An Innovative Framework for Effectively and Efficiently Supporting Big Data Analytics over Geo-Located Mobile Social Media

Alfredo Cuzzocrea
DIA Department, University of Trieste and ICAR-CNR
Italy
alfredo.cuzzocrea@dia.units.it

Giuseppe Psaila
DIGIP Department, University of Bergamo
Italy
psaila@unibg.it

Maurizio Toccu
DIGIP Department, University of Bergamo
Italy
maurizio.toccu@unibg.it

ABSTRACT

Mobile Social Media are gaining momentum in the broader context of *Big Data Analytics*, where the main issue is represented by the problem of extracting interesting and actionable knowledge from big data repositories. Mobile social media sources like *Twitter* and *Instagram* are indeed producing massive amounts of data (namely, *posts*) that represent a very rich source of knowledge for predictive analytics. In line with this emerging trend, this paper proposes *an innovative approach for effectively and efficiently supporting big data analytics over geo-localized mobile social media*, with particular emphasis with the context of *modern tourist information systems*. In this context, the innovative *FollowMe* suite, which implements the proposed methodology, is also described in details. We complement our analytical contribution with a real-life case study focusing on the EXPO 2015 event in Milan, Italy which clearly shows benefits and potentialities of our proposed big data analytics framework.

CCS Concepts

•Information systems → Information retrieval query processing;

Keywords

Big Data Analytics, Mobile Social Media, Big Data Frameworks

1. INTRODUCTION

Modern smartphones are enabling the concept of *Mobile Social Computing* (e.g., [19, 24, 14, 2]), a set of methodologies and related computational infrastructures and services for dealing with social information (referred as *Social Computing*), enhanced with capabilities of *mobile devices* (e.g., [5]), thus producing so-called *mobile social media* (e.g., [3, 21, 18]). In particular, GPS localization, provided by popular mobile devices, provides us with the following, critical contribution: with respect to the context of *social networks*, people can post *geo-localized messages and pictures*, giving this much more indirect information than non-localized posts (e.g., [16, 23]). Among the wide family of social networks adopting this approach, *Twitter* and *Instagram* are particularly attractive for the purpose of searching interesting messages and discovering useufl knowledge for analytics (e.g., [22, 28]). In fact, every user can see messages by other users without limitations. Nevertheless, observe that geo-localized posts represent a kind of voluntary contribution (e.g., [29]), because users voluntarily install the social network app, and voluntarily send posts.

Knowledge kept in such posts, hardly acquirable by means of traditional survey methods, can be very useful for public administrations, large-scale organizations, e-tourism companies and so forth, which aim at understanding how tourists travel on the region of reference, especially when the region is served by an international airport. In this specific application scenario, one typical, hard question to be answered is the following one:

where do tourists actually go?

The intuition beyond our research is that mobile social computing can indeed help to understand where travelers actually go, what they actually visit, and where they actually spend nights. Indeed, by gathering geo-localized posts they send during their travel, it should be possible to reconstruct their trips (e.g., [17, 20, 25, 27]). This aspect plays a critical role, especially when dealing with *Big Data Analytics* (e.g., [12, 10, 8, 9, 13, 4]), as powerful analytics can be built on top of such *big multimedia data*. Also, related *user adaptation issues* (e.g., [7, 6]) can be studied under this general framework.

Starting from these considerations, we proposed the *FollowMe* project, a real-life national project whose aim is that of developing innovative techniques for querying social networks in order to discover posts sent by travelers and trace them during their trips. These techniques are implemented within a software suite encompassing several tools, called *FollowMe suite*. At the actual stage of the project, we developed tools that work with *Twitter* and *Instagram*. In the next stages of the project, we will consider other social networks.

In this paper, we introduce and experimentally assess the *FollowMe* approach, as well as we provide a detailed description of the *FollowMe suite*. In addition to this, we provide a real-life case study focusing on the well-known *EXPO* 2015 *event in Milan, Italy* [1] which clearly shows benefits and potentialities of our proposed big data analytics framework.

The paper is organized as follows. Section 2 reports on the initiative which has originated the *FollowMe* project. Section 3 deals with the formal problem definition. Section 4 contains architec-

IDEAS '16, July 11-13, 2016, Montreal, QC, Canada

DOI: http://dx.doi.org/10.1145/2938503.2938517

ture and functionalities of the *FollowMe suite*. Section 5 provides the so-called *analysis dimensions*, which are the parameters used in the specific big data analytics provided by our approach. Section 6 reports on the case study focusing on the EXPO 2015 event. Section 7 contains a literature review of related works. Finally, Section 8 draws our conclusions and future work. A preliminary version of this paper appears in the short paper [11], where we introduced the basic idea. In this paper, we significantly extend models, algortihms and framework description, along with a comprehensive experimental evaluation and analysis.

2. THE INITIATIVE: RATIONALE OF THE *FOLLOWME* PROJECT

The airport located in Bergamo (Northern Italy) has been incredibly growing since a well-known low cost company chose it as its Italian hub. In a few year, it has become fourth Italian airport (in January 2014); in fact, the airport serves a range of 100 km in with 9 Million inhabitants and $800,000$ enterprises which produce the 21% of Italian GDP. The consequence on the territory in general and on Bergamo in particular is an augmented international visibility and an increment of foreign tourists that visit the town and its medieval center.

This new context should be an opportunity to exploit, in order to further increase the international visibility of Bergamo and its hinterland trying to address the interests of new and different types of tourism. With this goal, the University of Bergamo launched the initiative "Centrality of territories" proposes the regeneration of Bergamo's territory by the activation of a new type of tourism called s-Low (it combines low cost air mobility with sustainable use of the territory). In fact, combining air connectivity, cultural and natural resources and local artisan-ship it is possible to recover the centrality of the territories. Then, this project represents a new business model, in particular an innovative socio-economic system, based on micro entrepreneurship, environmental sustainability and participation of citizens. The project involves a cluster of 7 European towns that have similar characteristics and are connected each other by low cost flights.

Within this initiative, it is important to study where people that arrives at Bergamo airport from towns in the cluster actually spend their time in Italy. But it is not possible to perform a survey through interviews. Therefore, we started the *FollowMe* project.

The goal of the project is simple and ambitious: to apply the concept of Mobility and Social Computing to understand what is happening in the hinterland of Bergamo, with respect to passengers, tourists or not, that arrives at the airport of Bergamo. Social networks are a valuable source of information, because travelers feel the need to share geo-located posts about the places they are visiting, more or less in real time These posts could help to give answers to questions like “who goes, who is, who remains”. For example, how many tourists arrived at Bergamo airport visit the city of Bergamo? How many go to Lecco or Milan? What is the percentage of tourists that remain in the Province? What nationality are they? What do they say?

3. ADVANCED BIG DATA ANALYTICS OVER GEO-LOCATED MOBILE SOCIAL MEDIA: FOUNDATIONS

The aim of the *FollowMe* project is to build techniques and tools enabling the tracking of movements of tourists visiting a given region, in order to build suitable big data analytics over them. In this Section, we first provide the basic definitions that are required to formalize the main problem addressed by *FollowMe*, which is then introduced.

3.1 Basic Definitions

Basic definitions of our formal framework include: *Post*, *Hang Post*, *Tracked Post*, *Trip*. We introduce them in the following.

Definition 1: **Post** A *post* t is a tuple of kind:

$$t :<\textit{id}, \textit{userid}, \textit{date}, \textit{time}, \textit{text}, \textit{lat}, \textit{lon}>$$

where: (*i*) *id* uniquely identifies the post; (*ii*) *userid* is the identifier of the user that sent the post; (*iii*) *date* is the date the post was sent; (*iv*) *time* is the time the post was sent; (*v*) *text* is the text in the post; (*vi*) *lat* (*lon*, respectively) is the latitude (longitude, respectively) of the position where the user sent the post through a mobile device with geo-localization capabilities. □

Definition 2: **Hang Post** A *hang post* ht is a geo-localized post sent in an airport area. A hang post is obtained via submitting a spatial query to the social network (i.e., *Twitter* or *Instagram*), and it allows to discover *userid* of traveling users. Formally, with respect to a simple post, a *hang post* ht has an extra field named *origin*, which is the name of the airport area where the (hang) post was posted. □

Definition 3: **Tracked Post** A *Tracked Post* tt is a geo-localized post sent by a traveling user, provided that the user sent a hang post ht no more than 9 days before tt is posted. □

The above definitions are motivated by the following considerations. In order to discover travelers and retrieve enough information about their provenance, posts must be looked-for in the airport where trips of these travelers originate from; this is the role of *hang posts*, specifically. This way, it is possible to know *userid* and track their trips by querying their *timeline*, i.e. the reverted list of posted posts returned by *Twitter* or *Instagram* for a given user. Therefore, tracked posts are extracted from timelines of users discovered by means of hang posts. The temporal limit of 9 days is reasonable to consider travelers as tourists through airports.

Figure 1 provides an intuitive view of the main idea behind FollowMe. In this case, the goal is to analyze travelers visiting Lombardy, the Italian region with Milan (venue of the Expo 2015 event). Here, it is possible to find hang posts in some critical origin airports, such as Barcelona, London Stansted and Munich, and then follow those travelers that later send posts while they were in Lombardy.

Definition 4: **Trip** A *Trip* is a tuple of kind:

$$\textit{trip} := <\textit{userid}, \textit{date}, \textit{origin}, \textit{tSeq}: (tt_1, \ldots, tt_n)>$$

where: (*i*) *userid* is the identifier of the user that posted the tracked posts composing the trip; (*ii*) *date* is the date the track post was sent; (*iii*) *origin* is the airport where the trip originated from; (*iv*) *tSeq* is the actual sequence of *tracked posts* that compose the trips (posts are sorted by date and time). □

3.2 Problem Formulation

The *FollowMe* project addresses three main problems, namely *Hang Post Discovery*, *Tracked Post Gathering* and *Trip Extraction*, which are formally introduced next.

Problem 1: Hang Post Discovery Consider a set of *Origin Airports* OA, where each airport is described by a tuple $a :<\textit{name}, \textit{lat}, \textit{lon}, \textit{radius}>$, such that fields denote, respectively, the airport name, the latitude and longitude of the center of the airport and the radius of the search area in kilometers, and an *observation period* $op = [d_b, d_e]$, where d_b and d_e denote, respectively, the beginning and end date of the period. The *Hang Post Discovery* problem can be defined as follows: discover all hang posts ht posted in one of the airports in OA within the observation period *op*, and build the persistent set $HT = \{ht\}$ of discovered hang posts. □

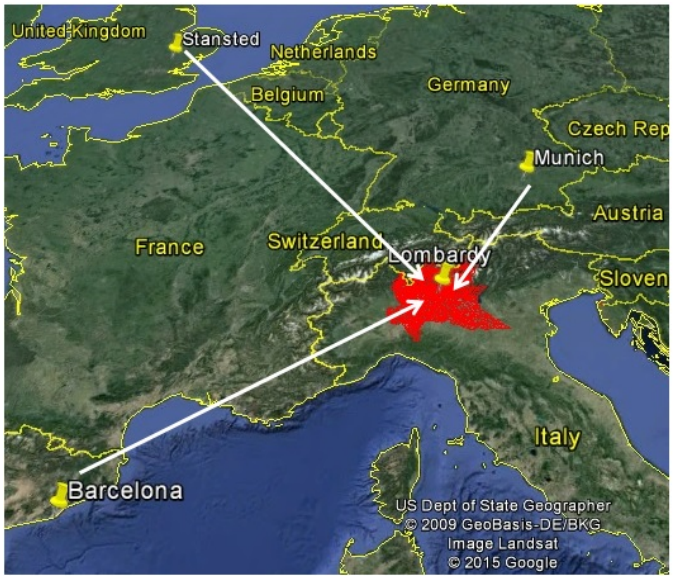


Figure 1: Movements and tracking of passengers from Lombardy by *Twitter* or *Instagram*.

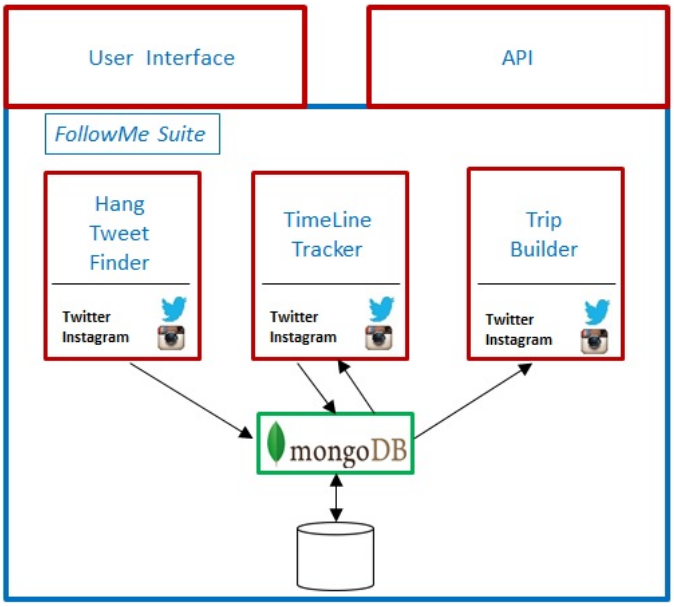


Figure 2: Architecture of the *FollowMe* Suite.

Problem 2: Tracked Post Gathering For each hang post $ht \in HT$, consider $ht.date$ and $ht.userid$. The *Tracked Post Gathering* problem can be defined as follows: gather all tracked posts tt from *Twitter* or *Instagram* such that there exists a hang post $ht \in HT$ for which the following property holds: $tt.userid$=$ht.userid$ and $tt.date \in$[ht.*userid*, $ht.userid$ + 8days]; build the persistent set $TT = \{tt\}$ of gathered tracked posts. □

The persistent sets HT (composed by hang posts) and TT (composed by tracked posts) makes it possible to define (and solve) the main problem, i.e. reconstructing trips of travelers.

Problem 3: Trip Extraction Consider a *Bounding Box* BB, defined as a tuple $BB =< lu_{lat}, lu_{lon}, rb_{lat}, rb_{lon} >$, where the fields model, respectively, latitude and longitude of the left upper corner, and latitude and longitude of the right bottom corner. The *Trip Extraction* problem can be defined as follows: given a bounding box BB and the sets HT and TT of hang posts and tracked posts, respectively, extract all trips such that there exists a hang post $ht \in HT$ for which there exist a set of tracked posts, denoted as *tSeq*: $(tt_1, \ldots, tt_n)$, in TT that are posted by the *same* user such that ($tt_i.lat$, $tt_i.lon$)$\in$ BB and $tt.date \in$[ht.*userid*, $ht.userid$ + 8days]. □

In order to address the three above-stated problems, the *FollowMe* project introduces the *FollowMe suite*, described in Section 4.

4. THE *FOLLOWME* SUITE

The *FollowMe* suite is an open pool of tools, each one devoted to a specific task. These solutions allow to easily add new components to existing ones. in fact, we very recently added component to get posts from *Instagram*. Since they are in the early stage of running, at the time the paper is written we are not able to report about them. Therefore, for these reasons and for the sake of clarity, we discuss only about *Twitter* posts (but the concepts can be straightforwardly extended to *Instagram* posts).

Hereafter, we describe in more details the software tools that are currently embedded in the suite, whose architecture is depicted in Figure 2.

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

- *Hang Post Finder*. This component is responsible of querying the *Twitter* API (and *Instagram* API) in order to discover *Hang Posts* (and solve Problem 1 – see Section 3). In fact, *Twitter* API (and *Instagram* API) provides the capability of searching for geo-located posts, given the coordinates of the center and the radius of an area of interest. Due to scalability reasons and also because of posts disappear from result sets returned by the API as they become too old, the *Hang Post Finder* is launched every day to find hang posts posted during the day before.

- *Timeline Tracker*. This component addresses Problem 2 (see Section 3), via gathering *Tracked Posts*. For each user identified by means of collected hang posts, the *Timeline Tracker* follows his/her timeline, i.e. the history of posts posted by the user, by querying *Twitter* API (and *Instagram* API), by taking only geo-located posts that comply with Definition 3. Due to scalability reasons, the *Timeline Tracker* is launched every day as well, following timelines one day backwards only, and only hang posts no older than 9 days are considered for user identification.

- *Trip Builder*. Problem 3 (see Section 3) is solved by the component named *Trip Builder*: given a bounding box BB, *Trip Builder* extracts all sequences of tracked posts that are posted in the bounded area by a user with a hang post. Section 4.1 illustrates in details how *Trip Builder* works.

- *User Interface*. A suitable user interface that allows analysts and administrators to manage the gathering process and run queries.

- *API*. Services provided by the *FollowMe* suite tools are exposed via to applications through suitable APIs.

4.1 Querying Trips

The *Trip Builder* component implements algorithm **Algorithm 1** for querying trips. Algorithm Algorithm **Algorithm 1** is divided into two main components: the main procedure, named *ExtractTrips*, and the supporting function, named *ExtractSeq*.

ExtractTrips is responsible for activating the process and extracting hang posts ht from within the (persistent) set HT. Focus the attention on algorithm Algorithm **Algorithm 1**. Here, Line 2. sorts hang posts by *date*, *userid* and *time*, so that it is possible to take only one hang post among all those posted by the same user in the same date (conditional instruction at Line 5.). For each hang post (for loop at Line 3.), function *ExtractSeq* is called, to get the sequence *tSeq* of tracked posts for that hang post in the area bounded by BB (Line 6.): if the sequence is not empty (Line 7.), the new trip descriptor is generated (Line 8.) and added to the result set (Line 9.).

ExtractSeq is responsible to look for tracked posts, provided the persistent set TT, a hand post ht and the bounding box BB. Every tracked post is checked to be eligible, i.e. posted by the same user that posted ht and no later than 9 days after the hang post (Line 12.). If the tracked post is eligible, it is added to the sequence *utSeq* (Line 13.), which models the unordered result sequence. Finally, at the end of the loop, if the sequence *utSeq* is not empty, it is sorted (Line 14.) and the final result sequence is returned (Line 15.). Function *ExtractSeq* does not actually run several times on the persistent TT sets, but rather it exploits querying and index capabilities provided by *MongoDB* to optimize the whole process.

Algorithm 1.

```
Procedure ExtractTrips
Input:HT: set of hang Posts
      TT: set of Tracked Posts
      BB =< lu_lat, lu_lon, rb_lat, rb_lon >: bounding box
Output:Trips: set of trip instances
begin
1.  Trips := ∅;
2.  SHT := sort HT by date,userid, time;
3.  for i := 0 to |SHT| do
    begin
4.    ht := SHT[i];
5.    if i = 1 or (i > 1 and SHT[i − 1].userid ≠ ht.userid) then
      begin
6.      tSeq := ExtractSeq(TT, BB, ht);
7.      if tSeq ≠ ∅ then
        begin
8.        trip := < ht.userid, ht.ate, ht.originAirport, tSeq >;
9.        Trips := Trips ∪ { trip };
        end if
      end if
    end for
  end Procedure

  Function ExtractSeq(TT, BB, ht, tSeq)
  begin
10. utSeq := ();
11. for each t ∈ TT do
    begin
12.   if t.userid = ht.userid and (t.lt, t.lon) in BB and
          t.date in [ht.date, ht.date + 8days] then
      begin
13.     utSeq := utSeq • (t);
      end if
    end for each
14. tSeq := sort utSeq by time;
15. return tSeq;
  end Function
```

4.2 Output Data Formats

The *Trip Builder* produces a file in the CSV format, which has to amenity of being processable by external tools, such as MatLab, Excel etc., in order to inspect and analyze trips. However, it is important at the same to visualize trips on maps, in order to permit visual analysis of discovered trips. For this reason, the *FollowMe* suite provides converters to several *KML representations* of the trips. KML is the input format accepted by *Google Earth* and by *Google Maps API*, hence achieving a full integration with *FollowMe*.

In particular, for analysis tasks, *Google Earth* turns to be a very powerful tool, as it permits to select information on items to be shown on the maps. In particular, KML files can contain (possibly nested) folders, which can be very useful to partition information items based on a specific property. For instance, an analyst could be interested in partitioning trips based on the origin airport.

In order to cope with a large variety of analysis tasks, several KML files are generated. A more complete discussion about that is provided in Section 5.

5. ANALYSIS DIMENSIONS

What kind of analysis can be performed on trips? This Section answers to that question, and the derived models are then exploited in the case study analysis provided in Section 6. Indeed, here we introduce suitable *analysis dimensions*, which are the basis for the big data analytics process supported by the *FollowMe* suite.

From a practical point of view, a graphical representation on the *Google Earth* map is straightforward, but the way trips are represented is not obvious, hence we introduce the following analysis dimensions.

- *Path.* For each user, the analysis of the path followed during the trip could reveal unexpected knowledge. For instance, discovering that a tourist attraction is often visited after the visit to a museum could suggest local governments focused to better organize public transportation services.

- *Origin Airports.* The origin airport of trips could let administrators to understand for which countries the target region is more attractive. This could lead to marketing actions for consolidating specific attraction factors, or understanding how to become more attractive for other countries.

- *Time Slots.* Depending on the daylight time, travelers do different activities. In particular, during the morning or the afternoon, they usually go around visiting places. During the evening, they usually look for a restaurant where having dinner. During the night, they probably are in their hotels. Therefore, posts could be grouped and analyzed based on precise daylight time slots, in order to discover, for instance, where they mostly spend nights.

- *Week Days.* Another relevant analysis 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 instance, this could suggest to open a specific museum on Sundays.

5.1 Post Alignment

In order to make effective path analysis and, more generally, to enable intermediate aggregations, each post in a trip is aligned on the basis of the distance between its date and the date of the beginning post of the trip. Post Alignment is performed by computing the *Post Trip Day* $t.td$, as follows:

$$t.td = (t.date - h.date) + 1$$

where h models the hang post of the trip (i.e., the post sent in the origin airport).

5.2 Daylight Time Partitioning

In order to enable the dimension analysis based on daylight time slots, each post is extended with the proper time slots. The following mapping has been adopted to this end.

1. $TS1$: 10 : 00am - 05 : 59am, Night;
2. $TS2$: 06 : 00am - 11 : 59am, Morning;
3. $TS3$: 00 : 00pm - 5 : 59pm, Afternoon;
4. $TS4$: 6 : 00pm - 9 : 59pm, Evening.

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

6. CASE STUDY: THE EXPO 2015 IN MILAN

In order to illustrate the effectiveness of our big data analytics framework, we built a case study focused on the well-known EXPO 2015 event in Milan, Italy, and on the basis of a set of geo-located posts gathered by the *FollowMe* suite.

The goal of the case study is to discover travelers coming to Lombardy, the region in the north-center of Italy where the main city is Milan, which was world-wide famous due to EXPO 2015. Therefore, we identified a pool of 30 European airports, which have been chosen on the basis of the presence of flights towards airports in Lombardy, and such that the number of posted posts in a single day is not huge, by following the goals of our case study.

We collected hang posts and timelines in the period between April 20, 2015, and June 29, 2015. By performing a query to discover trips in the bounding box of Lombardy, the *Trip Builder* generated a result set of 161 trips, formed by a total of 597 posts.

Discovered trips originated in 11 different airports, reported in Table 1. Here, for each airport, we report number of trips that originated from that airport (column *Trips*) as well as the total number of posts that constitute those trips (column *Sent Posts*). For instance, the 19 identified trips originated in Athens are composed by 73 posts. It is possible to notice that Spanish travelers use to post more than travelers coming from other countries: the system detected 47 trips from Barcelona and 30 from Munich, i.e. about 50% of the total.

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

Figure 4, which represents the dimension *Origin Airports*, shows the distribution of posts with respect to travelers coming from

Origin Airport	Trips	Sent Posts
Athens	19	73
Barcelona	47	176
Beauvais	4	39
Berlin	10	35
Charleroi	6	20
Copenhagen	15	47
Dublin	2	2
Munich	30	143
Stansted	12	41
Valencia	4	7
Warsaw	6	12

Table 1: Number of posts and travelers for Origin Airports

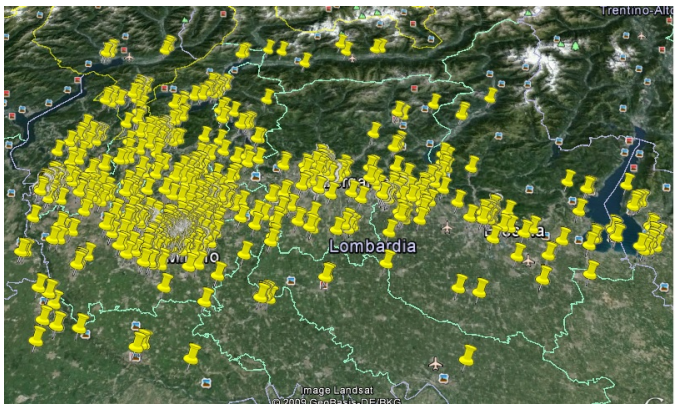


Figure 3: Post distribution on the Lombardy region.

Barcelona. It is possible to note that Spanish travelers concentrate their posts mainly in the area of Milan and beyond.

Moreover, some travelers that posted these posts arrived in Lombardy after EXPO 2015 started. For instance, in Figure 5 we report a post sent by a Spanish traveler in which the writer talks about EXPO 2015.

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

Figure 7 represents the distribution of posts with respect to the four Daylight Time Slots defined in Section 3 and the dimension *Time Slots*, respectively. It is possible to note how the distribution of posts is geographically more sparse in the Afternoon than in the others time slots, where the posts 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.

Figure 8 shows the distribution of posts with respect to hours of the day. Hours when the travelers post the most number of posts are: 14, 15, 20.

Figure 9 represents the distribution of posts with respect to days of the week. It is possible to note how the distribution of posts is geographically more concentrated on Monday around the Milan area than during the others days. Moreover, Tuesday and Friday are the days in which travelers frequently post their first post in

Figure 4: Post distribution with respect to travelers come from Barcelona.

Figure 5: Post of a Spanish traveler.

Lombardy. In conclusion, it is possible to note that the number of posts is similar for the first four days of the week, but Friday, Saturday and Sunday it increases significantly. Therefore, in these days the travelers post about 50% of posts. There are many reasons that explain this behavior, but one possible cause of this is that the travelers leave in weekend.

Finally, Figure 10 represents the distribution of the number of days elapsed since the day of the hang post. It is possible to note how a traveler posts his/her first post mainly within three days, in particular after 2 days. Moreover, on average, a traveler posts the first post in Lombardy 3.8 days after his/her hang post.

7. RELATED WORK

In this Section, we focus on state-of-the-art proposals that are correlated with our research.

In [17], Hawelka *et al.* analyze geo-located *Twitter* messages in order to uncover global patterns of human mobility. Based on a dataset of almost a billion tweets recorded in 2012, they estimate the volume of international travelers by country of residence. Mobility profiles of different nations were examined based on such characteristics as mobility rate, radius of gyration, diversity of destinations, and inflow-outflow balance. Temporal patterns disclose the universally valid seasons of increased international mobility and the particular character of international travels of different nations. Their analysis of the community structure of the *Twitter* mobility network reveals spatially cohesive regions that follow the regional division of the world. They validate their result using global tourism statistics and mobility models provided by other authors and argue that *Twitter* is exceptionally useful for understanding and quantifying global mobility patterns.

In [5], Bora *et al.* they try to understand how racial segregation of the geographic spaces of three major US cities (New York, Los Angeles and Chicago) affect the mobility patterns of people living in them. Collecting over 75 million geo-tagged tweets from these cities during a period of one year beginning October 2012 they identified home locations for over $30,000$ distinct users, and prepared models of travel patterns for each of them. Dividing the

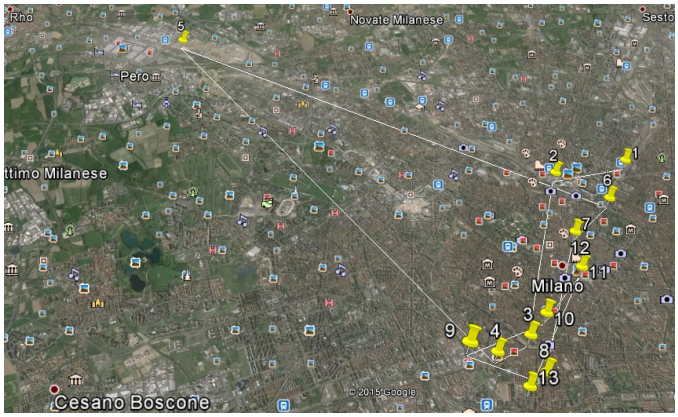
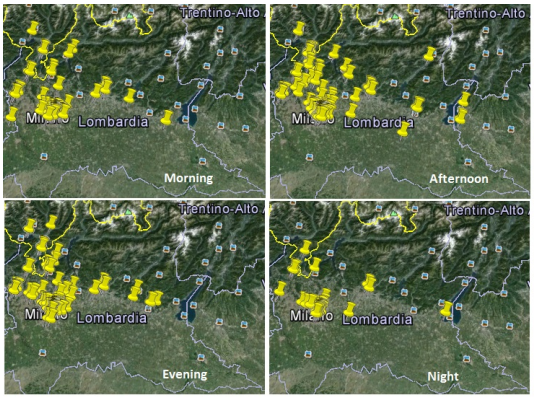


Figure 6: Path of a traveler.



Time Slot	Tweets Number
Morning	110
Afternoon	231
Evening	157
Night	113

Figure 7: Post distribution in Time Slots.

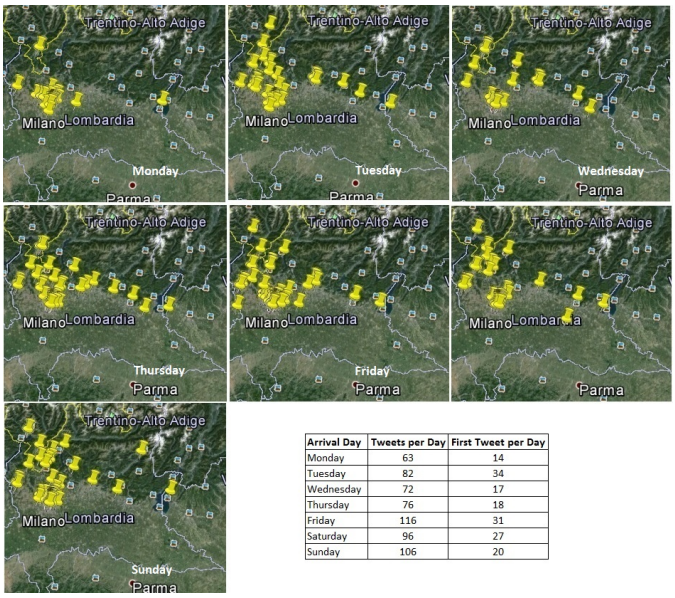
cities' geographic boundary into census tracts and grouping them according to racial segregation information they try to understand how the mobility of users living within an area of a particular predominant race correlate to those living in areas of similar race, and to those of a different race. While these cities still remain to be vastly segregated in the 2010 census data, they observe a compelling amount of deviation in travel patterns when compared to artificially generated ideal mobility. A common trend for all races is to visit areas populated by similar race more often. Also, blacks, Asians and Hispanics tend to travel less often to predominantly white census tracts, and similarly predominantly black tracts are less visited by other races.

In [27], Widner and Li use an advanced data-mining framework with a novel use of social media data retrieval and sentiment analysis to understand how geo-located tweets can be used to explore the prevalence of healthy and unhealthy food across the contiguous United States. Additionally, tweets are associated with spatial data provided by the US Department of Agriculture (USDA) of low-income, low-access census tracts (e.g. food deserts), to examine whether tweets about unhealthy foods are more common in these disadvantaged areas. Results show that these disadvantaged census tracts tend to have both a lower proportion of tweets about healthy foods with a positive sentiment, and a higher proportion of unhealthy tweets in general. These findings substantiate the methods used by the USDA to identify regions that are at risk of having low access to healthy foods.

In [25], Walther and Kaisser monitor all posts on *Twitter* issued

Morning		Afternoon		Evening		Night	
Hours	Tweets	Hours	Tweets	Hours	Tweets	Hours	Tweets
6	7	12	32	18	37	22	40
7	15	13	22	19	41	23	24
8	10	14	59	20	49	1	13
9	21	15	49	21	30	2	8
10	31	16	36	--	--	3	5
11	26	17	33	--	--	4	2
--	--	--	--	--	--	5	3

Figure 8: Post distribution in 24 hours.



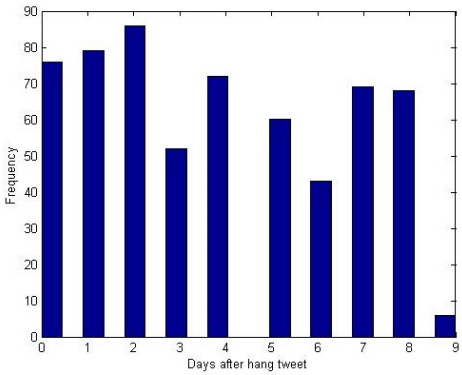
Arrival Day	Tweets per Day	First Tweet per Day
Monday	63	14
Tuesday	82	34
Wednesday	72	17
Thursday	76	18
Friday	116	31
Saturday	96	27
Sunday	106	20

Figure 9: Posts distribution in Days.

in a given geographic region and identify places that show a high amount of activity. In a second processing step, they analyze the resulting spatio-temporal clusters of posts with a Machine Learning component in order to detect whether they constitute real-world events or not. They show that this can be done with high precision and recall. The detected events are finally displayed to a user on a map, at the location where they happen and while they happen.

In [16], Grabovitch *et al.* present a study that focuses on these questions: Are users who are similar from the geo-spatial perspective (i.e., who send messages from nearby locations) also similar from the textual perspective (i.e., send messages with similar textual content)? Do posts with similar content have a spatial distribution similar to that of any random set of posts? The authors provide statistical tests to examine the correlation between textual content and geo-spatial locations in tweets. They show that although there is some correlation between locations and textual content, they provide different similarity measures, and combining these two properties for identification of users by their posts outperforms methods that merely use locations or only use the textual content, for identification.

In [23], Stephens and Poorthuis compare the social properties of *Twitter* users' networks with the spatial proximity of the networks. Using a comprehensive analysis of network density and network transitivity they found that the density of networks and the spatial clustering depends on the size of the network; smaller networks are more socially clustered and extend a smaller physical distance



Days	Frequency
0	76
1	79
2	86
3	52
4	72
5	60
6	43
7	69
8	68
9	6

Figure 10: Days elapsed after the first post.

and larger networks are physically more dispersed with less social clustering. Additionally, *Twitter* networks are more effective at transmitting information at the local level. For example, local triadic connections are more than twice as likely to be transitive than those extending more than 500 km. This implies that not only is distance important to the communities developed in online social networks, but scale is extremely pertinent to the nature of these networks. Even as technologies such as *Twitter* enable a larger volume of interaction between spaces, these interactions do not invent completely new social and spatial patterns, but instead replicate existing arrangements.

In [20], Lee and Sumiya aim to develop a geo-social event detection system by monitoring crowd behaviors indirectly via *Twitter*. In particular, they attempt to find out the occurrence of local events such as local festivals; a considerable number of *Twitter* users probably write many posts about these events. To detect such unusual geo-social events, we depend on geographical regularities deduced from the usual behavior patterns of crowds with geo-tagged microblogs. By comparing these regularities with the estimated ones, they decide whether there are any unusual events happening in the monitored geographical area. Finally, they describe the experimental results to evaluate the proposed detection method on the basis of geographical regularities obtained from a large number of geotagged tweets around Japan via *Twitter*.

8. CONCLUSIONS AND FUTURE WORK

Tourists are an important asset for the economy of the regions they visit. In particular, for public administrations, it is very useful to understand how tourists travel on the region they govern. In this paper, we have thus proposed a big data anlaytics framework that permits to follow traveling *Twitter* and *Instagram* users by tracking their geo-located messages they post during their trips. The comprehensive *FollowMe* suite, which implements the proposed framework, incorporates several tools that generate various outputs for the result set of reconstructed trips, so that several analysis dimensions (i.e., *Time Slot*, *Origin Airport*, *Path*, and others) can be built over them and exploited to analyze results. We provided architecture and functionalities of the propose suite, along with a case study focused on the EXPO 2015 event that clearly shows the benefits deriving from our proposed big data analytics framework. Our experiments, although limited in size, clearly witness the goodness of our approach and open the door to advanced scalable big data analytics tools over large-scale (big) data.

Future work is devoted to connect with other social networks and gather posts from them.This way, we should obtain a wider spec-

trum 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 close to the well-known *entity resolution* scientific area (e.g., [26, 15]).

9. REFERENCES

[1] Expo 2015 - milan. http://www.expo2015.org/.

[2] S. M. Allen, M. J. Chorley, G. Colombo, E. Jaho, M. Karaliopoulos, I. Stavrakakis, and R. M. Whitaker. Exploiting user interest similarity and social links for micro-blog forwarding in mobile opportunistic networks. *Pervasive and Mobile Computing*, 11:106–131, 2014.

[3] M. Balduini, A. Bozzon, E. D. Valle, Y. Huang, and G. Houben. Recommending venues using continuous predictive social media analytics. *IEEE Internet Computing*, 18(5):28–35, 2014.

[4] D. Bernstein. The emerging hadoop, analytics, stream stack for big data. *IEEE Cloud Computing*, 1(4):84–86, 2014.

[5] N. Bora, Y. Chang, and R. Maheswaran. Mobility patterns and user dynamics in racially segregated geographies of us cities. In *International Social Computing, Behavioral Modeling and Prediction*, April 2014.

[6] M. Cannataro, A. Cuzzocrea, C. Mastroianni, R. Ortale, and A. Pugliese. Modeling adaptive hypermedia with an object-oriented approach and xml. In *Proceedings of the Second International Workshop on Web Dynamics*, pages 35–44, 2002.

[7] M. Cannataro, A. Cuzzocrea, A. Pugliese, and V. P. Bucci. A probabilistic approach to model adaptive hypermedia systems. In *Proceedings of the First International Workshop for Web Dynamics*, pages 12–30, 2001.

[8] A. Cuzzocrea. 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, July 22-26, 2013*, pages 481–483, 2013.

[9] A. Cuzzocrea. Big data mining or turning data mining into predictive analytics from large-scale 3vs data: The future challenge for knowledge discovery. In *Model and Data Engineering - 4th International Conference, MEDI 2014, Larnaca, Cyprus, September 24-26, 2014. Proceedings*, pages 4–8, 2014.

[10] A. Cuzzocrea, L. Bellatreche, and I. Song. 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, October 28, 2013*, pages 67–70, 2013.

[11] A. Cuzzocrea, G. Psaila, and M. Toccu. Knowledge discovery from geo-located tweets for supporting advanced big data analytics: A real-life experience. In *Model and Data Engineering - 5th International Conference, MEDI 2015, Rhodes, Greece, September 26-28, 2015, Proceedings*, pages 285–294, 2015.

[12] A. Cuzzocrea, D. Saccà, and J. D. Ullman. Big data: a research agenda. In *17th International Database Engineering & Applications Symposium, IDEAS '13, Barcelona, Spain - October 09 - 11, 2013*, pages 198–203, 2013.

[13] A. Cuzzocrea and I. Song. 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, November 3-7, 2014*, pages 99–101, 2014.

[14] T. T. A. Dinh, M. Ganjoo, S. Braghin, and A. Datta. Mosco: a privacy-aware middleware for mobile social computing. *Journal of Systems and Software*, 92:20–31, 2014.

[15] D. Firmani, B. Saha, and D. Srivastava. Online entity resolution using an oracle. *PVLDB*, 9(5):384–395, 2016.

[16] I. Grabovitch, Y. Kanza, E. Kravi, and B. Pat. On the correlation between textual content and geospatial locations in microblogs. In *GeoRich14 June 23 2014, Snowbird, Utah (USA)*, June 2014.

[17] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(1):260–271, 2014.

[18] W. He. A survey of security risks of mobile social media through blog mining and an extensive literature search. *Inf. Manag. Comput. Security*, 21(5):381–400, 2013.

[19] O. Khalid, M. U. S. Khan, S. U. Khan, and A. Y. Zomaya. Omnisuggest: A ubiquitous cloud-based context-aware recommendation system for mobile social networks. *IEEE T. Services Computing*, 7(3):401–414, 2014.

[20] R. Lee and K. Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *ACM LBSN10, San Jose, CA, (USA)*, November 2010.

[21] Y. Liu, V. Lehdonvirta, T. Alexandrova, and T. Nakajima. Drawing on mobile crowds via social media - case ubiask: image based mobile social search across languages. *Multimedia Syst.*, 18(1):53–67, 2012.

[22] R. Nishi, T. Takaguchi, K. Oka, T. Maehara, M. Toyoda, K. Kawarabayashi, and N. Masuda. Reply trees in twitter: data analysis and branching process models. *Social Netw. Analys. Mining*, 6(1):26:1–26:13, 2016.

[23] M. Stephens and A. Poorthuis. Follow thy neighbor: Connecting the social and the spatial networks. *Computers, Environment and Urban Systems*, 41(1), 2014.

[24] N. Vastardis and K. Yang. An enhanced community-based mobility model for distributed mobile social networks. *J. Ambient Intelligence and Humanized Computing*, 5(1):65–75, 2014.

[25] M. Walther and M. Kaisser. Geo-spatial event detection in the twitter stream. In *35th European Conference on IR Research, ECIR 2013, Moscow (Russia)*, pages 356–367, March 2013.

[26] H. Wang, J. Li, and H. Gao. Efficient entity resolution based on subgraph cohesion. *Knowl. Inf. Syst.*, 46(2):285–314, 2016.

[27] M. Widener and W. Li. Using geolocated twitter data to monitor the prevalence of healthy and unhealthy food references across the us. *Applied Geography*, 54:189–197, 2014.

[28] Y. Yu, X. Huang, X. Zhu, and G. Wang. Camel: A journey group t-pattern mining system based on instagram trajectory data. In *Database Systems for Advanced Applications - 19th International Conference, DASFAA 2014, Bali, Indonesia, April 21-24, 2014. Proceedings, Part II*, pages 527–530, 2014.

[29] J. Zhang. Voluntary information disclosure on social media. *Decision Support Systems*, 73:28–36, 2015.