What Users Care about: A Framework for Social Content Alignment

Lei Hou¹, Juanzi Li¹, Xiaoli Li², Jiangfeng Qu¹, Xiaofei Guo¹, Ou Hui¹, Jie Tang¹

¹ Knowledge Engineering Group, Dept. of Computer Science and Technology, Tsinghua University
² Institute for Infocomm Research, A*STAR, Singapore
Outline

• Motivation & Challenges
• Related Work
• Approach
• Experiment
• Conclusion & Future Work
Motivation

78% of Internet users in China (461 million) read news online [Jun, 2013, CNNIC]

The average numbers of comments for top news in Yahoo! and Sina are 5684.6 and 9205.4 respectively (on Nov, 2012)

How to find what the users care about
Motivation

• How to achieve that?
  – Link sentences and comments \(\leftrightarrow\) Social Content Alignment

• How to align?

WASHINGTON—...
Boehner won the backing of 220 Republicans, who retained a majority in the chamber after November's election. But a handful of GOP members voted no or abstained. Most Democrats voted for House Minority Leader Nancy Pelosi.
Boehner's grasp on his speakership seemed tenuous going into the vote. ....
Several northeastern Republicans loudly criticized Boehner for stalling a $60 billion relief bill for states hit by Superstorm Sandy. Boehner has pledged to hold a vote on Sandy relief on Friday.
....
Once the votes were cast and Boehner was announced the winner, Republican and Democratic leaders joined the Ohio delegation in escorting Boehner to the speaker's chair, where he will serve for two more years. In his first speech to the 113th Congress, Boehner urged members to remain true to the Constitution and focused his remarks on the national debt. "Our government has built up too much debt. Our economy is not producing enough jobs. These are not separate problems," Boehner told the members in the chamber. "At $16 trillion and rising, our national debt is draining free enterprise and weakening the ship of state. "The American Dream is in peril so long as its namesake is weighed down by this anchor of debt. Break its hold, and we begin to set our economy free."
Challenges

- Sparse feature (average length <40)
- Non-uniform vocabulary (<10% in common)
- Lack of labeled data (thousands of comments)

- Similarity based method
- Supervised learning
Related Work - social content analysis

- **Readalong**: reading articles and comments together.
  - Dyut Kumar Sil, Srinivasan H. Sengamedu, and Chiranjib Bhattacharyya.
  - In WWW’11 (poster)

- **Supervised matching of comments with news article segments**.
  - Dyut Kumar Sil, Srinivasan H. Sengamedu, and Chiranjib Bhattacharyya.
  - In CIKM’11 (short paper)

- **Opinion integration through semi-supervised topic modeling**.
  - Yue Lu and Chengxiang Zhai.
  - In WWW’08
Related Work - topic modeling

• A time-dependent topic model for multiple text streams.
  – Liangjie Hong, Byron Dom, Siva Gurumurthy, and Kostas Tsioutsioulklis.
  – In KDD’11

• Multi-topic based query-oriented summarization.
  – Jie Tang, Limin Yao, and Dewei Chen
  – In SDM’09

• Cross-domain collaboration recommendation.
  – Jie Tang, Sen Wu, Jimeng Sun, and Hang Su.
  – In KDD’12,
Related Work—positive unlabeled learning

• Building text classifiers using positive and unlabeled examples.
  – Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S. Yu.
  – In ICDM’03

• Learning with positive and unlabeled examples using weighted logistic regression.
  – Wee Sun Lee and Bing Liu.
  – In ICML’03.

• Learning to classify texts using positive and unlabeled data.
  – Xiaoli Li and Bing Liu.
  – In IJCAI’03.

• Learning to identify unexpected instances in the test set.
  – Xiaoli Li, Bing Liu, and See-Kiong Ng.
  – In IJCAI’07.
Approach Framework

PHASE 1

Document Comment Topic Model

• Different vocabulary
• Sparse feature
• Dependency

PHASE 2

Learning from Positive and Unlabeled Data

• Unbalanced volume
• Lack of labeled data
Document-Comment Topic Model

Step 1: Initialize a standard LDA model over $S$;

Step 2:

The left only uses comments, and the right takes news as background.

Top words for topic *launch cost*

<table>
<thead>
<tr>
<th>Comment only</th>
<th>News only</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aid</td>
<td>Korea</td>
<td>Money</td>
</tr>
<tr>
<td>Stomach</td>
<td>America</td>
<td>Launch</td>
</tr>
<tr>
<td>America</td>
<td>Food</td>
<td>Food</td>
</tr>
</tbody>
</table>

Algorithm 1: Generative process for DCT model.

**Input:** the priors $\alpha$, $\beta$, $\gamma_c$, $\gamma_s$; $S$ and $C$

**Output:** estimated parameters $\theta_c$, $\theta_s$, $\lambda$ and $\phi$

foreach comment $c \in C$ do

foreach word $w_{ci} \in c$ do

Toss a coin $x_{ci}$ according to $bernoulli(x_{ci}) \sim beta(\gamma_s, \gamma_c)$, where $beta(.)$ is a beta distribution, and $\gamma_c$ and $\gamma_s$ are two parameters;

if $x_{ci} = 0$ then

Draw a topic $z_{ci} \sim multi(\theta_c)$ from a comment-specific topic mixture;

else

Draw a topic $z_{ci} \sim multi(\theta_s)$ from a document-related topic mixture;

end

draw a word $w_{ci} \sim multi(\phi_{z_{ci}})$ from $z_{ci}$-specific word distribution;

end

Algorithm 2: PU learning

Input: news sentences $S$, social contents $C$, topic distribution $\theta$, word distribution $\phi$

Output: A set of classifiers

for each topic do

1. Extract the positive and unlabeled example set;
2. Build first classifier:
   - calculate centroid and radius to construct a hyper-sphere
   - extract potential positive examples and negative examples
   - build first classifier using Ricchio
3. Build final classifier:
   - classify unlabeled data using first classifier
   - build final classifier using WSVM
end

<table>
<thead>
<tr>
<th>topic</th>
<th>vote</th>
<th>relief</th>
<th>...</th>
<th>debt</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>0.173</td>
<td>0.039</td>
<td>...</td>
<td>0.094</td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.082</td>
<td>0.127</td>
<td>...</td>
<td>0.077</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_M$</td>
<td>0.184</td>
<td>0.083</td>
<td>...</td>
<td>0.105</td>
</tr>
<tr>
<td>$C_1$</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$C_2$</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$C_N$</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Positive example for topic vote
1. But a handful of GOP members voted no or abstained.
2. Boehner’s ... seemed tenuous going into the vote.
3. Once the votes were cast and ... .
Algorithm 2: PU learning

**Input:** news sentences \( S \), social contents \( C \), topic distribution \( \theta \), word distribution \( \phi \)

**Output:** A set of classifiers

for each topic do

1. Extract the positive and unlabeled example set;
2. Build first classifier:
   - calculate centroid and radius to construct a hyper-sphere
   - extract potential positive examples and negative examples
   - build first classifier using *Ricchio*
3. Build final classifier:
   - classify unlabeled data using first classifier
   - build final classifier using WSVM
end

<table>
<thead>
<tr>
<th></th>
<th>( f_1 )</th>
<th>( f_2 )</th>
<th>( \ldots )</th>
<th>( f_K )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_1 )</td>
<td>0.043</td>
<td>0.019</td>
<td>( \ldots )</td>
<td>0.024</td>
</tr>
<tr>
<td>( P_2 )</td>
<td>0.052</td>
<td>0.037</td>
<td>( \ldots )</td>
<td>0.017</td>
</tr>
<tr>
<td>( \ldots )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( P_{</td>
<td>P</td>
<td>} )</td>
<td>0.054</td>
<td>0.033</td>
</tr>
</tbody>
</table>

Average \( \rightarrow \) Centroid
Outside \( \rightarrow \) Potential Negative
Max distance \( \rightarrow \) Radius

Inside \( \rightarrow \) Potential Positive
PU Learning

Algorithm 2: PU learning

Input: news sentences $S$, social contents $C$, topic distribution $\theta$, word distribution $\phi$

Output: A set of classifiers

for each topic do

1. Extract the positive and unlabeled example set;
2. Build first classifier:
   - calculate centroid and radius to construct a hyper-sphere
   - extract potential positive examples and negative examples
   - build first classifier using Ricchio
3. Build final classifier:
   - classify unlabeled data using first classifier
   - build final classifier using WSVM

end

Adjust the label according to $S_1$ and $S_2$, as well as assign a confidence score

$$L = \frac{\max(s_1, s_2)}{s_1 + s_2}$$
PU Learning

**Algorithm 2: PU learning**

**Input:** news sentences $S$, social contents $C$, topic distribution $\theta$, word distribution $\phi$

**Output:** A set of classifiers

for each topic do

1. Extract the positive and unlabeled example set;
2. Build first classifier:
   - calculate centroid and radius to construct a hyper-sphere
   - extract potential positive examples and negative examples
   - build first classifier using Ricchio
3. Build final classifier:
   - classify unlabeled data using first classifier
   - build final classifier using WSVM

end

<table>
<thead>
<tr>
<th></th>
<th>$L$</th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$\ldots$</th>
<th>$f_K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>1</td>
<td>0.043</td>
<td>0.019</td>
<td>$\ldots$</td>
<td>0.024</td>
</tr>
<tr>
<td>$P_2$</td>
<td>1</td>
<td>0.052</td>
<td>0.037</td>
<td>$\ldots$</td>
<td>0.017</td>
</tr>
<tr>
<td>$\ldots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$LP_1$</td>
<td>0.7</td>
<td>0.054</td>
<td>0.033</td>
<td>$\ldots$</td>
<td>0.015</td>
</tr>
<tr>
<td>$\ldots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$LN_1$</td>
<td>0.83</td>
<td>0.003</td>
<td>0.061</td>
<td>$\ldots$</td>
<td>0.055</td>
</tr>
<tr>
<td>$\ldots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Minimize: $\frac{1}{2} w^T w + C_P \sum_{i \in P} \xi_i + C_{LP} \sum_{j \in LP} \xi_j + C_{LN} \sum_{k \in LN} \xi_k$

subject to: $y_i(w^T \tilde{x}_i + b) \geq 1 - \xi_i, \ i = 1, 2, \ldots, n$
Data Set

• Sources (Chinese: Sina, English: Yahoo!)
• 22 news articles (10 Chinese, 12 English)
• 950 news sentences (516 in Chinese, 434 in English)
• 6,219 comments (4,069 in Chinese, 2,150 in English)

Table 1: Statistics on datasets

<table>
<thead>
<tr>
<th>Source</th>
<th>#Sen/Com</th>
<th>Words</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sina</td>
<td>Sen</td>
<td>516</td>
<td>8,932</td>
</tr>
<tr>
<td></td>
<td>Com</td>
<td>4,069</td>
<td>112,853</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>Sen</td>
<td>434</td>
<td>5,767</td>
</tr>
<tr>
<td></td>
<td>Com</td>
<td>2,150</td>
<td>39,917</td>
</tr>
</tbody>
</table>
Annotation

• Manually Annotation
  – 7 annotators (publish task online)
  – Confidence: 5 out of 7 agree
  – Results: 7,520 (cn) + 2,327 (en) links

• Annotated Data Observation
Baseline Methods & Metric

• Methods
  – unsupervised
    • **VSM**: tf-idf + cosine similarity
    • **DCT**: topic directly
  – supervised
    • **BSVM**: classifier on sentence
    • **T-SVM**: classifier on topic
  – Ours(T-PU): unsupervised classifier on topic

• Metric

\[\text{Precision} = \frac{|\bigcup_{i=1}^{N} \{c_i | r_i \cap \tilde{r}_i \neq \emptyset\}|}{|C|}\]

where \( r_i \) and \( \tilde{r}_i \) stands for the annotated alignments and the alignments that found by our method
Results

• Overall

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sina</td>
<td>75.3%</td>
<td>56.7%</td>
<td>64.7%</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>74.9%</td>
<td>63.4%</td>
<td>68.7%</td>
</tr>
</tbody>
</table>

• Comparison
  - best among unsupervised methods (VSM +7.9%)
  - BSVM (+25.9%), significant improvement
  - T-SVM, comparable results (-2.1% in Sina and -2.9% in Yahoo!)
Results

• What leads to failed alignment
  – comment chain (a series of comments issued by two or more users while discussion)
  – topic drift

• Example:
Conclusion

• 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

• Experimental results show the effectiveness of the proposed approach
Future Work

- Alignment over similar web documents
- Whether the social relationships influence the alignment
- Topic drift in the social content
Thanks!