Who is Retweeting the Tweeters? Modeling Originating and Promoting Behaviors in Twitter Network

Aek Palakorn Achananuparp, Ee-Peng Lim, Jing Jiang, and Tuan-Anh Hoang

Living Analytics Research Centre, Singapore Management University
Introduction

• Multitudes of tweets are not informative (Naaman et al. 2010)
  • Twitter was initially designed as a personal “status updating” platform

Source: http://www.buzzfeed.com/mattbuchanan/the-first-30-tweets-ever
Introduction

• How do we find interesting Twitter users and contents?
  • Interesting users **propagate**/are the sources of interesting items

• How to model information sharing/propagating behavior of Twitter users
  • Retweeting (explicit propagation)
  • **Originating:** Writing tweets interesting to others
  • **Promoting:** Passing on other users’ interesting tweets

• How do we apply the propagation behaviors to other application context?
Twitter Data → Key-Phrase Extraction → URL Extraction → Retweet Linkage → Behavior Modeling → Burst Detection → Events

Key-Phrase Extraction
URL Extraction
Hashtag Extraction

Behavior Modeling Framework
Behavior-Based Event Detection

Framework
Roadmap

Twitter Data → Key-Phrase Extraction

URL Extraction → Retweet Linkage

Hashtag Extraction → Behavior Modeling

Burst Detection → Events

3 min → 6 min → 8 min → 7 min → 24 min
Twitter Data and Key-Phrase Extraction

Twitter Data

Key-Phrase Extraction

URL Extraction

Hashtag Extraction

Retweet Linkage

Behavior Modeling

Burst Detection

Events
Dataset: Singapore Twitter Users

- Snowball sampling
- Twitter REST API
- Two-hop neighborhoods of the top 1000 Singapore-based Twitter users from Twitaholic.com
- Why Singapore?
- 18K unique users
- 2.6 million tweets published from Dec 2009 – Apr 2010
53.8% of users have more followees than followers

Low correlation between the number of tweets published and the numbers of followees/followers

From Dec 2009 – Apr 2010:
- The most active user published 3K tweets
- Popular users (>1K followers) published [10,1000] tweets
Keyphrase Extraction

- **TextRank** algorithm (Mihalcea and Tarau, 2004)
- Preprocessing
  - Remove common stop words
  - Remove non-content bearing words, e.g., @, RT, via, etc.
  - Remove internet slangs, e.g. lol, omg, rofl, etc.
- Represent the tweet data as an **undirected lexicon graph**
- **6.8k** candidate phrases (unigrams & bigrams) are extracted.
Examples of Keyphrases Extracted

- Mostly global and local named entities
- Justin Bieber
- Jack Neo
- Tiger Woods
- Apple iPhone
- Google China
- Singtel
- StarHub
- CNY
- Glee
- Youth Olympic Games
- Universal Studio
Retweet Linkage

Twitter Data → Key-Phrase Extraction → Retweet Linkage → Behavior Modeling → Burst Detection → Events

Key-Phrase Extraction
URL Extraction
Hashtag Extraction

User Tweet Item

… … … …
(Strict) Retweet Linkages

- Explicit propagation
- **Patterns:**
  “RT @[username]” or “via @[username]”
- **7,099** strict retweet relationships (<1% of the tweet data)
- Perhaps we may consider implicit propagation…
Weak Retweet Linkages

...Justin Bieber... ...Justin Bieber...

@Foo \rightarrow @Bar

- Implicit propagation (co-mentioning of interesting item)
- **Rules:**
  - $T_{Wi}$ and $T_{Wj}$ were published by $u_i$ and $u_j$, respectively
  - $T_{Wi}$ and $T_{Wj}$ **co-mention the same item** $d$
  - $T_{Wi}$ was published **within 24 hours** (empirically set) of $T_{Wj}$
  - No other tweets from the followees of $u_j$ containing $d$ and having published timestamp closer to $T_{Wi}$'s
Weak Retweet Linkages

• Keyphrases as interesting items

<table>
<thead>
<tr>
<th>Multi-Word</th>
<th># weak retweets</th>
<th>Single-Word</th>
<th># weak retweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Justin Bieber</td>
<td>194 (16)</td>
<td>CNY</td>
<td>635 (10)</td>
</tr>
<tr>
<td>Jack Neo</td>
<td>189 (19)</td>
<td>Jonghyun</td>
<td>487 (6)</td>
</tr>
<tr>
<td>Apple iPhone</td>
<td>114 (4)</td>
<td>David</td>
<td>239 (26)</td>
</tr>
<tr>
<td>Universal Studio</td>
<td>82 (13)</td>
<td>Glee</td>
<td>180 (11)</td>
</tr>
<tr>
<td>Star Awards</td>
<td>62 (7)</td>
<td>Idol</td>
<td>178 (14)</td>
</tr>
</tbody>
</table>

Parentheses indicate the numbers of strict retweets.

• Caveat: Homophily or globally popular phrases may bias linkage formation.
Comparison: Depth of Retweets
Comparison: Retweet Interaction Networks

Strict Retweet Linkages

Weak Retweet Linkages
Behavior Modeling

User  Tweet  Item
...  ...  ...

Twitter Data → Key-Phrase Extraction → URL Extraction → Hashtag Extraction → Retweet Linkage → Behavior Modeling → Burst Detection → Events
Behavior Models

• Basic model
  • Fractions of tweets/retweets
  • Fractions of followers/followeees interacting with the users

• Mutual dependency model
  • Fractions of tweets/retweets
  • Fractions of followers/followeees interacting with the users
  • **Behavior strength of the interacting users**.
  • HITS-like algorithm (Kleinberg 1998)

• Range-based model
  • Fractions of tweets/retweets
  • Fractions of followers/followeees interacting with the users
  • **Depths and reaches (numbers of users in the diffusion chains)**
Basic Model

A user exhibits a strong originating behavior if a large fraction of her tweets have been retweeted by a large fraction of her followers.

Originating (O)

User

Followers

Promoting (P)

Followees

User

A user exhibits a strong promoting behavior if she retweets a large fraction of interesting tweets, i.e., those which have been retweeted by others, she has seen from a large fraction of her followees.
Mutual Dependency Model

**Originating (O\text{m})**

A user is a strong originator when most tweets by her are retweeted by many strong promoters.

**Promoting (P\text{m})**

A user is a strong promoter when most retweets she published are based on tweets published by strong originators.
A user is a strong originator if a large fraction of her tweets have been retweeted by a large fraction of her followers and her original tweets tend to be propagated in great depth and reach by her extended followers.

A user is a strong promoter if she retweets a large fraction of interesting tweets she has seen from a large fraction of her followees and those interesting tweets tend to be propagated in greater depth and reach by the extended followers of her followees.
Formulas...

### Basic

Originating

\[ O_i = \sum_{j \in P_i} \frac{t_{ij}}{t_i} \times \frac{u_{i}^{in}}{|U_k F_{ik}^{in}|} \]

Promoting

\[ P_i = \sum_{j \in P_i} \frac{t_{ij}}{\sum_{k \in P_i} \sum_i t_{kl}} \times \frac{u_{o}^{out}}{|U_k F_{ik}^{out}|} \]

### Mutual Dependency

Originating

\[ O_i^m = \sum_{j \in P_i} \frac{t_{ij}}{t_i} \times \frac{u_{i}^{in}}{|U_k F_{ik}^{in}|} \]

Promoting

\[ P_i^m = \sum_{j \in P_i} \frac{t_{ij}}{\sum_{k \in P_i} \sum_i t_{kl}} \times \frac{u_{o}^{out}}{|U_k F_{ik}^{out}|} \]

### Range-Based

Originating

\[ O_i^r = \left( \sum_{j \in T_i} d_j \times \log \left| \bigcup_{k=1}^K U_{i k}^{in} \right| \right) \times O_i \]

Promoting

\[ P_i^r = \left( \sum_{j \in T_i} d_j \times \log \left| \bigcup_{j=1}^K \bigcup_{k=1}^K U_{i k}^{in} \right| \right) \times P_i \]
Kendall’s Tau Correlations of Different User Rankings

<table>
<thead>
<tr>
<th></th>
<th>O</th>
<th>P</th>
<th>O&lt;sup&gt;m&lt;/sup&gt;</th>
<th>P&lt;sup&gt;m&lt;/sup&gt;</th>
<th>O&lt;sup&gt;R&lt;/sup&gt;</th>
<th>P&lt;sup&gt;R&lt;/sup&gt;</th>
<th>TCount</th>
<th>RTCount</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>1</td>
<td>0.21</td>
<td>0.56</td>
<td>0.16</td>
<td>0.65</td>
<td>0.25</td>
<td>0.17</td>
<td>-0.11</td>
</tr>
<tr>
<td>P</td>
<td>1</td>
<td>0.21</td>
<td>0.54</td>
<td>0.16</td>
<td>0.68</td>
<td>0.11</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>O&lt;sup&gt;m&lt;/sup&gt;</td>
<td>1</td>
<td>0.17</td>
<td>0.46</td>
<td>0.24</td>
<td>0.15</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P&lt;sup&gt;m&lt;/sup&gt;</td>
<td>1</td>
<td>0.16</td>
<td>0.36</td>
<td>0.14</td>
<td>0.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O&lt;sup&gt;R&lt;/sup&gt;</td>
<td>1</td>
<td>0.10</td>
<td>0.01</td>
<td>-0.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P&lt;sup&gt;R&lt;/sup&gt;</td>
<td>1</td>
<td></td>
<td></td>
<td>0.24</td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCount</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.12</td>
</tr>
<tr>
<td>RTCount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

TCount = Total tweet count  
RTCount = Total retweet count
Comparison of the Top-10 Users

- **Power users** = those whose number of followers is at least two orders of magnitude greater than the number of followees

**News Media**

- The Straits Times
  - @STcom
  - Breaking news from the online site of Singapore’s most widely-read newspaper
  - Singapore: [http://www.straitstimes.com](http://www.straitstimes.com)

- Channel NewsAsia
  - @ChannelNewsAsia
  - The official space of Channel NewsAsia - a premier source of real time news and videos from Asia and the world.
  - Singapore: [http://www.channelnewsasia.com](http://www.channelnewsasia.com)

**Religious Organization**

- City Harvest Church
  - @chcs
  - The Official City Harvest Church Twitter. Loving God Loving People

**Popular Online Forum**

- sglatestnews
  - @sglatestnews
  - Singapore’s breaking news from newspaper in Singapore.
  - Singapore: [http://sglatestnews.com](http://sglatestnews.com)

- Singapore News
  - @SGnews
  - Singapore news consolidated and tweeted. Managed by @benfoc.
  - Singapore: [http://sgnews.com](http://sgnews.com)
## Comparison of the Top-10 Users

### Basic

<table>
<thead>
<tr>
<th>Rank</th>
<th>Originator</th>
<th>Promoter</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>stcom*</td>
<td>rochorbeancurd</td>
<td>Entrepreneur</td>
</tr>
<tr>
<td>2</td>
<td>ChannelNewsAsia*</td>
<td>alkanphel</td>
<td>Blogger</td>
</tr>
<tr>
<td>3</td>
<td>MediaAsia</td>
<td>charlesyeo</td>
<td>Entrepreneur</td>
</tr>
<tr>
<td>4</td>
<td>chess*</td>
<td>MediaAsia</td>
<td>Media</td>
</tr>
<tr>
<td>5</td>
<td>arieszulkarnain</td>
<td>CalvinTimo</td>
<td>Writer</td>
</tr>
<tr>
<td>6</td>
<td>mbrown</td>
<td>minzmint</td>
<td>Student</td>
</tr>
<tr>
<td>7</td>
<td>hardwarezone*</td>
<td>callsg</td>
<td>N/A</td>
</tr>
<tr>
<td>8</td>
<td>TaufikBatisah</td>
<td>LisaHK</td>
<td>Student</td>
</tr>
<tr>
<td>9</td>
<td>Seowhow</td>
<td>conguyuan</td>
<td>Student</td>
</tr>
<tr>
<td>10</td>
<td>lynettegoh</td>
<td>trevidzi</td>
<td>Student</td>
</tr>
</tbody>
</table>

### Mutual Dependency

<table>
<thead>
<tr>
<th>Rank</th>
<th>Mutual Originator</th>
<th>Mutual Promoter</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>stcom*</td>
<td>CalvinTimo</td>
<td>Writer</td>
</tr>
<tr>
<td>2</td>
<td>SingaporeClub</td>
<td>sgbreakingnews</td>
<td>Aggregator</td>
</tr>
<tr>
<td>3</td>
<td>mbrown</td>
<td>LisaHK</td>
<td>Student</td>
</tr>
<tr>
<td>4</td>
<td>lynettegoh</td>
<td>ZairoOng</td>
<td>Blogger</td>
</tr>
<tr>
<td>5</td>
<td>chess*</td>
<td>SingaporeClub</td>
<td>Aggregator</td>
</tr>
<tr>
<td>6</td>
<td>angjiehui</td>
<td>rochorbeancurd</td>
<td>Entrepreneur</td>
</tr>
<tr>
<td>7</td>
<td>sgfinomap</td>
<td>alkanphel</td>
<td>Blogger</td>
</tr>
<tr>
<td>8</td>
<td>eglatestnews*</td>
<td>synthiaschul</td>
<td>Student</td>
</tr>
<tr>
<td>9</td>
<td>patlaw</td>
<td>Seowhow</td>
<td>Pastor</td>
</tr>
<tr>
<td>10</td>
<td>enalatest*</td>
<td>prelius</td>
<td>Student</td>
</tr>
</tbody>
</table>

### Range-Based

<table>
<thead>
<tr>
<th>Rank</th>
<th>Range-Based Originator</th>
<th>Range-Based Promoter</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>stcom*</td>
<td>alkanphel</td>
<td>Blogger</td>
</tr>
<tr>
<td>2</td>
<td>MediaAsia</td>
<td>charlesyeo</td>
<td>Entrepreneur</td>
</tr>
<tr>
<td>3</td>
<td>chess*</td>
<td>LisaHK</td>
<td>Student</td>
</tr>
<tr>
<td>4</td>
<td>ChannelNewsAsia*</td>
<td>rochorbeancurd</td>
<td>Entrepreneur</td>
</tr>
<tr>
<td>5</td>
<td>hardwarezone*</td>
<td>AngMoGirl</td>
<td>Designer</td>
</tr>
<tr>
<td>6</td>
<td>angjiehui</td>
<td>CalvinTimo</td>
<td>Writer</td>
</tr>
<tr>
<td>7</td>
<td>enalatest*</td>
<td>moemasa</td>
<td>Singer</td>
</tr>
<tr>
<td>8</td>
<td>TaufikBatisah</td>
<td>deckstor</td>
<td>Student</td>
</tr>
<tr>
<td>9</td>
<td>SGnews*</td>
<td>JadeChenSS5144</td>
<td>Student</td>
</tr>
<tr>
<td>10</td>
<td>patlaw</td>
<td>trevidzi</td>
<td>Student</td>
</tr>
</tbody>
</table>
Comparison of the Top-10 Users

• **40%-60% of the top originators** are power users while there are **no power users** among the top promoters.

• The top originators generally consist of **news media, news aggregators, and local celebrities.**

• The top range-based originators contain **slightly more news media users** than the basic and mutual dependency originators.

• The top promoters generally consist of **students, bloggers, and entrepreneurs.**
Basic vs. Mutual Dependency Models

One-hop neighborhood (retweets)

Top-3 Basic Promoters

Top-3 Mutual Dependency Promoters
Strong Mutual Dependency among Religious Users

Churches and pastors

City Harvest Church
@chcsg
The Official City Harvest Church Twitter. Loving God Loving People
Singapore  http://www.chc.org.sg

Tan Seow How
@Seowhow
liveGo

Aries Zulkarnain
@arieszulkarnain
Work hard, play hard, be modest, do your job to the utmost & laugh at yourself
Singapore  http://www.chc.org.sg

Followers

K. Lisah
@LisahK
History-maker, world-changer, gate-keeper

Jing Wen
@aikanphel
Shouting wide open since 2010
Singapore  http://thanjajingwen.com/

Zaira Ong
@ZairaOng
everything about Singapore...
Singapore

congyuan
@congyuan
doesn't blog.com
I hard I hard!

preius
@preius

SMU
Singapore Management University

mda
Media Development Authority
Singapore

Carnegie Mellon
Behavior-Based Event Detection

User | Tweet | Item
--- | --- | ---
… | … | …

Twitter Data ➔

Key-Phrase Extraction

URL Extraction

Hashtag Extraction

Tweet | Retweet | Item
--- | --- | ---
… | … | …

Retweet Linkage ➔

Behavior Modeling

Burst Detection ➔

Events

User | Score
--- | ---
… | …
Detecting Interesting Real-World Events

• **Questions:**
  - Does a sudden surge of interests among Twitter users, measured in terms of their aggregated information propagation behaviors, possibly point to an interesting event?
  - How effective are the proposed behavior models perform in the event detection task, compared to the message-based approaches, such as surges of tweet and retweet volumes?

• **An event** is described as an occurrence \( e_i(s,t) \), specific to a **topic of interest** \( s \) and **time period** \( t \), that significantly differs from other occurrences \( e_j(s,t') \) determined by some baseline.

• Threshold-based approach
Burst Detection

\[ v(t) = \text{Sum of behavior scores of users who tweeted about } s \text{ during time period } t \]

Detect an event \( e(s,t) \) iff:
- \( v(t) - v(t-1) > 2 \times v(t-1) \)
- \( v(t) - v(t-1) > \text{threshold} \)
- \( \text{threshold} = \text{mean}(\delta(t')) + \text{SD}(\delta(t')) \)
Ground Truth Events

- 940 candidate events (47x20)
  - 47 frequent bigram phrases (mentioned in at least 10 retweets)
  - 20 weekly time periods (Dec 2009 – Apr 2010)
  - E.g. (Tiger Woods, March 14, 2010)

- 3 annotators (postdoctoral researchers)
  - Familiar with Twitter
  - Knowledgeable in global and local news
  - Each examines a set of tweets containing the keyphrase published within corresponding the week
  - Provides a binary judgment whether the tweets collectively express a common user interest or not

- Fleiss’ Kappa = 0.34 suggesting that event judgment is highly subjective
## Examples of Positive Ground Truth Events

<table>
<thead>
<tr>
<th>Keyphrase</th>
<th>Weekly Period</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Idol</td>
<td>Jan 10, 2010</td>
<td>The premiere of the 9th season of American Idol show</td>
</tr>
<tr>
<td>Google China</td>
<td>Jan 10, 2010</td>
<td>Announcement about Google’s operations in China</td>
</tr>
<tr>
<td>Apple iPad</td>
<td>Jan 24, 2010</td>
<td>Announcement of Apple iPad tablet device</td>
</tr>
<tr>
<td>Alexander McQueen</td>
<td>Feb 7, 2010</td>
<td>Breaking news about Alexander McQueen’s suicide</td>
</tr>
<tr>
<td>Jack Neo</td>
<td>Mar 7, 2010</td>
<td>Press conference by a local film director Jack Neo</td>
</tr>
<tr>
<td>Universal studio</td>
<td>Mar 14, 2010</td>
<td>Official opening of Universal Studio Singapore</td>
</tr>
</tbody>
</table>
Experiment Setup

• Event ranking (top-k) evaluation
• Two ground truth sets
  • G1: 21 positive (unanimous agreements) and 919 negative events
  • G2: 31 positive (agreements of at least two annotators) and 909 negative events
• 940 candidate events
• For each frequent keyphrase, construct:
  • Six time-series data; one for each behavior measure (O,P,Om,Pr,OR, and PR)
  • Two message-count baselines (TCount and RTCount)

• Combining multiple sets of events: Originator-promoter agreement:
  • O AND P – Intersection of the sets of events detected by O and P
  • O OR P – Union of the sets of events detected by O and P
  • Om AND Pm
  • Etc.
Evaluation Metrics

- Order predicted events by their relative importance ($v(t) - v(t-1)/\text{threshold}$)
- Average precision ($AP$)

$$AP = \frac{1}{|G|} \sum_{i=1}^{|G|} \text{Prec}(G_i)$$

- Precision at top-$k$ events ($\text{prec}@k$)
  - $\text{prec}@5$
  - $\text{prec}@10$

$$\text{Prec}@k = \frac{tp_k}{tp_k + fp_k}$$

- Both $AP$ and $\text{prec}@k$ outputs a value at $[0,1]$ range
## Results

**Performance on G1 ground truth set**

<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>Prec@5</th>
<th>Prec@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavior (P)</td>
<td>0.4608</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Behavior agreement (O OR P)</td>
<td><strong>0.6057</strong></td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>TCount</td>
<td>0.4672</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>RTCount</td>
<td>0.2792</td>
<td>0.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

**Performance on G2 ground truth set**

<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>Prec@5</th>
<th>Prec@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavior (P)</td>
<td>0.53</td>
<td>1</td>
<td>0.9</td>
</tr>
<tr>
<td>Behavior agreement (O OR P)</td>
<td><strong>0.6952</strong></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TCount</td>
<td>0.5418</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>RTCount</td>
<td>0.308</td>
<td>0.4</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Discussion

- Using signals from **both originating and promoting** behaviors is more helpful in detecting events.

- More elaborate models, i.e., the mutual dependency and range-based models, do not perform as well as the naive ones.

- In some cases, the behavior models perform better than the message-based models especially when the events did not generate enough surges in the volumes of tweets/retweets. These events include:
  - Youth Olympic Games 2010 in Singapore
  - The opening of Universal Studio Singapore
  - Star Awards 2010
Almost There…

Twitter Data

- Key-Phrase Extraction
- URL Extraction
- Hashtag Extraction

Tweet | Retweet | Item
--- | --- | ---
... | ... | ...

Retweet Linkage

Behavior Modeling

Burst Detection

Events

User | Score
--- | ---
... | ...

Event | Score
--- | ---
... | ...

User | Tweet | Item
--- | --- | ---
... | ... | ...

...
Conclusion

- Framework to model information propagation behaviors of Twitter users: **Originating** an **promoting** behaviors
- Notion of **weak retweet linkages** identified by co-mentioning of interesting items among tweets
- Comparisons of users ranked by different behavior models
- Application of the behavior models in detecting interesting real-world events from the Twitter streams
Aek Palakorn Achananuparp
Research Scientist, palakorna@smu.edu.sg