Conference Classifier and Meta Data Extraction

Course Project
CS-498
Prof. ChengXiang Zhai
Objective

- To extract useful information from the conference page. e.g., workshop
- Limiting search to specific domain
- Have high recall rate.
Block Diagram
• Conference URL’s manually obtained (~100)
• Negative training data obtained by web search for arbitrary queries (~100)
• Make dictionary.
Training Phase -- Preprocessing

- HTML Parsing, Stop Word Removal, Remove noisy words.
- Index data with different fields and extract Features.
Feature Extraction for Classification

- Vector space model is used for feature representation.
- Phrase = each word.
- Features: Title, H1, H2, H3
- Vector Weight = \( f(\text{TF},\text{IDF},\text{wt(Features)}) \)
Training Phase -- Step 3

- SVM
- Language Model
• Fetch query and do a web search.
Testing Phase – Step 2

- Fetch query and do a web search.
- Process the contents of the webpage.
Extracting Meta-Data

- Crawling to increase the set of URL. (Why?)
- URL Country code.
- Binary classification of keywords.
- Attempt at using Open-NLP.
Testing Phase – Step 3

Top Trigrams

Can be used for increasing recall rate.

but are not
Call For Papers 18
The conference will 17
…
## Results

- **SVM**
  - Training Set 92%
  - Test Set 36%

- **Language Model**

<table>
<thead>
<tr>
<th>Discounting Method</th>
<th>Average Positive Perplexity</th>
<th>Average Negative Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>608</td>
<td>2810</td>
</tr>
<tr>
<td>Absolute</td>
<td>594.97</td>
<td>3095.58</td>
</tr>
<tr>
<td><strong>Witten Bell</strong></td>
<td><strong>696.28</strong></td>
<td><strong>3408.71</strong></td>
</tr>
<tr>
<td>Good Turing</td>
<td>516.46</td>
<td>2453.26</td>
</tr>
</tbody>
</table>
In Retrospect

- Feature engineering.
- Look at it from retrieval point of view.
- Use Available Tools.
- Larger dataset poses serious problems.
- Many small issues…
Future Work

- Build a complete search engine on top of conference database.
- Extract more relevant data.
- Explore retrieval point of view.
- Clustering of conferences.
- Incorporate domain specific stemming.
Questions?

- Thank You