Ushio: Analyzing News Media and Public Trends in Twitter

Fangzhou Yao, Kevin Chen-Chuan Chang and Roy H. Campbell

Introduction

• Social Network Services (SNS) are evolving:
  • There are more than 500 million tweets sent per day in Twitter nowadays, while the number was only 200 million back to 2011.
  • Hashtags and geographical information are embedded in tweets and Facebook statuses.
  • Lots of news agencies are broadcasting their news via Twitter, and people would like to participate in these discussions.
A Mashed Up But Curated World

- Mashable develops news stories according to trending topics from SNS.
- Flipboard generates pages based on SNS and news topics.
- Pulse aggregates news and tailors them based on user’s LinkedIn professional interests.
- Yahoo presented News Digest, which summarizes top news from different sources, like multiple news agencies, Twitter and Wikipedia.
- New York Time follows and presented NYT Now.
Trending Topics

- Trends provided by many services do not differentiate which topics are covered more by news media, and which are discussed more by the public.

- Are there differences between the events that news media are more willing to cover and stories that people are really interested in?

- Also, can we discover any related news stories along with the trending topics?
We Would Like to Build…

- Monitoring informative streams from both news media and the public, which is able to extract meaningful data and use it to analyze trends in real-time

- Discovering the correspondence between the focuses of news media and people, as well as the leading roles in assorted topics, would be beneficial to help media in building a more relevant news coverage
How Can We Collect Good Data?

• Aggregation and Statistics

• We use Term-Frequency, as we want to have a simple aggregation which could take least system resources.

• Meaningful and Fine-Grained Entities

• Named Entity Extraction is an approach to detect meaningful words and phrases quickly.

• Data Reliability
Design

Twitter

- Sample of the Public Timeline
- Account Following News Media

Ushio

Database
- Media
- People

Parser

Ruby Java Bridge

Twitter Connector

Query Handler

User

NER Framework
Design

• Data Collection
  • Twitter Streaming API
    • As Twitter does not have an API for full public timeline streams, and we did not assign any topic or keyword to the API, therefore it returned us a random sample of all public statuses.
    • We are able to obtain the complete tweets in real-time with Twitter’s user account streaming API.
  • Named Entity Recognition
    • We used Stanford NER framework, which achieves a good accuracy and speed in extracting named entities from texts.
    • This framework tags every word with its possible properties, such as PERSON, ORGANIZATION, LOCATION and MISC.
• Database Schema
  • Tables are named as Media and People, respectively.
  • The tuple has four columns: entity, type, time and tweet_id.
Relation Data Model

• Finding Trending Topics

```sql
SELECT entity, count(*) AS count
FROM social.People
WHERE time > $a AND time < $b AND type = $t
GROUP BY entity ORDER BY count DESC;
```

• Finding Related Topics

```sql
SELECT social.People.entity AS name_entity,
    count(*) AS count FROM social.People
WHERE tweet_id IN
    (SELECT social.People.tweet_id FROM social.People
    WHERE entity = $e AND time > $a AND time < $b)
GROUP BY entity ORDER BY count DESC;
```
What A Busy Week!

• Top 10 trending topics from both the news media and the public in the week starting from April 28th to May 4th, 2014.

• News covers both Ukraine crisis and Sterling’s NBA discrimination speech, but people who like to talk about the latter one more. **Modi:13(152) and India:14(150)** presented by media, but only **Modi:79(1932) and India:33(3516)** by the public.

• Sports and entrainment topics are favored by the public more, but maybe not tech news, due to **Google:36(3411) and Apple:39(3179)**.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Media</th>
<th>#</th>
<th>Public</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ukraine</td>
<td>462</td>
<td>Chelsea</td>
<td>11282</td>
</tr>
<tr>
<td>2</td>
<td>China</td>
<td>369</td>
<td>EU</td>
<td>10524</td>
</tr>
<tr>
<td>3</td>
<td>Donald Sterling</td>
<td>363</td>
<td>God</td>
<td>9913</td>
</tr>
<tr>
<td>4</td>
<td>Obama</td>
<td>287</td>
<td>Tribez</td>
<td>9625</td>
</tr>
<tr>
<td>5</td>
<td>NBA</td>
<td>282</td>
<td>Justin</td>
<td>9132</td>
</tr>
<tr>
<td>6</td>
<td>US</td>
<td>281</td>
<td>Argentina</td>
<td>8848</td>
</tr>
<tr>
<td>7</td>
<td>Russia</td>
<td>220</td>
<td>Donald Sterling</td>
<td>8788</td>
</tr>
<tr>
<td>8</td>
<td>Clippers</td>
<td>173</td>
<td>Best</td>
<td>8586</td>
</tr>
<tr>
<td>9</td>
<td>Apple</td>
<td>168</td>
<td>NBA</td>
<td>6790</td>
</tr>
<tr>
<td>10</td>
<td>Oklahoma</td>
<td>153</td>
<td>London</td>
<td>6293</td>
</tr>
</tbody>
</table>
• On April 29th 2014, **Microsoft** was mentioned more than usual, and we discovered the related topics, which indicated they were about to start selling **Xbox One** in **China**.
Why Do They Care About It So Much?

- The **NBA** racism talk scandal by **Donald Sterling** and **Clippers** was the trend around that day, and hence all counts for these entities overwhelmed the ones about Microsoft.

![Bar charts showing media and public interest in NBA, Donald Sterling, Clippers, and Adam Silver](chart.png)
Who is the Winner?

• The figure shows correlation between media and the public by showing the PERSON type entities ranking of Donald Sterling from April 26th until May 7th, 2014.

• Are people leading the board?

• What about political news?
Future Work

• Conducting more experiments on assorted topics and gathering more data

• Deploying this system with a visualized interface for public accesses

• Adding segregation based on geographical information in tweets

• Using Map-Reduce and / or NoSQL database for data entities’ aggregation
Questions?

Thanks!