Information Sciences 340–341 (2016) 122–143

Contents lists available at ScienceDirect

Information Sciences

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

User-driven geo-temporal density-based exploration of periodic and not periodic events reported in social networks

Paolo Arcaini[a,*], Gloria Bordogna[b], Dino Ienco[c,d], Simone Sterlacchini[e]

[a] Charles University in Prague, Faculty of Mathematics and Physics – Czech Republic
[b] CNR IREA, Milano – Italy
[c] IRSTEA, UMR TETIS, Montpellier – France
[d] LIRMM, Montpellier – France
[e] CNR IDPA, Milano – Italy

ARTICLE INFO

Article history:
Received 11 April 2014
Revised 25 June 2015
Accepted 1 January 2016
Available online 11 January 2016

Keywords:
Crawling social networks
User Query
Density-based Clustering
Geo-temporal proximity measure

ABSTRACT

In this paper we propose a procedure consisting of a first collection phase of social network messages, a subsequent user query selection, and finally a clustering phase, defined by extending the density-based DBSCAN algorithm, for performing a geographic and temporal exploration of a collection of items, in order to reveal and map their latent spatio-temporal structure. Specifically, both several geo-temporal distance measures and a density-based geo-temporal clustering algorithm are proposed. The approach can be applied to social messages containing an explicit geographic and temporal location. The algorithm usage is exemplified to identify geographic regions where many geotagged Twitter messages about an event of interest have been created, possibly in the same time period in the case of non-periodic events (aperiodic events), or at regular timestamps in the case of periodic events. This allows discovering the spatio-temporal periodic and aperiodic characteristics of events occurring in specific geographic areas, and thus increasing the awareness of decision makers who are in charge of territorial planning. Several case studies are used to illustrate the proposed procedure.

1. Introduction

Web 2.0 applications, such as the social platforms Facebook, Twitter, Foursquare, LinkedIn, provide nowadays citizens with a direct means to spread information in the form of textual messages, images, videos, links and so on, to their communities about the most diverse topics. These messages can eventually report information and comments about events the authors have been testimonies of (e.g., natural disasters, traffic jams, protests, comments on daily news, workshops, call for jobs), which have spatio-temporal contextual information, explicitly stating where on Earth (geotags) and when (timestamp) the messages were created. These metadata (geotags and timestamp) are embedded within the content of the message, so that the analysis of an event of interest can be performed by applying a space–time classification.

Analyzing the latent spatio-temporal structure of a huge amount of messages, reporting information or comments on an event in a social network, can aid to discover the spatio-temporal characteristic of the event, which may be either *recurring* (i.e., *periodic*) or *non-periodic*, (i.e., *aperiodic*). An example of recurring event can be the occurrence of traffic jams

* Corresponding author. Tel.: +420 221 914 285; fax: +420 221 914 323.
E-mail address: arcaini@d3s.mff.cuni.cz (P. Arcaini).

http://dx.doi.org/10.1016/j.ins.2016.01.014
0020-0255/

in a specific place at specific time periods of the day; such periodic events can be expressed linguistically by a phrase such as "*every morning from 8 a.m. to 9 a.m. in Milan Central station there is traffic jam*"; such information can be used by the administration for planning alternative routes for public transports in that time interval. Another example of regularly recurring event is the presence of a huge number of tourists in a region in specific periods of the year; identifying such periods and interested areas can be useful to plan social events and provide tourist facilities. Conversely, an example of aperiodic event reported in messages can be related to the occurrence of an extraordinary natural disaster like the typhoon Haiyan in Philippines in 2013, whose analysis can benefit from the exploitation of the information contained in messages, to better understand the human perception and the consequences of the event. Another example of aperiodic long term event could be a social/political crisis, where the long term analysis of social networks can improve the awareness of the impact on the conditions of the populations and reasons of outbreaking of the crisis in specific geographic areas.

Providing means to explore the unknown spatio-temporal distribution and density of messages reporting information on events of interest can increase our ability to make predictions in several contexts about similar events that may occur in the future, and may support territorial and social planning and analysis.

In this paper, we propose a two steps procedure for analyzing events of interest to a user reported in social networks: firstly, messages about events of possible interest are collected based on crawling the social networks; in the second step, the user can specify some criteria in order to drive an original geo-temporal analysis of the messages dealing with an event of interest to verify an "a priori" hypothesis. In this second step, first a user queries the collection to filter messages about an event of interest that he/she wants to study, and finally a density-based clustering defined in the paper is applied for performing a geographic and temporal exploration of the filtered messages in order to reveal and map their latent spatio-temporal structure.

Our proposal is innovative since it is flexible for several aspects: it is user-driven, and allows making several geographic and temporal explorations to discover periodic and aperiodic events, at local or global scale.

After the seminal work on the evolution of topics in the news sphere [19], many approaches have been proposed which allow us to explore the contents of the messages created within social media such as Twitter, in order to analyze the polarity about some themes [8,22], and to identify the initiators and followers by applying in and out link analysis [33].

Other approaches have also considered the problem of geographic or temporal analysis of tweets, like in [16] where the authors' goal is to characterize the geographical spread of a communication network, in [34] where the authors examine information spread in the social network and across geographic regions, in [26] and [32] where the authors analyze the temporal spread and evolution of news stories, and in [17] where the temporal analysis of retweets is studied. Very few approaches have considered the problem of geographic and temporal analysis, like in [2] where temporal and spatial analyses are performed investigating the time-evolving properties of tweets and the geospatial characteristics of highly popular topics, or in [15], where the authors study the geographic distribution and propagation of tweets hashtags and the peak characteristics of their temporal distribution. These approaches mainly explore the geographic and the temporal correlations with contents of tweets separately, but do not apply geographic-temporal analysis.

For example, *Twitris* is an interactive web application that maps temporal summaries of tweets about a number of selected critical geographic natural and social events, such as Oklahoma tornado, India Floods, and "Occupy Wall Street" protests. The user can select the time period and one of the listed events to see the summary of the tweets about it within the time period, but does not apply spatial analysis of the tweets. Conversely, in [3] the spatio-temporal dynamics of tweets about the same topic are explored with a fixed timestamp of 1 h.

On the other side, there are crowdsourcing platforms accessible over the Internet that allow citizens to freely create messages of events they have been testimonies of, in the form of Volunteered Geographic Information (VGI) [7,12–14]: such applications perform some spatial and temporal analysis of the messages [1,5,21]. For example, the *Ushahidi* application [23] allows applying spatial clustering depending on the visualization scale of the map to group messages close in space, but completely disregards their content and timestamp. It also allows the temporal tracking of the frequency of VGI messages that deal with a given content, identified by a term within the textual message. This allows tracking the temporal evolution of a topic "popularity", irrespective of the geographic location.

In the recent approach in [10], a space–time scan statistics method to analyze tweets is proposed. It looks for clusters within the dataset across both space and time, regardless of the tweet content. By this approach, it is expected that clusters of tweets will emerge during spatio-temporal relevant events, as people will tweet more than expected in order to describe the event and spread information. The authors apply this approach to identify a disaster in London. Nevertheless, when different events occur in the same spatio-temporal region, this approach cannot tell them apart.

The objective of our proposal is defining a flexible user-centered exploratory framework for performing both geographic, temporal, and geo-temporal analysis of a collection of items having an explicit geo-temporal reference. The study of the spatio-temporal pattern of geotagged tweets has been recognized as providing important information for various applications, such as urban science, location-based services, targeted advertising, content delivery networks, and social media research [15].

The originality of our proposal is the flexibility of the approach that allows a user to specify both **what** is the event of interest he/she wants to explore, and **how** he/she intends to explore it, by choosing a temporal, geographic, or geo-temporal analysis. Specifically, the proposal permits to identify both geographic events occurring in specific regions (but not characterized by a specific timestamp such as social and economic crises during a long time period) and temporal events