What Users Care about: A Framework for Social Content Alignment

Lei Hou[1], Juanzi Li[2], Xiaoli Li[2], Jiangfeng Qu[1], Xiaofei Guo[1], Ou Hui[1], Jie Tang[1]

[1]Knowledge Engineering Group, Dept. of Computer Science and Technology, Tsinghua University, China

[2]Institute for Infocomm Research, Singapore

IJCAI-13

Tsinghua University

Institute for Infocomm Research A*STAR

Motivation and Problem Definition

Motivation

The rapid development of Web and social media often floods users with huge volume of information

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

Comment number for top news in Yahoo! and Sina are 5684.6 and 9205.4 [Nov, 2012].

To understand the web document and related social content

We Want Know

- What topics the document and social media talk about
- Which part in the news does the social content focuses
- What others discuss over the part that I'm interested in

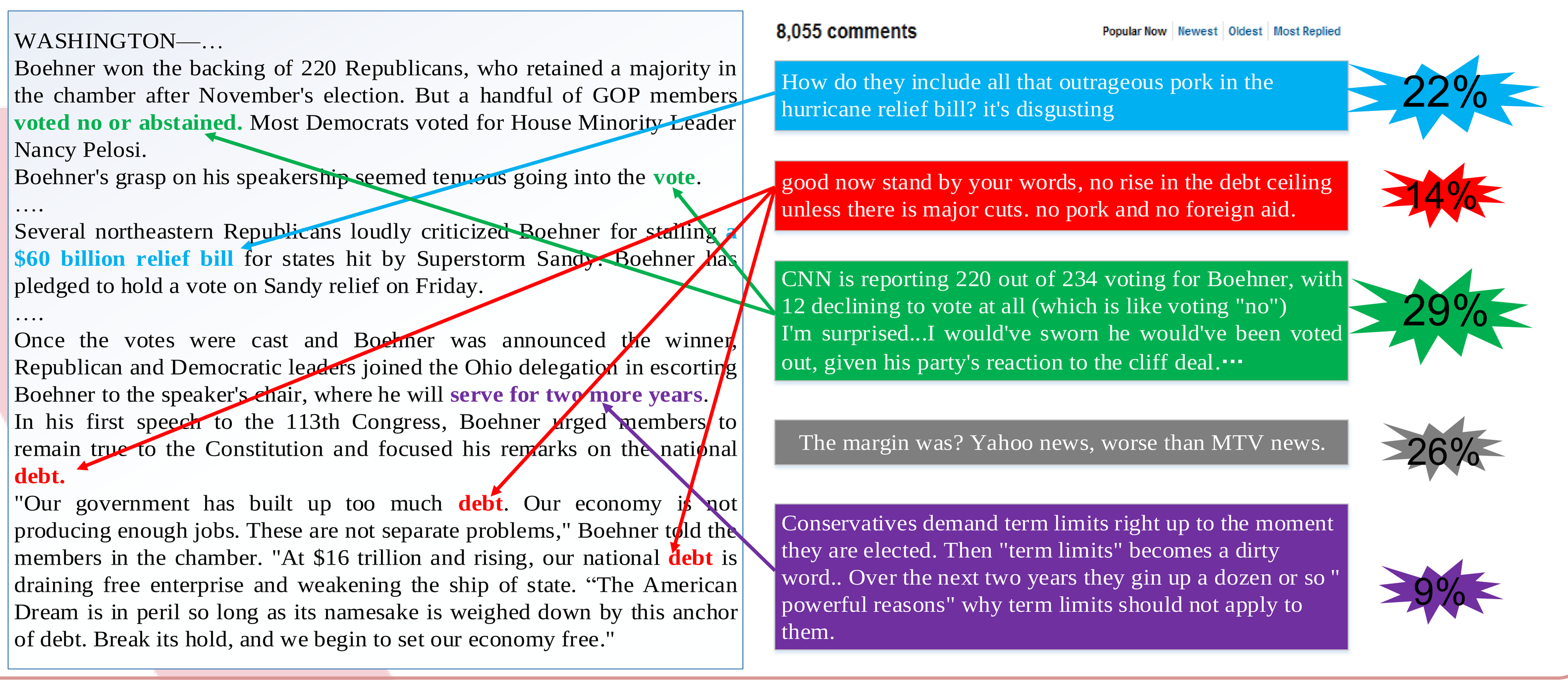
Problem Definition

d: web document; S: sentence set; C:comment set; T: topic set

d is consisting of S and associated with C, and all of them talks about T

Social Content Alignment is to generate a set of matching pairs <social content, topic>, namely $\{(c_i, t_j) \mid where\ c_i \in C, t_j \in T \cup \emptyset\}$, which means social content c_i discusses the specific topic t_j and $\emptyset$ means there is no such topic in the document.

Example



Approach

Framework

PHASE 1

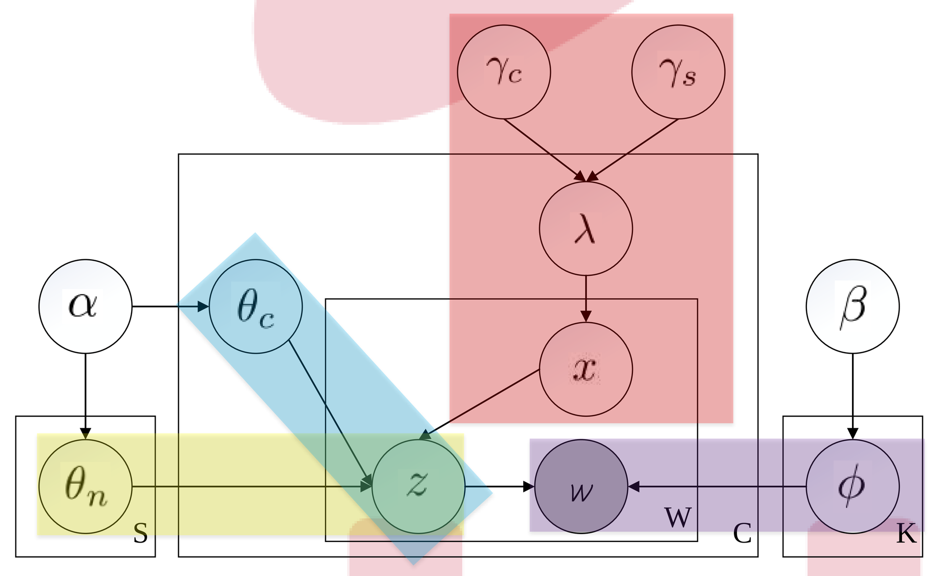
Document Comment Topic Model (DCT Model)

- Different vocabulary
- Sparse feature
- Dependency

PHASE 2

Learning from Positive and Unlabeled Data (PU Learning)

- Unbalanced volume
- Lack of labeled data



Graphical representation

DCT Model

Algorithm 1: Generative process for DCT model

Input: the priors α, β, γ_c, γ_n; S and C
Output: estimated parameters θ_n, θ_c, λ and ϕ
Initialize a standard LDA model over S;
foreach *document* $d \in C$ **do**
  foreach *word* $w_{di} \in d$ **do**
    Toss a coin x_{di} according to $bernoulli(x_{di}) \sim beta(\gamma_n, \gamma_c)$ where $beta(.)$ is a Beta distribution, and γ_c and γ_n are two parameters;
    if $x_{di} = 0$ **then**
      Draw a topic $z_{di} \sim multi(\theta_c)$ from a comment-specific topic mixture
    else
      Draw a topic $z_{di} \sim multi(\theta_n)$ from a document-related topic mixture
    end
    Draw a word $w_{di} \sim multi(\phi_{z_{di}})$ from z_{di}-specific word distribution
  end
end

Generative process

Top words for topic *launch cost*

Aid	Korea	Comment only
Stomach	Money	News only
America	Launch	Both
Food	America	
Korea	Food	

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

PU Learning

Goal: build a classifier to identify more accurate comments for a given topic

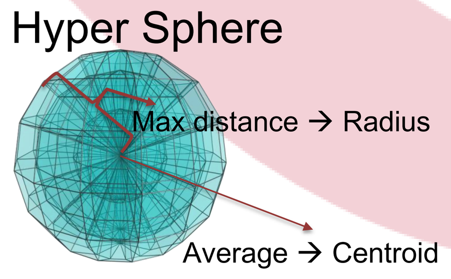
Assumption: the topic sentences in news can be used as positive examples

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 Positive examples

	f_1	f_2	…	f_K
P_1	0.043	0.019	…	0.024
P_2	0.052	0.037	…	0.017
…				
$P_{\lvert P \rvert}$	0.054	0.033	…	0.015

Hyper Sphere



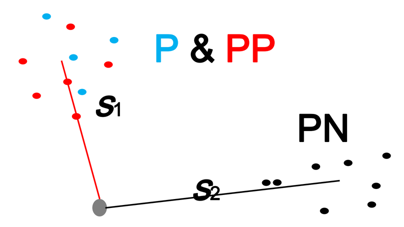
Classify

Unlabeled Data → Potential Positive / Potential Negative

2 Three example sets

Positive Examples (P)
Potential Positive (PP)
Potential Negative (PN)

Ricchio Classifier



Reclassify

Unlabeled Data → Likely Positive (LP) / Likely Negative (LN)

with Confidence $L = \max(s_1, s_2)/(s_1 + s_2)$

3 Training Examples With Different Confidences

	L	f_1	f_2	…	f_K
P_1	1	0.043	0.019	…	0.024
P_2	1	0.052	0.037	…	0.017
…					
LP_1	0.7	0.054	0.033	…	0.015
…					
LN_1	0.83	0.003	0.061	…	0.055
…					

Build the final classifier with Weighted Support Vector Machine, whose objective function is

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

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

Experiment

Dataset

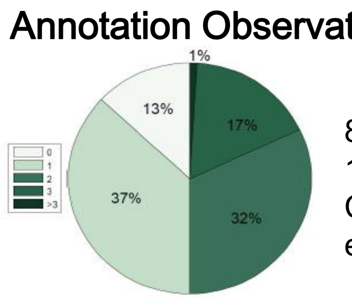
Basic Information[Total (cn + en)]:

- 22 (10 + 12) news
- 950 (516 + 434) sentences
- 6,219 (4,069 + 2,150) comments
- 7 annotators
- Confidence: 5 out 7 agree
- 9,847 (7,520+2,327) links

Statistics

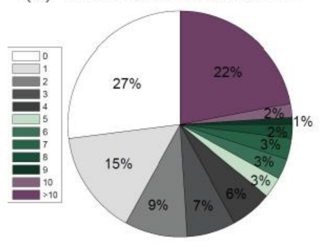
Source		Number of Sen/Com	Words	Vocabulary
Sina	Sen	516	8,932	2,772
	Com	4,069	112,853	13,891
Yahoo!	Sen	434	5,767	2,679
	Com	2,150	39,917	9,972

Annotation Observation



(a) Comment Distribution

87% ←→ one or multiple news sentences
13% ←→ no sentences
Conclusion: it is reasonable to make use of comments to enhance topic detection in DCT model.



(b) Sentence Distribution

22% ←→ more than 10 comments
27% ←→ no comments
Conclusion: automatic alignment is necessary; there are some sentences that simply provide some background of the news.

Result

Methods

- Unsupervised
  - **VSM:** TF-IDF + Cosine Similarity
  - **DCT:** DCT Model directly
- Supervised
  - **BSVM:** classifier on sentences
  - **T-SVM:** classifier on topics extracted by DCT
- Ours (T-PU): unsupervised, classifier on topics

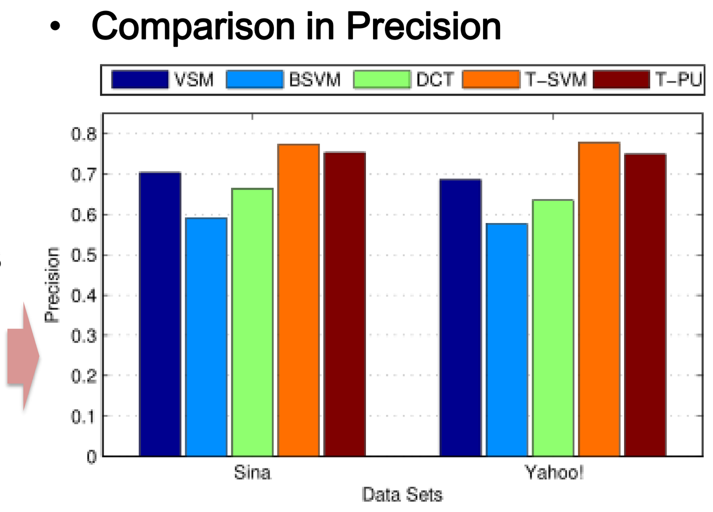
Result

- **Overall**

	Precision	Recall	F1-Measure
Sina	75.3%	56.7%	64.7%
Yahoo!	74.9%	63.4%	68.7%

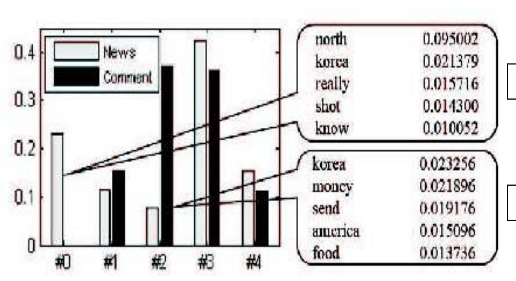
- Best among three unsupervised methods
- With supervised methods
  - BSVM: significant improvement(> 25%)
  - T-SVM: comparable result (-2.1% in Sina and -2.9% in Yahoo!)

- **Comparison in Precision**



Failed Alignment

- **Comment chain:** a series of comments issued by two or more users while discussion, many annotators assign same links for them
- **Topic drift:** Topics may changes



background topic & no comments

cost → food aid

Conclusion and Future Work

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
- 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@keg.cs.tsinghua.edu.cn

Aug. 6th, 2013 @BICC, Beijing, China