Analyzing the Semantic Modeling Capabilities of Google Sets

Azfar Khandoker
Ryan Miller

April 23, 2013
Background

- **Google Sets**
  - Google Labs (2002)
  - Semantics behind search queries
  - Discontinued in 2011
  - Lives on in Google Spreadsheets!
Background

- Building a star graph using Google sets
  - Start with seed word, $s$
  - Use Google Sets to retrieve result set $R = \{ r_1, \ldots, r_n \}$
Background

- **Other Semantic Networks**
  - **Wikipedia**
    - 5.7M nodes, 130M edges
    - Avg. Degree: 45.531 (max_\text{in} = 374934 = "U.S." )
    - Avg. Path Len: ???
  - **Amazon Co-purchasing**
    - 542,000 nodes, 1.2M edges
    - Avg. Degree: 4.538 (max = 118 = "Laura")
    - Avg. Path Len: 2.842

---


Other Semantic Model Research

- Small World Phenomenon (Milgram, 1967; Watts & Strogatz, 1998)
  - Small average path length, high clustering
  - Power law degree distribution

  - Cosine similarity

\[
p(x) \propto x^{-\alpha}
\]

\[
\sigma_{ij} = \cos \theta = \frac{n_{ij}}{\sqrt{d_i d_j}}
\]

\[
n_{ij} = \sum_k A_{ik} A_{kj}
\]
Implementation

● Automate Google Sets output
● Build network starting from a seed word

● When to stop?
  ○ Two ideas implemented:
    ■ Destination word (perfect to relate two seemingly-unrelated words)
    ■ Depth level (better for general network analysis)
Implementation

- Building a network using Google sets
  - With repeated iterations, we generate a graph
  - Nodes represent words
  - Edge \((i, j)\) means word \(j\) was a result of seed word \(i\)

Graph of "mpg" to "oil"
Studies

- Shortest Path Characteristics
- Comparison of Semantic Networks
- Semantic Similarity
Studies

- Shortest Path: "wine" --> "france"

Paths found:

<table>
<thead>
<tr>
<th>Network</th>
<th>Path Length</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Sets</td>
<td>4</td>
<td>&quot;wine&quot; --&gt; &quot;champagne&quot; --&gt; &quot;bordeaux&quot; --&gt; &quot;france&quot;</td>
</tr>
<tr>
<td>Wiki</td>
<td>1</td>
<td>&quot;Wine&quot; --&gt; &quot;France&quot;</td>
</tr>
</tbody>
</table>
| Amazon       | 4           | "Wine" -->
|              |             | "Italy (Culinaria)" -->
|              |             | "Culinaria: The United States: A Culinary Discovery (Culinaria)" -->
|              |             | "Culinaria France (Culinaria Series)"                               |
Studies

- Shortest Path Characteristics
  - Small average path length, high clustering coefficient

<table>
<thead>
<tr>
<th></th>
<th>Google Sets</th>
<th>Wiki</th>
<th>Amazon</th>
</tr>
</thead>
<tbody>
<tr>
<td># nodes</td>
<td>2,871</td>
<td>544</td>
<td>1,373</td>
</tr>
<tr>
<td># edges</td>
<td>8,962</td>
<td>25,403</td>
<td>3,432</td>
</tr>
<tr>
<td>Avg. degree</td>
<td>6.243</td>
<td>93.393</td>
<td>5.998</td>
</tr>
<tr>
<td>Avg. path length</td>
<td>4.329</td>
<td>2.344</td>
<td>4.287</td>
</tr>
<tr>
<td>Clustering coeff.</td>
<td>0.255</td>
<td>0.727</td>
<td>0.343</td>
</tr>
</tbody>
</table>
Studies

● Small World Characteristics
  ○ Power law degree distribution
Studies

- **Shortest Path Characteristics**
  - Power law degree distribution

Entire Wiki Network 🗒️
# Studies

- **Comparison of Semantic Networks - "Oreo"**

<table>
<thead>
<tr>
<th></th>
<th>Google Sets</th>
<th>Wiki</th>
<th>Amazon</th>
</tr>
</thead>
<tbody>
<tr>
<td># nodes</td>
<td>10,229</td>
<td>16,478</td>
<td>1,119</td>
</tr>
<tr>
<td># edges</td>
<td>33,081</td>
<td>1,252,227</td>
<td>1,118</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Unique for Google Sets</th>
<th>Not in Google Sets</th>
<th>Shared</th>
<th>Percentage Shared</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Google Sets vs. Wiki</strong></td>
<td>9,682</td>
<td>15,931</td>
<td>547</td>
<td>3.320 %</td>
</tr>
<tr>
<td><strong>Google Sets vs. Amazon</strong></td>
<td>9,919</td>
<td>809</td>
<td>310</td>
<td>27.703 %</td>
</tr>
</tbody>
</table>
Studies

- Semantic Similarity
  - Cosine similarity

\[
\sigma_{ij} = \cos \theta = \frac{n_{ij}}{\sqrt{d_i d_j}}
\]

<table>
<thead>
<tr>
<th></th>
<th>Google Sets</th>
<th>Wiki</th>
<th>Amazon</th>
</tr>
</thead>
<tbody>
<tr>
<td>wine -&gt; france</td>
<td>0.0</td>
<td>0.0058</td>
<td>0.0096</td>
</tr>
<tr>
<td>wine -&gt; grape</td>
<td>0.0</td>
<td>0.02229</td>
<td>0.0344</td>
</tr>
<tr>
<td>wine -&gt; alcohol</td>
<td>0.1427</td>
<td>0.00921</td>
<td>0.0094</td>
</tr>
<tr>
<td>wine -&gt; beer</td>
<td>0.8007</td>
<td>0.1987</td>
<td>0.6862</td>
</tr>
</tbody>
</table>

\[
n_{ij} = \sum_k A_{ik} A_{kj}
\]
Conclusions

- Findings
  - Google Sets and Amazon both exhibit small world behavior
  - Overlap between semantic models for Google Sets and Amazon implies a use of Google Sets data for product recommendation
  - Cosine similarity can further refine recommendations by semantic meanings
Conclusions

● Future Work
  ○ Compare against other networks
    ■ Twitter
    ■ Facebook
    ■ The New York Times
  ○ Psychology
    ■ Word association of Humans vs. Google Sets
Questions?