Interactive Clustering of Text Collections According to a User-Specified Criterion

Ron Bekkerman, *University of Massachusetts, Amherst*
Hema Raghavan, *University of Massachusetts, Amherst*
James Allan, *University of Massachusetts, Amherst*
Koji Eguchi, *National Institute of Informatics, Tokyo*

Email: ronb@cs.umass.edu
Outline

- Problem statement and motivation
- Underlying technology
  - Combinatorial Markov Random Fields
- Interactive clustering algorithm
- Results and conclusions
Data analysis

- Given a collection of documents
  - No labels
- Tell me *something* about its structure
  - What something? 😊
- Cluster it!
  - Group documents *by topic*
    - Where topic is some kind of word similarity
Topical clustering

- Example: two news titles:
  - “Seafood Benefits Outweigh Potential Risks”
  - “One Study Calls Fish a Lifesaver, Another Is More Cautious”

- Clustering won’t work here
  - The one based on BOW representation

- But what if you don’t care about the topic?
Non-topical clustering

- Clustering by author's mood at the writing time
  - Words like "happy", "upset", "tired" would help
  - Smilies!

- Clustering by author's attitude
  - Words like "great", "mediocre", "awful" do the work
  - Topical words wouldn't help

- Clustering by documents' genre
  - Not easy to come up with representative words
  - Syntactic features might help (e.g. POS tag n-grams)

- How trustable is this source of information?
  - Particular words / phrases / expressions
  - Layout details

- How easy is the text for reading?
  - Particular words can be chosen ("synchrophasotron")
  - Syntactic features are crucial (parse trees?)

Document representation makes a difference here!
Solution: interactive clustering

- A useful tool for on-the-fly data analysis
- The user comes up with *feature types*
  - Which are *modalities* of the data
- The user comes up with *feature examples*
  - For the modalities where it is possible
- We apply our technology
  - For multi-modal clustering
  - While enriching feature lists
Combinatorial MRF (Comraf)

Data modalities are represented with one random variable each
Which are nodes in a Comraf graph $G$
Edges are interactions between the modalities
Chef’s special inference method:
- Local combinatorial optimization for nodes
- Iterative Conditional Mode for traversing $G$

Bekkerman et al., ECML-2006
Multi-way distributional clustering

A specific inference algorithm used in Comrafs for multi-modal clustering

- Multi-modal clustering: simultaneously constructing $N$ clusterings of $N$ data modalities

Combination of:
- Top-down clustering for some modalities
- Bottom-up clustering for the others
- Local optimization at each iteration

Bekkerman et al., ICML-2005
Example: 4-way clustering
Example: interactive clustering

- User labeling
- Doc clustering
- Word clustering
- User correction
- Doc clustering
- Etc…
Clustering by genre

- Feature types:
  - BOW
  - POS tag n-grams (1- 2- 3- and 4-grams)
  - Both BOW and POS n-grams (2-grams)
- No feature examples can be given
- 2-way or 3-way distributional clustering used as it is
Genre: experimental setup

- **Dataset:**
  - British National Corpus
  - 21 categories (genres), 32 documents in each
  - Semi-manually POS-tagged (91 tags overall)
  - 64,000 unique words, 6000 POS 2-grams

- **Baselines:**
  - K-means (WEKA's implementation)
  - LDA (Xuerui Wang's implementation)
Genre: results

<table>
<thead>
<tr>
<th>Doc representation</th>
<th>K-means</th>
<th>LDA</th>
<th>Comraf</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW</td>
<td>9.1%</td>
<td>55.4±0.1%</td>
<td>55.7±0.2%</td>
</tr>
<tr>
<td>POS 2-grams</td>
<td>23.2%</td>
<td>44.7±0.2%</td>
<td>51.0±0.2%</td>
</tr>
<tr>
<td>BOW + POS 2-grams</td>
<td></td>
<td></td>
<td>58.5±0.6%</td>
</tr>
</tbody>
</table>

Comraf accuracy with POS ngrams

Comraf accuracy with BOW

Accuracy with threshold on low frequent words
Clustering by sentiment

- Application of interactive clustering of movie reviews
  - *Harry Potter and the Goblet of Fire*
  - 1613 reviews downloaded from IMDB.com
  - With original user ratings

- Ratings are translated to 4 categories:
  - Strongly disliked, somewhat disliked, somewhat liked, strongly liked
5 users were given a list of 563 words
- Marked between 26 and 58 words from it
- Revised word clusters after each iteration

Oracle:
- For each category $C$, 25 words were chosen
  - From a list of 4295 “sentimental” words
  - Their distribution over categories had a peak at $C$

Baseline (besides k-means and LDA):
- SVM trained on 22,546 reviews
  - To 46 popular Hollywood movies of 2005
## Sentiment: accuracy

<table>
<thead>
<tr>
<th>Doc represent.</th>
<th>K-means</th>
<th>LDA</th>
<th>SVM</th>
<th>Comraf</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW</td>
<td>28.2</td>
<td>37.0±0.2</td>
<td>39.1±0.3</td>
<td>40.3±0.8</td>
</tr>
<tr>
<td>Sentiment list</td>
<td>29.0</td>
<td>40.2±0.5</td>
<td>41.3±0.6</td>
<td>43.0±0.9</td>
</tr>
<tr>
<td>Interactive clustering (Oracle)</td>
<td></td>
<td></td>
<td></td>
<td><strong>47.1±0.2</strong></td>
</tr>
</tbody>
</table>

The table above compares the accuracy of different document representations and algorithms for sentiment analysis. The accuracy is measured in percentages with standard deviations (±). The Interactive clustering (Oracle) model achieves the highest accuracy of 47.1% with a standard deviation of 0.2%.
Sentiment: accuracy by user

![Bar chart showing accuracy by user with different correction steps and seed words.](chart.png)
Sentiment: accuracy by category
Conclusion

- One of the first works on non-topical clustering
- General approach proposed
  - Based on Comraf paradigm
- Scores are high for genre
  - Low for sentiment
    - Which shows how difficult the problem is
- Thank you!