Data Science Summit 2017

www.tue.nl/dsce/summit
Rakesh Agrawal
President, Data Insights Laboratories

Keynote - Big Data or Big Garbage: A Tale of a Quest for Business Intelligence from Social Data
Big Data or Big Garbage: A Tale of a Quest for Business Intelligence from Social Data

Rakesh Agrawal
President, Data Insights Laboratories, San Jose, CA, USA
Visiting Professor, DIAS Lab, EPFL, Switzerland
Fellow Travellers*

Abhimanyu Das
Stelios Paparizos
John Shafer

* Colleagues at the Microsoft Search Labs
Terra Incognita
Perception of Social Data

Noisy!

Worthless!

Garbage!!!

More than 60% of the Twitter sample stream is useless garbage...

Further 20% are trolls...

Pavlo Baron, QCON Int’l Software Development Conf., Nov 2012
Research Expedition

Can enterprises mine useful real-time business intelligence from social data, and in particular from Twitter stream?
Testing Waters
TwitterSphere was reacting to the product related events!
More sustained buzz around WP8!
(Coefficient of variation in tweet count: 0.52 for WP8, 0.81 for win8 and 1.35 for surface)
Weekly Tweet Count: Surface vs. iPad Mini (July 2012 – Jan 2013)

More buzz around iPad Mini than Surface!
Tweet Counts Across the World: Surface, WP8, & Win8

More buzz around WP8 than Surface in ROW!
Apps:
-Determined to like Microsoft Surface. But WinRT is feeling like beta. Don't understand why some apps couldn't be ported from WindowPhone.
-@vbandi Keyboard. Productivity with Office-type apps like Google Docs. Some people like that. Surface RT has it too w/ Office. iPad lacking.
-The native apps on the new Windows 8 surface tablet feel pretty good. News and travel are especially nicely done.

Price:
-Microsoft disappoints with Surface price and misses bigger opportunity http://t.co/9CVaWKvq via @siliconbeat
-@vegaobscura So an upper-tier product like the Surface Pro will definitely fall into the price category of a 4G LTE 64GB Retina iPad.
-Condition of Surface = Windows Phone.. great design ..good price but fail apps! Why the hell we should appreciate?

TouchScreen:
-Why do I love Surface RT (and hopefully Pro) so much? It's simple: keyboard + touchscreen = crazy delicious.
-I'm not impressed w/ the Surface tablet, but i am impressed with Win8 on the new touchscreen laptops. The Acer s7 is awesome.
-OK, I’m really impressed with NBC’s touchscreen system. Giant Microsoft Surface turned on its side?

Keyboard:
-RT @notch: Got to play with the Surface, and I quite like it so far. The keyboard cover thing is clever, and works great.
-Microsoft Surface users complain about Wi-Fi problems http://t.co/DpLZDjk5 - and remember the touch keyboard problem, not good
-#buildwin Jordan Rudess is killing it on stage on the keyboard and earlier on a freaking surface tablet. Digital Theremin woo!
Diverse Tweets for Surface

Display:
-Sure. Worse is better. Sums up M$. Microsoft: Surface screen better than iPad Retina display http://t.co/ACGuXtbF

-Finally got my hands on a Surface today at China Open Days. Now I really want one. Display looks great and I like the keyboard.

-In a dark room w/ a movie? iPad prob wins. But I'd like a display that works well in many conditions. Hopefully Surface has that.

Battery:
-What I like about #XPS10 over Surface, a true keyboard dock and additional battery in dock for full day productivity

-the battery is like fine on ipad the surface is actually 40 minutes less than ipad supposedly..

-It has great thoughtful design, useful keyboard dock, great performance and battery life. The BEST tab along with Microsoft SURFACE

Camera:
-Playing with my new Surface tablet. Verdict: Good, needs some refinement. Definitely not an iPad. 1MP camera blows.

-HOLY SHAKEY CAMERA MICROSOFT. Geez plus its like watching a copycat Apple launch. #getyourownstyle Surface product looks great

Commercial:
- I'm not gonna lie that Microsoft surface commercial is pretty cool

- What an iPad and Microsoft Surface Parody Commercial Looks Like: GIZMODO Cheery piano musical... http://t.co/yiYhrmcA
The Ship: WaveFour
Sample Analytics:
“Which countries are talking the most about Nokia Android”
“How has sentiment changed amongst users about XBox over the last few days”
“Most talked-about Superbowl ad URLs amongst English vs. Spanish speakers”

Alerts based on deviation detection
Early Sightings
Tami Reller’s interview, taken to imply the imminent iPad office release.
Most Talked About Super Bowl Ads (New Yorker vs. Social)

- Budweiser: "puppy love"
- Coca-cola America: the beautiful
- Honda: hugs with Bruce Willis
- Turbotax: big game
- Hyundai: dad’s sixth sense
- Doritos super bowl contest (not aired)
- Morpheus for Kia’s car
- Blitz Energy Drink: (not aired)
- Microsoft: empowering users
- Ellen’s dance for Beats headphones

---

**SQL Query**

```sql
SELECT url.value as url, count(DISTINCT posting.tweetid_userid) as volume
FROM
    [wavefrontweets].[dbo].[Posting]
JOIN [wavefrontweets].[dbo].[Language] ON language.tweetid = Posting.tweetid AND Language.userid = Posting.userid
JOIN [wavefrontweets].[dbo].[Url] ON url.tweetid = Posting.tweetid AND url.userid = Posting.userid
WHERE contains(tweettext, 'superbowl OR "super bowl"') AND (commercial OR ad OR ads OR advertisement)
AND Language.value = 'en'
GROUP BY url.value
ORDER BY volume desc
```
None points to the top traditional news sources!
A Bounty

A Journey with Behzad Golshan & Evangelos Papalexakis
Citizens should have access to diverse perspectives as exposure to different views is beneficial for the advancement of humanity
- The Fairness Doctrine, FCC 1949

See [AGP WWW2015, KDD 2015, TIST 2017] for details
Diversity in Search Results, Used to be...

- Low level of result overlap, 85% of the search results unique to one engine [Spink et al. 2005]
- 72% of users used more than one search engine [White et al. 2012]
Research Questions

❖ Have Google and Bing results become similar?
❖ Can search over Social Data (Twitter) provide different and useful results?
Social Pulse

- Built on WaveFour Infrastructure
- Given query $q$ find all tweets that contain $q$ and a URL
- Rank them by number of retweets
- Implementation:
  - Using 1% sample from Twitter API
  - Restrict results to last 24 hours
  - Also include popular tweets without URLs
Methodology

- Two sets of queries
  - **Trends**: “Head” queries
    - Google Trends queries for April 2014
  - **Manual**: “Trunk” queries
    - Handpicked queries we were familiar with
- Data Collection
  - Top-10 results every day around the same time during June-July 2014
- Result Representation
  - **Google & Bing**: Union of bag-of-words for all snippets of top-10 results
  - **SocialPulse**: Union of bag-of-words for top-10 tweets

See [AGP TIST 2017] for details
Non-negative tensor decomposition of (query, result, engine) tensor - PARAFAC

The \((i,j)\) entry of \(C\) gives the participation of search engine \(i\) in cluster \(j\)

- Plot participation values of the search engines for different clusters
- Values cluster around \((0.5, 0.5)\) => E1 and E2 similar
- Values cluster around \((0, 1)\) and \((1,0)\) => E1 and E2 dissimilar

\[
\begin{bmatrix}
0.5 & 0 & 0.9 \\
0.5 & 1 & 0.1
\end{bmatrix}
\]

Latent Clusters
CrossLearnCompare (Quantification)

**Key Idea:**
If $M_A$ predicts results of Engine B $\rightarrow A, B$ are similar

- **Use classification accuracy as the measure of similarity/overlap**
- Each query is a class
- Feature representation of query to train a multi-class model
- One-vs-all linear SVM classifier
TensorCompare: Google vs Bing

High overlap! (esp. for Trends)
CrossLearnCompare: Google vs Bing

High overlap! (esp. for Trends)
(a) TensorCompare for Trends

(b) TensorCompare for Manual

Low Overlap!
CrossLearnCompare: Google vs Social Pulse

(c) CrossLearnCompare for TRENDS

(d) CrossLearnCompare for MANUAL

Low Overlap!
User Study

- So far, Twitter results are different.
- Are they useful??
- Amazon Mechanical Turk user study
  - Take results for a single day
  - Ask Turkers to judge how informative the results are
Usefulness of Results (Trends Query Set)
Usefulness of Results (Manual Query Set)
Looking Over the Horizon
Latency Between a New Tweet and its Appearance in WaveFour

Only ~10mins delay

<table>
<thead>
<tr>
<th>tweetid</th>
<th>userid</th>
<th>twetext</th>
<th>Creation Timestamp</th>
<th>UTCTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>44672586322772760</td>
<td>ConorAggregate Just donated. My 3 yr old has cheered you on ...</td>
<td>2014-03-20 19:11:58</td>
<td>2014-03-20 19:22:27 8400973</td>
</tr>
<tr>
<td>3</td>
<td>44672586322189860</td>
<td>RT @myfiespichi: Y’re acuerdo de aquel día en que me declar...</td>
<td>2014-03-20 19:11:58</td>
<td>2014-03-20 19:22:27 8400973</td>
</tr>
<tr>
<td>6</td>
<td>44672586314330308</td>
<td>#32BinTibbiLaborant/HanaBekiboy @At_Erdogan @Meezzing ...</td>
<td>2014-03-20 19:11:58</td>
<td>2014-03-20 19:22:27 8400973</td>
</tr>
</tbody>
</table>

-- show latest tweets

```sql
SELECT TOP 1000 posting.tweetid, posting.userid, posting.tweettext
    ,CreationTimestamp.value as CreationTimestamp, SYSUTCDATETIME() as UTCTime
FROM [wavefourtweets].[dbo].[Posting]
join [wavefourtweets].[dbo].CreationTimestamp on CreationTimestamp.tweetid = Posting.tweetid and CreationTimestamp.userid = Posting.userid
order by CreationTimestamp.value desc
```
Assertions

WaveFour was a worthwhile research expedition

Big social data is not garbage

Real-time business intelligence using social data is feasible and valuable

At the Data Insights Laboratories, I am taking another stab at the problem, but from a different perspective!

Magic happens when what is technically feasible meets what is desperately needed
Thank You!

Questions?