Contextual Suggestion Track
TREC
Thaer Samar, Alejandro Bellogin, Jimmy Lin, Arjen P. de Vries, Alan Said
Summary

- Content-based recommendation
  - Computes the similarity between documents and users profiles

- Classifier (not submitted)
  - Training data:
    - + Yelp, tripadvisor, wikitravel, zagat, yahoo-travel, orbitz
    - - Random sample

- Using full ClueWeb12
Statistics:
- From February to May 2012
- 5.5 TB (compressed)
- 27.3 TB (uncompressed)
- 33,447 WARC files
- 733,019,372 documents

Hadoop cluster:
- 90 computing nodes
- 720 parallel map/reduce tasks
Profiles & Attractions Files

Generate Profiles

<userID, descriptions>

Dictionary

Transform Profiles

<userID, {termId, tf}>

Transform Profiles

<userID, contextId, docId, score>

Sim(Document, user)

Generate Desc & Titles

<userID, contextId, docId, desc, title>

<userID, contextId, docId, rank, desc, title>

Generate Ranked list

ClueWeb12 WARC Files

Find Context

<(contextId, docId), doc content>

Generate Dictionary

<term, id>

Transform Documents

<(contextId, docId), {termId, tf}>

Generate Ranked list

<userID, contextId, docId, rank, desc, title>
Find Context

- **Goal:** extract relevant documents for each context

- **How do we measure the relevance?**
  - Exact mention of the context (format: {City, ST})
    - Kennewick, WA
  - Exclude non related sentences
    - I am in Kennewick, washing ... 
  - Exclude documents that mention the city of interest but in different states
    - Greenville, NC and Greenville, SC

- **We found 13,548,982 documents out of 733,019,372 ClueWeb12 documents**
Generate profiles

- We used the description of attractions rated by the user to generate his profile

- Why descriptions not the attraction website
  - 7 urls were found with one-one matching
  - 35 were found considering hostname matches and url variation, i.e., http(s), www
  - ratings for the attraction's descriptions and websites were very similar
Documents & profiles representation

- Vector Space Model

- Elements of the vectors are <term, frequency> pairs

- Efficient in terms of:
  - Size
    - 918 GB (before)
    - 40 GB (after)
  - Processing speed

- More complete implementation in https://github.com/lintool/clueweb
Similarity

- Cosine similarity between profile and document vector space representation
Descriptions and final results

\( <\text{contextid,docid}, \text{doc_content}> \)

Break document content into sentences

Keep sentences related to context

Concatenate candidates

\( <\text{contextid,docid}, \text{description, title}> \)

\( <\text{contextid,docid}, \{\text{termid, tf}\}> \)

Compute Similarity

\( <\text{userid}, \{\text{termid, tf}\}> \)
From Hadoop Distributed Cache

\( <\text{userid, contextid, docid, score}> \)

join
Results
Analysis

- We asked the following questions
  - Effect of sub-collection creation (context finding)
  - Effect of similarity function
  - Rating bias in ClueWeb vs Open Web
Effect of sub-collection creation 1/2

- Re-run our approach on the sub-collection given by organizers
  - 27% of given sub-collection are in our sub-collection

<table>
<thead>
<tr>
<th>Method</th>
<th>MRR</th>
<th>MRR&lt;sub&gt;d&lt;/sub&gt;</th>
<th>P@5&lt;sub&gt;d&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBCosTop1</td>
<td>0.0559</td>
<td>0.0745</td>
<td>0.0587</td>
</tr>
<tr>
<td>IBCosTop1 (given)</td>
<td>0.0528</td>
<td><strong>0.0955</strong></td>
<td>0.0484</td>
</tr>
</tbody>
</table>
Effect of sub-collection creation 2/2

- Significant improvement when ignoring the geographical aspect \((P@5_g)\)
- Our method retrieves relevant documents for the user but not geographically appropriate
- The given sub-collection is more appropriate for the contexts

<table>
<thead>
<tr>
<th>Method</th>
<th>MRR</th>
<th>MRR(_d)</th>
<th>(P@5_d)</th>
<th>(P@5_{d\bar{g}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBCosTop1</td>
<td>0.0559</td>
<td>0.0745</td>
<td>0.0587</td>
<td>0.2202</td>
</tr>
<tr>
<td>IBCosTop1 (given)</td>
<td>0.0528</td>
<td>0.0955</td>
<td>0.0484</td>
<td>0.0780</td>
</tr>
</tbody>
</table>
Effect of ranking function

- (Low coverage of relevance assessment)
- 5-nearest neighbour outperform other k-neighbours
- Generating user profiles based on descriptions with negative rating gave the worst results

<table>
<thead>
<tr>
<th>Method</th>
<th>MRR</th>
<th>MRRₜₐ</th>
<th>P@5ₜₐ</th>
<th>P@5ₜₐ̅</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBCosTop1</td>
<td>0.0559</td>
<td>0.0745</td>
<td>0.0587</td>
<td>0.2202</td>
</tr>
<tr>
<td>IBCosTop1 + 5NN text cos</td>
<td>0.0455</td>
<td>0.0562</td>
<td>0.0330</td>
<td>0.1486</td>
</tr>
<tr>
<td>IBCosTop1 + 5NN text Jacc</td>
<td>0.0433</td>
<td>0.0521</td>
<td>0.0330</td>
<td>0.1294</td>
</tr>
<tr>
<td>IBCosTop1 + 5NN rating cos</td>
<td>0.0429</td>
<td>0.0553</td>
<td>0.0349</td>
<td>0.1477</td>
</tr>
<tr>
<td>IBCosTop1 + 5NN rating Pearson</td>
<td>0.0450</td>
<td>0.0580</td>
<td>0.0358</td>
<td>0.1560</td>
</tr>
<tr>
<td>Classifier + 5NN text cos</td>
<td>0.0045</td>
<td>0.0112</td>
<td>0.0036</td>
<td>0.0251</td>
</tr>
<tr>
<td>Classifier + 5NN text Jacc</td>
<td>0.0045</td>
<td>0.0121</td>
<td>0.0045</td>
<td>0.0260</td>
</tr>
<tr>
<td>Classifier + 5NN rating cos</td>
<td>0.0045</td>
<td>0.0090</td>
<td>0.0027</td>
<td>0.0242</td>
</tr>
<tr>
<td>Classifier + 5NN rating Pearson</td>
<td>0.0045</td>
<td>0.0067</td>
<td>0.0018</td>
<td>0.0233</td>
</tr>
<tr>
<td>Positive profile</td>
<td>0.0396</td>
<td>0.0588</td>
<td>0.0359</td>
<td>0.1498</td>
</tr>
<tr>
<td>Negative profile</td>
<td>0.0045</td>
<td>0.0045</td>
<td>0.0009</td>
<td>0.0152</td>
</tr>
<tr>
<td>Positive + 5NN text cos</td>
<td>0.0426</td>
<td>0.0572</td>
<td>0.0341</td>
<td>0.1399</td>
</tr>
</tbody>
</table>
Archive Web vs Open Web evaluation

**Open Web**
- Mean = 2.16
- Median = 3

**ClueWeb12**
- Median = 1
- Mean = 1.24
Thanks!