

Geocoding of traffic-related events from Twitter

Juan C. Salazar C, Miguel Torres-Ruiz, Clodoveu A Davis Jr., Marco Moreno-Ibarra

Instituto Politécnico Nacional, Centro de Investigación en Computación
UPALM-Zacatenco, 07738 – Mexico City – Mexico

b130126@sagitario.cic.ipn.mx, mtorres@cic.ipn.mx,
clodoveu@dcc.ufmg.br, marcomoreno@cic.ipn.mx

***Abstract.** Nowadays social networks provide information with high correlation with events that are occurring in the worldwide. Twitter is a microblogging network of real time posts in which people know different classes of events such as concerts, festivals, demonstrations, etc. Other relevant topic is traffic congestions; user-generated content is useful to assist drivers in avoiding crowded areas. Therefore, this work is oriented towards following specific steps focused on improving the geocoding of traffic-related events that are associated with a number of geographic elements. Preliminary results have increased the precision and recall of locating geographic elements, achieving 85% and 83%, from a baseline of 36% and 30% respectively.*

1. Introduction

Detection of road traffic congestions is a significant problem to be solved for large cities. Urban mobility can be improved if bad conditions and accidents that occur during the day are known. Thus, the possibility to avoid congested crossroads, street segments with heavy traffic, demonstrations, lane interruptions and so on, is highly desirable.

However, it is difficult to gather traffic-related information of all places in cities throughout the day, and to carry out this task with tracking devices could probably take a long time. Therefore, to find other sources of information and know what happens in the city, social networks like Twitter are very useful. Twitter has become quite popular in different countries. The ease of sharing content through this social network encourages users to post a great amount of information regarding events that are happening in the real world [Lee et al. 2013]. Information related with traffic conditions is also very common on Twitter. When people are moving around the city, users post information about traffic-related events using their devices. As a matter of fact, there are several accounts that post exclusively information about traffic-related events. Some of these accounts are from government

agencies, official profiles of public transportation, accounts managed by radio stations and independent users. On the other hand, information in Twitter rarely includes the coordinates of where the accident happened. Usually, only few tweets contain latitude and longitude, and sometimes the location at which people tweet is not the same place where the event occurred. Tweets without coordinates are frequently ambiguous, with abbreviations, nicknames and misspellings. Furthermore, tweets are limited to 140 characters. For that reason, geocoding tweets and the relationships between them are very important challenges in particular research areas.

In this paper, we propose an approach to geocode traffic-related events from Twitter, improving the accuracy of representations by means of the number of geographic elements involved. We present a case study based on data from Mexico City. Preliminary results show a precision and recall rates of 85% and 83% respectively.

The remainder of the paper follows. Section 2 presents related work. The methodology to identify and locate geographic elements is outlined in Section 3. A description of the tweet dataset is presented in Section 4. Section 5 describes experimental results and a discussion related to the analysis. Finally, conclusions and future work are described in Section 6.

2. Related work

There are various methodologies for retrieving, processing and displaying geographic information from the web. The following approaches identify different geographic components from text. The Traffic Observatory [Ribeiro et al. 2012] proposes geocoding tweets using a gazetteer called GEODICT. This dictionary contains a collection of thoroughfare segments, street crossings, abbreviations, nicknames, neighborhoods and landmarks, along with their geographic representations. The Traffic Observatory uses exact and approximate string matching functions on gazetteer data to geolocate the streets mentioned in Twitter's stream.

Delboni et al. [2007] proposed a method to retrieve information from the web, by using natural language processing techniques, thereby recognizing positioning expressions formed by landmarks and spatial relations. Davis Jr. et al. (2011) proposed a methodology based on user relationships to infer the location of messages in Twitter. A network is created taking into account the follower-following relationships. Starting from known locations of users in the network, it infers the location of others. Working with Facebook data, Backstrom et al. (2010) showed that there is a strong connection between social relationships and geography. People that interact daily almost always live near each other, and each user has at least 10 friends with shared locations. With these assumptions, the methodology infers the most probable user location.

3. Geocoding traffic-related events from tweets

The proposed methodology involves automatic and manual steps in order to search geographic elements in Twitter (streets, neighborhoods, public transportation stations, places, and others) using a gazetteer. The initial gazetteer was obtained from GeoNames, the dataset is composed of 36,236 different streets (more than 150,000 street segments) from the Mexico City. They are in upper case, and they contain accent marks (Spanish language). They have blank rows or names with default values (*e.g.* geometries without name or with ‘NO NAME’ assigned) and many of them have abbreviations, such ‘as’, ‘ave’, ‘st’, ‘rd’, etc. The geocoding process consists of the following steps: (1) collect information from the tweet dataset, (2) create dictionaries and equivalent road axis names, (3) divide the gazetteer, (4) standardize, and (5) identify and locate traffic-related events. Each of these steps will be covered next.

3.1. Information from the tweet dataset

From the tweet dataset, it is important to go deeper on what people are talking about. Many tweets talk about traffic-related events, since that is the purpose of the selected accounts, but we found mentions to popular streets, common nicknames, common abbreviations, popular places and popular historical monuments. According to these findings, a script for determining the most common words in the tweet dataset was developed. The script finds the most common N-grams. A N-gram is a N-word slice of a longer statement [Cavnar et al. 1994]. From each tweet the n-grams are obtained, and the number of occurrences orders them. The most repetitive n-grams are selected (the threshold established was more than 100 mentions). Even though a N-gram includes the notion of any combination of characters or words in a sequence, in this case the script only considers contiguous slices of N-words. As an illustration, Figure 1 shows how the script works to get N-grams:

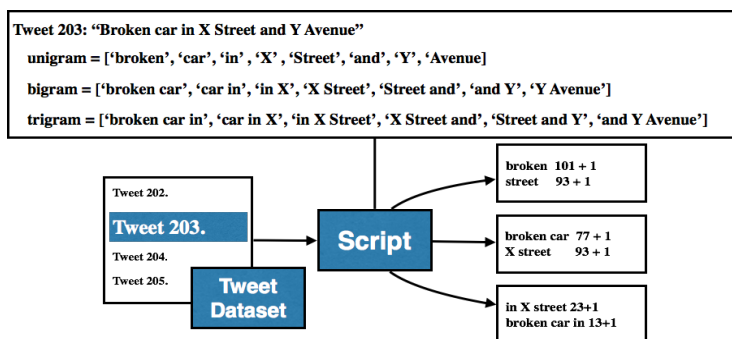


Figure 1. Creation of the lists of most frequent n-grams by the script.

From the most frequent unigram, bigram and trigram lists, we have identified by hand 456 common streets, 150 common traffic-related events, 135 common hashtags, 69 common nicknames, 65 common buildings, places and monuments, 34 common abbreviations and 26 common combinations of prepositions. Figure 2 represents all the collected information from the tweet dataset.

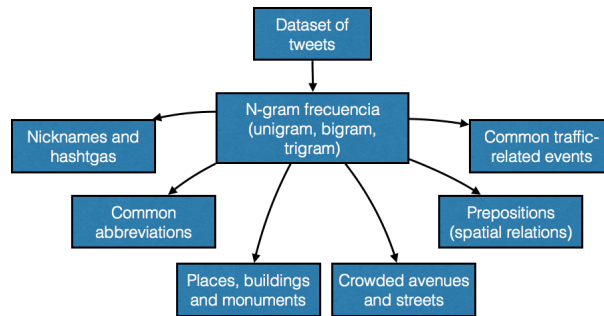


Figure 2. The collected information from the n-gram analysis.

This information was divided from the results of the frequency n-gram script. Each group of selected elements was saved in a CSV file, in order to facilitate further processing.

3.2. Dictionaries of geographic elements and equivalent road axis names

From the results obtained in the previous step, data from Open Street Map and from the National Institute of Statistics and Geography (INEGI for its acronym in Spanish) were used to generate some dictionaries meant to enrich the gazetteer (see Figure 3). The proposed dictionaries are the following: dictionary of abbreviations, dictionary of nicknames, dictionary of hashtags, dictionary of traffic-related events, dictionary of public transportation, dictionary of principal streets (only streets that appear in tweets and exist in the gazetteer), dictionary of places, buildings and monuments, and dictionary of neighborhoods. The last four dictionaries have a geographical component, which is used to spatially map the geographic features. We call them *dictionaries of geographic elements*. A dictionary of spatial relations could have been created, but we are classifying traffic-related events by using the number of geographic elements identified in tweets.

About the equivalent road axis names, it is frequent that streets can be named for more than one official name. Mexico City has 31 road axes and 2 circuits that cover more than 10 thousand kilometers of length. Axes and circuits change their names along their way when crossed with other streets. For that reason, sometimes people call them by the principal name, by its name in a certain segment (second name) or name them together (principal name + second name). However, all these options are valid, thus all alternatives must be

searched in the tweets. With the purpose of solving this issue, a dictionary of equivalent road axis names is added to the named collection.

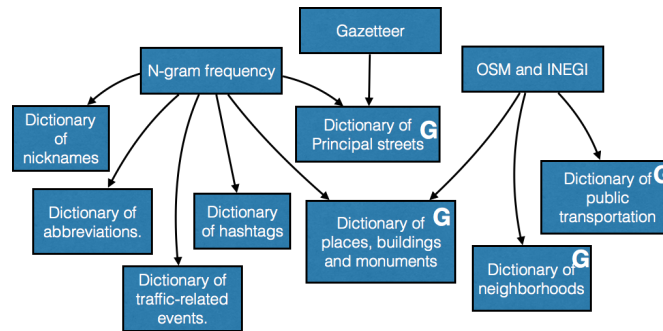


Figure 3. Creation of geographic and non-geographic dictionaries.

3.3. Division of the gazetteer

There is frequently a small group of streets that concentrate traffic-related events. Based on the N-gram frequencies, we found that only 19% of the streets in the complete gazetteer appear in tweets. In consequence, the gazetteer is split in two parts, the first part formed by the frequently named streets in tweets that exist in the gazetteer (some streets posted in tweets are outside of Mexico City, and are not considered) and the second part is composed of the remaining streets. Although the reduction does not improve the precision and recall of geocoding, the performance of the identification and location steps have increased.

3.4. Standardization

In order to improve the standardization process, dictionaries of non-geographic elements (dictionary of abbreviations, dictionary of nicknames and dictionary of hashtags) are used. In our gazetteer, the street names contained in geographic dictionaries include abbreviations, names in uppercase, names with accent marks, and even blank rows or default values. Hence, this process changes each street name to lowercase and removes accent marks (e.g. TALISMÁN ST. - talisman st.). Moreover, using the dictionary of common abbreviations, they are replaced with the complete word (talisman st. - talisman street). Finally, blank spaces and streets with default values are deleted. Other problems detected in tweets are links and mentions to other accounts, nicknames, misspellings and hashtags (e.g. <http://t.co/hAN0K0WS>, @OVIACDMX, ‘The angel’, ‘circuito interior street’, #insurgentesavenue). So, in order to solve these new issues, the dictionaries of nicknames and hashtags are used to replace them with the official name in tweets (e.g. ‘The angel’ - ‘angel of independence’ and #insurgentesavenue - ‘insurgentes avenue’). Links and mentions to other accounts are deleted and misspellings are not solved yet, and left for future work (using fuzzy matching or a frequent misspellings dictionary).

In addition, in both tweets and gazetteer, stop words have to be filtered out. Stop words are words that do not add meaning to the statement and they are the most common words in a language (e.g. articles, pronouns and prepositions). There is not a universal list of stop words used in natural language processing. For this case, the list of stop words has been defined by the Natural Language Toolkit Library [Bird, 2006].

3.5. Identification and location of traffic-related events

The identification of geographic elements in tweets has been carried out using all the previously described dictionaries of geographic elements: frequently named streets, uncommonly named streets (only if a considerable number of geographic elements was not found), neighborhoods, public transportation, places, buildings and monuments.

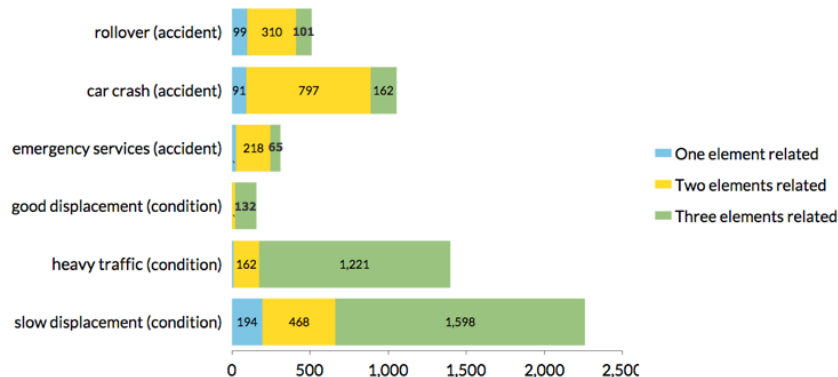


Figure 4. This bar chart shows examples of common traffic-related events that appear on tweets and the number of geographic elements detected.

Traffic-related tweets from the selected accounts frequently talk about accidents, bad or good traffic conditions. For example, traffic-related events such as accidents are described as ‘**car crash**’, ‘**rollover**’, ‘**emergency services**’, bad conditions are ‘**slow displacement**’, ‘**road closure**’, ‘**settlements**’, etc., and good conditions are ‘**still moving**’, ‘**good displacement**’ and so on. Although there is a considerable amount of advertising, questions and safety recommendations, they are easily filtered out because of the absence of any geographical element and due the short length of tweets. Based on the tweet dataset of and the N-gram frequency analysis, an accident is considered an event that happens in a certain (point) space with one or two geographic elements involved. Examples include, ‘**a car crash at x street and z street**’, ‘**broken traffic signal at the intersection of x street and y street**’, ‘**rollover in front of x subway station**’, etc. The description of a bad or good condition is considered as the actual situation of a street segment. In such descriptions, commonly one, two or three geographic elements are included. For example, ‘**settlements on x street between y street and z street**’, ‘**good displacement on x street from y street**

to z neighborhood’, ‘heavy traffic in x street on z neighborhood’, ‘raining over x neighborhood’, and so on. A fragment of the lists of accidents and conditions identified in the approach are presented in Table 1. Since tweets can only contain 140 characters, it is difficult to post a mention, a link, a traffic-related event and more than three geographic elements. Therefore, we identified that the number of geographic elements included in the tweet has a strong relation with the kind of traffic-related event (see Figure 4).

Table 1. Common accidents and conditions mentioned on Twitter.

Accident	Frequency	Condition	Frequency
emergency service	378	blocked road	4377
rollover	612	still close	1053
accident	1162	heavy traffic	1423
flooding	432	slow displacement	2779
car crash	1312	road work	1225
emergency in place	508	road close	2521
broken car	1002	traffic jam	1423
Vehicular congestion	570	bumper to bumper	2246
traffic signals out of service	241	gridlock	1101

Each dictionary of geographic elements has a primitive geographic representation (point, line, polygon). Thus, the dictionary of public transportation, which is depicted by a set of points, the dictionary of streets is characterized by a set of lines and the dictionaries of neighborhoods, places, buildings and monuments are represented by polygons. Therefore, as a result of searching geographic elements from dictionaries in tweets, we obtained a collection of geographic primitive elements (see Figure 5).

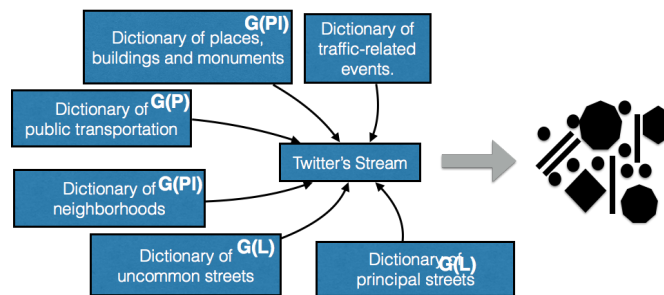


Figure 5. Result of identification process on Twitter’s stream.

Assuming that there can be 1, 2 or 3 references to places in a tweet, the number of possible relations that can happen among them follows the formula for combinations with replacement (see Equation 1).

$$CR_m^n = \binom{m+n-1}{n} = \frac{(m+n-1)!}{n!(m-1)!}, \quad \text{Equation 1.}$$

Where **m** is the number of possible elements to select, in this case point, line or polygon, and **n** is the number of elements found. Therefore, for 1 element we have: [(point), (line), (polygon)], 2 elements identified: [(point, point), (point, line), (point, polygon), (line, line), (line, polygon), (polygon, polygon)], and for 3 elements identified: [(point, point, point), (point, point, line), (point, point, polygon), (point, line, line), (point, line, polygon), (point, polygon, polygon), (line, line, line), (line, line, polygon), (line, polygon, polygon), (polygon, polygon, polygon)]. Many of these relationships of the geographic primitives are not accurate with respect to the location or the relation frequency in tweets is not high, thus they can be discarded. For this preliminary research work, the following relationships were considered: [(point, line), (line, line), (point, point, line), (point, line, line), (line, line, line), (line, line, polygon)]. The relations listed above were clearly identified in the tweet dataset, the assumptions of each relationship is described as follows:

- (point) represents an accident in a public transportation station.
- (point, line) represents a condition of a street segment in front of a public transportation station.
- (line, line) represents an accident in a street intersection.
- (point, point, line) represents a condition of a street segment delimited by two public transportation stations.
- (point, line, line) represents a condition of a street segment delimited by another street and a public transportation station.
- (line, line, line) represents a condition of a street segment delimited by two streets.
- (line, line, polygon) represents a street segment delimited by a street and a place, building or historical monument.

Although it is quite probable that there are other situations involving these groups of geographic elements, the situations described above occur more frequently. In order to obtain the result of these assumptions, three spatial operations have been devised:

1. Find the street intersection.
2. Find the closest point of the geographic element (polygon or line).
3. Find the bounding box (or convex hull) of the line segment.

The spatial operations were executed using PostGIS functions such as ST_Intersection, ST_ClosestPoint, ST_Envelope and ST_ConvexHull. The result is a geometric element, where the traffic-related event took place. For example, the process to find the (line, line, line) relation is the following:

- While all the possible combinations of elements are not tested:
 - Is there intersection of A element and B element: (operation 1)
 - Save it.
- Are there two intersections?
 - Yes: Is there an element in common from the two intersections?
 - Yes: Find the bounding box (or convex hull) of the element in common delimited by the two intersections. (operation 3)
 - No: Check (line, line) relation.
 - No: Check (line, line) relation.

4. Description of the tweet dataset

Our tweet dataset contains 64,250 tweets collected over a period of six months, from July 07, 2014 until December 24, 2014, without considering retweets and posts with blank spaces. Tweets are collected from reliable Twitter profiles that correspond to known services and institutions. Such accounts have been selected considering some features: account location, account creation date, number of followers, average number of tweets posted per day, if the account belongs to a government agency and if the account has its own website (see Table 2).

Table 2. Traffic-related Twitter accounts covering Mexico City

Twitter Account	Location	Creation date	Followers	Number of tweets	Belongs to government	Website
SSPDFVIAL	Mexico City	07.14.2010	369,115	154.65	Yes	sup.df.gob.mx
PolloVial	Mexico City	01.31.2013	667	71.91	No	No website
Trafico889	Mexico City	05.14.2009	137,099	90.54	No	siempre889.com/trafico
Alertux	Mexico City	10.16.2012	179,574	35.59	No	www.alertux.com
072AvialCDMX	Mexico City	10.20.2010	83,535	134.71	Yes	www.agu.df.gob.mx
RedVial	Mexico City	03.09.2010	63,702	44.81	No	rvial.mx

Most tweets that come from @SSPDFVIAL and @072AvialCDMX profiles belong to government agencies. The number of followers of these profiles keeps growing, and they have a specific behavior to explain traffic-related events. Hence, it is easier to geolocate their tweets. @Trafico889 and @RedVial belong to radio stations. They post information about weather and traffic conditions, and show their information in their websites. @Alertux and @PolloVial gather volunteered information and retweet information from other Twitter accounts.

5. Experiments and results

In order to measure the accuracy of this methodology, a test dataset was put together using 652 tweets geocoded by hand. We identified streets, public transportation stations, neighborhoods, places, buildings and monuments. The test dataset was compared with elements identified by our methodology, and we computed precision and recall. The methodology consists of the standardization process (this process includes the non-geographic dictionaries), the equivalent axis names and the dictionaries of geographic elements. At first, we compared with a baseline using only the gazetteer with part of the standardization process (only lowercase). Then, we compared with the baseline plus the full standardization process, and finally with the baseline plus the standardization process plus the equivalent axis names. When the system identifies all the elements of the solution, a hit is considered. When the system identifies at least one element of the solution, a partial hit is taken into consideration. Mistakes were also counted when the system did not find any element of the solution. Thus, precision and recall use true positives, true negatives and false negatives and were computed by applying Equation 2 for both cases respectively.

$$P = \frac{T_p}{T_p + F_p}; R = \frac{T_p}{T_p + F_N} \quad \text{Equation 2}$$

True positives are the geographic elements that were found by the methodology and belong to the gold standard, true negatives are geographic elements found by the methodology that do not belong to the gold standard and false negatives are geographic elements that belong to the gold standard that were not found. The precision and recall were computed for each tweet and we obtained the average of each test. Results are shown in Table 3.

In Figure 6, the behavior in Twitter and its relation with traffic-related events in the real world is shown. The number of tweets posted at 18, 19 and 20 hours is higher. This is the time of the day with the highest level of participation. Another relevant period is in the morning, around 8AM, with another peak in participation. This behavior corresponds to the rush hours in the city.

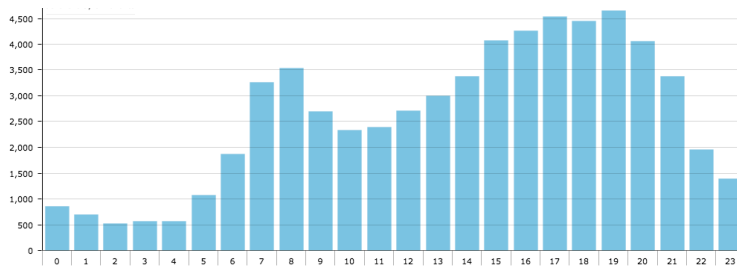


Figure 6. Relation between the behavior in Twitter and the traffic-related events.

Table 3. Results obtained by the methodology.

	Baseline	Standardization	Standardization + equivalent axis names	Standardization + equivalent axis names + dictionaries	Test Dataset
All elements found	152	152	427	456	652
At least one element found	289	388	599	608	652
Mistakes	363	264	53	44	0
Precision	0.39	0.43	0.83	0.85	1.0
Recall	0.31	0.39	0.80	0.83	1.0

As part of this work, a visualization of the results by means of a web-mapping application is presented. We have created a system that shows accidents and conditions in real time; therefore, we need two different ways to represent them (points and lines). According to that, different representations of traffic events, based on the number of geographic elements detected (see Figure 7).

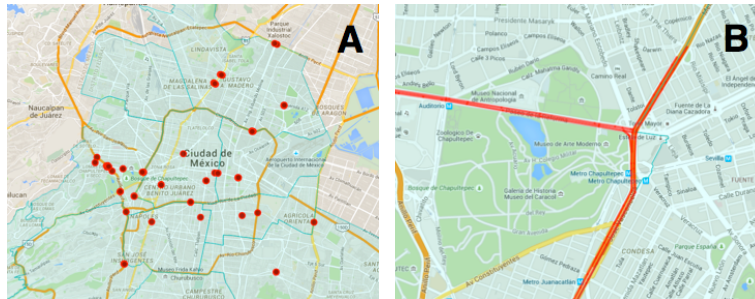


Figure 7. Based on the number of geographic elements detected, image A shows events classified like accidents, therefore their representation are points, in the other hand, image B shows events that were classified as conditions, so they are represented as lines.

6. Conclusions and future work

In this paper, a novel methodology to geolocate traffic-related events in Twitter is proposed. We improve considerably the geocoding, using a gazetteer enriched with information from Twitter’s stream. We also discovered how to divide traffic-related events in order to give more accurate representations, and found that the number of geographic elements in a tweet has a relation with the kind of traffic-related event. This research demonstrated that there is a relation between Twitter participation and the rush hours in the city. Geocoding of traffic-related events in Twitter or even in the web is an useful resource to know the behavior of city conditions, such as how to avoid crowded areas, assign traffic

polices or detect broken traffic lights. This information can be used for training a machine learning approach and making predictions about the city conditions throughout the day.

Future works are focused on discovering other relationships among geographic elements and finding other cases that can occur independently of the assumptions established in this work. Our approach does not consider the direction of the traffic-related event. So, it is necessary to define a method for inferring the direction where the event is located. The nearest areas outside Mexico City have a lot of urban mobility, so it is necessary to expand this methodology to the nearest areas beyond city boundaries. A temporal analysis is required to establish reasonable time duration for an accident or condition. A baseline is set up to each traffic-related event. So, a time duration t could be assigned to them and if the event is mention again, then the time is reset. Moreover, Twitter accounts always post with the same structure, thus a machine learning method could be implemented to learn features.

Acknowledgments

This work was partially sponsored by the Instituto Politécnico Nacional under grants 20151176, and 20151652. Additionally, we are thankful to the INEGI for the free access to its cartography, CNPq and FAPEMIG, Brazilian agencies in charge of fostering scientific development.

References

- Backstrom, L., Sun, E., & Marlow, C. (2010) "Find me if you can: improving geographical prediction with social and spatial proximity." In: Proceedings of the 19th international conference on World wide web (pp. 61-70). ACM.
- Bird, S. (2006) "NLTK: the natural language toolkit." In: Proceedings of the COLING/ACL on Interactive presentation sessions (pp. 69-72). Association for Computational Linguistics.
- Cavnar, W. B., & Trenkle, J. M. (1994) "N-gram-based text categorization." In: Ann Arbor MI, 48113(2), 161-175.
- Davis Jr, C. A., Pappa, G. L., de Oliveira, D. R. R., & de L Arcanjo, F. (2011) "Inferring the location of twitter messages based on user relationships." In: Transactions in GIS, 15(6), 735-751.
- Delboni, T., Borges, K. A., Laender, A. H., & Davis, C. A. (2007) "Semantic expansion of geographic web queries based on natural language positioning expressions." In: Transactions in GIS, 11(3), 377-397.
- Lee, R., Wakamiya, S., & Sumiya, K. (2013) "Urban area characterization based on crowd behavioral lifelogs over Twitter." In: Personal and Ubiquitous Computing, 17(4), 605-620.
- Ribeiro Jr, S. S., Davis Jr, C. A., Oliveira, D. R. R., Meira Jr, W., Gonçalves, T. S., & Pappa, G. L. (2012) "Traffic observatory: a system to detect and locate traffic events and conditions using Twitter", In: Proceedings of the 5th International Workshop on Location-Based Social Networks (pp. 5-11). ACM