Knowledge Discovery from Geo-Located Tweets for Supporting Advanced Big Data Analytics: A Real-Life Experience

Alfredo Cuzzocrea[1](✉), Giuseppe Psaila[2], and Maurizio Toccu[2]

[1] DIA Department, University of and ICAR-CNR, Trieste, Italy
alfredo.cuzzocrea@dia.units.it

[2] Department of Engineering, University of Bergamo, Bergamo, Italy
{psaila,maurizio.toccu}@unibg.it

Abstract. Tourists are an important asset for the economy of the regions they visit. The answer to the question "where do tourists actually go?" could be really useful for public administrators and local governments. In particular, they need to understand what tourists actually visit, where they actually spend nights, and so on and so forth.

In this paper, we introduce an original approach that exploits geo-located messages posted by *Twitter* users through their smartphones when they travel. Tools developed within the *FollowMe* suite track movements of *Twitter* users that post tweets in an airport and reconstruct their trips within an observed area. To illustrate the potentiality of our method, we present a simple case study in which trips are traced on the map (through KML layers shown in *Google Earth*) based on different analysis dimensions.

1 Introduction

Modern smartphones are enabling the concept of *Mobile Social Computing*, i.e. the capability of exploiting computation services to deal with social information (*Social Computing*) enhanced with capabilities of mobile devices (e.g., [1]). In particular GPS localization, provided by most of mobile devices, gives an important contribution: (with respect to social networks) people can post geo-localized messages and pictures, giving this way much more indirect information than non-localized posts (e.g., [7,10]). Among all social networks, *Twitter* (as well as other social networks that adopt the same approach) is particularly attractive for the purpose of searching interesting messages: in fact, every user can see messages by other users without limitations. Nevertheless, observe that geo-localized posts represent a kind of voluntary contribution, because users voluntarily install the (*Twitter*) app and voluntarily post messages (tweets).

These posts' knowledge, hardly acquirable with traditional survey methods, can be very useful for public administrations that like to understand how tourists travel on the region they govern, especially when the region is served by an International Airport. One typical hard question to answer is:

Where do tourists actually go?

L. Bellatreche and Y. Manolopoulos (Eds.): MEDI 2015, LNCS 9344, pp. 285–294, 2015.
DOI: 10.1007/978-3-319-23781-7_23

The intuition is that *Mobile Social Computing* can help to understand where travelers actually go, what they actually visit, where they actually spend nights: in fact, by gathering the geo-localized tweets they post during their travel, it should be possible to reconstruct their trips (e.g., [8,9,11,12]). This aspect plays a critical role, especially when dealing with *Big Data Analytics* (e.g., [2–6]).

The *FollowMe* project originated from these considerations. The aim is to develop techniques and build a suite of tools (the so-called *FollowMe* suite) that query social networks to discover posts sent by travelers and trace them during their trip. At the moment, we developed tools working with *Twitter* and tweets; in the next stages of the project we will consider other social networks. In this paper, we present: the approach we followed, the way reconstructed trips can be analyzed, the *FollowMe* suite from a technical point of view and a case study built with the initial data sets we collected.

The paper is organized as follows. Section 2 deals with problem definition and analysis dimensions. Section 3 reports about architecture of *FollowMe* suite. Section 4 reports the case study. Finally, Sect. 5 draws our conclusions and future work.

2 Problem Definition and Analysis Dimensions

The aim of the project is to build techniques and tools that permit to study the movements of tourists visiting a given region. The choice of *Twitter* is motivated by the fact that messages are short and visible to every user, without limitations.

2.1 Problem Definition

A key point was to find a way to identify traveling users: in fact, it is not feasible to detect them simply asking *Twitter* API to retrieve geo-located post in a given area: who is actually traveling? who resides in the area?

The answer can be found by observing the typical behavior of travelers, depicted in Fig. 1.

While they are waiting for boarding, travelers have time to post tweets, notifying friends that their trip is beginning. After the flight, they transit through the arrival airport (represented by the cue ball reached by the dashed arrow), but here they do not post tweets; in particular, this happens in small airports for passengers with hand-baggage only. Instead, they usually post tweets when they are visiting some wonderful place/tourist attraction or in the hotel (represented by the other cue balls). This gives us the solution to the above mentioned problem. It is necessary to find travelers that (potentially) reached the region of interest by retrieving tweets posted in the departing airport connected with airports close to the region of interest.

Tweet Gathering. The *Gathering Problem* can then be stated as follows.

Problem 1. Given one or more regions of interest R, identify the airports A_R that serve R. Then, identify the airports A_O which flight having destination in A_R originate from.

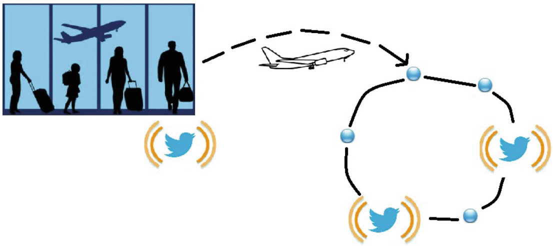


Fig. 1. Movements and tracking of passengers by *Twitter*.

Every day d, collect the set $H(d, A_O)$ of *hang tweets*, i.e., geo-loacetd tweets posted in the area of an airport in A_O; the union H of the daily gathered tweets is $H = \bigcup_d H(d, A_O)$.

Them for each user u having a tweet $h \in H$, gather all geo-located tweets posted by user u (i.e., his/her *timeline*) in the next 8 days after $h.date$, i.e., the date of the hang tweet h, denoted as $T(u, t.date)$. The set of overall collected timelines is $\mathcal{T} = \bigcup_{u,d} T(u, d)$. □

Trip Querying. Once hang tweets and timelines are collected (Problem 1), it is necessary to extract trips from $\mathcal{T}$, as far as they touch the region of interest. This is the *Trip Querying* problem and is stated as follows.

Problem 2. Consider the set $\mathcal{T} = \bigcup_{u,d} T(u, d)$ of gathered timelines. Given a region of interest $\overline{R}$, a query q is the pair $q = (\mathcal{T}, \overline{R})$.

The *Trip Result Set* $R_q = \{\overline{T}(u, d)\}$ such that for each $T(u, d) \in \mathcal{T}$ for which $\overline{T}(u, d) \subseteq T(u, d)$ and each $t \in \overline{T}(u, d)$ is geo-located within $\overline{R}$. □

In order to address the two above stated problems, we developed tools described in Sect. 3.

2.2 Analysis Dimensions

What kind of analysis can be performed on trips? Certainly a graphical representation on the Earth map is straightforward, but the way trips are represented is not obvious.

We identified several *Analysis Dimensions*.

- *Path.* For each user, the analysis of the path followed during the trip could reveal unexpected knowledge. For example, discovering that a tourist attraction is often visited after the visit to a museum, could suggest local governments to better organize public transportation services.
- *Origin Airports.* The origin airport of trips could let administrators to understand for which countries the governed region is more attractive. This could lead to marketing actions to consolidate the attraction factors, or to understand how to become more attractive for other countries.

- *Time Slots.* Depending on the daylight time, travelers do different activities. In particular, in the morning or in the afternoon they go around visiting places; in the evening usually they look for a restaurant where to have dinner; in the night they probably are in their hotel room. Thus, tweets could be grouped and analyzed based on precise daylight time slots, to discover, e.g., where they mostly spend nights.
- *Week Days.* Another important dimension concerning time is the week day. In fact, it is likely that the specific week day can influence the places visited by tourists. For example, this could suggest to open a Museum on Sundays.

Tweet Alignment. In order to make effective path analysis and, in general, to enable intermediate aggregations, each tweet in a trip is aligned based on the distance between its date and the date of the beginning tweet of the trip. Tweet Alignment is performed by computing the *Tweet Trip Day* $t.td$ as $t.td = (t.date - h.date) + 1$, where h is the hang tweet of the trip (the tweet posted in the origin airport).

Daylight Time Partitioning. In order to enable the dimension analysis based on daylight time slots, each tweet is extended with the proper time slots. We decided to adopt the following mapping:

1. $TS1$: 22:00am – 05:59am, Night;
2. $TS2$: 06:00am – 11:59am, Morning;
3. $TS3$: 12:00pm – 17:59pm, Afternoon;
4. $TS4$: 18:00pm – 21:59pm, Evening.

In particular, $TS1$ can provide information about where travelers sleep. Instead, likely, $TS4$ can provide information about where travelers have dinner. Finally, $TS2$ and $TS3$ can provide information about the activities of our travelers within the region of interest.

3 The *FollowMe* Suite

The *FollowMe* suite is an open pool of tools, each one devoted to a specific task. In fact, at the current stage of development of the project, we only gather messages from *Twitter*, but the long term goal of the project is to collect posts coming from various social networks. Consequently, new components must be easily added and the data storage service must flexibly deal with semi-structured and text-based documents. Hereafter, we describe in more details the software tools currently in the suite, which are depicted in Fig. 2.

- *MongoDB.* The storage service is responsibility of *MongoDB*, a recent and very famous No-SQL DBMS. It is designed to deal with collections of documents, where each document is represented as a JSON object. The main advantage in using MongoDB is the ability to manage documents with different structures within the same collections, this way overtaking the concept of schema in tables, that obstacle the adoption of traditional relational technology where documents with variable structures must be stored.

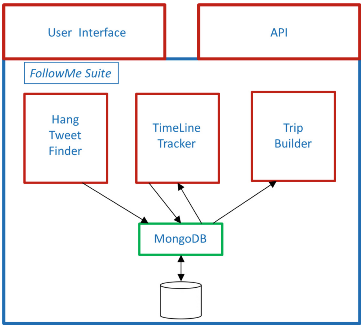


Fig. 2. Architecture of the *FollowMe* suite.

- *Hang Tweet Finder*. This component is responsible to query the *Twitter* API to look for tweets posted in the area of the monitored airports. In fact, *Twitter* API provides the capability to search for geo-located tweets, given the coordinates of the center and the radius of an area of interest.
  These tweets are called *hang tweets*, because they are the hang to identify users to follow in their trip.
- *Timeline Tracker*. For each user identified by means of hang tweets, the *Timeline Tracker* follows his/her timeline, i.e., the history of tweets posted by the user. In particular, the *Timeline Tracker* considers only geo-localized tweets posted in the next 8 days after the date of the hang tweet.
- *Trip Builder*. While the *Hang Tweet Finder* and the *Timeline Tracker* collect potentially interesting tweets from *Twitter*, the *Trip Builder* actually reconstruct trips by querying the storage area. In particular, the *Trip Builder* is launched by specifying the bounding box of the geographical area in which we want to discover trips.
- Finally, the *FollowMe* suite is completed by a *user interface*, that allows analysts and administrators to manage the gathering process and start queries. Furthermore, services provided by tools within the suite can be exploited by applications through suitable APIs.

Output Data Formats. The *Trip Builder* represents trips as sequences of tweets. In order to allow an easy exploitation by external tools, such as MatLab, Excel, etc., trips are generated as CSV (comma separated value) files. For each tweet the user identifier, the data and time, the latitude and longitude are reported; furthermore, for each identified trip the origin airport (i.e., the airport where the hang tweet was posted) the date of the last tweet in the monitored area and the duration of the stay in the same area are reported, among all. Table 1 describes above attributes.

Geographical Layer. When geographical data are concerned, visualization on a map is an important issue. This is even more important in our project, where analysts need to understand where travelers mostly spend their time in the area. For this reason, we also generate several KML representations of the trips.

Table 1. Attributes for each tweet

Attributes	Description
SNet	Identifier of each social network
TweetId	Identifier of each tweet
UserName	Identifier of each user (traveler)
Date	Date of tweet publication
Time	Time of publication of tweet
Latitude	Latitude at which the tweet was posted
Longitude	Longitude at which the tweet was posted
OriginAirport	Departure airports for each user (traveler)

KML is the input format accepted by *Google Earth* and by Google Maps API; in particular, in Google Maps API-based web applications information layers described as KML files can be added to maps. However, for analysis tasks, *Google Earth* is a very powerful tool, because it permits to select information item to show. In particular, KML files can contain (possibly nested) folders, that can be very useful to partition information items based on a specific property. For example, an analyst could interested in partition trips based on the airport where trips originated from. Since the analysis needs could be manifold, several KML files are generated. They are reported in the following list.

- Locations partitioned by origin airport.
- Locations partitioned by time slot.
- Trips depicted as polylines and partitioned by users.

In this way, the analyst can view the trips by several perspectives, and better understand the dynamics of trips.

4 Case Study: The EXPO 2015 in Milan

In order to illustrate the effectiveness of our approach, we built a simple case study on the basis of a small set of geo-located tweets gathered by the *FollowMe* suite. The goal of the case study is to discover travelers coming to Lombardy, the region in the center of northern Italy where the main city is Milan, that in these days is world wide famous due to EXPO 2015. Therefore, we identified a pool of 30 European airports; they were chosen based on the presence of flights to airports in Lombardy and such that the number of posted tweets in a single day is not huge (Madrid and Lisbon were discarded because more than 1500 tweets a day were posted in the area of the two airports).

We collected hang tweets and timelines in the period between April 20, 2015, and May 11, 2015. By performing a query to discover trips in the bounding box of Lombardy, the *Trip Builder* generated a result set of 50 trips, formed by a total of 168 trips.

Table 2. Number of tweets and travelers for origin airports

Origin airport	Trips	Posted tweets
Athens	8	32
Barcelona	16	65
Beauvais	1	2
Berlin	3	5
Bucharest	1	1
Charleroi	2	8
Copenhagen	7	27
Frankfurt	1	6
Munich	3	6
Stansted	7	14
Stansted	1	2

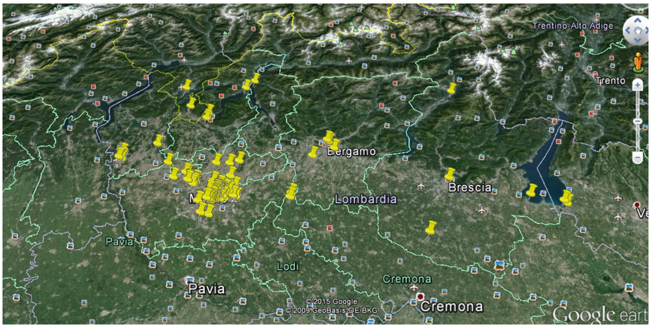


Fig. 3. Tweets distribution on the Lombardy region.

Discovered trips originated in 11 different airports, reported in. In Table 2 we show the 11 Origin Airports. Besides each airport name, we report number of trips that originated from that airport (column *Trips*), as well as the total number of tweets that constitute those trips (column *Posted Tweets*). For example, the 8 identified trips originated in Athens are composed of 32 tweets. It is possible to notice that Spanish travelers use to post a more tweets than travelers coming from other countries: the system detected 16 trips from Barcelona, that is, more than the sum of trips from Athens and Copenhagen.

The KML layers describing the discovered trips, were analyzed by means of *Google Earth*. Figure 3 shows the distribution of tweets that travelers posted within the Lombardy region. It is possible to note that these tweets are mainly concentrated in Milan area. The presence of travelers in this area are likely conditioned by EXPO 2015.

Figure 4, that represents the dimension *Origin Airports*, shows the distribution of tweets with respect to travelers coming from Barcelona. It is possible to

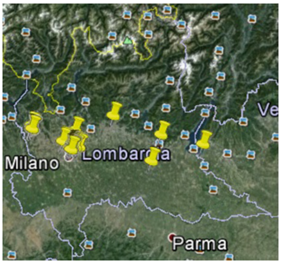
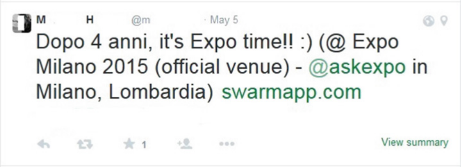


Fig. 4. Tweets distribution with respect to travelers come from Barcelona.



M H @m · May 5

Dopo 4 anni, it's Expo time!! :) (@ Expo Milano 2015 (official venue) - @askexpo in Milano, Lombardia) swarmapp.com

1

View summary

Fig. 5. Tweet of a Spanish traveler.

note that Spanish travelers concentrate their tweets mainly in the area of Milan and beyond.

Moreover, some travelers that posted these tweets arrived in Lombardy after EXPO started. For example, in Fig. 5 we report a tweet posted by a Spanish traveler in which the writer talks about EXPO 2015.

Figure 6, that represents the *Path* dimension, shows the full trip of the same Spanish traveler, that is, his/her route in Lombardy region. It is possible to note two interesting things. The first one is that the traveler posted his/her tweets mainly in the city of Milan. The second one is pushpin 5, that represents the post reported in Fig. 5; the pushpin shows that the traveler was actually in EXPO 2015 area.

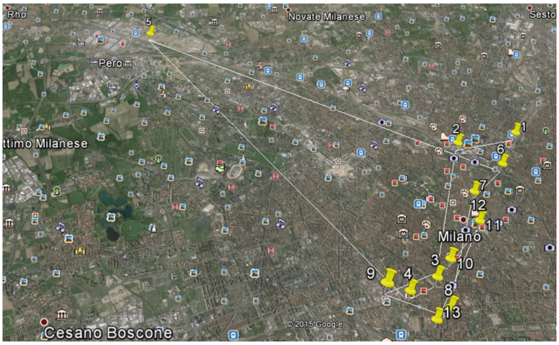
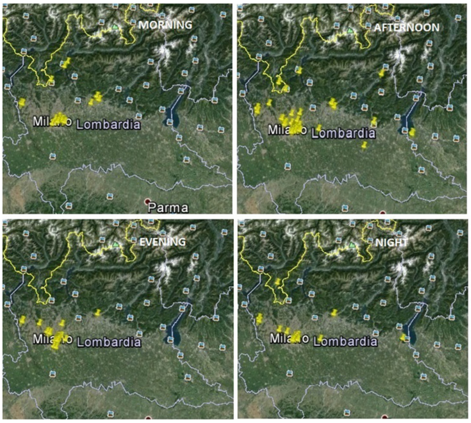


Fig. 6. Path of one traveler.



Time Slot	Tweets Number
Afternoon	66
Evening	43
Morning	32
Night	27

Fig. 7. Tweets distribution in time slots.

Finally, Fig. 7 represents the distribution of tweets with respect to the four Daylight Time Slots defined in Sect. 2 and, moreover, represents the *Time Slots* dimension. It is possible to note how the distribution of tweets is geographically more sparse in the Afternoon than in the others time slots, where the tweets are concentrated in the Milan area. There are many reasons that explain this behavior, but one possible cause of this is that the travelers have their base in city of Milan and prefer to visit the Lombardy area after lunch.

5 Conclusions and Future Work

As stated at the beginning, tourists are an important asset for the economy of the regions they visit. In particular, for public administrations is very useful to understand how tourists travel on the region they govern.

Therefore, in this work, we developed an original approach that permits to follow traveling *Twitter* users by tracking their geo-located messages they post on *Twitter* during their trips. Tools in the *FollowMe* suite generate several outputs for the result set of reconstructed trips, so that several analysis dimensions (*Time Slot*, *Origin Airport*, *Path*) can be exploited to analyze results.

As far as future work is concerned, we have to consider that the project is only at the beginning steps. The main efforts will be devoted to connect with other social networks and gather posts from them.This way, we should obtain a wider spectrum of information, by integrating several sources of information. For this purpose, the main problem is that users use different ids on different social networks, so the hardest, yet exciting challenge, will be to find techniques to recognize different ids belonging to the same user.

References

1. Bora, N., Chang, Y.-H., Maheswaran, R.: Mobility patterns and user dynamics in racially segregated geographies of US cities. In: Kennedy, W.G., Agarwal, N., Yang, S.J. (eds.) SBP 2014. LNCS, vol. 8393, pp. 11–18. Springer, Heidelberg (2014)
2. Cuzzocrea, A.: Analytics over big data: Exploring the convergence of datawarehousing, OLAP and data-intensive cloud infrastructures. In: 37th Annual IEEE Computer Software and Applications Conference, COMPSAC 2013, Kyoto, Japan, 22–26 July 2013, pp. 481–483 (2013)
3. Cuzzocrea, A.: Big data mining or turning data mining into predictive analytics from large-scale 3Vs data: the future challenge for knowledge discovery. In: Ait Ameur, Y., Bellatreche, L., Papadopoulos, G.A. (eds.) MEDI 2014. LNCS, vol. 8748, pp. 4–8. Springer, Heidelberg (2014)
4. Cuzzocrea, A., Bellatreche, L., Song, I.-Y.: Data warehousing and OLAP over big data: current challenges and future research directions. In: Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP, DOLAP 2013, San Francisco, CA, USA, 28 October 2013, pp. 67–70 (2013)
5. Cuzzocrea, A., Saccà, D., Ullman, J.D.: Big data: a research agenda. In: 17th International Database Engineering & Applications Symposium, IDEAS 2013, Barcelona, Spain, 09–11 October 2013, pp. 198–203 (2013)
6. Cuzzocrea, A., Song, I.-Y.: Big graph analytics: The state of the art and future research agenda. In: Proceedings of the 17th International Workshop on Data Warehousing and OLAP, DOLAP 2014, Shanghai, China, 3–7 November 2014, pp. 99–101 (2014)
7. Grabovitch, I., Kanza, Y., Kravi, E., Pat, B.: On the correlation between textual content and geospatial locations in microblogs. In: GeoRich 2014, Snowbird, Utah (USA), 23 June 2014, June 2014
8. Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C.: Geo-located twitter as proxy for global mobility patterns. Cartography Geogr. Inf. Sci. **41**(1), 260–271 (2014)
9. Lee, R., Sumiya, K.: Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In: ACM LBSN 2010, San Jose, CA, (USA), November 2010
10. Stephens, M., Poorthuis, A.: Follow thy neighbor: connecting the social and the spatial networks on Twitter. Comput. Environ. Urban Syst. **41**(1) (2014). doi:10.1016/j.compenvurbsys.2014.07.002
11. Walther, M., Kaisser, M.: Geo-spatial event detection in the twitter stream. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) ECIR 2013. LNCS, vol. 7814, pp. 356–367. Springer, Heidelberg (2013)
12. Widener, M., Li, W.: Using geolocated twitter data to monitor the prevalence of healthy and unhealthy food references across the us. Appl. Geogr. **54**, 189–197 (2014)