Social Media-Driven News Personalization

O'Banion, Birnbaum, Hammond
Knight News Innovation Lab
Northwestern University
Goal

News and media content recommendation based on an analysis of users’ social media activity.

Contributions:

1. User-modeling technique based on analysis of Twitter content.
2. News story recommendation system providing stories from The Huffington Post.
3. Unique evaluation technique using existing Twitter data as a basis for comparison.
Twitter as an Information Source

- Popularity
  - 100+ million active users, 200+ million tweets per day

- Unique practices promote frequent and concise posts

- Continuous and real-time information stream

---

Approach

- Identify named entities in text.

(Indiana Pacers, Organization, Sports)

“Indiana Pacers guard Paul George throws down a 360-degree, through-the-legs dunk”

(Paul George, Person, Sports)
Approach

- Build vertically aligned lexicons.
  - E.g., cooking terms (ingredients, equipment, actions)

"Recipe: Vegan White Cheese Party Dip [bit.ly/U93r3s]"

"Futures dip on manufacturing data: U.S. stock market futures are falling after the Commerce Department reported ... [bit.ly/OrorOY]"

Problems:
- Ambiguity
- Not always “newsworthy”
Approach

- Identify **categories** and **tags** in tweets.
  - Derived from the The Huffington Post’s taxonomy.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>People, Organizations, Topics, etc.</td>
</tr>
<tr>
<td>Books</td>
<td>Chicago Bulls</td>
</tr>
<tr>
<td>Business</td>
<td>Mitt Romney</td>
</tr>
<tr>
<td>College</td>
<td>Entrepreneurship</td>
</tr>
<tr>
<td>Education</td>
<td>Music</td>
</tr>
<tr>
<td>Entertainment</td>
<td>Beer</td>
</tr>
<tr>
<td>Food</td>
<td></td>
</tr>
<tr>
<td>Green</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Dataset

• The Huffington Post news story corpus.

Facebook Stock Plunges To New Low

SAN FRANCISCO — Facebook's stock plunged to a new low Thursday as some of the social networking leader's early backers got their first chance to sell their shares since the company's initial public offering went awry.

Analysts interpreted the unusually high trading volume as a clear sign that at least a few of the insiders were seizing on a fresh opportunity to cash out.
Classifiers

- Centroid-based classification algorithm\(^1\).
- A category/tag is represented as a term vector with tfidf weights.
- Inverted index maps distinct terms to weighted categories for quick lookup.

"Reading now: Matt Bai's @NYTimes magazine story on the Boehner/Obama debt deal. Fascinating."

"reading now matt bai's the new york times magazine story on the boehner obama debt deal fascinating"

\[\text{debt} \rightarrow [(\text{BUSINESS}, 0.00123), (\text{POLITICS}, 0.0004), ...] \]
Classifiers

"Reading now: Matt Bai's @NYTimes magazine story on the Boehner/Obama debt deal. Fascinating."

Category: Business

Tags: John Boehner, Debt Ceiling, Barack Obama, Debt Limit
User Profiling

A profile for user $u$ consists of categories $c$ and tags $t$:

$$P(u) = \{(c, f(c)) \mid c \in C\}, \{(t, f(t), \text{maxdate}(t)) \mid t \in T\}$$

Each tag $t$ is assigned a score:

$$TS(u, t) = f(t) \times 0.9^d$$

Each category $c$ is assigned a score:

$$CS(u, c) = \sum_{t \in TC} TS(u, t)$$
User Profiling

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food</td>
<td>40</td>
</tr>
<tr>
<td>World</td>
<td>7</td>
</tr>
<tr>
<td>Sports</td>
<td>39</td>
</tr>
<tr>
<td>Politics</td>
<td>6</td>
</tr>
<tr>
<td>Entertainment</td>
<td>17</td>
</tr>
<tr>
<td>Books</td>
<td>5</td>
</tr>
<tr>
<td>College</td>
<td>16</td>
</tr>
<tr>
<td>Arts</td>
<td>4</td>
</tr>
<tr>
<td>Technology</td>
<td>15</td>
</tr>
<tr>
<td>Travel</td>
<td>4</td>
</tr>
<tr>
<td>Education</td>
<td>12</td>
</tr>
<tr>
<td>Style</td>
<td>3</td>
</tr>
<tr>
<td>Healthy living</td>
<td>11</td>
</tr>
<tr>
<td>Business</td>
<td>3</td>
</tr>
<tr>
<td>Media</td>
<td>9</td>
</tr>
<tr>
<td>Religion</td>
<td>3</td>
</tr>
<tr>
<td>Green</td>
<td>7</td>
</tr>
<tr>
<td>Impact</td>
<td>2</td>
</tr>
</tbody>
</table>
A story $s$ with tags $t < T$ is assigned a score for user $u$: $\sum_{t \in T} TS(u, t)$.
Evaluation

- Gathering data on user preferences is difficult.

- Field Study
  - Requires a deployed system with significant user-base.

- Controlled Experiment
  - Potentially time consuming and costly.
Evaluation

- Our solution: Mine existing data for explicit indications of interest.

- For example:
## Twitter Dataset

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>1,000</td>
</tr>
<tr>
<td>Total user tweets</td>
<td>1,082,441</td>
</tr>
<tr>
<td>Avg. tweets per user</td>
<td>1,082</td>
</tr>
<tr>
<td>Avg. valid tweets per user</td>
<td>774</td>
</tr>
<tr>
<td>Distinct stories</td>
<td>729</td>
</tr>
<tr>
<td>Story tweet date range</td>
<td>1/12/12 - 2/11/12</td>
</tr>
<tr>
<td>Avg. tags per story</td>
<td>11.69</td>
</tr>
<tr>
<td>Categories represented</td>
<td>18</td>
</tr>
</tbody>
</table>
Evaluation

1. Identify explicit indications of interest.
2. Perform retroactive analysis.
3. Answer the question:
   - Would our system have recommended this story or a similar one to this user?

Similar stories:
- Overlapping Tags
- High Cosine Similarity (represented as term vectors with tfidf weighting)
Results

- Recall: Fraction of instances where a correct story was recommended
- Baseline: Stories featured on the “most popular” section of the Huffington Post.

![Recall Chart]

- "Popular" Stories: 15%
- Tweeted Story: 38%
- TS + Overlapping Tags: 62%
- TS + High Cosine Similarity: 69%
Acknowledgements

I’m grateful for the support of:
- National Science Foundation
- John S. and James L. Knight Foundation
- Cognitive Science Fellowship from Northwestern University

Website:
- http://twxray.knightlabprojects.com/
Related Work

- **User Modeling based on Social Media**
  - J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams.

- **News Media Content Recommender Systems**