Research Journal of Applied Sciences 11 (10): 1083-1088, 2016
ISSN: 1815-932X

Extracting Implicit Geolocation Based on Google Maps Geocoding API of Social Media Networks

[1]Haider M. Habeeb and [2]Nabeel Al-A'araji
[1]Department of Information Networks, College of IT, University of Babylon, Hillah, Iraq
[2]Ministry of Higher Education and Scientific Research, Hillah, Iraq

Abstract: Social Media such as Twitter has become one of the essential bases for supporting and evolving applications in several life fields such as commerce, manufacture, education and health. In the health section, the extraction of geolocation can be exploited in detecting epidemic outbreaks. This research aims at proposing a framework to improve the academic and scientific knowledge about a population health or any other fields interested in geolocation retrieval. The method has been implemented by incorporating google maps geocoding API into a dataset of social media networks for retrieving implicit geolocation. It has been applied on thousands tweets that have not had an explicit geolocation in the metadata. The findings obtained inclusion most dataset collected rather than ignore the users that have no explicit geolocation. The evaluation showed that there is a noticeable improvement in the dataset by including about 80% of users who did not have an explicit geolocation. Thus, their data can be used instead of ignoring it.

Key words: Geolocation, social media network, geocoding API, Twitter

INTRODUCTION

Including geolocation in social media has become a useful technique for many sections that use location-based applications. Twitter as a social network is filled by millions Tweets (messages) that are updated by people from all around the world. For instance, the analysis of geolocation associated with Tweets can help in detecting earthquakes quickly (Sakaki *et al*., 2010). The main mission of the surprised statistics of Twitter focuses on Twitter as microblogging-messages. It has more than 310 million active users and one billion visitors monthly. These features motivated us to analyze tweets of Twitter users and extract their implicit geolocation.

To improve the academic and scientific knowledge about a population health, social networks represent a valuable resource for the medical informatics and public health fields (Habeeb and Aaraji, 2015). Data of Social Networks can be studied to track the spread and occurrence of diseases learn about trends in tobacco and drug use (Prier *et al*., 2011). and understand pain and other ailments (Heaivilin *et al*., 2011). Knowing the geolocation of social networks users became an essential part of this kind of applications.

Social networks applications can include two types of geolocation. The first one is the explicit geolocation where it can be directly found in the user profile such as his/her country, city and/or full address. Another type is the implicit location which can be extracted based on some attributes or objects associated with the user and message. Twitter as a Social Network includes an object/attribute called location which allows for users to write their location instead of extracting it by the GPS technique in the PC or Mobile devices. Most of these values of "location" objects refer to the real location of users. Many users don't prefer to be tracked by applications, so they disable the GPS of their mobile or PC. In this case, such applications are unable to record the geolocation of users based on their metadata.

Developing a method for extracting the full address of users in accordance with the implicit location is the key objective of the present research. Our framework aims at proposing an easy way for extracting implicit geolocation and avoiding its ignorance. Geocoding API on Google Map was adopted in order to address the aforementioned issue and identify the implicit geolocation in the Social Networks.

Literature review: This study includes some of the relevant works that investigated the use of social networks for extracting geolocation. Also it explains the current method of social networks such as Twitter in determining geolocation of user and tweet as well. As such, it is divided into two sub-sections; related works and determining geolocation in social networks.

Corresponding Author: Haider M. Habeeb, Department of Information Networks, College of IT, University of Babylon, Hillah, Iraq

Literature review: Carmen is a system aims to allocate a position to each tweet based on a database of structured location information. Carmen researchers built an alias set by using two stages. The first one is eliciting common profile locations over a huge number of tweets, while the second stage is yielding the actual places such as "Oxford," "California") as well as the false places such as "Earth", "Neverland". A combination of automated filters and a manual review removed invalid places and combined duplicates (e.g. "New York City" and "NYC".). The list resulted from this process was geolocated by using Yahoo's PlaceFinder API. Because a location for false queries can also be returned by the API, the list was further trimmed and aliases were combined in accordance with the returned location. The final list includes 4811 individual places. The list was called "Human Curated". The aliases were added based on a new resource from, whereas according to the observed social network, the observed attributes of a huge number of Twitter users were clustered. Information included in clusters is first names, last names and locations. This process was also used in to augment the alias list. It has the ability of discovering the unknown user-provided locations on Twitter, for example, bmore and balto. Users of Twitter with locations such as balto or bmore frequently have a communication with users with the location Baltimore, where this process works based on the observation of their location. This list is known the "Automatically Extended". The system in the study conducted by, 90% of accuracy was achieved for predicting the correct countries of tweets based on the location that was explicitly provided by the user.

In (Sakaki *et al*., 2010), earthquakes as the real-time interaction of events have been investigated in Twitter. An algorithm was proposed to monitor tweets for detecting a target event. Subsequently, a classifier of tweets was devised based on many features. For example, the study used the number of words, the keywords in each tweet and the words context. As such, a probabilistic spatiotemporal model was produced for the target event. It was used for two purposes. The center of the event location can be found and its trajectory. The study considered each Twitter user as a sensor. Both Kalman and particle filtering were applied due to their popularity for location estimation in ubiquitous/pervasive measurement. The findings indicated that the particle filter achieved better results in comparison to other methods used in the estimation of the centers of earthquakes as well as the typhoons trajectories. The earthquake reporting system was constructed in Japan as a practical application. This is because earthquakes widely happen in this country and due to the numerous number of Twitter users, the application had high predictability of earthquakes based on monitoring users' tweets (96%). The detection of earthquakes in accordance with this system is promptly, where it sends e-mails to all users registered. Further results indicated that the notification is delivered faster than the traditional announcements of the Japan Meteorological Agency (JMA) (Sakaki *et al*., 2010).

The researchers in Han *et al*. (2014) have examined many central issues which are associated with text-based geolocation deducing for users of Twitter. Many feature selection methods were applied to highlight the location indicative words (LIWs) and demonstrate the efficiency of this method on global (WORLD) and regional (NA) datasets. The research then was extended by analyzing the influence of non-geotagged data and the effect of language as well as information of the complementary geographical in metadata of users. The researchers showed that the use of the explicit technique can promote the accuracy of predicting geolocation, in comparison to the use of the full feature set. The feasibility of modeling and inferencing geotagged and non-geotagged data was indicated as well. Accordingly, the research concluded that predicting "new" data based on a trained one "old" is plausible. In the estimation of the prediction confidence, the probability ratio can be measured as well as selecting only users in a case of obtaining high accuracy in the system prediction (Han *et al*., 2014).

Determining geolocation in social networks: Most social networks include the geolocation of their users and posts by using different techniques. In this study, Twitter was used because it represents one of the most popular Social Networks. Our research shows how Social Networks deal with the geolocation, it more especially focuses on the Twitter techniques.

Twitter provides various data APIs and tweets come as JSON objects that include the tweet text alongside the metadata such as time, location (coordinates) (if it is provided by the user) and user profile information. A user profile includes optional user-provided information such as his/her real name and location. These metadata can be used to geolocate the tweets. In particular, there are four primary ways in which geolocation is commonly determined on Twitter users and messages. These methods are briefly discussed below:

Tweets place object: Many tweets are delivered by the Twitter API comprise a JSON "Place" object. Its role is to encode a location linked with the user's tweet. These cover different fields such as the city linked with the place, the country, as well as the coordinates. The place types encompass finer-grained information such as the

number of known places, street addresses and business names. Twitter provides its users the ability of providing their locations or not. Based on the user's GPS location, the tagging can be automatically performed as well in a case that the user has activated this method.

Tweets coordinates: According to GPS locations of Twitter users, the coordinates of these users can be determined in order to geolocate their tweets.

Profile locations: Public locations are provided directly by some Twitter users in their profiles as a free form with string values. It is quite often to use such strings in order to resolve and structure the positions based on the existing map APIs. These locations are coarse-grained and the most of them are static, corresponding to the user's primary location rather than the location at the tweeting time which may be different if the user is traveling. A high number of Twitter users have profile locations instead of the geocoordinates.

Content-based geolocation: another technique to determine the Geolocation is based on a message or set of messages in accordance with the textual content of the messages. This technique is one of the common applied methods. According to users' dialect, their primary position can be inferred. In a case that users do not provide their explicit locations, these techniques can be adopted.

Google Maps Geocoding API: One of the most remarkable developed tools is Google Maps API. It is easy to use and incorporate into web applications. In order to provide and reverse geocoding of addresses, the Google Maps Geocoding API service is used. Geocodin is defined as the method of converting the user profile information such as a city address into latitude and longitude which is known the geographic coordinates. The use of geographic coordinates can lead to placing markers on a map, or positioning the map. Reverse geocoding, on the other hand, refers to converting the geographic coordinates into a human-readable form. The Google Maps Geocoding API's reverses the geocoding service and let to locate the address for a certain place ID.

A limitation in Google maps geocoding API can be considered where it allows for only 2500 free requests per day, otherwise payment has to be done for extra requests. In some cases, the same name can have more than one region . In this case, Google maps geocoding AIP returns geographic information for all regions. Up to date, the determination of exact information is still an open problem that requires further research. For example, the inquiry about "waterloo" returns geographic information for two countries which are USA and Canada as shown in Fig. 1. However, if the value of the "location" attribute is

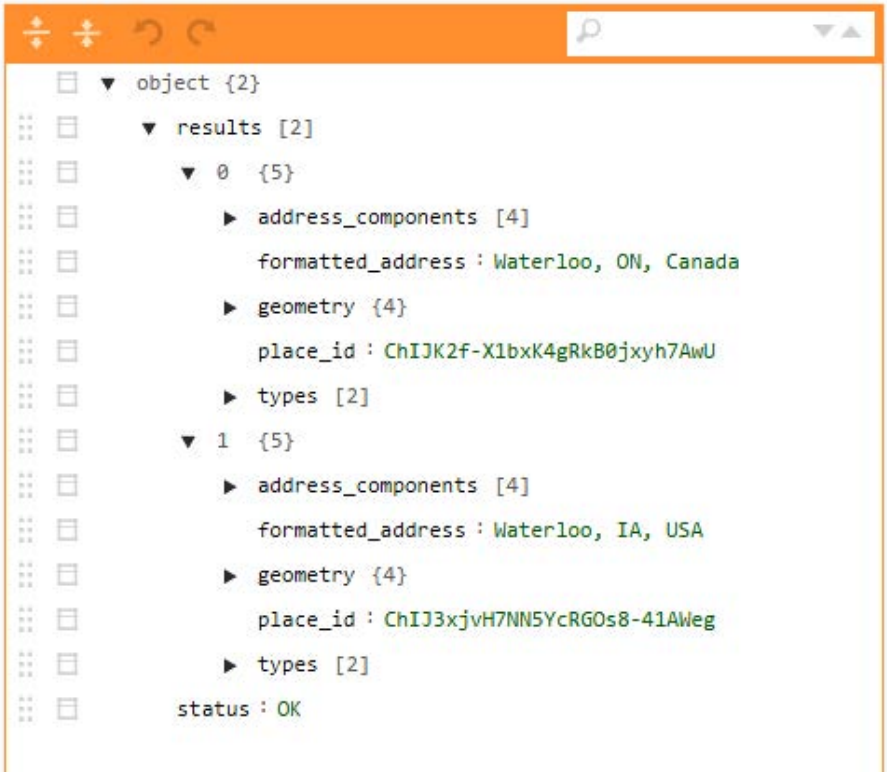

Fig. 1: Structured JSON format for returning results of "Waterloo" word

specific, Google map geocoding API can return the exact result. For instance, when the value is "waterloo, on", the result will be Canada only.

MATERIALS AND METHODS

Extracting implicit geolocation method: In this research, the Extracting Implicit Geolocation (EIG) method is proposed. It aims at extracting implicit geolocation for targeted population on the Social Media Networks. The study also evaluates the accuracy of the proposed method (EGI) and shows its usefulness for geolocation-based applications. Figure 2 explains the framework of extracting implicit geolocation by incorporating the Social Media Networks with Google Maps API services.

Any Social Media has its own metadata such as the post, tweet, message and comment. Therefore, a part of this metadata can be associated in somehow with the explicit or implicit geolocation. Our framework plays an essential role in finding the geolocation for an implicit geolocation. It is obvious that the explicit geolocation has already full information about the geolocation of the tagged user.

Twitter as a social media network has an attribute/object in its metadata which is called 'location' and it can be found in the user profile. It is an optional for users to write anything refers to the undetermined location name which mostly is a part of the address such as street name, neighbor name, city name, or country. Sometimes, the user writes unrelated words in 'location' attribute, so the EIG method will ignore it. The EGI sends the location name as a request to the Google Maps Geocoding API and receives geocoding of the location name as a response. Algorithm 1 describes the steps of the EIG method.

Algorithm 1: Extraction Implicit Geolocation

Input: T, L where T is Tweet, L is an attribute contains word(s) refer to place.
Geocoding AIP: online geocoding AIP from Google Map.
Output: Aggregated geolocation database
Processing:
Step 1: Read Tweet
Step 2: If T has explicit geocoding, go to End.
Step 3: If T[L] is empty, ignore T.
Step 4: for T[L] has location name, fetch geographic information from geocoding API.
Step 5: Match extracted geolocation with the real geolocation of the same T.
Step 6: Evaluation; Accuracy = Number of Matching Geolocation / Total of Real Geolocation
Step 6: End

As demonstrated in the framework structure (Fig. 2), there are explicit and implicit modes in the social media networks metadata such as Twitter. The real geolocation of users in the dataset was used for testing the accuracy of the proposed method.

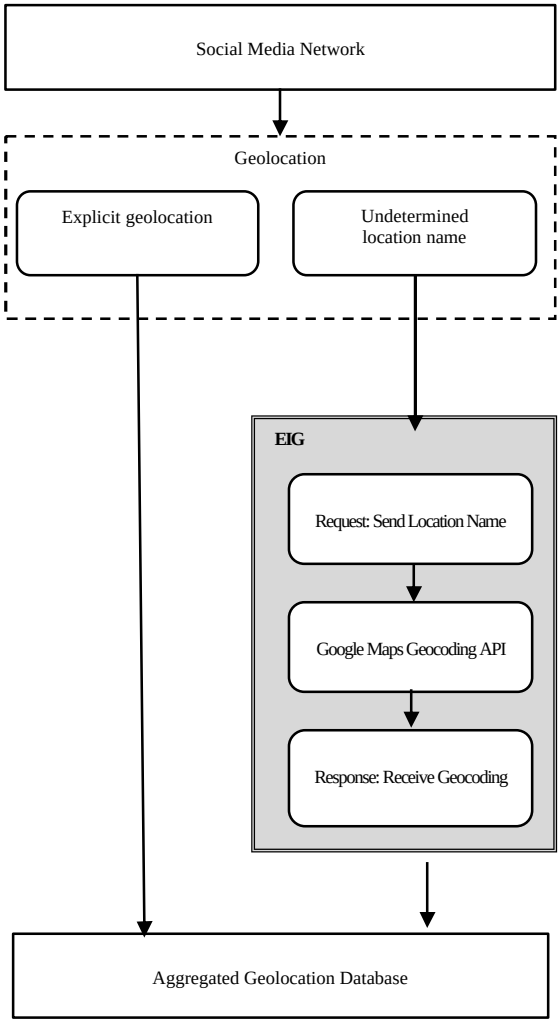

Fig. 2: Framework of extracting implicit geolocation

RESULTS AND DISCUSSION

More than 10000 metadata of tweets were collected to test the accuracy of the proposed framework for extracting implicit geolocation by using Google Maps Geocoding API. Users who have real geolocation (explicit geocoding) and location attributes in their profile were used in the test stage.

Extracting implicit geolocation was conducted by feeding the Geocoding API project which was created on the Google Maps. Feeding process in our method depends on the content (value) of the 'location' attribute in the user's profile. After extracting geolocation by the proposed method, a matching process was performed between the extracted geolocation of our method and the explicit geocode of each user. The results confirmed the accuracy of the proposed technique which exceeded 80%.

Table 1: Sample of results

User ID	Location	Explicit country	Extracted country by EIG	Matching
2534654015	La Quinta, Ca.	United States	United States	Yes
326037276	Plainfield, Indiana	United States	United States	Yes
2246072656	MofoloNorth (Soweto)	South Africa	South Africa	Yes
344922401	nj	United States	India	No
1079581686	Ukraine	Ukraine	Ukraine	Yes
425520312	Damansara - Kajang	Malaysia	Malaysia	Yes
451386980	+65	Singapore	United States	No
538080444	still not at hogwarts	Philippines	ZERO_RESULTS	No
392719858	Edinburgh	United Kingdom	United Kingdom	No
456919839	Lagos Nigeria	Nigeria	Nigeria	Yes
128883516	Diantara pria-pria romantis~	Indonesia	ZERO-RESULTS	No
982248684	Cebu City	Philippines	Philippines	Yes
315237388	Blantyre, Malawi	Malawi	Malawi	Yes

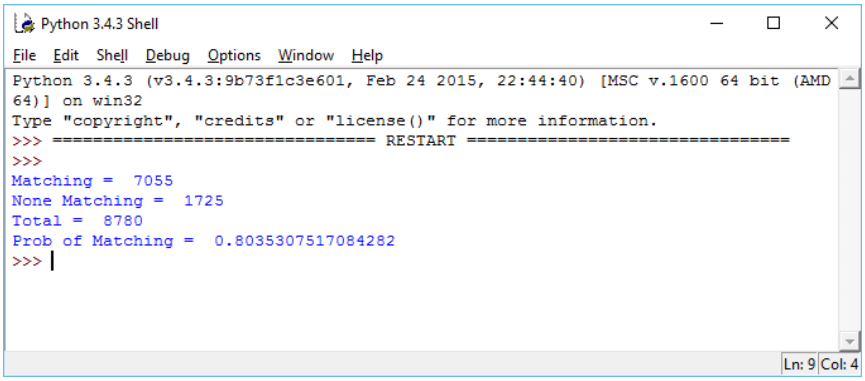

Fig. 3: Some statistical numbers of results analyzing

Table 2: Statistical information of framework

Total Tweets	10961
ZERO_RESULTS	2181
Net Tweets	8780
Matching	7055
Not Matching	1725
Accuracy	0.8035307517084282

Some findings returned "ZERO-RESULTS" because there is no real region name in the "location" attribute as shown in Table 1 where the content was "still not at hogwarts".

Therefore, the geocoding API project couldn't recognize the country name of the "location" content. Some of the results obtained were not identical with the original country. This problem is attributed to name similarity of many regions in different countries. Hence, the EIG method based on the first object that comes from Google map geocoding AIP. As long as there is no explicit geolocation, ZERO-RESULTS should be ignored from the EIG method.

Figure 3 illustrates some statistical information resulted from the proposed framework application. It is noteworthy that the proposed method was programmed by using Python 3.4. Table 2 briefs more statistical information of our framework regarding the collected data.

CONCLUSION

Social Media Networks such as Twitter come with explicit and/or implicit geolocation. Geocoding AIP service that comes with Google Maps is very useful tool for extracting implicit geolocation. The proposed framework in this research achieved more than 80% of accuracy. The accuracy have been computed as illustrated in algorithm 1 (step 6).

This outcome suggests that using Google map geocoding AIP in the location-based application is an efficient technique to be used in different fields such as the health system.

Our framework guarantees 80% of population that have implicit geolocation will be included in the determination of the geolocation for controlling any possible epidemic outbreaks and only 20% of them may be ignored at variance of other systems that based on explicit geolocation must exclude 100% of population that have implicit geolocation.

REFERENCES

Habeeb, H.M. and N.A. Aaraji, 2015. An overview on the use of data mining and linguistics techniques for building microblog-based early detection systems in the healthcare sector. Int. J. Comput. Sci. Inf. Technol., 7: 143-155.

Han, B., P. Cook and T. Baldwin, 2014. Text-based twitter user geolocation prediction. J. Artif. Intell. Res., 49: 451-500.

Heaivilin, N., B. Gerbert, J.E. Page and J.L. Gibbs, 2011. Public health surveillance of dental pain via Twitter. J. Dent. Res., 90: 1047-1051.

Prier, K.W., M.S. Smith, C.G. Carrier and C.L. Hanson, 2011. Identifying health-related topics on twitter. In: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction, Salerno, J., J.Y. Shanchieh, N. Dana and S.K. Chai (Eds.). Springer, Berlin, Germany, ISBN: 978-3-642-19656-0, pp: 18-25.

Sakaki, T., M. Okazaki and Y. Matsuo, 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. Proceedings of the 19th International Conference on World Wide Web, April 26-30, 2010, Raleigh NC., USA., pp: 851-860.