Real-time event detection in social network data streams

FEUP ProDEI – 7th Edition
Mário Miguel Fernandes Cordeiro
pro11001@fe.up.pt

Supervisor:
Dr. João Gama, FEP, LIAAD, Universidade do Porto, Portugal

Co-supervisors:
Dr. Ricardo Morla, FEUP, INESC Porto, Universidade do Porto, Portugal
Dr. Miles Osborne, School of Informatics, University of Edinburgh, UK

17/03/2014
Outline

- Introduction
  - Motivation
  - Problem statement
- Real-time Event Detection
- Proposal
  - Hypothesis
  - Research questions
  - Evaluation
- Planning
- Discussion
How Our News Sources Changed

Source: Timeline: How Our News Sources Changed in the Last 200+ Years

The Revolution Will Be Tweeted.


Image Source: http://www.foreignpolicy.com/articles/2011/06/20/the_revolution_will_be_tweeted
Online social networking services properties

ubiquity
available in any internet connected device
Online social networking services properties

- Ease of use
- Short messages, pictures, shortened links
Online social networking services properties

messages spread very fast on the network

Online social networking services properties

Exchanged messages follow a time decay pattern

Lardinois, F. (2010)

*Readwritesocial: The short lifespan of a tweet: Retweets only happen within the first hour.*


*The origin of bursts and heavy tails in human dynamics.*
Social networks text stream data differs substantially from general text stream data.
rich social connections

Source: Twitter Company conversations mapped
rich social connections

temporal attributes of each text piece

Source: Twitter Company conversations mapped
rich social connections

temporal attributes of each text piece

more context sensitive
Following Followers
Retweet, Reply, Mentions
Each OSN user is regarded as a sensor of the real world; each message as sensory information.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: real-time event detection by social sensors*
Using social sensor data to monitor unforeseen patterns:
Using social sensor data to monitor unforeseen patterns: Event Detection
Importance of real-time Event Detection

Data Value Chain

Value of an Individual Data Item vs. Data Age

Value of Data in Aggregate vs. Data Age

Data Value

Data Age

Interactive Real time Analytics Record Lookup Historical Analytics Exploratory Analytics

Source: VoxtDB, Inc.

Allied with 3 + 1 V´s of Big Data

Volume
Data at Rest
Terabytes to exabytes of existing data to process

Velocity
Data in Motion
Streaming data, milliseconds to seconds to respond

Variety
Data in Many Forms
Structured, unstructured, text, multimedia

Veracity*
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Exact vs Approximate Answers

Volume
Velocity

Exact Solution

Quick Response

slower response time

approximated solution

low data volume and/or velocity

Pick Two

Availability

MySQL, PostgreSQL, Vertica etc.

Consistency

Merge, Rebase, gigascale, mem etc.

Gigascale, Vertica, Hive, Dynamic etc.

Partition
Tolerance

MyISAM, HBase, gigascale, mem etc.

Availability

MySQL, PostgreSQL, Vertica etc.
Communities vs Globally
Event detection state-of-the-art

Topic detection and tracking (TDT):

- event detection

- first story detection / novelty detection
Event detection state-of-the-art

Advent and massification OSNs and big data era:

▪ first story detection:

▪ survey event detection:
Real-time social network text stream event detection algorithm

should be able to mine continuously, high-volume, open-ended social network data stream documents as they arrive, interpret their network relations and be ready to detect new events at any time
Research areas

- Data Stream Mining
- Social network analysis
- Natural Language Processing

Data Mining » Machine Learning » Unsupervised Learning
Hypothesis

In social networks real-time event detection using data stream algorithms, major events are better predicted by correlating the observation of peaks in a specific set of topic mentions contained in the text stream, and the spontaneous creation or growth of their network linked communities.
Research Questions

Question 1

Can a data stream algorithm provide a robust community identification and tracking in dynamic social networks?
Research Questions

Question 2

Is the abrupt increase of topic mentions in a social network text stream representative of the occurrence of an event?
Research Questions

Question 3

Can the **accuracy** of a Social Network event detection algorithm be **enhanced** with the **dynamics of the network** and its information spreading patterns?
Evaluation

Reference systems:
- dynamic community detection
  - Louvain method (Blondel et al., 2008)
- event detection
  - UMASS system (Allan et al., 2000b)
  - LSH, (Petrovic, 2012)

Datasets:
- FSD twitter corpus
  - 50 million tweets
  - 27 manually annotated events
  - 3035 tweets were labeled as being on-topic for one of the 27 events (Osborne et al., 2012).
Work Plan

- 2013: State-of-the-art
- 2014: Detection, Tracking and Community Dynamics
  - CP1 (CD) PhD proposal
  - CP2 (CD+DS)
- 2015: Topic Detection and Text Summarization
  - R1
  - CP3 (ED)
  - R2
- 2016: Data Streaming: Text Mining, Social Networks and Event Detection
  - CP4 (ED+CD+DS)
  - R3
- 2016: Writing the PhD Thesis

- Literature review:
  - R1: Update SNA and DS state-of-the-art
  - R2: Update NLP and DS state-of-the-art
  - R3: Final literature review

- Conferences:
  - CP1: Community Detection paper
  - CP2: Community Detection using Data Streams paper
  - CP3: Event Detection paper
  - CP4: Prof of concept paper (Event Detection + Community Detection + Data Streams)

- Journals:
  - JP1: Journal paper
  - JP2: Journal Paper
Current Activities

Dynamic Community Detection Algorithm:
- Based Louvain method (Blondel et al., 2008)
- Adding removing modes and edges

Image Source: https://sites.google.com/site/findcommunities/
Papers

- **2014:**

- **2012:**
Past Events

- November 2013:
  - Big Data Spain
    - http://www.bigdataspain.org
  - Strata Conf EU

- July 2013:
  - 3rd Lisbon Machine Learning School
    - http://lxmls.it.pt/2013
Other Activities

- **Oporto MongoDB User Group:**
  - Founder of the user group
  - Community with 140 members
  - Total 3 meetups (average 35 participants)
Bibliography

- Books:
Discussion
**Story**
“a topically cohesive segment of news that includes two or more declarative independent clauses about a single event.”

**Event**
“something that happens at some specific time and place along with all necessary preconditions and unavoidable consequences.”

**Topic**
“a seminal event or activity, along with all directly related events and activities.”
Data Stream Mining

State-of-the-art

• Properties:
  – **approximate answer**, dependent on chosen accuracy
  – models based on a **summary** or "**sketch**" of the data stream in memory

• Requirements:
  – Process an **example at a time**, inspect it only one
  – Use **limited amount of memory**
  – Work in a **limited amount of time**
  – Be ready to predict at any time
State-of-the-art

• Community detection:
  – Based on modularity
  – Spectral Analysis

• Network is not static, evolves over time
  – Creation, growth and disband of communities

• Group Formation:
  – exploring the principles by which groups develop and evolve in large-scale social networks

• Information spreading:
  – Identification of “social sensors” that pass information quickly
  – Cascading behavior (in Blogs)
State-of-the-art

• Text representation models:
  – **unstructured text**: vector space model (VSM);
  – **feature extraction**: bag-of-words, entity recognition, summarization, sentiment analysis

• Text analysis:
  – **term trend approach**: trends in text streams (frequencies)
  – **semantic space approach** (category found in the collection)

• Topic extraction:
  – Latent Dirichlet Allocation (LDA), Dirichlet Compound Multinomial (DCM) mixtures and von-Mises Fisher (vMF) mixture models

• Event detection:
  – statistical methods (LSH), wavelets, topic models (LDA)
# FSD twitter corpus Results

<table>
<thead>
<tr>
<th>Topic description</th>
<th>On-topic</th>
<th>Broad topic type</th>
<th>P/U</th>
<th>Lag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amy Winehouse dies</td>
<td>1021</td>
<td>Celebrity/Human interest news</td>
<td>U</td>
<td>1h17m</td>
</tr>
<tr>
<td>Atlantis shuttle lands</td>
<td>49</td>
<td>Science and discovery news</td>
<td>P</td>
<td>0m</td>
</tr>
<tr>
<td>Betty Ford dies</td>
<td>14</td>
<td>Celebrity/Human interest news</td>
<td>U</td>
<td>N/A</td>
</tr>
<tr>
<td>Richard Bowes killed in riots in England</td>
<td>39</td>
<td>Acts of violence or war</td>
<td>U</td>
<td>39m</td>
</tr>
<tr>
<td>Flight 4896 crash</td>
<td>11</td>
<td>Accidents</td>
<td>U</td>
<td>1h52m</td>
</tr>
<tr>
<td>S&amp;P downgrade US credit rating</td>
<td>334</td>
<td>Financial news</td>
<td>P</td>
<td>N/A</td>
</tr>
<tr>
<td>US increases debt ceiling</td>
<td>89</td>
<td>Financial news, also New laws</td>
<td>P</td>
<td>0m</td>
</tr>
<tr>
<td>Terrorist attack in Delhi</td>
<td>39</td>
<td>Acts of violence or war</td>
<td>U</td>
<td>9m</td>
</tr>
<tr>
<td>Earthquake in Virginia</td>
<td>318</td>
<td>Natural disasters</td>
<td>U</td>
<td>2m</td>
</tr>
<tr>
<td>First victim of London riots dies</td>
<td>85</td>
<td>Acts of violence or war</td>
<td>U</td>
<td>0m</td>
</tr>
<tr>
<td>War criminal Goran Hadžić arrested</td>
<td>2</td>
<td>Legal/Criminal cases</td>
<td>U</td>
<td>22h18m</td>
</tr>
<tr>
<td>India and Bangladesh sign a border pact</td>
<td>4</td>
<td>Political and diplomatic meetings</td>
<td>P</td>
<td>N/A</td>
</tr>
<tr>
<td>Plane with Russian hockey team Lokomotiv crashes</td>
<td>277</td>
<td>Accidents</td>
<td>U</td>
<td>54m</td>
</tr>
<tr>
<td>Explosion in French nuclear plant in Marcoule</td>
<td>162</td>
<td>Accidents</td>
<td>U</td>
<td>57m</td>
</tr>
<tr>
<td>NASA announces there might be water on Mars</td>
<td>127</td>
<td>Science and discovery news</td>
<td>P</td>
<td>8m</td>
</tr>
</tbody>
</table>