Exploring Spatiotemporal Characteristics of Twitter data Using Topic Modelling Techniques

Young-Hoon Kim [a, *], Hyun-Jee Woo [b]

[a, b] *Department of Geography Education, Korea National University of Education, Cheongju, Republic of Korea, gis@knue.ac.kr*

* Corresponding author

Abstract:

This research purpose aims to explore the spatiotemporal aspects of social media data with Twitter data by using topic modelling techniques. The spatiotemporal limits are restricted in two areas of the Republic of Korea: Seoul and Jeju Island. This paper searches the semantics and geographical place characteristics of the Twitter data, and the semantics and place characteristics are regarded as topics in the topic modelling. This paper also discusses the temporal intensity over different spatial areas and visualizes the spatiotemporal patterns with GIS techniques.

As Twitter mobility message contains a user's interests and behavioural patterns in the geo-tagged data corresponding to its location, it is possible to explore geographical locality and the user's mobility over space by using textual ontological techniques such as topic modelling. Therefore, this paper attempts keywords searching and textural classification to classify the shared spatial activity patterns of the Twitter users. Consequently, two main analysis themes are explored: the tourist activity patterns attracting the visitors in Jeju over time and temporal periodicity for shopping and meal preference in Seoul, respectively.

In conclusion, this research represents a potential of the social network data that enables to fill the gap of spatiotemporal patterns of human beings over the online and mobile environment. Furthermore, our study confirms social data analysis techniques as an alternative geographical data source that can complement and replace the roles of spatial data, which could not be analysed in the conventional offline data.

Background

Nowadays, one can access social media anytime and anywhere with mobile devices. Social Network Services (SNS), in particular, provide new spatial data and information through channels such as Twitter. Social media data contain various socio-economic information over space and time and volunteer information from users. For instance, a Twitter message contains a user's interests and behavioral patterns in the geo-tagged data corresponding to its location. It is also possible to explore geographical locality and user mobility over space by using textual ontological techniques such as topic modeling. Conventional discussions for Twitter's spatiotemporal patterns have been mainly articulated on temporal variations of hotspots and their visualization, with comparisons of weekday versus weekend periods. One can identify the hotspots and map their clustering of Twitter data, which allow analyzing patterns and variations over space and time, as characterized by spatiotemporal patterns.

Since people are questioning why tweets concentrate on some specific hotspots, for what certain reasons they occur, and what people tweet in those hotspots, this research's objectives are as follows. First, exploring the spatiotemporal aspects of social media data with Twitter data by using topic modelling techniques is aimed. Next, temporal intensity over different spatial areas is discussed and spatiotemporal patterns are visualized with GIS techniques. Keyword searching and textural classification to classify the shared spatial activity patterns of the Twitter users are also attempted while two main analysis themes are explored: the tourist activity patterns attracting the visitors in Jeju Island over time and the temporal periodicity for shopping and meal preference in Seoul.

Analysis Results on Jeju Island

The data collection periods of tweets for the topic modelling of Jeju Island were from June to September, 2016, and July to September 2017. The total number of cases collected was 4,472 geotagged tweets of Jeju Island. Jeju Island's tweets were mainly made by the tourists, and their visited places were remarkable in the tweets. This study searched the most

relevant topics through topic modeling, exploring the common behavioral characteristics of the tourists through inter-relational network analysis. Jeju Island's tweets mostly concentrated around Jeju Airport, around Jeju East Gate, and the shorelines of Jeju Island due to the summer season.

The places where one could clearly find the themes of the Twitter through topic modelling reflected the preference of the tourists. These places were a tourist destination or an area where popularity had increased recently among young people. This place preference corresponded to the results of the probabilistic topic modeling analysis. For example, topic one among 15 topic groups generated from the topic modeling included several places of those coastal sites in Jeju. This was the result of the preference in action of Twitter users who had participated or became interested in the festival. By using topic modeling methods with Twitter data, it became possible to search the characteristics of a small-scaled area that newly constituted the geographical features of the area.

The spatiotemporal distribution of the Twitter data showed similar patterns with that of a previous survey on hourly result based on Jeju population flow undertaken by the researchers. It would be a new potential on the use of Twitter data that can replace the conventional survey methods of hour-based population flow data. In addition, the accumulation of tweet data with flexible temporal criteria would assist the enhancement of the tweets' data analysis accuracy. It was also possible to assume the tourist preference that was related to a specific topic, such as why a specific place was popular, when tourists were concentrated, and which themes were correlated with tourist concentration at a specific place.

Analysis Results of Seoul

For the Twitter users' pattern of food-related themes in Seoul, the total number of GPS-tagged Tweets was 7,653 cases, and the tweets occurred intensively 1) near Seoul city hall, 2) around Gangnam area, and 3) near Hongil University subway station area, where the restaurants and shops populate the vicinity, emerging as food-related topics. This study identified those three place preferences and hotspots of food tweets which were correlated statistically. In those places where the density of food-related Tweets was high, patterns similar to actual restaurant and shop distribution were shown.

Meanwhile, there were some places where the density of eating and drinking establishments was low, but the occurrence of tweets was concentrated, relatively paying attention to the shops and restaurants. Twitter users had more interest in distinctive eating places such as having unique features of recipes and cuisines compared with regular restaurants; that is, with Twitter, it was easy to search places that interested the general public and that explored the geographical characteristics of the places or areas having unique features. The following was the analysis result of tweets on a food theme in Seoul.

First, the temporal pattern of the food tweets showed a distinct difference from the spatial pattern. The tweets on the morning, afternoon, and evening differed in terms of main topics in the textual messages such as when the tweet occurrence was high in the morning, afternoon, or evening. Further, many tweets referred to meal periods in both lunch and dinner times. Although tweet occurrence was high in lunchtime, it was not possible to identify which places were most popular or highly demanded by the people who were looking for lunch instead of seeking for a coffee break or doing shopping or sports activities.

Topic modeling and its text mining techniques were applied to search the places and indicated business districts in the city centre and suburban areas as hotspots for lunch-related themes. Popular areas for dinner were downtown Seoul and those near Hongdae subway station areas which were attractive to tourists. A total of 661 tweets cases of "lunch" as keyword were collected, and topics containing "Myung-dong," an old city centre in Seoul, were accompanied by several keywords referring to "late lunch meal." Thus, one can assume that the lunch meal at Myung-dong was closer to "late lunch" than site-seeking or shops for lunch.

Thus it was assumed that "Myung-dong" was preferred as "a place for late lunch meal" instead of night-time shopping or dinner meal. Although the location being Myung-dong was not clear, it was also possible to assume where in Seoul city it was popular to eat noodle or pizza at lunch time. In this way, the topic model analysis was effective for searching where high correlated keywords were clustered. In those areas, it was easy to extract adequate themes that explained the area in detail.

Discussion and conclusion

This paper demonstrates a potential possibility of a new place searched for exploration. One can infer the places where general public attention has been growing recently (i.e., Stepping Stone Festival in Jeju, UNPLUGGED in Seoul). This paper also represents a potential of the social network data that enables to fill the gap of spatiotemporal patterns of human beings over the online and mobile environments. As supplementary data, this paper shows that social network data can supplement or replace the roles of spatial data, which could not be analyzed in the conventional offline data. In addition, the outcomes of this paper confirm several advantages of Topic modelling for Twitter data analysis over conventional

text mining techniques. First, as every single tweet is geotagged, containing x and y coordinates of the occurrence, it is possible to explore spatiotemporal behavioral patterns among Twitter users. Second, it calculates the correlation between the Twitter texts containing location information. Third, this correlation makes it possible to search for clusters of a topic's keywords that are extracted by topic models. Finally, the clustered topics are classified with several keywords, and geographical characteristics of the clustering are explored.