ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-1009

Social network data analysis for event detection

Dario Garcia-Gasulla and **Sergio Alvarez-Napagao** and **Arturo Tejeda-Gómez** and **Luis Oliva-Felipe** and **Ignasi Gómez-Sebastià** and **Javier Bejar** and **Javier Vázquez-Salceda** [1]

Abstract. Cities concentrate enough Social Network (SN) activity to empower rich models. We present an approach to event discovery based on the information provided by three SN, minimizing the data properties used to maximize the total amount of usable data. We build a model of the normal city behavior which we use to detect abnormal situations (events). After collecting half a year of data we show examples of the events detected and introduce some applications.

1 Introduction

Large amounts of strongly distributed data are available today through the use of Social Network (SN) applications in smart mobile devices. That information combined becomes a large real-time stream of situational data. The computation of large amounts of data typically entails a complexity limitation, as detailed models are often unable to process large data sets in acceptable time. In that regard it has been argued that simpler models and larger datasets can outperform more complex models based on less data [4]. Following this argument in this paper we develop a simple model to capture a complex domain (the events in a city) using huge amounts of simple data.

Our goal is to detect ongoing events in a city. An event can be defined as something non-trivial that happens at a particular time and place. Several research fields in Artificial Intelligence have been employed for the solution of the event discovery problem, such as Natural Language Processing (NLP, for topic detection and tracking), Knowledge Representation (through ontologies and taxonomies), Data Mining (for prediction and tracking) and Visual Analytics. In contrast we predict events in a city in real time through the frequency and distribution of SN activity. We we only require that the activity has a temporal and spatial grounding. In our approach each SN user provides equally relevant data, and through the aggregated data of all users we build a growing model to represent what the normal behavior of a city looks like. Then we identify where and when the city is behaving abnormally to identify ongoing events in real time.

2 Related Work

In [7], authors identify earthquakes, typhoons and traffic jams based on tweets from Twitter. Their model assumes that messages represent an exponential distribution as users post the most after a given time. In [9], authors propose an algorithm based on clustering Wavelet-based signals. Their algorithm builds signals for single words and captures the bursts ones to measure cross-correlations. In [2] authors use heat-maps visualizations and density strips to enable the detection of anomalous data. The TEDAS system [6] solves analytical queries and generates visual results that rank tweets and extract patterns for online queries. Tweevent [5], retrieves tweets from the Twitter stream, segmenting them in consecutive phrases. It applies a clustering algorithm to group segments, each representing a detected event. Most of these proposals use internal features of the SN activity (*e.g.*, text of tweets) to build their model. As a result, the main challenges they must face are related with NLP (*e.g.*, solving ambiguities, finding stop-words, *etc.*). Our proposal is different in that it does not consider properties of SN activity, only the existence of the activity itself and its temporal and spatial grounding.

3 Data Properties and Processing

We use three SN sources: Twitter for tweets generated in an area, Foursquare to know where users are currently located, and Instagram for the location and time of picture uploads. Since semantics of each SN activity are hard to process we focus on their most basic shared properties: a happening at a time and place. By simplifying the problem we increase the size of the dataset (Twitter + Foursquare + Instagram) and increase its resistance to variations and outliers. The simplicity of the data available forces us to characterize events in a simple manner, using only time to represent the persistence of events. This allows us to study how events are created and how they fade away over time. The other dimension of our event model is certainty, which our system approximates based on the evidence available.

We split time in 15 minutes time-windows, allowing us to detect all events lasting longer than that. We collected data since July 2013, and at the time of writing this paper we are still collecting it. That amounts to over 22,000 time-windows. Regarding space, we focus on the Barcelona metropolitan area, with an area of $633km^2$ and a population of 3.2M people. We split the city into 6 characters geohash sectors [3], each representing $0.55km^2$. As a result we have over 2,000 land sectors with data. We aggregate the SN by computing a combined activity density for each specific sector and time-window. We perform a second aggregation to obtain a separate behavioral model of each sector for each 15 minutes interval of a week (Monday to Sunday). Given the activity density in a land sector for a particular weekly interval (*e.g.*, Tuesday, between 10:15 and 10:30) we define normality measures using the median and the *inter-quartile range* (iqr) [8]. We define *normal activity* as any under 1.5 iqr, *deviated activity* as any between 1.5 and 3 iqr, and *abnormal activity* as any above 3 iqr. Finally, we define an *ongoing event* in a sector as a consecutive time gap in such that all its time-windows correspond to deviated activity and at least one corresponds to abnormal activity. See an example in top image of Figure 1. We model events as quasi-trapezoids where the horizontal axis represents time and the vertical axis certainty (see an example in bottom image of Figure 1). In this representation we identify three components in each event: its beginning (an ascending slope), body (a horizontal top line) and end-

[1] Universitat Politècnica de Catalunya, Barcelona, Spain, email: {dariog,salvarez,jatejeda,loliva,igomez,bejar,jvazquez}@lsi.upc.edu

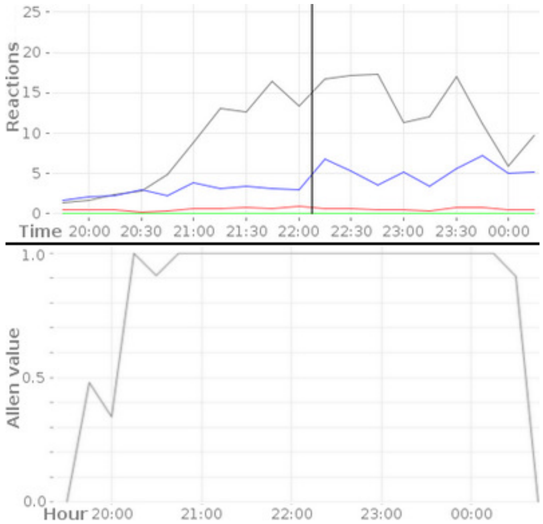

Figure 1. Model of a concert event. Top image shows SN activity (top line is the event, middle line is the 3 iqr for this sector and week interval). Bottom image shows its trapezoid representation, with certainty on the y axis.

ing (a descending slope). This three part decomposition of events is frequent in the bibliography and allows for temporal reasoning [1].

4 Experimentation

In 229 days of captured data our system detected 712 events. We checked for known events taking place in the same place and time as those our system detected to validate its precision. In Table 1 we present the top 20 events captured by impact. 12 of those correspond to games of Football Club Barcelona (FCB), its stadium having a capacity for 98,000 people. Within the top events there are also 4 concerts and 3 events in the airport. Football games events indicate that the popularity of the event is linked with our detection strength. The most popular game is against the main rivals of FCB, and Champions League games (the most important competition FCB plays) are highly ranked (vs Milan, vs Celtic). This effect is also found in concerts, which rank based on the popularity of the performer. There are other interesting events which we have identified within the total of 712 detected, and its variety gives an idea of the capabilities of the methodology. Some of those are: the iPhone 6 release day, a strike in the train service and congresses such as Smart Cities. The relation between event size and detection strength motivated a formal correlation analysis. We computed a linear regression (see Figure 2) between detection strength of the top 100 events and the actual event attendance (obtained afterwards from various sources). The Pearson correlation coefficient is 0.82.

1	FCB vs Madrid	11	3 nearby concerts
2	FCB vs Elche	12	Airport
3	FCB vs Malaga	13	Michael Buble concert
4	FCB vs RCDE	14	Arctic Monkeys concert
5	FCB vs Milan	15	New Year @ Park Guell
6	FCB vs Granada	16	Airport
7	FCB vs Valencia	17	Bruno Mars concert
8	FCB vs Villareal	18	FCB vs Efes (Basketball)
9	FCB vs Celtic	19	FCB vs Real Sociedad
10	FCB vs Cartagena	20	Airport

Table 1: Top 20 events by impact

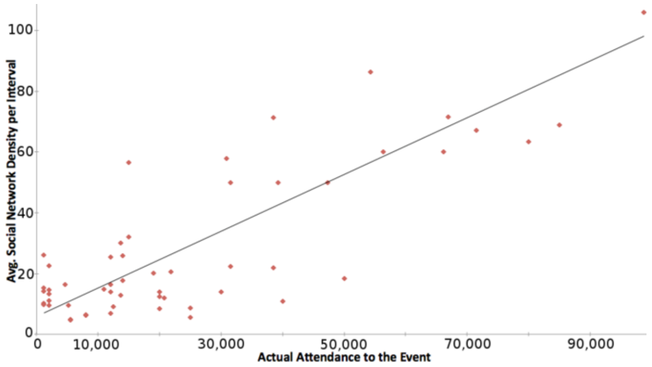

Figure 2. Captured data vs. actual attendance for the Top 100 events.

5 Conclusions and Future Work

The behavior of a city is a complex domain to model. To simplify we aggregate lots of small information units (activity from SN at a time and place) and obtain a simple function of what can be considered normal. That function is then used as baseline against ongoing activity patterns to identify when and where are abnormal situations happening. By applying statistical analysis on the data we prune events and keep those which are certainly relevant. The continuous training and test methodology implemented allows our system to become more outlier-proof with time. The result is a system capable of detecting a wide variety of events with excellent precision (*i.e.*, we do not detect irrelevant events) in real time [2].

This work could be used by city authorities to detect and react to ongoing incidents. The profile of events (*e.g.*, how fast it forms) could be used to adapt safety and mobility actions. Public transport authorities can use the density analysis to alter their schedule and optimize their service. Finally, a future spatio-temporal correlation analysis of events could be useful also to private enterprises.

REFERENCES

[1] James F Allen, 'Maintaining knowledge about temporal intervals', *Communications of the ACM*, **26**(11), 832–843, (1983).

[2] D Cheng, P Schretlen, N Kronenfeld, N Bozowsky, and W Wright, 'Tile based visual analytics for Twitter big data exploratory analysis', 2–4, (October 2013).

[3] A Fox, C Eichelberger, J Hughes, and S Lyon, 'Spatio-temporal indexing in non-relational distributed databases', in *Big Data, 2013 IEEE International Conference on*, pp. 291–299, (2013).

[4] Alon Halevy, Peter Norvig, and Fernando Pereira, 'The unreasonable effectiveness of data', *Intelligent Systems, IEEE*, **24**(2), 8–12, (2009).

[5] Chenliang Li, Aixin Sun, and Anwitaman Datta, 'Twevent: Segment-based Event Detection from Tweets', 155–164, (2012).

[6] Rui Li, Kin Hou Lei, R Khadiwala, and K C C Chang, 'TEDAS: a twitter-based event detection and analysis system', 1273–1276, (April 2012).

[7] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo, 'Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors', 851–860, (2010).

[8] John W Tukey. Exploratory data analysis, 1977.

[9] Jianshu Weng and Bu-Sung Lee, 'Event Detection in Twitter', in *Proceedings of the Fifth International Conference on Weblogs and Social Media*, Barcelona, Spain, (2011). The AAAI Press.

[2] This work has been partially supported by the SUPERHUB project (SUstainable and PERsuasive Human Users moBility in future cities, Cofinanced by the European Commission ICT-FP7-289067)