EVENT DETECTION IN TWITTER

Jianshu Weng and Bu-Sung Lee, Francis
HP Labs Singapore
AGENDA

– Motivation

– Proposed Method
  • Wavelet analysis
  • Event Detection with Clustering of Wavelet-based Signals (EDCoW)

– Evaluation
  • Experimental study
  • General Election case study

– Conclusion
MOTIVATION

Benefits

Different perspective from the traditional media

Disaster early warning

Challenges

200 million Tweets per day
Thursday, June 30, 2011

- Winner -
“Pointless Babble”

2nd
“Conversational”

3rd
“Pass-Along Value”
MOTIVATION (CONT’D)

– We need a way to detect “events” out of “babble”-flooded twitter content stream.
  • Fast
  • Low storage requirement
  • Can tell how “significant” the event is.

– Existing techniques: usually first-order signals, e.g. TFIDF/DFIDF
  • Redundant information
  • Many features would seem to have similar waveforms when we apply feature-pivot methods
PROPOSED METHOD

– EDCoW: Event Detection with Clustering of Wavelet-based Signals

– Steps:
  • Construct 1\textsuperscript{st}-order signals for individual features based on their DFIDF scores, for a given interval
  • Wavelet analysis on the 1\textsuperscript{st}-order signals to construct 2\textsuperscript{nd}-order signals
  • Filtering
  • Clustering with modularity-based graph cut
  • Measuring the event “significance”
SIGNAL CONSTRUCTION

– First order: DFIDF

\[ y_x(t) = \frac{N_x(t)}{N(t)} \times \log\left( \frac{\sum_{i=1}^{T_c} N(i)}{\sum_{i=1}^{T_c} N_x(i)} \right) \]

– \( N_x(i) \) is the number of tweets containing feature \( x \) at time point \( i \), and \( N(i) \) is the total number of tweets at time point \( i \).
SIGNAL CONSTRUCTION (CONT’D)

– Second stage:
  • applying wavelet analysis to compress the DFIDF signals to keep only the changes within a number of sample points in the first stage.
    – H is the normalized wavelet entropy in a sliding window
    – Each 2nd-stage sampling point captures how much the change in the wavelet entropy is

\[
S'_w(t') = \begin{cases} 
  \frac{H_{t^*} - H_{t'-1}}{H_{t'-1}} & \text{if } (H_{t^*} > H_{t'-1}); \\
  0 & \text{otherwise}
\end{cases}
\]

• Why 2nd-stage signal?
  – Make the change in usage pattern more salient, easier to measure the similarity between features
  – Compression: only non-trivial changes are stored
SIGNAL CONSTRUCTION (CONT’D)

1\textsuperscript{st} stage \quad t = 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad \ldots \ldots \quad T_c

\begin{align*}
D_1 & \\
D_2 & \\
D_2^* & \\
\end{align*}

2\textsuperscript{nd} stage, \quad t' = 0 \quad 1 \quad 2 \quad 3 \quad \ldots \ldots \quad T'_c

\Delta = 3
SIGNAL CONSTRUCTION (CONT’D)

1\textsuperscript{st}-order

2\textsuperscript{nd}-order
FILTERING OF TRIVIAL FEATURES

- Measure the auto-correlation of the signal corresponding to a feature
  \[(f \ast g)(t) = \sum f \ast (\tau)g(t + \tau)\]

- Filtering with median absolute deviation (MAD)
  \[MAD(S^T) = \text{median}(|A^T_i - \text{median}(A^T_i)|)\]

- 5% features remain
MEASURE THE SIMILARITY BETWEEN FEATURES

– Similarity measured as the cross-correlation between signals
  \[(f \star g)(t) = \sum f \star (\tau)g(t + \tau)\]

– Ignore the similarity between features if it is too weak
  • Using Median Absolute Deviation

– \(O(n^2)\) complexity, but still tractable with small portion of remaining features
CLUSTERING FEATURES TO DETECT EVENTS

– Organize the cross-correlation between features as a matrix $M$

  • Adjacency matrix of a graph $G$.

\[
\begin{pmatrix}
0 & 0.8 & 0.5 \\
0.8 & 0 & 0.6 \\
0.5 & 0.6 & 0
\end{pmatrix}
\]

\[
G \quad M
\]
CLUSTERING FEATURES TO DETECT EVENTS (CONT’D)

– Event detection can be formulated as a graph partitioning problem.

– Modularity can be applied to measure the quality of the partitioning

  • the sum of weights of all the edges that fall within sub-graphs (after partitioning) subtracted by the expected edge weight sum if edges were placed at random

\[
Q = \frac{1}{2m} \sum_{ij} (w_{ij} - \frac{d_i \cdot d_j}{2m}) \delta_{c_i,c_j}
\]

• The goal is to partition the graph so that \( Q \) value is maximized.
MEASURE THE SIGNIFICANCE OF THE EVENT

\[- \epsilon = (\sum w_{ij}^c) \times \frac{e^{1.5n}}{(2n)!}, \quad n = |V^c|\]

– The first part \(\sum w_{ij}^c\) sums up all the cross correlation values between signals (i.e. features) associated with an event.

– The second part \(\frac{e^{1.5n}}{(2n)!}\) discounts the significance if the event is associated with too many features.
  
  • it is not reasonable for an event to be associated with too many words.
EVALUATION

- Experimental

- Data

  • Obtain the top 1000 Singapore-based Twitter users with the most followers from http://twitaholic.com/. $U$
  • For each user in $U$, include her Singapore-based followers and friends within 2 hops. $U^*$
  • For each user in $U^*$, collect the tweets published in June 2010.

- Remove stop words, apply stemming
<table>
<thead>
<tr>
<th>Day</th>
<th>Event</th>
<th>$\epsilon$ value</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3</td>
<td>democrat, naoto</td>
<td>0.417</td>
<td>Ruling Democratic Party of Japan elected Naoto Kan as chief.</td>
</tr>
<tr>
<td>2.</td>
<td>ss501, suju</td>
<td>0.414</td>
<td>Korean popular bands Super Junior’s and SS501’s performance on mubank.</td>
</tr>
<tr>
<td>4.</td>
<td>music, mubank</td>
<td>0.401</td>
<td>Related to Event 2, mubank is a popular KBS entertainment show.</td>
</tr>
<tr>
<td>5.</td>
<td>shindong, youngsaeng</td>
<td>0.365</td>
<td>Related to Event 2, Shindong and Youngsaeng are member of the two bands.</td>
</tr>
<tr>
<td>5.</td>
<td>junior, eunhyuk</td>
<td>0.124</td>
<td>Related to Event 2, Eunhyuk is a member of super junior.</td>
</tr>
<tr>
<td>5</td>
<td>robben, break</td>
<td>0.404</td>
<td>No clear corresponding real-life event</td>
</tr>
<tr>
<td>7</td>
<td>kobe, kristen</td>
<td>0.417</td>
<td>Two events: Kristen Stewart won some MTV awards, and Kobe Bryant in a NBA match.</td>
</tr>
<tr>
<td>8.</td>
<td>#iphone4, ios4, iphone</td>
<td>0.416</td>
<td>iPhone 4 released during WWDC 2010</td>
</tr>
<tr>
<td>9</td>
<td>reformat, hamilton</td>
<td>0.391</td>
<td>No clear corresponding real-life event</td>
</tr>
<tr>
<td>10</td>
<td>avocado, commence, ongoing</td>
<td>0.124</td>
<td>No clear corresponding real-life event</td>
</tr>
<tr>
<td>11</td>
<td>#failwhale, twitter</td>
<td>0.360</td>
<td>A number of users complained they could not use twitter due to over-capacity.</td>
</tr>
<tr>
<td>12</td>
<td>vuvuzela, soccer</td>
<td>0.387</td>
<td>People started to talk about world cup.</td>
</tr>
<tr>
<td>13</td>
<td>#svk, #svn</td>
<td>0.418</td>
<td>#svk and #svn represent Team Slovakia and Slovenia in World Cup 2010.</td>
</tr>
<tr>
<td>14</td>
<td>#kor, greec, #gre</td>
<td>0.102</td>
<td>A match between South Korea and Greece in World Cup 2010.</td>
</tr>
<tr>
<td>15</td>
<td>whale, twitter</td>
<td>0.417</td>
<td>Similar as Event 10.</td>
</tr>
<tr>
<td>16</td>
<td>lippi, italy</td>
<td>0.326</td>
<td>Italy football team coach Marcello Lippi made some comments after a match in World Cup 2010.</td>
</tr>
<tr>
<td>17</td>
<td>drogba, ivory</td>
<td>0.417</td>
<td>Football player Drogba from Ivory Coast is given special permission to play in World Cup 2010.</td>
</tr>
<tr>
<td>18</td>
<td>#prk, #bra, north</td>
<td>0.114</td>
<td>A match between North Korea and Brazil in World Cup 2010.</td>
</tr>
<tr>
<td>19</td>
<td>orchard, flood</td>
<td>0.357</td>
<td>Flood in Orchard Road.</td>
</tr>
<tr>
<td>17</td>
<td>greec, #gre, nigeria</td>
<td>0.122</td>
<td>A match between Greece and Nigeria in World Cup 2010.</td>
</tr>
<tr>
<td>18</td>
<td>#srb, podolski</td>
<td>0.403</td>
<td>Podolski is a member of Team Germany in World Cup 2010.</td>
</tr>
</tbody>
</table>

Table 1: All Events Detected by EDCoW in June 2010
EVALUATION (CONT’D)

– Practical case study

– Singapore General Election 2011
  • 7 parties
  • Most Singaporeans vote for the first time
  • Social media is allowed as a campaign platform for the first time
Volume Monitoring
Online popularity alone won’t get you elected

Being a hot topic in cyberspace will not determine whether you win or lose an election.
By JASMINE OSADA

The General Election 2011, which saw candidates and voters engage each other directly on Facebook and Twitter pages, was a watershed for social media.

Data from HP Labs Singapore the computer maker’s research arm showed that social media chatter increased 30 times between April 13, when the online activity was first tracked, and Nomination Day on April 27.

Due to the relaxation of online campaigning rules and the ease in which information can be shared through social platforms, the online realm had created overnight celebrities out of some candidates while subjecting others to scrutiny and criticism.

Ms Tin Pei Ling, from the People’s Action Party (PAP) who won the contest for Marine Parade GRC, bore the brunt of online criticisms. At least nine Facebook groups were created against the newly elected Member of Parliament. The largest, called “I Do Not Want Tin Pei Ling In Parliament” had some 45,000 members as of Monday.

On the other hand, candidates like the National Solidarity Party’s (NSP) Ms Nicole Seah, shot to fame with the help of social media. The 24-year-old candidate, who was part of the NSP team which contested in Marine Parade GRC, was Singapore’s most “liked” political figure by press time. She had 96,500 fans on the social networking site. Coming in second was Minister Mentor Lee Kuan Yew’s Facebook page with 85,610 fans.

However, being well-liked online does not necessarily mean that one will score a victory at the polls. Social networking posts analyzed by HP Labs Singapore’s sentiment analysis software indicated that the NSP had garnered the most number of positive posts of all the seven political parties. Yet, the party failed to win a single contest at the polls. The ruling party PAP, which won 81 of 87 seats, came in at third place in HP’s list after the Workers’ Party.

HP’s sentiment analysis software makes use of data from both Facebook and Twitter to determine whether a certain topic is discussed positively or negatively online.

A large percentage of the social networking posts were made during the party rallies, as well as netizens blogging and uploading photos of the rallies to their Facebook pages. Telcos here have reported a marked increase in the use of mobile data during the two-week campaign at rally sites, where networks experienced congestion issues from too many users trying to access data services at the same time.

Mr Ivan Lim, M1’s deputy director of corporate communications and investor relations, told Digital Life that the telco saw network traffic increase two to four times at the more popular rallies, especially at those held by the Workers’ Party.

Ms Cassie Fong, StarHub’s corporate communications manager, said the telco saw a traffic increase of up to 15 per cent at certain rally sites as well.

Both telcos said there had been some congestion issues, such as users having to make more than one attempt to establish a voice call or experiencing slow data speeds. However, they noted that services promptly returned to normal once the crowd had thinned out.

A SingTel spokesman, who also acknowledged a significant increase in mobile broadband traffic, said traffic peaked between 7pm and 11pm during election rallies.

osadaj@sph.com.sg

WHAT NETIZENS WERE INTERESTED IN

Political parties ranked based on the amount of positive posts they received in social networking sites:
1. National Solidarity Party
2. Workers’ Party
3. People’s Action Party
4. Singapore Democratic Party
5. Singapore Democratic Alliance
6. Singapore People’s Party
7. Reform Party

(Source: HP Labs Singapore Sentiment analysis software)
SOME TWEETS

• PAP is changing! @PAPSingapore: Join PM Lee's live chat at the PAP FB page TONIGHT at 8pm. http://on.fb.me/kg4qiS #sgelections #sgelection

• say Yes to Tin Pei Ling & get a free Kate Spade wallet

• Potong Pasir comes back to big boss PAP. Let them upgrade already, 5 years later go back to SPP. Smart tactic residents.
CONCLUSION

- Text analysis on Twitter content reveals discussion about “real-life” events

- Spaces for improvement:
  - Combine discussion-based similarity (signal cross correlation) with other factors
    - N-Gram
    - Semantic similarity
  - Further exploit the social relationship among users
  - Larger scale study
Q&A