**DOI: 10.35595/2414-9179-2020-1-26-464-470**

**Mohamad Hasan**[1]

# USING SOCIAL MEDIA DATA TO MAP THE AREAS MOST AFFECTED BY ISIS IN SYRIA

## ABSTRACT

This paper presents a model to collect, save, geocode, and analyze social media data. The model is used to collect and process the social media data concerned with the ISIS terrorist group (the Islamic State in Iraq and Syria), and to map the areas in Syria most affected by ISIS accordingly to the social media data. Mapping process is assumed automated compilation of a density map for the geocoded tweets. Data mined from social media (e.g., Twitter and Facebook) is recognized as dynamic and easily accessible resources that can be used as a data source in spatial analysis and geographical information system. Social media data can be represented as a topic data and geocoding data basing on the text of the mined from social media and processed using Natural Language Processing (NLP) methods. NLP is a subdomain of artificial intelligence concerned with the programming computers to analyze natural human language and texts. NLP allows identifying words used as an initial data by developed geocoding algorithm. In this study, identifying the needed words using NLP was done using two corpora. First corpus contained the names of populated places in Syria. The second corpus was composed in result of statistical analysis of the number of tweets and picking the words that have a location meaning (i.e., schools, temples, etc.). After identifying the words, the algorithm used Google Maps geocoding API in order to obtain the coordinates for posts.

**KEYWORDS:** ISIS, GIS, data mining, geocoding, NLP

## INTRODUCTION

The introduction of Web 2.0 in the early 2000s allowed the creation and exchange of user-generated content. Social media can be understood as web applications which allow creating social networks (e.g., Facebook), blogs and microblogs (e.g., Twitter), and photo, audio, and video-sharing services (e.g., YouTube and Flickr) [*Kaplan, Haenlein,* 2010]. Now Social media applications have hundreds of millions of users and generate petabytes of data. Twitter, for example, has rapidly gained approximately by the end of 2018 about 326 active million users worldwide[2], while Facebook had 2.32 billion monthly active users[3]. In fact, the Library of Congress between 2010 and 2017 archived both Twitter feeds and Facebook posts until the library could not keep up with the increase of data size.

Since social media data users generate data continuously and dynamically, extensive studies with significant impacts have been conducted on using social media data on a wide range of subjects, such as predicting disaster responses [*Huang, Xiao*, 2015], infectious disease tracking [*Hay et al.*, 2013], earthquakes [*Sakaki, Okazaki*, 2010] measuring public opinion and political sentiment without explicit surveys [*Soulis et al*., 2013], and predicting stock market value [*Jin et al.*, 2017]. Moreover, social media data are used to explain many geographical phenomena, either by analyzing the geotagged social media data [*Hecht et al.*, 2011] or geocoding social media data to get posts' coordinates [*Alexander*, 2014]. Researchers found, that in times of emergency in either natural disasters or terrorist attacks number of posts increases, with such an increase in post numbers could help to study human actions during disasters [*Simon et al.*, 2015]. Therefore, social

[1] Saint Petersburg State University, Institute of Earth Sciences, Department of Cartography and Geoinformatics, 10th line of Vasilievsky island, 31−33, 199178, St. Petersburg, Russia; *e-mail*: **mo-hasan89@hotmail.com**

[2] Statsoft Inc. https:// www.statista.com/statistics/282087/number-of-monthly-active-twitter-users

[3] Statsoft Inc. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide

media data sources become valuable in time-critical cases, such as disaster response [*Goodchild, Glennon,* 2010] and disaster relief [*Barbier et al.,* 2011]. Social media data also are being used to mapping and analyzing crimes and crime predictions [*Ming-Hsiang,* 2015]. Moreover, combing social media data with other geospatial data types can be used to counter-terrorism applications [*Alharith, Samak,* 2018]. With such wide applications, social media data can be used to analyze ISIS's effects on the population in Syria, since more social media data in a certain area means a larger effect for ISIS on the population for that area.

## MATERIALS AND METHODS OF RESEARCHES

Since most social media services provide full access to the data generated on their sites on commercial basis, it may be too expensive for most researchers. Most social media services have an application programming interface (API) that allows some sort of limited access to the data on the sites for free. Twitter's free API allows collecting data only up to 7 days, while Facebook's graph API allows only access on public pages and only with their page id. A solution to this restriction is to use a third-party application, which allows to collect data requesting the data through browsers, but such applications return a subset of the full data requested by the twitter's API, only the data which are shown on the browser. A Python script was used to interact with social media API and store the data in the JSON format in PostgreSQL, an open-source SQL database with JSON processing capabilities [*Hasan et al.,* 2019]. In order to set a location to the social media data, Google Maps geocoding API is used. The input of Google Maps geocoding API is tokenized words which have a location meaning. Usually, the tokenized location word is collected using a special Python library which compares the text to a set of collected words that have geolocation meaning, but there is no such library for the Arabic language. In order to geocode the social media a library of Arabic location meaning words was created by analyzing 10000 tweets/posts basing on the frequency of each word and selecting which words have a location meaning. Finally, most GIS software is not designed to parse JSON files and display it, so the JSON data was parsed into PostgreSQL view, in order to use the data in GIS software. The components used are presented in fig. 1.

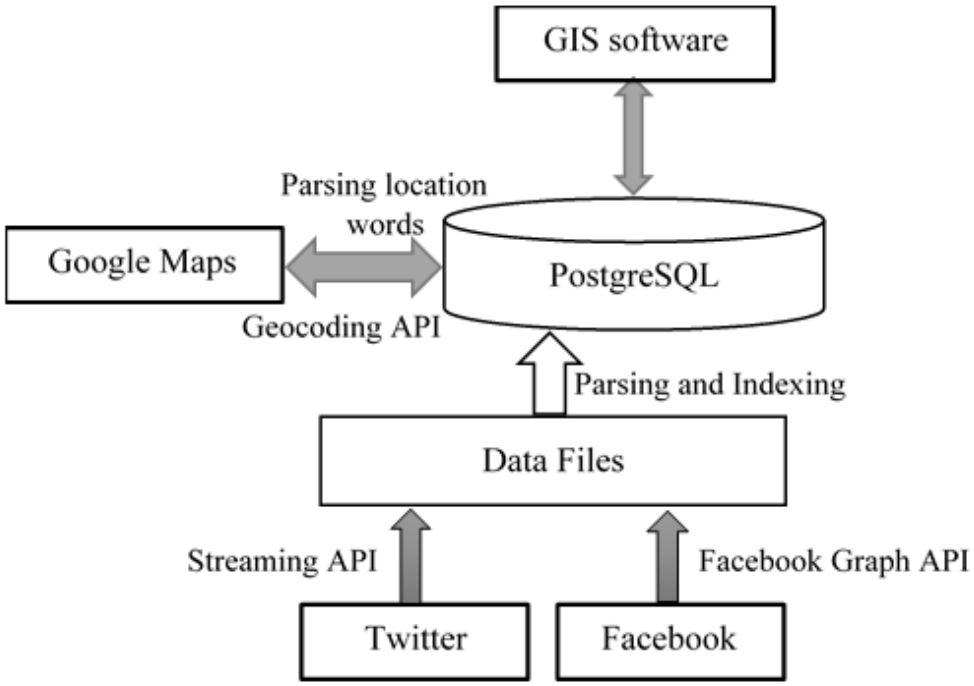


*Fig. 1. GIS architecture for collecting and using social media data*

The geocoding process forms a basic fundamental part of spatial analysis in a variety of research disciplines, especially in extracting spatial data from social media [*Goldberg et al.*, 2014]. In the model, the geocoding problem was solved using an algorithm similar to gazetteer algorithm [*Hill*, 2000]. Gazetteer is ideal to identify keywords from texts that have a geospatial meaning, especially for texts from social media. Since the gazetteer typically does not contain the functionality to generate the geocodes, Google Maps geocoder returns the coordinate.

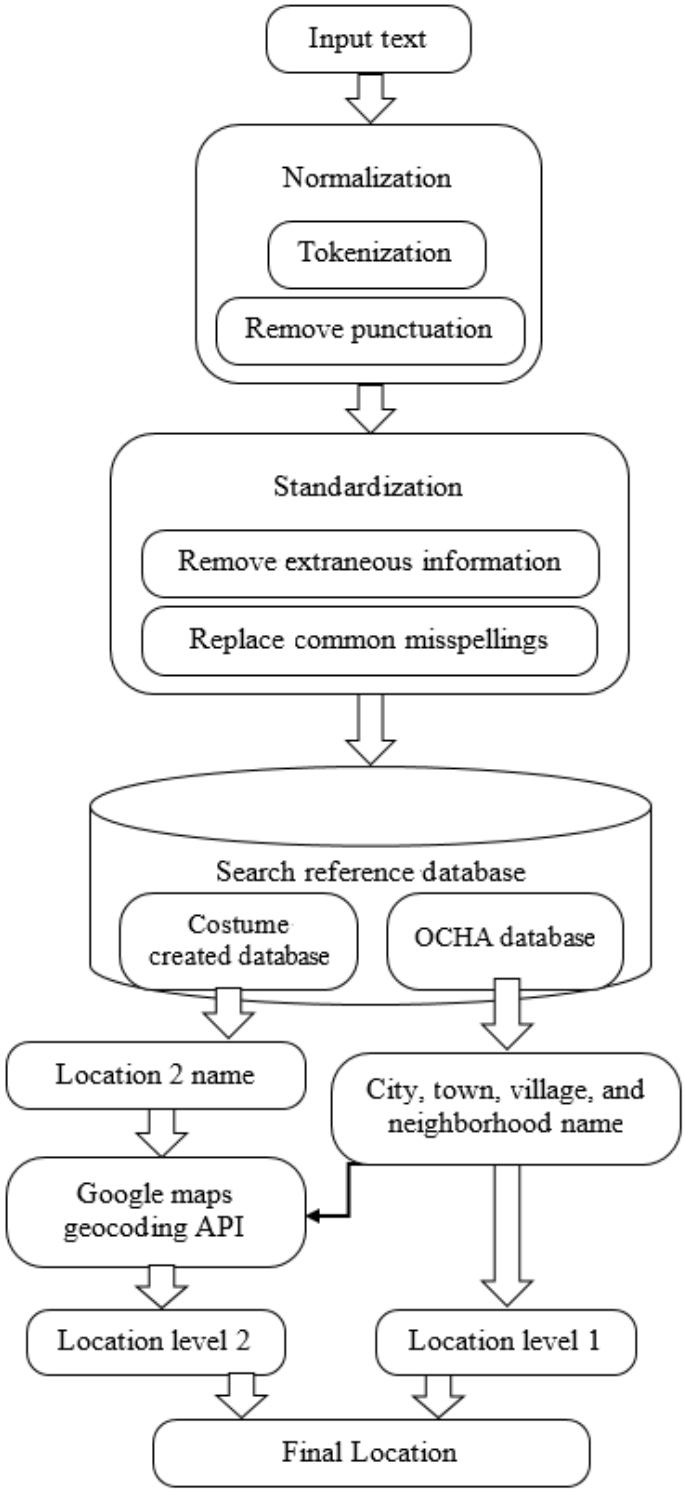


*Fig 2. Schematic of geocoding social media data*

Fig. 2 describes the processing algorithm that standardizes and normalizes the input text data. The key role of standardization and normalization is to determine which tokenized words in the input and to turn each into versions consistent with those in the reference dataset. Then the algorithm picks the words from a search reference database composed of two corpora. The first corpus derived from the OCHA database (Office for the Coordination of Humanitarian Affairs) for Syrian populated places, i.e. cities, towns and villages. The second corpus consists of the most frequent words that may indicate an address. The statistical analysis was done on 10 000 of the mined tweets. The content of the second created corpus is shown in fig. 3.

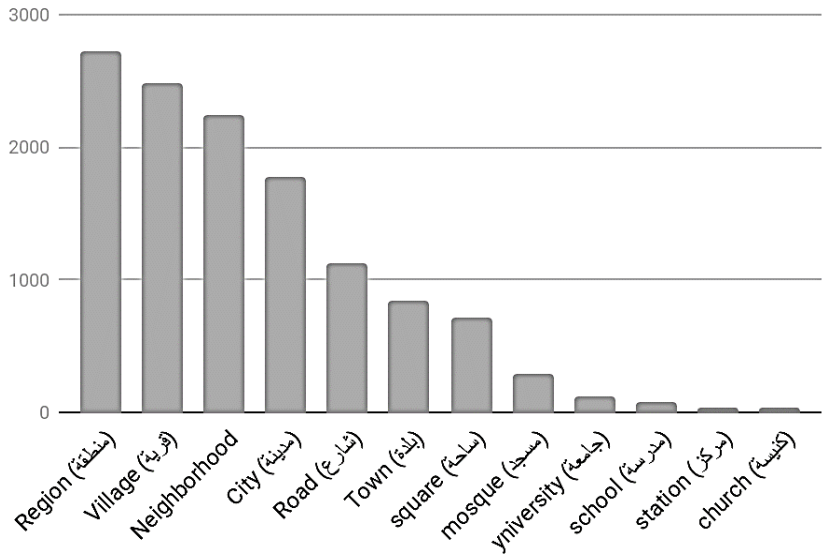


*Fig 3. Frequency of location words in 10 000 posts*

In this dataset, there are words in Arabic and Syrian dialect such as school, road, hospital, etc. If the algorithm allocates words from the second datasets it picks also two words after the found word to pick the name of the point. After the normalization and standardization to be searchable on Google Maps by geocoding API, the algorithm queries the text through the API to derive the final output coordinates.

Mining the data from social media about ISIS was implemented by two methods. Mining data from Twitter was done using multiple keywords referring ISIS in the Arabic language, such as “Daesh”, “The Islamic State of Iraq and the Levant”, and “Islamic State”. Mining data from Facebook graph API does not allow keywords use; therefore, mining data was done on about 80 Facebook pages that publish posts about local news and official news network pages, after mining all the data from the Facebook pages it was filtered to separate ISIS data. Both Streaming API and Facebook Graph API extracted data form social media on the whole territory of Syria. Approximately 14 mn tweets were extracted and 1 million Facebook posts gathered from 80 local news network pages. The social media data were generated over the span of 5 years since 2014 and amassed over 25 GB.

## RESULTS OF RESEARCHES AND THEIR DISCUSSION

All the data was processed by the created geocoding algorithm. The geocoding algorithm found coordinates for 27 % for Twitter data and 63 % of the Facebook data, this can be explained by mining technology. Since Facebook data was mined from news networks, it contained more information about the activities of ISIS, while Twitter data was collected from public users where it had much more opinion tweets. The results of the geocoding are shown in fig. 4.

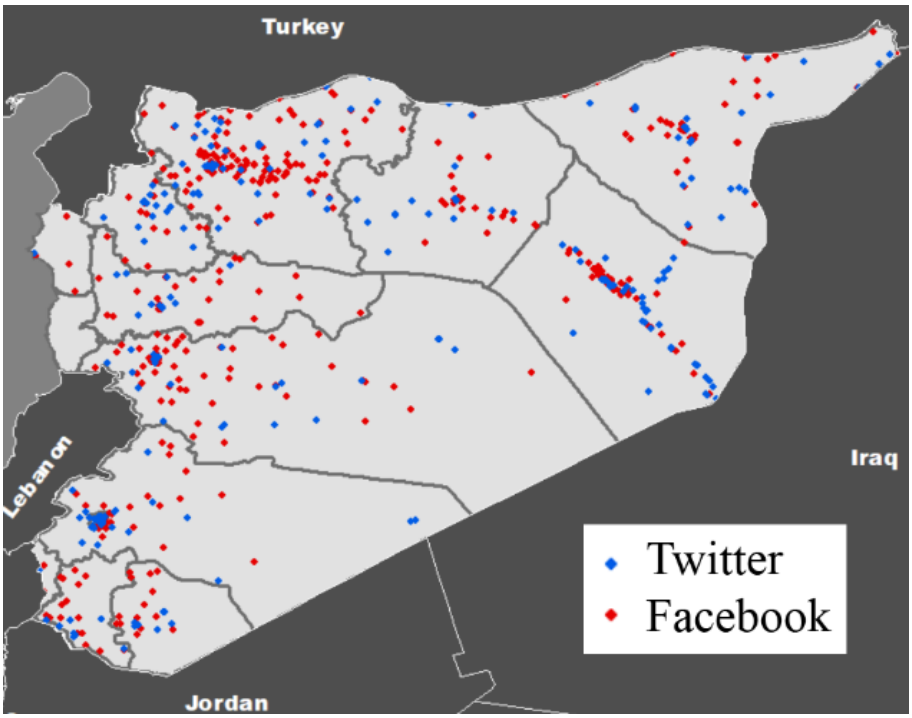


*Fig. 4. Social media posts about ISIS in Syria*

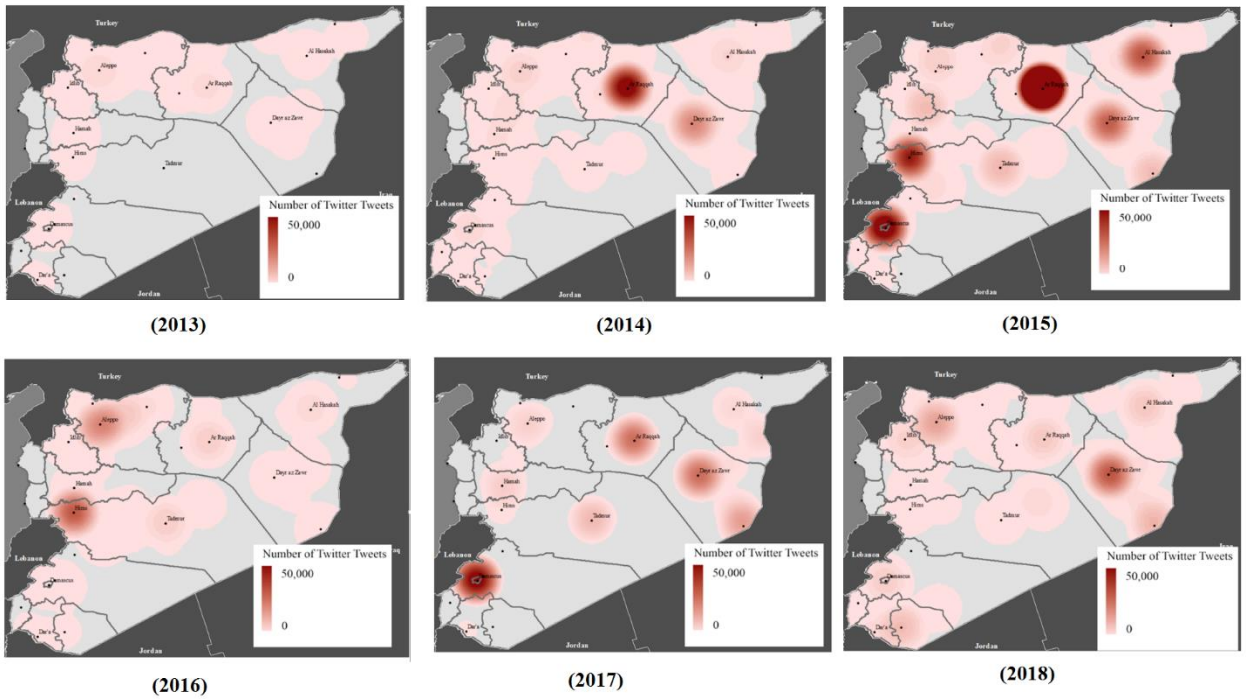


*Fig. 5. Density map of social media data about ISIS in Syria by year*

Fig. 5 presents the results of data yearly collected and processed since the emergence of ISIS in 2013 until its loss of its territories in 2018. The density maps show the residential areas in the northern and northeastern territories that are controlled by ISIS, the cells were highlighted with red color. The maps show that higher populated areas in the territories controlled by ISIS are the most affected.

Following the data retrieval and geocoding, mapping of the words associated with ISIS is done. The word mapping (fig. 6) requires development of a corpus with words associated with

ISIS activities. The corpus was created by analyzing the tokenized mined texts of social media by choosing the most frequent tokens and picking the words associated with ISIS and terrorist activities and removing the stop words. Stop words are the most commonly used words that do not add information to the text. The social media data were clustered by the smallest administrative division or Nahiyah. The size of the word is associated with the frequency of the word in social media data in each cluster.

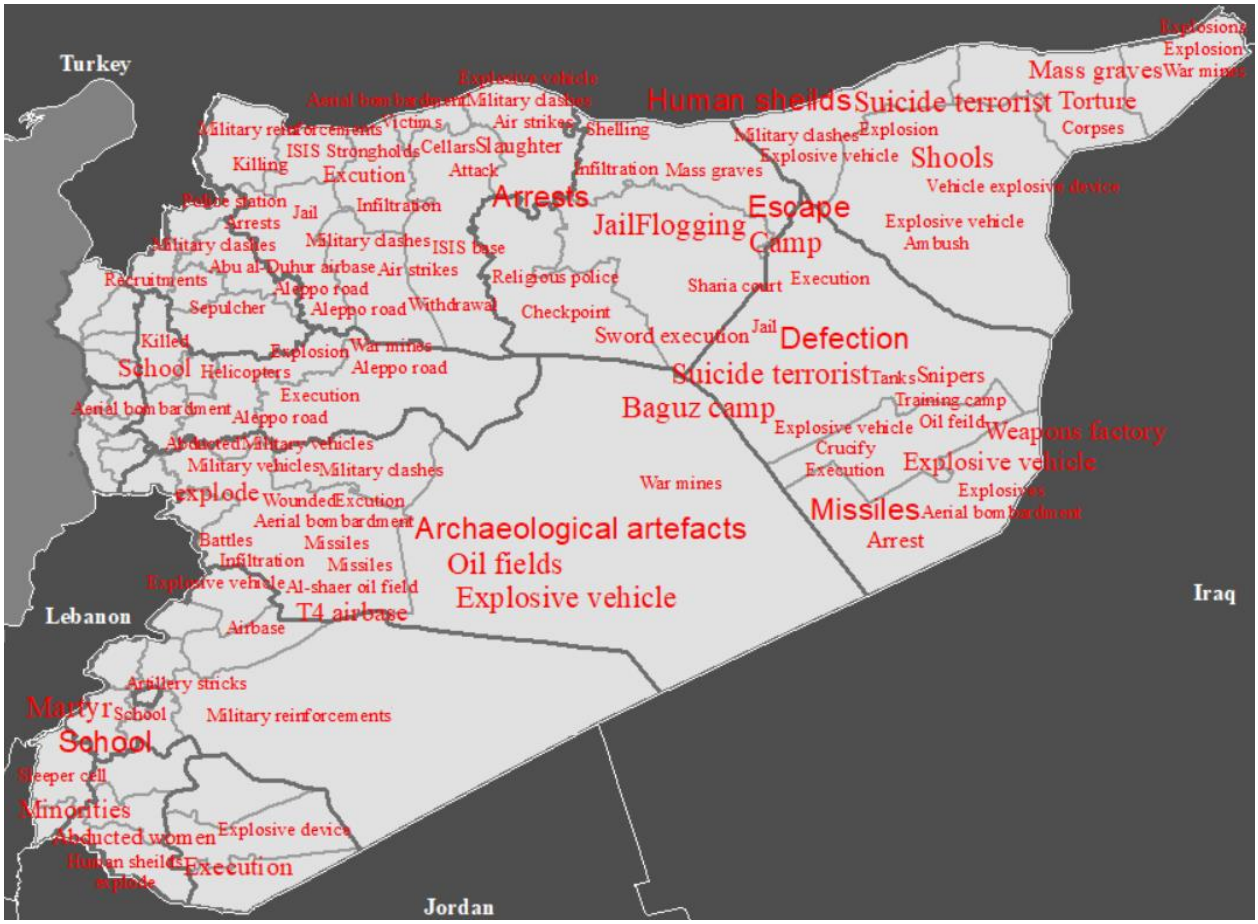


*Fig. 6. Words associated with ISIS based on location*

## CONCLUSIONS

Using of social media data in GIS provides a better understanding of complex spatial dynamics of the social and environmental systems by enabling individual-level dynamic observations from social media users. Through using the model developed, social media data collection service based on Twitter and Facebook can provide GIS with a highly interactive and user-friendly process of location-based social media data harvesting.

## REFERENCES

1. *Alexander D.E.* Social media in disaster risk reduction and crisis management. Science and Engineering, 2014. V. 20. No 3. P. 717−733.
2. *Alharith A., Samak Y.* Fighting terrorism more effectively with the aid of GIS: Kingdom of Saudi Arabia case study. American Journal of Geographic Information System, 2018. V. 7 No 1. P. 15−31.
3. *Barbier G., Goolsby R., Gao H.* Harnessing the crowdsourcing power of social media for disaster relief. IEEE Intelligent Systems, 2011. V. 26. No 3. P. 10−14.

4. *Goldberg D.W., Wilson J.P., Knoblock C.A.* From text to geographic coordinates: the current state of geocoding. Journal of Spatial Information Science, 2014. V. 9. P. 37−70.
5. *Golubovic N., Krintz C., Wolski R., Lafia S., Hervey T., Kuhn W.* Extracting spatial information from social media in support of agricultural management decisions. GIR'16. Proceedings of the 10th Workshop on Geographic Information Retrieval, October 2016. 2017. Article No 4. P. 1–2.
6. *Goodchild M., Glennon A.* Crowdsourcing geographic information for disaster response: A research frontier. International Journal of Digital Earth, 2010. V. 3. No 3. P. 231−241.
7. *Hasan M., Panidi E., Badenko V.* Comparative evaluation of NoSQL and relational databases performance while analyzing semi-structured geospatial data. 5th International Scientific Conference GEOBALCANICA 2019. 2019. P. 541−549. DOI:10.18509/GBP.2019.64.
8. *Hay S.I., George D.B., Moyes C.L., Brownstein J.S.* Big data opportunities for global infectious disease surveillance. PLoS Medicine. 2013. V. 10, No 4. Article No e1001413.
9. *Hecht B., Hong L., Suh B.* Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. 29th Annual CHI Conference on Human Factors in Computing Systems, 2011. P. 237−246.
10. *Hill L.* Core elements of digital gazetteers: placenames, categories, and footprints. Research and Advanced Technology for Digital Libraries, 2000. V. 1923. P. 280−290.
11. *Huang Q., Xiao Y.* Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. ISPRS International Journal of Geo-Information, 2015. V. 4. No 3. P. 1549−1568.
12. *Jin F., Wang W., Chakraborty P., Self N., Chen F., Ramakrishnan N.* Tracking multiple social media for stock market event prediction. Advances in Data Mining. Applications and Theoretical Aspects, 2017. V. 10357. P. 16−30.
13. *Kaplan A.M., Haenlein M.* Users of the world, unite! The challenges and opportunities of social media. Business Horizons, 2010. V. 53. No 1. P. 59−68. DOI: 10.1016/j.bushor.2009.09.003.
14. *Ming-Hsiang T.* Research challenges and opportunities in mapping social media and big data. Cartography and Geographic Information Science, 2015. V. 42. No 1. P. 70−74.
15. *Sakaki T., Okazaki M.* Earthquake shakes twitter users: Real-time event detection by social sensors. Proceedings of the Nineteenth International Conference on WWW, 2010. P. 851−860.
16. *Simon T., Goldberg A., Adini B.* Socializing in emergencies, a review of the use of social media in emergency situations. International Journal of Information Management, 2015. V. 35. No 1. P. 609−619.
17. *Soulis K., Varlamis I., Giannakoulopoulos A., Charatsev F.* A tool for the visualization of public opinion. International Journal of Electronic Governance, 2013. V. 6. No 3. P. 218−231.
18. *Wang S., Hu H., Lin T., Liu Y., Padmanabhan A., Soltani K.* CyberGIS for data-intensive knowledge discovery. SIGSPATIAL Special, 2014. V. 6. No 2. P. 26−33.