On-line Relevant Anomaly Detection in the Twitter Stream: An Efficient Bursty Keyword Detection Model

PRISMA Research Group
Computer Science Department – University of Chile

Authors: Jheser Guzman
Bárbara Poblete

August 2013
¿What is the breaking news?
¿What is happening in the world right now?
Online Social Networks (OSN)
Microblogs

Motivation

too much information
online processing
Blogs, WebPages, Forums, etc.
Motivation

Microblogs
Main Problem: Detect emerging topics
Specific Problem: Efficient bursty keyword detection.

Challenges
- Large quantity of Keywords (Twitter stream).
- Memory limitation.
- Fast detection (reporting).
- Scalability.
Burst Detection Algorithm (BD)

**Tweet:** Short message up to 140 characters

Relevance: “The probability of occurrence of a term in a Window”.

Windows (20 min each)
Burst Detection Algorithm (BD)

KeyWord-1

KeyWord-2

…

KeyWord-N

Time

Ranked Keywords
Results

Matches Comparison


- BD/LDA: 91% Keywords matches
- TM/LDA: 3% Keywords matches
- EDCoW/LDA: 22%

Comparison with other algorithms

- Complexity: Our solution in worst case runs in time $O(n \log(n))$.
- Parameters: Simple to parameterize.
- Stopword Filter based on minimization of the noise and displaying results quickly (each 20 min).
Conclusions

• The Window-Slicing technique is **scalable** and **efficient** using less computational resources in the detection of Bursty Keywords in Stream of Messages.
• Scalable (Memory and Order of Growth).
• Map-reduce compatibility.

Future Work
Detect Bursty Keywords using the track of *Std.Deviation* and the *mean* of each Keyword in on-line processing.
Demo

Scientific Collaboration and Creativity Lab – University of Washington
http://depts.washington.edu/sccl
### Table: Examples of bursty keywords detected for the SuperBowl 2013 event, February 3rd and 4th, 2013

<table>
<thead>
<tr>
<th>Window Time</th>
<th>Top 10 Keywords</th>
<th>Tweet Examples</th>
</tr>
</thead>
</table>
| Feb 3rd 21:20 GMT| ravens, beyonce, 49ers, #superbowlsunday, @mrbbutterfield, forward, hours, #superbowlxlvii, xlvii, halftime | • @ArianaGrande: @glitteryariana: @ArianaGrande 49ers or ravens? Beyonce lol made my day.  
• RT @NiallOfficial: Who do I follow in superbowl tonight american fans? 49ers or ravens? I don’t know much about american football  
• 49ers: 27 Ravens: 24 Goes into quadruple OT. Monday declared national holiday. |
| Feb 3rd 23:40 GMT| ravens, touchdown, commercial, #thekiss, kaepemick, commercials, boldin, godaddy, tackles, @godaddy | • RT @lifestyle: #TheKiss? The Worst! @GoDaddy #NotBuyingIt. Wish they’d spend that money on customer service instead of wasting it on a  
• RT @PiperJones: That GoDaddy commercial was just…ew. |
| Feb 4th 01:40 GMT| lights, power, outage, superdome, 49ers, ravens, turned off, stadium, luz, #lightsout | • RT @YourAnonNews: NFL can control everything except power in Superdome. What an embarrassment. 7 mins and counting.  
• RT @ShamelessProbs: The lights went out because we don’t need football after Beyonce.  
• RT @billmarchil3: @TheShieldWWE turned off the lights and are making their way to take out #Flacco! @WWE #Superbowl |
On-line Relevant Anomaly Detection in the Twitter Stream: An Efficient Bursty Keyword Detection Model

PRISMA Research Group
Computer Science Department – University of Chile

Authors: Jheser Guzman
Bárbara Poblete

Thank you