2019-03

Localising social network users and profiling their movement

Pellet, H

10.1016/j.cose.2018.10.009
Computers and Security
Elsevier

Localising social network users and profiling their movement

Hector Pellet[a], Stavros Shiaeles[a] and Stavros Stavrou[b]

[a] *Centre for Security, Communications and Network Research (CSCAN). University of Plymouth, Plymouth, UK*
[b] *Faculty of Pure & Applied Sciences, Open University of Cyprus, Nicosia, Cyprus*

hector.pellet@postgrad.plymouth.ac.uk, stavros.shiaeles@plymouth.ac.uk, stavros.stavrou@ouc.ac.cy

Abstract

Open-source intelligence (OSINT) is intelligence collected from publicly available sources to meet specific intelligence requirements. This paper proposes a new method to localise and profile the movement of social network users through OSINT and machine learning techniques. Analysis of obtained OSINT social networks posts data from targeted users, suggests that it is possible to extract information such as their approximate location, leading also to the profiling of their movement, without using any supported Global Navigation Satellite System functionality which may be passed to the social network through a capable smart device. The ability to profile a target's movement activity could allow anyone to track a social network user or predict his or her future location. Moreover, in this work, we also demonstrate that information from social networks can be extracted relatively in real time, thus targeted users are prone to lose any sense of physical privacy.

Introduction

Personal data has become omnipresent in the past two decades, originating from the vast usage of the internet and its social networks. Since various organisations rely on this data for business intelligence, or other uses, data processing time changed from periodic to real time. Under this context, the emergence of collection and aggregation of personal data, and, location information, gave rise among others, to innovative services such as Uber, Tinder, Facebook, Foursquare etc. Analysis of the available data can be used to produce relevant user patterns, help to profile users or even anticipate their future behaviour (Gritzalis et al., 2014). For example, geolocation technology which is the foundation for location-positioning services and location-aware applications (apps), used among others in social media, can be used to localize assets and persons of interest (Yang et al., 2012).

Geolocation has become an essential aspect of our daily life (Mansfield et al., 2017), used in navigational platforms or even social networks. With geolocation functionality, one can determine the location of an object or person, assisting the functioning of contemporary daily life. The widespread and usage of smartphones, the generated social network data, the ability to provide real-time analysis of natural language and friend-based inferring, and the usage of machine learning, can assist in the geolocation of social network users. Collection and manipulation of such information can be classified as part of an Open-Source INTelligence (OSINT) process (Heuer, 1999). OSINT differs from traditional intelligence since it collects data through public accessible resources such as social networks, media, blogs and web communities (Heuer, 1999). Social Networks trigger

users to rapidly share data between widely used networks. Users can post content in a variety of formats, which can be instantly made available to their entire social network (Giordano et al., 2015; Cho and Choi, 2014). Thus, OSINT has the advantage of being accessed without any limitations. Its main purpose is to serve specific intelligence needs, through the collection, processing and correlation of open source information. Intelligence should always provide credible information, positively contributing to the information analysis process. Continuous gathering and delivery of reliable and precise information can lead to precise answers and correct strategic decisions.
Social networks provide an important platform for disseminating open-source information, including any web content, opinion sharing, discussions and debates. The large volume of public data that flows through social networks has the potential to deliver valuable new insights, regarding user behaviour, and the analysis of such data is of interest to the academic community, marketing agencies, or other organisations involved with the analysis and understanding of user online behaviour and social trends (Puzis et al., 2009; Giordano et al., 2015).

The dissemination of this open source information, which includes personal data, other sensitive data, or even metadata, is amplified through the use of smartphones over cellular networks. Nowadays, social network users can connect anytime to their social network platforms and disseminate in real time any personal information. Nevertheless, availability of such real-time, or almost real-time personal information, may jeopardise a user's privacy or even physical security. The rise of security risks because of the spread of such information highlights that privacy protection is one of the most alarming issues in the contemporary society (Tan et al., 2012). Nowadays, the volume of personal and sensitive data is derived mostly from social networking platforms. Facebook for example holds records for 2 billion users, which on average they interact 4 million times per minute (Brandwatch, 2018). Needless to say, it is very difficult to anticipate and track around the world the information disseminated via the internet (Schoen et al., 2013).

On the protection of personal data, social media platforms provide their users with some basic privacy setting tools. Cognizance by the content owner regarding the management of digital information plays an essential role in the protection of personal data. Currently, Twitter allows restricting the visibility of the user's information to his followers and not all Twitter users. On Instagram, the visibility of a user content can be limited only if the platform is being used in the E.U. (Twitter, 2018; Instagram, 2018). Additionally, many social networks ameliorate their advertising policies using the personal data they have already received, and they put ads on the site according to the personal interests of the users. Use of personal information for such reasons is clearly stated in the user agreement, accepted by the user during registration into the social platform. Furthermore, due to their dominant position and the fact that they provide free access to their social network platforms, until recently, the service providers can change the privacy agreement without having the confirmation of the end user (Kalampokis et al., 2013)

With regards to localization, on Twitter and Facebook profiles, location data can be made available to third parties and this information can be extracted from the user's profile, leading to the visualization of the user's location on a map. On Instagram, location can be provided in the form of longitude and latitude information (Perwitasari et al., 2015; Xie et al., 2013). Since a vast amount of information is exchanged through social networks, one may ask how all this publicly available information can be used to track a user's location. This work examines the above question by collecting and analysing data from three social networks, namely Twitter, Facebook and Instagram. To prove the possibility of localising a social network user based on relevant OSINT, a toolkit was developed, consisting of a number of modules for data collection, spatial storage of offline data, spatial data retrieval and analysis, full-text search, and geolocation data mapping.

This work discusses the toolkit architecture, its modules and the processes involved in profiling and predicting the location of a social network user.

Related work

User geolocation is an important field of study as an attacker can invade a user's privacy. Geolocation can be obtained through different methods such as an app that can infer users' location and travelled routes without the users' knowledge, using gyroscope, accelerometer, and magnetometer information (Narain et al., 2016) or through Social Networks. Our research focusses on the Social Networks feeds and close friends. In this respect, Davis Jr et al. 2011 main goal was to acquire in real-time the geographical distribution of a disease discussion to identify regions of disease outbreaks. To locate targets, they extracted users' location from a social graph. The non-located users' locations were set to the most popular location on a simple voting scheme. This meant that each friend's location had the same weight as the final result. The number of votes and the number of followers of a friend may avoid this location inference: too many followers can mean being a celebrity and not enough may lead to no confidence in the data provided. Under this knowledge, the authors iterated through the social graph until no progress was possible. This method was designed to provide only one location per user, where practically multiple locations per user over time should be spanned.

In a similar context, Jurgens 2013 used label propagation for inference user's location based on a friend network. In this case, the place names were propagated instead of the coordinates. Two problems may arise when propagating place names. Firstly, no calculation may be executed on the labels, and this could limit the geolocation process. Secondly, ambiguity may appear with place names, which cannot appear with coordinates. The ideal solution would be to propagate the exact piece of an element that infers the location. However, this element might not be comparable with other information (e.g. a place name is not fully comparable with coordinates). To conserve a trace of the type of element used to locate a tweet, a column source was used to indicate how this location was obtained.

Under the model of Compton et al., 2014 ground truth is only based on GPS information, and friends' role in that model is to only propagate location. At the start of the process, the locations that do not provide GPS geotag on their tweets are marked as unknown. For each iteration, the location of a user is updated according to the multivariate median of their friend's location. In this method, it is important to provide enough friends to propagate the location of the targets. In our context, it is considered that the above social graph method may not contain enough relations in a relatively short time. Also, propagation of a friend's location helps to locate users according to the relationships it maintains, so to improve this method, an approach should be used that classifies friends according to their closeness.

McGee et al., 2013 used a CART (Classification And Regression Trees) algorithm trying to distinguish between friends that are closer than others by using factors such as the number of followers, the amount of communication, whether the account is private or not, the proximity of friends, the location they were and the type of contact: mentioning, following, being followed, reciprocal following, reciprocal mention. The authors themselves recommended combining their method with natural language to improve the precision of the geolocation. Building on this, Kong et al., 2014, examined the linkage between the probability of being a friend and the distance between two people. Cosine similarity is used to measure the social closeness. The closer is a user to one of his friends, the more influence this friend will have on the location of the user. However, their work was not very precise as estimated location and real location marked as correct if less than 100 miles.

Locating people based on indices, such as place names, or from where a post originates, is called "natural language processing". Roller et al., 2012 examined already geolocated documents (not specifically Tweets or Posts but longer texts), where they extracted place names from the text. Then the distribution of those place names was compared to the text they wanted to locate. The most similar document (similarity based on the distribution of words) indicated the location of the targeted document. However, their method cannot be applied for real-time user tracking, and the results regarding accuracy are close to 51% within 100 miles.

An equivalent technique was used by Ghahremanlou et al., 2015 and was applied to tweets. In their work, they clustered tweets based on crisis events and geolocation information (e.g. tweets related to an earthquake in California). They first gathered already-geolocated tweets and from those tweets by using GeoNames, a gazetteer, 'local lexicons' are constructed. "Local lexicons" are distributions of place names in function of the places the text is supposed to originate. However, since different places can have the same names without having the same coordinates, the place names need to be disambiguated, so some heuristic rules were generated. As evoked in this paper, tweets are limited in length and as such so is the number of place names. With Facebook, this effect is expected to diminish since users tend to write longer posts. Since this work does not focus on crisis management but instead is focused on tracking social network users, this method is expected to be applicable since tweets will be located individually, instead of being located as a group related to the same crisis.

Mapping of people's location, interactions, and their social ties, was investigated by Sadilek et al., 2012. In their work, they used a decision tree to represent a series of questions. This process assisted in estimating the probability of friendship between two users and its closeness. Then the location history of one user was modelled with a dynamic Bayesian network. Their method was mainly based on friendship and did not take into account other parameters for locating users. Ahmed et al., 2013 decided to model both location and topics in the same hierarchical tree. They intertwined location and topical distributions into a joint model, trading off between improved spatial accuracy and a better content description.

Other methods are based on a set of unprecise or widespread clues. Backstrom et al., 2010 based their work on a maximum likelihood approach to predict a user's location. Their research work

aimed at retrieving the home address where only one location was considered per user. Crandall et al., 2010 tried to infer social ties based on co-occurrences: by placing users on a grid map over time and by counting the time's people were on the same cell, they determined the probability of existing relationships between people that visited the same places. This method can be applied to data retrieved from multiple social platforms since it only requires the identification and analysis of relevant data. Kotzias et al., 2015 also tried to locate tweets based on a maximum likelihood approach. Their method was based on awareness of the tweets location habits. They first extracted the potential place names, then calculated the probability the tweet was posted at every location in the areas of the geolocated words. Then the one that maximizes the likelihood was picked. The method relies on the splitting of the map into grid cells. The smaller the cells, the more accurate estimation is expected, but multiple iterations will be required to reach a solution. The method was tested with data from three big cities.

Looking into prediction models, Serdyukov et al., 2009 used the description of a photo in Flickr to generate a ranking list of locations. They then used spatial minimality to disambiguate places. With spatial minimality, one aims to retrieve the different sets of the location which contain all the place names found in the text, in the smaller possible spatial area. A naive Bayes classifier is also applied to the results. This method was only used on image tags, but nothing prevents it from being used on wider texts from Twitter or Facebook. In addition, it is expected that this method should apply successfully to Instagram since Instagram and Flickr are very similar social platforms.

Regarding mobility, Cho et al. 2011 divided the human mobility pattern in two categories, the periodic model and the social one. Based on that classification they managed to identify 60% to 90% of the user mobility via only the social network. However, this model is limited to Brightkite and Gowalla and only takes into account, the home and work addresses. The work presented in this paper intends to geolocate the user at any location in the world.

Finally, Cranshaw et al., 2010 analysed the social context of a geographic region, to measure the diversity of unique visitors at a location. They predicted friendship using various techniques: using GPS, Wi-Fi and IP, via specific web and mobile applications. The main drawback of the method is the required use of specific applications. Since users may decline or not participate through these applications limits the scope of the work.

Based on the literature survey, it appears that researchers have developed different approaches that can be used to track/localize a user, based on social network data. However, none of these approaches combines the simultaneous usage of social network APIs, real-time crawlers, machine learning and natural language processing. The proposed methodology in this work combines all the above elements in an effort to achieve better user localisation profiling. Obtained results suggest the usefulness of the proposed approach.

Proposed methodology

The goal of this work is to localise a user based on the social network available information. Firstly, relevant social network data is gathered and the basic features of that data are extracted and placed in a database for further analysis. Location information, if available, is not the only feature that has to be extracted: the relationship between users has to be retrieved, as well as IP direction. Even the username may be useful, for retrieving other accounts belonging to the same person. In this paper, the category defined as "friends" does not relate to friends in its pure etymology, but instead, it defines any person with which the user is directly connected.

To efficiently aggregate and analyse the gathered data, an appropriate model is required. To achieve this, we examine data from Twitter, Instagram and Facebook in order to understand which relations link the different entities, so we can efficiently apply our model.

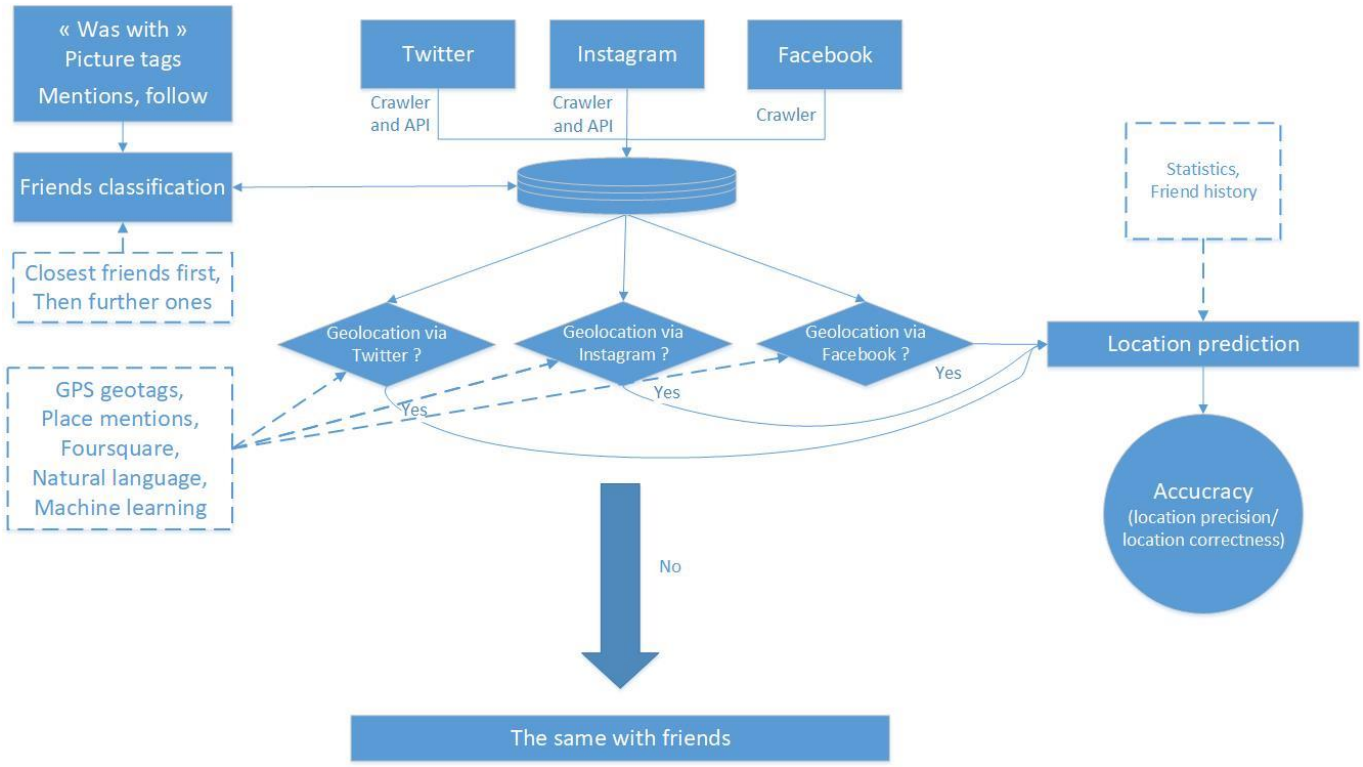

Figure 1: Proposed Methodology – Platform Architecture

Data gathering

It has to be noted that participants initially were friends of the research team and then it was expanded to friends of friends who gave their consent through email, to store and analyse their data shared on Social Networks for a given timeframe and only for this research. We followed all the appropriate procedures to store data using AES-256 encryption and also the raw data where sanitise using Random Data Method to protect participants privacy after the completion of this research.
Data was captured and analysed through three different social platforms: Facebook, Instagram and Twitter. The data was gathered via a crawling process and/or through an API whenever such an API was available. The data was then stored in a database (Figure 2). If the geolocation information is available in the captured information and reflects the user's real-time location, then the user is located in real time. If the user cannot be located through the previous process, the method proceeds to examine 'friend location' to localize the target. It is worth mentioning that the proposed method starts localizing users from Twitter and expands its search through other social networks

based on the information found. Through the processing of a growing amount of location data, the users' location should be retrieved.

Concerning the data gathering process, the use of crawlers and APIs are different, since the crawler gathers information from the timeline, while the API if one is available, is used to gather information for one user at a time, and limitations may apply. Friends of a user will be first identified by crawlers, and only then any existing API will be utilised. This process provides more flexibility in case restrictions be imposed by the API. In our implementation, crawlers are python scripts using selenium for browsing the website of the social platform of interest (Seleniumhq.org, 2018). The usage of crawlers also allows identical steps to be followed on all social network platforms: first, the crawler propagates to the news feed to extract available information. A java script function scrolls down the webpage until all news feeds are collected. Taking Twitter as an example, same principles apply to the other two social networks, when the process reaches the end of the news feed, it splits it into blocks of HTML that represent the inter-independence of posts/Tweets. Then a java script is executed by selenium to extract the post text, the time it was published, the name of the user, the mentions, and the location and the image information if those are present. The mentions in the posts are embedded as HTML links and can be easily identified. By checking the class of the HTML element (link tag), it is possible to identify whether it is a mention or just a link. Also, under Twitter, a '@' is present, preceded by whitespace (or the '@' can be the first character of the Tweet), followed by alphanumeric characters. Location usually appears under the name of the user. However, instead of the name of a user, it can appear in groups of conversation, which eliminates the location information.

As aforementioned, every piece of information can be extracted stored and analysed to locate a user. Images from the posts/tweets can be used by extracting the metadata (mostly the embedded geolocation in EXIF metadata), however, during the testing phase, we discovered the metadata was wiped out except the one relative to the colours and the MAC times. Thus pictures were not utilised in the current research.

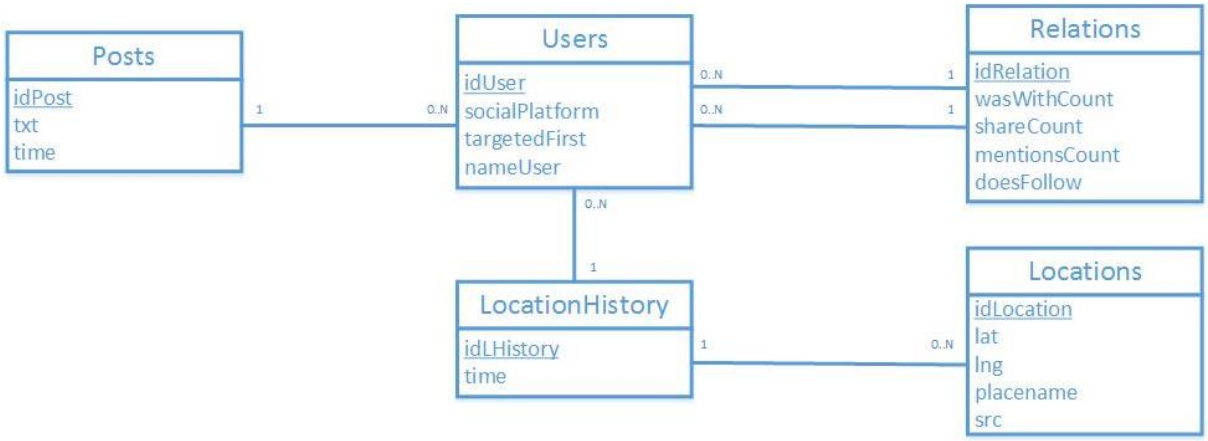

Figure 2 : Database tables and relations

Analysis of relationships and location retrieving

As stated by Mitrou et al. 2014, accessing a user's social media data can be performed indirectly via other users that communicate with the target user. This is the keystone of user ties geo-inference. This section introduces the techniques used to build the social graph and more precisely, retrieve relationships between users. It has to be noted that the difference between this

work and the work of Mitrou et al. 2014, is that images are not used directly, but instead, their tags are used to identify if target users were at the same place. In this case, the crawler identifies it, and the relation between the two identified users is updated to increase the count of 'was with' field in the database.

After enumerating the database with who has mentioned the target, or retweeted content from him or her, we update the relation. Knowledge of the relationships a target maintains with his friends can assist in correlating their location and estimating the location of the target. Finally, the location that mostly appeared through friends is identified as the location of the target.

During the analysis, we used Geotags embedded in the tweets/posts were used and natural language tools such as CLAVIN (Cartographic Location And Vicinity INdexer) and Geograpy. The process looks for place mentions in order to find the location from where the tweet was emitted. In its most basic form, it uses Geograpy to look through the words that compose the tweet/post and returns an array of possible places. This is the most brute way for extracting place names since no calculation is made to locate globally the tweet/post. However, it provides a base for further estimations, which can make use of statistics or machine learning. Foursquare was also used: each Twitter user, was examined if he/she has any check-ins since multiple formats of geolocated check-ins exist. Finally, CLAVIN, an open-source software package for document geotagging and geoparsing was used. This tool not only lists the different place names but also uses heuristic strategies from natural language for identifying place names. It is also capable of extracting place names over multiple words (Île-de-France, United States of America, etc.). Moreover, it uses machine learning to improve the outcome of the analysis. It was found to provide the best compromise between accuracy (as it is based on multiple place name spotting) and compatibility (applicable to every post/tweet that contains text).

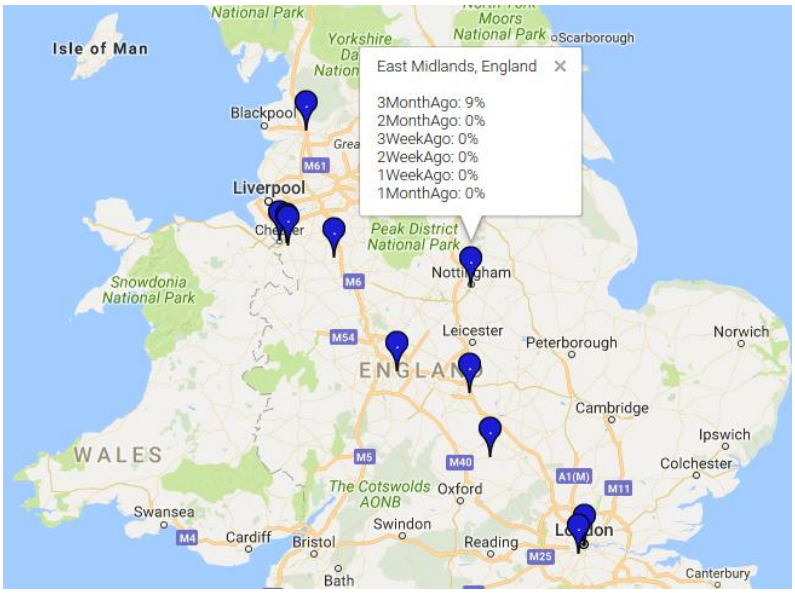

Figure 3 : frequency map of the place visited by a user

Location prediction

By knowing the location history of a user, it is possible to determine the number of times the user visited a certain place. It then becomes possible to calculate and assign a 'future visit' probability to every place the user already visited. Based on the above, a map was created that

presents every single place a user has visited and provides a percentage of times he or she went there, compared to the total number of his / her location records.

Findings and discussion

After gathering data over the three aforementioned social platforms (Facebook, Instagram, and Twitter), a dataset of more than 6500 users and 5447 locations were collected for those users. With this dataset, we tried to predict the location of a user at a certain point in time. Table 1 presents the result analysis of this work compared with other approaches.

	Our method	**Davis Jr et al. (2011)**	**Backstrom et al. (2010)**	**Ahmed et al. (2013)**
Real-time	Yes	Yes	No	Yes [1]
Predict Accuracy	77.72%	40.3%	25%	64.35%[2]
Natural language	Yes	No	Yes	Yes
Machine learning name	Logistic regression	No machine learning	No machine learning	Gibbs Sampling
Social platform	Twitter, Facebook, Instagram	Twitter	Facebook	Twitter
Friends for location inference	Yes	Yes	Yes	No

Table 1 : Methods comparison

Table 1 suggests that the achieved accuracy from the suggested methodology is higher than the accuracies achieved by other methods presented in the above table.. It is believed that this is due to the additional features we developed that were not included in the other methodologies (multi-platform, machine learning usage, usage of natural language).

	Twitter	Instagram	Facebook	TOTAL
Foursquare Check-ins	8	0	0	8
Geotag (latitude longitude)	11	0	0	11
Geotag (Place name)	1534	171	120	1825
Natural Language (Geograpy)	2500	106	222	2828
Natural Language (CLAVIN)	754	5	16	775
TOTAL	4807	282	358	5447

Table 2 : location source distribution

Table 2 suggests that the location source is not homogeneously distributed. Since most of our data came from Twitter, the location data will mostly be linked to this social platform. Since Foursquare

[1] The location is not really in real time but the geolocation is considered for each tweet rather than for each user.
[2] See Ahmed et al., 2013. The accuracy is calculated on the DS2 findings: ((regions –average_error)/regions)

social media is linked to Twitter, Foursquare check-ins have very few chances to be linked to other Social Medias accounts like Facebook or Instagram. This is the reason why there are no Foursquare check-ins under Instagram or Facebook. Since the Facebook API was not used, some features are not available on the gathered data, geotagged with latitude/longitude. Moreover, Instagram does not provide GPS-geotag, but still, place-names-geotag was exported through the Instagram API. With respect to Geograpy and CLAVIN, the first one picks every place name cited in the post/tweet, while the other one is trained to calculate one location from every single place name presented in the post/tweet. Actually, CLAVIN is based on OpenNLP, which is a machine learning algorithm, so it is possible to increase the number of spotting places or to limit the recognized place names to a geographic region.

One example of a relation created by using the developed system is the favourite place of a user. The location history is classified according to the last time the user went to this place. This allows us to make estimations w.r.t. the place he will be, even if nothing was posted/tweeted. The more 'sighting incidents' he has during a period, means there is a higher probability for the user to be at a specific place. Also, for each user, and for each place he visited, we extract the moment he was there and calculate the interval between each visit. Then, by analysing the intervals between visits, we can estimate the next visit to this place. We also extracted whether two users were located at the same place to enrich the available list of, 'was with'. The case of two people being at the same place is important and has to be taken into account when trying to geolocate people using their friends. This helps us to understand the frequency of a determined behaviour, e.g. when trying to predict if a 'target' is going from home to work and vice-versa. In this sense, the application of fuzzy logic can improve the effectiveness of geolocation, through the calculation of the probability that someone has been at a certain place. Analysing social media information that present where the user or 'friend' have been geotagged together, helps to create probabilistic models to better understand the migration flux.

Extracted features can be used to feed machine learning training algorithms to predict user behaviour, and thus predict the user's current or future locations. To predict a future location, a machine learning algorithm can be fed with the habits of a user over a given time frame, e.g. the number of times he went to specific places. The expected result can be the number of times he may go to those places over a subsequent time frame.

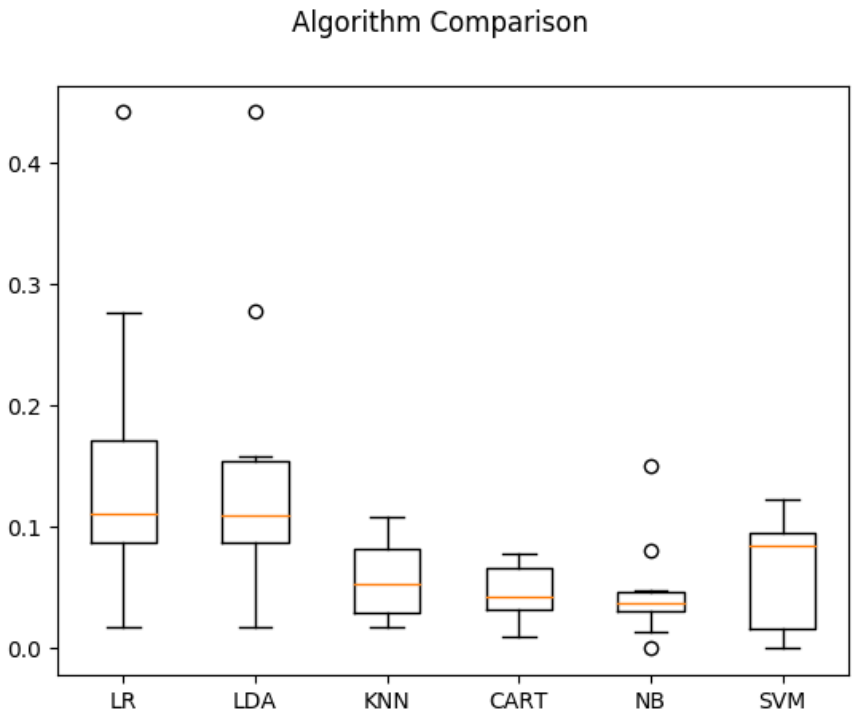

Figure 4: algorithm comparison for location prediction

In order to predict the user location, we tested various machine learning algorithms (Logistic Regression, Linear Discriminant Analysis, K Neighbours Classifier, Decision Tree Classifier (CART), Gaussian Naive Bayes, c-Support Vector Classification (SVM)), and we cross-validated them. Figure 4 presents the performance analysis between the different machine learning algorithms for the dataset used.

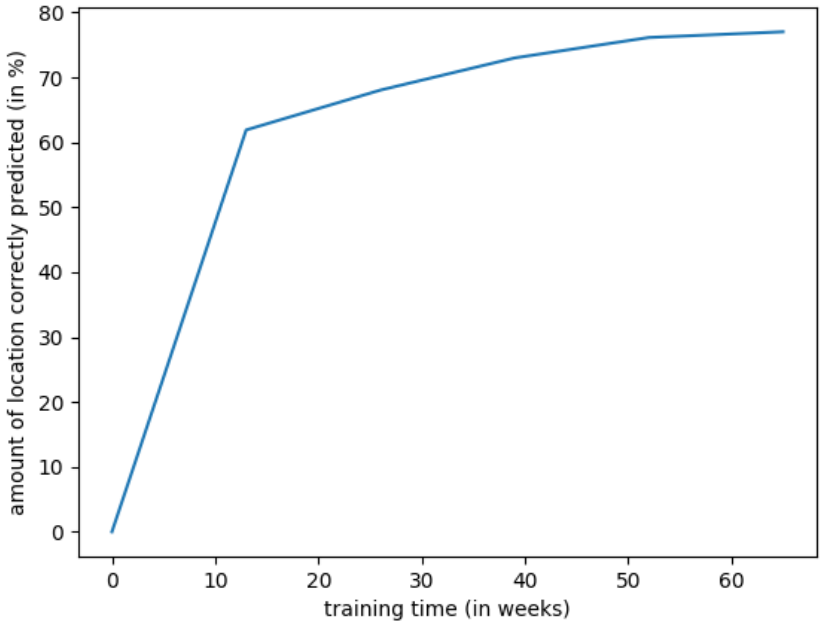

Figure 5: Accuracy of prediction as a function of time

Using logistic regression, we have managed to create accurate user location predictions at a success ratio of 77.72%. The rate was calculated by comparing the true location of a user, based on the actual responses of users involved in the experiment, with the place predicted by the system. Figure 5 also suggests that with further training the suggested framework can achieve even better predictions since after 60 weeks of training the graph did not reach a steady state.

Limitations

Limitations wise, it should be acknowledged that the current work focused on users utilising the three social networks, Twitter, Facebook and Instagram. With respect to that, if the target user does not have an account on these social networks, then they cannot be monitored with the proposed method, although aspects of it can be applied in a general sense.

Future work

To further improve geolocation, image processing can be performed on image content collected from social network accounts. Images may add a real added-value, by face/landscape recognition, or just by searching for equivalent images; the description of the image provided by google image or in the “alt” attribute of “img” tags from social networks webpages may also provide feedback on the target user and his friends. Images may also help to retrieve other accounts of the same person: if the same image appears as a Twitter and as a Flickr avatar, there is a high probability the two accounts belong to the same person.

Our prototype toolkit is capable of displaying the successive locations of a user on a map, but for now, it is not capable of displaying multiple users’ tracks.. In this work, we concentrated on extracting the successive locations of a user, but with further analysis, it is also possible to establish a psychological profile of a user. A psychological profile may be constructed according to the words a user uses (register of sadness, poetry citations…), the distribution of places a user goes to (and if this distribution brutally changes, it could mark important changes in his life). The method for creating a psychological profile may be found in Tang et al., 2014. The possibility of psychological profiling through the social platform for acquiring political affiliation was also shown in Kandias et al., 2013.

Conclusion

Our work aimed to primarily profile the movement and to locate social platform users by processing their posts, thus raising the privacy issues involved. Even if a user does not disclose his or her location, it becomes possible to predict its future location using social network data. A social graph method was developed, aiming to process personal data in an automated way and subsequently localize a user. Results were favourable, suggesting an estimated accuracy of 77,72% highlighting that users can be geolocated, through OSINT and machine learning without GPS/GNSS related information.

The very nature of social media means that we cannot fully protect end users from such tracking actions since such actions are based on the analysis of content shared by the users themselves. The use of common sense, the separation and the preservation of sensitive personal information, and the raising of awareness on security issues amongst the user community can help to significantly reduce the personal exposure.

References

[1] Ahmed, A., Hong, L., & Smola, A. J. (2013,). Hierarchical geographical modelling of user locations from social media posts. *In Proceedings of the 22nd international conference on World Wide We*b (pp. 25-36). ACM.

[2] Backstrom, L., Sun, E., and 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.

[3] Bo, H., Cook, P., and Baldwin, T. (2012). Geolocation prediction in social media data by finding location indicative words. *In Proceedings of COLING* (pp. 1045-1062).

[4] Brandwatch. (2018). 116 Amazing Social Media Statistics and Facts. [online] Available at: https://www.brandwatch.com/blog/96-amazing-social-media-statistics-and-facts/ [Accessed 25 Jul. 2018].

[5] Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. *In Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 759-768). ACM.

[6] Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1082-1090). ACM.

[7] Cho, H. R., and Choi, M. (2014). Replacing Socket Communication by REST Open API for Acquisition Tax Analyzer Development. In *Advanced Information Networking and Applications Workshops (WAINA), 2014 28th International Conference on* (pp. 462-468). IEEE.

[8] Compton, R., Jurgens, D., & Allen, D. (2014, October). Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on* (pp. 393-401). IEEE.

[9] Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., & Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52), 22436-22441.

[10] Cranshaw, J., Toch, E., Hong, J., Kittur, A. and Sadeh, N. (2010). Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing* (pp. 119-128). ACM.

[11] 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. *Transactions in GIS*, *15*(6), 735-751.

[12] Ghahremanlou, L., Sherchan, W., & Thom, J. A. (2015). Geotagging twitter messages in crisis management. *The Computer Journal*, 58(9), 1937-1954.

[13] Giordano, A., Spezzano, G., Sunarsa, H., and Vinci, A. (2015). Twitter to integrate human and Smart Objects by a Web of Things architecture. In *Computer Supported Cooperative Work in Design (CSCWD), 2015 IEEE 19th International Conference on* (pp. 355-361). IEEE.

[14] Gritzalis, D., Kandias, M., Stavrou, V., & Mitrou, L. (2014). History of information: the case of privacy and security in social media. In *Proc. of the History of Information Conference* (pp. 283-310).

[15] Highfield, T., & Leaver, T. (2014). A methodology for mapping Instagram hashtags. *First Monday*, *20*(1).

[16] Instagram (2018). Available from https://www.instagram.com/about/legal/privacy/ [Accessed 1 August 2018]

[17]Jurgens, D. (2013). That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. *ICWSM*, *13*(13), 273-282.

[18]Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., and Ruths, D. (2015). Geolocation prediction in Twitter using social networks: a critical analysis and review of current practice. In *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media* (ICWSM).

[19]Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, *23*(5), 544-559.

[20]Kandias, M., Mitrou, L., Stavrou, V., & Gritzalis, D. (2013, July). Which side are you on? A new Panopticon vs privacy. In *Security and Cryptography (SECRYPT), 2013 International Conference on* (pp. 1-13). IEEE.

[21]Kinsella, S., Murdock, V., and O'Hare, N. (2011). I'm eating a sandwich in Glasgow: modelling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents* (pp. 61-68). ACM.

[22]Kong, L., Liu, Z., & Huang, Y. (2014). Spot: Locating social media users based on social network context. *Proceedings of the VLDB Endowment*, *7*(13), 1681-1684.

[23]Kotzias, D., Lappas, T., & Gunopulos, D. (2016). Home is where your friends are: Utilizing the social graph to locate Twitter users in a city. *Information Systems*, *57*, 77-87.

[24]Li, R., Wang, S., & Chang, K. C. C. (2012). Multiple location profiling for users and relationships from social network and content. *Proceedings of the VLDB Endowment*, *5*(11), 1603-1614.

[25]Li, R., Wang, S., Deng, H., Wang, R., & Chang, K. C. C. (2012, August). Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1023-1031). ACM.

[26]Mansfield, T.O., Ghita, B.V. and Ambroze, M.A., 2017. Signals of opportunity geolocation methods for urban and indoor environments. *Annals of Telecommunications*, *72*(3-4), pp.145-155.

[27]McGee, J., Caverlee, J., & Cheng, Z. (2013). Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 459-468). ACM.

[28]Mitrou, L., Kandias, M., Stavrou, V., & Gritzalis, D. (2014, April). Social media profiling: A Panopticon or Omniopticon tool?. In *Proc. of the 6th Conference of the Surveillance Studies Network*.

[29]Moorosi N. and Marivate V. (2015), Privacy in mining crime data from Social Media: A South African perspective, In *Second International Conference on Information Security and Cyber Forensics (InfoSec)*, Cape Town, 2015, (pp. 171-175).

[30]Narain, S., Vo-Huu, T. D., Block, K., & Noubir, G. (2016, May). Inferring user routes and locations using zero-permission mobile sensors. In *Security and Privacy (SP), 2016 IEEE Symposium on* (pp. 397-413). IEEE.

[31]Neri, F., and Geraci, P. (2009). Mining Textual Data to boost Information Access in OSINT. In *Information Visualisation, 2009 13th International Conference* (pp. 427-432). IEEE.

[32]Perwitasari, A., Akbar, S., & Saptawati, G. P. (2015, November). Software architecture for social media data analytics. In *2015 International Conference on Data and Software Engineering (ICoDSE)* (pp. 208-213). IEEE.

[33]Puzis, R., Yagil, D., Elovici, Y., & Braha, D. (2009). Collaborative attack on Internet users' anonymity. *Internet Research*, *19*(1), 60-77.

[34]Roller, S., Speriosu, M., Rallapalli, S., Wing, B., & Baldridge, J. (2012). Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1500-1510). Association for Computational Linguistics.

[35]Rout, D., Bontcheva, K., Preoţiuc-Pietro, D., & Cohn, T. (2013), Where's@ wally?: A classification approach to geolocating users based on their social ties. In P*roceedings of the 24th ACM Conference on Hypertext and Social Media* (pp. 11-20). ACM.

[36]Sadilek, A., Kautz, H. and Bigham, J.P. (2012), February. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 723-732). ACM.

[37]Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Internet Research*, *23*(5), 528-543.

[38]Seleniumhq.org. (2018). Selenium - Web Browser Automation. [online] Available at: *http://www.seleniumhq.org/* [Accessed 31 Jan. 2018].

[39]Serdyukov, P., Murdock, V., & Van Zwol, R. (2009). Placing Flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 484-491). ACM.

[40]Shaik, A., Borgaonkar, R., Asokan, N., Niemi, V., & Seifert, J. P. (2015). Practical attacks against privacy and availability in 4G/LTE mobile communication systems. *arXiv preprint arXiv:1510.07563*.

[41]Tan, X., Qin, L., Kim, Y., & Hsu, J. (2012). Impact of privacy concern in social networking websites. *Internet Research*, *22*(2), 211-233.

[42]Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1555-1565).

[43]Twitter (2018). Available from https://support.twitter.com/articles/20169886 [Accessed 1 August 2018]

[44]Xie, Y., Chen, Z., Cheng, Y., Zhang, K., Agrawal, A., Liao, W. K., & Choudhary, A. (2013). Detecting and tracking disease outbreaks by mining social media data. *Dimensions*, *17*(16), 16-70.

[45]Yang, Y., Lutes, J., Li, F., Luo, B. and Liu, P., 2012, February. Stalking online: on user privacy in social networks. In *Proceedings of the second ACM conference on Data and Application Security and Privacy* (pp. 37-48). ACM.

[46]Zheng J, Liu S. and M. Ni L., (2014). User characterization from geographic topic analysis in online social media, In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference* on, Beijing, 2014, (pp. 464-471).