#FluxFlow: Visual Analysis of Anomalous

Jian Zhao, Nan Cao, Zhen Wen, Yale Song, Yu–Ru Lin, Christopher Collins.

Presenter: Keqian Li
What: SOCIAL MEDIA

SOCIAL MEDIA VISITOR GROWTH

REGISTERED USERS

BUSINESSES AND MARKETERS LOVE SOCIAL MEDIA!

0
93%
70%
70%

OF MARKETERS USE SOCIAL MEDIA FOR BUSINESS
OF BRANDS HAVE A PRESENCE ON GOOGLE+
4% GROWTH FROM Q4 2012
OF MARKETERS HAVE USED FACEBOOK TO SUCCESSFULLY GAIN NEW CUSTOMERS

0
34%

OF MARKETERS HAVE USED TWITTER TO SUCCESSFULLY GENERATE LEADS

FACEBOOK, TWITTER, AND GOOGLE ARE THE TOP 3 SOCIAL MEDIA SITES USED BY MARKETERS

FLICKR, TUMBLR, AND STUMBLEUPON ARE THE LEAST POPULAR SITES USED BY MARKETERS

GMAIL USERS
Why: Abnormal conversational threads

- Rioters attack London zoo and release animals
- Rioters cook their own food in McDonald's
- Police "beat a 16-year-old girl"
- London Eye set on fire
- Rioters attack a children's hospital in Birmingham
- Army deployed in Bank
- Miss Selfridge set on fire
How: FluxFlow

Preprocessing and Storage:
- Feature Extraction
- Threads Reconstruction
- Data Filtering

Analysis:
- Anomalous Threads Detection
- Multidimensional Scaling
- Hierarchical Content Clustering
- User Interaction Graph Extraction

Visualization

Data flows from Hadoop and Twitter into a Data Store, which then feeds into the Preprocessing and Storage phase. The output from this phase goes to the Analysis phase, which then outputs to the Visualization.
Abnormal Retweet Threads Detection: A Data mining approach

• One-Class Conditional Random Fields Model (OCCRF)
  – temporal dependency, due to mechanism in RT time series data
  – one-class nature. There is little to no example (or even a clear definition) of true anomalies
  – contains a set of hidden variables to capture the underlying sub-structure of the sequential data

• Extracted Feature for each single retweet
  – User profile features: counts of followers, friends, status
  – User network features: in-degree and out-degree
  – Temporal features: intervals between two adjacent tweets in the sequence
Data mining pipeline

RT threads

Feature vector time-series

Ranked RT threads with abnormal scores

Hidden states

s1 s2 ... s8
RT Thread Visualization: RT Thread Glyph

- Ending Time
- Anomaly Score
- Sentiment Score
- Starting Time
- User Volume

Sentiment Score: Negative  |  Positive
Anomaly Score: Low       |  High
RT Thread Visualization: RT Thread Timeline
System interface
Hierarchical cluster of RT threads by topics
MDS view of threads from high dimensional feature space
User social connections at the intra- or inter-thread level

User relationship graphs
Deep-Level Information for Input feature vectors, model hidden states, raw tweets
### Visualization techniques summary

<table>
<thead>
<tr>
<th>How: Encode</th>
<th>Glyph, Thread Timelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>How–Facet</td>
<td>Multiform, Overview/Detail, linked highlighting.</td>
</tr>
<tr>
<td>How: Reduce</td>
<td>Item filtering, Item aggregation, Attribute aggregation, Elide, Superimpose</td>
</tr>
<tr>
<td>How: Manipulate</td>
<td>Highlighting, Project, Zoom</td>
</tr>
</tbody>
</table>
Task Summary

• T1 Summarizing and aggregating important features of retweeting threads.
  – Glyph, Cluster View, MDS View
• T2 Indicating characteristics and connections of involving users.
  – User relationship graphs
• T3 Revealing temporal patterns of information spreading.
  – Thread Timeline
• T4 Facilitating visual data comparisons and correlations.
  – Cluster View, MDS View
• T5 Accessing deep-level information of the model and input.
  – Thread Timeline, Features View, Status View, Tweets View
Evaluation

• Datasets: two 10% Twitter feed datasets collected during two significant events:
  – 2012 Hurricane Sandy (52 million tweets)
  – 2013 Boston Marathon Bombing (242 million tweets)

• Baseline: One-Class SVM (OCSVM) [Scholkopf et al., 2001]

• Ground truth: manually labeled by three annotators based on reports after the events
Comparison Results

Accuracies of OCCRF and OCSVM in correctly detecting rumors in the top-K retweeting threads ranked by the models in datasets: a) Hurricane Sandy, and b) Boston Bombing.
Case Study of Hurricane Sandy
Critiques

• Data
  – Incorporate further content attribute (e.g., topics, tags, deeper semantic analysis)

• Data mining algorithm
  – Improve on algorithm scalability and response time
  – Decouple with specific models
  – More insights about the model beyond hidden states, e.g., interactions of model parameters

• Visualization
  – Timeline visualization need better reducing techniques to be scalable for real social network data
  – Better to show the “chain” of retweeting, and influence between users

• Evaluations
  – Stronger ground truth for quantitative evaluation
Thank you