Audience Segmentation Based on Topic Profiles

Ljubljana, 9 October 2017

Matic Kladnik, Luka Stopar, Blaz Fortuna, Dunja Mladenić
Motivation

• Web portals can offer frequently updated content:
  • News articles
  • Market data
  • Financial data

• First-time Users -> Returning Users -> Loyal Users

• Serve interesting content and advertisement to users
Motivation

• Audience Segmentation: dividing (website’s) audience into smaller groups

• Audience segmentation a key activity in audience analysis
Outline

• Data description

• Approach overview

• Architecture of the proposed approach

• Evaluation
Data Description

• Visit logs of users

• Content from almost 3000 pages (crawled)

• Anonymized user data from over 500k users from visit logs

• All page content (text) in English
Data Description

• Each web page represented as a Bag-of-Words (BoW) with TF-IDF weights

• Content Labels for each page
  • Assigned by editorial team
  • Used to annotate users
  • Examples: Brexit, jobs, Europe, markets

• User properties
  • Demographics
  • Page content labels
  • (pages visited)
Approach Overview

• Audience segmentation commonly groups users by their common interest, behavior, etc.

• However:
  • Users may have several interests
  • Exhibit different behavior (current focus)

• This approach might group users that only share some interests with a third user
Approach Overview

• Our approach: audience segmentation based on similarity of topics that users are interested in

• Obtain segments of users through topics of the visited web pages

• Same user can appear in several segments
Architecture of the Proposed Approach
Page Classification

• DMOZ classifier with Custom Taxonomy

• Classify each web page into a hierarchical content topic

• Pages classified to any level of hierarchy
Page Classification

• Upper level topics give context
  - Science/Technology/Aerospace
  - Business/Aerospace_and_Defense/Aeronautical

• More information about content topic
Features for Audience Segmentation

- In experimental evaluation we combine two sources of data

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
<th>No. of values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web page</td>
<td>BoW - Words from the Web pages</td>
<td>59929</td>
</tr>
<tr>
<td>User interest</td>
<td>Content Labels of the visited Web pages</td>
<td>1268</td>
</tr>
</tbody>
</table>
Cluster Dispersity Measure

• Used to compare influence of different feature sources

• Weighted average distance between examples and their centroid normalized by average distance to the global mean

\[
D = \frac{\sum_{i=1}^{k} \frac{n_i}{n} \sum_{j=1}^{n_i} \frac{1}{n_i} d(\mu_i, x_j)}{\frac{1}{n} \sum_{j=1}^{n} d(\mu, x_j)} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} d(\mu_i, x_j)}{\sum_{j=1}^{n} d(\mu, x_j)}
\]

• Examples in more compact clusters lie closer to the centroid
Evaluation

• Comparing relative mean distances for different data representations over different number of clusters (k)
Evaluation: example clusters

- From 50 clusters
- Some clusters with a broad range of topics

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Cluster 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business/Biotechnology_and_Pharmaaceuticals</td>
<td>Recreation/Models</td>
<td>Home/Personal_Finance</td>
<td>Health/Addictions</td>
<td>Business/Financial_Services/Venture_Capital/Regional</td>
</tr>
<tr>
<td>Health/Child_Health</td>
<td>Science/Astronomy</td>
<td>Recreation/Drugs</td>
<td></td>
<td>Business/Transportation_and_Logistics/Bus</td>
</tr>
<tr>
<td>Health/Conditions_and_Diseases/Cancer</td>
<td></td>
<td></td>
<td></td>
<td>Government/Agencies</td>
</tr>
<tr>
<td>Health/Conditions_and_Diseases/Immune_Disorders</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health/Conditions_and_Diseases/Infectious_Diseases</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health/Pharmacy</td>
<td></td>
<td></td>
<td></td>
<td>Science/Environment</td>
</tr>
<tr>
<td>Science/Astronomy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Science/Environment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation

• Comparing average size of the Audience Segments in relation to the granularity of segmentation

<table>
<thead>
<tr>
<th>No. of segments</th>
<th>Average size</th>
<th>Median size</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>43592.35</td>
<td>589.5</td>
</tr>
<tr>
<td>50</td>
<td>17436.94</td>
<td>301</td>
</tr>
<tr>
<td>100</td>
<td>8718.47</td>
<td>229.5</td>
</tr>
<tr>
<td>200</td>
<td>4359.235</td>
<td>193</td>
</tr>
</tbody>
</table>
Conclusion

• Topic profiles: user properties + content of web page

• Audience segmentation based on topic profiles of web pages

• Topics obtained by classifying pages with custom taxonomy

• Small scale experiments show some promise

• Future work:
  • Large scale experiments
  • Add additional topic features as in proposed architecture: user demographics etc.