TopicSketch: Real-time Bursty Topic Detection from Twitter

Wei Xie, Feida Zhu, Jing Jiang, Ee-Peng Lim and Ke Wang*
Living Analytics Research Centre
Singapore Management University

* Ke Wang is from Simon Fraser University, and this work was done when the author was visiting Living Analytics Research Centre in Singapore Management University.
Twitter as News Media

• Twitter works as a huge news media.

• For some topics, especially bursty topics, news first appears in Twitter, rather than traditional news media.

• It is interesting and also useful to detect bursty topics from Twitter.
Handling Tweet Stream is Challenging

- **Large Volume**
  Number of tweets per day : 340 million

- **Large Velocity**
  Number of tweets per second : 9,000 (average) / 143,000 (peak)

- **Large Variety**
  All kinds of activities and topics appear in Twitter
Motivation

Related Work

Proposed Method

- Intuition
- Indicator of burst
- Assumptions
- Solution
- Framework
- Dimension reduction

Experiment

Conclusion
Related Work
Related Work

• **Topic Modelling**
  — Liangjie Hong, et al. A time-dependent topic model for multiple text streams. KDD 2011
Related Work

• **Topic Modelling**
  —Liangjie Hong, et al. A time-dependent topic model for multiple text streams. KDD 2011

• **Topic Detection & Tacking**
  —Sasa Petrovic, et al. Streaming First Story Detection with application to Twitter. HLT-NAACL 2010
  —Chenliang Li, et al. Twevent: segment-based event detection from tweets. CIKM 2012
Related Work

• **Topic Modelling**
  — Liangjie Hong, et al. A time-dependent topic model for multiple text streams. KDD 2011

• **Topic Detection & Tacking**
  — Sasa Petrovic, et al. Streaming First Story Detection with application to Twitter. HLT-NAACL 2010
  — Chenliang Li, et al. Twevent: segment-based event detection from tweets. CIKM 2012

Both of them face difficulty to handle large tweet stream, as they need to process very huge historical data.
Intuition
• Rather than keep the big historical data, maybe we can take a snapshot of the current data stream.
• Rather than keep the big historical data, maybe we can take a snapshot of the current data stream.

• At least, it takes much smaller space and hopefully we can efficiently infer topics from it.
• Rather than keep the big historical data, maybe we can take a snapshot of the current data stream.

• At least, it takes much smaller space and hopefully we can efficiently infer topics from it.

But How?
Acceleration as an Indicator

Adopt the concepts in physics:
Acceleration as an Indicator

Adopt the concepts in physics:

**Velocity**
Acceleration as an Indicator

Adopt the concepts in physics:

**Velocity** the rate of change of the volume of tweet stream
Acceleration as an Indicator

Adopt the concepts in physics:

**Velocity**  
the rate of change of the volume of tweet stream

\[ v = \frac{\Delta x}{\Delta t} \]
Acceleration as an Indicator

Adopt the concepts in physics:

**Velocity**  the rate of change of the volume of tweet stream

$$v = \frac{\Delta x}{\Delta t}$$

**Acceleration**
Acceleration as an Indicator

Adopt the concepts in physics:

**Velocity**  
the rate of change of the volume of tweet stream

\[ v = \frac{\Delta x}{\Delta t} \]

**Acceleration**  
the rate of change of the velocity of tweet stream
Acceleration as an Indicator

Adopt the concepts in physics:

**Velocity**  
the rate of change of the volume of tweet stream

\[ v = \frac{\Delta x}{\Delta t} \]

**Acceleration**  
the rate of change of the velocity of tweet stream

\[ a = \frac{\Delta v}{\Delta t} \]
Acceleration as an Indicator

Acceleration: a very good early indicator of burst.
Acceleration as an Indicator

Usually we can observe the peak of acceleration earlier than the peak of velocity.
Usually we can observe the peak of acceleration earlier than the peak of velocity.
Acceleration as an Indicator

Usually we can observe the peak of acceleration earlier than the peak of velocity.
Acceleration as an Indicator

Usually we can observe the peak of acceleration earlier than the peak of velocity.
Acceleration as an Indicator

Usually we can observe the peak of acceleration earlier than the peak of velocity.
Acceleration as an Indicator

Usually we can observe the peak of acceleration earlier than the peak of velocity.
Acceleration as an Indicator

Usually we can observe the peak of acceleration earlier than the peak of velocity.
Acceleration as an Indicator

Acceleration: a very good early indicator of burst.
Acceleration: a very good early indicator of burst.

1. Is there any burst at all?
Acceleration as an Indicator

Acceleration: a very good early indicator of burst.

1. Is there any burst at all?

The acceleration of the whole tweet stream.
Acceleration as an Indicator

Acceleration: a very good early indicator of burst.

1. Is there any burst at all?

   The acceleration of the whole tweet stream.

2. Is there any word bursting?
Acceleration as an Indicator

Acceleration: a very good early indicator of burst.

1. Is there any burst at all?
   The acceleration of the whole tweet stream.

2. Is there any word bursting?
   The acceleration of each word in the tweet stream.
Acceleration as an Indicator

Acceleration: a very good early indicator of burst.

1. Is there any burst at all?

   The acceleration of the whole tweet stream.

2. Is there any word bursting?

   The acceleration of each word in the tweet stream.

3. Is there any topic bursting?
Acceleration as an Indicator

Acceleration: a very good early indicator of burst.

1. Is there any burst at all?
   
   The acceleration of the whole tweet stream.

2. Is there any word bursting?
   
   The acceleration of each word in the tweet stream.

3. Is there any topic bursting?
   
   The acceleration of each pair of words in the tweet stream.
Assumptions

• Each topic is represented as a distribution over words $p_k$. 
Assumptions

- Each topic is represented as a distribution over words $p_k$.
- Tweet stream is modelled as a mixture of multiple latent topic streams. The stream of topic $k$ has velocity $v_k(t)$ and acceleration $a_k(t)$.
Assumptions

• Each topic is represented as a distribution over words $p_k$.

• Tweet stream is modelled as a mixture of multiple latent topic streams. The stream of topic $k$ has velocity $v_k(t)$ and acceleration $a_k(t)$.

• Each tweet is related to only one topic.
Assumptions

- Each topic is represented as a distribution over words $p_k$.

- Tweet stream is modelled as a mixture of multiple latent topic streams. The stream of topic $k$ has velocity $v_k(t)$ and acceleration $a_k(t)$.

- Each tweet is related to only one topic.

The final goal is to discover these unknown $p_k$ and $a_k(t)$ from a snapshot of the tweet stream.
(1). $\hat{S}''(t)$: The acceleration of the total number of tweets in $D(t)$, $\hat{S}(t) = |D(t)|$.

(2). $\hat{X}''(t)$: The acceleration of each word, $\hat{X}(t)$ is a $N$-dimension vector such that $\hat{X}_i(t) = \sum_{d \in D(t)} \frac{d(i)}{|d|}$, $(1 \leq i \leq N)$.

(3). $\hat{Y}''(t)$: The acceleration of each pair of words, $\hat{Y}(t)$ is a $N \times N$ matrix such that

$$\hat{Y}_{i,j}(t) = \begin{cases} 
\sum_{d \in D(t)} \frac{d(i)^2-d(i)}{|d|(|d|-1)} & , \quad i = j \\
\sum_{d \in D(t)} \frac{d(i)d(j)}{|d|(|d|-1)} & , \quad i \neq j
\end{cases}$$

$(1 \leq i \leq N, 1 \leq j \leq N)$. 
Properties

\[ S''(t) = \sum_{k=1}^{K} a_k(t) \]  

(1)

\[ E[X''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \]  

(2)

\[ E[Y''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \cdot p_k^T \]  

(3)
Properties

\[ S''(t) = \sum_{k=1}^{K} a_k(t) \]  \hspace{1cm} (1)

\[ E[\bar{X}''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \]  \hspace{1cm} (2)

\[ E[\bar{Y}''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \cdot p_k^T \]  \hspace{1cm} (3)
The topics with small accelerations will be filtered out.
Properties

\[ S''(t) = \sum_{k=1}^{K} a_k(t) \] (1)

\[ E[X''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \] (2)

\[ E[Y''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \cdot p_k^T \] (3)

The topics with small accelerations will be filtered out.

Minimise the difference between observation and expectation.
Real-time Framework

word vector

current tweet $d$

tweet stream $D(t)$

monitor

estimator

reporter

sketch $X''(t)$ $Y''(t)$

$S''(t)$ $N$

$LARC$
Real-time Framework

word vector

N

current tweet

$\text{d}$

tweet stream

$D(t)$

N is very large

sketch

$X''(t)$

$Y''(t)$

$N$

$S''(t)$

$N$

monitor

$S''(t)$

estimator

reporter

14
Dimension Reduction

From $O(N^2)$ to $O(HB^2)$, $B<<N$, $H<<N$

Efficiency Evaluation

**Dataset** : Singapore based Twitter data, which contains over 30 millions tweets. We use these tweets to simulate a live tweet stream.
Dataset: Singapore based Twitter data, which contains over 30 millions tweets. We use these tweets to simulate a live tweet stream.

Throughput
**Dataset**: Singapore based Twitter data, which contains over 30 millions tweets. We use these tweets to simulate a live tweet stream.
Effectiveness Evaluation

• Compare with Twevent

• Use the same dataset which contain over 4 million tweets

• List all the events detected by both algorithms between June 7, 2010 to June 12, 2010, in which period several big events happened.
<table>
<thead>
<tr>
<th>Event</th>
<th>Sub-event</th>
<th>TopicSketch</th>
<th>Twevent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steve Jobs released iPhone 4 during WWDC2010</td>
<td>Farmville client for iPhone 4 was demonstrated.</td>
<td>#wwdc, iphone, farmville</td>
<td>steve jobs, iMovie, wwdc, iphone, wifi</td>
</tr>
<tr>
<td></td>
<td>Retina display of iPhone 4 was introduced.</td>
<td>iphone, 4, #wwdc, display, retina</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iMovie for iPhone 4 was demonstrated.</td>
<td>iphone, 4, imovie, #wwdc</td>
<td></td>
</tr>
<tr>
<td></td>
<td>New iPhone 4 was available in Singapore in July.</td>
<td>iphone, 4, singapore, july</td>
<td></td>
</tr>
<tr>
<td>Event</td>
<td>Sub-event</td>
<td>TopicSketch</td>
<td>Twevent</td>
</tr>
<tr>
<td>-------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Steve Jobs released iPhone 4 during WWDC2010</td>
<td>Farmville client for iPhone 4 was demonstrated.</td>
<td>#wwdc, iphone, farmville</td>
<td>steve jobs, iMovie, wwdc, iphone, wifi</td>
</tr>
<tr>
<td></td>
<td>Retina display of iPhone 4 was introduced.</td>
<td>iphone, 4, #wwdc display, retina</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iMovie for iPhone 4 was demonstrated.</td>
<td>iphone, 4, imovie, #wwdc</td>
<td></td>
</tr>
<tr>
<td></td>
<td>New iPhone 4 was available in Singapore in July.</td>
<td>iphone, 4, singapore, july</td>
<td></td>
</tr>
</tbody>
</table>
Effectiveness Evaluation
Event
Effectiveness Evaluation

Event

Sub-event
Effectiveness Evaluation

Event Detection

Sub-event
• We proposed **TopicSketch** a framework for real-time detection of bursty topics from Twitter.

• We developed a concept of “sketch” which provides a “snapshot” of the current tweet stream. It can be updated efficiently. And we can find bursty topics from it efficiently.

• TopicSketch provides a temporally-ordered sub-events to describe the event, which is more informative than the traditional methods.
Thanks