Exploiting Social Networks for Disease Surveillance
Web Observatory for Dengue

Wagner Meira Jr.

Universidade Federal de Minas Gerais, Brazil
InWeb - Brazilian National Institute of Science and Technology for the Web

Sep 12, 2011
INWeb – Brazilian National Institute of Science and Technology for the Web

To develop models, algorithms and technologies to contribute to the integration of the Web with our society. As a result, we expect more effective and secure distribution of information, more efficient and useful applications, so that the Web can become a vector for social and economic changes in the country.
INWeb’s Integrated View
Research Challenges

1. Identification, characterization and modeling of user interests and behavioral patterns on the web as well as of the social networks established among them.

2. Treatment of the information that circulates on the various networks of the web.

3. Delivery of information in a satisfying way regardless of time and place.
Research Tracks

1. Social Networks
   (Coordinator: Virgilio Almeida)

2. User Behavior and Interaction Modeling
   (Coordinator: Jussara Almeida)

3. Information Retrieval
   (Coordinator: Nivio Ziviani)

4. Web Data Management
   (Coordinator: Alberto Laender)

5. Parallel and Distributed Systems
   (Coordinator: Dorgival Guedes)

6. Knowledge Discovery
   (Coordinator: Wagner Meira Jr.)
Motivation

- There is an increasing use of the Web in events of overall interest such as politics and sports.
- Major motivations are the lack of a central control and the fast information propagation.
- Recently, there has been an emphasis on "what you are doing" instead of "who you are".

Challenge

Qualify, quantify, and summarize the content being exchanged in the various Internet-related media on line and evaluate its impact on specific events.
Web Observatory

On line tool for capturing, analyzing and presenting the dynamics of a given scenario on the Web.

Research issues:

- Capture and process information on line
- Source diversity
- Information extraction and entity resolution
- Effectiveness of metrics and metaphors
- Scalability and efficiency
Scenarios

- Soccer World Cup
- Brazilian Presidential Elections
- Dengue Epidemics
- Brazilian National Soccer League
- Public Safety
- Brand reputation
Twitter is a unique social media channel, in the sense that users discuss and talk about the most diverse topics, including their health conditions.

Traditional disease surveillance comprises a set of epidemiological procedures that monitor the spread of a disease and determine how it is spreading.

Social media channels, such as Twitter, offer a continuous source of epidemic information, arming public health agencies with the ability to perform real-time surveillance.
Background on dengue

- Dengue is a mosquito-borne infection that causes a severe flu-like illness, and sometimes a potentially lethal complication.
- Approximately 2 billion people from more than 100 countries are at risk of infection and about 50 million infections occur every year worldwide.
- Outbreaks tend to occur every year during the rainy season but there is large variation of the degree of the epidemic in areas with similar rainfall.
Background on dengue

- Current strategies for prediction of dengue epidemics are based on surveillance of insects, which provide only a rough estimate of cases.
- Once disease outbreaks are detected in a certain area, efforts need to be concentrated to avoid further cases and to optimize treatment and staff - number of cases may reach several hundred thousands.
- In Brazil, where there is a functional disease reporting system, detection of important outbreaks may take a few weeks, leading to loss of precious time to address the epidemic.
Goals

- To analyze how dengue epidemics manifests in Twitter and to what extent that information can be used for surveillance.
- To design and implement an active surveillance framework that analyzes how social media reflects epidemics based on a combination of four dimensions: volume, location, time, and public perception.
- To exploit user generated content available in online social media to predict the dengue epidemics.
Datasets

Data from two different sources:

- official dengue reports from the Brazilian Health Ministry
  - contains the number of dengue cases per city, notified between 2007 and 2011

- twitter messages mentioning the word “dengue”
  - from 2006 to July 2009: 27,658 tweets, out of which 90.27% are from 2009
  - from December 2010 to April 2011: 465,444 tweets
Active dengue surveillance based on four dimensions:
- Public perception
- Volume
- Location
- Time

Methodology steps
- Content analysis
- Correlation analysis
- Spatio-temporal analysis
- Surveillance
Content analysis

- Content analysis is employed in order to
  - understand user’s attitude associated with the tweets
  - reduce noise by focusing only on tweets that are related to dengue cases
- Classification techniques may be used to estimate sentiments expressed in tweets

Steps:

1. Determine the sentiment categories
2. Create a representative training dataset
3. Build a classification model
4. Classify tweets
Content analysis

Social Networks for Disease Surveillance

InWeb

Web Observatory

Dengue’s Observatory

Methodology

Case Study

Surveillance

Summary

Sep 12, 2011

Meira
Content analysis
Content analysis

- Determine the sentiment categories
  - **Personal experience:** “You know I have had dengue?”
  - **Ironic/sarcastic tweets:** “My life looks like a dengue-prone steady water”
  - **Opinion:** “The campaign against dengue is very cool”
  - **Resource:** “Dengue virus type 4 in circulation”
  - **Marketing:** “Everybody must fight dengue. Brazil relies on you”
Content analysis

- Create a representative training dataset
  - Selective sampling strategy: small and representative training dataset

- Classify tweets
  - Lazy associative classifier (LAC)
  - LAC estimates the likelihood of each sentiment in a tweet
Content analysis

Sentiment distribution over time

Sentiment Distribution

% of total

Months

Marketing
Resource
Opinion
Ironic/sarcastic
Personal Experience
Is personal experience a good indicator of dengue’s incidence?
Correlation analysis

- Does Twitter traffic correlates with dengue’s incidence?
- Linear regression model
- Variables:
  1. The volume of tweets related to dengue, posted by Brazilian users ($\#tweets$)
  2. The ratio of tweets expressing personal experience, posted by Brazilian users ($PTPE$)
Correlation analysis

- Does Twitter traffic correlates with dengue’s incidence?
- Linear regression model
- Variables:
  1. The volume of tweets related to dengue, posted by Brazilian users (#tweets)
  2. The ratio of tweets expressing personal experience, posted by Brazilian users (PTPE)

Linear regression models

\[
\#cases_t = \beta_0 + \beta_1 \times \#tweets_t + \beta_2 \times \#tweets_{t-1} + \epsilon
\]

\[
\#cases_t = \beta_0 + \beta_1 \times PTPE_t + \beta_2 \times PTPE_{t-1} + \epsilon
\]
Correlation analysis

- Does Twitter traffic correlates with dengue’s incidence?
- Linear regression model
- Variables:
  1. The volume of tweets related to dengue, posted by Brazilian users (#tweets)
  2. The ratio of tweets expressing personal experience, posted by Brazilian users (PTPE)
- Linear regression models

\[
\text{#cases}_t = \beta_0 + \beta_1 \times \text{#tweets}_t + \beta_2 \times \text{#tweets}_{t-1} + \epsilon
\]
\[R^2 = 0.7829\]

\[
\text{#cases}_t = \beta_0 + \beta_1 \times \text{PTPE}_t + \beta_2 \times \text{PTPE}_{t-1} + \epsilon
\]
\[R^2 = 0.9578\]
Correlation Analysis

Manaus

Personal experience, notifications and symptom perception

From November, 2010 to May, 2011
Manaus

Cross-correlation between personal experience and symptom perception from November, 2010 to May, 2011
Correlation Analysis

Rio de Janeiro

Personal experience, notifications and symptom perception

From November, 2010 to May, 2011
Correlation Analysis

Rio de Janeiro

Cross-correlation between personal experience and symptom perception from November, 2010 to May, 2011
Spatio-temporal analysis

- Determine groups of cities that are **near** from each other and have **similar** dengue incidence rates in a given time **period**
- Info enables government agencies to timely concentrate efforts on critical areas
- ST-DBSCAN is a density-based clustering algorithm and is used for clustering spatial-temporal data

Steps:

1. Calculate the incidence rate associated with each city
2. Determine the input parameters and run ST-DBSCAN
3. Assess clustering quality through Rand Index
Spatio-temporal analysis

Observatório da Dengue

Menções à dengue no Twitter no mês de fev/2011

Clique nos pontos do mapa para informações

<table>
<thead>
<tr>
<th>Cidade</th>
<th>Pop.</th>
<th>Tweets</th>
<th>Tx.Inc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belém</td>
<td>377547</td>
<td>14</td>
<td>1.8230e-04</td>
</tr>
<tr>
<td>Brumadinho</td>
<td>34013</td>
<td>1</td>
<td>1.4289e-04</td>
</tr>
<tr>
<td>Contagem</td>
<td>603048</td>
<td>22</td>
<td>1.7922e-04</td>
</tr>
<tr>
<td>Ibirapuera</td>
<td>159026</td>
<td>4</td>
<td>1.2110e-04</td>
</tr>
<tr>
<td>Itabira</td>
<td>109551</td>
<td>6</td>
<td>2.7305e-04</td>
</tr>
<tr>
<td>Ilha Bela</td>
<td>85396</td>
<td>1</td>
<td>5.2124e-05</td>
</tr>
<tr>
<td>João Paulo</td>
<td>73451</td>
<td>4</td>
<td>2.7146e-04</td>
</tr>
<tr>
<td>Montes Claros</td>
<td>32973</td>
<td>1</td>
<td>1.4765e-04</td>
</tr>
<tr>
<td>Ribeirão das Neves</td>
<td>296376</td>
<td>2</td>
<td>2.6665e-05</td>
</tr>
<tr>
<td>Sabará</td>
<td>126219</td>
<td>3</td>
<td>1.1399e-04</td>
</tr>
<tr>
<td>Santa Barbara</td>
<td>27850</td>
<td>1</td>
<td>1.7627e-04</td>
</tr>
</tbody>
</table>
Spatio-temporal analysis

- Evaluated two metrics
  - the volume of tweets
  - the PTPE value
Spatio-temporal analysis

- Evaluated two metrics
  - the volume of tweets  \( \text{Rand Index} = 0.8506 \)
  - the PTPE value  \( \text{Rand Index} = 0.8914 \)
Spatio-temporal analysis

- Evaluated two metrics
  - the volume of tweets \( \text{Rand Index} = 0.8506 \)
  - the PTPE value \( \text{Rand Index} = 0.8914 \)
Surveillance

- **Strategy**: Analyze the proportion of tweets expressing personal experience in a weekly basis
- **Intuition**: an abrupt increase in PTPE may indicate outbreaks
- **Presentation**:
  - heat maps showing PTPE variations
  - temporal graphs
z-score based alarm system

Surveillance

Rio de Janeiro/RJ, 14/11/10
Summary

- Twitter data is useful for disease surveillance
- Four-dimension evaluation methodology: volume, location, time and content
- Personal experience tweets are highly correlated with dengue incidence
- Even simple alarm systems may help for early detection of outbreaks
- Dengue Observatory has been adopted by the Brazilian government as a rumour surveillance tool
Thank you!

Wagner Meira Jr.
meira@dcc.ufmg.br

http://observatorio.inweb.org.br/dengue/