LASTA: Large Scale Topic Assignment on Multiple Social Networks

Nemanja Spasojevic, Jinyun Yan, Adithya Rao, Prantik Bhattacharyya
Klout, Inc.
{nemanja, jinyun, adithya, prantik}@klout.com
Highlights and Contributions

- **Fully deployed production system** to assign topics at scale
  - ~10,000 topics assigned to hundreds of millions of users daily
  - Reactive to fresh data

- Data from **multiple social networks** used to create an aggregated profile for a user:
  - Twitter, Facebook, LinkedIn, Google+, Wikipedia
  - User activity, profiles, connections

- **Feature engineering** approach that uses following categories:
  - Original generated content
  - Reactions to original content
  - Indirect attributions to user
  - Graph based features

- **Cross-Network information** leads to:
  - More topics assigned per user
  - More users who can be assigned topics
  - Better user-topic associations compared to using a single network
Klout

- Klout is a social influence measurement tool.
- Users register on Klout.com and connect their social network accounts.
- Klout collects authorized/public information from connected networks.
- Klout derives influence scores and topics for users from collected data.
Motivation

- Assign topics to the **long tail**

- Focus on **socially recognizable topics of interest**
  - Warren Buffett may be interested in *Ukulele* and *Online Bridge*, but is known for his recognizable interests like *Business* and *Money*.

- Applications in **Recommendation** and **Targeting** systems:
  - Content recommendations
  - User targeting
  - Question Answering

- **Extensibility** in terms of data sources.
Challenges in social data

- **Message size:**
  - Overall data size may be huge, but message size per user may be small.

- **Text Sparsity:**
  - Many users may be passive consumers of content.

- **Noise:**
  - Social content abounds in colloquial language, slang, grammatical errors, abbreviations.

- **Context:**
  - Need to expand context to get more information
Why use data from multiple networks?

- Phrase usage on different social networks is different
- Phrase overlap across social networks is small
- Combination of networks provides more quantity and diversity of phrases used.

Fig. 1. Verbosity distribution across social networks
Fig. 2. Phrase overlap on social networks
**Data Pipeline**

- **Facebook**: Authored status updates, shared URLs, commented and liked posts, text and tags associated with videos and pictures.
- **Twitter**: Authored tweets, retweets, mentions and replies on other tweets, shared URLs, created and joined lists.
- **LinkedIn**: Authored posts, comments, skills stated by the user and endorsed by connections.
- **Google+**: Authored messages, re-shares, comments, shared URLs and plus-ones.
- **Wikipedia**: Wikipedia pages for well known personalities.
System Details

- Topic Assignment runs as a bulk job on the Hadoop MapReduce stack
  - HDFS, Hive, HBase

- Exploded Resource footprint (uncompressed reads/writes from HDFS):
  - Feature Generation: 55.42 CPU days, 6.66 PB reads, 2.33 PB writes
  - Score generation: 11.33 CPU days, 3.78 PB reads, 1.09 PB writes

- Hive User Defined Functions (UDFs) implement utilities for data aggregation and transformation
  - https://github.com/klout/brickhouse

- Machine Learned models are trained offline and improved regularly.
A **Topic Feature** in the pipeline is represented as a bag-of-topics derived in a specific manner.
- eg. TW_MSG_TEXT => { (topic1, 1.0), (topic2, 3.0), … (topicN, 1.0) }
- A particular topic may occur in multiple bags of topics.

Data sources are attributed to users as:
- **Generated**: Original text, urls created and shared by a user.
- **Reacted**: Reactions to original content from a user.
- **Credited**: Attributions that do not depend directly on the activity of a user.
- **Graph**: Topics derived for friends, followers, connections.
Feature Engineering

- Each Feature is encoded as `<network>_<data-source>_<time-window>_<attribution>`

- Extensibility to create new features is important for experimentation and prototyping
  - e.g. Add a new time window, or a new data source

Network
- Data Source
- Time Window
- Attribution
Since we want socially recognizable topics, members in a user’s social graph evaluate topics for the user. Order is not considered during labeling.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># of participants</td>
<td>43</td>
</tr>
<tr>
<td># of evaluated users</td>
<td>766</td>
</tr>
<tr>
<td># of (user, topic) labels</td>
<td>32,264</td>
</tr>
<tr>
<td># of positive (user, topic) labels</td>
<td>17,208</td>
</tr>
<tr>
<td># of negative (user, topic) labels</td>
<td>15,056</td>
</tr>
</tbody>
</table>
Evaluation and results

Training:
- Transform bag of topics to a feature vector for each topic user pair (ti, u).
- Train a Binary Classification model using ground truth data.

Evaluation on test set:
- Single Network comparison
- Attribution Comparison
- Most Predictive Features

<table>
<thead>
<tr>
<th>User</th>
<th>Top 10 Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marissa Mayer</td>
<td>yahoo, google, technology, business, twitter, social-media, flickr, design,</td>
</tr>
<tr>
<td></td>
<td>marketing, seo, gmail</td>
</tr>
<tr>
<td>Lady Gaga</td>
<td>music, lady-gaga, celebrities, art, fashion, born-this-way, venus, entertainment, radio</td>
</tr>
<tr>
<td>Barack Obama</td>
<td>politics, affordable-care-act, healthcare, new-york-times, congress, chicago, twitter, washington, illinois</td>
</tr>
</tbody>
</table>
LASTA vs single networks

Better long tail performance:
LASTA assigns topics to a higher percentage of users, compared to using a single network.

More comprehensive per user:
LASTA assigns more topics per user than a single network.
### Cross-Network topics

#### Table 9: Super-topic percentage distribution across different networks

<table>
<thead>
<tr>
<th>Super-topic</th>
<th>LASTA</th>
<th>TW</th>
<th>FB</th>
<th>LI</th>
<th>GP</th>
<th>WIKI</th>
</tr>
</thead>
<tbody>
<tr>
<td>technology</td>
<td>23.972</td>
<td>19.706</td>
<td>11.559</td>
<td>33.420</td>
<td>22.822</td>
<td>8.247</td>
</tr>
<tr>
<td>business</td>
<td>15.893</td>
<td>10.628</td>
<td>7.567</td>
<td>41.053</td>
<td>12.857</td>
<td>10.937</td>
</tr>
<tr>
<td>lifestyle</td>
<td>7.910</td>
<td>7.403</td>
<td>11.409</td>
<td>2.328</td>
<td>7.969</td>
<td>4.810</td>
</tr>
<tr>
<td>government-and-politics</td>
<td>3.547</td>
<td>4.763</td>
<td>4.388</td>
<td>2.182</td>
<td>3.534</td>
<td>5.261</td>
</tr>
<tr>
<td>sports-and-recreation</td>
<td>4.379</td>
<td>7.503</td>
<td>7.591</td>
<td>0.659</td>
<td>4.913</td>
<td>7.921</td>
</tr>
<tr>
<td>food-and-drink</td>
<td>2.671</td>
<td>7.228</td>
<td>11.863</td>
<td>0.819</td>
<td>7.255</td>
<td>2.142</td>
</tr>
<tr>
<td>health-and-wellness</td>
<td>1.976</td>
<td>3.894</td>
<td>5.150</td>
<td>1.691</td>
<td>4.083</td>
<td>1.867</td>
</tr>
<tr>
<td>fashion</td>
<td>1.439</td>
<td>2.645</td>
<td>2.945</td>
<td>0.732</td>
<td>2.776</td>
<td>2.203</td>
</tr>
<tr>
<td>education</td>
<td>1.443</td>
<td>2.375</td>
<td>3.485</td>
<td>3.369</td>
<td>2.170</td>
<td>4.058</td>
</tr>
<tr>
<td>news-and-media</td>
<td>0.966</td>
<td>1.722</td>
<td>0.899</td>
<td>2.597</td>
<td>1.060</td>
<td>4.366</td>
</tr>
<tr>
<td>travel-and-tourism</td>
<td>0.535</td>
<td>0.779</td>
<td>1.155</td>
<td>0.614</td>
<td>1.041</td>
<td>0.654</td>
</tr>
<tr>
<td>hobbies</td>
<td>0.246</td>
<td>0.543</td>
<td>0.683</td>
<td>0.100</td>
<td>1.070</td>
<td>0.285</td>
</tr>
</tbody>
</table>
Key Takeaways

- Do not ignore the long tail.
- Using more than one social network offers the opportunity to get a deeper understanding of users.
- Expanding context is important for topic derivation.
- If you are designing a production system, ensure it has the following characteristics:
  - It is extensible
  - It allows fast experimentation and prototyping
Questions?