

Establishing the Process of Spatial Informatization Using Data from Social Network Services

Eo, Seung-Won¹⁾ · Lee, Youngmin²⁾ · Yu, Kiyun³⁾ · Park, Woojin⁴⁾

Abstract

Prior knowledge about the SNS (Social Network Services) datasets is often required to conduct valuable analysis using social media data. Understanding the characteristics of the information extracted from SNS datasets leaves much to be desired in many ways. This paper purposes on analyzing the detail of the target social network services, Twitter, Instagram, and YouTube to establish the spatial informatization process to integrate social media information with existing spatial datasets. In this study, valuable information in SNS datasets have been selected and total 12,938 data have been collected in Seoul via Open API. The dataset has been geo-coded and turned into the point form. We also removed the overlapped values of the dataset to conduct spatial integration with the existing building layers. The resultant of this spatial integration process will be utilized in various industries and become a fundamental resource to further studies related to geospatial integration using social media datasets.

Keywords : Social Network Service, Spatial Informatization, Spatial Integration, Data Conversion

1. Introduction

The data acquired from SNS (Social Network Service) commonly have various information such as text, location, and time information describing specific life events. There are several typical methodologies to utilize these information. First, the technique called 'text mining' is needed to extract the contents including texts related to the certain topics or events. The attempts made by Mei *et al.* (2006) and Dhawan *et al.* (2014) to analyze social trends and human emotion using the text contents of social media are the studies related to mining the text.

Another one is the technique to conduct the spatio-temporal analysis using the location and time information extracted from SNS datasets. The studies by Hughes and Palen (2009), and Cheng and Wicks (2014) examined the

spatio-temporal changes of a specific event using Twitter. A lot of the existing studies using SNS datasets have focused on these two techniques. These kinds of studies try to extract the keyword related to certain topics or events and sentimental words from the datasets.

The text mining using SNS data means not only for analyzing the keyword but also spatially valuable. Cheng *et al.* (2010) and Dashti *et al.* (2014) utilized textual contents of the SNS data to gather the spatial information for their studies. Although these efforts have been made, understanding the characteristics of valuable information extracted from SNS datasets leaves much to be desired in many ways. The reason is that there are various ways to process the datasets depending on the purpose of analysis. GIS (Geographic Information Science) is one of the fields to fulfill the desire to process the datasets with various approach. The GIS industry

Received 2015. 12. 01, Revised 2015. 12. 16, Accepted 2016. 01. 13

1) Dept. of Civil and Environmental Engineering, Seoul National University (E-mail: esw1026@snu.ac.kr)

2) Dept. of Civil and Environmental Engineering, Seoul National University (E-mail: daldanka@snu.ac.kr)

3) Member, Dept. of Civil and Environmental Engineering, Seoul National University (E-mail: kiyun@snu.ac.kr)

4) Corresponding Author, Member, Korea Land and Geospatial InformatiX Corporation (E-mail: wjpark@lx.or.kr)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

interprets social phenomenon or events with geographic traits and tries to integrate the information extracted from SNS datasets with spatial information. The works by Sakaki *et al.* (2010), Gomide *et al.* (2011) and Noulas *et al.* (2011) are the studies that can be classified in this range.

As importance to perform the GIS related-works using SNS data increases, the integration with social media becomes more emphasized (Sui and Goodchild, 2011). In this regard, prior knowledge of the details of SNS datasets is required in order to utilize them properly. To apply SNS datasets to various analysis, the process to convert the datasets into the form of spatially valuable datasets should be defined to analyze not only text information but also the information about location and time for spatio-temporal analysis.

There is also a lack of efforts to spatially join the information from social media with existing urban environment such as building layers. All the studies mentioned above have extracted location information to analyze a certain events or patterns like the study by Lwin *et al.* (2015). There are several studies in the traditional spatial data integration field. Flowerdew (1991) tried to solve the incompatibilities between the spatial entities and Mohammadi *et al.* (2006) tried to handle the spatial data integration problems, but their interest heavily relied on the major policies that interrupt the

application of spatial data on built environment. Cruz (2004) also focused on geospatial data integration, but did not pay attention to social media database.

In this regard this paper aims on analyzing the detail of the SNS datasets to establish the process to integrate social media information with existing spatial datasets. This process is called 'spatial informatization' in this study. Fig. 1. represents the work flow of the entire process.

In the next section we explains the most important part of this study, analysis of the fields of SNS datasets in every detail which is required for the analysis related to the studies in GIS. Section 3 elaborates the spatial informatization process with the collected SNS datasets and conduct the spatial join (data integration process) with existing spatial datasets. We also try to visualize the results of the process. Finally in section 4 we discuss the spatial integration process established through section 2 to 3 and emphasize the significance of the fields of the SNS datasets that we have chosen. We also discuss the difficulties handling SNS datasets in this section.

2. Understanding of SNS Datasets

2.1 Target SNS

The data acquisition process from SNS is different from each other depending on the regulations of each service, and most services offer some parts of their data via Open API (Application Programming Interface) for free. We selected three social network services, Twitter, Instagram, and YouTube, among the services frequently used in related researches. The following is the selection of the fields of each SNS dataset. We've performed the spatial join with the building layers dataset using the SNS dataset which has been extracted based on the selection of the fields.

2.2 Twitter

Twitter offers Search, REST⁵⁾, and Streaming APIs. Each API provides different options to collect data. Search API is the API to collect the information about the Twitter contents using the query type such as a user information or

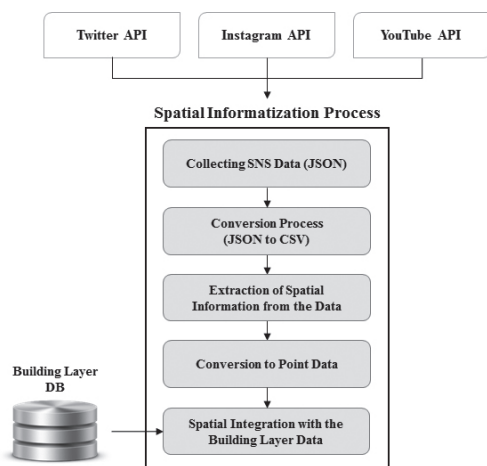


Fig. 1. Workflow of this study

5) REST (REpresentational State Transfer) : One of the software architecture styles in the World Wide Web

a certain keyword. REST API is able to collect and access to the Twitter’s core information like timeline, status update, and user information. This API also offers a function to read and write which gives an opportunity to develop applications based on Twitter. Both Search, and REST APIs, however, have a limit on calls and it is impossible to make 180 calls per fifteen minuets using the APIs. Streaming API offers low latency access to Twitter’s global stream of Tweet data, and collects Tweets in real-time. Streaming API is also preferred by many who to conduct data mining or keyword analysis, because using this API can quickly collect the massive amount of data. Nevertheless, this API is not easy to use personally, because it requires expertise to install the infrastructure for the data collection.

In this study, REST API has been used to collect geo-tagged data which is in the form of JSON⁶⁾ because the API is affordable for CRUD⁷⁾ functions. The process of data

collection is described in Fig. 2. Twitter offers four different objects which are Users, Tweets, Entities and Places. Each object has the fields with unique values, and each field divides into sub-fields.

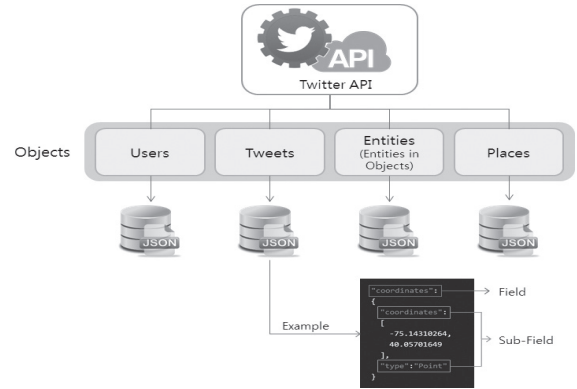


Fig. 2. Flow chart of collecting the data via Twitter API

Table 1. The selected fields of Twitter dataset

Objects	Field	Explanation	Data Type	
Users	id	The user ID	integer	
Tweets	created_at	The Time that the Tweet has been created. Based on UTC (Universal Time Coordinated)	string	
	id	The ID of the Tweet	integer	
	text	The text information of the tweet based on UTF-8 standards	string	
	retweet_count	The number of times the Tweet has been retweeted	integer	
	favorite_count	Represents how many times the tweet has been ‘liked’ by other Twitter users	integer	
	coordinates	coordinates	The location information of the Tweet. Represented in geoJSON (lng, lat)	float
Entities	hashtags	text	The Text information included in a hashtag	string
Places	id	The ID of the Place’s location	string	
	country	The name of the country	string	
	country_code	The code to represent nation	string	
	place_type	The type of the place (ex: city)	string	
	name	The short name of the place (ex: Paris)	string	
	full_name	The full name of the place (ex: San Francisco, CA)	string	
	url	The URL address, including the additional information about the place	string	

6) JSON (JavaScript Object Notation) : a readable format for structuring data as an alternative to XML

7) CRUD (Create, Read, Update, Delete) : the fundamental functions of data processing

Among the various fields, the selected fields in this study are described in Table 1. 'Id' of the 'Users' object, 'created_at', 'id', 'text' of the 'Tweets' object and 'hashtags_text' of the 'Entities' object are required to analyze the contents of Twitter datasets. The both 'id's have a role of identifier to classify users and places. 'text' field and 'hashtags_text' are often regarded as the most important resources to conduct text and sentiment analysis. On the other hand, 'coordinates_coordinates' of the 'Tweets' object and all the fields that the 'Places' object possesses are the essential fields to be acquired for spatial analysis.

2.3 Instagram

Instagram has increasing growth of use and popularity during recent years. Images or videos shared by its users are challenging to be mined because the Open API offered by it has a limit to collect data, so that only twenty images can be collected by the API at one call. In this reason, the iterative module to collect data should be developed. Endpoints, an individual data entity with a unique identifier, have to be used to collect data via Instagram API. Instagram currently offers the eight types of endpoints in total. Among those endpoints, spatial information is included in 'media', 'tags', and 'locations.' In this study, media endpoint is chosen because it contains the media information which is the core content of Instagram. There are four different types of media endpoints and the characteristics of them are described in Table 2.

The four media endpoints provide the information of the geo-tagged data only, but '/media/search' offers the parameters to select the areas for data collection which means it is the most valuable endpoint for the researches using Instagram datasets. For example, Hochman and Schwartz (2012) analyzed how Instagram is used to communicate the visitor's experiences while visiting a museum of natural history. They used the parameters offered from '/media/search' to collect the images created only in the two cities, New York and Tokyo, to conduct the analysis. When making a request to Instagram API, '/media/search' offers five different parameters⁸⁾ to acquire the data, which are latitude (LAT), Longitude (LNT), minimum time (MIN_TIMESTAMP), maximum time (MAX_TIMESTAMP), and distance (DISTANCE) (Table 3). The process to collect the data using these parameters is described in Fig. 3.

Among the available fields that can be collected via Instagram API, the selected fields in this study are presented in Table 4. Among these selected fields, 'id', 'tags', 'created_time', 'images_url', and 'caption_text' are needed to analyze the contents of the data. Especially 'tags' has attracted many researchers because it indicates 'hashtags' which provides a link to related posts with the same hashtag. A study of Karimkhani et al. (2014) that collected dermatology-related contents on Instagram focused on analyzing the hashtags related to their topic. 'Location' is the mandatory field for spatial analysis.

Table 2. The types of media endpoint and its characteristics

Media Endpoints	Characteristics
/media/media-id	Provides the information of the media object, and the return value, 'type', represents whether the collected data is an image or a video
/media/shortcode/shortcode	Provides the same information as media-id offers. (ex. if the link URL of some data is http://instagram.com/p/D/ , D means shortcode)
/media/search	Provides the information of the media data in a given area selected by a user. It is basically set up to search the data created within 5 days. Although a user changes the setting to search data, it won't be over 7 days
/media/popular	Provides the most popular media data when collects

8) Instagram API rate-limits and behaviors have been newly updated on Nov 17, 2015. The updated /media/search endpoint offers the 4 parameters: access token (ACCESS_TOKEN), latitude (LAT), longitude (LNG), and distance (DISTANCE). The 3 parameters except access token function exactly the same as described in Table 3. The access token parameter provides a valid access token.

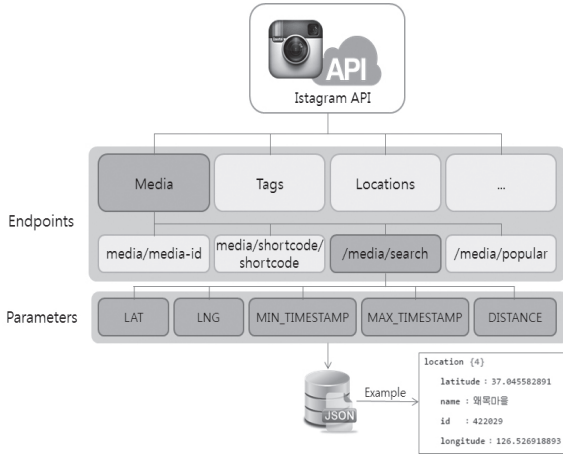


Fig. 3. Flow chart of collecting the data via Instagram API

2.4 YouTube

YouTube has become one of the most successful social media websites providing of short video sharing service since its establishment in 2005. Resources, which have exactly the same traits as the endpoints in Instagram, are the individual data entity with a unique identifier. Resources are represented in the form of JSON, and give an opportunity to integrate the functions that YouTube generally provides to user’s website or mobile applications.

YouTube currently offers total eleven resources including ‘activity’, ‘channel’, ‘playlist’, ‘video’ and etc. Among the resources, ‘video’ has been used in this study because it

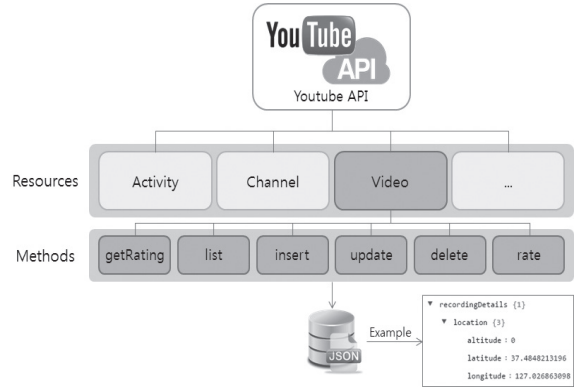


Fig. 4. Flow chart of collecting data via YouTube API

contains the main contents (video) of YouTube. YouTube API also provides a command called ‘method’ which makes users insert, update or delete their contents. The types of applicable methods are different in each resource, and Table 5. describes the characteristics of the video resource and its methods. The process to collect YouTube data using the video resource with its methods is presented in Fig. 4.

Among the various fields that can be collected using YouTube API, the selected fields to be analyzed are listed in Table 6. Among the fields, ‘recordingDetails’, ‘recordingDetails.locationDescription’, ‘recordingDetails.location’ and ‘fileDetails.recordingLocation’ are required to conduct spatial analysis, and the rest of the fields are needed for the analysis of the contents.

Table 3. The characteristics of the parameters offered by /media/search in Instagram

Parameters	Characteristics
LAT	Latitude of the center search coordinates. If used, LNG is required
LNG	Longitude of the center search coordinates. If used, LAT is required
MIN_TIMESTAMP	Indicates the minimum time to collect the data. If this value is empty, 5 days of the search period will be automatically set up
MAX_TIMESTAMP	Indicates the maximum time to collect the data. Cannot be set up over 7 days
DISTANCE	Indicates the distance from the center search coordinates. Default is 1km, and max distance is 5km

Table 4. The selected fields of Instagram dataset

Endpoint	Fields	Characteristics	Data Type	
/media/search	id	The ID of the data	string	
	tags	The tag that the data is included	array of string	
	location	The location information about the data	string	
	created_time	The created time of the data	integer	
	images	url	The link address about the picture information included in the data	string
	caption	text	The text information of the data	string

Table 5. The characteristics of the video resource and its methods in YouTube

Methods	Characteristics
getRating	Retrieves the ratings that the authorized user gave to a list of specified videos
list	Offers a list of videos that match the API request parameters
insert	Uploads a video to YouTube and optionally sets the video's metadata
update	Uploads a video's metadata
delete	Deletes a YouTube video
rate	Adds a like or dislike rating to a video or remove a rating from a video

Table 6. The selected fields of YouTube dataset

Resource	Fields	Characteristics	Data Type
video	snippet.publishedAt	The date and time that the video was published	datetime
	snippet.description	The description of the video	string
	contentDetails.duration	The length of the video	string
	statistics.viewCount	The number of times the video has been viewed	unsigned long
	statistics.likeCount	The number of users who have indicated that they liked the video	unsigned long
	statistics.dislikeCount	The number of users who have indicated that they disliked the video	unsigned long
	recordingDetails.locationDescription	The text description of the location where the video was recorded	string
	recordingDetails.recordingDate	The date and time when the video was recorded	datetime
	fileDetails.creationTime	The date and time when the uploaded video file was created	string
suggestions.tagSuggestions[].tag	The keyword tag suggested for the video	string	

3. Spatial Informatization

As presented in Fig. 1, the brief spatial informatization process in this study is divided into several phases. But the detail of the process consists of total nine phases described in Fig. 7.

3.1 Preparation

3.1.1 Data collection

The datasets from each SNS have been collected in Seoul where latitude is 37.541° and longitude is 126.986°. The collection period of the datasets had to be adjusted because the API regulations of each SNS have been considered to achieve the total number of each SNS data similar to each other. YouTube, first, has been collected from June 1st to 30th in 2015. Twitter has been collected from July 7th to 22nd in 2015. Instagram has been lastly collected from July 27th to Oct 2nd in the same year. Totally 12,938 data have been collected as described in Fig. 5.

Building layer dataset was created based on the road name-based address. It was first established in 2013 and this study used new version of the dataset developed in Oct, 2015 (Fig. 6).

3.1.2 Geographic coordinate system

To conduct the integration process, we tried to turn the different datasets into an unified coordinate system,

WGS84 (World Geodetic System 1984). In this study we chose WGS84 as a default coordinate system. Thus the SNS and building layers datasets have been converted into the coordinate system.



Fig. 6. The building layers dataset

3.1.3 Geo-coding

The SNS dataset has been geo-coded based on the field with X, Y coordinates in the dataset. We conducted geo-coding process with this geographic information using ‘XrGeocoder’⁹⁾ and ‘XrOldAddressToNew’ tools. The first one convert coordinates to parcel-based addresses. the another one performed to convert parcel-based addresses to road-name addresses. The fields for the both types of address has been created in the attribute table of the dataset.




Twitter	Instagram	YouTube
2015.07.20 ~ 2015.07.22	2015.07.27 ~ 2015.10.02	2015.06.01 ~ 2015.06.30
5,719	2,709	4,510
		

Fig. 5. The display of each SNS dataset

9) Open source address conversion tool developed by GeoService. ‘XrOldAddressToNew’ was also developed by the same company.

3.2 Spatial integration (Spatial join)

The fields selection as informed in section 2 has been conducted and the data collection has also been implemented to store the datasets as JSON format. The JSON data should be transformed into the CSV¹⁰⁾ form to be utilized through various spatial information tools such as ArcGIS or QGIS. After the transformation, ‘Table to dBASE (multiple)’ conversion tool offered by ArcGIS was used to change the format of the data from CSVs to DBF¹¹⁾s. ‘Merge’ tool in ArcGIS was also used to generate an unified dataset of the DBFs. Finally ‘Find Identical’ and ‘Delete Identical’ both helped to remove the overlapped values in the unified dataset (Fig. 7).

The process to extract coordinates information from the

data is conducted to convert into the point form dataset. it would be possible to perform the spatial integration process with the existing spatial datasets such as building layers. The spatial integration, then, leads to measure the number of spatialized data in a building. The unified SNS dataset was spatially joined with the existing building layers through ArcGIS. The attributes in the SNS dataset was combined with the attributes in building layers dataset as a resultant of this process.

4. Discussion

Spatial integration with the dataset collected from SNS

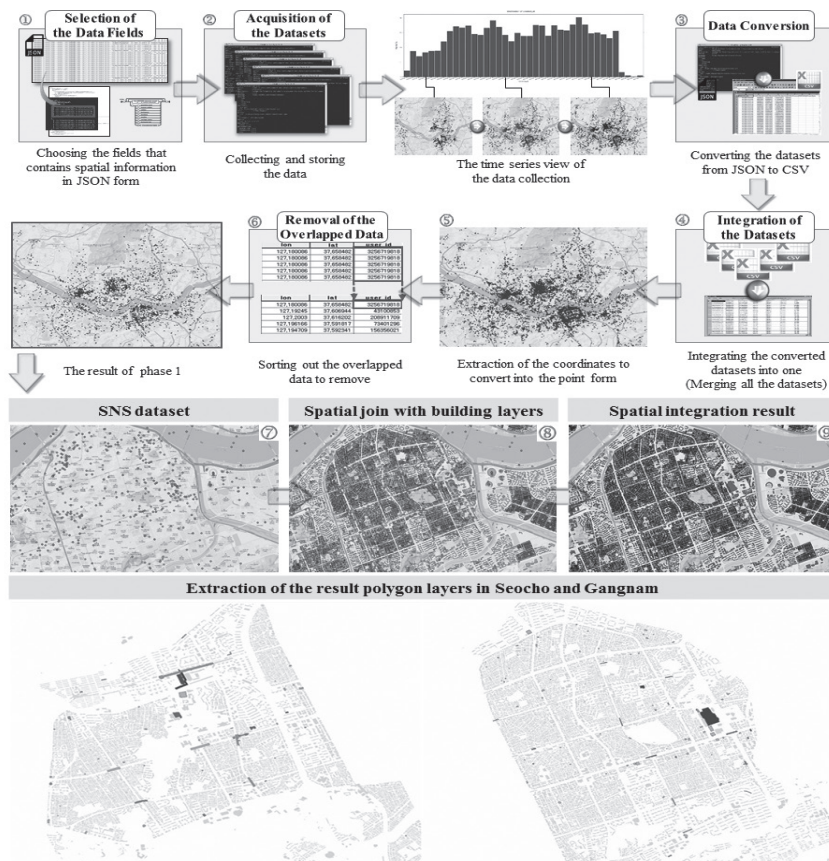


Fig. 7. Spatial informatization process

10) CSV (Comma Separated Value) file contains the values in a table as a series of ASCII text lines organized that each column value is separated by a comma from the next column's value and each row starts a new line.

11) DBF (DataBase File) is a file format typically used by database software.

has been made through this study. First the fields with spatial information in SNS dataset has been selected and then a request sent to Open API has also been made to collect the datasets. The conversion process to transform the data format has been conducted to extract the geographic information. The extracted spatial traits of the dataset has been converted into the point form and spatially integrated with existing building layers. The resultant of this integration was the building information combined with SNS information that represent the spatial experiences of the users at the specific location.

On the other hand, there are some downside conducting spatial integration using SNS dataset. The first one is the difficulty of getting useful information. There were a lot of overlapped data that were written by same person (user_id distinguished) or the people on purpose to propagate. Especially the latter has made some parts of spatial integration useless. It means that the location or place name that the existing spatial data (building layers) also had not been the same as the name extracted from the SNS dataset sometimes. In this case we had to rely on the coordinates given in the both datasets.

The missing addresses were also one of the problems to make difficulties on finding the appropriate location. The most of the SNS messages tend to explain the visiting experience to certain locations in the spatial perspectives through the text contents. In this case text mining will be helpful to extract information related to a certain location or place (Mostafa, 2013).

Understanding the characteristics of the location information extracted from the SNS data leaves much to be desired in many ways because of the various directions to use the datasets. The resultant of analysis will be also different depending on how the datasets have been processed. In this regard the spatial informatization process is worthy to be established and it will be the fundamental resource for the further studies related to spatial integration.

Acknowledgments

This research, "Geospatial Big Data Management, Analysis and Service Platform Technology Development,"

was supported by the MOLIT (The Ministry of Land, Infrastructure and Transport), Korea, under the national spatial information research program supervised by the KAIA (Korea Agency for Infrastructure Technology Advancement) (15NSIP-B081011-02).

References

- Antoine, É., Jatowt, A., Wakamiya, S., Kawai, Y., and Akiyama, T. (2015), Portraying Collective Spatial Attention in Twitter, *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 10-13 August, Sydney, Australia, pp. 39-48.
- Cheng T. and Wicks, T. (2014), Event Detection using Twitter: A Spatio-Temporal Approach, *PLoS ONE*, Vol. 9, Issue. 6, e97807.
- Cheng, Z., Caverlee, J., and Lee, K. (2010), You are where you tweet: a content-based approach to geo-locating twitter users, *In Proceedings of the 19th ACM international conference on Information and knowledge management*, 26-30 October, Toronto, Canada, pp. 759-768.
- Cruz, I.F. (2004), *Geospatial Data Integration*, Chicago: ADVIS Lab, Dept. of Computer Science, University of Illinois, pp.73-77.
- Dashti, S., Palen, L., Heris, M.P., Anderson, K.M., Anderson, S., and Anderson, S. (2014), Supporting disaster reconnaissance with social media data: a design-oriented case study of the 2013 colorado floods, *Proc. of ISCRAM*, 24 May, Pennsylvania, USA, pp. 632-641.
- Dhawan, S., Singh, K., and Sehrawat, D. (2014), Emotion Mining Techniques in Social Networking Sites, *International Journal of Information & Computation Technology*. Vol. 4, No. 12, pp. 1175-1153.
- Flowerdew, R. (1991), *Spatial data integration*, Geographical information systems, Colorado, pp. 375-387.
- Gomide, J., Veloso, A., Meira Jr, W., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. (2011), Dengue surveillance based on a computational model of spatio-temporal locality of Twitter, *In Proceedings of the 3rd International Web Science Conference*, 11-14 June, Koblenz, Germany, pp. 3-10.
- Hochman, N. and Schwartz, R. (2012), Visualizing instagram: Tracing cultural visual rhythms, *In Proceedings of the*

- Workshop on Social Media Visualization (SocMedVis) in conjunction with the Sixth International AAAI Conference on Weblogs and Social Media*, 8-11 May, Dublin, Ireland, pp. 6-9.
- Hughes, A.L. and Palen, L. (2009), Twitter adoption and use in mass convergence and emergency events, *International Journal of Emergency Management*, Vol. 6, Issue 3, pp. 248-260.
- Instagram developers. (2015), Instagram developers' documentation, *Instagram*, USA, <https://www.instagram.com/developer/endpoints/> (last date accessed: 26 November 2015).
- Karimkhani, C., Connett, J., Boyers, L., Quest, T., and Dellavalle, R. P. (2014), Dermatology on instagram, *Dermatology online journal*, Vol. 20, No. 7, doi 23129.
- Kim, Y., H. (2015), *KISDI STAT Report: Analysis of SNS usage in Korea*, Annual report, Korea Information Society Development Institute, Seoul.
- Kwak, H., Lee, C., Park, H., and Moon, S. (2010), What is Twitter, a social network or a news media?, *In Proceedings of the 19th international conference on World wide web*, 26-30 April, Raleigh, USA, pp. 591-600.
- Lwin, K.K., Zettsu, K., and Sugiura, K. (2015), Geovisualization and correlation analysis between geotagged Twitter and JMA rainfall data: Case of heavy rain disaster in Hiroshima, *In Spatial Data Mining and Geographical Knowledge Services*, 8-10 July, Fuzhou, China, pp. 71-76.
- Ma, Z., Sun, A., and Cong, G. (2013), On predicting the popularity of newly emerging hashtags in twitter, *Journal of the American Society for Information Science and Technology*, Vol. 64, No. 7, pp. 1399-1410.
- Mei, Q., Liu, C., Su, H., and Zhai, C. (2006), A probabilistic approach to spatiotemporal theme pattern mining on weblogs, *In Proceedings of the 15th international conference on World Wide Web*, 22-26 May, Edinburgh, UK, pp. 533-542.
- Mohammadi, H., Binns, A., Rajabifard, A., and Williamson, I.P. (2006), Spatial data integration, 17th UN Regional Cartographic Conference for Asia and the Pacific, 18-22 September, Bangkok, Thailand, pp. 1-11.
- Mostafa, M.M. (2013), More than words: Social networks' text mining for consumer brand sentiments, *Expert Systems with Applications*, Vol. 40, No. 10, pp. 4241-4251.
- Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011), An Empirical Study of Geographic User Activity Patterns in Foursquare, *ICWSM II*, 17-21 July, Barcelona, Spain, Vol. 11, pp. 70-73.
- Sakaki, T., Okazaki, M., and Matsuo, Y. (2010), Earthquake shakes Twitter users: real-time event detection by social sensors, *In Proceedings of the 19th international conference on World wide web*, 26-30 April, Raleigh, USA, pp. 851-860.
- Sui, D. and Goodchild, M. (2011), The convergence of GIS and social media: challenges for GIScience, *International Journal of Geographical Information Science*, Vol. 25, No. 11, pp. 1737-1748.
- Twitter developers. (2015), Twitter Platform and documentation, *Twitter*, USA, <https://dev.twitter.com/overview/api> (last date accessed: 26 November 2015).
- YouTube for Developers. (2015), API Resources and documentation, *YouTube*, USA, <https://www.youtube.com/yt/dev/api-resources.html> (last date accessed: 26 November 2015).