Motivation and Problem Definition

**Motivation**

The rapid development of Web and social media often floods users with huge volumes of information. To understand the web documents and related social content, we need to know what topics the document and social media talk about, which part in the news does the social content focuses, and what others discuss over the part that I’m interested in.

**Problem Definition**

Given the document $d$ and associated with $C$ and all of them talks about $T$, Social Content Alignment is to generate a set of matching pairs $<\text{social content},\text{topic}>$, namely $\{(c_i, t_j)\}$ where $c_i \in C, t_j \in T \cup \emptyset$, which means social content $c$ discusses the specific topic $t$ and 0 means there is no such topic in the document.

**Example**

<table>
<thead>
<tr>
<th>Positive examples</th>
<th>Hyper Sphere</th>
<th>Classify</th>
</tr>
</thead>
<tbody>
<tr>
<td>Com</td>
<td>SVM</td>
<td>Reclassify</td>
</tr>
<tr>
<td>Sina</td>
<td>0.980</td>
<td>PP &amp; PP</td>
</tr>
<tr>
<td>Com</td>
<td>0.984</td>
<td>PP &amp; PP</td>
</tr>
<tr>
<td>Com</td>
<td>0.984</td>
<td>PP &amp; PP</td>
</tr>
<tr>
<td>Com</td>
<td>0.984</td>
<td>PP &amp; PP</td>
</tr>
<tr>
<td>Com</td>
<td>0.984</td>
<td>PP &amp; PP</td>
</tr>
</tbody>
</table>

**Preliminary Work**

Data preparation is an important step. The desired dataset contains 461 million (78% of internet users) in China and 4.069 million (9.4%) in the United States [CNNIC, Jun. 2013]. Comment number for top news in Yahoo and Sina are 5964.4 and 9325.4 [Nov. 2012].

**Framework**

- Document Comment Topic Model (DCT Model)
- Unbalanced volume
- Lack of labeled data
- Different vocabulary
- Sparse feature
- Repetition

**PU Learning**

- Goal: build a classifier to identify more accurate comments for a given topic
- Assumption: the topic sentences in news can be used as positive examples
- Core Idea: due to it is difficult to build an accurate classifier with very few positive and noise negative examples, we try to extend the positive example set as well as purify the negative set in three steps:
  1. Three example sets
  2. Ricchio Classifier
  3. Reclassify

**Experiment**

**Dataset**

Basic Information:
- Total (cn + en): 22 (10 + 12) news
- 1950 (516 + 434) sentences
- 8.219 (4,969 + 2,150) comments
- 7 annotators
- Confidence: 5 out 7 agree
- 9.847 (7,520 + 2,327) links

Statistics

<table>
<thead>
<tr>
<th>Source</th>
<th>Number of Sen/Comm</th>
<th>Words</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sina</td>
<td>516</td>
<td>8,932</td>
<td>2,772</td>
</tr>
<tr>
<td>Com</td>
<td>4,069</td>
<td>11,283</td>
<td>3,155</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>434</td>
<td>5,767</td>
<td>2,679</td>
</tr>
<tr>
<td>Com</td>
<td>2,679</td>
<td>39,917</td>
<td>9,972</td>
</tr>
</tbody>
</table>

Annotation Observation

- 87% $\rightarrow$ one or more news sentences
- 13% $\rightarrow$ no sentences
- Conclusion: it is reasonable to make use of comments to enhance topic detection in DCT model.

**Conclusion and Future Work**

- Study the social content alignment problem and present a two-phase framework to address it
- Propose DCT model which exploits Web document, social content, and their dependency
- Employ PU learning algorithm for alignment
- Experiments show the effectiveness of the proposed approach

**Future Work**

- Social content alignment over similar web documents
- Investigate whether the social relationships influence the alignment
- Study topic drift in the social content

Email: houlei@se.google.com
Aug. 6, 2013 @ BICC, Beijing, China