 분석, 연관 분석을 통해, 호텔 분야의 국제 연구 동향 파악을 목적으로 하였다. 또한 사회적 이슈에 대한 비정형 데이터를 대상으로 하여 모형을 구축하고 분석하였다. 여러 가지 의미 있는 시사점을 도출하였음에도 불구하고, 텍스트마이닝 기법으로 인한 재현가능성 관점의 한계가 존재한다고 판단하였다. 오피니언 마이닝은 주로 다양한 웹사이트와 소셜미디어에서 특정 주제에 대한 사회적 사건, 정치 이슈 등에 대한 대중의 의견을 분석하는 데 적용되어 왔으며, 관련 연구 동향 또한 이러한 의견들의 추출, 평가, 분류, 이해를 위한 방법론에 관한 논의를 주로 다루고 있다(Chen and Zimbra, 2010). 텍스트 마이닝의 하위 분야인 오피니언 마이닝은 데이터의 수집 및 분석을 위한 전산언어학적 방법, 자연어 처리 기법들과 더불어 텍스트 마이닝 기법들이 적용된다. 의견들의 평가, 감정 등을 분석하고 대부분의 오피니언 마이닝 연구는 분류하는 방식으로 이루어진다. Song and Yang (2017)은 2000년 4월부터 2017년 2월까지 네이버 뉴스에서 유통되었던 뉴스 기사 80,428,892건을 수집하여 분석하였다. 포털 뉴스 유통을 네 가지 시기로 구분하여, 플랫폼으로서 포털이 어떤 뉴스를 보여주어야 하는지에 대한 논의가 필요함을 제시하였으며, 데이터를 통해 확인함으로써 뉴스 유통 경향을 파악하는 것에 대해 집중하였으나, 기존 온라인 뉴스 유통 현황을 정리하고 포털에 관한 다양성 연구를 위한 기초 자료를 제공한다는 의의를 가진다. Lee *et al*. (2017)은 빅데이터 분석을 소셜 네트워크에 적용하여 2018 평창올림픽에 대한 키워드를 살펴보았다. 네이버와 다음에서 제공하는 웹 문서, 카페, 지식인(팁), 뉴스, 블로그를 분석 채널로 선정하였으며, 빅데이터 자료 검색을 통해 키워드를 선정하

여 주제어로 추출하였다. 자료수집 및 분석을 위해 텍스톰(소셜 메트릭스 프로그램)을 통해 빈도 및 메트릭스 데이터를 추출하였으며, 단어 간 연결 정도 중심성과 연결 구조를 분석하여, 관계의 정도를 Ucinet6를 사용하여 계량화하였다. 추가로, CONCOR분석을 실시하여 유사성을 가진 단어들이 형성하는 군집을 도출하였다. Kim and Lee (2021)은 '코로나19'와 관련된 사회적 관심사를 연구하기 위해 클릭 수가 높은 연관검색어 230개를 선정하여 자기중심 네트워크를 구성하였다. 네트워크 그룹핑 결과에 따라, 총 5개의 관심 영역을 파악하고, 네트워크 분석 방법론으로 결과를 도출하였다. Kim (2020)은 코로나와 관련된 뉴스 기사 빅데이터로 사회적으로 형성하고 있는 주요 의제를 20개로 구분하여 추출하였다. 본 연구와 관련된 수공학분야의 경우 본 연구에서 분석하고자 하는 누수사고를 대상으로 비정형데이터를 활용한 사례 및 연구는 존재하지 않는다. 다만 수공학분야의 경우 최근 홍수 및 가뭄사상에 대한 모니터링 자료와 비정형 소셜 네트워크 분석 자료를 연계하여 비교분석하는 연구가 Lee and Hwang (2019)와 Jung *et al*. (2020)에 의해 각각 수행된바 있다.

본 연구에서는 웹 크롤링 방법을 이용한 자료수집, 텍스트 마이닝을 활용한 데이터 분석과 같은 빅데이터 분석기법을 이용하여 국내 상수도 누수 사고에 대한 정보를 추출하고, 추출된 결과에 대한 정확도 비교분석, 새로운 정보 추출 등을 수행하고자 하였다. 상수도 시스템의 누수사고 빅데이터 뉴스의 추출을 위한 웹크롤링 기법을 적용하고 정확한 누수사고 뉴스를 획득하고자 알고리즘을 절차화하여 제시하였다. 또한 추출된 누수사고 기사에서 발생일시, 발생원인, 발생지점, 피해시설, 피해영향 등과 같은 추가적인 정보의 획득이 가능하도록 상수도 누수사고 정보 분석에 적합한 데이터 분석 기법을 개발하였으며 그에 따른 적용결과를 제시하였다.

2. 상수도 시스템 누수 사고 웹 크롤링 모형

공식적으로 발표 및 배포되는 사회기반시설물 관련 정보와 심도 있는 연구 분석이 필요한 정보는 접근이 여전히 제한적이다. 사회기반시설물인 상수도 시스템은 대부분 국가중요시설로 지정되어 있어 다양한 정보를 획득하고 분석하는데 제약이 존재한다. 관련 국가통계인 상수도통계에서는 누수 사고 등과 같은 비정상적 상황에 대한 사고지점, 원인 등과 같은 세부 정보는 제공하고 있지 않다. 따라서, 본 연구에서는 일반인에게 제공되는 상수도 누수 사고 관련 뉴스 빅데이터를 분석하여 유의미한 가치를 추출하는 방법론을 제시하고, 과거 일정 기간 발생한 지자체의 상수도 누수 사고 관련 뉴스를 전수조사하고 도출된 사고 건수를 국가 공인 정보인 상수도통계 자료와 비교 분석을 실시하였다. 독립적인 누수 사고 기사를 추출하기 위해서 중복기사의 제거, 누수 관련 키워드 정립, 상수도 분야 이외의 관련 기사 제거 등의 절차가 필요하며, 이와 같은 기법을 R 프로그래밍을 통해 구현하였다. 추가적으로 뉴스 기사의 자연어 처리기반 정보추출기법을 통해 누수 사고 건수뿐만 아니라 사고발생일, 사고지역, 관의 크기, 피해 정도, 사고원인 등을 획득하여 상수도통계에서 제시하고 있는 정보보다 많은 가치를 추출하여 분석하였다. 본 연구에서 수행한 누수 기사 분석의 전체적인 흐름도는 Fig. 1과 같다. 개발된 상수도 시스템 누수사고 웹 크롤링 모델의 적용 및 검증을 위해 2017년부터 2019년도까지 총 3년에 대한 네이버 뉴스 검색을 통해 서울특별시에서 발생된 누수와 관련된 키워드를 검색하여 도출되는 네이버 뉴스 기사와 제안된 모형에 의한 웹 크롤링 뉴스데이터의 비교분석을 실시하였다.

2.1 비정형 기사 웹 크롤링

Fig. 1의 연구흐름도와 같이 대상지역의 누수사고정보와 일치도가 높은 기사를 수집하기 위한 웹 크롤링 과정의 단계는 다음과 같이 크게 네 단계로 구분된다. 첫 번째, 데이터를 수집하고자 하는 검색 웹사이트 선정, 두 번째, 수집하고자 하는 데이터에 관련된 검색 키워드 선정, 세 번째, 수집할 데이터에 관련된 날짜(기간) 설정, 마지막으로 앞서 검색 키워드 선정 및 날짜(기간) 설정에 맞는 뉴스 기사를 크롤링하여 데이터를 수집한다. 데이터 자료수집과 분석을 위해 프로그래밍 언어인 'R'을 활용하였다. R은 다양한 통계(선형 및 비선형 모델링, 고전적인 통계 테스트, 시계열 분석, 분류, 클러스터링 등) 및 그래픽 기법을 제공하며 확장성이 매우 높으며, 최근 R을 활용한 연구 등 많은 빅데이터를 활용하여 다양한 분야에서 활용되고 있다. 비정형 데이터의 유형은 크게 이미지, 텍스트, 영상과 음성, 로그파일 등으로 구분할 수 있는데, 네이버 뉴스에서 제공하는 뉴스 기사는 텍스트에 해당한다. 본 연구에서는 뉴스 기사 내의 뉴스 작성, URL, 뉴스가 작성된 일자, 언론사 URL, 기사 제목 및 본문 내용을 뉴스 기사와 직접 관련된 정보만을 선택적으로 수집하는 방식인 '웹 스크래핑' 기법을 활용하였다.

웹 스크래핑은 비정형 데이터를 수집하는 방법 중 하나이고, 인터넷상에 존재하는 데이터를 컴퓨터 프로그램을 통하여 자동화된 방법으로 웹에서 데이터를 수집하는 기법이다. 따라서 웹 스크래핑 기법을 이용하여 분석하고자 하는 관련된 뉴스 기사들을 수집하기 위해서는 수집할 기사의 주소(URL)

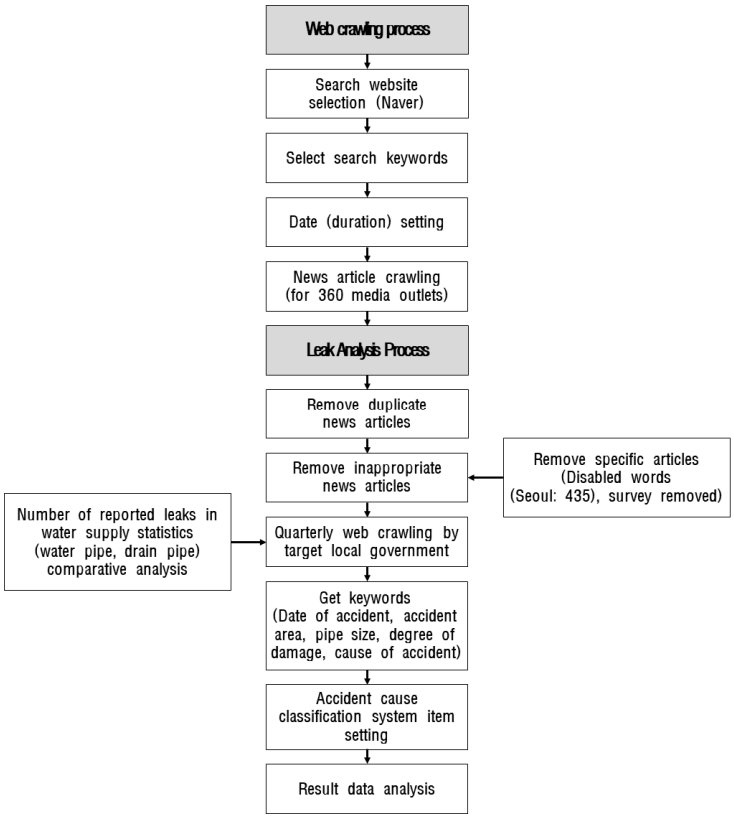

Fig. 1. Flow chart of proposed method

를 알아야 한다. 해당 URL은 네이버 뉴스 홈페이지에서 검색 키워드 선정을 통한 검색 키워드와 수집하고자 하는 기간에 대한 기간을 입력하면 컴퓨터상으로 확인할 수 있다. URL을 추출하였다면, 해당 기사에 대한 뉴스가 작성된 일자, 언론사 URL, 기사 제목 및 본문 내용 정보를 추출할 수 있다. URL을 바탕으로 통계 계산과 그래픽에 활용하는 프로그래밍 언어인 R을 활용하여, 수집한 HTML 문서의 익스플로러 개발자도구(F12)를 이용하여 태그(tag)라고 칭해지는 문법적 표시를 이용하여 수집된 각각의 기사들의 '보도 일자', '언론사 URL', '제목', '네이버 뉴스 URL' 관련 정보를 추출하였다.

2.1.1 검색 웹 사이트 선정

본 연구에 신문, TV, SNS 등 다양한 언론 매체 중 '인터넷 포털'을 선택한 것은 인터넷 언론사가 인터넷 뉴스에 대한 선호도가 증가하는 시점에서 매체 시장에서 급부상하고 있다는 점을 반영한 결과다. 필요한 데이터를 수집하고자 하는 검색 웹사이트 선정은 INTERNET TREND에 의해 작성된 국내 포털 사이트 점유율 순위에 의해 선정하였다. 국내 월 활성 사용자(MAU) 웹사이트 행동 데이터를 바탕으로 생성된 데이터를 제공하는 인터넷트렌드(http://www.internettrend.co.kr/trendForward.tsp)의 데이터에 따르면, 2012년 5월부터 2021년 4월까지 네이버가 가장 많은 73.7%의 점유율을 차지하는 것으로 나타났다. 따라서 네이버에서 제공되는 기사들이 대표성을 지닌다고 판단하여, 본 연구에서는 네이버를 검색대상 인터넷 포털로 선정하였다.

2.1.2 검색 키워드 설정

본 연구에서 활용한 검색 키워드는 누수에 관련한 사고 건수에 대해 수집할 지자체(1개의 지자체)의 이름과 사고로, 즉 "서울 누수"와 같이, 지자체명과 누수라는 단어를 합하여 검색 키워드를 선정하였다. 위와 같이 검색 키워드를 선정한 이유는 다음과 같다. 네이버 검색창의 검색 결과 알고리즘은 관련도순, 최신순, 과거순의 총 3가지 방식으로 결과를 표출한다. 관련도순은 검색어와 뉴스 간의 연관성을 파악하여 제공하는 것이고, 최신순은 뉴스 검색 결과를 최신 날짜 순서대로 정렬하여 제공하는 것이다. 마지막으로 과거순은 1990년 이후 기사를 오래된 순으로 제공하는 것이다. 검색 키워드를 "서울 누수"로 설정하는 경우 네이버 검색 결과로 도출되는 모든 기사에는 "서울" 그리고 "누수" 두 가지의 단어가 모두 들어감을 확인할 수 있다. 즉 네이버의 검색 결과 알고리즘은 키워드의 합집합이 아닌 교

Table 1. Article search results using all search keyword

No.	News article title
1	In the 'Mulrae-dong turbid tap water incident'... the Seoul Metropolitan Government Officials Union "needs to recruit more people"
2	Seoul Civil Servants Union "Reduced 605 people during 10 years of waterworks manpower… Recruitment is necessary"
3	[Editorial] Aging underground SOC that threatens public safety must be repaired quickly
Keyword	Seoul city leakage, Seoulcityleakage, Seoul leakage, Seoulleakage, Seoul metropolitan city leakage, Seoulmetropolitancityleakage

Table 2. Article extraction result by search keyword

Keyword	News article (No. in Table 1)
Seoul city leakage	1, 2
Seoulcityleakage	1, 2
Seoul leakage	1, 2, 3
Seoulleakage	1, 2, 3
Seoul metropolitan city leakage	1, 2
Seoulmetropolitancityleakage	1, 2

집합으로 검색 결과가 도출된다. 또한 네이버는 검색 키워드에서 단어를 서로 붙여 기입하더라도 자동적으로 띄어쓰기 된 단어를 모두 고려하여 검색 결과를 반영하기 때문에 "서울누수"와 "서울 누수" 두 가지의 키워드를 모두 사용할 수 있다. 그렇지만 한국어 표기법에 맞춰 띄어쓰기를 한 "서울 누수"를 키워드로 선정하는 것이 합리적이다. Table 1은 "서울시누수", "서울시 누수", "서울누수", "서울 누수", "서울특별시 누수", "서울특별시 누수"와 같이 총 6개의 검색 키워드를 모두 적용하였을 경우 검색되는 모든 뉴스기사를 나타낸 것이며, Table 2는 6개 각각 키워드 검색에 따라 도출되는 뉴스 기사의 번호를 나타내었다. "서울 누수"에 대한 기사에 대해서는 모든 뉴스 개수인 3개의 기사가 검색되는 것을 확인할 수 있었고, 나머지 5개의 키워드에 대해서는 "서울 누수"로 검색되어 수집된 기사 결과에 모두 포함되는 것으로 확인되었다.

2.1.3 날짜(기간) 설정 및 누수관련 뉴스기사 웹 크롤링

본 연구에서는 웹크롤링 결과의 비교를 위해 환경부에서 발행하는 2017년부터 2019년의 정보를 담은 "상수도통계(ME, 2018; 2019; 2020)"의 자료를 활용하고자 하였으며, 상수도통계는 지자체의 1분기 단위의 누수건수 결과를 제시하고 있다. 따라서 본 연구에서는 누수 사고 건수의 분석 기간과 적용대상 지역은 2017~2019년 3년에 해당하는 특별시 1개(서울특별시)로 설정하였고, 각 연도에 대하여 총 4분기로 구분하여(1분기(1월 1일~3월 31일), 2분기(4월 1일~6월 30일), 3분기(7월 1일~9월 30일), 4분기(10월 1일~12월 31일) 분석을 진행하였다. 웹 크롤링을 위한 기사의 확보를 위해 국내 포털 사이트 점유율 순위(월간 활성 사용자 기준)에 의하여 선정된 네이버의 뉴스 홈페이지를 활용하여 누수 사고 분석을 진행하였다. 네이버에서 제공하는 언론사 중 스포츠, 불교 등 관련도가 높지 않은 언론사를 제외한 총 360개를 대상으로 분석을 진행하였다.

2.2 중복/부적합 뉴스 기사 제거

뉴스 제목이 60~100% 일치하거나 기사전문의 명사를 추출하였을 때 60~100% 일치하는 경우 두 기사는 같은 주제의 내용을 전달하고 있다고 가정하여 중복 제거하였다. 또한 네이버 뉴스 기사 내용 중 본 연구에서 추출하고자 하는 상수도 누수기사와 관련 없는 내용은 제외하였다. 이와 같은 중복기사 및 관련도가 낮은정보의 제거와 같이 수집한 데이터를 분석하기 위해 용이한 형태로 처리하는 것은 모든 데이터에서 공통적으로 요구되는 작업이지만, 이 작업은 정형 데이터보다 비정형 데이터에서 더욱 필수적이다. 이와 같은 과정의 상세 내용은 다음과 같다.

중복기사를 제외하는 단계에서의 방법으로는 자카드 유사법을 이용하였다. 수집한 뉴스 기사들은 자카드 유사도(Jaccard Index)를 이용해 합집합에 대한 교집합의 비율로 유사도를 측정한다. 교집합의 비율은 기사내용 및 기사제목에 대해 각각 0.37, 0.1로 설정하였으며, 자카드 유사도 식은 다음 Eq. (1)과 같다.

$$J(A,B)=\frac{n(A\cap B)}{n(A\cup B)}=\frac{n(A\cap B)}{n(A)+n(B)-n(A\cap B)} \quad (1)$$

'불용어 처리' 단계에서는 수집된 데이터 중 분석할 때 불필요한 단어와 어구를 삭제하는 작업을 수행한다. 또한 마침표, 쉼표, 따옴표, 묶음표 등의 문장 부호와 각종 기호로 표현된 특수 문자, 의존명사 및 기타 분석할 때 의미를 부여할 필요가 없는 불용어(stopword)들을 제거해야 한다. 또한 본 연구에서는 누수 사고 기사의 정확한 추출을 위해 분석대상 지자체인 서울특별시에 대해 데이터 전처리 단계(기사 내용 및 기사 제목)에서 활용된 불용어를 총 435개로 설정하여 과정을 수행하였다.

Table 3. Five detailed factors information extraction from web-crawling articles

Detailed Information	Contents (Example)
Date of accident	2018.07.17.
Effect of damage	interruption, inundation, interruption, control, sea of water, disconnection
Accident area	447 (si/gun) by location (si/gun/gu)
Cause of accident	Aging, pipe aging, valve aging, accessory material aging, water supply pressure, high water pressure, maintenance, heavy vehicles, traffic load, temperature change, freeze and burst, expansion and contraction, typhoon, flood, disaster, corrosion, corrosiveness, corrosion soil, corrosive soil , local corrosion, ground subsidence, river scour, perforation, coating damage, bolt tightening, poor welding, poor electrical method, backfilling, materials, foundation treatment, soft ground, heterogeneous foundation, construction
Facility/Accident size	OOmm/m

Table 4. Example of extracting detailed information from leak accident article

'Water crisis' around Chungmuro Station due to water pipe rupture				
[Reporter Lee Young-ho, Korea Economy TV] At around 1:26 pm on the 17th, at the intersection of Chungmuro Station on Subway Line 3 in Jung-gu, Seoul, the water supply ruptured and water poured out from under the asphalt road. As nearby roads and sidewalks were submerged in muddy water up to knee height, vehicles and citizens I had trouble with this passage. The Seoul Waterworks Headquarters has identified that the 300mm water supply pipe has been ruptured, and is carrying out emergency restoration while controlling two lanes in the direction of Namdaemun-Dongdaemun Station. Water supply to some water pipes has been cut off, but other water pipes are connected to nearby houses. There was no water outage as water was supplied to the shopping mall. Repairs to the leak are expected to be completed around 10 p.m.				
Date of accident	Accident area	Facility/Accident size	Effect of damage	Cause of accident
17th at 1:00	Seoul Jung-gu	300 mm	controlled	-

2.3 누수 세부정보 획득

분석하고자 하는 뉴스 기사를 앞서 분석 기간과 선정한 검색 키워드를 이용하여 수집하고, 중복 및 부적합 뉴스 기사의 정제를 수행하다. 이 후, Table 3과 같이 누수 사고 관련 5가지 상세정보(발생일시, 피해영향, 발생지점, 피해원인, 피해시설)를 자연어 처리기법인 결합, 추출, 분리, 대체 기능을 활용하여 추출하고 실제 해당 기사와 비교·분석하여 5가지 정보의 정확한 추출 여부를 추출율로 제시하여 분석결과를 도출하였다. Table 4는 웹크롤링된 서울시 누수사고에 대한 비정형정보를 자연어처리기법을 통해 추출하여, 5가지 상세정보를 도출한 결과예시를 보여주고 있다. 총 5가지 인자 중 4개의 인자를 도출하였으므로 추출율을 80%가 된다.

3. 적용 및 결과

3.1 서울, 누수 키워드 기반 언론기사 웹크롤링

제안한 방법론을 서울특별시에 적용하여 결과를 분석하였다. 2017년 1월 1일부터 2019년 12월 31일까지 네이버의 뉴스 홈페이지를 활용해 서울특별시의 누수 사고 관련 뉴스 기사를 수집하였다. 각 분기는 1분기(1월 1일~3월 31일), 2분기(4월 1일~6월 30일), 3분기(7월 1일~9월 30일), 4분기(10월 1일~12월 31일)로 지정하였다. 대상 지자체(서울특별시)에 대해 누수 사고 관련 기사를 수집한 건수는 Fig. 2와 같다. 웹크롤링을 이용한 서울특별시의 누수 사고 관련 기사 건수는 4,706건이 수집된 것을 확인할 수 있다. 모든 뉴스기사는 "서울 누수"로 검색되어 수집된 기사는 총 4,706건으로 모든 기사에는 "서울" 그리고 "누수" 두 가지의 단어가 모두 들어감을 확인할 수 있다. 따라서 뉴스기사에서의 사고 지역에 "서울"이라는 단어가 포함됨을 알 수 있다. 동일기간 상수도통계자료에서 보고된 서울특별시 지역의 송배수관로의 누수사고 건수는 Table 5에 제시된 바와 같이 총 2,633건이다.

3.2 실제 누수사고 웹크롤링 및 결과비교

대상지자체에 대한 키워드 설정을 통한 분기별 웹 크롤링을 진행하고 도출된 누수 기사 건수를 상수도통계 자료와 우선적으로 비교하였다. 또한 Table 4에서 제시된 바와 같이 누수사고 관련 5가지 상세정보(발생일시, 피해영향, 발생지점, 피해원인, 피해시설)를 자연어 처리 기법인 결합, 추출, 분리, 대체 기능을 활용하여 추출하고 실제 해당 기사와 비교·분석하여 5가지 정보의 정확한 추출 여부를 추출율로 제시하여 분석을 실시하였다. 또한, 도출된 사고 발생 원인에 대해 상수관로의 누수사고 원인 분류로 공식적으로 활용하고 있는 6개의 항목 및 15개 세부 항목으로 자동분류 될 수 있도록 구현하였다.

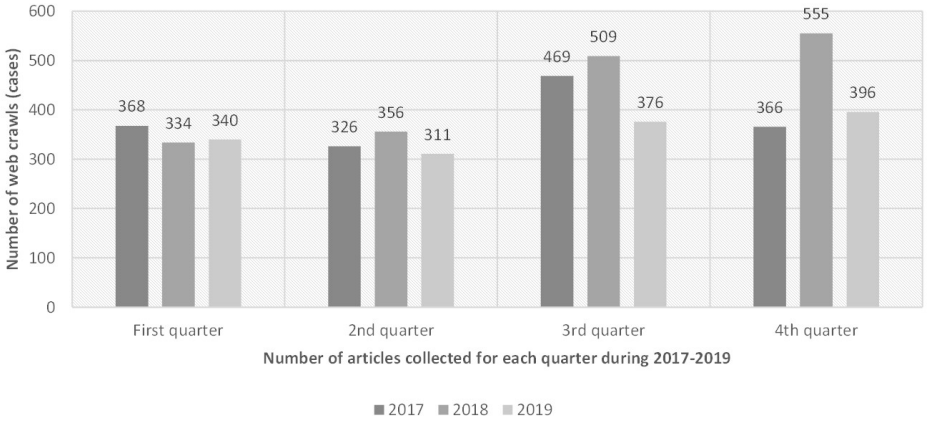

Fig. 2. Number of articles collected about municipal leak incidents for each quarter

Table 5. Comparison with number of leaks from web-crawling and number of reported leaks from waterworks statistics in Seoul

Year	Quarter	Seoul			
		Number of reported leaks (Transmission and distribution pipe)	Number of leaks	Rate (%)	5 factor information extraction rate (%)
2017	First	183	0	0.0	84
	2nd	169	0	0.0	
	3rd	239	3	1.3	
	4th	252	2	0.8	
2018	First	251	4	2.0	80
	2nd	161	1	0.6	
	3rd	214	3	1.4	
	4th	240	2	0.8	
2019	First	234	2	1.3	68
	2nd	203	0	0.0	
	3rd	231	1	0.4	
	4th	256	2	0.8	
Total		2,633	20	0.8	77.3

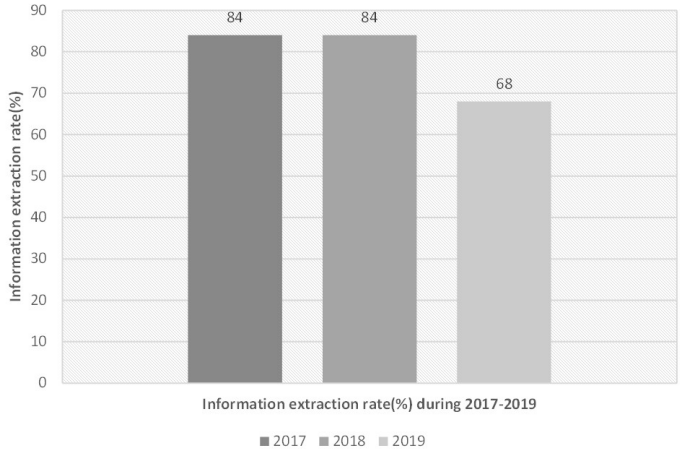

Fig. 3. Information extraction rate (%) during 2017-2019

2017년부터 2019년까지 분기별로 진행된 웹크롤링에 따른 실제 기사회된 누수건수와 상수도통계에서 제시된 자료의 비교 결과는 Table 5과 같으며, 실제 확인된 누수사고기사에서 획득된 정보의 추출비율을 연도별로 나타낸 그림은 Fig. 3과

Table 6. Acquisition and analysis results of leaked articles in Seoul (2017 year)

No.	Date (2017)	News article title	Date of accident	Accident area	Facility/ Accident size	Effect of damage	Cause of accident	Information extraction rate (%)
1	08.21	Road subsidence due to water pipe rupture in Jeongneung-dong··· Vehicle entry control, etc.	12 hours around 5	Seoul		Water cut-off Cotrol	construction	80
2	08.31	Dansu, Seongbuk-gu, Seoul, targeting 5,500 households in 7 buildings... Bottle water Arisu support	30th 31st	Seongbuk-gu Seoul	600 mm	Water cut-off	construction	100
3	09.22	5,000 households cut off water due to water pipe rupture in Mapo-gu, Seoul	from 6 p.m. on the 22nd	Mapo-gu, Seoul	600 mm	Water cut-off	construction	100
4	11.19	Water pipes freeze due to cold...'Seoullo 7017' flood		Seoul Station Seoul-ro			frozen	40
5	11.30	Gayang Station Street, temporary traffic delay on the way to work due to water pipe rupture repair	at 1:30 on the 30th	Gangseo-gu, Seoul	70 mm	Water cut-off Cotrol	restoration work	100

Table 7. Acquisition and analysis results of leaked articles in Seoul (2018 year)

No.	Date (2018)	News article title	Date of accident	Accident area	Pipe size	Degree of damage	Cause of accident	Information extraction rate (%)
1	01.26	900 households in Nogosan-dong, Seoul, 'single water'... Water pipe ruptured during construction	26th at 4:30	Mapo-gu, Seoul	200 mm	water interruption	under construction	100
2	01.29	Gyeongbokgung Station Water Pipe Leak... Recover by 6am on the 30th	29th	Jongno-gu Seoul		control the water supply	stretching action	80
3	03.02	A 'sinkhole' in the middle of Gangnam, Seoul... Road control and surrounding heating disruptions	1st 10:00 a.m.	Gangnam-gu, Seoul		control	construction	80
4	03.11	Water pipe leak in Jongno-gu, Seoul... Road inundation, 40 households cut off	10th at 11:00	Jongno-gu, Seoul		innundation stop	construction	80
5	06.11	University of Seoul Intersection Ground is off... Water pipe leak	4 o'clock on the 10th	Seoul Dongdaemun		control		60
6	07.17	Controlled by two cars due to flooding in Chungmuro··· 10:00 PM Recovery work completed (2 reports total)	1 o'clock	Dongdaemun Jung-gu Seoul	300 mm	control water cut-off flood	under construction	100
7	07.22	A road in Gangnam-daero in Seoul is turned off due to a leak in the water supply pipe	22nd	Seoul		control	deteriorated	80
8	07.23	A pothole with a diameter of 60 cm in Gyeomjae-ro, Jungnang-gu, Seoul··· "Recovery"	around 4 o'clock	Jungnang-gu, Seoul		control		60
9	11.02	The situation in Seoul Station where the water supply broke yesterday (the 1st) evening	around 9 o'clock on the 1st	Seoul Station Seoul		innundation	deteriorated construction	80
10	11.11	Hundreds of households cut off water supply due to water pipe damage... Residents evacuated due to villa fire	2 hours 1 hour	Seoul Guro-dong		water interruption	Restoration work	80

Table 8. Acquisition and analysis results of 5 leaked articles in Seoul (2019 year)

No.	Date (2019)	News article title	Date of accident	Accident area	Pipe size	Degree of damage	Cause of accident	Information extraction rate (%)
1	01.01	Road flooded due to leaking water pipe... Sunrise car crash kills 1	8 o'clock, 2 hours	Seongbuk-gu, Seoul		Flooding stop		60
2	02.14	'Water crisis' in the middle of the night when water pipe burst near Seoul Forest Station	13th at 11:00	Seoul Forest Seongdong-gu		Flooding control		60
3	09.20	In the process of repairing the water supply rupture in front of the National Police Agency... "Scheduled to be past 12:00 at night"	at 11:30 on the 20th	Seodaemun Seoul Station Seoul		controlled	backfill	80
4	11.21	Rupture of an old water supply pipe in front of the triangle area... Recovery complete	around 5 o'clock on the 20th	Yongsan-gu, Seoul		water cut-off		60
5	11.22	Water pipe burst in Sungin-dong, Seoul... Road flooded	22nd at 12:00	Seoul	300 mm	Flooding		80

같다. 웹크롤링을 통해 실제 누수사고로 인지된 건수는 3년간 총 20건으로 상수도통계에서 제시된 전체 누수건수(2,633건)의 약 0.8%에 불과한 것을 확인하였다. 이와 같은 결과는 누수사고의 경우 일정규모 이상의 시공간적 피해나 인명피해가 발생되지 않을 경우 기사화되지 않고 실시간적인 공표나 공지는 이루어지지 않음을 의미한다. 다만, 온수관로 파손에 따른 인명피해가 발생한 2018년의 경우 국민들이 체감하는 관로 파손에 따른 사고영향의 관심도가 높아, 상대적으로 다른해에 비하여 다소 높은 비율을 나타내었지만 절대적으로는 여전히 낮은 값으로 도출되었다. 이처럼 상수도 시스템에서 발생되는 누수나 수질사고의 경우 인간의 삶에 있어서 매우 중요한 먹는물을 공급하는 역할을 함에도 불구하고, 실시간적 사고나 동향정보가 활발히 제공되지 않는다는 것은 디지털과 데이터 시대를 선도하는 대한민국의 기술력을 고려할 경우 상수도 데이터의 개방이 매우 필요함을 역설적으로 보여준다고 판단된다.

각 년도 별 웹크롤링을 통해 실제 누수사고로 도출되고 전수조사를 통해 확인된 기사 내역과 5가지 세부정보의 추출결과는 Tables 6~8과 같다. 2017년부터 2019년까지 각각 5, 10, 5건의 누수사고가 확인된 것을 알 수 있다.

또한 5가지 정보의 추출율은 2017년의 경우 84%, 2018년에는 80%, 2019년에는 68%로 나타났다. 그 상세 내용을 살펴보면, 누수사고가 기사화 될 경우 5가지 정보 중 최소 2-3가지의 누수사고 정보를 정확하게 추출할 수 있음을 확인가능하다. 물론 누수사고의 정확한 원인과 더 심도깊은 사고상황 및 전개양상에 대한 정보는 기사를 통해 추출불가하지만, 단순한 누수사고의 건수보다 발생일시, 피해영향, 발생지점, 피해원인, 피해시설를 자연어처리기법을 통해 도출하는 것은 보다 정확한 누수사고의 양상과 경향을 파악하는데 도움이 된다. 특히 누수사고의 원인과 피해영향의 정보는 20개의 누수사고 중 14개(70%)는 80%이상의 정보를 정확히 추출한 것을 확인할 수 있어, 본 연구에서 제안한 누수사고 정보 추출 모형이 단순한 웹기사정보의 도출과 개수의 산정뿐만 아니라 추가적인 정보와 가치의 추출이 가능함을 알 수 있다.

4. 결 론

본 연구에서는 웹 크롤링 방법을 이용한 자료수집, 텍스트 마이닝을 활용한 데이터 분석과 같은 빅데이터 분석기법을 이용하여 국내 상수도 누수사고에 대한 정보를 추출하고, 추출된 결과에 대한 정확도 비교분석, 새로운 정보 추출을 수행하였다. 상수도 시스템의 누수사고 빅데이터 뉴스의 추출을 위한 웹크롤링 기법을 적용하고 정확한 누수사고 뉴스를 획득하고자 알고리즘을 절차화하여 제시하였다. 또한 추출된 누수사고 기사에서 발생원인, 발생지점, 피해범위 등과 같은 추가적인 정보의 획득이 가능하도록 상수도 누수사고 정보 분석에 적합한 데이터 분석 기법을 개발하였다.

제안된 웹크롤링 건수를 상수도통계자료의 누수건수와 비교할 경우 그 비율이 상대적으로 매우 낮게 나타났는데 그 원인은 다음과 같다. 뉴스 기사 중 사진으로만 보도된 뉴스 기사, 짧은 글의 뉴스 기사 등이 큰 비중을 차지하였다. 또한 지자체 누수사고의 경우 실제 기사화되는 비율이 매우 낮으며, 누수사고의 정보가 공개되지 않고 지자체 차원에서 주로 관리되고

있음을 확인할 수 있었다. 따라서 향후 상수도 사고와 관련된 더 많은 정보가 공개된다면, 일반인이 인식하는 먹는물 공급의 중요성과 지속적인 노후관 개량의 필요성이 높아질것으로 보이며, 궁극적으로는 상수도 누수 관련 정보의 투명한 공개에 따른 상호효과성이 향상될 것으로 판단된다.

본 연구에서 제시한 방법론과 방법론의 적용 결과에 따른 논문의 의의는 다음과 같다. 첫째, 웹크롤링 분석 결과를 통해 상수도 시스템의 누수사고에 대한 언론 매체의 공개와 사고 발생 시 대응, 복구에 대한 관련 정보가 뉴스기사를 통해 빠른 제공이 되고 있지 않음을 수치적으로 확인할 수 있었다. 이것은 상수도시설의 운영 및 유지관리에 있어 4차 산업혁명의 가속화에 따른 디지털 전환의 성공요인은 정보의 투명한 공개에서부터 시작할수 있다는 점을 의미한다. 즉, 사고 상황에 대한 투명한 공개를 통한 소비자와의 양방향 소통이 반드시 필요하며, 이를 통한 공급자 위주의 서비스 공급이 아닌 소비자 중심의 서비스수준의 결정 및 이해가 가능하다 할 수 있다. 둘째, 앞서 도출된 결과에 따라 누수사고와 같은 소비자에게 직접적인 영향을 미치는 정보의 공개가 제한된 상황에서, 본 연구에서는 일반 뉴스기사를 통해 누수사고에 대한 추가적인 정보의 추출이 가능함을 방법론을 통해 확인하였다. 즉, 일반뉴스기사로부터 상수도통계에서도 제시되고 있지 않은 누수사고의 원인, 피해규모, 그리고 피해의 확산정도 등에 대한 가치를 추출할 수 있었다. 이것은 위기상황에 대한 정보의 공개가 제한되는 현 상황에서 실시간적인 사고의 분석 및 인지, 대응을 위한 실시간 웹크롤링 분석의 상수도시스템 적용 가능성을 보여준다 할 수 있다. 또한, 이와 같은 분석결과의 제시를 통해 사고와 같은 정보를 대중에 조금더 알려야하는 필요성을 제시하고, 사고 발생시 빠른 대처가 가능할 수 있는 전파 및 대응 체계를 마련하는데 연계활용할 수 있다.

본 연구결과를 기반으로 향후 연구에서는 추가적인 지자체 적용을 통해 특·광역시와 일반 지자체와의 정보공개의 차이점에 대한 분석이 필요하다. 또한, 누수사고의 원인에 대한 추가적인 분류기준을 도출하고 뉴스에서 활용되는 일반적 용어보다 상수도분야 전문적인 용어를 기반으로 한 분석결과가 연계되어 도출 될 수 있는 고도화 연구가 필요하다. 또한, 다양한 누수사고 사례를 분석하여 원인에 따른 사고의 전개양상에 대한 분류와 상세분석이 진행될 필요가 있다.

감사의 글

본 결과물은 환경부의 재원으로 한국환경산업기술원의 지능형 도시수자원 관리사업의 지원을 받아 연구되었습니다 (2019002950002).

Conflicts of Interest

The authors declare no conflict of interest.

References

Chen, H., and Zimbra, D. (2010) “AI and opinion mining.” *IEEE Intelligent Systems*, Vol. 25. No. 3, pp. 74-80.

Jung, J., Park, D.H., and Ahn, J. (2020). “Drought evaluation using unstructured data: A case study for Boryeong area.” *Journal of Korea Water Resources Association*, Vol. 53, No. 12, pp. 1203-1210.

Kim, H.S., and Lee, K.S. (2021). “Virtual travel according to the development of information media and the changes in tourism after COVID-19.” *Journal of Korean Geographical Society*, Vol. 56, No. 1, pp. 1-14.

Kim, T.J. (2020). “COVID-19 news analysis using news big data: Focusing on topic modeling analysis.” *International Journal of Contents*, Vol 20, No. 5 pp. 457-466.

Lee, J., and Hwang, S. (2019) “A study on the application of social network service data for monitoring flood damage.” *Journal of Korean Society of Hazard Mitigation*, Vol. 19, No. 7. pp. 77-85.

Lee, J.H., Lee, J.M., and Jang, Y.S. (2017). “Analysis of 2018 Pyeong-Chang Olympic keywords using social network big data analysis.” *Journal of Korean Society for Sport Management*, Vol. 22, No. 6, pp. 73-89.

Ministry of Environment (ME) (2018) *Waterworks statistics*.

Ministry of Environment (ME) (2019) *Waterworks statistics*.

Ministry of Environment (ME) (2020) *Waterworks statistics*.

Ministry of Environment (ME) (2021) *Waterworks statistics*.

Park, J.S., Kim, C.S., and Kwak, K.Y. (2016) “Investigation of research trend in hotel domain using text mining and social network analysis.” *Journal of Tourrism and Leisure Research*, Vol. 28, No. 9, pp. 209-226.

Song, H.Y., and Yang, J.H. (2017). “Changes in portal news service and news distribution: 2000-2017 naver news big data analysis.” *Korean Journal of Journalism and Communication Studies*, Vol. 61, No. 4, pp. 74-109.