Analyzing Social Media Data to Discover Mobility Patterns at EXPO 2015: Methodology and Results

Eugenio Cesario
DtoK Lab Srl, Italy
ICAR-CNR, Italy
Email: cesario@icar.cnr.it

Andrea Raffaele Iannazzo
DtoK Lab Srl, Italy
Email: iannazzo@dtoklab.com

Fabrizio Marozzo
DtoK Lab Srl, Italy
DIMES, University of Calabria, Italy
Email: fmarozzo@dimes.unical.it

Fabrizio Morello
DtoK Lab Srl, Italy
Email: morello@dtoklab.com

Gianni Riotta
Princeton University, NJ, USA
Email: griotta@princeton.edu

Alessandra Spada
Alkemy Tech Srl, Italy
Email: alessandra.spada@alkemy.com

Domenico Talia
DtoK Lab Srl, Italy
DIMES, University of Calabria, Italy
Email: talia@dimes.unical.it

Paolo Trunfio
DtoK Lab Srl, Italy
DIMES, University of Calabria, Italy
Email: trunfio@dimes.unical.it

***Abstract*—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 presents a methodology and discusses results of a study aimed at discovering behavior and mobility patterns of 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 2015, 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. To cope with this large amount of data, the whole process - from data collection to data mining - was implemented on a high-performance cloud platform that provided the necessary storage and compute resources. 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. A strong correlation (Pearson coefficient 0.7) was measured between official visitor numbers and the visit trends produced by our analysis, which assessed the effectiveness of the proposed methodology and confirmed the reliability of results.**

I. 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 data 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 understanding the behavior of groups of people or the dynamics of public opinion [1]. 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 georeferenced data that embed extensive knowledge about human dynamics and mobility behaviors. In the latest years, there has been a growing interest in the extraction of trajectories from geotagged social data using trajectory mining techniques [2].

This paper describes a methodology and the results of a study aimed at discovering behavior and mobility patterns of Instagram users attending EXPO 2015, the Universal Exposition held under the theme "Feeding the Planet, Energy for Life," which was hosted in Milan, Italy, from May 1st to October 31st, 2015. The data source is composed of the geotagged posts published by the Instagram users who visited EXPO. Instagram is a sharply rising service for social networking, mobile photo-sharing and video-sharing, which allows users, after taking a photo or video, to share it, possibly on other platforms such as Facebook, Twitter, Tumblr and Flickr. Each media (picture or video) can be accompanied by text, hashtags, mentions to other users, and a location field that allows to geographically tag the photo or video.

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 from one month before to one month after their visit to EXPO. The main goal of the analysis was to discover the mobility patterns of the visitors inside the EXPO pavillons and outside EXPO. In particular, users' positions were tracked before, during and after the visits to the exhibition area. To this end, data were processed through algorithms and methodologies for associative analysis (associative pattern mining) and sequential analysis (sequential pattern mining). Associative analysis was used to discover the most frequent sets of visited pavilions, while sequential pattern analysis allowed us to discover the origin of visitors, and the

main flows of destination of foreign visitors towards Italian regions and cities in the days after their visit to EXPO. To cope with the large amount of data and the complexity of associative and sequential pattern mining tasks, the whole data collection and mining process was implemented on a high-performance cloud platform - Microsoft Azure - that provided the necessary scalability to storage and compute resources. A strong correlation (Pearson coefficient 0.7) was measured between official visitor numbers and the visit trends obtained by our analysis, which confirms the reliability of the results as well as the effectiveness of the methodology.

The plan of the paper is as follows. Section II provides some definitions and the objectives of the study. Section III describes our methodology for data collection and analysis. Section IV presents the main results of the analysis. Section V discusses related work. Section VI concludes the paper.

II. Definitions and objectives

The study is aimed at discovering behavior and mobility patterns of Instagram users visiting the exhibition sections of EXPO 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.

A. Preliminary definitions

We provide here some definitions of the main concepts that will be used in the remainder of the paper.

- IP: Set of georeferenced Instagram posts published inside EXPO in the period May 1st - October 31st, 2015.
- VIS: Set of EXPO visitors who published at least one of the Instagram posts in IP.
- OP: Set of georeferenced Instagram posts published outside EXPO by all the users in VIS, from one month before to one month after their visit to EXPO.
- pt_i: An Instagram post in IP or OP, characterized by the following properties: *user* who posted pt_i, *location* (user-defined name of the place), *latitude* and *longitude* (coordinates of the place from which pt_i was sent), *ts* (timestamp of the post) and *text*.
- pn_i: *Position name* of pt_i, a string that uniquely represents the place from which pt_i was published, which is associated to pt_i based on its location, coordinates or text properties.
- CN: Set containing the position name of all the countries in the world but Italy ("USA", "China", etc.).
- RN: Set containing the position name of all the Italian regions ("Lazio", "Lombardy", etc.).
- MN: Set containing the position name of the main Italian cities ("Rome", "Milan", etc.).
- PAV: Set of EXPO pavilions, including both country pavilions and organization/company ones.
- PN: Set containing the position name of every country pavilion ("USA Pavilion", "China Pavilion", etc.) and of every organization/company pavilion ("UN Pavilion", etc.) in PAV.
- gn: Generic position name ("EXPO 2015") for all the places that are inside EXPO but are not associated with any pavilion in PAV.
- P_{ij} = $<pn_{ij}$,$ts_{ij}>$ is a *geo-temporal descriptor* containing *position name* and *timestamp* of the j-th post published by a user $v_i \in VIS$, considering all the posts present in IP and OP.
- T = $\{T_1, T_2, ..., T_m\}$, where m is the size of VIS, is a dataset containing the geo-temporal descriptors of all the georeferenced posts published by the Instagram users who visited EXPO. In particular, the i-th tuple T_i of T is a pair:

$$< v_i, \{P_{i1}, P_{i2}, ..., P_{ik}\} >$$

where v_i is a user in VIS who published k of the posts present in IP and OP.

B. Objectives of the analysis

The main goals of the study are as follows.

1) *Analysis of visit trends.* We studied how the number of visitors changed over time. The number of visitors in a day d_i is the number of tuples in T containing at least one geo-temporal descriptor with position name in $PN \cup gn$ and timestamp in d_i. The results are compared with the official visitor numbers to measure the correlation between them, in order to assess the obtained results.
2) *Discovery of the most visited pavilions.* We analyzed the observed data to learn the list of pavilions that have been most visited by the Instagram users. The number of visitors to a pavilion identified by a position name pn in PN is given by the number of tuples in T containing at least one geo-temporal descriptor with position name pn.
3) *Discovery of the most frequent sets of visited pavilions.* We extracted the sets of pavilions that are most frequently visited together by the Instagram users. The problem is modeled as an associative pattern mining instance, where a frequent associative pattern *fap* with support s,

$$fap = \{pav_i, pav_j, \ldots, pav_k\}\ (s)$$

is a subset of the pavilions in PAV, and s is the percentage of tuples in T containing the position names of *fap*. The final goal of this analysis is to discover all the sets of pavilions whose support s is equal to or greater than a given minimum support count s_{min}, i.e. $s \geq s_{min}$.
4) *Discovery of the origin of visitors.* We studied the mobility flows to EXPO, evaluating which countries visitors came from. In case of visitors coming from Italy, we also discovered city and region of origin. This task

is formulated as a sequential pattern mining problem, where a frequent sequential pattern *fsp* with support s,

$$fsp =< site_1, site_2 > \ (s)$$

is a sequence with $site_1 \in CN \cup RN \cup MN$ and $site_2 \in PN \cup gn$, and s is the percentage of tuples in T containing *fsp*. Therefore, the goal is to discover all the sequences whose support s is equal to or greater than a given minimum support count s_{min}, i.e. $s \geq s_{min}$.

5) *Analysis of the impact on local territory.* We evaluated the main flows of destination of foreign EXPO visitors to Italian regions and cities, in the days after their visit to EXPO. Also this task is modeled as a sequential pattern mining problem, where the first site in the sequence represents a location inside EXPO, and the second site represents an Italian region or city.

III. Methodology

The whole data collection and analysis process is composed of five steps: A) identification of the Instagram locations inside EXPO; B) collection of Instagram posts inside EXPO and identification of visitors; C) collection of Instagram posts published by visitors outside EXPO; D) creation of the input dataset; E) trajectory mining. A description of the steps is given in the remainder of the section. For the Reader's convenience, Table I lists the main symbols that will be used throughout the section.

TABLE I
Main symbols used throughout the section.

Name	Meaning
IP	Instagram posts published by visitors inside EXPO
VIS	Instagram users who published at least one post in IP
OP	Instagram posts published outside EXPO by users in VIS
CN	Position names of all the countries in the world but Italy
RN	Position names of all the Italian regions
MN	Position names of the main Italian cities
PN	Position names of all the EXPO pavilions

A. Identification of the Instagram locations inside EXPO

The goal of this step is to identify the set EL of Instagram locations inside the EXPO area. The Instagram API provides a service that, given a geographical point c and a distance r, returns the set of Instagram locations falling in the circle centered at c of radius r. The set returned by the service may not contain all the locations in the specified circle, because there is a limit in the number of elements returned by the Instagram API. To overcome this limitation, we proceeded as follows. We partitioned the EXPO area in a grid, as shown in Figure 1.

Let $C = \{c_1, c_2, ..., c_n\}$ be the set of grid intersection points laying within the EXPO area. For each c_i in C, we asked the Instagram API to return the set L_i of locations within a given distance r from c_i. The set EL of Instagram locations inside EXPO is the union of all L_i. EL was then cleaned to filter out the locations with coordinates outside the EXPO area, which were returned starting from grid points at the borders of the area. At the end of data collection and cleaning, EL contained 2,890 locations. Every l_i in EL is a tuple <*id*, *name*, *latitude*, *longitude*>, where *latitude* and *longitude* are the geographical coordinates of l_i. Note that if two locations have the same *id*, they have also the same coordinates, but can have different names.

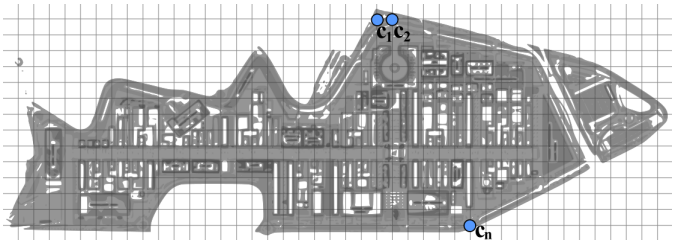


Fig. 1. Partitioning the EXPO area in a grid.

B. Collection of Instagram posts inside EXPO and identification of visitors

The goal of this step is to collect the set IP of Instagram posts published inside the EXPO area in the period May 1st - October 31st, 2015, and to identify the set VIS of Instagram users who visited EXPO. To this end, given the set EL of Instagram locations inside EXPO identified in the previous step, we proceeded as follows. For each l_i in EL, we submitted to the Instagram API a query that, given l_i and the period specified above, returns the set PT_i of Instagram posts published from l_i during the period. By making the union of all the PT_i, we obtain the set IP of all the Instagram posts published inside EXPO. Then, we extracted the set VIS of the distinct users who published at least one post in IP. The number of posts in IP is equal to 570,973, while the number of users in VIS is equal to 238,196, with an average of 2.4 posts per user.

C. Collection of Instagram posts published by visitors outside EXPO

After having identified the Instagram users VIS who visited EXPO, and having collected their posts IP during the visits, we performed this step to collect the set OP of Instagram posts published by the users in VIS in the timeframe from one month before to one month after their visit at EXPO. We created OP by retrieving, for each user v_i in VIS who visited EXPO in a day d_i, all the georeferenced posts published by v_i in the period starting one month before d_i and ending one month after d_i. We made this operation using a service of the Instagram API that allows retrieving all the posts of a given user in a specified period. The total number of posts in OP resulted in about 2.63 millions.

D. Creation of the input dataset

The goal of this fourth step is the creation of the dataset T used as input for the analysis. We recall that the i-th tuple T_i of T is a pair $<v_i, \{P_{i1}, P_{i2}, ..., P_{ik}\}>$, where P_{ij} contains

position name and *timestamp* of the j-th post published by user v_i. The position name associated to a post depends on whether it belongs to IP or to OP.

To assign a position name to the posts in IP, we created and used two dictionaries: First Level Dictionary (FLD) and Second Level Dictionary (SLD). FLD receives the Instagram location of a post and returns the corresponding position name in PN, if a mapping is found. To this end, FLD contains a set of tuples $<term_a, term_b>$, where $term_a$ is an Instagram location in EL, and $term_b$ is a position name in PN. There can be multiple tuples having the same $term_b$ but different $term_a$. For instance, the FLD tuples <"US Pavilion","USA Pavilion"> and <"Padiglione USA","USA Pavilion"> map the Instagram locations "US Pavilion" and "Padiglione USA" to the same position name "USA Pavilion". In practice, FLD allows us to convert different versions of the same location (as defined by multiple Instagram users) to a unique position name.

Using FLD, we were able to assign a position name to 160,307 posts, which represent the 28.08% of the posts in IP. The remaining posts were passed to SLD, which receives the text field of a post and returns the corresponding position name in PN, if a mapping is found. SLD contains a set of tuples $<term_a, term_b>$, where $term_a$ is a word or phrase that may be found in the text of a post, and $term_b$ is a position name in PN. An example is <"Enjoi #Emirates #Expo","UAE Pavilion">. Using SLD, we were able to assign a position name to 195,699 posts, which represent an additional 34.27% of the posts in IP. After having used FLD and SLD, 214,967 posts (corresponding to the 37.65% of all the posts in IP) did not receive a specific position name inside EXPO and, therefore, were labeled with the generic position name gn "EXPO 2015".

To assign a position name to the posts in OP, we considered their geographical coordinates (recall that all the posts in OP are georeferenced). If the coordinates of a post are within the borders of a country different from Italy, then its position name is the name of that country as listed in CN. If the coordinates are within the borders of a main Italian city, the position name is the name of that city as listed in MN. Finally, if the coordinates are in Italy but not in a main city, the position name is the name of the region, as listed in RN, where those coordinates lie. Note that we went into finer granularity on Italian places to better evaluate the mobility impact of foreign visitors on the different locations of the country that hosted EXPO.

E. Trajectory mining

Finally, after the creation of the dataset, it was analyzed through algorithms and techniques for associative analysis (associative pattern mining) and sequential analysis (sequential pattern mining).

Associative analysis algorithms have the goal of discovering (inside data) the values of attributes 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 $A \implies B$ (if event A occurs, then it is likely that also event B occurs). Applied to the EXPO data, this type of analysis has allowed us to extract the sets of pavilion that are most frequently visited together by the Instagram users. We performed this task using FP-Growth, an optimized frequent pattern mining algorithm that exploits a special data structure named FP-tree [3].

Algorithms for sequence analysis 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 the chronological order in which the values appear in the data. As part of the analysis that was carried out on the EXPO data, it was possible to discover the origin of the Instagram users who visited EXPO, and their destination to Italian region and/or cities after their visit. This task has been performed using SPADE (Sequential PAttern Discovery using Equivalence classes), a frequent sequence mining algorithm proposed in [4].

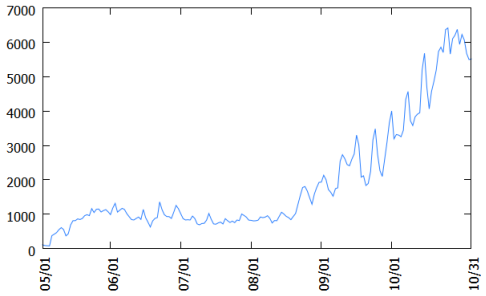


Fig. 2. Number of daily visitors.

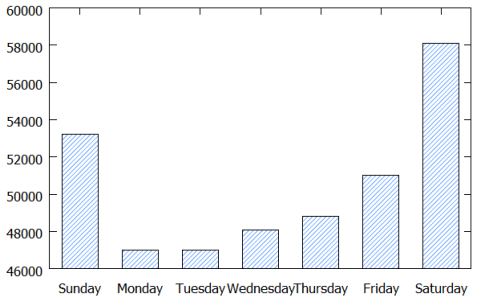


Fig. 3. Number of visitors per week day.

IV. Results

This section presents the main results of the study, divided in five parts: A) visit trends; B) most visited pavilions; C) most frequent sets of visited pavilions; D) origin of visitors; E) impact on local territory.

A. Visit trends

The goal of this analysis is to find how the number of visitors changed over time. Figure 2 shows the number of daily visits to EXPO of Instagram users. The trends are quite evident: initially (May and June) the visitors are relatively few;

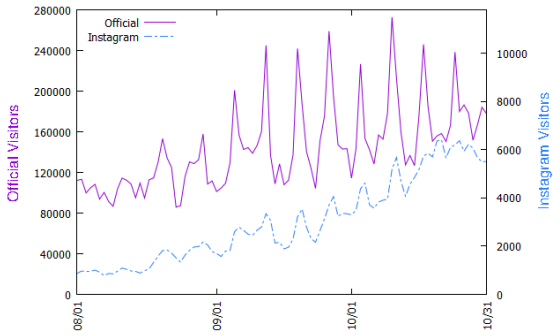


Fig. 4. Comparison with the official visitor numbers.

then, they grow significantly during the months of September and October. Figure 3 aggregates the visitor numbers based on the week day. The results clearly show that during the week-end days there is a peak of visits, with the highest number of visitors registered on Saturdays.

Figure 4 presents a comparison between trends and numbers of the Instagram visitors we tracked, and the official visitors published in the EXPO website for the period August 1st - October 31st (official numbers have not been published for the period before August). 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. By looking at the trends in the figure, it can be noted a strong correlation (Pearson coefficient 0.7) between official visitor numbers and those obtained from our analysis, which confirms the reliability of the results we obtained.

B. Most visited pavilions

We analyzed collected data to find the list of pavilions that have been most visited by Instagram users. Table II presents those results, restricted to 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%.

C. Most frequent sets of visited pavilions

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 reported in Table III. 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>.

TABLE II
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 III
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%

D. Origin of visitors

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 5 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 6(a) shows the regions from which the 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.

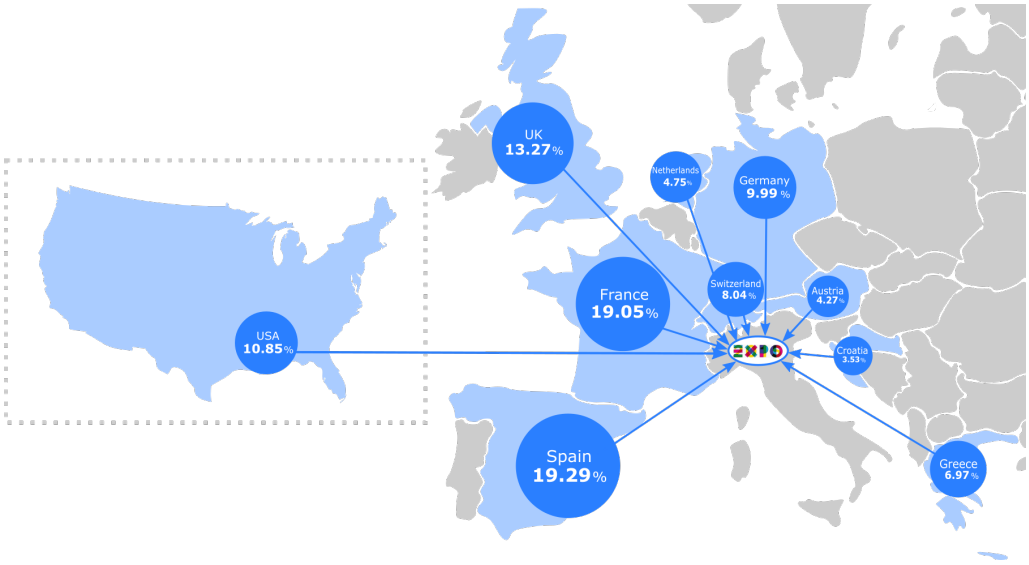


Fig. 5. Main origins of foreign visitors.

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 6(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%).

E. Impact on local territory

Finally, 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.

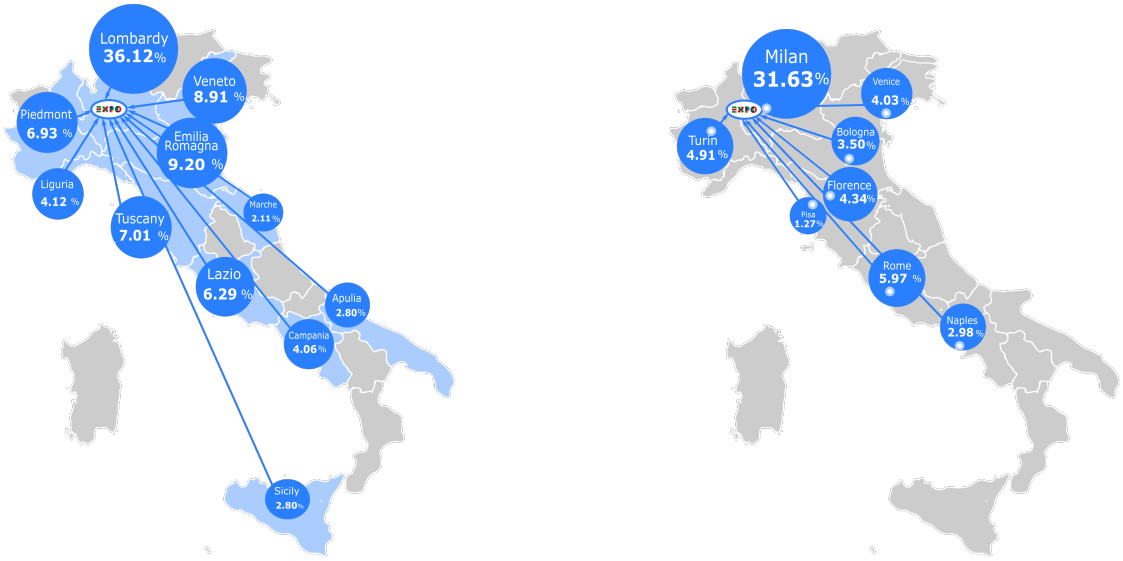
Figure 7(a) shows in which Italian regions the foreign Instagram users went in the days after their visit to EXPO. Only flows greater than 1% are reported. As shown in the figure, 63.24% of the foreigners who attended EXPO visited Lombardy (the region of Milan). Other regions with significant flows were Veneto (10.85%), Tuscany (8.18%) and Lazio (7.97%). Figure 7(b) presents the flows of destinations to the main Italian cities. As expected, the most visited city was Milan (59.06%), followed by Rome (7.18%), Venice (7.17%) and Florence (5.21%). By looking at Figures 7(a) and 7(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).

V. Related work

Several approaches concerning mobility analysis from social network data have been proposed in literature [2]. Some of them are focused on trajectory pattern mining from spatio-temporal data, while others are more related to the analysis of geotagged photo data. In this section we will briefly review some of the most representative researches in both areas.

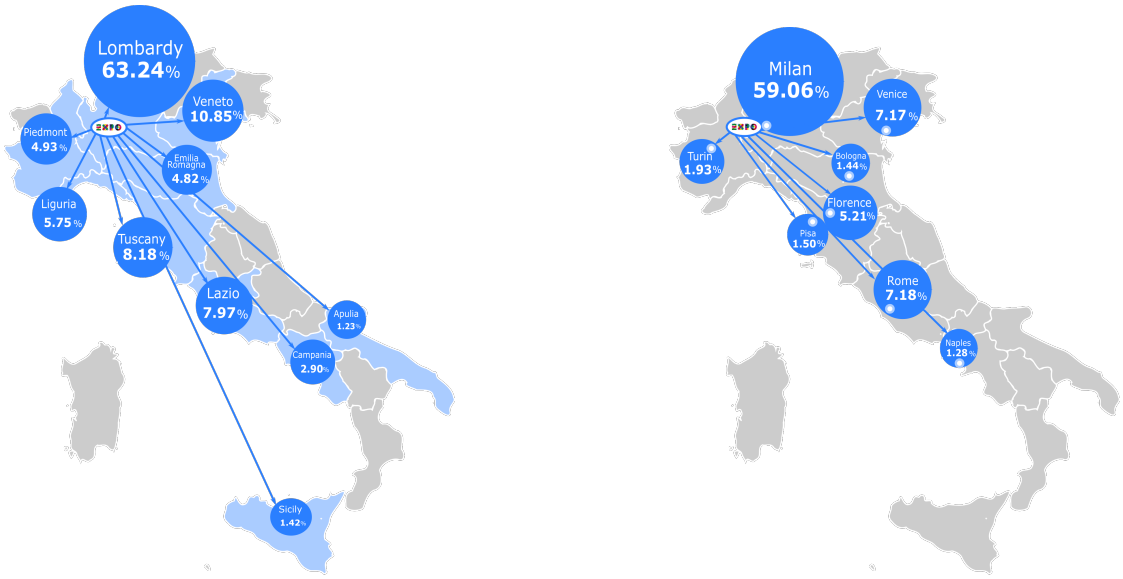
A trajectory pattern mining algorithm to discover mobility patterns, modeled as sequences of visited dense regions with travel time, is proposed in [5]. The authors extended sequential pattern mining models to analyze moving objects, and evaluated the proposed approach over real data and synthetic benchmarks. In [6], a clustering algorithm is proposed to discover local urban area dynamics. By analyzing social media data of residents in urban areas, the algorithm discovers the local assets of a city, such as municipal borders, demographics, development and geographic resources. Reference [7] describes the analysis of geotagged tweets to discover the most frequent movements of fans attending the 2014 FIFA World Cup. By adopting a trajectory pattern mining methodology, several results in terms of number of matches attended by groups of fans, clusters of most attended matches and most frequented stadiums, are reported. In [8] is studied how large events, e.g. Olympic Games, could influence the flow movement of urban population. Analyzing data from a large location-based social service, the authors created a supervised learning algorithm, exploiting a combination of both geographic and mobility features, for predicting businesses thrive of local retailers during large events. The framework described in [9] is aimed at performing social data analysis for smart cities, in particular to support smart mobility tasks. A Cloud-based platform for urban computing is proposed in [10], [11], 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 framework, has been performed on a real-world dataset concerning mobility of citizens within the Beijing urban area. The framework presented in [12] is devoted to spatio-temporal analysis of massive georeferenced social media data, in which the activities of social media users are modeled as space-time trajectories.

Another challenging task faced in literature is the analysis



(a) (b)

Fig. 6. Main origins of Italian visitors: (a) Regions; (b) Cities.



(a) (b)

Fig. 7. Local destinations of foreign visitors: (a) Regions; (b) Cities.

of geotagged photos published by social media systems, to discover Regions-of-Interest (RoIs) and frequent trajectory patterns for travel recommendation. In [13], tourist photos as held by Flickr are exploited to estimate the probability that a tourist will be visiting a landmark. The system proposed in [14] exploits a trip model based on canonical mobility sequences among touristic place clusters. Reference [15] describes a system that interactively helps users in travel routes planning, by exploiting the popular destinations to visit, the visiting order of destinations, the visiting time, etc. In [16], the authors exploit a Markov chain model to discover tourist routes among different RoIs and propose also an algorithm for topological analysis of personalized travel routes for different tourists. References [17] and [18] present two approaches for sequential pattern mining from Flickr datasets. The algorithm proposed in [17] is aimed at detecting fine and accurate arbitrary RoI shapes that allow to discover meaningful landmarks and interesting places, while the approach described in [18] discovers RoIs using both space and time simultaneously, and thus allows not only to find major patterns, but also to see at what times these patterns are occurring. In [19] a novel neighbourhood detection algorithm from geotagged data is proposed. As a

case study, the authors created a recommendation system where neighbourhoods are suggested to Twitter users based on textual profile information. The algorithm proposed in [20] is aimed at discovering associative RoI patterns from geotagged photos, by performing RoI clustering and association rules mining. Reference [21] presents a strategy to exploit 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 method for ranking trajectory patterns mined from geotagged photos is described in [22]. In particular, the authors proposed an algorithm exploiting relationships among users, locations and trajectories, to assign an importance score to the discovered trajectory patterns. In [23] an algorithm is proposed to classify (i.e., to discover the semantic type of) a place in a city (e.g. schools, hospitals, train stations and restaurants), based on data collected from Flickr photos and tweets posted in the urban area. Finally, in [24] data collected from Foursquare are exploited to investigate the properties and the growth patterns of urban place networks formed by the check-in patterns across a set of 100 cities around the globe.

VI. Conclusion

This paper described a methodology and main results of an experimental study aimed at discovering behavior and mobility patterns of Instagram users visiting EXPO 2015. The study demonstrated how the huge amount of data posted by social media users attending a popular event can be analyzed to infer patterns and trends about people behaviors related to that event on a very large scale. In particular, when social media posts are tagged with geographical coordinates or other information that allows identifying the positions of users, it is possible to perform mobility pattern analysis using trajectory mining techniques.

We collected and analyzed the geotagged posts published by about 238,000 Instagram users who visited EXPO. The analysis allowed us to discover how the number of visitors changed over time, identify the most frequent sets of visited pavilions, which countries the visitors came from, and the main flows of destination of foreign visitors to Italian regions and cities after their visit to EXPO. A strong correlation was registered between official visitor numbers and the visit trends produced by our analysis, thus confirming the significance of input data, the reliability of obtained results as well as the effectiveness of the methodology.

Acknowledgments

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

References

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

[2] Y. Zheng, "Trajectory data mining: An overview," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, p. 29, 2015.

[3] J. Han, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., 2005.

[4] M. J. Zaki, "Spade: An efficient algorithm for mining frequent sequences," *Machine Learning*, vol. 42, no. 1, pp. 31–60.

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

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

[7] E. Cesario, C. Congedo, F. Marozzo, G. Riotta, A. Spada, D. Talia, P. Trunfio, and C. Turri, "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, July 2015, pp. 33–38, iSBN 978-1- 4799-7748-2.

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

[9] L. You, G. Motta, D. Sacco, and T. Ma, "Social data analysis framework in cloud and mobility analyzer for smarter cities," in *Service Operations and Logistics, and Informatics (SOLI), 2014 IEEE International Conference on*, Oct 2014, pp. 96–101.

[10] E. Cesario, C. Comito, and D. Talia, "Towards a cloud-based framework for urban computing, the trajectory analysis case," in *2013 International Conference on Cloud and Green Computing, Karlsruhe, Germany, September 30 - October 2, 2013*, 2013, pp. 16–23.

[11] A. Altomare, E. Cesario, C. Comito, F. Marozzo, and D. Talia, "Trajectory pattern mining over a cloud-based framework for urban computing," in *Proc. of the 16th International Conference on High Performance Computing and Communications (HPCC 2014)*. Paris, France: IEEE, 2014, pp. 367–374.

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

[13] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura, "Travel route recommendation using geotags in photo sharing sites," in *Proc. of the 19th ACM Int. Conf. on Information and Knowledge Management*, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 579–588.

[14] K. Okuyama and K. Yanai, "A travel planning system based on travel trajectories extracted from a large number of geotagged photos on the web," in *The era of interactive media*. Springer, 2013, pp. 657–670.

[15] X. Lu, C. Wang, J.-M. Yang, Y. Pang, and L. Zhang, "Photo2trip: generating travel routes from geo-tagged photos for trip planning," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 143–152.

[16] Y.-T. Zheng, Z.-J. Zha, and T.-S. Chua, "Mining travel patterns from geotagged photos," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, pp. 56:1–56:18, May 2012.

[17] G. Cai, C. Hio, L. Bermingham, K. Lee, and I. Lee, "Sequential pattern mining of geo-tagged photos with an arbitrary regions-of-interest detection method," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3514 – 3526, 2014.

[18] L. Bermingham and I. Lee, "Spatio-temporal sequential pattern mining for tourism sciences," *Procedia Computer Science*, vol. 29, no. 0, pp. 379 – 389, 2014, 2014 International Conference on Computational Science.

[19] A. X. Zhang, A. Noulas, S. Scellato, and C. Mascolo, "Hoodsquare: Modeling and recommending neighborhoods in location-based social networks," *CoRR*, vol. abs/1308.3657, 2013.

[20] I. Lee, G. Cai, and K. Lee, "Exploration of geo-tagged photos through data mining approaches," *Expert Systems with Applications*, vol. 41, no. 2, pp. 397 – 405, 2014.

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

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

[23] S. Van Canneyt, S. Schockaert, and B. Dhoedt, "Discovering and characterizing places of interest using flickr and twitter," *Hospitality, Travel, and Tourism: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*, p. 393, 2014.

[24] A. Noulas, B. Shaw, R. Lambiotte, and C. Mascolo, "Topological properties and temporal dynamics of place networks in urban environments," *CoRR*, vol. abs/1502.07979, 2015. [Online]. Available: http://arxiv.org/abs/1502.07979