# 웹 뉴스 빅데이터를 이용한 태풍 상황정보의 인터렉티브 지도 기반 시공간 시각화 방안

이지애*, 김준철*
*서울기술연구원 스마트도시연구실
jiaelee@sit.re.k, kjc@sit.re.kr

# Interactive Map-based Spatio-Temporal Visualization of Typhoon Situation using Web News BigData

Jiae Lee*, Junchul Kim*
*Department of Smart City, Seoul Institute of Technology
jiaelee@sit.re.kr, kjc@sit.re.kr

요 약

웹 뉴스 기사는 태풍과 같은 재해 발생상황에 대한 신속하고 정확한 정보를 포함하고 있다. 예를 들어, 태풍의 발생시점, 이동·예측경로, 피해·사고 현황 등 유용한 정보를 텍스트, 이미지, 동영상의 형태로 관련 상황정보를 전달한다. 그러나 대부분의 재해재난 관련 뉴스 기사는 특정 시점의 정보만을 웹페이지 형태로 제공하므로, 시계열 측면의 연결성을 지니는 기사들에 대한 정보를 전달하기 어렵다. 또한 시간적 변화에 따라 기사 내용에 포함된 장소, 지역, 건물 등의 지명에 대한 공간적 정보를 지도와 연계하여 정보를 전달하는데 한계가 있어, 시공간적 변화에 따른 특정 재해재난 상황정보에 대한 전체적인 현황파악이 어렵다. 따라서, 본 논문에서는 데이터 시각화 측면에서 이러한 한계를 극복하기 위해, 1) 웹크롤링을 통해 구축된 뉴스 빅데이터를 자연어 처리를 통해 태풍과 관련된 뉴스 기사들을 추출하였고, 2) 시공간적 관련 정보를 지식그래프로 구축하였고, 이를 통해 최근 발생한 태풍 사건들과 관련된 뉴스 정보를 시계열 특성을 고려하여 3) 인터렉티브 지도 기반의 태풍 상황정보를 시각화하는 방안을 연구하였다.

## 1. 서론

웹 뉴스 기사는 재해재난과 관련된 일련의 사건들에 대한 시간, 위치, 상황 등을 포함한 객관적이고 정확한 정보를 신속성 있게 제공한다. 사용자는 보도된 기사 정보를 통해서 재해재난 등에 대한 정보를 신속하게 파악할 수 있다.

그러나 보도된 정보는 특정 시점에 대한 정보를 웹페이지 형태로 제공하기 때문에, 변화하는 상황에 따라 추가 보도되는 뉴스 정보를 시계열 측면에서 상호연계하는 데이터 시각화 방안이 필요하다. 실시간 뉴스 정보는 인터넷 미디어 또는 검색포털을 통해 사용자, 기업, 기관 간 공유되고 있다. 특히, 국내에서는 네이버와 다음이 대표적인 뉴스 공유매체로서, 다양한 기사를 공유 및 제공하기 때문에 사용자 이용률이 높고, 뉴스 기사 수집을 통한 연구 분야에서 활용도가 높다.

기존 선행연구를 살펴보면, 재해재난 분야에서는 소셜미디어(뉴스 등)와 SNS(트위터 등) 데이터를 이용한 연구가 진행되어왔다. 국립재난안전연구원은 소셜빅데이터를 실시간으로 모니터링하고 재난사건을 탐지하는 스마트빅보드를 개발했다[1-2]. 홍수 재해에 따른 재난 연구중에는 SNS 데이터에서 홍수 관련 키워드들의 빈도수를 분석하여 시공간 위치를 분석하는 연구가 수행되기도 했다[3]. Thom *et al.*은 트위터 메시지 정보를 공간에 매핑하여 이상탐지를 하는 연구를 진행했다[4]. 그러나, 기존 연구들은 트위터와 같은 주관적인 정보를 포함하는 매체 데이터만을 사용하거나 특정 키워드에 대한 빈도분석, 사

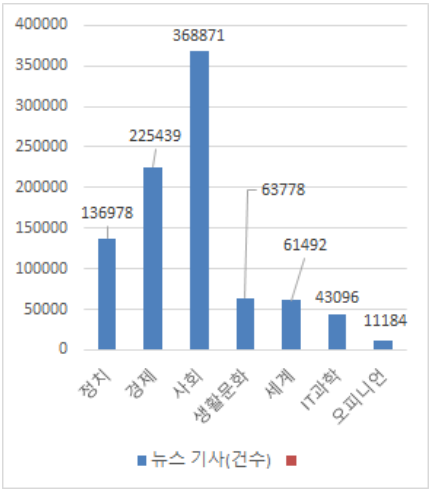
건 상황정보와 시공간 정보를 동시에 시각화하여 분석하기 힘들다는 한계가 있다.

이에 본 논문에서는 재해재난의 시공간 분석과 시각화를 위해 뉴스 기사를 이용하며, 특정 재해 사건에 관한 시공간 정보를 뉴스 본문에서 추출하고, 시계열에 따른 공간정보를 지오코딩하여 인터렉티브 지도 기반에 태풍 상황정보를 시각화하는 방안을 제시하고자 하였다. 본 연구의 실험에서는, 최근 발생된 태풍 재해 사건들을 중심으로, 웹 뉴스 빅데이터를 이용한 태풍 상황정보의 인터렉티브 시공간 시각화를 수행하였다. 특히, 재해재난에서 실시간 모니터링을 위해서는 재해재난 사건의 위치와 시간정보가 중요하므로, 시간에 따라 상황변화를 지도와 연계하여 이를 추적 가능하게 하였다.

## 2. 데이터 수집 및 처리

1) 원천데이터

웹 뉴스 빅데이터는 네이버 뉴스를 웹크롤링하여 데이터베이스로 구축하였다(KoreaNewsCrawler[1]를 이용). 최근 태풍 발생시점을 고려하여 2020년 7월부터 9월 초까지 기간에 대해서 선별하였으며, 본 실험을 위해 사용된 뉴스 기사 7개 분야별 건수는 (그림 1)과 같다.




(그림 1) 뉴스 기사 분야별 건수

2) 방법론

뉴스 기사의 제목과 본문에서 재해재난 사건 키워드를 검색하여 추출하였다. 추출된 뉴스 본문에서 자연어처리 모델 BERT-CRF(Bidirectional Encoder Representation Transformers-Condition Random Field)로 <표 1>과 같이 시간과 공간에 관련된 개체명 인식(NER, named entity relation)을 통해 6개 태그(tag)에 대해 추출하였다. 추출한 개체들은 시공간 관계를 구성하여 분석을 효율적으로 수행하기 위해 지식그래프(knowledge graph)로 구축되었다(그림 2). 각 노드(node)는 각 태그로 추출된 개체를 의미하며 간선(edge)의 굵기는 기사에서 한 문장안에서 언급된 개체들간의 빈도수를 가중치화하여 상호관계를 연결하여 표현한 것이다.

<표 1> 개체명 인식 관련 시공간 범주와 태그

| 범주 | 시간 | 공간 |
|---|---|---|
| 태그 (tag) | DAT(date): 날짜 | LOC(location): 지명 |
| | TIM(time): 시간 | ORG(organization): 기관명 |
| | DUR(duration): 기간 | POH : 기타 |

(그림 2) 태풍 관련 공간 키워드 그래프 일부(8월 27일)

추출한 공간 단어는 한 단어 노드 간선의 수를 측정해 각 노드의 총 빈도수를 구한다. 언급이 많은 지명을 내림차 정렬하여 랭킹을 정한다. 순위가 높은 장소는 OpenStreetMap API[2]로 위도/경도 정보를 지오코딩해 수집하였다. 본 실험에서는 태풍과 관련된 249개 지명이 인식되었고 182개가 성공적으로 매칭되어 약 73.1%의 정확도(Precision=TP/(TP+FP))로 나타났다. 지명에 대한 지리좌표를 이용하여 오픈소스 기반 Processing[3]과 Unfoldingmaps[4]을 활용해 지도 위에 마커로 가시화한다. 위치 마커는 3가지 색상 스케일(적-녹-파)로 시각화한다. 빈도(노드 간선 수)를 최소/최대 정규화(min-max normalization)로 정규화한다. 마커는 원형으로 표현하고 빈도수는 크기에 비례하

1) KoreaNewsCrawler( https://github.com/lumyjuwon/KoreaNewsCrawler)

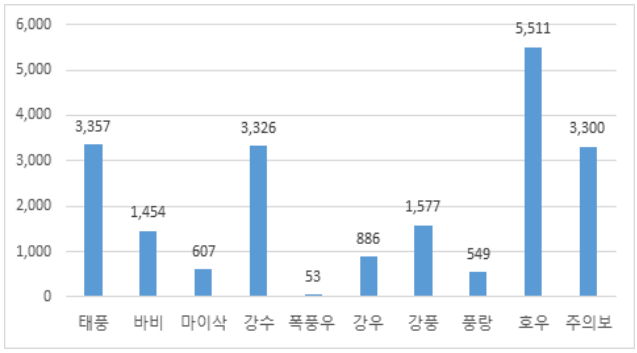
2) OpenStreetMap(https://www.openstreetmap.org/)
3) Processing (https://processing.org/)
4) Unfoldingmaps(http://unfoldingmaps.org/)

고 빈도가 높을수록 적색으로 표현한다.

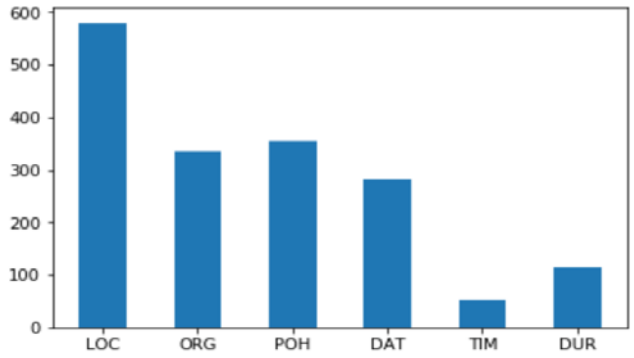
## 3. 실험 및 결과

태풍 발생 기간인 7월 1일부터 9월 2일까지 태풍과 관련된 뉴스 기사는 5월 163건, 6월 161건, 7월 140건, 8월 2,923건, 2020년 총9 월초까지 4,249건 중 8월은 68%의 비중을 차지한다<표 2>.

키워드는 적정성 확인을 위해 연관어와 함께 뉴스 기사 검색하고 통계낸다. 연관어는 이름(바비, 마이삭), 강수, 폭풍우, 강우, 강풍, 풍랑, 호우, 주의보[5] 이고, 이를 포함하는 기사와 본문은 10,600건이다. 높은 빈도의 키워드는 호우, 주의보, 태풍이다(그림 3).




(그림 3) 7~9월 전체 키워드 및 연관키워드 포함 뉴스




(그림 4) 7월 태풍 키워드 포함 기사의 태그 빈도

1) 7월 뉴스 기사 월 분석

7월 뉴스 기사는 태풍을 키워드로 한 달 전체에 대한 연관어 1,718개를 추출했다. 빈도수는 LOC 579개, ORG 335개, POH 354개, DAT 282개, TIM 53개, DUR 115개이고, LOC의 빈도가 가장 높다(그림 4).

(그림 5)는 7월 중 가장 많은 빈도수를 보인 공간정보 20개<표 3> 중 노이즈값을 제외하고 지도에 표현한다. 서울과 제주 부근에서 빈도수가 가장 높은 것으로 보인다. 7월은 제주도가 가장 많이 언급되었고 반도 등의 POH 값을 포함한다<표 3>.

5) 표준국어대사전(http://stdict.korean.go.kr)

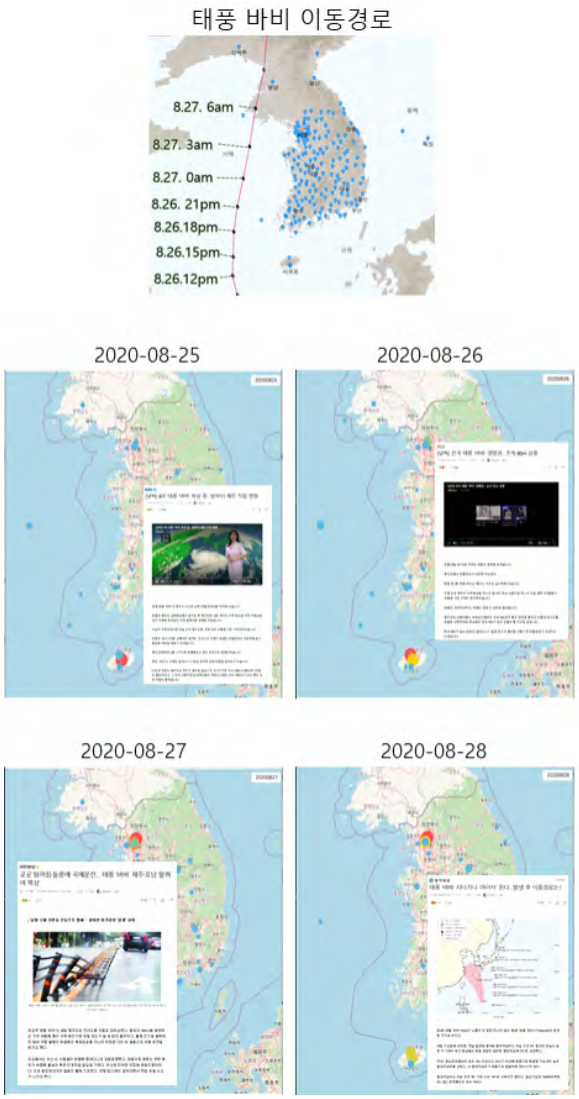
<표 2> 태풍별 발생 및 소멸 시간[5]

| 태풍명 | 발생-소멸시간(KST) |
|---|---|
| 장미 | '20.08.09 03:00 ~ '20.08.10 17:00 |
| 바비 | '20.08.22 09:00 ~ '20.08.27 15:00 |
| 마이삭 | '20.08.28 15:00 ~ '20.09.03 12:00 |
| 하이선 | '20.09.01 21:00 ~ '20.09.07 21:00 |


(그림 5) 7월 전체 건수 중 장소 빈도 표현

<표 3> 7월 태풍 키워드 검색 결과

| 순위 | 개체명 | 빈도수 | 태그 | 날짜 |
|---|---|---|---|---|
| 1 | 제주 | 9,855 | POH | 20200703 |
| 2 | 서울 | 8,305 | LOC | 20200701 |
| 3 | 부산 | 7,464 | LOC | 20200701 |
| 4 | 반도 | 6,934 | POH | 20200712 |
| 5 | 영동 | 6,541 | LOC | 20200701 |
| 6 | 코로나19 | 4,442 | POH | 20200701 |
| 7 | 안마도 | 4,128 | LOC | 20200714 |
| 8 | 제주도 | 3,772 | LOC | 20200704 |
| 9 | 한국 | 3,604 | LOC | 20200704 |
| 10 | 한반도 | 3,515 | LOC | 20200701 |
| 11 | 수도권 | 3,320 | LOC | 20200701 |
| 12 | 충청 | 3,256 | LOC | 20200709 |
| 13 | 강릉 | 3,140 | LOC | 20200701 |
| 14 | 기상청 | 3,106 | ORG | 20200701 |
| 15 | 영덕 | 3,061 | LOC | 20200724 |
| 16 | 동해안 | 2,993 | LOC | 20200701 |
| 17 | 거제도 | 2,874 | LOC | 20200707 |
| 18 | 난곡 | 2,736 | LOC | 20200704 |
| 19 | 한라산 | 2,360 | LOC | 20200713 |
| 20 | 남해안 | 2,253 | LOC | 20200709 |




(그림 6) 태풍 바비 2020년 8월 26일부터 27일까지 실제 이동경로와 뉴스 빅데이터 기반 인터렉티브 지도 연계 태풍 상황정보 시각화

2) 8월 뉴스 기사 시계열 분석

뉴스 기사로부터 태풍의 상황정보를 시공간 시각화하기 위해, 지오코딩 결과의 공간범위를 한반도인 위도/경도(33.0~38.62/124.0~132.0)로 설정하여 최상위 빈도수를 갖는 지명들을 추출하여 시각화하였다. (그림 6)은 태풍 바비의 8월 26일부터 27일까지의 이동경로와 동일 기간에 대한 뉴스 빅데이터로부터 추출된 상위 30개의 공간개체를 지도와 연계하여 태풍 상황에 대한 뉴스정보를 시각화한 결과를 보여준다. 제주도지역에서 태풍 바비의 직접 영향권에 들어오는 동일 지역 및 지명이 높은 빈도를 보였으며, 서울에서는 여러 지명을 포함한 단어가 각각 낮은 빈도수로 나타났다. 최근 9월 2일에 수집된 뉴스 빅데이터로부터는 태풍 마이삭이 상륙한 시점으로 부산지역의 지명이 다수 추출되었다. 마이삭 이동 경로와 직접적으로 인접한 부산이 높은 빈도를 보였다.

따라서, 실험 결과로부터, 뉴스 빅데이터 기반 태풍 상황정보의 시각화 방안이 시계열 측면의 연결성을 지니는 기사들에 대한 정보를 효과적으로 전달하고, 시간적 변화에 따라 기사 내용에 포함된 장소, 지역, 건물 등의 지명에 대한 공간적 정보를 지도와 상호연계하여 정보를 전달하는 것이 실현 가능함이 검증되었다.

## 4. 결론

본 논문에서는 뉴스 기사에 대한 자연어 처리 및 개체명 인식을 통해 시간 및 공간정보를 추출하여 시공간적 분석이 가능한 지식그래프를 구축하였으며, 이를 통해 재해재난에 대한 상황정보를 효율적으로 시각화하는 방안을 제시하였다. 결과적으로 재해재난 사건의 상황정보를 인터렉티브 지도 기반에서 시공간 시각화하는 방안의 실현 가능성을 실험을 통해 검증하였다. 향후 재해재난 알림서비스 및 SNS 데이터 등의 데이터를 추가 활용하여 분, 시간 단위별 상세 정보추출을 자동화하여 실시간 정보 분석도 가능할 것으로 기대된다.

## Acknowledgement

본 논문은 서울기술연구원(2020-AD-001, 서울 대도시권 데이터 사이언스 체계 구축방안)의 지원을 받아 수행된 연구임.

## 참고문헌

[1] 최선화, 배병걸, “소셜 빅데이터로부터의 재난이슈 탐지 모델”, 한국정보과학회:컴퓨팅의 실제 및 레터, Vol. 20, No. 5, pp. 286-290, 2014.

[2] 최선화, 배병걸, 이보람, “소셜 빅데이터 실시간 재난 모니터링, 소셜 빅데이터 실시간 재난 모니터링”, 대한토목학회 학술대회, pp. 255-256, 2014.

[3] 이정하, 황석환, “홍수 피해 발생 감시를 위한 소셜 네트워크 서비스 데이터 활용 방안 연구”, 한국방재학회 논문집, Vol. 19, No. 7, pp. 77-85, 2019.

[4] Dennis Thom *et al.*, "Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages.", 2012 IEEE Pacific Visualization Symposium, IEEE, pp. 41-48, 2012.

[5] 기상청(https://www.weather.go.kr/)