CBMI - Internationl Workshop on Content-Based Multimedia Indexing
19-21 June 2017, Firenze

MICC
Media Integration and Communication Center

Separating the Wheat from the Chaff
Events Detection in Twitter Data

Andrea Ferracani, Daniele Pezzatini, Lea Landucci, Giuseppe Becchi, Alberto Del Bimbo
MICC - University of Florence, Italy

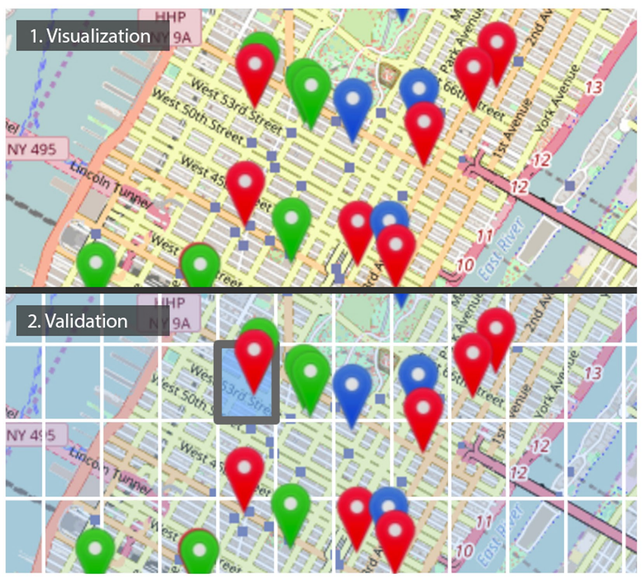
The System

In this paper we present a system for the **detection and validation of macro and micro-events** in cities (e.g. concerts, business meetings, car accidents) through the analysis of geolocalized messages from **Twitter**. A simple but effective method is proposed for unknown event detection designed to alleviate computational issues in traditional approaches.

The method is exploited by a **web interface** that in addition to visualizing the results of the automatic computation exposes interactive tools to inspect, validate the data and refine the processing pipeline.

Researchers can exploit the web application for the rapid creation of macro and micro-events **datasets of geolocalized messages** currently unavailable and needed to improve supervised and unsupervised events classification on Twitter. The system has been evaluated in terms of **precision**.

The proposed processing pipeline used in our system contemplates three main steps:
a) **tweets extraction**; b) **abnormalities detection**; c) **mining and visualization**.



a) tweets extraction: tweets are extracted daily through the **Twitter API** using the Java library twitter4j2 and stored in a mySQL database for post-processing;

b) abnormalities detection: we use a grid-based method to analyze all the geo-referenced data in a certain urban area. The area is divided arbitrarily in cells of predefined dimensions.
Tweets in each cell of the grid are analyzed with our method which combines **temporal and spatial density analysis**;

c) mining and visualization: Once the event clusters has been identified by the DBSCAN algorithm, the event is visualized on the web interface and positioned in its center of mass with respect to the geolocalization of each tweet.

Abnormalities detection

Our method for abnormalities and event detection which consists in a pipeline of three core algorithms:

1) **DTW**: To detect windows of time in a cell of the gridwhere the distribution of tweets is unusual, we consider the **average of the distance** and the **standard deviation**.

$$d_i > \overline{d} + \sigma_d \quad \| \quad d_i < \overline{d} - \sigma_d$$

2) **Crest-Detection**: A peak is detected at time *t* if:

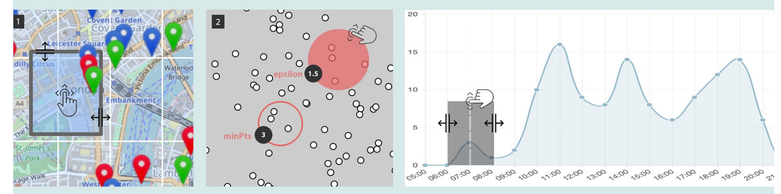
$$v(t) > \sum_{j=1,..,\Gamma_t} v(t-j) + \epsilon \quad \&\& \quad v(t) > \sum_{j=1,..,\Gamma_t} v(t+j) + \epsilon$$

3) **DBSCAN**: given a set of geo-located tweets in the space groups together the points that are closer in the distribution. An **event is proposed** by the system **if at least one cluster is detected** by DBSCAN for the tweets posted during the temporal interval of the peaks detected in step 2.

Interface and evaluation

The interface visualise on a map the results of the automatic event detection pipeline.

The interface provides **two main views** both mapbased with two different access levels:
1) the **visualization view** shows on the map the events automatically detected bythe system;
2) the **validation view** allows an authorized user to confirm (or not) the correctness of the detection and to customize the search.

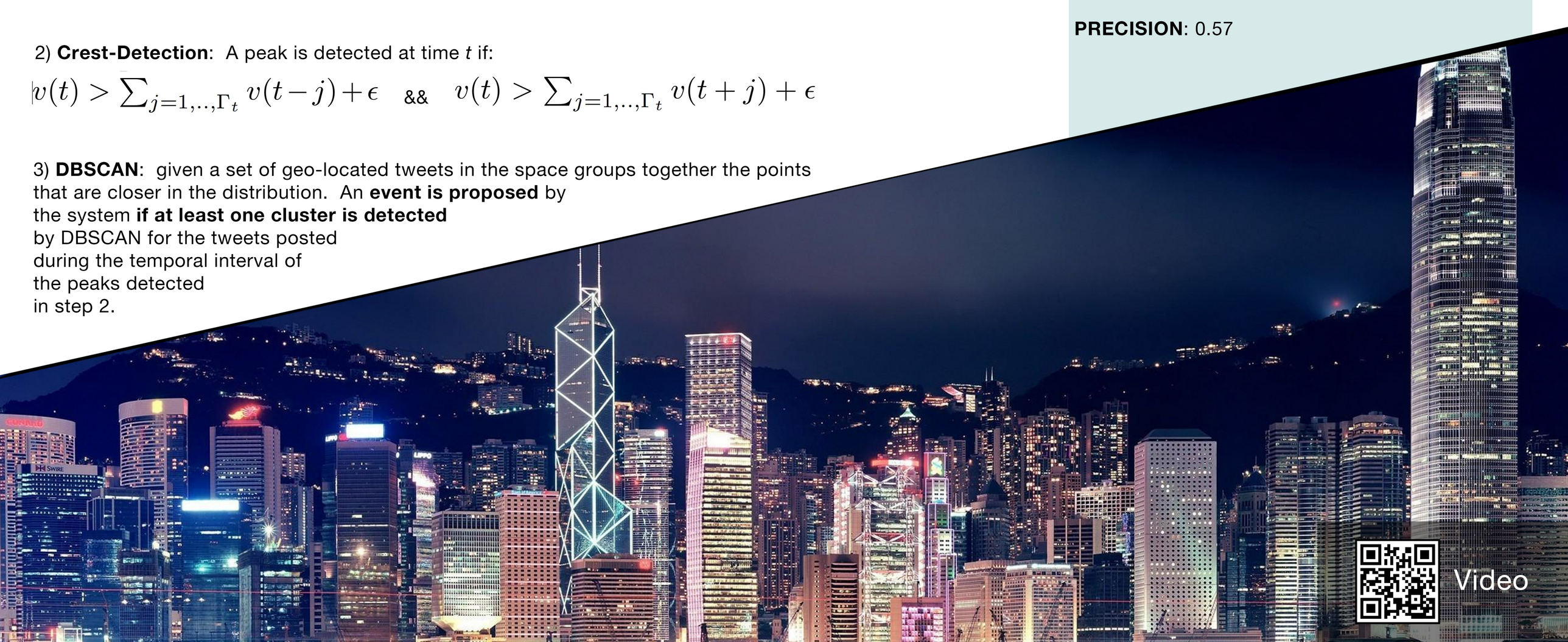


Each event is represented by a marker that can have three color: blue, red and green.
Blue pins: events detected not still validated
Red pins: events respectively misclassified
Green pins: correctly classified as confirmed by a user.

In the validation view the user can adjust the **parameters** used in abnormalities detection.
1) **position and dimension** of the **grid cell**
2) **time interval** of event detection;
3) **time period** of the average and standard deviation of the **DTW**;
4) **threshold value** over the average for reporting an abnormality;
5) **time window** and **delta** for **crest-detection**;
6) **minPts** and the **epsilon** of **DBSCAN**.
He can also validate and categorize detected events.

Evaluation: tweets published in New York and LondonMarch 31 and April 9, 2016

LND: 17176 users/44932 tweets > **190/340 events**
NY: 17378 users/43186 tweets > **516/900 events**
PRECISION: 0.57



Video