홍수 피해 발생 감시를 위한 소셜 네트워크 서비스 데이터 활용 방안 연구

A Study on the Application of Social Network Service Data for Monitoring Flood Damage

이정하* · 황석환**

Lee, Jeongha*, and Hwang, SeokHwan**

Abstract

Recently, as social network services (SNS) have developed, various data have been produced. In particular, people serve as a sensor in the event of a flood, and data including the fluidity of the population are being produced. For this reason, there are several cases overseas of use of social network data as a monitoring method in the event of a disaster. However, Korea does not use such data because the data are unstructured data. Therefore, this study proposes the use of the unstructured data generated by SNS as a monitoring method for flood occurrence. To this end, we have developed a web crawler program that extracts SNS data and derived ways to utilize social network service data in the event of a disaster such as a flood through comparative analysis of rainfall events and patterns.

Key words : Flood, Big Data, Flood Monitoring, Web Crawling, Flood Damage Prediction

요 지

최근 소셜 네트워크 서비스(SNS)가 발달함에 따라 다양한 데이터들이 생산되고 있다. 특히 홍수와 같은 수난(水難) 발생 시 사람들이 하나의 센서 역할을 함으로써 인구의 유동성을 포함한 데이터가 생산되고 있다. 이러한 이유로 재난 발생 시 모니터링 방법으로 활용된 해외 사례가 있으나 우리나라에서는 비정형데이터라는 이유로 소셜네트워크 서비스 데이터를 활용하고 있지 않고 있는 실정이다. 따라서 본 연구에서는 SNS으로 생성된 비정형 데이터들을 홍수 발생 시 모니터링 방법의 하나로 사용하는 방안을 제안하고자 한다. 이를 위해 SNS 데이터를 추출해내는 웹 크롤러 프로그램을 개발하였으며 이를 활용한 강우 사상과 패턴을 비교 분석을 통해 홍수와 같은 재난 발생 시 소셜 네트워크 서비스 데이터 활용 방안을 도출하고자 한다.

핵심용어 : 홍수, 빅데이터, 홍수 모니터링, 웹 크롤링, 홍수 피해 예측

1. 서 론

높은 스마트 폰 보급률과 함께 급진적으로 발달한 소셜 네트워크 서비스(SNS)는 서비스의 형태를 벗어나 새로운 정보를 생산하는 채널로 이용되고 있다. 이로써 생산된 빅데이터(Big Data)는 다양한 분야에서 활용되고 있으며 특히

재난과 관련된 상황에서도 유의하게 사용되고 있다. 한 예로 2013년에 캐나다에서 발생한 알버타 강 범람이나 인도네시아의 자카르타 홍수가 발생했을 때 Google 위기 대응팀은 이러한 데이터를 수집하여 위급한 상황을 실시간으로 알리는 서비스를 제공하였으며 대피 구역 및 긴급 피난처 위치 등을 나타낸 위기 지도를 제공하였다. 현재 국내 홍수 예·경

*정회원, 과학기술연합대학원 대학교 KICT School 박사과정(E-mail: leejungha100@kict.re.kr)

Member, Ph.D Candidate, University Of Science and Technology KICT School

**교신저자, 정회원, 한국건설기술연구원 돌발홍수연구센터 센터장(Tel: +82-31-910-0241, Fax: +82-31-910-0251, E-mail: sukany@kict.re.kr)

Corresponding Author, Member, Professor, University Of Science and Technology KICT School

보 시스템은 계측기를 통한 정형화된 수치 데이터만을 활용하여 예경보가 이뤄지고 있어 미계측 지역이 많은 소하천지역이나 도심하천의 경우 늦은 예경보가 이루어져 인명사고가 일어나는 등의 한계점이 존재한다.

따라서, 본 연구에서는 기존의 정형데이터 뿐만 아니라 인구의 유동성을 고려한 SNS 데이터 등의 비정형 데이터의 효용성을 분석해보고 이를 통해 홍수와 같은 수난(水難) 발생 시 소셜 미디어 데이터와 같은 비정형데이터의 활용 방안을 도출하는 것을 목표로 한다.

국외의 선행연구로 Taniguchi (2012)는 태풍으로 인한 폭우가 내리는 동안 트위터(Twitter)가 재난 정보를 효과적으로 제공하는 도구로서의 성능을 확인하였다. Zhang et al. (2016)은 재난 및 사건, 사고 등의 이벤트들에 대한 트위터의 흐름을 실시간으로 모니터링이 가능한 GEOBURST 프로그램을 제작하였으며 이벤트의 규모를 보다 빠르게 전달이 가능함을 확인하였다. Johansson et al. (2012)은 유럽의 Alert4All 연구 프로젝트를 통해 소셜 미디어를 활용한 위기 모니터링 및 대응에 효과적임을 확인하였다.

2010년 Murthy and Longwell (2013)은 파키스탄에 발생한 홍수에서 트위터의 효용성을 확인하기 위해 키워드 홍수와 관련된 키워드를 기반으로 트위터 데이터를 추출 후 사람들의 기존 뉴스와 같은 전통 미디어 매체와의 활용성과 선호도를 비교하였다. 그 결과 홍수와 같은 재난을 겪는 동안 사람들은 기존의 매체보다 트위터가 활성화된 것을 확인하여 재난 발생 시 활용방안을 제안하였다.

기존 국내 선행연구에서 Kim (2011) 국가 재난 대응에 있어서의 소셜 미디어 활용 방안에 대한 연구가 진행되었다. 소셜 미디어의 신속성, 경제성 및 다양성은 재난 대응에서의 의사소통 역할이 가능하다는 것을 실제의 활용 예시와 함께

Fig. 1. Schematic of the Web Scraping and Analysis Process

긴급 상황에서의 소셜 미디어의 활용 방안을 제안하였다. Kim et al. (2017)은 실시간의 기상 데이터 및 뉴스 데이터, 키워드 데이터를 기반으로 하는 홍수 재난 안전 시스템 연구가 진행된 바 있다. 따라서 본 연구에서는 Fig. 1과 같이 웹 스크래핑 프로그램을 개발하여 데이터를 그림과 같이 추출하려 한다. 또한, 이를 통해 수집한 소셜 네트워크 서비스 데이터(SNS)를 분석 및 패턴 해석을 하여 홍수 발생 예측에 있어서 SNS 데이터 활용 가능성을 확인하고자 한다.

2. SNS 크롤러 개발 개요

2.1 SNS 크롤러 개발 목적 및 기능

소셜 네트워크 서비스 데이터에는 다양한 데이터가 혼재되어있으며 홍수 관련 데이터만 추출하기 위해서는 키워드를 기반으로 하는 웹 크롤러(Web Crawler) 개발이 요구된다. Web Crawler란 Web Scraping이라고도 불리는데 각종 웹 사이트에서 데이터를 자동적으로 추출할 수 있도록 하는 일종의 컴퓨터 프로그램이다. 본 연구에서는 홍수와 관련된 수난(水難)과 관련된 데이터만을 추출하기 위해서 키워드를 설정한 후 키워드를 바탕으로 수집하도록 프로그램을 설계하였다. 본 연구에서 개발한 키워드 기반 웹 크롤러의 기능은 Table 1과 같으며 구성 화면은 Fig. 2와 같다.

Table 1. Function of SNS Crawler

	Function	Explanation
1	Keyword Selection	Keyword can be changed. (Currently Set to 18 keywords related to flood)
2	Location Extraction	Area names in the title or body can be extracted. (Can be extracted to the 'Gun' Unit)
3	Time Extraction	Can Select the time period want to crawl and extract the publishing Time up to Minutes
4	Data Analysis Result	Visualize the results of statistical analysis after selecting date range, region and keywords
5	Database Creation	The database is created and can be extracted in several formats

(a) WebCrawler Drive Screen

(b) WebCrawler Drive Results

Fig. 2. Web Crawler System

특정한 이벤트가 발생한 날짜 지정하여 수집이 가능하도록 설계하였으며 수집 대상 포털 사이트와 SNS 채널을 선택할 수 있다. 본 연구에서 지정한 수집 대상 사이트는 네이버(Naver), 다음(Daum), 트위터(Twitter) 와 인스타그램(Instagram)이며 지정된 키워드를 중심으로 데이터 수집을 수행한다. 수집된 데이터는 마리아 데이터베이스(MariaDB)에 저장되며 후 웹크롤러 프로그램을 이용하여 키워드별과 지역별 통계 분석이 가능하도록 설계하였다. MariaDB는 오픈 소스의 관계형 데이터베이스 관리 시스템(RDBMS)이다. MySQL과 동일한 소스 코드를 기반으로 하며, GPL v2 라이선스를 따른다.

2.2 SNS크롤러 개발 과정

2.2.1 웹 사이트 및 소셜 네트워크 채널의 선정

웹 크롤러를 통한 데이터 추출을 위해서는 먼저 수집하고자 하는 웹사이트의 선정이 우선적이다. 본 연구에서는 SNS 및 수집 대상 웹사이트 선정에 있어 다음과 같은 기준을 설정하였다.

- 첫 째, 많은 이용자를 확보하고 있는 포털 사이트 및 소셜 네트워크 서비스 채널
- 둘 째, 개방형 SNS 채널

첫째, 많은 이용자를 확보하고 있는 포털 사이트 및 소셜 네트워킹 서비스 채널이어야 한다. 데이터의 절대적 양은 데이터 분석과 결과에 많은 영향을 미치게 되기 때문에 가능한 많은 양의 데이터 확보가 필요하다. 따라서, 이용자들이 많은 포털 사이트 및 네트워킹 서비스 채널을 선정하고자 하였으며, 주이용 SNS 채널에 대한 조사하였다. Table 2는 주요 이용 SNS 채널을 조사한 결과를 나타낸 것으로 통계청의 자료에 따르면, 가장 많이 사용되는 SNS 채널 유형은 카카오톡이며 페이스북, 트위터, 인스타그램 순으로 많이 사용한다는 결과가 나왔다. 본 조사결과는 2015년 기준으로 작성된 것이며 1, 2 순위 중복 선택이 가능하도록 조사된 것이다.

둘째, 개방형의 SNS 채널이어야 한다. 개방형의 채널은 게시된 콘텐츠에 누구나 접근이 가능하나 폐쇄형 채널의 경우, 서로 소셜 네트워크 서비스 상에서 ‘팔로우’를 하거나 ‘친구’로 추가되어야지만 타인의 게시물에 접근이 가능하기 때문이다.

주요 이용 소셜 네트워크 서비스 유형인 카카오톡이나 페이스북은 텍스트, 사진과 동영상 등의 다양한 콘텐츠를 갖고 있으며 폐쇄형 SNS 채널의 성격을 갖고 있는 것으로 조사되었다.

반면 트위터는 텍스트 형태의 콘텐츠만 게시가능하며 최대 140자로 제한되어있어 서술식 형태의 데이터보다 개조식으로 작성되어 있는 경우가 많다. 인스타그램은 사진과 최대 15초로 제한한 동영상을 업로드할 수 있는 SNS 플랫폼으로 공개여부를 선택할 수 있는 반개방형 유형이다. 본 과제에서는 실시간으로 임의의 웹 데이터를 추출이 불가능할 것으로 판단되는 폐쇄형 채널(카카오톡, 페이스북)은 웹 크롤러 수집 대상에서 제외하였으며, 개방형 채널인 트위터와 인스타그램을 대상으로 크롤링 범위 선정을 하였다.

좀 더 많고 정제된 데이터를 추출하기 위해 포털 사이트도 웹 크롤링 수집 대상으로 하였다. 국내 주요 포털 사이트인 Naver와 Daum에서는 각 사이트 뉴스를 대상으로 수집하도록 알고리즘을 작성하였다. 실제로 현장에서 사람들이 실제로 체감하는 정도는 SNS 상의 데이터에 많이 분포되어있어 정보를 실시간으로 얻기 위해서는 SNS 데이터 수집이 필수적이다. 하지만 두 포털 사이트는 사실 위주와 구조화된 데이터를 갖고 있어 SNS 데이터보다 정확하며 분석이 보다 쉽다는 장점이 있다.

최종적으로 선정한 수집 대상 웹사이트 및 SNS 채널은 총 4가지로, 포털사이트인 Naver와 Daum 그리고 SNS인 Twitter와 Instagram이다. 또한, 실시간 데이터 수집을 위해서는 일정한 주기로 수집이 가능하도록 설계가 필요하여 오늘 날짜와 특정 기간 동안 작성된 웹문서를 선택적으로 크롤링을 수행하도록 알고리즘을 작성하였다.

Table 2. Web Crawler's Website to Collect and Each Website's Characteristics

SNS Channels	SNS Channel Type	Characteristic
NAVER	Open	<ul style="list-style-type: none"> ■ The most used portal site in Korea ■ Collection of news articles provided by this site
Daum		
Twitter		<ul style="list-style-type: none"> ■ Fewer ads ■ SNS platform with text ■ Social network service channels that function as messengers, blogs, etc
Instagram	Selectable	<ul style="list-style-type: none"> ■ Social network service channel for sharing photos and videos ■ Main content with photos and short videos ■ Search function using “#” is possible

Table 3. Keyword Specified in Connection with Water Disaster

Keyword	Meaning
홍수	- 비가 많이 와서 강이나 개천에 갑자기 크게 불은 물
홍수주의보	- 홍수예보 발령지점의 수위가 계속 상승하여 경계홍수위를 초과할 것이 예상되는 경우에 발표하는 기상주의보 - 경계홍수위는 제방·수문·교량 등에 대한 경계가 필요한 수위로서, 하천부속물의 설치계획을 수립할 때에 그 기준으로 정하는 홍수량인 계획홍수량의 100분의 50에 해당하는 유량이 흐를 때의 수위 또는 2년이나 3년에 한번정도의 횡수로 도달하는 수위를 의미
홍수경보	- 홍수예보 발령지점의 수위가 계속 상승하여 위험홍수위를 초과할 것이 예상되는 경우에 발표하는 기상주의보 - 위험홍수위는 제방·수문·교량 등에 대한 붕괴의 위험이 예상되는 수위로서, 계획홍수량의 100분의 70에 해당하는 유량이 흐를 때의 수위 또는 5년 내지 10년에 한번정도의 횡수로 도달하는 수위를 의미
호우특보	- 홍수에 의하여 피해가 발생하거나 예상될 때 특별히 하는 보도
폭우	- 갑자기 세차게 쏟아지는 비
장대비	- 장대처럼 굵고 거세게 짹짹 내리는 비
비	- 대기 중의 수증기가 높은 곳에서 찬 공기를 만나 식어서 엉기어 땅 위로 떨어지는 물방울
침수피해	- 물에 잠김으로 인한 피해
물폭탄	- 폭우 등으로 물이 폭탄처럼 쏟아지는 것을 비유적으로 이르는 말
물난리	- 큰물이나 그 밖의 원인으로 많은 물이 넘쳐서 일어난 혼란
잠김	- 깊숙하게 박히거나 폭 묻히다는 뜻의 ‘잠기다’의 활용형
홍수피해	- 홍수에 의한 피해를 이르는 말
기습폭우	- 갑작스럽게 세차게 쏟아지는 비

Note. 본고의 키워드는 국립국어원 표준국어대사전을 참고해 작성하였으며, 영문으로 번역시 내용의 의미가 왜곡될 수 있어 한국어로 표기함(이하 그림 내 국문 키워드 동일 사유 적용).

2.2.2 데이터 분류를 위한 키워드 선정

수난 예측 및 예보를 위해서는 leadtime 확보가 매우 중요하다. 웹 사이트에는 수많은 데이터가 있어 모든 데이터를 추출하기에는 많은 시간이 소요된다. 따라서, 원하는 웹 데이터만을 단시간 안에 효율적으로 추출하기 위해서는 키워드 위주의 데이터 수집이 필요하다. 키워드별 수집을 위해서 본 연구에서 수난과 관련된 키워드를 지정하였다. 지정된 키워드와 각각의 의미는 Table 3과 같다. 본 연구에서 선정한 키워드는 이전에 수행한 웹문서 분석과정에서 ‘홍수’ 키워드를 대상으로 웹 문서 데이터를 수집하였을 때 제목이나 본문에 ‘홍수’와 관련하여 등장한 빈도수가 높은 키워드를 토대로 정리하였으며 본 과제에서 수난(水難)과 밀접한 단어를 추가하여 작성한 것이다. 수집 키워드에 대한 의미는 국립국어원 표준국어대사전(<http://stdweb2.korean.go.kr/main.jsp>)을 참고하였다.

2.2.3 시공간적 분석을 고려한 데이터 추출 항목 선정

각 사이트에서 키워드를 중심으로 추출할 데이터 항목은 Table 4와 같다. 키워드가 포함된 게시된 글의 url, 제목, 본문, 작성일 및 게시일, 웹문서에 포함된 지역 명, 수집 키워드이다. 게시 및 작성된 일자는 분 단위로 수집하도록 설정하였으며 SNS 채널에서의 본문 내용은 태그(Tag, #)를 이용하여 작성한 내용도 포함이 된다. 더불어 공간적 정보 확보를 위해 지역 명을 디셔너리 형태로 분류가 가능하도록

작성하여 시·도명 및 시군구명을 포함했을 경우 이를 추출하도록 설정하였다.

Table 4. Extract Data Item

	Explanation
Extract Data	<ul style="list-style-type: none"> • (Article) title • Publish date and creation date • Web document URL • Collected Keywords in Web Documents • Text in web document • Region names included in web documents (City, District)

3. SNS 크롤러 홍수 사상 검증

3.1 검증 방법

SNS크롤러에 대한 홍수 사상 검증에는 SNS 데이터와 강우량의 패턴분석이 요구된다. 데이터의 범주가 다른 두 자료들의 패턴을 분석해야하므로 정규화(Normalization) 후 상관관계 분석을 통해 패턴 분석을 진행하였으며 사용된 정규화 공식은 Eq. (1)과 같다.

$$x = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

정규화 공식은 범주를 0과 1사이로 일치시켜 범주가 다른 두 데이터를 분석하기 위해 주로 사용되는 공식이다. $x_{\min}$ 은 최솟값을 $x_{\max}$ 는 최대값을 지칭한다.

정규화 이후 상관계수(Correlation Coefficient; CC) 값을 비교하여 SNS데이터의 패턴과 강우량의 상관관계의 강도를 분석하였다. 상관계수의 공식은 Eq. (2)와 같다.

$$CC = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2)$$

상관관계가 강할수록 1 또는 -1에 가까우며 연관성이 떨어질수록 0에 가까운 값을 나타낸다. n 은 전체 자료의 개수를, $\bar{x}$, $\bar{y}$ 는 각각의 평균값을 지칭한다.

4. 실시간 웹크롤러 기반 SNS 데이터와 홍수 발생 간의 패턴 분석 및 결과

SNS 데이터와 홍수 발생 간의 시·공간적 패턴을 분석을 하고자 2018년에 발생한 강우 사상과 태풍의 사례에 대한 비교 분석하였다. 개발한 웹 스크래핑 프로그램을 이용하여 각 사상에 대한 데이터를 수집한 후 데이터 분석을 진행하였다. 분석에 사용한 홍수 사상은 Table 5와 같다.

총 두 개의 강우 사상을 이용하여 SNS 데이터와 비교 분석하였으며 이는 각각 2018년 5월 16일 서울 성북구 정릉천에서 발생한 홍수사상과 같은 해 10월 6일에 발생한 태풍 콩레이 홍수 사상이다.

4.1 2018년 5월 16일 정릉천 홍수

2018년 5월 16일 낮 12시 20분쯤 서울 성북구 하월곡동 정릉천 자전거도로에서 불어난 물살에 휩쓸려 1명이 사망하는 사고가 발생하였다.

정릉천 일대인 월계2교에서 12시 20분 쯤 수위가 증가하였다. 도심에서의 SNS 데이터와 패턴을 비교 해보고자 2018년 5월 15일부터 18일까지 크라우드 소싱 데이터를 기반으로 분석을 진행하였다.

그래프 Fig. 3은 키워드 ‘호우’와 ‘폭우’와 관련하여 수집

된 성북구 지역의 데이터와 당시의 일강우량을 비교해 본 결과이다. 인명 피해가 발생한 16일은 키워드 ‘폭우’와 관련된 데이터의 양이 가장 높았으며 ‘호우’와 관련된 데이터 또한 비슷한 양상을 보였다. ‘호우’는 일강우량과 동일한 그래프 양상을 보이는 것으로 확인되었으며 ‘폭우’와 관련된 데이터 또한 일별 강우량과 대체적으로 비슷한 그래프를 보여 홍수 탐지기의 결과가 강우의 흐름과 시·공간적으로 일치함을 알 수 있다.

각 키워드별 정릉천 홍수 관련 SNS 데이터들에 대한 키워드별 빈도수를 확인하여 Fig. 4와 같은 그래프를 얻을 수 있었다. 정릉천 홍수 사상에서는 ‘비’, ‘폭우’, ‘침수’ 등의 순으로 빈도가 높았다. 그리고 16일부터 18일까지 3일간의 일강우량과 키워드별 상관계수(Pearson correlation coefficient)를 비교한 결과는 Fig. 5와 같다. ‘비’와 같은 키워드는 상관계수가 1에 근접하게 나타나 실제 강우사상과 높은 상관관계를 보이는 것을 알 수 있었다. 이에 반해 ‘침수피해’ 같은 키워드들은 음의 상관관계를 나타냈다. 주목할 점은 높은 양의 상관관계를 나타내는 키워드는 강수와 관련된 키워드(비, 폭우, 장대비, 태풍, 호우특보, 장마, 기습폭우, 물폭탄) 들이고 높은 음의 상관관계를 나타내는 키워드는 홍수나 홍수피해와 관련된 키워드(침수피해, 침수, 홍수, 물난리, 범람)라는 점이다. 이는 강수나 홍수 등의 실제 발생 시점과 키워드 검색 시점의 일치 여부에 의해 발생하는 것으로 추정된다. 비와 관련된 키워드 검색은 비가 내리는 시점에서 증가하고 홍수나 홍수피해와 관련된 시점은 비로 인한 홍수피해가 발생한 시점 이후에 증가한다고 해석이 가능하다. 특히 홍수피해와 관련된 키워드들은 뉴스에서 많이 쓰이는 키워드들로 홍수 발생 이후 피해에 대한 뉴스기사들로 인해 높은 음의 상관계수가 나타난 것으로 확인되었다.

이러한 측면에서 호우의 발생시점과 인간 활동 시점(주간, 야간, 심야 등)에 따른 뉴스 노출시점의 영향을 분석해 보았다. SNS데이터와 시 강우량과의 비교·분석을 진행해본 결과 Fig. 6과 같은 그래프를 얻을 수 있었다. 인구의 유동이 높은 시간대인 출근시간대와 퇴근시간대는 강우량과 데이터 증감 패턴의 시간적 차이가 거의 없는 반면 새벽시간대는 피크 점의 시간차 있는 것을 확인 할 수 있었다. 이는 체감 가능 시점에 호우가 발생하는 경우 상관성이 높지만 야간과 같이 인간 활동이 뜸해지는 경우에는 SNS데이터와 강우량

Table 5. Data Summary of Flood Used for Analysis

Rainfall Events	Area	Precipitation Data	Daily Precipitation Data Time Zone	SNS Datas	Number of SNS Data Used for Verification
2018-05-16 Jeongneungcheon flood	Jeongneungcheon, Seongbuk-gu, Seoul	Hourly, Daily Precipitation	0 h ~ 23 h	Naver, Daum, Twitter, Instagram	4,175
2018-10-06 Typhoon Conlay	Seoul, Daejeon, Busan				5,115

Fig. 3. Comparison of Data Volume and Daily Precipitation

Fig. 4. Compare SNS Data Frequency by Keyword

Fig. 5. Comparison of Correlation Coefficient Between Daily Precipitation and Keywords

Fig. 6. Comparison of Seoul SNS Data Volume and Hourly Precipitation

과의 상관이 낮다는 의미로 향후 이러한 점을 구분하여 분석이 필요함을 시사한다.

Table 6에서 확인할 수 있듯, 일강우량과 시강우량의 상관계수를 확인한 결과 각각 약 93%와 46%의 상관성을 나타내었다. 이는 Fig. 6의 결과와 같이 인간 활동 시점이 구분되는 시강우량의 경우는 인간 활동 시점에 의해 상관이 떨어지는 시점이 있는 반면 일강우량의 경우는 이러한 시점의 구분이 나타나지 않기 때문에 상관이 매우 높게 나타나는 것임은 방증한다.

Table 6. Correlation Coefficient Between SNS Seoul Data Volume and Rainfall

	Daily Precipitation	Hourly Precipitation
Correlation Coefficients	93%	46%

4.2 2018년 10월 6일 태풍 콩레이 홍수

태풍 콩레이에 대한 소셜 네트워크 서비스 데이터의 효용성을 확인하기 위해 2018년 10월 6일 기준으로 데이터를 수집하여 그 상관성을 비교하였다. 공간적 상관성을 확인하고자 서울, 대전, 부산으로 지역을 설정하여 피해가 가장

크게 났던 10월 6일의 SNS 데이터의 양과 강우량을 비교·분석하였다. 당시 태풍의 경로 우측에 위치한 부산은 서울과 대전과 비교하였을 때 영향력이 커 피해규모가 가장 컸다.

10월 6일의 대전, 서울과 부산의 SNS 데이터량과 각 지역별 강우량을 비교한 결과 Fig. 7과 같은 그래프를 얻을 수 있었다. 태풍이 상륙한 당일 가장 피해가 컸던 부산에서는 350건이 넘는 관련 데이터가 수집되었으나 서울과 대전에서는 50개미만으로 수집되었다. 태풍이 상륙한 같은 시점에서의 지역별 데이터의 양과 태풍의 경로 및 피해규모와 밀접한 연관성이 있음을 확인할 수 있었으며 Fig. 7에서 확인할 수 있듯 일강우량의 패턴과도 유사함을 확인할 수 있었다.

Fig. 8에서 확인할 수 있듯 ‘물폭탄’과 ‘장마’ 키워드를 제외하고는 모두 1과 근접한 CC값을 나타내었다. ‘물폭탄’과 ‘장마’는 빈도수가 10개미만인 반면 중복된 기사가 다음날에도 다시 업로드 되어 상관계수가 비교적 낮은 것으로 확인되었다.

또한 태풍 콩레이 관련 SNS 데이터들에 대한 키워드별 빈도수를 확인하여 Fig. 9와 같은 그래프를 얻을 수 있었다. 태풍 콩레이 사상에서는 ‘태풍’, ‘비’, ‘침수’, ‘폭우’ 등의 순으로 빈도가 높았으며 6일부터 8일까지 3일간의 일강우량과 키워드별 상관계수(CC 값)를 비교한 결과는 Fig. 8의 그래프와 같다.

Fig. 7. Comparison of Data Volume and Daily Precipitation by Keyword and Region

Fig. 8. Comparison of Correlation Coefficient Between Daily Precipitation and Keywords

Fig. 9. Compare SNS Data Frequency by Keyword

Fig. 10. Comparison of Busan SNS Data Volume and Hourly Precipitation

태풍 피해와의 시간적 상관성 분석을 위해 가장 많은 피해가 있었던 부산의 SNS 데이터와 일강우량과 시강우량의 상관계수는 Table 7과 같이 나타났다. 부산의 소셜네트워크 서비스 데이터와 일강우량 및 시강우량의 상관계수 각각 약 98%와 54%로 나타났다. 일강우량은 98%로 매우 높으나 시강우량과의 상관성은 54%로 비교적 낮게 나왔다. 이전의 홍수 사상에서도 확인할 수 있었듯이 사람의 패턴의 증감의 시간차로 인해 낮아지는 것을 알 수 있다.

Table 7. Correlation Coefficient Between SNS Busan Data Volume and Rainfall

	Daily Precipitation	Hourly Precipitation
Correlation Coefficients	98%	54%

또한 부산 SNS 데이터량과 시강우량을 비교한 그래프 Fig 10에서 확인할 수 있듯 강우량이 있을 때 유사한 패턴으로 증가하다가 강우량이 끝나고 점차 줄어들지만 지속적으로 데이터의 양이 있는 것을 볼 수 있다. 이는 피해 발생 이후 피해현상에 대한 기사들로 인해 상관계수가 그의 절반으로 나타난 것으로 확인하였다. 강수가 끝나는 시점에서 강수와 기사 노출 시간의 역관계는 앞서 분석한 시간 단위 분석 시 높은 음의 상관관계를 보이는 결과를 잘 설명해 준다.

5. 결 론

현재 우리나라 홍수 예보는 계측기를 활용한 정형데이터인 수치 데이터를 주로 활용하고 있다. 홍수로 인한 인명 피해 발생 예측은 인구의 유동성 고려가 요구되지만 비정형 데이터의 불확실성으로 인해 우리나라에서는 활용을 못하고 있는 실정이다. 반면 해외에서는 홍수나 지진과 같은 재해 및 재난 발생 시 다양한 채널의 소셜 네트워크 서비스를 통하여 재난 발생 시 대피로나 인명구조에 활용된 사례가 있다. 따라서 본 연구에서는 우리나라의 홍수 발생 시 비정형 데이터를 활용하는 방안을 제안하고자 클라우드 소싱 기반 웹크롤러를 개발하였다. 키워드를 기반으로 하여 웹사이트 데이터 추출을 함으로써 홍수와 관련성 높은 데이터를 얻을 수 있었으며 실제 홍수 사례와 비교하여 시·공간적 상관성을 분석하였다. 2018년에 발생한 두가지 홍수 사상인 정통천 홍수사례와 태풍 콩레이 사례를 이용하였으며 두 사상 모두 SNS 데이터의 양과 일일 강우량이 대체적으로 비슷한 경향을 보였다. 키워드별 빈도수는 사상마다 차이가 있으나 대체적으로 유사한 결과가 나왔으며 빈도수가 높은 키워드의 경우 지역별 데이터의 양에 대한 상관계수가 매우 높게 나오으로써 시·공간적 상관성이 매우 높은 것으로 확인되었다. 낮은 상관성을 보이는 키워드는 주로 신문기사에서 쓰이는 데이터로 나타났다. 보다 정형화되어있는 기사보다 SNS데이터만을 분석에 활용하는 것 또한 고려해 보아야할

것이다. 이를 통하여 SNS 데이터와 국부적으로 발생하는 강우사상 뿐만 아니라 전 지역에 걸쳐 발생하는 태풍 피해 예측에도 SNS 데이터와 같은 비정형데이터의 적용이 가능할 것이라고 사료된다. 더불어 본 결과는 홍수 외에도 미세먼지, 화재, 범죄, 지진, 산사태, 산불, 가뭄 등 인간이 체감 가능한 어떠한 재해 재난에도 적용할 수 있고 피해 정도를 실시간으로 검증할 수 있다는 점에서 매우 활용성이 높을 것으로 판단되므로 향후 보다 다양한 연구가 필요하다.

감사의 글

본 연구는 한국건설기술연구원 ‘커뮤니티 빅데이터 패턴 해석을 통한 수난(水難) 발생 및 규모 예측 기술 개발’ 연구사업의 연구비 지원에 의해 수행되었습니다.

References

Johansson, F., Brynielsson, J., and Quijano, M.N. (2012). Estimating citizen alertness in crises using social media monitoring and analysis. *2012 European Intelligence and Security Informatics Conference*, IEEE, pp. 189-196.

Kim, H.G. (2011). Application of social media for responding to a national disaster. *Journal of the Korea Entertainment Industry Association*, Vol. 5, No. 4,

pp. 147-153.

Kim, Y.W., Kim, B.H., Ko, G.S., Choi, M.W., Song, H.S., Kim, G.H., et al. (2017). Design and implementation of a flood disaster safety system using realtime weather big data. *The Journal of the Korea Contents Association*, Vol. 17, No. 1, pp. 351-362.

Murthy, D., and Longwell, S.A. (2013). Twitter and disasters: The uses of Twitter during the 2010 Pakistan floods. *Information, Communication & Society*, Vol. 16, No. 6, pp. 837-855.

Taniguchi, S. (2012). About the usefulness of twitter at the time of disaster. *Disaster Information*, Japan Society for Disaster Information Studies, Vol. 10, pp. 56-67.

Zhang, C., Zhou, G., Yuan, Q., Zhuang, H., Zheng, Y., Kaplan, L., et al. (2016). GeoBurst: Real-time local event detection in geo-tagged tweet streams. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 513-522.

Received	October 31, 2019
Revised	October 31, 2019
Accepted	November 8, 2019