Understanding the Diversity of Tweets in the Time of Outbreaks

Nattiya Kanhabua and Wolfgang Nejdl

L3S Research Center
Leibniz Universität Hannover, Germany
http://www.L3S.de
Love in the Time of Cholera

Book by Gabriel García Márquez


Published: 1985
Author: Gabriel García Márquez
Genre: Novel
Characters: Fermina Daza, Florentino Ariza, Dr. Juvenal Urbino, Lorenzo Daza, Aunt Escolástica

People also search for

One Hundred Years of Solitude
Gabriel García Márquez

Chronicle of a Death Foretold
Gabriel García Márquez

Of Love and Other Demons
Gabriel García Márquez

No One Writes to the Colonel
Gabriel García Márquez

The Autumn of the Patria...
Gabriel García Márquez

Search result from Google retrieved on 12 May 2013
Tweets in the Time of Outbreaks

Paper by Nattiya Kanhabua and Wolfgang Nejdl


Published: 1985
Author: Gabriel García Márquez
Genre: Novel
Characters: Fermina Daza, Florentino Ariza, Dr. Juvenal Urbino, Lorenzo Daza, Aunt Escolástica

People also search for

One Hundred Years of... Gabriel García Márquez
Chronicle of a Death Foretold Gabriel García Márquez
Of Love and Other Demons Gabriel García Márquez
No One Writes to the Colonel Gabriel García Márquez
The Autumn of the Patria... Gabriel García Márquez

Search result from Google retrieved on 12 May 2013
Motivation

• Numerous works use Twitter to infer the existence and magnitude of *real-world events* in real-time
  – Earthquake [Sakaki et al., 2010]
  – Predicting financial time series [Ruiz et al., 2012]
  – Influenza epidemics [Culotta, 2010; Lampos et al., 2011; Paul et al., 2011]

• In the medical domain, there has been a surge in detecting health related tweets for *early warning*
  – Allow a rapid response from authorities [Diaz-Aviles et al., 2012]
Health related tweets

- User status updates or news related to public health are common in Twitter
  - I have the mumps...am I alone?
  - my baby girl has a Gastroenteritis so great!! Please do not give it to mee
  - #Cholera breaks out in #Dadaab refugee camp in #Kenya http://t.co/....
  - As many as 16 people have been found infected with Anthrax in Shahjadpur upazila of the Sirajganj district in Bangladesh.
Web Observatory Application

Outbreak

twitter

Application
Challenge I. Noisy data

- Ambiguity
  - having several meanings
  - used in different contexts

- Incompleteness
  - missing or under-reported events
  - data processing errors
# Challenge I. Noisy data

<table>
<thead>
<tr>
<th>Category</th>
<th>Example tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature</td>
<td><em>A two hour train journey, Love In the Time of Cholera</em> ...</td>
</tr>
<tr>
<td>Music</td>
<td><em>Dengue Fever’s “Uku,” Mixed by Paul Dreux Smith Universal Audio</em>...</td>
</tr>
<tr>
<td>Marketing</td>
<td><em>Exclusive distributor of high quality #HIV/AIDS Blood &amp; Urine and #Hepatitis #Self -testers.</em></td>
</tr>
<tr>
<td>General</td>
<td><em>Identification of genotype 4 Hepatitis E virus binding proteins on swine liver cells: Hepatitis E virus</em>...</td>
</tr>
<tr>
<td>Negative</td>
<td><em>i dont have sniffles and no real coughing..well its coughing but not like an influenza cough.</em></td>
</tr>
<tr>
<td>Joke</td>
<td><em>Thought I had Bieber Fever. Ends up I just had a combo of the mumps, mono, measles &amp; the hershey squ</em>...</td>
</tr>
</tbody>
</table>
Challenge II. Dynamics

- **Time**
  - *seasonal* infectious diseases
  - *rare* and *spontaneous* outbreaks

- **Place**
  - *frequency* and *duration*
  - *levels of prevalence* or severity
Challenge II. Dynamics

- **Time**
  - Seasonal infectious diseases
  - Rare and spontaneous outbreaks

- **Place**
  - Frequency and duration
  - Levels of prevalence or severity

---

Number of RASFF notifications
(10^1 for salmonella, mycotoxins and heavy metals)

Hazards (level of notification)
- Mycotoxins (high)
- Salmonella (high)
- Heavy metals (high)
- Dioxins (medium)
- Aluminium (medium)
- Melamine (medium)
- Radioactivity (medium)
- DDT (low)
- Vibrio cholera (low)

[Ortais et al., 2010 in Journal of Food Research International]
Challenge II. Dynamics

• Time
  – **seasonal** infectious diseases
  – **rare** and **spontaneous** outbreaks

• Place
  – **frequency** and **duration**
  – **levels of prevalence** or severity
Challenge II. Dynamics

Figure 1
Total number of cholera outbreaks, 1974–2005.

[Emch et al., 2008 in International Journal of Health Geographics]
Problem Statement

• How to detect outbreaks for **general diseases**?
  – Previous works focus on a **limited number of diseases**, i.e., influenza or dengue, based on **supervised learning**

• How to take into account **temporal and spatial diversities** for outbreak detection?
  – Previous works **do not explicitly model** the diversity dimension
Contributions

• We conduct the **first study of temporal diversity** in Twitter

• A method to extract topic dynamics for outbreaks used as an **estimate of real-world statistics**

• A **correlation analysis** of temporal diversity and estimate statistics for 14 outbreak ground truths
System Framework

• Part I. Ground truth creation
  – Official outbreak reports
    • World Health Organization\(^1\)
    • ProMED-mail\(^2\)

• Part II. Creating Twitter time series
  1. medical condition
    • disease name, synonyms, pathogens, symptoms
  2. location
    • geographic expressions, geo-location, or user profile
    • 3 levels: country, continent, latitude

\(^1\)http://www.who.int
\(^2\)http://www.promedmail.org/
Ground Truths

- Extract events in a pipeline fashion
- Annotated documents
  - named entities (diseases, victims and locations)
  - temporal expressions
  - a set of sentences
- Event \( e: (v, m, l, t_e) \)
  - who (victim \( v \)) was infected
  - what (disease \( m \)) causes
  - where (location \( l \))
  - when (time \( t_e \))

[Kanhabua et al., 2012a]
Event Extraction

• An event is a **sentence containing two entities**
  – (1) medical condition and (2) geographic expression
  – A minimum requirement by domain experts

• A **victim** and the **time** of an event can be identified from the sentence itself, or its surrounding context

• Output: a set of **event candidates**

---

*Reported by World Health Organization (WHO) on 29 July 2012 about an ongoing *Ebola* outbreak in *Uganda* since the beginning of July 2012*
List of 14 Outbreaks

<table>
<thead>
<tr>
<th>ID</th>
<th>Disease</th>
<th>Country</th>
<th>Event Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>anthrax</td>
<td>Bangladesh</td>
<td>[11-May,18-Jun]</td>
</tr>
<tr>
<td>2</td>
<td>anthrax</td>
<td>India</td>
<td>[03-Jun,22-Jun]</td>
</tr>
<tr>
<td>3</td>
<td>botulism</td>
<td>Finland</td>
<td>[17-Oct,01-Nov]</td>
</tr>
<tr>
<td>4</td>
<td>botulism</td>
<td>France</td>
<td>[01-Sept,10-Sept]</td>
</tr>
<tr>
<td>5</td>
<td>cholera</td>
<td>Kenya</td>
<td>[11-Nov,03-Dec]</td>
</tr>
<tr>
<td>6</td>
<td>ebola</td>
<td>Uganda</td>
<td>[13-May,30-Dec]</td>
</tr>
<tr>
<td>7</td>
<td>ehec</td>
<td>Germany</td>
<td>[05-May,30-Jun]</td>
</tr>
<tr>
<td>8</td>
<td>leptospirosis</td>
<td>Denmark</td>
<td>[02-Jul,23-Jul]</td>
</tr>
<tr>
<td>9</td>
<td>leptospirosis</td>
<td>Philippines</td>
<td>[27-Jun,15-Jul]</td>
</tr>
<tr>
<td>10</td>
<td>mumps</td>
<td>Canada</td>
<td>[10-Jun,17-Aug]</td>
</tr>
<tr>
<td>11</td>
<td>mumps</td>
<td>United States</td>
<td>[27-Sept,11-Oct]</td>
</tr>
<tr>
<td>12</td>
<td>norovirus</td>
<td>France</td>
<td>[16-Jul,25-Jul]</td>
</tr>
<tr>
<td>13</td>
<td>rubella</td>
<td>Fiji</td>
<td>[26-Jul,09-Aug]</td>
</tr>
<tr>
<td>14</td>
<td>rubella</td>
<td>New Zealand</td>
<td>[15-Aug,19-Aug]</td>
</tr>
</tbody>
</table>
Matching Tweets

1. The user enters a query for an event: <medical condition, location, time, normalization>

2. The system retrieves and displays results related to the event.

3. Summary of the event: including estimated dates and victims/cases

4. Time series visualization for different locations

5. Cross correlation results of Twitter and official health report data

6. The system returns the list of all documents related to the event

[Kanhabua et al., 2012b]
Matching Tweets

(1) Estimate of event magnitude

(2) Twitter time series data

[Kanhabua et al., 2012b]
Identifying Topic Dynamics

- **Input:** time series data of *relevant* tweets
- For each time $t_k$, *unsupervised clustering* by topic
- Filter result topics by *cluster quality*
- **Output:** *outbreak-related topic* time series
### Outbreak Negative Terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>anthrax</td>
<td>concert, fly, song, castle, film, novel, book, music, band</td>
</tr>
<tr>
<td>botulism</td>
<td>plastic+surgery, cosmetic, dermatologic, soccer, film, book, football, show, concert, band, movie, music</td>
</tr>
<tr>
<td>cholera</td>
<td>love, album, band, music, concert, music, song, book</td>
</tr>
<tr>
<td>ebola</td>
<td>film, book, football, show, concert, band, movie, music</td>
</tr>
<tr>
<td>mumps</td>
<td>multifrontal+massively, programming+system, processing+system, concert, film, novel, book, music, band, movie,</td>
</tr>
<tr>
<td>norovirus</td>
<td>film, book, football, show, concert, band, movie, music</td>
</tr>
</tbody>
</table>
## Outbreak Topic Dynamics

<table>
<thead>
<tr>
<th>Cluster Name</th>
<th>Representative Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tapenade</td>
<td>tapenade linked eight botulism cases france</td>
</tr>
<tr>
<td>Recall</td>
<td>botulism france the official recall with photo</td>
</tr>
<tr>
<td>Serious</td>
<td>france people seriously ill after botulism life support infected tapenade products produced cavaillon provence</td>
</tr>
<tr>
<td>Food Agency</td>
<td>warning botulism outbreak france the food standards agency warning people not consume certain brand</td>
</tr>
<tr>
<td>Botulism France</td>
<td>botulism toll climbs france company not inadequate processing</td>
</tr>
<tr>
<td>Life Support</td>
<td>people sick and life support with botulism france the culprit olive tapenade from unlicensed vendor</td>
</tr>
</tbody>
</table>

07 Sep 2011

08 Sep 2011
Diversity Metric

• Refined Jaccard Index (RDJ-index)
  – average Jaccard similarity of all object pairs

\[ RDJ = \frac{2}{n(n-1)} \sum_{i < j} JS(O_i, O_j) \]

1 \leq i < j \leq n

• **Note**: lower RDJ corresponds to **higher diversity**
• **Problem**: “All-Pair comparison”
• **Solution**: Estimation algorithms with probabilistic error bound guarantees

[Deng et al., 2012]
Diversity Metric

- Refined Jaccard Index (RDJ-index)
  - average Jaccard similarity of all object pairs

\[
RDJ = \frac{2}{n(n-1)} \sum_{i<j} JS(O_i, O_j)
\]

1 ≤ i < j ≤ n

- **Note**: lower RDJ corresponds to higher diversity
- **Problem**: “All-Pair comparison”
- **Solution**: Estimation algorithms with probabilistic error bound guarantees

[Notes: lower RDJ corresponds to higher diversity, Problem: “All-Pair comparison”, Solution: Estimation algorithms with probabilistic error bound guarantees]
Estimate Algorithms

- **Input**: Relative error $e$, accuracy confidence $d$
- **Output**: Estimated RDJ value

$$
\Pr \left[ \left| \frac{\hat{RDJ} - RDJ}{RDJ} \right| > \varepsilon \right] < \delta
$$

- **Algorithms**: SampleDJ, TrackDJ (claims and proofs in [Deng et al., 2012])

Temporal Diversity

\[ \text{div}_{\text{mix}}(t_k) = \alpha \cdot \text{div}_{\text{entity}}(t_k) + (1 - \alpha) \cdot \text{div}_{\text{term}}(t_k) \]

- where \( \alpha \) underlines the importance of both metrics. The value will be empirically determined.
Temporal Diversity

Temporal diversity of 'botulism' tweets

number of tweets

date


R.D.-index

#tweet  div(andy)  div(arm)
Experimental Settings

• Official outbreak reports
  – ~3,000 ProMED-mail reports from 2011

• Twitter data
  – ~1,200 health-related terms
  – Over 112 millions of tweets from 2011

• Series of NLP tools including
  – OpenNLP (tokenization, sentence splitting, POS tagging)
  – OpenCalais (named entity recognition)
  – HeidelTime (temporal expression extraction)
Results

- Identified topics show similar trends during the known time periods of real-world outbreaks.
- Diversity reflects how the language (i.e., terms and locations) are used differently.
- Div(entity) highly correlates with topic dynamics for some diseases, i.e., mumps, ebola, botulism and ehec.
- Div(term) shows correlation with topic dynamics for cholera, anthrax and rubella.
Conclusions

- Study of detecting real-world outbreaks in Twitter
- Proposed method to compute temporal diversity
- Correlation analysis of temporal diversity and estimate magnitude of outbreaks
- Future work: improve diversity measures
  1. new representations for tweets, e.g., using other types of entities
  2. employ a semantic-based similarity measurement
References


