SMA4TD: A Social Media Analysis Methodology for Trajectory Discovery in Large-Scale Events

Eugenio Cesario[a], Fabrizio Marozzo[b], Domenico Talia[b], Paolo Trunfio[b]

[a]*ICAR-CNR, Rende, Italy*
[b]*DIMES, University of Calabria, Rende, Italy*

Abstract

The widespread use of social media platforms allows scientists to collect huge amount of data posted by people interested in a given topic or event. This data can be analyzed to infer patterns and trends about people behaviors related to a topic or an event on a very large scale. Social media posts are often tagged with geographical coordinates or other information that allows identifying user positions, this way enabling mobility pattern analysis using trajectory mining techniques. This paper describes SMA4TD, a methodology for discovering behavior and mobility patterns of users attending large-scale public events, by analyzing social media posts. The methodology is demonstrated through two case studies. The first one is an analysis of geotagged tweets for learning the behavior of people attending the 2014 FIFA World Cup. The second one is a mobility pattern analysis on the Instagram users who visited EXPO 2015. In both cases, a very high correlation (Pearson coefficient 0.7-0.9) was measured between official attendee numbers and those produced by our analysis. This result shows the effectiveness of the proposed methodology and confirms its accuracy.

Keywords: Mobility patterns, Urban Computing, Trajectory Mining, Social network, Social media posts, Geotagged data.

1. Introduction

The huge volume of user-generated data in social media platforms, such as Facebook, Twitter and Instagram, can be exploited to extract valuable information concerning human dynamics and behaviors. Social media analysis is a fast growing research area aimed at extracting useful information from this large amount of data. It is used for the analysis of collective sentiments, for

Email addresses: eugenio.cesario@icar.cnr.it (Eugenio Cesario), fmarozzo@dimes.unical.it (Fabrizio Marozzo), talia@dimes.unical.it (Domenico Talia), trunfio@dimes.unical.it (Paolo Trunfio)

understanding the behavior of groups of people or the dynamics of public opinion [17]. Social media posts are often tagged with geographical coordinates or other information (e.g., text, photos) that allows identifying users' positions. Therefore, social media users moving through a set of locations produce a huge amount of geo-referenced data that embed extensive knowledge about human dynamics and mobility behaviors. In fact, in the latest years, there has been a growing interest in the extraction of trajectories from geotagged social data using trajectory mining techniques [21].

This paper describes SMA4TD (Social Media Analysis for Trajectory Discovery), a methodology aimed at discovering behavior and mobility patterns of users attending large-scale public events. The methodology is composed of seven steps: *i*) identification of the set of events; *ii*) identification of places-of-interests where the events take place; *iii*) collection of geotagged items related to events and pre-processing; *iv*) identification of users who published at least one of the geotagged items; *v*) pre-processing and creation of the input dataset; *vi*) data analysis and trajectory mining; and *vii*) results visualization.

As a first case study, we present an analysis of geotagged tweets that we carried out to discover the behavior of people attending the 2014 FIFA World Cup. We monitored the Twitter users attending the World Cup matches to discover the most frequent movements of fans during the competition. The data source is represented by 526,000 tweets collected during the 64 matches of the World Cup from June 12 to July 13, 2014. For each match we considered only the geotagged tweets whose coordinates fallen within the area of stadiums, during the matches. Then, we carried out a trajectory pattern mining analysis on the set of the tweets considered. Original results were obtained in terms of number of matches attended by groups of fans, clusters of most attended matches, and most frequented stadiums. A strong correlation (Pearson coefficient 0.9) was measured between official attendee numbers and the number of Twitter users identified by our analysis.

A second case study presented in this paper is a mobility pattern analysis that we carried out on Instagram users who visited EXPO 2015, the Universal Exposition hosted in Milan, Italy, from May to October 2015. We collected and analyzed geotagged posts published by about 238,000 Instagram users who visited EXPO, including more than 570,000 posts published during the visits, and 2.63 million posts published by them from one month before to one month after their visit to EXPO. The analysis allowed us to discover how the number of visitors changed over time, which were the sets of most frequently visited pavilions, which countries the visitors came from, and the main flows of destination of visitors towards Italian cities and regions in the days after their visit to EXPO. Also in this case, a high correlation (Pearson coefficient 0.7) was measured between official visitor numbers and the visit trends produced by our analysis. This result shows the effectiveness of the proposed methodology and confirms the accuracy of the approach.

The structure of the paper is as follows. Section 2 provides some definitions and details the objectives of SMA4TD. Section 3 describes the methodology proposed in this paper. Section 4 and Section 5 describe how the methodology has been exploited on the two case studies introduced above. Section 6 discusses related work. Finally, Section 7 concludes the paper.

2. Definitions and objectives

This section provides a definition of the main concepts underlying the problem and the objectives of the analysis.

2.1. Preliminary Definitions

Let $\mathcal{P} = \{p_1, p_2, ...\}$ be a set of *places-of-interest* (*PoIs*), where each p_i is a specific area that is considered interesting for a community (during a given time period). For instance, a PoI could refer to a business location (e.g., shopping mall), a tourist attraction (e.g., theater, museum, park, bridge) or some particular location (square, pavilion, stadium) that is relevant during specific events. The concept of PoI considered in this work is not limited to a single geographical point or a street, but it refers to an area bounded by a polygon over a map. For this reason, PoIs can also be referred as *Regions-of-Interest* (*RoIs*), where the region represents the boundaries of the PoI's area [10].

Let $\mathcal{E} = \{e_1, e_2, ...\}$ be a set of events involving a massive presence of people, where an event $e_i = \langle p_i, [t_i^{begin}, t_i^{end}]\rangle$ has occurred in a place $p_i \in \mathcal{P}$ during the time interval $[t_i^{begin}, t_i^{end}]$. For instance, a single event e_i may be (*i*) a sport match played in a stadium, or (*ii*) a concert held in a theater or in a square, or (*iii*) a showcase hosted in a pavilion, as part of (*i*) a world-wide sport tournament, or (*ii*) a music tour of an artist, or (*iii*) a trade fair involving industry partners and customers, respectively. Two different events e_r and e_s can occur in the same place p_k, in two different time intervals. An event can have some additional descriptive properties, related to the specific domain.

Let $\mathcal{G} = \{g_1, g_2, ...\}$ be a set of geotagged items, where a *geotagged item* g_i is a social media content (e.g. tweet, post, photograph, video, link) posted by a user during an event $e_i \in \mathcal{E}$ from the place p_i where e_i was held. Specifically, a geotagged item g_i includes the following fields:

- $user_{ID}$, containing the identifier of the user who posted g_i;
- *coordinates*, consisting of *latitude* and *longitude* of the place where g_i was sent from;
- *timestamp*, indicating when (date and time) g_i was posted;
- *text*, containing a textual description of g_i;
- *tags*, containing the tags associated to g_i.

Let $\mathcal{U} = \{u_1, u_2, ...\}$ be a set of users, where each user u_i has published at least one geotagged item in $\mathcal{G}$.

2.2. Objectives of the analysis

The SMA4TD methodology is aimed at discovering behavior rules, correlations and mobility patterns of visitors attending large-scale events, trough the analysis of a large number of social media posts. In particular, the main goals of the methodology are as follows.

1. *Discovery of most visited places and most attended events.* We analyze the collected data to discover the places that have been most visited by users, and the events that have been most attended by visitors during the observed period.
2. *Discovery of most frequent sets of visited places and most frequent sets of attended events.* We extract the sets of places that are most frequently visited together by users, and the events that have been most attended by visitors during the observed period.
3. *Discovery of most frequent mobility patterns among places and most frequent sequences of attended events.* We analyze the collected data to discover mobility behaviors among the places, and to extract useful knowledge (i.e. patterns, rules and regularities) about the attended events.
4. *Discovery of the origin and destination of visitors.* We study the mobility flows of people attending the events, evaluating which countries visitors came from and which countries they moved after the events. In some cases, this information can give some insights about the touristic impact on the local territory.

The third and fourth objectives are the core goals of the proposed methodology, since they focus specifically on mobility pattern analysis. Even if, from a mobility analysis perspective, the first and second objectives are less important, they can provide useful insights about the popularity of places measured through social network users activity. This data may be compared with official data - when available - to validate the significance of the analysis. Indeed, in the use cases discussed in this paper, we registered a high degree of correlation between official attendees numbers and those provided by our analysis.

It is worth noting that some queries can be performed over the whole dataset, while others can be executed by analyzing specific subsets of data. For example, in some cases it is suitable to filter users with respect to their nationality, in order to perform a more detailed analysis about citizenship-related mobility. In other cases, events sharing common features (i.e., shows involving the same actor, matches related to the same team, etc.) can be grouped together, to discover topic-dependent patterns. Such choices strictly depend on the goals of the analysis and on the type of collected data as well.

3. Methodology

The SMA4TD methodology includes seven main steps:

1. Definition of the set of events $\mathcal{E}$.

2. Definition of the places-of-interests $\mathcal{P}$ where the events in $\mathcal{E}$ are held.
3. Collection and pre-processing of the geotagged items $\mathcal{G}$ related to the events in $\mathcal{E}$.
4. Identification of the users $\mathcal{U}$ who published at least one of the geotagged items in $\mathcal{G}$.
5. Creation of the input dataset $\mathcal{D}$.
6. Data and trajectory mining on $\mathcal{D}$.
7. Results visualization.

3.1. Steps 1-2: Definition of events and places-of-interest

The first two steps aim at defining the events $\mathcal{E}$ and the corresponding places-of-interest $\mathcal{P}$. Specifically, during step 1, each event is described by the id of the place-of-interest (PoI) where it is located, starting/ending time of the event, and other optional data (e.g., free/paid event, type of event, etc.). Step 2 is aimed at defining the geographical boundaries of the PoIs in $\mathcal{P}$. This can be done in two ways: *i*) *manually* defining the boundaries of the PoIs (e.g., as polygons on a map); *ii*) *automatically*, using external services (e.g., cadastral maps [10]), or public web services providing the geographical boundaries of a place given its name (e.g., OpenStreetMap[1]).

3.2. Steps 3-4-5: Collection and pre-processing of geotagged items, identification of users and creation of the input dataset

The goal of step 3 is to collect all the geotagged items $\mathcal{G}$ posted during each event $e_i \in \mathcal{E}$ from the place p_i where e_i was held. Data collection is done by using the publicly available APIs provided by most social media platforms. The $\mathcal{G}$ dataset is pre-processed in order to clean, select and transform data to make it suitable for analysis. In particular, we first clean the collected data by removing all items with unreliable positions (e.g., items with coordinates that have been manually set by users or applications). Then, we proceed by selecting only the geotagged items posted by users who actually attended an event, by removing replies and favorites posted by other users. Finally, we transform data by keeping one item per user per event, because we are interested to know only if a user attended an event or not. The identification of users is the goal of step 4. This is done by extracting the set $\mathcal{U}$ of distinct users who published at least one geotagged item in $\mathcal{G}$.

Step 5 creates the input datasets $\mathcal{D} = \{d_1, d_2, ...\}$, where d_i is a tuple

$$< u_i, \{e_{i1}, e_{i2}, ..., e_{ik}\}, optFields >$$

in which e_{ij} is the j^{th} *event* attended by user u_i, and *optFields* are optional descriptive fields (e.g., nationality, interests).

[1] https://www.openstreetmap.org/

3.3. Step 6: Data and trajectory mining

After having built the input dataset $\mathcal{D}$, it is analyzed for discovering behaviour and mobility patterns of users attending the large-scale event under investigation. Specifically, we perform both *associative* and *sequential analysis*, as described in the following.

Associative Analysis. Associative analysis is exploited with the goal of discovering (inside data) the item values that occur together with a high frequency. The mechanisms of association allow identifying the conditions that tend to occur simultaneously, or the patterns that repeat in certain conditions. This analysis also allows to derive implication rules like $e_s \implies e_r$ (if event e_s occurs, then it is likely that also event e_r occurs). Applied to dataset $\mathcal{D}$, we perform two associative mobility mining tasks: (*i*) *frequent event sets discovery*, aimed at extracting the sets of events (places) that are most frequently attended (visited) together by visitors during the whole observed large-scale event; and (*ii*) *frequent event rules extraction*, devoted to discover frequent associative rules among the events.

The pseudo-code implementing the frequent event sets discovery task is shown in Figure 1. The procedure receives in input the dataset $\mathcal{D}$, the event set $\mathcal{E}$ and the minimum support sup_{min}. The procedure initially computes the support $\sigma(e)$ of each event e, by making a single pass over D; upon completion of this step, the set of all frequent 1-event sets (i.e., single events) E_1 will be known (lines 1 and 2). Next, the algorithm iteratively generates new candidate k-event sets (C_k) using the frequent $(k-1)$-event sets found in the previous step (line 5); such a task is performed by exploiting the a-priori paradigm (well-known in literature [1]), based on candidate generation and pruning steps. Then, the support counting of the candidates is performed through an additional pass over the dataset (lines 6-10), by computing the support of each subset C_d of C_k that is contained in each transaction d (for efficiency reasons, we implement it by using a hash tree, whose description is out of the scope of this paper). After counting their supports, all candidate event sets whose support count is higher than sup_{min} (line 12) are considered as frequent event sets. The loop terminates when there are no new frequent event sets generated, i.e. $E_k = \emptyset$. The result of the procedure is a set $\mathcal{FE}$ of frequent events, computed as the union of all frequent event sets generated, i.e. $\mathcal{FE}= \bigcup_k E_k$.

The pseudo code used to discovery frequent event rules is shown in Figure 2. The procedure receives in input the set $\mathcal{FE}$ of frequent events and the minimum confidence $conf_{min}$. Initially, for each frequent event set $E_k \in \mathcal{FE}$, all non empty subsets of E_k are generated (lines 2-3) and stored in $\mathcal{H}_k$. Then, for each rule consequent candidate $h \in \mathcal{H}_k$, the procedure attempts to generate rules having $E_k - h$ as antecedent and h as consequent (lines 4-12). Specifically, if the confidence of the generated rule is higher than the minimum confidence, the new rule $\mathcal{R}$ = "$E_k - h \Rightarrow h$" is generated (lines 6-10) and added to the frequent rule set $\mathcal{FR}$, otherwise it is discarded.

Sequential Analysis. Sequential analysis algorithms are intended to discover the sequences of elements that occur most frequently in the data. Unlike associative analysis, in sequential analysis are fundamental the time dimension and

```
Discover-Frequent-Event-Sets($\mathcal{D}, \mathcal{E}, sup_{min}$)
$k \leftarrow 1$
$E_k \leftarrow \{e | e \in \mathcal{E}\ and\ \sigma(\{e\}) \geq sup_{min} * |\mathcal{D}|\}$
repeat
    $k \leftarrow k + 1$
    $C_k \leftarrow$ GenerateCandidates($E_{k-1}$);
    for each transaction $d \in \mathcal{D}$ do
        $C_d \leftarrow subset(C_k, d)$
        for each candidate itemset $c \in C_d$ do
            $\sigma(c) \leftarrow \sigma(c) + 1$
        end for
    end for
    $E_k \leftarrow \{c | c \in C_k\ and\ \sigma(c) \geq sup_{min} * |\mathcal{D}|\}$
until $E_k = \emptyset$
$\mathcal{FE} \leftarrow \bigcup_k E_k$
return $\mathcal{FE}$
```

Figure 1: Pseudo code for frequent event sets discovery.

```
Detect-Frequent-Event-Rules($\mathcal{FE}, conf_{min}$)
$\mathcal{FR} \leftarrow \emptyset$
for each frequent k-event set $E_k \in \mathcal{FE}$ do
    $\mathcal{H}_k \leftarrow subsets(E_k)$
    for each $h \in \mathcal{H}_k$ do
        $conf \leftarrow \sigma(E_k)/\sigma(E_k - h)$;
        if $conf \geq conf_{min}$ then
            create the antecedent $A = E_k - h$;
            create the consequent $C = h$;
            $\mathcal{R} \leftarrow$ "$A \Rightarrow C$";
            add $\mathcal{R}$ to $\mathcal{FR}$
        end if
    end for
end for
return $\mathcal{FR}$
```

Figure 2: Pseudo code for frequent event rules discovery.

the chronological order in which the values appear in the data. In our case, this type of analysis is useful to discover the most frequent mobility patterns among the places, and/or the most frequent sequences of attended events. Moreover, if the observed period is extended to some days (or weeks) before/after the event time, we can also discover the origin/destination (i.e., country, city) of visitors and which countries visitors came from/move after the event (i.e., to infer touristic insights). The pseudo code used to discovery frequent sequential events is shown in Figure 3. The procedure receives in input the set $\mathcal{FE}$ of frequent events (obtained though the procedure reported in Figure 1) and returns a set of frequent sequential events $\mathcal{FSE}$. First, each frequent event set $fes_k \in \mathcal{FE}$ is transformed into an ordered list (line 2), sorting each event according to its start time (i.e. the t^{begin} timestamp). Then, the chronological temporal order of the events in fes_k is verified, by guaranteeing that each event is finished before the subsequent event will start (lines 5-8). If such a condition is satisfied, the event list can be considered a sequential pattern (lines 10-11), otherwise not.

```
DISCOVER-EVENT-SEQUENCES($\mathcal{FE}$)
 1: for each frequent event set $fes_k \in \mathcal{FE}$ do
 2:    $el \leftarrow sort(fes_k)$
 3:    sequence = true
 4:    for $(i = 1, ..., |el| - 1)$ do
 5:       if $el_i.t^{end} > el_{i+1}.t^{begin}$ then
 6:          sequence=false
 7:          break
 8:       end if
 9:    end for
10:    if (sequence==true) then
11:       add $el$ to $\mathcal{FSE}$
12:    end if
13: end for
    return $\mathcal{FSE}$
```

Figure 3: Algorithm for sequential event detection

3.4. Step 7: Results visualization

Finally, results visualization is performed by the creation of info-graphics aimed at presenting the results in a way that is easy to understand to the general public, without providing complex statistical details that may be hard to understand to the intended audience. The graphic project is grounded on some of the most acknowledged and ever-working principles underpinning a 'good' info-graphic piece. In particular, we follow three main design guidelines: i) preferring a visual representation of the quantitative information to the written one; ii) minimizing the cognitive efforts necessary to decoding each system of signs; iii) structuring the whole proposed elements into graphic hierarchies.

Displaying quantitative information by visual means instead of just using numeric symbols - or at least a combination of the two approaches - has been proven extremely useful in providing a kind of sensory evidence to the inherent abstraction of numbers, because this allows everybody to instantly grasp similarities and differences among values. In fact, basic visual metaphors (e.g., the largest is the greatest, the thickest is the highest) enable more natural ways of understanding and relating sets of quantities [19].

In order to reduce the cognitive load necessary to information decoding and absorbing, several paradigms have been employed. In this regard, the most relevant are: i) aiming at the simplicity of the visual language (by using flat and monochromatic icons, for example); ii) limiting the number of different signs to the necessary; iii) sorting and arranging colors as syntactic elements; iv) using every visual component with coherency throughout each chart [15].

A proper hierarchy of the presented material (graphic images, written text and numbers, symbols, etc.) is a crucial factor that helps the readers in identifying which are the core issues to focus on and what is auxiliary or complementary to them. Since the chart reading process can start randomly everywhere, it is important to create visual affordances capable of attracting the observers gaze and so inducing their understanding pathways to begin from the most proper points. These reading patterns are mostly achieved by adjusting and combining visual features such as dimension, position, color, composition [18].

According to the principles introduced here, we use a high-level visualization model that helps readers to easily catch the main concepts and the key meaning of the knowledge extracted by the data mining process.

3.5. A Running Example

To better explain how the SMA4TD methodology works, we describe it using a running example. Let us consider a large-scale cultural event in the center of Rome composed by free or paid events like concerts, guided tours, outdoor exhibits spread over three days (e.g., from November, 1^{st} to November, 3^{rd}, 2016).

At *Step 1*, the following events are defined:

$$e_1 = \langle Colosseum, (\text{2016-11-01T19:00}, \text{2016-11-01T23:00}), \text{Paid, Concert}\rangle$$
$$e_2 = \langle RomanForum, (\text{2016-11-02T10:00}, \text{2016-11-02T19:00}), \text{Free, Guided tour}\rangle$$
$$e_3 = \langle TiberIsland, (\text{2016-11-02T10:00}, \text{2016-11-02T19:00}), \text{Free, Guided tour}\rangle$$
$$e_4 = \langle PiazzaVenezia, (\text{2016-11-03T09:00}, \text{2016-11-03T18:00}), \text{Free, Outdoor exhibit}\rangle$$
$$e_5 = \langle CircusMaximus, (\text{2016-11-03T19:00}, \text{2016-11-03T23:00}), \text{Paid, Concert}\rangle$$

For each event, place id, date and time, and some additional fields (free or paid participation, and type of event) are specified. For example, event e_1 is held at the *Colosseum* on $November\ 1st$ (from 19:00 to 23:00), and it is a concert with paid entrance. Event e_2 is at the $RomanForum$ on $November\ 2nd$ (from 10:00 to 19:00), and it is a free guided tour.

At *Step 2*, the places-of-interests of each event are defined. In our example, the events took place in the following sites (whose boundaries are shown in Figure 4(a)): *Colosseum*, $RomanForum$, $TiberIsland$, $PiazzaVenezia$ and $CircusMaximus$.

Figure 4(b) shows on a map the geotagged items associated to the five events under analysis (*Step 3*). As an example, g_1 is a geotagged item related to event e_1:

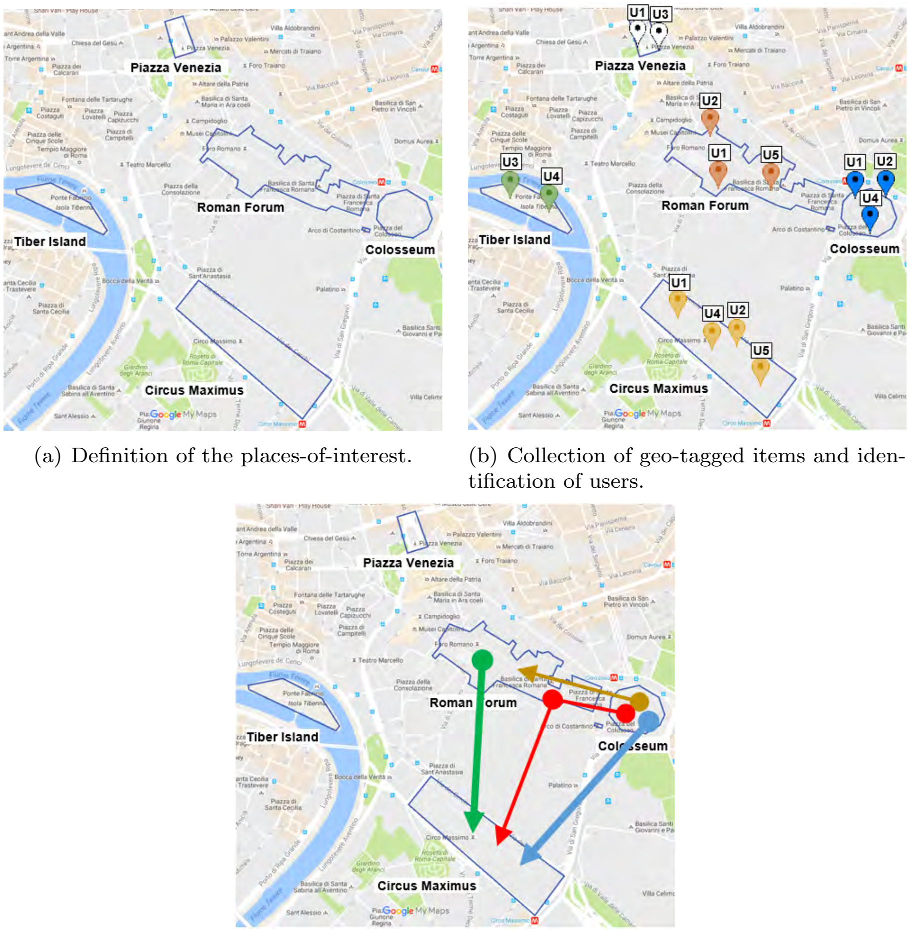
$$g_1 = \langle \text{``}u1\text{''}, [41.889995, 12.492157], \text{2016-11-01T20:35}, \\ \text{``Unforgettable experience at the Coliseum.''}, [\#Coliseum, \#Art, \#History]\rangle$$

where $u1$ is the id of the user who has posted g_1, $[41.889995, 12.492157]$ are the coordinates of g_1, *2016-11-01T20:35* is the timestamp indicating when it was posted, *"Unforgettable experience at the Coliseum."* is the text and $[\#Coliseum, \#Art, \#History]$ are the tags. Each geotagged item in Figure 4(b) is represented as a pointer. The items related to events e_1, e_2, e_3, e_4 and e_5 are colored respectively as blue, orange, green, white and yellow. On top of each pointer, is reported the id of user who has posted the item. For example, user u_1 posted items during the events e_1, e_2, e_4, e_5, and user u_2 posted items during the events e_1 and e_5. Therefore, as a result of *Step 4*, five users have been identified starting from the collected geotagged items.

After *Step 5*, the following input dataset has been generated, in which the collected geotagged items are aggregated on a per-user basis:

$< u_1, \{e_1, e_2, e_4, e_5\}, Italian, art >$
$< u_2, \{e_1, e_2\}, German, music >$
$< u_3, \{e_3, e_4, e_5\}, French, art >$
$< u_4, \{e_1, e_3, e_5\}, German, music >$
$< u_5, \{e_2, e_5\}, Italian, music >$

For each user, the input dataset lists the events in which he/she participated



(a) Definition of the places-of-interest.

(b) Collection of geo-tagged items and identification of users.

(c) Trajectory mining and results visualization.

Figure 4: Main steps of SMA4TD on a map.

and some optional fields like, for example, his/her nationality and main interest. For example, user $u1$ attended the events $e1$, $e2$, $e4$, $e5$, is $Italian$ and interested in art, while user $u2$ attended the events $e1$ and $e2$, is $German$ and interested in $music$.

At *Step 6*, the following rules are found by trajectory mining:

$(e_1, e_5), 3$
$(e_2, e_5), 3$
$(e_1, e_2), 2$
$(e_1, e_2, e_5), 2$

For example, the first rule says that 3 users attended the events $< e1, e5 >$, while the last rule says that 2 users attended the events $< e1, e2, e5 >$.

Finally, Figure 4(c) represents the trajectories identified by the rules as directed lines (*Step 7*), where different line sizes are used to represent the support of each rule.

It is worth noticing that exploiting the optional fields in the input dataset, we can analyze behavior of specific classes of users. For example, we can analyze the movement patterns of the $German$ users, or the patterns of users interested in art, or those of users attended only $free$ events.

4. First case study: FIFA World Cup

In this section we present the results obtained by the analysis of geotagged tweets of people attending the FIFA World Cup 2014. Specifically, we monitored the Twitter users attending the World Cup matches and analyzed such data using the SMA4TD methodology to discover behavior and most frequent movements of fans during the competition [6].

The 2014 FIFA World Cup was the 20^{th} edition of the quadrennial world championship for national football teams organized by FIFA. It took place in Brazil from 12 June to 13 July 2014. Thirty-two national teams participated to the competition, with a total of 64 matches that have been played in 12 stadiums distributed across Brazil. Cumulatively, the whole competition gathered more than five million people who attended soccer matches and other related events[2], and represented one of the most important large-scale event happened in the year 2014.

4.1. Steps 1-2: Definition of events and places of interest

The set of events $\mathcal{E}$ is composed by the 64 matches played during the World Cup. Specifically, $\mathcal{E} = \{e_1, e_2, ..., e_{64}\}$, where each match e_i is described by the following properties:

$$e_i = \langle p_i, [t_i^{begin}, t_i^{end}], team1, team2, phase\rangle$$

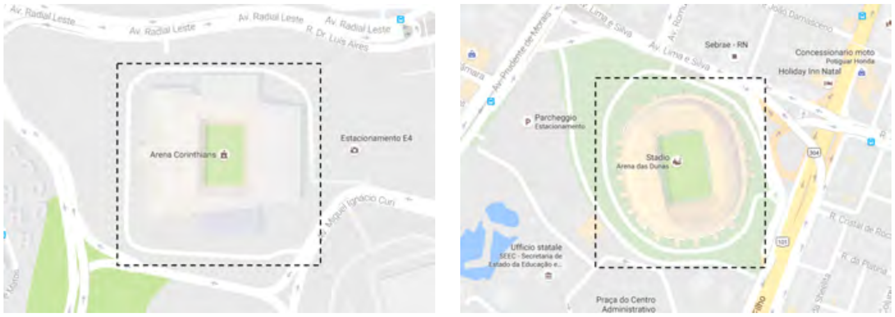
[2] http://www.fifa.com/worldcup/news/y=2014/m=9/news=2014-fifa-world-cup-braziltm-in-numbers-2443025.html

where p_i is the stadium, t_i^{begin} is three hours before the start of the match, t_i^{end} is three hours after the begin of the match, $team1$ is the first team and $team2$ is the second team, and $phase$ is one of: *’opening match’* (match no. 1), *’group stage’* (matches no. 2-48), *’round of 16’* (matches no. 49-56), *’quarter finals’* (matches no. 57-60),*’semi-finals’* (matches no. 61-62), *’final’* (match no. 64). For example, the first two matches e_1 and e_2 are represented by

$$e_1 = \langle SãoPaulo, [\text{2014-06-12T14:00}, \text{2014-06-12T20:00}], Brasil, Croatia, OpeningMatch\rangle$$
$$e_2 = \langle Natal, [\text{2014-06-13T10:00}, \text{2014-06-13T16:00}], Mexico, Cameroon, GroupStage\rangle$$

showing that the first match was played in *São Paulo* (i.e., *Arena de São Paulo* stadium) at *17:00* local time (from *14:00* to *20:00*), by the teams of *Brasil* and *Croatia*, and was the *opening match*, while the second match was played in *Natal* (i.e., *Estadio das Dunas* stadium) at 13:00 local time (from *10:00* to *16:00*), by *Mexico* and *Cameroon*, and was a *’group stage’* match.

Therefore, the places-of-interest, in this case study, are the stadiums in which the World Cup matches have been played. Specifically, we defined the set of PoIs $\mathcal{P} = \{p_1, p_2, ..., p_{12}, \}$, where each p_i is a stadium that hosted at least one match during the competition. The corresponding RoIs have been manually defined from a map as the smallest rectangles fully containing the boundaries of each stadium. For example, Figure 5 shows the RoIs traced for two stadiums: Figure 5(a) depicts the RoI of the Arena de São Paulo, while Figure 5(b) shows the RoI of the Estadio das Dunas at Natal.



(a) RoI of the Arena de São Paulo.

(b) RoI of the Estadio das Dunas (Natal).

Figure 5: Two examples of RoIs.

4.2. Steps 3-4-5: Collection and pre-processing of geotagged items, identification of users and creation of the input dataset

After having defined events and PoIs, we collected all the geotagged tweets sent by fans during the World Cup matches. For each match we considered only the tweets posted from coordinates falling within the above defined RoIs during the matches. Totally, the number of tweets that have been collected (from June 12 to July 13, 2014) amounted to about 526,000. Formally, we gathered the set of geotagged tweets $\mathcal{G} = \{g_1, g_2, ...\}$, where each tweet g_i contains $user_{ID}$, *coordinates*, *timestamp*, *text*, *tags*, and *application* (optional field) used to generate g_i. An example of geotagged item related to the opening match e_1 is

$$g_i = \langle 11223344, [-23.545531, -46.473372], \text{2014-06-12T16:59} \\ \text{“Proud to be here.”}, [\#BRAvsCRO, \#Brasil2014], TwitterForAndroid\rangle$$

A pre-processing step on $\mathcal{G}$ has been performed to clean, select and transform data to make it suitable for analysis. First, we cleaned collected data by removing all the tweets with unreliable position (e.g., tweets with coordinates manually set by users or applications). Then, we selected only tweets written by users attending the matches, by removing re-tweets and favorites posted by other users. Finally, we transformed data by keeping one tweet per user per match, because we were interested to know only if a user attended a match or not. After the preprocessing step, the final number of tweets amounts to about 151,000. Then, we built the final dataset $\mathcal{D}$ that results composed of about 10,000 transactions, each one containing the list of matches attended by a single Twitter user. Formally,

$$\mathcal{D} = T_1, T_2, \ldots, T_n$$

where $Ti =< u_i, m_{i1}, m_{i2}, , m_{ik} >$ and $m_{i1}, m_{i2}, , m_{ik}$ are the matches attended by a user u_i.

4.3. Steps 6-7: Data and trajectory mining and results visualization

The main goals of the analyses conducted on $\mathcal{D}$ were:

- Estimating the number of people attending the matches over time.
- Estimating the number of matches attended by fans during the competition.
- Identifying the most frequent sequences of matches attended by fans, either in the same stadium or to follow a given soccer team.
- Finding the most frequent movement patterns obtained by grouping matches based on the phase in which they were played.

In this case, all the events (matches) are ordered chronologically. Therefore, a sequential analysis was carried out to discover the sequences of elements that occured most frequently in the data, as discussed in Section 3.3.

Finally, results visualization was implemented by the creation of infographics aimed at presenting the data and trajectory mining results in a compact and easy-to-understand way.

4.4. Results

In the following we present the main results of the data and trajectory mining analyses discussed above.

First, we describe how the number of people attending the matches changed over time. To do that, we report in Figure 6 trends and numbers *(i)* of Twitter users we tracked attending at the matches during the World Cup, and *(ii)* of attendees officially published by the FIFA website[3]. We used different scales for Twitter and the official fans: on the right is the scale of the formers, while on the left is the scale of the latter ones. Specifically, Figure 6(a) shows a time plot of the collected attendance data, in which the number of attendees is plotted versus the number of matches. It clearly shows that there are several peaks of participation during the competition, probably corresponding to some matches that have attracted more attention with respect to other ones. Interestingly, in some cases Twitter data peaks are equivalent to the official attendance ones. For example, peaks corresponding to the matches n. 11, 19, 44, 50 and 57 occur in both curves. By observing the trend lines (dotted lines), computed as second-order polynomial curves, it can be noted that the number of fans are pretty high for the first matches, then there is a decreasing during the middle part of the World Cup and it steadily increases from the round of 16 to the final match.

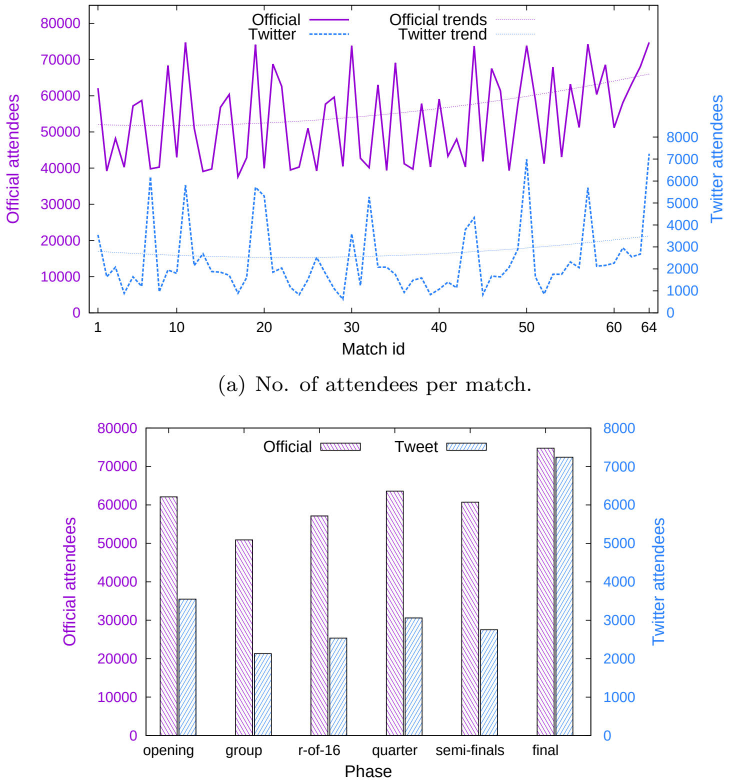
Figure 6(b) shows the average distribution of attendees per match grouped by World Cup phases, for both Twitter and official data. Histograms show that the number of attendees are highly skewed for different phases (periods) of the competition. Interestingly, in both cases the average number of attendees decreased during the *group stage* phase, corresponding to troughs in both data distributions. The Pearson correlation between official attendee numbers and those collected by monitoring Twitter users is equal to 0.9, which confirms the accuracy of the results we obtained.

Now, let us describe the participation of fans to the matches. Table 1 groups fans based on the number of matches attended during the whole World Cup. The results show that 71.3% of the fans attended a single match, 16% attended two matches, 6% attended three matches, and only 3% attended four matches. It is worth noticing that 3.7% of the spectators attended five or more matches. Looking at the Twitter profiles of the spectators of the latest set, we found that many of them were journalists.

Table 2 provides a general classification of the paths through the Brazil 2014 stadiums followed by fans who attended at least two matches. Focusing on fans who attended two or three matches, the table shows that:

- Among the spectators who attended two matches, 62.9% attended matches played in the same stadium, while 22.2% attended matches played by the

[3] http://www.fifa.com/worldcup/archive/brazil2014



(a) No. of attendees per match.

(b) Average no. of attendees per match, grouped by World Cup phases.

Figure 6: Statistics about attendees, comparing Twitter users and official attendee numbers.

same team.

- Among the spectators who attended three matches, 48.8% attended matches played in the same stadium, while 11.8% attended matches played by the same team.

In general, the results show that most of who attended multiple matches did it staying in the same city.

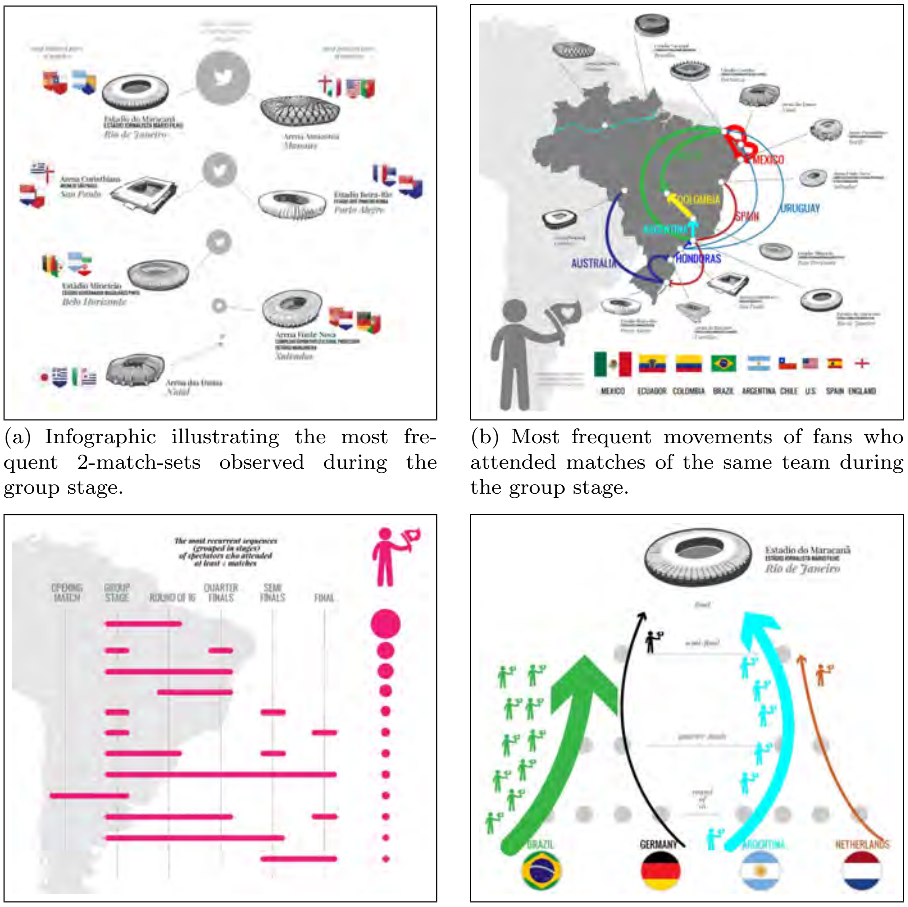
Figure 7(a) shows the most frequent 2-match-sets observed during the group stage, from June 12 to June 26, 2014. The set with the highest support was the couple of matches ⟨*England-Italy*, *USA-Portugal*⟩ played in Manaus, followed at short distance by ⟨*Argentina-Bosnia*, *Spain-Chile*⟩ played in Rio de Janeiro, and by ⟨*Uruguay-England*, *Netherlands-Chile*⟩ played in São Paulo. The size of each Twitter icon in the figure is proportional to the number of Twitter users who attended the pairs of matches besides the icon. The largest icon refers to two pairs of matches (England-Italy and USA-Portugal; Argentina-Bosnia

Table 1: Number of matches attended.

No. of matches	Spectators
1	71.3%
2	16.0%
3	6.0%
4	3.0%
5 or more	3.7%

Table 2: Classification of the paths followed by fans

No. of matches	Same stadium	Same team
2	62.9%	22.2%
3	48.8%	11.8%
4	41.0%	7.2%
5	37.0%	8.4%
6	33.7%	4.8%



(a) Infographic illustrating the most frequent 2-match-sets observed during the group stage.

(b) Most frequent movements of fans who attended matches of the same team during the group stage.

(c) Patterns of movements by grouping matches in phases.

(d) Comparison among the flows of the semi-finalists fans.

Figure 7: Infographic about user movements during FIFA World Cup 2014 Brazil.

and Spain-Chile) because these two pairs registered approximatively the same number of attendees. The same applies to the second largest icon.

Figure 7(b) shows the most frequent paths of fans who attended two or three matches of the same team during the group stage. The most frequent 2-match-set was ⟨*Colombia-Greece*, *Colombia-Cote dIvoire*⟩, followed by ⟨*Brazil-Mexico*, *Croatia-Mexico*⟩, and by ⟨*Argentina-Bosnia*, *Argentina-Iran*⟩, i.e., matches likely attended by fans of Colombia, Mexico and Argentina. The most frequent 3-match-set was ⟨*Mexico-Cameroon*, *Brazil-Mexico*, *Croatia-Mexico*⟩, followed by ⟨*Brazil-Croatia*, *Brazil-Mexico*, *Cameroon-Brazil*⟩, and by ⟨*Chile-Australia*, *Australia-Netherlands*, *Australia-Spain*⟩. Looking at their nationality, spectators were likely fans of Mexico, Brazil and Australia. The figure illustrates these frequent movements as directed arcs linking the stadiums in which the matches above were played. Different line sizes are used for the arcs to represent the support of each match-set.

At the end of the group stage, we carried out a specific analysis on the Twitter users who were present at the opening match ⟨*Brazil-Croatia*⟩ played on June 12, 2014 in São Paulo. The results showed that, among these fans:

- 50.4% did not attend other matches after the opening one;
- 13.7%, after the opening match, moved to Rio de Janeiro to attend other matches;
- 9.5% attended other matches in the same stadium in São Paulo;
- 7.0% attended another match played by the Brazilian team, either ⟨*Brazil-Mexico*⟩ played in Fortaleza or ⟨*Cameroon-Brazil*⟩ played in Brasilia;
- 2.8% attended both the following matches played by the Brazilian team, i.e. ⟨*Brazil-Mexico*⟩ and ⟨*Cameroon-Brazil*⟩.

Figure 7(c) shows the patterns of movements based on the grouping above, and the relative frequency (support) of these patterns. The most frequent pattern is represented by fans who attended at least one match of the group stage and one match of the round of 16. The second most frequent pattern are fans who attended a match of the group stage and one of the quarter finals. The third most frequent pattern includes spectators of at least one match of group stage, group of 16 and quarter finals. The relative frequency of each pattern is represented by a circle: the larger the circle, the higher the frequency. The least frequent pattern was that of fans who attended one semi-final and the final match.

Finally, Figure 7(d) compares the flows of the semi-finalists fans, who attended at least two matches of their team. Also in this case, the figure uses different line sizes to represent the relative number of users in each flow. The gray dots in the background indicate the number of matches at a given phase (e.g., two dots in the semi-finals), while the number of human cliparts is proportional to the number of Twitter users who followed a given team up to the semi-final.

5. Second case study: EXPO 2015

This section describes the results obtained using the SMA4TD methodology for analyzing mobility patterns of people attending EXPO 2015. Specifically, we monitored Instagram users who visited EXPO pavilions, to discover mobility patterns inside the exhibition area, correlations among visits to pavilions and the main flows of origin/destination of visitors [7].

EXPO 2015[4] was a Universal Exposition held under the theme "Feeding the Planet, Energy for Life", which was hosted in Milan, Italy, from May 1^{st} to October 31^{st}, 2015. Exhibitors were individual countries, international organizations, civil society organizations and companies, for a total of 188 exhibition spaces. Some of the exhibitors were hosted inside individual (self-built) pavilions, while others were hosted inside shared pavilions. For the sake of uniformity, in this paper we will use the term pavilion to indicate both an individual pavilion and a distinct area (assigned to a given exhibitor) of a shared pavilion. Cumulatively, about 22.2 million people visited the EXPO area and its pavilions during the six months of the whole exposition, making it the world-wide largest event of the year 2015.

Visitors at EXPO used various social network to share their experience with friends and followers. In particular, as one of the most common activities was image sharing, Instagram users resulted very active in doing that[5].

5.1. Steps 1-2: Definition of events and places of interest

The set of events $\mathcal{E}$ considered for this scenario is composed by the showcases (each one organized by a country or organization/company) exhibited in the exposition spaces (generally referred as pavilions in the following). Specifically, let us consider $\mathcal{E}$=$\{e_1, e_2, ..., e_{188}\}$, where each e_i is described by the following properties:

$e_i = \langle p_i, [t_i^{begin}, t_i^{end}] \rangle$

where p_i is the pavilion, t_i^{begin} is May 1st and t_i^{end} is October 31st.

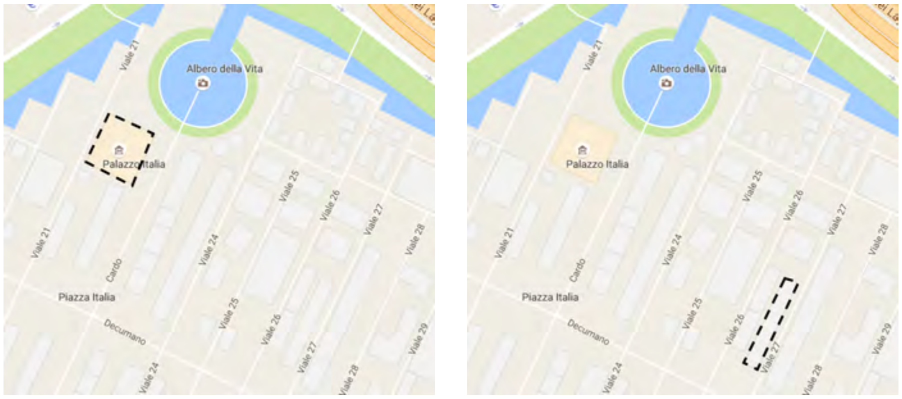
The places-of-interest to be considered are the pavilions. Specifically, we defined the PoI set $\mathcal{P} = \{p_1, p_2, ..., p_{188}\}$, where each p_i is a pavilion that has been used as exhibition area during the EXPO 2015. For each PoI, we drew its corresponding RoI as a rectangle bounding the pavilion area. For example, Figure 8 shows the RoIs of two EXPO pavilions: Italy pavilion in Figure 8(a) and USA pavilion in Figure 8(b).

5.2. Steps 3-4-5: Collection and pre-processing of geotagged items, identification of users and creation of the input dataset

After having fixed events and PoIs, we performed data acquisition by collecting all geotagged posts published by Instagram users who visited at least

[4] http://www.expo2015.org/

[5] http://volunteer.expo2015.org/en/news/expo-and-social-media-big-success-also-thanks-volunteers



(a) RoI of Italy pavilion.

(b) RoI of USA pavilion.

Figure 8: Two example of pavilions RoIs.

one pavilion during the EXPO. Specifically, we collected posts with coordinates falling within the above-defined RoIs (i.e., rectangles bounding the pavilions). In addition, we gathered also posts that were pre-related and post-related to the EXPO visits, that is, posts published by users from one month before to one month after their visit at EXPO. Totally, we collected and analyzed the geotagged posts published by about 238,000 Instagram users who visited EXPO, resulting in more than 570,000 posts published during the visits, and 2.63 million posts published by users one month before to one month after their visit.

Overall, we collected the set $\mathcal{G} = \{g_1, g_2, ...\}$ that is composed of geotagged Instagram post, where each post g_i contains $user_{ID}$, *coordinates*, *timestamp*, *text* and *tags*. An example of a geotagged item, related to the Italy pavilion, is:

$$\begin{aligned} g_i = \langle & 111122222, [-23.545531, -46.473372], \text{2015-09-01T15:03}, \\ & \text{“Discovering the Italian excellence in food”}, [\#Italy, \#food, \#excellence] \rangle \end{aligned}$$

Data have been preprocessed by keeping one Instagram post per user per pavilion per day, because we were interested to know only if a user visited a pavilion (or not) in a given day. The final dataset $\mathcal{D}$ contains about 570,000 transactions, each one containing the list of PoIs visited by a single Instagram user. Formally,

$$\mathcal{D} = T_1, T_2, , T_n$$

where the i-th tuple is $T_i =< u_i, \{p_{i1}, p_{i2}, ..., p_{ik}\} >$, where p_{ij} contains *event* (and *timestamp*, as additional field) of the j-th post published by user u_i. The PoI associated to a post depends whether it is a current, pre- or post-related post. Specifically, the PoI of a post sent during the EXPO visit corresponds to the visited pavilion, while the PoI of a post sent before or after the visit corresponds to a city/region/state (outside EXPO).

Let $\mathcal{U} = \{u_1, u_2, ...\}$ be a set of users, where each user u_i contains a $user_{ID}$ and its *nationality*. The nationality of a user u_i have been achieved analyzing the posts published outside EXPO from one month before to one month after her/his visit to EXPO. In case of visitors coming from Italy, we also discovered city and region of origin.

5.3. Steps 6-7: Data and trajectory mining and results visualization

The main goals of the analyses carried out in this case study were:

- Estimating the number of people visiting EXPO over time.
- Identifying the most visited pavilions.
- Identifying the most frequent sets of visited pavilions and the trajectories among pavilions.
- Identifying origin and destination of visitors.

The input dataset $\mathcal{D}$ has been analyzed using both associative and sequential analysis (see Section 3.3). In particular, associative analysis was used to identify the most frequent sets of visited pavilions (without considering the visit order), while sequential analysis was carried out to identify the most frequent trajectories among pavilions and the origin and destination of visitors.

Also in this case, according to the principles introduced in Section 3.4, we prepared some infographics to help readers to easily catch the main concepts and the key meaning of the knowledge extracted by the data mining process.

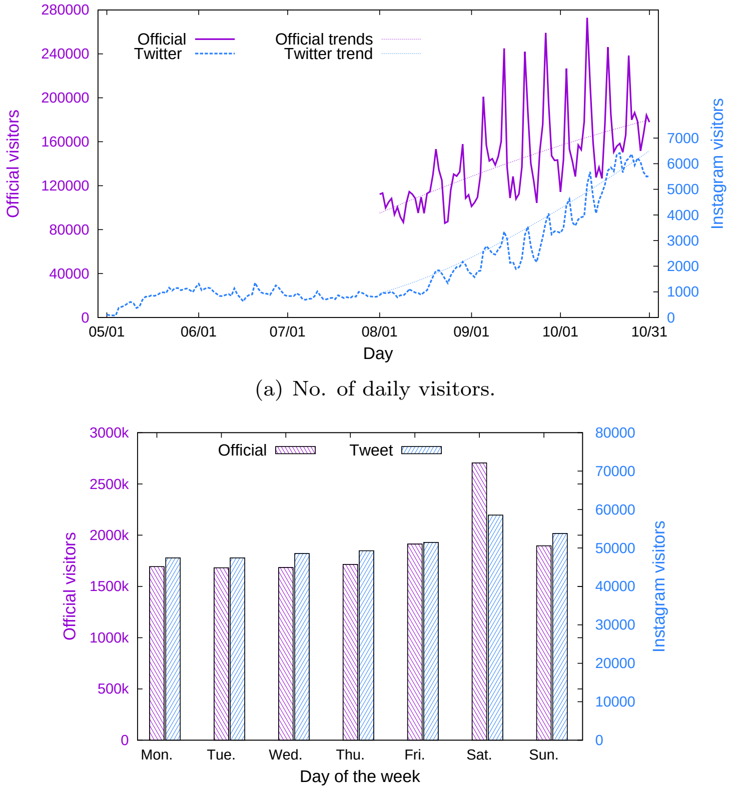
5.4. Results

This section presents the main results of the data and trajectory mining analysis introduced above.

Figure 9 shows a comparison between trends and numbers of the Instagram visitors we tracked, and the official visitors published on the EXPO website[6]. The observed period is August 1^{st} - October 31^{st}, but official numbers have been published only for the period starting in August, thus the corresponding curve has been traced only for the last three months. We used different scales for Instagram visitor numbers and the EXPO visitor ones: on the right is the scale of the formers, while on the left is the scale of the latter ones. In particular, Figure 9(a) shows a time plot of the daily visits to EXPO. The trends are quite evident: initially (May and June) the visitors are relatively few; then, they grow significantly during the months of September and October. Moreover, there are several peaks of attendance, corresponding to visits occurred during the week-end days. By looking at the trends in the figure, it can be noted a strong correlation (Pearson coefficient 0.7) between official visitor numbers and

[6] http://www.expo2015.org/

those obtained from our analysis, which confirms the reliability of the results we obtained. Figure 9(b) compares Instagram and official visitor numbers, aggregated by the week day (Pearson correlation 0.94). The results clearly show that during the week-end days there is a peak of visits, with the highest number of people registered on Saturdays.



(a) No. of daily visitors.

(b) No. of visitors per week day.

Figure 9: Statistics about visitors, comparing Instagram users and official attendee numbers.

Table 3 presents the pavilions that were visited by at least 6% of the users. The table shows that the most visited pavilion, according to the Instagram posts that were analyzed, was that of China (more than 20% of visitors), followed by the pavilions of Japan, UK, Korea, Russia, Brazil and USA. All other pavilions had percentages of visitors below 10%.

Instagram data were also analyzed to extract the sets of pavilions that were most frequently visited together by the Instagram users who attended EXPO. The outcomes of the analysis are the sets of pavilions whose support s is equal to or greater than a given minimum support count s_{min}. We used $s_{min} = 2\%$, which led to the extraction of sets of length 2 (all the sets of length greater than 2 had $s < s_{min}$). The sets of length 2 (pairs) extracted from this analysis are

Table 3: Most visited pavilions (visitors $\geq 6\%$).

Rank	Pavilion	Visitors
1	China	20.77%
2	Japan	16.38%
3	UK	14.40%
4	Korea	13.03%
5	Russia	13.02%
6	Brazil	12.56%
7	USA	11.75%
8	UAE	9.32%
9	Qatar	9.09%
10	Italy	8.59%
11	Netherlands	7.75%
12	Austria	7.72%
13	Spain	6.94%
14	Thailand	6.78%
15	Azerbaijan	6.48%
16	Poland	6.46%
17	Vietnam	6.38%
18	Nepal	6.35%
19	France	6.08%

Table 4: Most visited pairs of pavilions (support $\geq 2\%$).

Rank	Pavilion 1	Pavilion 2	Support
1	China	Japan	3.77%
2	China	UK	3.75%
3	China	Korea	3.29%
4	China	Russia	3.04%
5	Brazil	China	3.03%
6	Japan	UK	2.92%
7	China	USA	2.76%
8	Japan	Russia	2.57%
9	Japan	Korea	2.51%
10	Korea	UK	2.42%
11	Japan	USA	2.39%
12	China	UAE	2.23%
13	Russia	USA	2.20%
14	Russia	UK	2.19%
15	Brazil	Japan	2.14%
16	Brazil	UK	2.13%
17	China	Qatar	2.09%
18	China	Thailand	2.04%
19	Japan	UAE	2.03%

reported in Table 4. As shown in the table, 3.77% of users visited the pavilions of China and Japan, while 3.75% visited the pavilions of China and the UK. The other pairs of pavilions with a percentage of visits over 3%, were ⟨China, Korea⟩, ⟨China, Russia⟩ and ⟨Brazil, China⟩.

Another result of our analysis was discovering the sequences of visits to the pavilions, i.e., the sequential order with which the pavilions have been visited. Table 5 shows that the 2.3% of people visited the pavilion of China and then that of Japan, which is the most frequent sequence of visits monitored. It is interesting to note that the second and third most common sequences involve the pavilions of China and the UK, but in different order: the 2.2% of people visited first the pavilion of China then they moved to visit that of the UK, while the 1.92% of people visited first the pavilion of the UK and then that of China.

Table 5: Most visited sequences of pavilions (support $\geq 1.5\%$).

Rank	Sequence	Visitors
1	China→Japan	2.34%
2	China→UK	2.20%
3	UK→China	1.92%
4	Korea→China	1.90%
5	China→Russia	1.86%
6	Japan→China	1.83%
7	China→Korea	1.83%
8	Brasil→China	1.83%
9	UK→Japan	1.74%
10	China→USA	1.62%
11	Korea→Japan	1.55%

We studied the mobility flows to EXPO, analyzing which countries the visitors come from according to the Instagram data. In case of visitors coming from Italy, we also discovered the city and region of origin. According to our analysis, 81.7% of the Instagram posts were published by users coming from Italy, while 18.3% by users coming from other countries. Figure 10 shows the percentages of mobility flows from outside Italy to EXPO. It can be seen that the largest inflows originated from Spain and France (19.29% and 19.05%, respectively), followed by the UK (13.27%) and USA (10.85%).

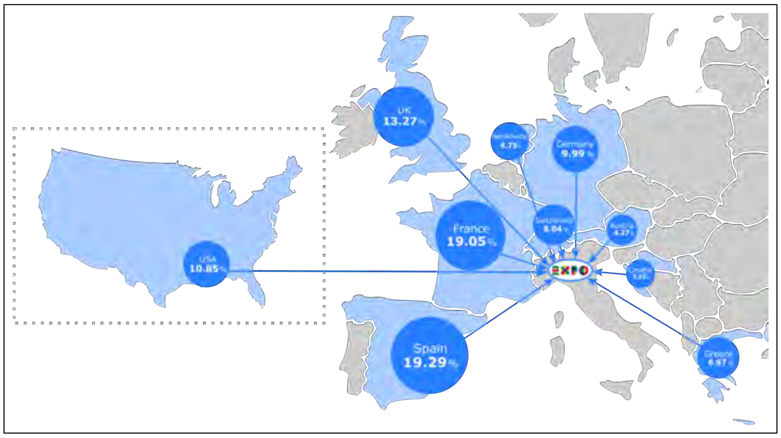
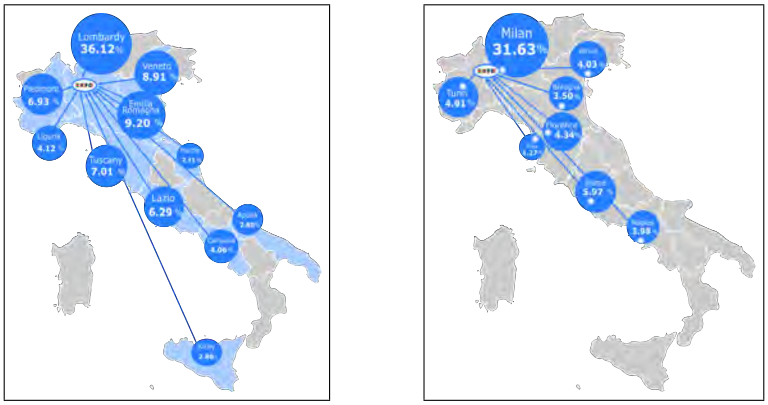


Figure 10: Main origins of foreign visitors.

In addition to the international mobility, we also focused our analysis to discover some insights about the mobility within Italian territory. Figure 11(a) shows the regions from which Italian visitors came. Only flows greater than 2% are reported. As shown in the figure, more than two-thirds of the total flow of Italian visitors to EXPO originated from five center-north regions: Lombardy, Emilia-Romagna, Veneto, Tuscany and Piedmont. In particular, 36.12% of the visitors came from Lombardy, which is the region where is located Milan, the city that hosted EXPO 2015. Figure 11(b) shows the main cities of origin of the Italian visitors. As expected, the greatest flow was registered from Milan (31.63%), followed by Rome (5.97%), Turin (4.91%) and Florence (4.34%).

Finally, we show some insights about the impact that EXPO had on the local territory. To do that, we analyzed data to discover the main flows of destination of foreign EXPO visitors to Italian regions and cities, in the days after their visit to EXPO. This is useful to understand the touristic impact of EXPO on the different parts of the country that hosted it. Figure 12(a) shows in which Italian regions the foreign Instagram users went in the days after their visit to EXPO. Only flows greater than 1% are reported. As shown in the figure, 63.24% of the foreigners who attended EXPO visited Lombardy (the region of Milan). Other regions with significant flows were Veneto (10.85%), Tuscany (8.18%) and Lazio (7.97%). Figure 12(b) presents the flows of destinations to the main Italian cities. As expected, the most visited city was Milan (59.06%),

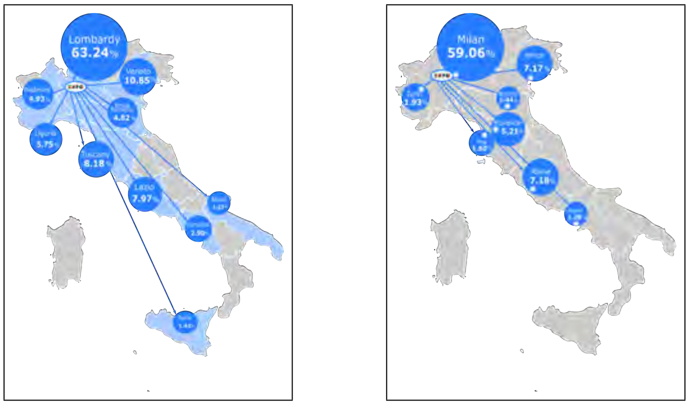


(a) Regions.

(b) Cities.

Figure 11: Main origins of Italian visitors.

followed by Rome (7.18%), Venice (7.17%) and Florence (5.21%). By looking at Figures 12(a) and 12(b) together, we can see that in some regions most of the flows was directed to their main cities. For example, in Lombardy most visitors went to Milan (59.06%, out of 63.24% registered in the whole region).



(a) Regions.

(b) Cities.

Figure 12: Local destinations of foreign visitors.

6. Related work

Several algorithms and techniques for mobility analysis of social media data have been proposed in the literature. An accurate survey of main contributions is presented in [21]. Despite the different algorithms and techniques introduced until today, no complete methodologies like the one we present here have been proposed for trajectory discovery in large-scale events. In this section we briefly review some of the most related research work in trajectory pattern mining from spatio-temporal data discussing differences and similarities with the methodology we designed.

A trajectory pattern mining algorithm for discovering mobility patterns, modeled as sequences of visited dense regions with travel time, is presented in [12]. The authors extended sequential pattern mining models to analyze moving objects, and evaluated the proposed approach over real data and synthetic benchmarks. In the application domains we consider, dense regions are already available, so it is not required to find them, however, the trajectory mining algorithm proposed in [12] could be integrated in the methodology we designed.

In [11] the 2012 Summer Olympics has been analyzed to study how a large event can influence the flow movement of urban population. Analyzing data from a large location-based social service (i.e., Foursquare), the authors created a supervised learning algorithm for predicting businesses thrive of local retailers by exploiting a combination of both geographic and mobility features. With respect to our work, this is less general since it is only specialized in providing a predictable algorithm for what businesses will increase their customers during a large event. Similar considerations can be made for [14], that presented a strategy for exploiting data collected from location-based social media, in order to forecast the area where a retail store may attract the maximum number of customers.

A Cloud-based platform for urban computing is proposed in [2], which allows users to execute workflow-based parallel tasks for discovering patterns and rules from trajectory data. The experimental evaluation, aimed at demonstrating scalability and efficiency properties of the Cloud framework used (i.e., DMCF [16]), has been performed on a real-world dataset concerning mobility of citizens within the Beijing urban area. This paper focuses mainly on mobility and for this reason it requires as input a trajectory dataset. Our work, on the other hand, is a methodology that can be applied to social data for finding the trajectories of individual users and extracting the most common routes.

The framework presented in [5] is devoted to spatio-temporal analysis of massive geo-referenced social media data, in which the activities of social media users are modeled as space-time trajectories. The aim of the work is different from ours, as that work provides a real-time framework for exploring people movements at multiple scales, instead we focus on infer mobility patterns of people related to an event.

A method for ranking trajectory patterns mined from geotagged photos is described in [20]. In particular, the authors proposed an algorithm exploiting relationships among users, locations and trajectories, to assign an importance score to the discovered trajectory patterns. This paper finds locations by clustering GPS photo coordinates with the Mean-shift algorithm [8] and extracts trajectories using the PrefixSpan algorithm [13]. The two mining algorithms proposed can be integrated in the methodology we designed.

There are different papers that focus on defining the geographical boundaries of the places-of-interests. For instance, in [9] a clustering algorithm has been proposed to discover local urban area dynamics (livehoods) of a city. In [3] a technique that exploits the indications contained in geotagged social media items is used to discover regions-of-interests with a high accuracy. In [4] an algorithm is proposed for detecting fine and accurate arbitrary shapes in order to discover meaningful landmarks and interesting places. All these three papers cover only the Step 2 of our methodology and may be considered as alternative algorithms for defining the geographical boundaries of the places-of-interests.

7. Conclusion

SMA4TD (Social Media Analysis for Trajectory Discovery) is a methodology for discovering behavior and mobility patterns of users attending large-scale public events. The methodology is composed of seven steps: *i*) identification of the set of events; *ii*) identification of places-of-interests where the events take place; *iii*) collection of geotagged items related to events and pre-processing; *iv*) identification of users who published at least one of the geotagged items; *v*) pre-processing and creation of the input dataset; *vi*) data analysis and trajectory mining; and *vii*) results visualization.

The methodology is validated through two case studies. The first one is an analysis of geotagged tweets for understanding the behavior of people attending the 2014 FIFA World Cup. The second one is a mobility pattern analysis on the Instagram users who visited EXPO 2015. In both cases, a very high correlation (Pearson coefficient 0.7-0.9) was measured between official attendee numbers and those produced by our analysis, which assess the effectiveness of the proposed methodology and confirm the reliability of the results.

The use of our methodology applied to advanced social network functionalities can be applied to the study of future events, for example helping event planning and organization, providing full access to a wide range of information that can be decisive for the monitoring and management of key services like transports, security, logistics, and others. Using the methodology discussed in this paper, large communities of people can be effectively analyzed to understand complex human behaviors and social dynamics. This new use of learning technology is very promising, as it provides critical information and high-quality knowledge that are fundamental for the growth of organization systems. Despite the fact that users send posts at irregular intervals and with many different reasons and motivations, the accurate analysis of a large number of posts can provide an effective base for learning people behavior and movements. Such learners are significant components of complex geospatial knowledge discovery systems.

Acknowledgement

This work was partly supported by project *DEEP - Data Enrichment for Engaging People* funded by PO FESR 2007-2013 Regione Sardegna.

[1] Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. VLDB '94. pp. 487–499.

[2] Altomare, A., Cesario, E., Comito, C., Marozzo, F., Talia, D., 2017. Trajectory pattern mining for urban computing in the cloud. IEEE Transactions on Parallel and Distributed Systems 28 (2), 586–599.

[3] Belcastro, L., Marozzo, F., Talia, D., Trunfio, P., October 2017. G-roi: Automatic region-of-interest detection driven by geotagged social media data. ACM Transactions on Knowledge Discovery from Data.

[4] Cai, G., Hio, C., Bermingham, L., Lee, K., Lee, I., 2014. Sequential pattern mining of geo-tagged photos with an arbitrary regions-of-interest detection method. Expert Systems with Applications 41 (7), 3514 – 3526.

[5] Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., Soltani, K., 2014. A scalable framework for spatiotemporal analysis of location-based social media data. CoRR abs/1409.2826.

[6] Cesario, E., Congedo, C., Marozzo, F., Riotta, G., Spada, A., Talia, D., Trunfio, P., Turri, C., July 2015. Following soccer fans from geotagged tweets at fifa world cup 2014. In: Proc. of the 2nd IEEE Conference on Spatial Data Mining and Geographical Knowledge Services. Fuzhou, China, pp. 33–38, iSBN 978-1- 4799-7748-2.

[7] Cesario, E., Iannazzo, A. R., Marozzo, F., Morello, F., Riotta, G., Spada, A., Talia, D., Trunfio, P., 18-22 July 2016. Analyzing social media data to discover mobility patterns at expo 2015: Methodology and results. In: The 2016 International Conference on High Performance Computing and Simulation (HPCS 2016). Innsbruck, Austria.

[8] Cheng, Y., 1995. Mean shift, mode seeking, and clustering. IEEE transactions on pattern analysis and machine intelligence 17 (8), 790–799.

[9] Cranshaw, J., Schwartz, R., Hong, J. I., Sadeh, N. M., 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. In: ICWSM.

[10] de Graaff, V., de By, R. A., van Keulen, M., Flokstra, J., 2013. Point of interest to region of interest conversion. In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. SIGSPATIAL'13. ACM, New York, NY, USA, pp. 388–391.
URL http://doi.acm.org/10.1145/2525314.2525442

[11] Georgiev, P., Noulas, A., Mascolo, C., 2014. Where businesses thrive: Predicting the impact of the olympic games on local retailers through location-based services data. CoRR abs/1403.7654.

[12] Giannotti, F., Nanni, M., Pinelli, F., Pedreschi, D., 2007. Trajectory pattern mining. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '07. ACM, New York, NY, USA, pp. 330–339.

[13] Han, J., Pei, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M., 2001. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: proceedings of the 17th international conference on data engineering. pp. 215–224.

[14] Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., Mascolo, C., 2013. Geo-spotting: Mining online location-based services for optimal retail store placement. CoRR abs/1306.1704.

[15] Maeda, J., 2006. The Laws of Simplicity. The MIT Press.

[16] Marozzo, F., Talia, D., Trunfio, P., 2015. Js4cloud: Script-based workflow programming for scalable data analysis on cloud platforms. Concurrency Computation 27 (17), 5214–5237.

[17] Talia, D., Trunfio, P., Marozzo, F., October 2015. Data Analysis in the Cloud. Elsevier.

[18] Tufte, E., 1990. Envisioning Information. Graphics Press, Cheshire, CT, USA.

[19] Tufte, E. R., 1986. The Visual Display of Quantitative Information. Graphics Press, Cheshire, CT, USA.

[20] Yin, Z., Cao, L., Han, J., Luo, J., Huang, T. S., 2011. Diversified trajectory pattern ranking in geo-tagged social media. In: SDM. SIAM, pp. 980–991.

[21] Zheng, Y., 2015. Trajectory data mining: An overview. ACM Transactions on Intelligent Systems and Technology (TIST) 6 (3), 29.