Grouper: A Dynamic Clustering Interface to Web Search Results

Oren Zamir and Oren Etzioni
Department of Computer Science and Engineering
University of Washington

Doğan Altunbay and M. Burak Şenol
Introduction

• long ordered list of document «snippets»
• an example: Google

• main goal of the paper is to make search engine results easy to browse by clustering them

• post retrieval clustering is used

• an interface to the HuskySearch meta-search service
STC Algorithm

• Suffix Tree Clustering
  – works in linear time
  – based on identifying phrases
  – can create overlapping clusters
  – works incrementally
  – robust
  – does not need the number of clusters as an input
**Query: israel**

**Documents: 272, Clusters: 15, Average Cluster Size: 15.1 documents**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Size</th>
<th>Shared Phrases and Sample Document Titles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>Society and Culture (56%), Faiths and Practices (56%), Judaism (69%), Spirituality (56%); Religion (56%), organizations (43%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Ahavot Israel – The Amazing Jewish Website!</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Israel and Judaism</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Judaica Collection</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>Ministry of Foreign Affairs (33%), Ministry (87%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Publications and Data of the BANK OF ISRAEL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Consulate General of Israel to the Mid-Atlantic Region</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- The Friends of Israel Gospel Ministry</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>Israel Tourism (36%), Comprehensive Israel (36%), Tourism (64%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Interactive Israel tourism guide - Jerusalem</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Ambassade d'Israel</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Travel to Israel Opportunities</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>Middle East (57%), History (57%); WAR (42%), Region (42%), Complete (42%), Listing (42%), country (42%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Israel at Fifty: Our Introduction to The Six Day War</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Mashal - Volunteers in the Israel’s War of Independence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- HISTORY: The State of Israel</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
<td>Economy (68%), Companies (55%), Travel (55%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Israel Hotel Association</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Israel Association of Electronics Industries</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Focus Capital Group - Israel</td>
</tr>
</tbody>
</table>
Speed

![Graph showing the average clustering time (sec.) vs. number of retrieved documents. The graph is a line that increases as the number of retrieved documents increases.]
Evaluation

• The system is evaluated using session logs
  – Logs store the behavior of both search Uis, Grouper and HuskySearch

• Evaluation Perspectives
  – Cluster Coherence
    • How many clusters are visited after a particular number of documents are followed?
  – Grouper UI vs. Ranked List
    • How does the clustering affect efficiency of retrieval?
Cluster Coherence

• **Hypothesis:** Users will tend to follow documents from relatively few clusters.

• **Metric:** Average number of followed clusters as a function of the number of documents followed in the session.
  
  • Grouper’s clusters
  • Random clustering
  • STC clustering on HuskySearch results
  • K-means clustering on HuskySearch results
Cluster Coherence

The graph illustrates the number of followed clusters versus the number of documents followed thus far in the session. The lines represent different clustering methods:

- **Grouper**
- **HS data - STC clusters**
- **HS data - K-Means clusters**
- **Random clustering**
Grouper UI vs Ranked List

• **Number of documents followed**
  – A document is followed if the user clicked on it.

• **Time Spent**
  – Elapsed time between two successive document requests, including network latency, reading time, and traversal in results.

• **Click Distance**
  – **Ranked List:** Number of documents between two clicks.
  – **Grouper Interface:** Number of clusters and snippets between two clicks.
**Grouper vs. Ranked List**

<table>
<thead>
<tr>
<th>Num. of Docs. Followed:</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8+</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of HuskySearch sessions</td>
<td>53.0</td>
<td>26.9</td>
<td>8.4</td>
<td>4.2</td>
<td>2.3</td>
<td>1.6</td>
<td>1.1</td>
<td>0.7</td>
<td>1.9</td>
</tr>
<tr>
<td>% of Grouper sessions</td>
<td>46.0</td>
<td>25.2</td>
<td>10.2</td>
<td>6.0</td>
<td>3.9</td>
<td>2.4</td>
<td>1.8</td>
<td>1.4</td>
<td>3.2</td>
</tr>
</tbody>
</table>

**Graphs:**
- Left: Time spent on document (sec.)
  - HuskySearch
  - Grouper

- Right: Click distance
  - HuskySearch
  - Grouper - snippets
  - Grouper - total
Conclusion

• The paper introduces a clustering interface to HuskySearch meta search engine.

• Two issues are forwarded unresolved to Grouper II:
  – Grouper should provide a view of non-merged base clusters, which may be helpful for novice users.
  – For scaling considerations, clusters should be presented hierarchically so users can navigate the results more efficiently.