Analysis of the Value of Public Geotagged Data from Twitter from the Perspective of Providing Situational Awareness

Aragats Amirkhanyan, Christoph Meinel

Analysis of the Value of Public Geotagged Data from Twitter from the Perspective of Providing Situational Awareness

Aragats Amirkhanyan and Christoph Meinel

Hasso Plattner Institute (HPI), University of Potsdam
{Aragats.Amirkhanyan,Christoph.Meinel}@hpi.de
https://hpi.de

Abstract. In the era of social networks, we have a huge amount of social geotagged data that reflect the real world. These data can be used to provide or to enhance situational and public safety awareness. It can be reached by the way of analysis and visualization of geotagged data that can help to better understand the situation around and to detect local geo-spatial threats. One of the challenges in the way of reaching this goal is providing valuable statistics and advanced methods for filtering data. Therefore, in the scope of this paper, we collect sufficient amount of public social geotagged data from Twitter, build different valuable statistics and analyze them. Also, we try to find valuable parameters and propose the useful filters based on these parameters that can filter data from invaluable data and, by this way, support analysis of geotagged data from the perspective of providing situational awareness.

Keywords: analysis, statistics, big data, location-based social networks, geotagged data, georeferenced data, situational awareness, public safety awareness

1 Introduction

Nowadays, social networks are an essential part of modern life for millions of users around the word. Users use social networks to communicate with friends and share what happens with them, what they feel and so on. One of the biggest social networks is Twitter that, according to the recent statistics [3], has more than 300 million active monthly users, who post every day more than 600 million tweets (messages) [6]. Twitter is not only the social network, but it is the source of real-time news feeds and it is the place where breaking news appear firstly, before we read it in the newspapers, listen at the radio or watch on TV [12]. The amount of data produced by social networks increases dramatically every year. And, also, there is a trend, that users post geotagged messages that reflect the situation around them. For research, it is important, because it gives us more possibilities for visualization and analysis of social data, since we can be interested not only in the content of messages but also in the location, from where these messages were posted.

Analysis of public social geotagged data is a big topic and there are many papers and enterprise solutions, which cover different aspects and challenges of social data analysis. One of the most popular cases is to use geotagged data for analyzing and detecting natural disasters. For example, Sakaki et al. [17] used Twitter users as social sensors to detect earthquake shakes. They investigated the real-time interaction of events, such as earthquakes, in Twitter and proposed an algorithm to monitor tweets and to detect an earthquake target event. Another use case of using public social data is to analyze natural disasters was presented by De Longueville et al. [10]. In their paper, they showed how location-based social networks (LBSN) can be used as a reliable source of spatio-temporal information, by analyzing the temporal, spatial and social dynamics of Twitter activity during a major forest fire event in the South of France in July 2009. Later in 2013, the amount of public social geotagged data increased and researchers had a possibility to detect not only global but, also, local geo-spatial events [18].

In 2013, Kalev Leetaru et al. [13] presented their deep study of geography of Twitter. They analyzed data from Twitter posted around the world and built many statistics, such as total tweets per day, average tweets per hour, all exact location coordinates and the top 20 cities by percent of georeferenced tweets. Also, their study includes linguistic, textual and user profiles analysis of data. In 2014, Muhammad Adnan et al. [7] presented their results of analysis geotagged data from Twitter. In their work, their concentrated on social dynamics of Twitter usage based on data from London, Paris and New York City. They showed the areas of tweeting activity, they provided the results of ethnicity analysis of Twitter users, geography of tweets of different ethnic groups and gender analysis. Another analysis of geotagged data, Diansheng Guo et al. [11] presented in 2014. In their work, they were aimed to detect non-personal and spam users on geotagged Twitter network. Their approach contains extracting user characteristics, constructing training datasets, conducting supervised classification for detecting non-personal users and the evaluation of the approach. In 2015, Umashanthi Pavalanathan et al. [16] presented interesting results. They showed that young people and women more often write geotagged tweets; users, who geotag their tweets, tend to write more, making them easier to geolocate; and text-based geo location is significantly more accurate for men and for older people.

Based on papers of previous years including mentioned in this section, we see how geotagged data are valuable and what we can obtain from analysis of them. Therefore, in this paper, we provide our research results of analysis of geotagged data in the scope of our specific challenge, which is analysis of the value of geotagged data from the perspective of providing situational awareness. For that, we start with Section 2, in which we provide the description of our research project, in the scope of which we address the current challenge, and our motivation for that. The remainder of the paper is organized as follows: in Section 3, we describe how we collect data and what data we have for analysis. Afterwards, in Section 4, we provide statistics and our analytic results of collected data. Based on that, we try to propose methods for making data more

valuable for analysis of situational awareness and public safety awareness. We conclude the paper and provide future directions of research in Section 5.

2 Project Scope and Motivation

In the scope of our main research project, we are aimed to analyze and visualize in real-time publicly available social geotagged data to provide situational and public safety awareness. This challenge requires a complex solution, therefore, during work on the project, we face with many additional challenges that come from the practice [8] [9], such as real-time clustering and visualization of massive geotagged data, provision of advanced methods for searching and filtering, provision of real-time valuable statistics and so on. Nowadays, there are a lot of social analytics tools, such as TweetDeck, Twitonomy, Hootsuite, Tweepsmap, Geofeedia and so on. Many of them provide powerful functionality for analysis and visualization of data for different purposes and solve some challenges mentioned above. But mostly, they are aimed to support marketing and business, therefore, their solutions of filtering and providing real-time statistics are not always can be applied to our research project that has other focus.

We have our specific goal, which is provision of situational and public safety awareness. To achieve this goal, we want to work with social geotagged data that describe the situation around the place, from which they were posted. But it is not always the case, and we have to work with a huge amount of data that are useless for our analysis and they do not reflect valuable information and do not describe the situation around. Therefore, we face with the challenge that presented in this paper. This challenge is analysis of the value of public geotagged data from Twitter from the perspective of providing situational awareness. In the scope of this challenge, we want to understand what and which parameters can help us to recognize whether some concrete message does have any value in describing the situation around or not, based on which characteristics we can fully exclude invaluable data from our dataset and which statistics of data can help us to better understand the situation around. All together should support our research project. Therefore, we are aimed to collect social geotagged data, build statistics and analyze them. Results of these statistics can help us to exclude invaluable data and develop advanced filters to support analysis of data from the perspective of providing situational awareness.

3 Data

Data is an essential part of any research. For our research, we use public social geotagged data from Twitter. Twitter provides a quite powerful API to fetch required data [1] [5]. One of them is commercial Twitter's Firehose, which guarantees access to 100% of tweets, and other one is public Streaming API, which does not guarantee the percentage of receiving data in real-time. More comparisons between data from Twitter's Streaming API and Twitter's Firehose, you can find in the paper of Fred Morstatter et al. [15]. But we want to mention that,

according to some studies [1], using Twitter's Streaming API users can expect to receive anywhere from 1% of the tweets to over 40% of tweets in near real-time. Also, you can increase the percentage of receiving tweets by applying more strict criteria. The criteria can be keywords, usernames, locations, named places, etc. In our case, we use location as a criteria and we collect data from London, because London is one of the world's most active Twitter cities [4]. To collect data, we use the Java-based application that connects to Twitter's Streaming API, specifies needed criteria and starts to receive tweets in real-time. For our experiments we decided to collect data from the area that covers London and its neighborhood. The coordinates of monitored area are {51.247948, -0.569042; 51.727184, 0.303813}, where the first pair is latitude and longitude coordinates of the south west and the second pair is latitude and longitude coordinates of the north east. We parse all received data, normalize them and save them into the database for further analysis. For research and further analysis, we collected 1 million tweets that cover about 12 days: from the 28th of January 11:37:30 AM until the 8th of February 12:41:48 PM.

4 Analysis

This section is the main part of the paper. In this section, we provide different valuable statistics and full description and analysis of them. Also, we try to find out how our results could be used to support analysis of situational and public safety awareness. It means that based on our statistic and statistics results, we try to propose (1) methods for reducing the amount of data by excluding (removing) invaluable data and (2) filters that can support analysis of geotagged data. Both proposals (excluding data and filters) are aimed to support analysis of situational awareness based on social geotagged data.

Table 1: Tweets' languages statistics

English	Undefined	Spanish	Arabic	Portuguese	French	Others
83.80%	6.76%	1.66%	1.37%	1.16%	0.92%	4.33%

We start with the statistics in Table 1. In this table, you can see the most popular languages of tweets posted from London. Information about languages we take from the tweet object, which is provided by Twitter [5]. Firstly, we can see that the official language is in the first place. About 83.80% of all tweets are posted in English. Then we can see that for 6.76% of tweets, Twitter was not able to identify the language. In most cases, it means that tweets without the language do not contain sentences or even words. They are just a set of hashtags, user mentions, symbols and URLs. So, such tweets do not have semantic sense. In the next places, we have Spanish (1.66%), Arabic (1.37%), Portuguese (1.16%), French (0.92%) and so on. Now, we need to consider how we can use obtained

information to support analysis of data. Firstly, we definitely can exclude all tweets with the undefined language, because they bring no sense and we can not use them to analyze situational awareness. By this way, we can reduce the amount of data by 6.76%. Other thing that we can do, we can use the knowledge about used languages to filter dataset by languages or filter dataset from tweets in other foreign languages. It can support analysis of data, because we would fetch only relevant data.

Table 2: Tweets' place types statistics

City	Admin	Country
79.08%	18.50%	2.42%

Usually tweets contain the place information, because users are asked to attached a geographic place to the tweet before to publish it. Additional, Twitter asks users to attach exact coordinates. So, the final tweet could have the exact geo coordinates and attached place. Places could be *city*, administrative area (*admin*), *country* or some concrete place of interest (*poi*), for example, Big Ben in London. In our dataset, almost all tweets have attached places (only 5 tweets do not have a place), but only 11.53% of data have exact coordinates. In Table 2, you can see which place types are usually attached to tweets. In the first place we have the *city* with about 79.08%, then the *admin* (administrative area) with about 18.50% and in the last place we have the *country* place type with about 2.42%. In our research, we are interested only in the tweets that have the *city* place type or more narrow as a place of interest (*poi*). Place types, such as *country* and administrative area (*admin*), are too big areas and we can not use such data for analyzing situational awareness. It means that we can easily exclude such data from our dataset, but with one assumption. We exclude tweets with irrelevant place types only if these tweets do not have exact geo coordinates. Because if the tweet contains exact geo coordinates, we should not consider to which place the tweet is attached. So, if we exclude tweets with *country* and *admin* place types, we can reduce dataset by maximum 20.92%.

Table 3: Statistics of external sources of tweets

twitter.com	instagram.com	bit.ly	swarmapp.com	goo.gl	trendinalia.com	youtube.com
8.46%	6.54%	1.51%	0.98%	0.47%	0.44%	0.44%

Not all content published in Twitter is original. Some tweets come from other social networks or have links to the external web sites or services, such

as Facebook, Instagram, Swarm (Foursquare) and son on. If the tweet contains several links, we consider only the last link, because usually the last link refers to the original source of information. According to Table 3, the main external source of tweets is twitter.com (8.46%). It means that users post tweets that contain links to other tweets. Then we have instagram.com (6.54%), bit.ly (1.51%), swarmapp.com (0.98%) and so on. We can suppose that if user posts a tweet about what happens around him, he does not include links to external websites into the tweet, otherwise, user likely does not really describe the situation around him. But this statement does not work if the tweet has link to another social network, because, in this case, it could be that user posted about the situation around in Instagram but then he reposted the message to Twitter. Therefore, it is not always obviously, which data we can exclude from dataset. It requires more detailed analysis. But some of them we can definitely exclude. For example, we can exclude tweets that refer to Foursquare, Swarm or Yelp, to social networks that have nothing with describing the situation around. Usually, tweets, which refer to Swarm, have information, such as "Hi, I am in London" or something similar to it. So, if we exclude tweets at least only from Swarm, we can reduce dataset by about 0.98%.

Table 4: Statistics of the most demanding geo coordinates

gcpe6rh4k4j9	gcpvjc9kxvpg	gcpuuqtyeztv	u120jz6zfbbe	gcpusn4djt0k	gcpvnqvtp20z
5.79%	5.75%	4.93%	4.33%	4.15%	3.34%

We have 1 million tweets but only 11.53% of tweets have exact geo coordinates, others have geo coordinates based on an attached geographic place (the center of the place). It means that if tweets have the same attached places then they have the same coordinates. Based on our dataset, we calculated that from 1 million possible geo coordinates, we have only 37650 unique geo coordinates, which is about 3.76% of all possible coordinates. This percentage is less than the percentage of tweets with exact coordinates. So, we can conclude that tweets with exact geo coordinates can have the same coordinates. Usual case for that is when tweets are reposted from other social networks. Additionally, we analyzed which geo coordinates are the most demanding. In Table 4, you can find the calculated statistics. The coordinates are presented in geohash[1] form to simplify the represent. They can be easily converted back to latitude and longitude by the geohash function. According to our statistics, about 5.79% of tweets have coordinates *gcpe6rh4k4j9* (51.23513843, -0.59857568), which is in the area of the city Guildford. The next, about 5.75% of tweets have coordinates *gcpvjc9kxvpg* (51.51294588, -0.09681718), which is the center of London. By this statistics, we can see which places are the most attractive by Twitter users. Above we mentioned that about 79.08% of tweets contain attached places of the *city* type.

[1] `https://en.wikipedia.org/wiki/Geohash`

But these places are different. If user posts a tweet in London and he attaches the place as London then his tweet would have coordinates of the center of London. But if he attaches some concrete district of the city then the tweet would have more concrete location. For example, such district of London can be Barnet, Hackney, Lambeth and so on. Actually, Twitter suggests the closest and the most appropriate place when user wants to attach the place to the tweet, and it helps more correctly determine the location for the tweet even if user did not attach the exact geo coordinates. Now, the question is what we can do with this statistics to support analysis of situational awareness. We can use this statistics to design a filter. This filter would use statistics from Table 4 and provide filtering data from tweets, which were posted from the most demanding geo coordinates. In some cases, it can facilitate visual analysis of situation.

In Section 3, we mentioned that we subscribe for data from London and its neighborhood defined by the following coordinates {51.247948, -0.569042; 51.727184, 0.303813}. But Twitter usually returns not only data from the specified area but also from areas that overlap that area. For example, we receive, also, tweets with coordinates of the entire country UK. Therefore, additionally, we calculated the percentage of tweets that have coordinates inside of the specified area. The result is 82.51%, which is close to the percentage of tweets containing a *city* place type. If we exclude tweets outside the monitored area, we can reduce the amount of data by 17.49%. But this percentage will be less, if we, firstly, exclude tweets with inappropriate place types *admin* and *country*.

In Introduction, we mentioned that some tweets contain location information, such as hashtags or words of locations, and they can be geolocated. Therefore, we want to find out how many tweets in our dataset contain additional location information. In some cases, it can help to identify the location more precisely than it is specified. For example, we could have the tweet "The house is on fire in Carnaby Street" and this tweet could have an attached place as the entire London, which has coordinates of the city center. But in this case, this message would be more valuable if it would have more concrete location. We can see this concrete location in the text message "in Carnaby Street". Therefore, we want to find out how many tweets in our dataset contain such additional location information. To find out it, we used Stanford Named Entity Recognizer (NER) [2]. And after applying this library for our dataset, we obtained that about 9.51% of tweets contain additional location information in their text content. It means that potentially 9.51% of tweets could be more accurate geolocated than it is specified.

Table 5: The most active Twitter users

tegrenade	hesjkr94	trendinaliagb	orgetorix	blankiiam015	don_jide	a_rockas	biggucci_idz
1.40%	0.61%	0.50%	0.26%	0.23%	0.22%	0.22%	0.20%

The next thing, in which we are interested in, is how many unique users we have. We calculated that in our dataset we have 107105 unique users, and it means that we have about 9.37 posts per user. Not all users are equally active. Therefore, we want to find out which users produce more tweets than others. In Table 5, you can find partial statistic results. In the first place, we have user *tegrenade*, who produced about 1.40% of all tweets from our dataset. It is about 14004 tweets during 12 days, which means 1167 tweets per day. The Twitter accounts in the next palaces produced also huge amount of data, you can see it in Table 5. Some of these active users we can consider as a non-personal or spam users. It is an interesting challenge to identify it and Diansheng Guo et al. [11] presented their approach of detecting non-personal and spam users. But in the scope of our paper, we would like to use found statistics to design a filter. And this filter would be aimed to filter data from the most active users by using statistics from Table 5. In some case, it can facilitate analysis of data.

Now, we consider hashtags, user mentions, symbols, URLs and media objects in tweets, and how they are mentioned by users. We start with symbols. The symbol is the character started with *$* dollar sign and it is often used in the financial area, for example, to put price information or to include the name of the company in the stock market. Examples of symbols are AMZN, GOOG, FB, GBPUSD, where AMZN - Amazon, GOOG - Google, FB - Facebook, and GBPUSD - British Pound to Dollar. So, we can suppose that tweets with symbols have nothing with describing the situation around, therefore, we can exclude them. But they constitute just about 0.01%, so the benefit is small.

Table 6: The percentage of tweets that contain URLs in their text content

has URL	1 URL	2 URLs	3 URLs	4 or more URLs
25.70%	23.96%	1.71%	0.03%	0.003%

Many tweets contain URLs. According to our statistics, about 25.70% of tweets contain URL. Mostly, tweets contain only one URL (23.96%), but some of them have more links. Details statistics you can find in Table 6. Such URLs could be links to other social networks or to some external websites with some news. We can suppose that if user posts the tweet about what he sees now, likely, he will not include a link to the external resources, but with the exception when user reposts the tweet from other social networks. Therefore, we can assume only one or maximum two links. Others tweets we can exclude from dataset.

Some tweets contain media objects: image or video. According to our statistics, 14.24% of tweets contain media object, and 14.238% of them contain only one media object and 0.002% of them contain 2 media objects. Media-based tweets are produced by 35.81% of users. So, it is common to include one media object to the tweet, but existence of media objects in tweets does not tell us

about the relevance of the tweets from the perspective of describing the situation around. Therefore, we should not consider this parameter for filtering data.

Table 7: The percentage of tweets with user mentions in their text content

has mention	1 mention	2 mentions	3 mentions	4 or more mentions
51.68%	36.35%	9.58%	3.12%	2.63%

Users often mention other users in their tweets to start or to keep discussion. For that, they use @ character and the account name of the user. From Table 7, we can see that about 51.68% of tweets contain user mentions. We could suppose that the number of user mentions can affect how the tweet describes the situation around. More user mentions in the tweet then more likely that this tweet is just a part of the discussion, but not describing the situation around. Therefore, we are interested in a filter that can filter data by number of user mentions. For example, we could want to filter data from tweets that contain 3 or more user mentions. It can reduce dataset by 5.74% for further analysis.

Table 8: The percentage of tweets with hashtags in their text content

has hashtag	1 hashtag	2 hashtags	3 hashtags	4 or more hashtags
22.71%	12.04%	5.04%	2.23%	3.40%

In Table 8, you can see the statistics of hashtags in tweets. According to our results, about 22.71% of tweets contain hashtags. About 12.04% of them contain only one hashtag, about 5.04% of them contain two hashtags, about 2.23% of them contain 3 hashtags and about 3.40% of them contain 4 or more hashtags. We can suppose that when user wants quickly to post a tweet about the situation around, he would not consider about including a huge amount of hashtags into the tweet. But we can not fully rely on this statement. We should consider another additional parameter.

Table 9: The count of hashtags and the count of days when these hashtags appear

12 (100%)	11 (91.66%)	10 (83.33%)	9 (75.00%)	8 (66.66%)	7 (58.33%)	6 (50.00%)
799	580	548	606	720	988	1286

In Table 9, we provide the statistics of the most utilized hashtags during the time periods of collected data. 799 hashtags appear every day (12 days) in our

dataset, 580 hashtags appear in 11 days (91.66% of the entire date range), 548 appear in 10 days (83.33%) and so on. Some hashtags, which appear every day, could be useless hahstags from our perspective, for example, the hashtag *#happybirthday*. Whereas among these hashtags, there are hashtags that represent geographic places, for example, *#london*. The hashtag *#london* appears every day, but we assume that this hashtag can be included into the tweet to describe the situation around. Therefore, we need to consider an additional parameter that can help us to understand the value of tweets with popular hashtags. Therefore, we built the statistics of the distribution of the most popular hashtags among users. You can find this statistics in Table 10. From that statistics, we can see that, for example, the hashtag *#london* is used by 5.07% of users. This hashtag is used by significant amount of users and it appears every day, therefore, we can not exclude tweets with this hashtag. Whereas, the hashtag *#happybirthday*, which appears also every day, is used only by 76 users (0.071% of users). This fact could give us a guess that such hashtag does not bring any situational information into the tweet. Therefore, we can think about filtering them.

Table 10: Statistics of usage concrete hashtags by the percentage of users

london	love	cbb	superbowl	uk	sb50	art	fridayfeeling	valentinesday	tbt
5.07%	0.71%	0.70%	0.60%	0.53%	0.52%	0.50%	0.49%	0.44%	0.43%

We have three parameters related to the hashtags: the number of hashtags in the tweet, the distribution of hashtags during the date period of dataset and the distribution of hashtags among users. We want to filter tweets that do not describe the situation around. For that, we need to use all three parameters. Therefore, our filter should assume that an invalid tweet should have more than n hashtags, at least m of these hashtags should appear every day during the monitored period d (in our case it is 12 days, but it can be customized), and less than p percents of users should use these m hashtags. We need to point out that we do not consider which values of parameters n, m, p, d to choose to obtain the best filtering results, but we have only showed the parameters that should be considered. Finding the concrete optimal values is a part of future work.

5 Conclusion and Future Work

We devoted this paper for analysis of the value of social geotagged data from the perspective of providing situational awareness. We were motivated to do it because we have to work with a huge amount of invaluable data and we wanted to reduce this amount of invaluable data, design filters and build statistics that can support analysis. During work on analysis of data, we built statistics by different parameters and different compound parameters, and we presented the most valuable found results from our perspective.

It is always important to point out that we do analysis of social geotagged data from the perspective of providing situational awareness and, all statistics which we built, we analyzed for our concert use case that makes our research different from many existing social analytics tools focused mostly on business and marketing. For example, in the statistics of the most popular languages in Table 1, we tried to find out how this statistics can help us to remove partially invaluable data. And we found out that the indicator of invaluable data, from the perspective of providing situational information, can be the undefined language. We went further and we tried to heuristically find out how number of different entities in the text content can affect the situational value of data. We had an assumption that if the tweet has many URLs, user mentions, hashtags and symbols then such tweet has less the situational value. Based on this assumption, we built statistics by mentioned parameters that showed us what the benefit we can obtain if we exclude much littered tweets. Another example of analysis of geotagged data from perspective of providing situational awareness is the statistics of original sources of tweets. We found out that some sources of tweets can be indicators of invaluable data, but it requires manual analysis and making the list of irrelevant sources, such as Swarm and Foursquare, which do not bring situational information in their content.

To evaluate the benefit from analysis of data and built statistics, we applied some recommendations for our research project. If we remove data with inappropriate place types (*country*, *admin*), data outside the specified monitored area, data from irrelevant services: Swarm and Foursquare, data with symbols and data with undefined languages, then we exclude 28.86% of invaluable data. Meanwhile, we can expect the higher percentage of removed invaluable data, if we apply more advanced methods for filtering based on compound parameters from built statistics. Therefore, we would like to continue research and, for that, we determine some future work directions.

As a future work, we would like to build and visualize in real-time presented in the paper statistics. Also, we would like to use statistics from this paper to develop proposed filters. These filters can help to better analyze data and provide more accurate situational awareness. Such filters could be filters of the spam hashtags, inappropriate URLs, inappropriate geo location coordinates, tweets that contain too much user mentions and so on. Also, we need to find the optimal values of the parameters for filters from the perspective of providing the optimal filtering to have mostly data that describe the situation around. With results in this paper and future work mentioned in this section, we plan to go further to achieve the main goal of the research project - real-time situational and public safety awareness based on public social geotagged data.

References

1. Bright planet. `http://www.brightplanet.com/2013/06`, last visited on 08.03.2016
2. Stanford Named Entity Recognizer (NER). `http://nlp.stanford.edu/software/CRF-NER.shtml`, last visited on 08.03.2016

3. The number of monthly active Twitter users worldwide. http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/, last visited on 08.03.2016
4. The world's most active Twitter cities. http://www.forbes.com/sites/victorlipman/2012/12/30/the-worlds-most-active-twitter-city-you-wont-guess-it/, last visited on 08.03.2016
5. Twitter api documentation. https://dev.twitter.com/overview/api, last visited on 08.03.2016
6. Twitter statistics. http://www.internetlivestats.com/one-second/#tweets-band, last visited on 08.03.2016
7. Adnan, M., Longley, P.A., Khan, S.M.: Social dynamics of twitter usage in london, paris, and new york city. First Monday 19(5) (2014), http://firstmonday.org/ojs/index.php/fm/article/view/4820
8. Amirkhanyan, A., Cheng, F., Meinel, C.: Real-time clustering of massive geodata for online maps to improve visual analysis. In: Innovations in Information Technology (IIT), 2015 11th International Conference on. pp. 308–313 (Nov 2015)
9. Amirkhanyan, A., Meinel, C.: Visualization and analysis of public social geodata to provide situational awareness. In: 2016 Eighth International Conference on Advanced Computational Intelligence (ICACI). pp. 68–73 (Feb 2016)
10. De Longueville, B., Smith, R.S., Luraschi, G.: Omg, from here, i can see the flames!": A use case of mining location based social networks to acquire spatio-temporal data on forest fires. In: Proceedings of the 2009 International Workshop on Location Based Social Networks. pp. 73–80. LBSN '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1629890.1629907
11. Guo, D., Chen, C.: Detecting non-personal and spam users on geo-tagged twitter network. T. GIS 18(3), 370–384 (2014), http://dx.doi.org/10.1111/tgis.12101
12. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web. pp. 591–600. WWW '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1772690.1772751
13. Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., Shook, E.: Mapping the global twitter heartbeat: The geography of twitter. First Monday 18(5) (2013), http://firstmonday.org/ojs/index.php/fm/article/view/4366
14. Mao, H., Shuai, X., Kapadia, A.: Loose tweets: An analysis of privacy leaks on twitter. In: Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society. pp. 1–12. WPES '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2046556.2046558
15. Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the sample good enough? comparing data from twitter's streaming API with twitter's firehose. CoRR abs/1306.5204 (2013), http://arxiv.org/abs/1306.5204
16. Pavalanathan, U., Eisenstein, J.: Confounds and consequences in geotagged twitter data. CoRR abs/1506.02275 (2015), http://arxiv.org/abs/1506.02275
17. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: Real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web. pp. 851–860. WWW '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1772690.1772777
18. Walther, M., Kaisser, M.: Geo-spatial event detection in the twitter stream. In: Proceedings of the 35th European Conference on Advances in Information Retrieval. pp. 356–367. ECIR'13, Springer-Verlag, Berlin, Heidelberg (2013), http://dx.doi.org/10.1007/978-3-642-36973-5_30