Expert Systems With Applications 95 (2018) 32–42

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Discovering geo-dependent stories by combining density-based clustering and thread-based aggregation techniques

Héctor Cerezo-Costas[a,*], Ana Fernández-Vilas[b], Manuela Martín-Vicente[a], Rebeca P. Díaz-Redondo[b]

[a] *Gradiant, Edificio CITEXVI, local 14, Universidade de Vigo, Spain*
[b] *Information & Computing Lab., AtlantTIC Research Center, School of Telecommunications Engineering, Universidade de Vigo, Spain*

ARTICLE INFO

Article history:
Received 11 July 2017
Revised 8 November 2017
Accepted 9 November 2017
Available online 10 November 2017

Keywords:
Data mining
Crowd detection
Density-based clustering
Content aggregation
Event detection

ABSTRACT

Citizens are actively interacting with their surroundings, especially through social media. Not only do shared posts give important information about what is happening (from the users' perspective), but also the metadata linked to these posts offer relevant data, such as the GPS-location in Location-based Social Networks (LBSNs). In this paper we introduce a global analysis of the geo-tagged posts in social media which supports (i) the detection of unexpected behavior in the city and (ii) the analysis of the posts to infer what is happening. The former is obtained by applying density-based clustering techniques, whereas the latter is consequence of applying content aggregation techniques. We have applied our methodology to a dataset obtained from Instagram activity in New York City for seven months obtaining promising results. The developed algorithms require very low resources, being able to analyze millions of data-points in commodity hardware in less than one hour without applying complex parallelization techniques. Furthermore, the solution can be easily adapted to other geo-tagged data sources without extra effort.

1. Introduction

Nowadays, users are the main source of alternative sensor information in a city, although this huge source of information is often overlooked. Being ubiquitously connected to the Internet with their mobile phones, they intensively use services which promote user generated content such as Online Social Networks (OSNs), one of the most massively alternatives employed. Content in OSNs is a combination of text/images (e.g. a user post, a reply to other users posts, etc.) and meta-data information (number of likes, stars of user posts, number of posts made by the user, GPS-location, etc.). When using a GPS-enabled device, users also add a very valuable information: from where the post is shared. Thus, by analyzing the geo-located posts it is possible to know what is happening and where it is happening (Adedoyin-Olowe, Gaber, Dancausa, Stahl, & Gomes, 2016; Hua, Chen, Zhao, Lu, & Ramakrishnan, 2016). This is especially relevant in OSNs adapted for fast consumption (e.g. microblogging or image messaging) in which the time lapse between an event and its appearance in the platform is very low.

Our previous work (Domínguez, Redondo, Vilas, & Khalifa, 2017) introduces an approach to take advantage of the information given by the geo-located posts shared in social media. Abnormal location patterns were detected in the urban area under study, such as unusual city states or dynamics. The input data of the model was restricted to the posts' geolocation. This information was employed in order to find out when the shared posts in a specific area at that time of the day and that day of the week can be considered usual or an outlier (too much or too many). For this to be possible, a density-based clustering technique was applied. After a training stage which obtained the usual pulse of the city, the technique allowed the detection of abnormal behaviors on-the-fly. The work introduced in this paper supplements our previous findings. Here we set up a two-folded approach. Once the abnormal location pattern is detected, it identifies what is going on, where and when. Taking the set of posts which lead to a geo-anomaly as an activity seed, our new proposal enlarges the focus to all those posts which are considered linked to the seed. Opposite to pure NLP (Natural Language Processing) for geo-dependent topic modeling (Capdevila, Cerquides, Nin, & Torres, 2017; Xia, Hu, Zhu, & Naaman, 2015), we apply Content Aggregation Models as the one

[*] Corresponding author.
E-mail addresses: hcerezo@gradiant.org (H. Cerezo-Costas), avilas@det.uvigo.es (A. Fernández-Vilas), mmartin@gradiant.org (M. Martín-Vicente), rebeca@det.uvigo.es (R. P. Díaz-Redondo).

https://doi.org/10.1016/j.eswa.2017.11.019

in Hodson, Wilkes, Daellenbach et al. (2015) to identify meaningful threads of content that reflect what is happening in the area under study in a timely fashion. We do this in order to react to potential threats as soon as possible.

The paper is organized as follows. Section 2 summarizes other research proposal that are relevant for our work. Section 3 overviews the proposed methodology of the events detection system. In Section 4 we describe the dataset and the reasons behind our selection of Instagram as data source. Section 5 details the main aspects that have focused the evaluation of our proposal, whereas in Section 6 we enumerate the obtained results after the different experiments performed. The results are discussed in Section 7 and, finally, in Section 8 we outline the conclusions and future work.

2. Related work

Analyses of data gathered from social media (text and location linked to geo-tagged posts) have been recently applied for different and interesting purposes related to mobility patterns. In Hawelka et al. (2014) for instance, a worldwide analysis of travelers is performed by using geo-located tweets. The approach was validated by comparing the results to global tourism information, showing a strong correlation. This travelers flow enables the detection of different communities in different countries, reflecting a regional division of the world. Another interesting approach is introduced in Frank, Mitchell, Dodds, and Danforth (2013), where sentiment analysis techniques are applied to about 180,000 geo-located tweets to infer the relationship between happiness and movements within a city.

In this paper, we focus our attention to another interesting field: the detection of crowds and events in urban areas. With this aim, information gathered from shared posts supports the application of different analysis techniques applied to both the text and the location of the geo-tagged posts.

2.1. Crowds and events detection

With an approach which constructs clusters of tweets according to their number in a given area (density), the detection of local events is the main aim in Walther and Kaisser (2013). Afterwards, these clusters are scored according to different criteria: textual content, number of users, number of tweets, etc. Quite similarly, the authors in Dokuz and Celik (2017) developed a customized density algorithm to obtain socially interesting locations in a city using geo-located tweets. In the first stage, they obtain the prevalence of locations for each user. After that, interesting locations, from the point of view of group behavior, are discovered combining the per-user results. In Ranneries et al. (2016) posts from both Twitter and Instagram are clustered according to their hashtags. After that, the density-based clustering algorithm DBSCAN is applied to these clusters in order to associate a single place to each cluster. A different clustering approach is presented in Lee and Sumiya (2010), where k-means is used to group the geolocated tweets and define Regions of Interest (RoI). Over these regions, the number of tweets is analyzed in order to detect outliers. The objective is to develop a geo-social event detection to monitor crowd behaviors and local events. The approach introduced in Lee (2012) tries to infer spatio-temporal information about the events mentioned in the shared tweets. Authors applied text mining techniques (a density-based online clustering method), with the aim of detecting events in urban areas. In this approach, the location of an event is extracted directly from the text content when the geo-tagged information linked to the tweets is not available.

In Ferrari, Rosi, Mamei, and Zambonelli (2011), most visited locations are detected applying the EM-Algorithm to the location of tweets in intervals of two hours. These popular places are associated to a ZIP code. Those ZIP codes are processed using Latent Dirichlet Allocation (LDA) to find patterns in the movements of the crowds and track events with a strong relation with the city. LDA is also applied in Chae et al. (2012), in this case to the text content of the tweets, in order to find popular topics. Then an abnormality estimation is calculated using Seasonal Trend Decomposition based on Loess smoothing (STL), in an iterative process which requires expert human supervision. Other LDA-based approach is detailed in Yuan, Zheng, and Xie (2012) to relate topics and regions. Once the topics are obtained, a clustering technique is used to aggregate regions with similar topic distributions. Topic distribution is also the base of the approach in Watanabe, Ochi, Okabe, and Onai (2011), where local events are detected from analyzing microblogging data. Geohash application[1] is used for clustering location data and authors. They apply keyword frequency as the discriminatory factor for aggregating content. Keywords are associated with regions when both appear jointly more than three times in the dataset to identify local events in a region. Although the simplicity of this approach is suitable for online analysis, the event extraction schema is too naive to provide the filtering capabilities needed for anomaly incident detection.

2.1.1. Clustering and outlier techniques

There are multiple clustering methods, which are generally classified in four groups: partitioning approaches (where the number of clusters is pre-assigned), grid-based (where the object space is divided into a pre-assigned number of cells), hierarchical (where the data is organized in multiple levels) or density-based (where density notion is considered). For our purpose, density-based algorithms are the most suitable since they are able to (i) discover clusters of arbitrary shapes, (ii) handle sparse regions (which are considered as noisy regions) and (iii) work without knowing the number of clusters in advance. Among the different proposals in the literature (NafeesAhmed & Abdul Razak, 2014), we selected DBSCAN (Ester, Kriegel, Sander, & Xu, 1996) (Density-Based Spatial Clustering of Applications with Noise). Two parameters are necessary in DBSCAN to define the density measure to obtain the clusters: the radius of a circle around the data point (ϵ) and the minimum number of points that should be in this circle in order to be considered a cluster (*minPoints*). The algorithm is very sensitive to both parameters, so it is essential to select their values properly. Our estimation algorithm, detailed in Domínguez et al. (2017), is adaptive (since it is based on the nature of the dataset) and has less time complexity than other approaches in the literature.

After being able to detect groups of geographically close citizens with activity in social media (crowds) by using DBSCAN, the second step is defining the conditions under which these crowds are considered outliers. According to Hawkins "*an outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*" (Hawkins, 1980). As described in Domínguez et al. (2017), we treat a cluster as an outlier whenever the number of points differs from the number of points of other clusters found in a similar location, day and hour. We also differentiate between mild and extreme outliers (Tukey, 1977): the former lies outside the interval $(Q_1 - 1.5IQR, Q_3 + 1.5IQR)$, whereas the latter lies outside the interval $(Q_1 - 3IQR, Q_3 + 3IQR)$, being $IQR = Q_3 - Q_1$ the Interquartile Range.

2.2. Content aggregation models

Content aggregation usually involves one-to-one similarity comparisons of records (with $O(n^2)$ complexity). In applications that

[1] http://geohash.org.