A Compression Framework for User Profiles

Xiaoxiao Shi, Kevin Chang, Vijay K. Narayanan, Vanja Josifovski, Alexander J. Smola
Yahoo! Research, Santa Clara, CA
• User profiling
• Compression Framework
  • Term extraction using information geometry
  • Compression for clustering
  • Hierarchical formulation
• Experiments
• Discussion
Why user profiling
Why user profiling

Webpage Ranking

d → r → q
Why user profiling

Display Advertising

Webpage Ranking
Why user profiling

Display Advertising

Webpage Ranking

Answers
Why user profiling

Display Advertising

Answers

News

Webpage Ranking
Why user profiling

Display Advertising

Answers

News

Webpage Ranking
User profiling data

- Page views
- Queries
- Comments
- Clicks
- Timestamps
Previous work

• Bag of tokens (unordered set of activities)
  • “page views about banking”
  • “queries about cars”

• Represent user as distribution over tokens
  • Laplace smoothing for events
  • Background distribution is uniform over users

• Kullback Leibler divergence weighting
  (Konopnicki et al., 2010)
  • KL divergence between user distribution and background (#bits to encode u)
  • Term weight proportional to KL contribution

\[
D(p \parallel q) = \sum_t p(t) \log \frac{p(t)}{q(t)}
\]
Compression Framework

• **Basic Idea**
  - Encode objects by most meaningful subset of activities relative to background
  - Do this hierarchically

• **Hierarchical probabilistic model**

• **Encoding**

\[
p(x) = \int p(x|\alpha)p(\alpha|\beta)p(\beta|\gamma) \ldots d\alpha d\beta d\gamma
\]

• **Stagewise compression for key terms**

\[
D(p(x|\gamma)||p(x))
\]

\[
D(p(x|\beta)||p(x|\gamma))
\]
Compression Framework

General model

x → α → β → γ

Compression Framework

General model

Konopnicki et al. 2010
One stage encoding

encode user tokens relative to background
Compression Framework

General model

Konopnicki et al. 2010
One stage encoding
encode user tokens relative to background

Our model
Two stage encoding
encode cluster tokens relative to background

encode user tokens relative to cluster

x α β γ
Why?

- **Meaningful features**
  - Cluster ID is often meaningless
  - Activity / tokens often user interpretable
  - Retain user interpretable representation
  - ‘tokens are the best features’

- **Simplicity**

- **Flexibility**
  - Extend this to general hierarchical models
  - Smoothing via hierarchy of latent parameters

- **Data reduction**
  (think information theoretic frequent item sets)
• 1 million users at Yahoo!
• 55 day period (1/1-2/25/2010)
• Sample data (user interpretable)
  User ID,
  cpv_Technology/Online Community=20,
  cpv_Technology/Internet Services=35,
  cpv_Technology=40
Concentration by clusters

- Many users well represented by clusters
- Still meaningful divergence beyond clusters (clustering would oversimplify distribution)
Concentration by clusters

- Clusters capture overall distribution well
- Small set of users with significant deviation beyond cluster model (KL > 3.0)
Meaningful cluster features

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Important User Behaviors</th>
<th>KL Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>cpv-Technology/Internet Services</td>
<td>0.3883</td>
</tr>
<tr>
<td>(Technology Group)</td>
<td>cpv-Technology/Internet Services/Online Community</td>
<td>0.3854</td>
</tr>
<tr>
<td></td>
<td>cpv-Technology</td>
<td>0.3840</td>
</tr>
<tr>
<td></td>
<td>cpv-Technology/Internet Services/Online Community/Email</td>
<td>0.2829</td>
</tr>
<tr>
<td></td>
<td>cpv-Technology/Internet Services/Online Community/Portals</td>
<td>0.2806</td>
</tr>
<tr>
<td></td>
<td>cpv-Technology/Internet Services/Online Community/Photos</td>
<td>0.0122</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>cadv-Finance</td>
<td>0.0365</td>
</tr>
<tr>
<td>(Finance Group)</td>
<td>cadv-Finance/Credit Services</td>
<td>0.0274</td>
</tr>
<tr>
<td></td>
<td>cadv-Finance/Insurance</td>
<td>0.0145</td>
</tr>
<tr>
<td></td>
<td>cadv-Finance/Insurance/Automobile</td>
<td>0.0119</td>
</tr>
<tr>
<td></td>
<td>cadv-Telecommunications/Cellular and Wireless Services</td>
<td>0.0081</td>
</tr>
<tr>
<td></td>
<td>cadv-Technology/Consumer Electronics/Comms/Mobile/Cellular Telephones</td>
<td>0.0078</td>
</tr>
</tbody>
</table>

- Features ordered by relevance
- Represent cluster well
  - sparse encoding
# Meaningful user features

<table>
<thead>
<tr>
<th>User Index</th>
<th>Behavior set</th>
</tr>
</thead>
</table>
| **User 1** Cluster 1 | cadv-Technology/Internet Services/Online Community:134  
 cpv-Technology/Internet Services/Online Community:190  
 cpv-Technology/Internet Services:192  
 cpv-Technology:192  
 cpv-Technology/Internet Services/Online Community/Email:190  
 cadv-Technology:150  
 cadv-Technology/Internet Services/Online Community/Email:132  
 cadv-Technology/Internet Services:142  
 cadv-Finance/Insurance:52  
 cadv-Finance:52  
 cpv-Sports/Soccer:21  
 cpv-Sports/Auto Racing:32 |
| **User 2** Cluster 1 | cpv-Technology:212  
 cpv-Technology/Internet Services/Online Community:212  
 cpv-Technology/Internet Services:212  
 cpv-Technology/Internet Services/Online Community/Email:212  
 cadv-Consumer Packaged Goods:18  
 cadv-Consumer Packaged Goods/Beauty and Personal Care:14  
 cadv-Life Stages/Parenting and Children/Baby:10  
 cadv-Life Stages:10 |
| **User 3** Cluster 2 | cadv-Finance/Credit Services:72  
 cadv-Finance:190  
 cadv-Finance/Investment/Discount Brokerages:44  
 cadv-Technology/Consumer Electronics/Communication/Mobile/Cellular Telephones:34  
 cadv-Technology/Consumer Electronics/Communication/Mobile:104  
 cadv-Small Business and B2B:44  
 cadv-Life Stages:32  
 cadv-Retail:24 |
Advertising results

- Cluster smoothing helps most for users with little click data
- Subset at least as efficient as full set of features
- ... very preliminary results ...
• General compression framework
• Retains meaningful tokens
• Generalizes existing methods
• Todo: generative models (a la sequence memoizer for better representation)