Detecting Sentiment Change in Twitter Streaming Data

Albert Bifet, Geoff Holmes, Bernhard Pfahringer and Ricard Gavaldà

University of Waikato
Hamilton, New Zealand

Laboratory for Relational Algorithmics, Complexity and Learning LARCA
UPC-Barcelona Tech, Catalonia

Castro, 19 October 2011
WAPA 2011
Detecting Sentiment Change in Twitter Streaming Data
Detecting Sentiment Change in Twitter Streaming Data

**Problem**: How to analyze Twitter data on real time
Detecting Sentiment Change in Twitter Streaming Data

Twitter Streaming API: API for accessing Twitter in real-time
Detecting Sentiment Change in Twitter Streaming Data

MOA is an open source project for data stream mining, for analyzing big data on real time
Detecting Sentiment Change in Twitter Streaming Data

**Sentiment analysis**: analyze tweets with positive :) or negative :( tweets
Detecting Sentiment Change in Twitter Streaming Data

**Problem:** We need to convert tweet texts into a sparse vector of features on real-time.
Detecting Sentiment Change in Twitter Streaming Data

Real-time means (i) change adaption (ii) fast: can not store tweets on memory
Detecting Sentiment Change in Twitter Streaming Data

**Solution:** MOA-TWEETREADER, a package to connect MOA with Twitter
Detecting Sentiment Change in Twitter Streaming Data

MOA-TWETREADER consists in (i) Adaptive Twitter filter (ii) Frequent item miner (iii) Change Detector
Detecting Sentiment Change in Twitter Streaming Data

\[ f_{i,j} = \frac{\text{freq}_{i,j}}{\sum_{\ell} \text{freq}_{\ell,j}} \]  
(number of times a word appears in the document)

\[ \text{idf}_i = \log \frac{N}{n_i} \]  
(inverse frequency of the word in the corpus)

\[ w_{i,q} = f_{i,j} \cdot \text{idf}_i \]

**MOA-TWEETREADER** Adaptive Twitter filter: online tf-idf
Space Saving (Metwally et al.)

1. $T \leftarrow \emptyset$
2. for every term $i$
3.   do if $i \in T$
4.       then $freq[i] \leftarrow freq[i] + 1$
5.   else if $|T| < k$
6.       then $T \leftarrow T \cup \{i\}$
7.       $freq[i] \leftarrow 1$
8. else $\triangleright$ Replace the item with lower freq.
9.   $j \leftarrow \arg\min_{j \in T} freq[j]$
10.  $T \leftarrow T \cup \{i\} \setminus \{j\}$
11.  $freq[j] \leftarrow freq[j] + 1$

Twitter

MOA-TweetReader Frequent item miner : Space Saving
Detecting Sentiment Change in Twitter Streaming Data

Space Saving (Metwally et al.)

```
T ← ∅
for every term i
do if i ∈ T
    then freq[i] ← freq[i] + 1
else if |T| < k
    then ▷ Add a new item
        T ← T ∪ {i}
        freq[i] ← 1
else ▷ Replace the item with lower freq.
    j ← arg min_{j ∈ T} freq[j]
    T ← T ∪ {i} \ {j}
    freq[j] ← freq[j] + 1
```

Space Saving is the frequent item algorithm for streams with best performance results
Detecting Sentiment Change in Twitter Streaming Data

**SPACE SAVING exponentially decayed (Cormode et al)**

1. $T \leftarrow \emptyset$
2. for every term $i$ with timestamp $t_i$
3.   do if $i \in T$
4.     then $freq[i] \leftarrow freq[i] + \exp(\lambda t_i)$
5.   else if $|T| < k$
6.     then $\triangleright$ Add a new item
7.       $T \leftarrow T \cup \{i\}$
8.       $freq[i] \leftarrow 1$
9.   else $\triangleright$ Replace the item with lower freq.
10.  $j \leftarrow \arg\min_{j \in T} freq[j]$
11.  $T \leftarrow T \cup \{i\} \setminus \{j\}$
12.  $freq[j] \leftarrow freq[j] + \exp(\lambda t_j)$

Improvement to **SPACE SAVING**: space saving with exponential decay, or using **ADWIN**
Detecting Sentiment Change in Twitter Streaming Data

**ADWIN: Adaptive Windowing Algorithm**

1. Initialize Window $W$
2. for each $t > 0$
   3. do $W \leftarrow W \cup \{x_t\}$ (i.e., add $x_t$ to the head of $W$)
   4. repeat Drop elements from the tail of $W$
   5. until $|\hat{\mu}_W - \hat{\mu}_W| < \varepsilon_c$ holds
   6. for every split of $W$ into $W = W_0 \cdot W_1$
   7. output $\hat{\mu}_W$

---

Improvement to *Space Saving*: space saving with exponential decay, or using ADWIN
Detecting Sentiment Change in Twitter Streaming Data

**SPACE SAVING ADWIN**

1. $T \leftarrow \emptyset$
2. for every term $i$ with timestamp $t_i$
   
3.     do if $i \in T$
4.         then Insert 1 into $\text{ADWIN}[i]$ and 0 to other $\text{ADWINs}$
5.     else if $|T| < k$
6.         then ▷ Add a new item
7.             $T \leftarrow T \cup \{i\}$
8.             Init $\text{ADWIN}[i]$
9.             Insert 1 into $\text{ADWIN}[i]$ and 0 to other $\text{ADWINs}$
10. else ▷ Replace the item with lower freq.
11.     $j \leftarrow \arg\min_{j \in T} \text{freq}[j]$
12.     $T \leftarrow T \cup \{i\} \setminus \{j\}$
13.     Insert 1 into $\text{ADWIN}[j]$ and 0 to other $\text{ADWINs}$

**Improvement to SPACE SAVING:** space saving with exponential decay, or using $\text{ADWIN}$
Detecting Sentiment Change in Twitter Streaming Data

Experiments: Frequency and ranking of twitter data follows a Zipf distribution.
Detecting Sentiment Change in Twitter Streaming Data

Space Saving Adwin is able to adapt automatically
Detecting Sentiment Change in Twitter Streaming Data

Toyota crisis: during end of 2009 and beginning of 2010 Toyota had problems with accelerator pedals and had to recall millions of cars
Detecting Sentiment Change in Twitter Streaming Data

Recommended reading “Toyota under fire”
There was a gap between the time that our U.S. colleagues realised that this was an urgent situation and the time that we realised here in Japan that there was an urgent situation going on in the U.S. It took three months for us to recognise that this had turned into a crisis. In Japan, unfortunately, until the middle of January we did not think that this was really a crisis.

Akio Toyoda

Recommended reading “Toyota under fire”
## Detecting Sentiment Change in Twitter Streaming Data

Following twitter data sentiment, and changes in MOA-TWEETREADER it is possible to know faster when problem starts.

<table>
<thead>
<tr>
<th>Term</th>
<th>Before</th>
<th>After</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>gas</td>
<td>0.122</td>
<td>0.484</td>
<td>0.363</td>
</tr>
<tr>
<td>pedals</td>
<td>0.129</td>
<td>0.438</td>
<td>0.309</td>
</tr>
<tr>
<td>wonder</td>
<td>0.017</td>
<td>0.214</td>
<td>0.198</td>
</tr>
<tr>
<td>problem</td>
<td>0.163</td>
<td>0.357</td>
<td>0.194</td>
</tr>
<tr>
<td>good</td>
<td>0.016</td>
<td>0.205</td>
<td>0.190</td>
</tr>
<tr>
<td>recalling</td>
<td>0.012</td>
<td>0.106</td>
<td>0.095</td>
</tr>
<tr>
<td>gm</td>
<td>0.011</td>
<td>0.089</td>
<td>0.077</td>
</tr>
<tr>
<td>#heard_on_the_street</td>
<td>0.040</td>
<td>0.113</td>
<td>0.073</td>
</tr>
<tr>
<td>social</td>
<td>0.031</td>
<td>0.099</td>
<td>0.068</td>
</tr>
<tr>
<td>sticking</td>
<td>0.070</td>
<td>0.125</td>
<td>0.055</td>
</tr>
<tr>
<td>fix</td>
<td>0.026</td>
<td>0.076</td>
<td>0.050</td>
</tr>
<tr>
<td>popularity</td>
<td>0.016</td>
<td>0.037</td>
<td>0.021</td>
</tr>
<tr>
<td>love</td>
<td>0.017</td>
<td>0.024</td>
<td>0.008</td>
</tr>
</tbody>
</table>
Detecting Sentiment Change in Twitter Streaming Data

A tool like MOA-TWEEETREADER would have helped Toyota to understand the crisis sooner and to respond more appropriately.
Detecting Sentiment Change in Twitter Streaming Data

CONCLUSIONS. Our goal: how to do real time analysis of twitter data. Our proposal: MOA-TWEETREADER