Online Social Networks and Media 3–4 (2017) 49–62

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

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*

ARTICLE INFO

Article history:
Received 13 September 2017
Revised 16 October 2017
Accepted 18 October 2017
Available online 10 November 2017

Keywords:
Mobility patterns
Urban computing
Trajectory mining
Social network
Social media posts
Geotagged data

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.

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 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 georeferenced 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

[*] Corresponding author.
E-mail addresses: eugenio.cesario@icar.cnr.it (E. Cesario), fmarozzo@dimes.unical.it (F. Marozzo), talia@dimes.unical.it (D. Talia), trunfio@dimes.unical.it (P. Trunfio).

https://doi.org/10.1016/j.osnem.2017.10.002
2468-6964/© 2017 Elsevier B.V. All rights reserved.

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. Sections 4 and 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 $\mathcal{P} = \{p_1, p_2, \ldots\}$ 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 [12].

Let $\mathcal{E} = \{e_1, e_2, \ldots\}$ 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 $\mathcal{G} = \{g_1, g_2, \ldots\}$ 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 $\mathcal{U} = \{u_1, u_2, \ldots\}$ 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.