Event Detection on Social Media Streams

Antonia Saravanou, Ph.D. Student

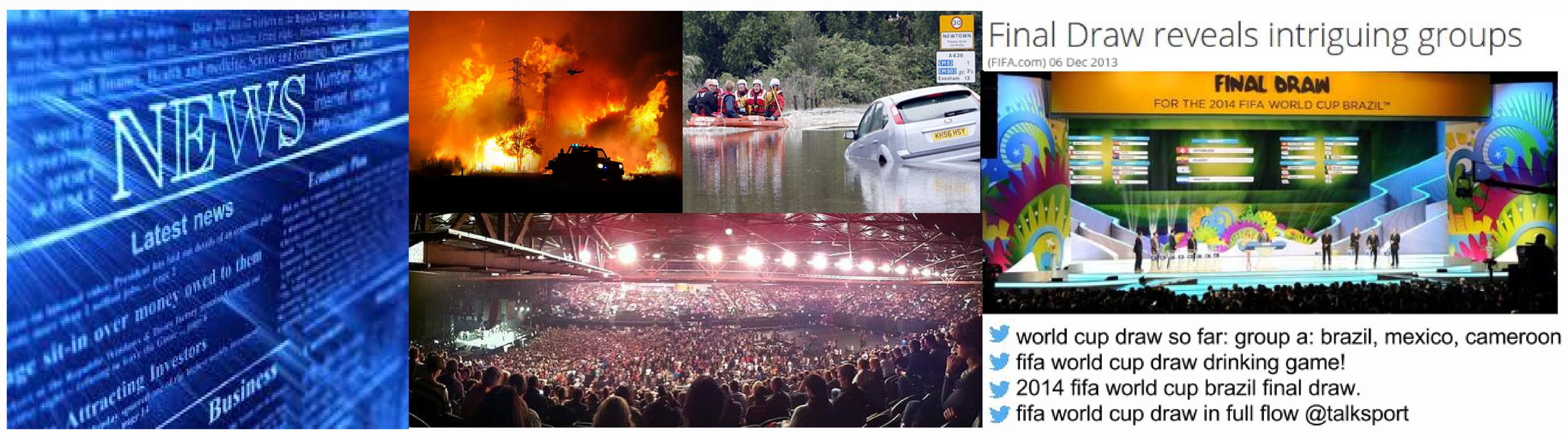
antoniasar@di.uoa.gr, www.di.uoa.gr/~antoniasar

KDDLab, dept. of Informatics and Telecommunications, University of Athens, Greece

Motivation

We study the problem of event detection in a social stream

facebook, FOURSQUARE, LinkedIn, YouTube, Google+, tumblr., Pinterest



Events could be news, disasters, concerts, sports happenings, etc.

Our goals:

- Identify the event
- Monitor the *evolution*, the *duration* and the *location* of the event
- Inform users

Content-based event detection

Identifying topics in a stream of text

A) Filtering Step:

Create Flood-Related Lexicon & Extract Flood-Related text

seed set: *rain, flood, storm, showers, ...* → tokens that contain >1 word of the seed set as a substring; e.g., raining, floods, #ukfloods, etc. → **Flood-Related Lexicon** **+** **Dataset**

$\underbrace{\text{Flood-Related Lexicon + Dataset}}_{\text{Flood-Related text}}$

Flood-Related text

B) Clustering Step:

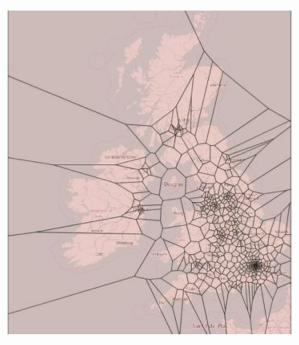
- Find areas

K-Means using the GPS coords → Clusters as Voronoi polygons

- Cluster areas

K-Means using:

1. ***count***(*d*)
2. ***ratio***(*d*) = *count*(*d*) / Σ *count*(*d'*), *forall d'*
3. ***speed***(*d*) = *ratio*(*d*) - *ratio*(*d*-1)



***K* = 500**

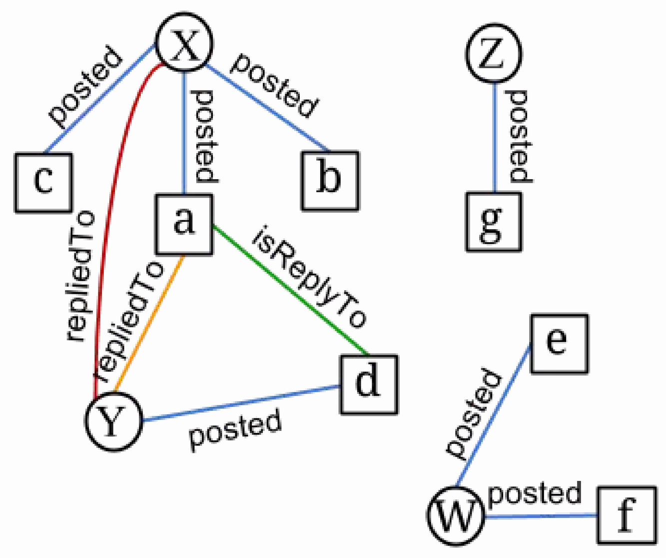
Graph-based event detection

Model network as a graph

A) Building Step:

Construct the network graph

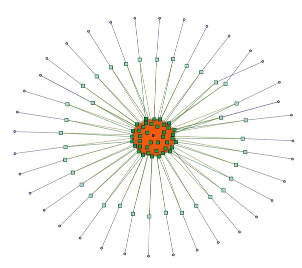
- userX posted tweetA
- userY posted tweetD
- tweetD isReplyTo tweetA
- userY repliedTo userX
- userY repliedTo tweetA

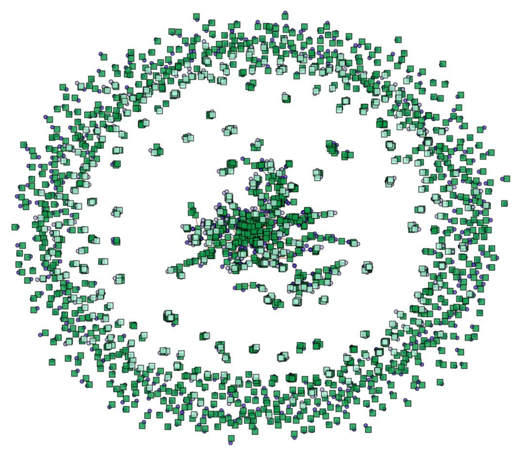


B) Filtering Step:

- ***Large Connected Components*** indicate large-scale conversations: *LCCs* → event candidates

prune *spam* LCCs → star graph structure e.g.,



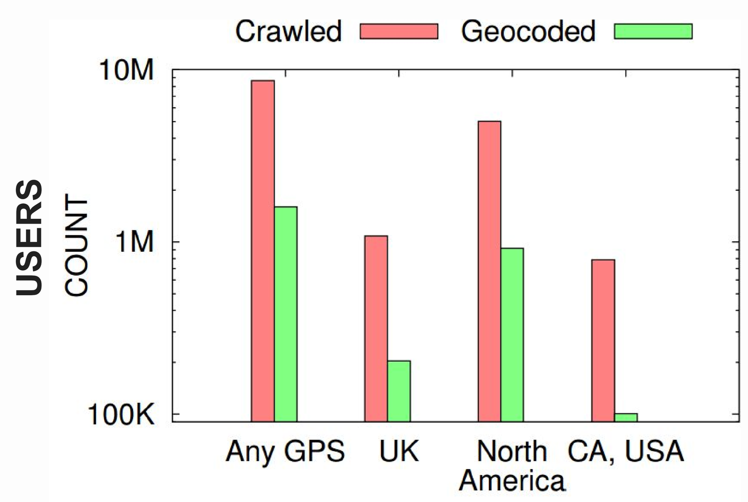


Facet Crawler

To collect data from Twitter, we use our custom made crawler

different queries → different constraints from Twitter API
e.g., limited number of queries per 15 min timecap

Our crawler can handle different query types with different parameters



example:

Crawl tweets with *location*:

- 2D bounding box with GPS (**green**), GPS enabled
- custom geocoding (**red**), e.g., location in user profile

Geocoding can extract an additional **10%**

Experiments on content-based

- Collection of **public** tweets from the UK
- > 2.3M geotagged tweets

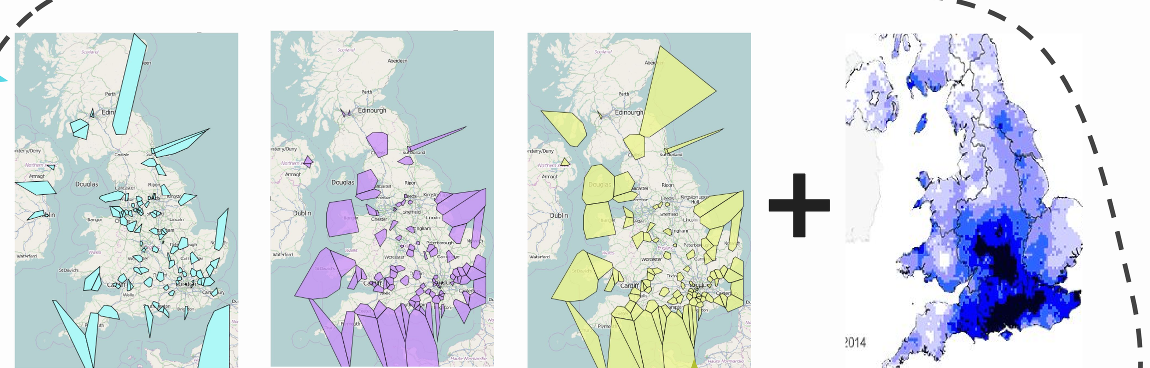
Top 100 areas based on:

a) number of **all** tweets

b) number of **flood-related** tweets

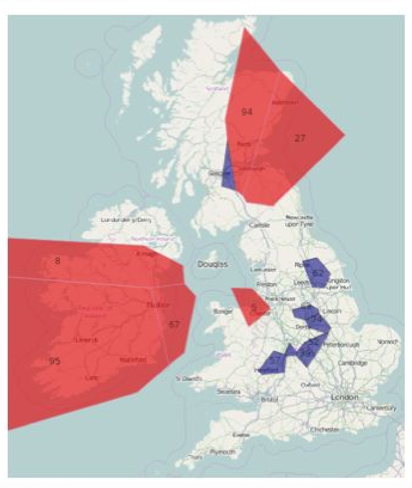
c) signal-to-noise ratio:

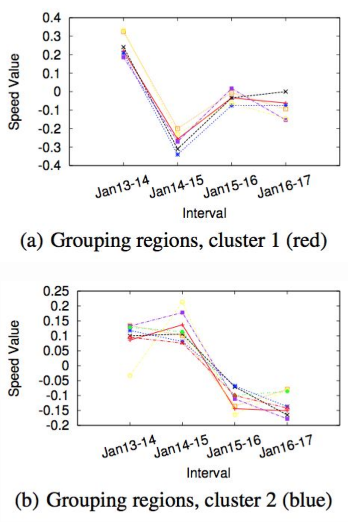
$$score(r) = \frac{\#\texttt{flood-related tweets in r}}{\#\texttt{tweets in r}}$$



Ground Truth
UK MetOffice

Visualization of 2 clusters



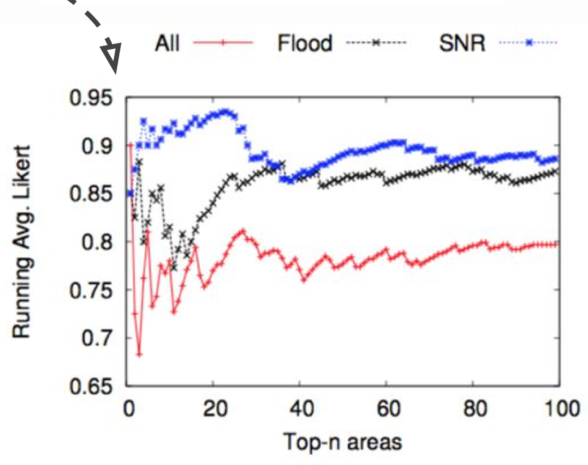


(a) Grouping regions, cluster 1 (red)

(b) Grouping regions, cluster 2 (blue)

- *Speed* feature has better performance
- Red *cluster*: mostly unaffected, *speed* decreases
- Blue *cluster*: affected, *speed* increases

$$value_i = \frac{\sum_{j=1}^{i} \frac{likert\_score(j)}{5}}{i} \rightarrow [0, 5]$$



(c) 1000 clusters

Experiments on graph-based

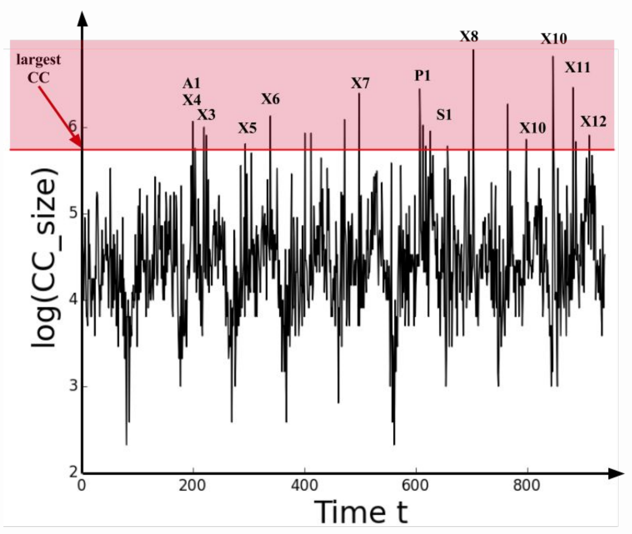
- Collection of **public** tweets from London
- ~ 700K geotagged tweets

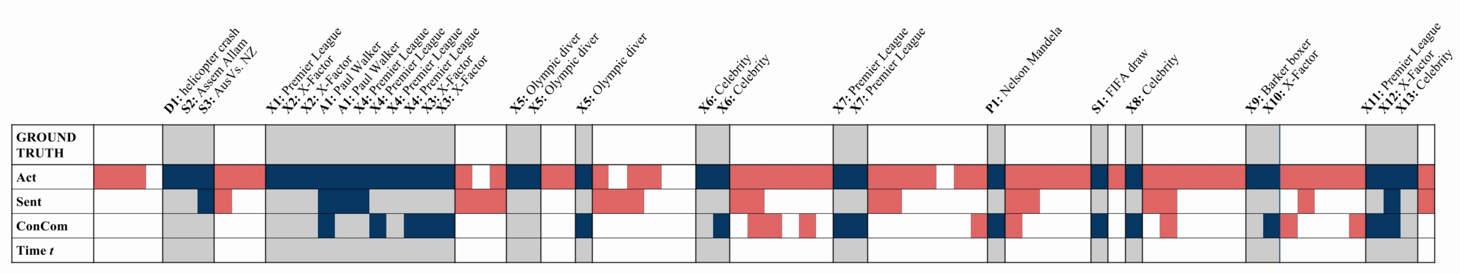
ground truth:

- *automatically* from Wikipedia
- manual investigation

data into 15-min segments

12 out of 19 events ✓





Comparison:

- **Act:** number of tweets
- **Sent:** negative / positive
- **ConCom:** our method

	jaccard similarity	precision	recall	f-score
Act	0.5	0.42	0.86	0.56
Sent	0.47	0.23	0.17	0.2
ConCom	0.72	0.65	0.52	0.58

A. Saravanou, G. Valkanas, D. Gunopulos, G. Andrienko, "**Twitter Floods when it Rains: A Case Study of the UK Floods in early 2014**", SWDM'15 (WWW'15 Workshop)

G. Valkanas, **A. Saravanou**, D. Gunopulos, "**A Faceted Crawler for the Twitter Service**", WISE 2014

ConCom is part of a joint work with I. Katakis and D. Gunopulos