GAZOUILLE: Detecting and Illustrating Local Events from Geolocalized Social Media Streams

Pierre Houdyer[1], Albrecht Zimmerman[2], Mehdi Kaytoue[2(✉)], Marc Plantevit[3], Joseph Mitchell[1], and Céline Robardet[2]

[1] Tapastreet Ltd., 36 Dame Street, Dublin, Ireland
{pierre,joe}@tapastreet.com
[2] INSA-Lyon, CNRS, LIRIS UMR5205, 69621 Lyon, France
{albrecht.zimmerman,mehdi.kaytoue,celine.robardet}@liris.cnrs.fr
[3] Université Lyon 1, CNRS, LIRIS UMR5205, 69622 Lyon, France
marc.plantevit@liris.cnrs.fr

Abstract. We present GAZOUILLE, a system for discovering local events in geo-localized social media streams. The system is based on three core modules: (i) social networks data acquisition on several urban areas, (ii) event detection through time series analysis, and (iii) a Web user interface to present events discovered in real-time in a city, associated to a gallery of social media that characterize the event.

1 Introduction

Social networks (such as Twitter, Instagram, ...) are rich sources of information that can be used to build a huge number of applications and services for certain end-users (b2c), for companies, e.g. with analytics platforms (b2b), but also to help governments and charitable organizations. Through several public APIs, one can access streams of messages, often provided with text (including hashtags, user mentions and URIs), media (images or video) and geo-tags indicating the position of the user emitting the message (called *post* in the sequel).

One way of exploiting such data is to discover global trends and detecting events in the streams of posts. The motivations are manifold: disaster detection, epidemic surveillance, identification of newsworthy events that traditional media are slow to pick up, identification of trends, monitoring of brand perception, etc. The question of how to identify events in streams of text data has been a research topic for more than a decade now, starting from e-mail, via blog posts, to location-based social networks data [2]. The general idea underlying most of that work is identifying "bursty" topics (mentioned significantly more often during a time period than in the period preceding it).

Whereas most of the existing systems identify global trends, only a few take into account the geo-localization of the posts for detecting local events [3]. This is actually the goal of GAZOUILLE: harvesting data from urban areas, the system is able to detect spatially circumscribed events in real-time and to intelligibly

This work was supported by the project GRAISearch (FP7-PEOPLE-2013-IAPP).

A. Bifet et al. (Eds.): ECML PKDD 2015, Part III, LNAI 9286, pp. 276–280, 2015.
DOI: 10.1007/978-3-319-23461-8_29

characterize them (with their periodicity, users, key-words, and via a media gallery, e.g. in Figure 2). In what follows, we present an overview of GAZOUILLE (Section 2) and a use-case on the city of New-York (Section 3).

2 System Overview

The system architecture is illustrated in Figure 1: a crawler engine harvests social networks (e.g. Twitter and Instagram) in specific locations (cities) and populates a database with posts. An event detection module is running continuously with a sliding window, and enters detected events into the database. Finally, a Web user interface allows to choose a time window, and explore the most highly expressed events with an intuitive characterization (see elements in Figures 2 and 3). We explain the different modules now.

Data acquisition. For acquiring data, trackers are set on a selected city, on which a grid is defined: each cell gives rise to a geo-tagged query on each social network every 5 minutes (default refresh rate). The back-end is realized in *Ruby*, connects to data providers' official APIs, and stores data in a *PostgreSQL* DBMS.

Data preparation. From each geo-tagged post, we extract meaningful key-words (e.g. hashtags and user mentions from *tweets*). Natural language processing tools could also be used here for stemming, lematization, etc. In the end, a city gives rise to a single stream of pairs $(timestamp, \{word_1, ..., word_n\})$.

Event detection. The detection is performed in a window of a given size. Each time the trackers refresh, the window is right-slid and a new detection is performed. Based on our review of the state-of-the-art, we have selected a light-weight method for bursty term detection [1] and implemented it with several modifications to handle texts from social networks (the method was originally designed analyzing news corpus). The method described in [1] transforms each terms' *document frequency - inverse document frequency* (DFIDF) scores over time via a Discrete Fourier Transform and derives its *periodicity* and *strength of expression* from the resulting periodogram. Originally, stopwords are used to identify "irrelevant" terms but since those are not available in our settings (and difficult to derive for Twitter in general), we classify all terms with less than average expression thus (denoted as 'L' for *Low* in Figures 2 and 3, 'H' for *high* otherwise). Individual bursts are modeled as *Gaussians*.

Front-end. The user selects a window of time in which events are detected. Bursty terms are ranked w.r.t. their strength of expression. The user selects the

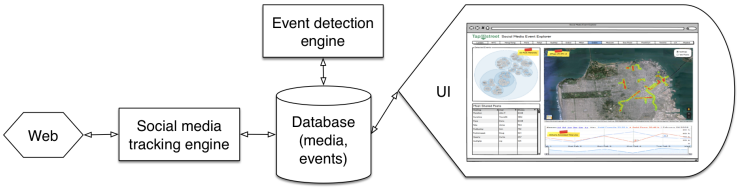

Fig. 1. System architecture

terms he is interested in and the rest of the interface updates (details hereafter). We used *Google Charts*[1] and *magic wall*[2] UI components.

3 Use Case: New York City

We collected tweets posted from October the 31^{st} to November the 4^{th} 2014 in New York City, covering the districts of Brooklyn, Manhattan, Queen's, and Staten Island (geo-tagged queries to the Twitter API every five minutes). We accordingly tracked 200,000 geo-localized tweets in a period containing the 2014 NYC marathon (November 2). Our goal is to validate that this event will be discovered in time and space, but we should also be able to identify several other events that happened in this area and time frame. The event detector engine is run with a sliding window of 128 time stamps, each corresponding to ten minutes.

As an end user, the window in which discovered events should be given can be selected and moved. We set the window to the day of the marathon (note that dates are in Dublin GMT timezone), see Figure 2. The ordered list of bursty terms is then refreshed automatically. In this case the most strongly expressed terms are *#nycmarathon* and *@nycmarathon*. Expression is given in percentage of expression w.r.t. the maximum, i.e. the most bursty term. We then select those two first terms and all tweets gathered by our initial trackers that involve these terms are displayed on the map. The shape of these geo-localized tweets strongly resembles the known marathon course (left map bordered in black). The third bursty terms also concerns the marathon. The fourth term is the user mention of a singer from Harlem that released a song that day, freely available on the Web, thus shared and discussed on Twitter. Finally, the last given bursty term concerns *@CRinQC* who we identified as a Republican tweeter whose criticism of President Obama was re-tweeted by noted New York business and former Republican presidential hopeful Donald J. Trump, amplifying his expression. The user can then discover other terms in the list.

When shifting the time window one day later (see Figure 3), the interface updates. We now have a new list of bursty terms. Whereas tags related to the New York marathon are still present, they are not the burstiest terms anymore. We select the four most strongly expressed terms (*@dalailamatruth*, *#dalailamabeacon*, *#religiousfreedom*, and *#dalailama*) since one may assume they concern the Dalai Lama. Plotting the concerned geo-localized tweets on the map result in a very concentrated area: the Beacon theater where the Daila Lama was giving a lecture on November 3 and 4 2014, in front of which protesters gathered, as the media gallery related to these posts suggests.

[1] https://developers.google.com/chart/interactive/docs/gallery/timeline
[2] http://teefouad.com/plugins/magicwall

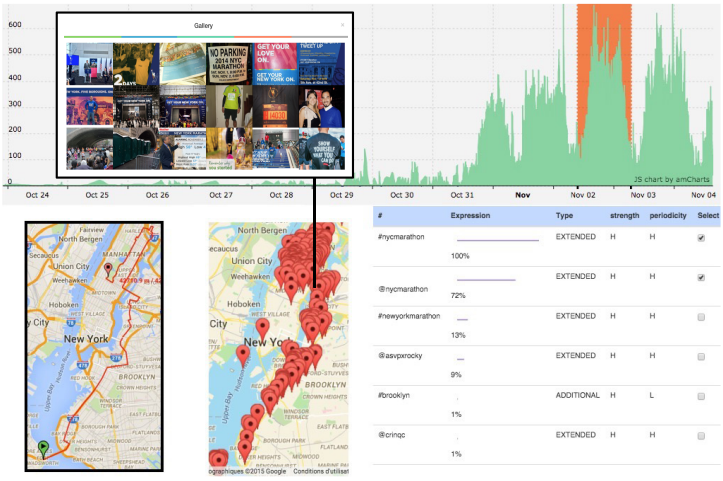

Fig. 2. Detecting events in New-York on November 2 with selected bursty terms *#nycmarathon* and *@nycmarathon*, corresponding geo-localized tweets (right map), and the known NYC 2014 marathon course (left map). In the table, expression is given in percentage with respect to the top expressed term, while expression and periodicity are given as high ('H') and low ('L'), i.e. above/below their average value.

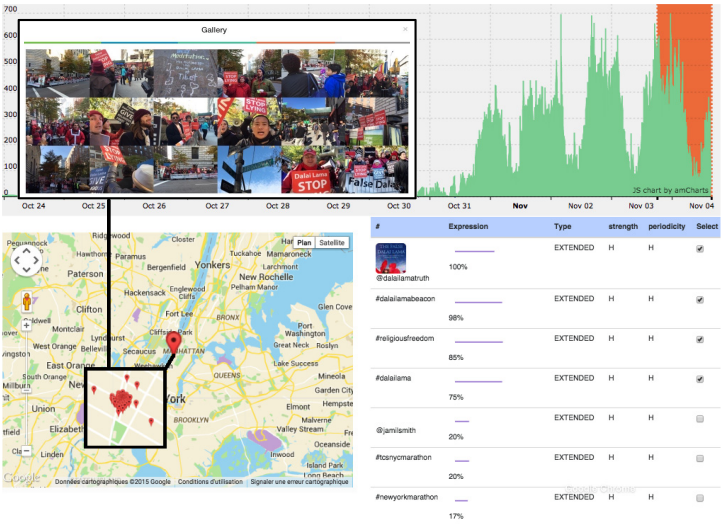

Fig. 3. Detecting events in New-York on November 3 with selected bursty terms *@dalailamatruth*, *#dalailamabeacon*, *#religiousfreedom*, and *#dalailama* and corresponding geo-localized tweets all concentrated around the Beacon Theater. A gallery of the corresponding media is also given.

References

1. He, Q., Chang, K., Lim, E.: Analyzing feature trajectories for event detection. In: SIGIR 2007, pp. 207–214. ACM (2007)
2. Symeonidis, P., Ntempos, D., Manolopoulos, Y.: Location-based social networks. Recommender Systems for Location-based Social Networks. Springer Briefs in Electrical and Computer Engineering, pp. 35–48. Springer, New York (2014)
3. Xia, C., Schwartz, R., Xie, K.E., Krebs, A., Langdon, A., Ting, J., Naaman, M.: Citybeat: real-time social media visualization of hyper-local city data. In: WWW 2014, Companion Volume, pp. 167–170. ACM (2014)