What Users Care about: A Framework for Social Content Alignment

Lei Hou[†], **Juanzi Li**[†], **Xiaoli Li**[‡], **Jiangfeng Qu**[†], **Xiaofei Guo**[†], **Ou Hui**[†], **Jie Tang**[†]
[†] Dept. of Computer Science and Technology, Tsinghua University, China 100084
[‡] Data Analytics Department, Institute for Infocomm Research, Singapore 138632
{houlei,ljz,qujf,guoxiaofei,xiou,tangjie}@keg.cs.tsinghua.edu.cn, xlli@i2r.a-star.edu.sg

Abstract

With the rapid proliferation of social media, more and more people freely express their opinions (or comments) on news, products, and movies through online services such as forums, discussion groups, and microblogs. Those comments may be concerned with different aspects (topics) of the target Web document (e.g., a news page). It would be interesting to align the social comments to the corresponding subtopics contained in the Web document. In this paper, we propose a novel framework that is able to automatically detect the subtopics from a given Web document, and also align the associated social comments with the detected subtopics. This provides a new view of the Web *standard* document and its associated user generated content through topics, which facilitates the readers to quickly focus on those *hot* topics or grasp topics that they are interested in. Extensive experiments show that our proposed framework significantly outperforms the existing state-of-the-art methods in social content alignment.

1 Introduction

With the rapid development of social media, more and more people express their opinions and comments on daily news, products, movies and various Web documents through different social media platforms. For example, in many news portals, users can leave their comments on various aspects/topics of a news article. On Twitter, users post the URL of a Web document together with their personal views, followed by streams of posts generated by friends and other people. Figure 1 illustrates a news article about *Boehner*[1] in Yahoo! News, as well as corresponding comments dedicated to a specific topic in the article, such as *vote*, *relief bill*, *tenure of office*, *national debt*, etc.

A popular document often attracts a bulk of social discussions. Our preliminary statistics show that the average numbers of comments for top news in Yahoo![2] and Sina[3] are 5684.6 and 9205.4 respectively (on Nov, 2012). What kinds of topics did the users discuss? How to align the user generated comments to the corresponding topics contained in the Web document? The objective of this research is to design a principled framework to automatically align the social content (user generated comments) to the topics contained in the Web documents, briefly referred to as topic level social content alignment. This can facilitate users to quickly focus on those topics that they are interested in and explore the user generated content at topic level.

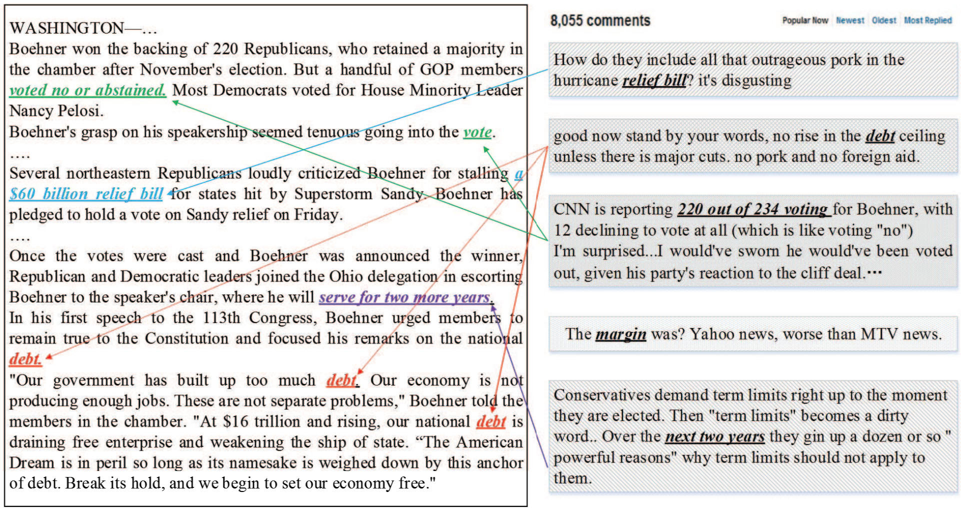
We propose a two-phase framework to address the social content alignment issue. In the first phase, we extract topics from both Web documents and social content. One straightforward idea is making use of probabilistic topic model on documents and social content independently, and then aligning the extracted topics on both sides, but it is difficult to guarantee the consistency between the two different topic sets. An alternative method is that we can model the Web documents and social content together, but there is a risk that topic bias may occur because of the extremely unbalanced volumes between them. Moreover, both methods ignore the dependency between social content and the Web document. We observe that the user generated comments are usually based on the content of Web document. In this paper, we propose a novel document-comment topic model which can effectively exploit the dependency between Web document and social content by using two correlated generative processes. Particularly, we first employ standard LDA to model sentences in Web documents, and for comments modeling, we use a Bernoulli distribution to determine whether it is generated from a document-related topic or a comment-specific topic.

The second phase aligns the topics detected in the first phase with the comments, using our proposed positive unlabeled learning technique. Intuitively, two kinds of machine learning methods [Sil *et al.*, 2011b], namely unsupervised methods and supervised methods, can be employed for social content alignment. The unsupervised methods define and extract features (mostly uses terms) from both Web documents and social content, and assign weights to them (TF-IDF, ESA,

[1] http://news.yahoo.com/blogs/ticket/john-boehner-elected-speaker-house-190301689–politics.html

[2] http://news.yahoo.com/most-popular/

[3] http://news.sina.com.cn/hotnews/



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."

8,055 comments

How do they include all that outrageous pork in the hurricane ***relief bill***? it's disgusting

good now stand by your words, no rise in the ***debt*** ceiling unless there is major cuts. no pork and no foreign aid.

CNN is reporting ***220 out of 234 voting*** for Boehner, with 12 declining to vote at all (which is like voting "no")
I'm surprised...I would've sworn he would've been voted out, given his party's reaction to the cliff deal.⋯

The ***margin*** was? Yahoo news, worse than MTV news.

Conservatives demand term limits right up to the moment they are elected. Then "term limits" becomes a dirty word.. Over the ***next two years*** they gin up a dozen or so " powerful reasons" why term limits should not apply to them.

Figure 1: An alignment example from Yahoo! News.

etc.). Then based on the weights and features, we can use the method to calculate similarities between a document and a comment. Finally, for each topic extracted in the first phase, we select comments with the highest similarities. However, social content is usually quite short, informal, not well structured and it could be written using quite different vocabulary compared with Web documents, unsupervised methods thus cannot effectively match the topic and comments. Supervised learning methods, on the other hand, need users to provide training data, i.e. which sentences in Web documents or comments belong to which topics (treat each topic as a class). Then a classification model can be built to classify all the comments into one of the topic classes. However, in many real-life applications, it is difficult to have sufficient training examples to build accurate classifiers. As such, supervised learning methods will not work as well. In this paper, we propose a novel positive unlabeled learning method (PU learning) which can effectively leverage rich unlabeled data, i.e. those unlabeled comments for building an accurate classifier. Particularly, in our PU learning setting, given a topic, those sentences associated with the topic will be treated as positive data P, all the other sentences and comments are unlabeled data U. Our proposed algorithm is able to partition U into potential positive and negative data and build more accurate classification model. Note our PU learning model is built automatically and thus effectively facilitates the social content alignment process.

The main contributions in this paper can be summarized as follows:

- We formally define the social content alignment problem and present a novel two-phase framework to address it.
- We propose an innovative topic model which can exploit Web document, social content and their dependency for accurately extracting latent topics and unifying both of them in the topic space.
- We propose a novel PU learning algorithm to build classifiers by exploiting the positive topic data identified by our topic model and a large amount of unlabeled data, which addresses the limited labeled topic data issue for effective classification.
- Experimental results show that our proposed framework can effectively address the challenging issues in the social content alignment problem, significantly outperforming the existing state-of-the-art methods.

The rest of the paper is organized as follows. We define the social content alignment problem and present our proposed algorithm in Section 2. Experimental results from extensive experiments are reported in Section 3. Finally, Section 4 concludes the paper.

2 Problem and Techniques

In this section, we first formalize the social content alignment problem, and then demonstrate our proposed techniques.

2.1 Problem Definition

While social content can be expressed through various ways, in this paper we focus on the most commonly used textual information, namely those social content dedicated to news, blogs and other Web documents posted through social media applications.

Preliminary Let d denote a Web document, consisting of a set of sentences $S = \{s_1, s_2, \ldots, s_M\}$, and associated with a social content set $C = \{c_1, c_2, \ldots, c_N\}$. Both the document and its associated social content cover several topics $T = \{t_1, t_2, \ldots, t_K\}$. Each *s*/*c* corresponds to a specific topic t, and contains a vector $\mathbf{w}_{s/c}$ of $N_{s/c}$ words, where each word $w_{s/ci}$ is chosen from a vocabulary of size V.

Definition 1: Social Content Alignment (SCA) Given a Web document with the sentence set S and the corresponding social content set C, the goal of social content alignment is to generate a set of matching pairs <social content, topic>, namely $\{(c_i, t_j) | where\ c_i \in C, t_j \in T \cup \emptyset\}$ which means social content c_i discusses the specific topic t_j.

As shown in Figure 1, the left is a news article entitled *John Boehner re-elected as speaker of the House* in Yahoo! News, which discusses topics such as *vote*, *relief bill*, *tenure of office* and *national debt*. The right part lists a few examples of comments posted by users, where the arrows link them to the representative sentences of the topics in the document.

There are a number of work related to current research. For example, [Lu and Zhai, 2008] studied how to automatically integrate opinions by a well-written expert with lots of opinions scattering in various sources such as blogs, spaces and forums. A semi-supervised model was proposed to deal with it. [Yang *et al.*, 2011] studied to leverage social information for Web document summarization and proposed a dual wing factor graph for summarizing news incorporated Twitter. The objectives of the two related work mentioned above are different from ours since they target at data integration and document summarization by exploiting additional social content. Recently [Sil *et al.*, 2011a; 2011b] proposed to allow users to read news along with relevant comments and presented a supervised learning method for linking comments to news segments. As mentioned before, supervised learning methods typically need time-consuming effort to hand-label the training set. In addition, the training set labeled for one particular document and corresponding social content can only be used once.

In this paper, we propose to design a generic positive unlabeled learning method to automatically align the document with social content. Furthermore, our method is topic based which is more flexible than the segment based ones because all the sentences within a segment must be together in the document, while the sentences within a topic could distribute across the whole document. More specifically, we present a two-phase framework, a multi-source probabilistic topic model and a PU learning method, some of which are novel with respect to this task. The rest of this section demonstrates the details.

2.2 Document-Comment Topic Model

To tackle the sparse and non-uniform feature problem, one idea is enriching features with the aid of large scale data collections like Wikipedia [Phan *et al.*, 2008], but it may introduce noise and often fails on newly-generated content. Considering such a scenario: when a user reads Web documents, he may have opinions on some inside topics, and he subsequently posts social content on the interested internal topics. Although he may use his own words, with the number of comments on this topic grows, the generality of them will be reflected. Therefore, extracting topic features is another direction to solve the problem.

For modeling documents from different sources, [Blei and Jordan, 2003] modeled pictures and their annotations, [Blei and McAuliffe, 2007] developed supervised topic models, where each document is paired with a response, to infer latent topics predictive of the response. Based on the previous works, [Wang *et al.*, 2009] took the category information into consideration for picture modeling and classification. [Tang *et al.*, 2009] proposed qLDA to extract an informative summary from a document collection for a given query, and [Tang *et al.*, 2012] developed CTL to learn and differentiate collaboration topics from other topics. [Hong *et al.*, 2011] extended standard topic models by allowing each text stream to have both local topics and shared topics.

Through observation (see Section 3.1), we find the social content heavily leverages the Web documents, and the latent topics build a bridge between them. The Web documents behave like a kind of background knowledge to guide the generation of social content. But all the existing works pay little attention on the dependencies, so here we develop a document comment topic model to model Web documents and social content simultaneously.

The basic idea is to use two correlated generative processes to model Web documents and comments. The first process is to model sentences in Web documents using standard LDA, and the second process is to model comments. For each word in comments, we use a Bernoulli distribution to determine whether it is generated from a document-related topic or a comment-specific topic. Figure 2 shows the graphical structure of the DCT model (For simplicity, we omit the modeling part for Web documents and focus on the modeling of comments).

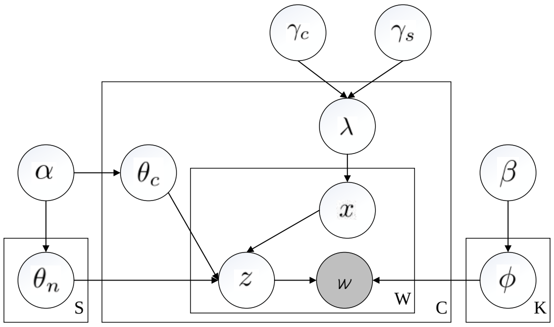


Figure 2: Graphical representation of DCT model.

Let us briefly introduce notations. θ_s, θ_c are topic models from documents and comments; x is a binary variable indicating whether the current word inherits the topic from document-related ($x = 1$) or by comment-specific topic ($x = 0$); α, β are the Dirichlet hyper parameters; λ is a parameter for sampling the binary variable x; γ_c, γ_s are *beta* parameters to generate λ.

Formally, the generative process is described in Algorithm 1: we extract topics from document according to the distribution $p(\theta_s|\alpha)$, while for word w_{ci} in comment c, a coin x is tossed according to $p(x|c) \sim beta(\gamma_c, \gamma_s)$ to decide whether w_{ci} is sampled from a document-related topic or comment-specific topic.

To estimate the parameters for this model, we take the widely-used Gibbs sampling technique. First we sample the coin x according to the posterior probability:

$$p(x_i = 0|\mathbf{x}_{\neg i}, \mathbf{z}, \cdot) = \frac{n^{\neg ci}_{cx_0} + \gamma_c}{n^{\neg ci}_{cx_0} + n^{\neg ci}_{cx_1} + \gamma_c + \gamma_s} \times \frac{n^{\neg ci}_{z_{ci}} + \alpha}{\sum_z (n^{\neg ci}_z + \alpha)} \quad (1)$$

where n_{cx_0} is the number of times that $x = 0$ has been sampled in c; $n_{z_{ci}}$ is the number of times that topic z has been

Algorithm 1: Generative process for DCT model.

```
Input: the priors $\alpha$, $\beta$, $\gamma_c$, $\gamma_s$; $S$ and $C$
Output: estimated parameters $\theta_s$, $\theta_c$, $\lambda$ and $\phi$
Initialize a standard LDA model over $S$;
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
end
```

sampled from c and ”$\neg$” indicates excluding that instance from counting. $p(x_i = 1|\cdot)$ can be analogously defined as Equation 1.

Then the posterior probability of topic z is defined as:

$$p(z_{ci}|x_{ci} = 1, \mathbf{x}, \mathbf{z}_{\neg ci}, \cdot) = \frac{n^{\neg ci}_{z_{ci}w_{ci}} + m_{z_{ci}w_{ci}} + \beta}{\sum_w (n^{\neg ci}_{z_{ci}w} + m_{z_{ci}w} + \beta)} \times \frac{n^{\neg ci}_{cz_{ci}} + m_{cz_{ci}} + \alpha}{\sum_z (n^{\neg ci}_{cz} + m_{cz} + \alpha)} \quad (2)$$

where $n^{\neg ci}_{z_{ci}w_{ci}}$, $m_{z_{ci}w_{ci}}$ denote the number of times that word w_{ci} has been generated by topic z_{ci} in comment and Web document respectively, and $n^{\neg ci}_{cz_{ci}}$ and $m_{cz_{ci}}$ are the number of times that topic z_{ci} have been sampled from comment-specific or document-related topic distribution.

During the parameter estimation, the algorithm keeps track of a $(|S| + |C|) \times K$ (sentence+comment by topic), a $|C| \times 2$ (comment by coin), a $2 \times |K|$ (coin by topic) count matrixes, and a $K \times W$ (topic by word) count matrix. Given these matrixes, we can estimate the probabilities θ_s, θ_c, λ and ϕ.

After that, each sentence is associated with a probability distribution over topics, and we can obtain representative sentences for each topic by selecting those with highest probability over the given topic.

2.3 Learning from Positive and Unlabeled Data

We are now ready to present our work on how to design a PU learning method to align the social content with the topics identified in the Section 2.2. Given a particular topic t_j, all the sentences in S belonging to t_j, will be treated as the positive set P. All the other sentences in $S - P$ will be treated as unlabeled set U_1, and all the social content will be treated as unlabeled set U_2, as shown in Equation 3. We then build a PU learning model to classify all the social content, into either positive or negative class. Particularly, those social content that are classified into positive class will be aligned to the topic t_j.

$$\begin{aligned} &P = \{s_i|\theta_{sij} = \max_{1\le k\le K} \theta_{sik}\},\ i = 1, 2, \ldots, N \\ &U_1 = S - P, \quad U_2 = C \end{aligned} \quad (3)$$

Learning from positive and unlabeled data (PU learning) was first proposed in [Liu *et al.*, 2002]. The core idea can be characterized as the following two steps, namely: (1) identify a set of reliable negative examples from the unlabeled set U; and then (2) build a classifier using EM or SVM iteratively. The difference among the existing algorithms, [Liu *et al.*, 2003], [Lee and Liu, 2003], [Li and Liu, 2003], [Li *et al.*, 2007], [Li *et al.*, 2009], [Li *et al.*, 2010] and [Nguyen *et al.*, 2011] lies in the specific algorithms used in these two steps.

The first step of existing PU learning methods, typically uses U as initial negative set to build a classifier with P to exact reliable negatives. However, since U contains hidden (false) positive data, the extracted negatives are not reliable. Different from the existing work, in this paper, we present a three-step PU learning method. We aim to first partition the U into potential positive set PP and potential negative set PN. In particular, we construct a hyper-sphere by calculating the centroid $\vec{o}$ of all the positive examples, and the average distance r (viewed as radius) between $\vec{o}$ and all the positive examples. Those examples which fall into the hyper-sphere will be treated as potential positive examples, while the others outside hyper-sphere are treated as potential negative examples, as shown in Equation 4. Note that the set PN will be much pure to serve as a negative set than the original unlabeled set U since we have taken out the potential positives PP in U. At the same time, the positive set P is relatively small, PP can be used to enhance the limited positive set P.

$$\begin{aligned} &\vec{o} = \frac{\sum_{d\in P} \vec{d}}{|P|} \\ &r = \frac{\sum_{d\in P} dist(\vec{d}, \vec{o})}{|P|} \\ &PP = \{b|b \in U_1 \cup U_2\ and\ dist(\vec{b}, \vec{o}) \le r\} \\ &PN = U_1 + U_2 - PP \end{aligned} \quad (4)$$

The second step of our method is to build a Rocchio classification model where we use $P \cup PP$ as positive training examples, PN as negative training examples. We construct positive and negative prototype vectors $\vec{p}$ and $\vec{n}$ respectively for the positive and negative classes:

$$\begin{aligned} \vec{p} &= \mu \frac{1}{|P \cup PP|} \sum_{d\in P\cup PP} \frac{\vec{d}}{\|\vec{d}\|} - \nu \frac{1}{|PN|} \sum_{d\in PN} \frac{\vec{d}}{\|\vec{d}\|} \\ \vec{n} &= \mu \frac{1}{|PN|} \sum_{d\in PN} \frac{\vec{d}}{\|\vec{d}\|} - \nu \frac{1}{|P \cup PP|} \sum_{d\in P\cup PP} \frac{\vec{d}}{\|\vec{d}\|} \end{aligned} \quad (5)$$

where $\mu = 16$, $\nu = 4$ as recommended in [Buckley *et al.*, 1994].

For each unlabeled example $u \in U_1 \cup U_2$, we calculate its cosine similarities with $\vec{p}$ and $\vec{n}$: if $sim(u, \vec{p}) > sim(u, \vec{n})$, then u will be added to likely positive set LP; otherwise, it is added to likely negative set LN. We also compute the confidence score l as:

$$l = \frac{\max(cosine(\vec{u},\vec{p}), cosine(\vec{u},\vec{n}))}{cosine(\vec{u},\vec{p}) + cosine(\vec{u},\vec{n})} \tag{6}$$

Note that we give the confidence score 1 to all the examples in the original positive set P.

Finally, we build the final classifier using Weighted Support Vector Machine (WSVM), whose optimizing goal is:

$$\begin{aligned} Minimize: &\frac{1}{2}\mathbf{w}^T\mathbf{w} + C_P \sum_{i \in P} \xi_i + \\ &C_{LP} \sum_{j \in LP} \xi_j + C_{LN} \sum_{k \in LN} \xi_k \\ subject\ to: &y_i(\mathbf{w}^T \vec{x}_i + b) \geq 1 - \xi_i,\ i = 1, 2, ..., n \end{aligned}$$

where C_P, C_{LP} and C_{LN} represent the penalty factors of misclassification for three types of training examples, namely, original positive set P, likely positive set LP and likely negative set LN. We use the confidence score directly for each example in P, LP and LN as C_P , C_{LP} and C_{LN} since we are more confident with positive set P than the likely positive set LP and likely negative set LN. Correspondingly, we give a larger penalty if examples from P are classified as negative class than if examples from LP are classified as negative or examples from LN are classified as positive.

After that, we use the our weighted SVM model to classify all the social content in C and those social content are classified as positive class will be aligned to the given topic t_j. Apparently, we can accomplish the alignment task by repeating it on all extracted topics.

3 Empirical Evaluation

In this section, we evaluate the proposed alignment method using our manually created news data. We will first introduce the data set in Section 3.1 and then present the experimental results in Section 3.2.

3.1 Data Preparation

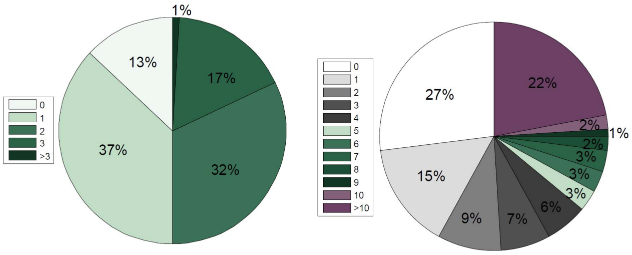
To the best of our knowledge, there is no existing benchmark dataset which can be directly used for our experiments. As such, we crawled news articles and corresponding social content from two popular news websites, namely Sina (China) and Yahoo!. We have chosen top 10 Chinese news from Sina and 12 English news from Yahoo! from Dec. 1st to Dec. 11th 2012 which have the most social content generated. Then we invited 7 annotators to build gold-standard link sets between the sentences in the news and social content/comments where we only include those links that majority of people agree (5 out of 7).

For both datasets, we perform preprocessing, such as remove stop words, and filter low-frequency words (frequency<= 3). Some statistics of the datasets after preprocessing are summarized in Table 1. Then we investigate the annotation results. Figure 3 shows the distribution of comments(or sentences) with respect to the number of sentences (or comments) they are linked to. We can observe that:

Table 1: Statistics on datasets

Source		#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

- 87% comments are linked to one or multiple news sentences, while the remaining 13% of comments are irrelevant to any sentences, indicating that it is reasonable to make use of comments to enhance topic detection in our document comment topic model.
- 22% news sentences can attract more than 10 comments, signifying it is important to automatically mine the relevant comments from large amount of social content so that we can quickly read through them to effectively understand what other people care about. We also notice there are 27% of sentences without any related comments. We found these sentences simply provide some background of the news and thus lead to ”no comments” scenario.



(a) Comment Distribution (b) Sentence Distribution

Figure 3: Annotation Analysis

3.2 Experiment

Baseline Methods. We use T-PU to denote the alignment technique that employs DCT model for topic extraction and weighted-SVM for the final classification. To test the effectiveness of T-PU, we compare it with four baseline methods:

VSM: a simplest similarity based method by using ***TF-IDF*** document representation method and cosine similarity.

BSVM: a classification based method. We build binary classifiers on sentences using the labeled data with libsvm[4] and test with five-folder cross-validation. The numbers of positive examples range from 0 to 78 (3.93 in average) while negative examples can be more than 100.

DCT: a straightforward method which classifies the comments directly by using the distribution obtained by DCT model.

T-SVM: a supervised method which needs users provide manually-labeled training examples. The only difference from BSVM is that classifiers here are built on the topics extracted by DCT model instead of individual sentences.

[4] http://www.csie.ntu.edu.tw/˜cjlin/libsvm/

Table 2: Overall results on two datasets

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

Metrics. The performance of text classification is typically evaluated based on *precision* and *recall*, but in our scenario *precision* is more important than *recall* because users prefer to read few accurate comments rather than many noisy comments.

For a sentence-comment set pair (S, C), let $r_i \subseteq S$ be the set of aligned sentences (labeled by different annotators) for comment c_i. If $|r_i| > 1$, then c_i has multiple related sentences while $|r_i| = 0$ indicates c_i has no related sentences.

Let $t_i \subseteq S$ denote the result topic for c_i found by our method, and t_i is associated with several sentences $\tilde{r}_i$. We consider it to be correct if $r_i \cap \tilde{r}_i \neq \emptyset$.

So the precision can be defined as:

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

To make the comparison fair, we use the DCT results for those methods without topic information.

Result and Analysis. Table 2 gives the overall performance of our method on the two datasets, and the comparison with other four techniques in terms of *Precision* is shown in Figure 4. We can see that:

- Our proposed method T-PU outperforms the baseline methods VSM and DCT on both datasets. This is because our T-PU method utilizes topic level features and word-level features in two steps while the other two use only one of them.
- Compared with BSVM, our method has significant improvement because it is difficult to build an accurate classifier with very few positive examples and many negative examples (contains noise inside). And our method can split the latter into potential positive and negatives examples, which can enhance the limited positive set as well as purify the negative set.
- T-PU is very close (-2.1% in Sina and -2.9% in Yahoo!) to T-SVM which performs best among all the methods. Note that T-SVM is a supervised method and its performance depends much on the quality and quantity of labeled data while our T-PU can achieve comparable results without using any labeled examples, so our method is more appropriate for the social content alignment task.

Furthermore, we investigate those comments that our method failed to find an appropriate topic for, and discover that the main reason is the comment chain and the topic drift caused by it. A drift example is shown in Figure 5. It is a report about *rocket launch* in *North Korea*. We can see that there is no comment on Topic 0 (background topic), and Topic 2 should have been talking about the cost but through the chain, it turns into *food aid*. Similar problem occurs in other news, which contributes most of the failed comments.

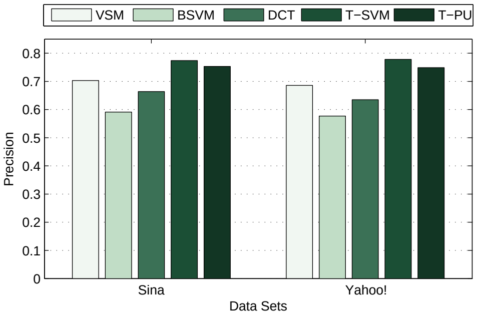
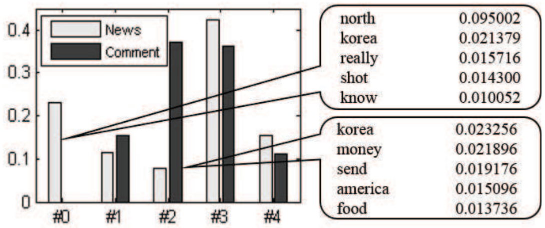


Figure 4: Results comparison (in *Precision*).



north	0.095002
korea	0.021379
really	0.015716
shot	0.014300
know	0.010052

korea	0.023256
money	0.021896
send	0.019176
america	0.015096
food	0.013736

Figure 5: Topic distribution in news and comments

Hyperparameters. Following [Rosen-Zvi *et al.*, 2004]'s guide, we take a fixed value for the hyper parameters α and β where $\alpha = 50/K, \beta = 0.1$. As for the Beta parameters γ_c, γ_s, we tried two methods, calculating the common words ratio and referring the comment-sentence distribution in Section 3.1, and the results show DCT model is not very sensitive to the prior. And the number of topics is set to be 5 for all news and it works well in most cases.

4 Conclusions

In this paper, we address the most fundamental problem in social content analysis, that is, finding the corresponding article sentences for each piece of social content. We propose a novel two-phase framework to accomplish this task automatically. Specifically, we present a document comment model to extract topics from both the document and comments and introduce a positive and unlabeled learning method to build an accurate classifier which does not need users provide labeled examples but can achieve comparable good results compared with the supervised learning methods.

Social content alignment is a challenging yet interesting problem. Thus there are many potential future directions of this work. For example, we can study the alignment over similar Web documents (i.e. news reports on same event) and explore the comment patterns they may share. We can also investigate whether users' relationships can help the alignment, and study the topic drift in the comment stream.

Acknowledgments. Thanks anonymous reviewers for their valuable suggestions that help us improve the quality of the paper. The work is supported by NSFC (No. 61035004, 61222212, 61073073), NSFC-ANR(No. 61261130588), 863 High Technology Program (2011AA01A207), FP7-288342, and THU-NUS NExT Co-Lab.

References

[Blei and Jordan, 2003] David M. Blei and Michael I. Jordan. Modeling annotated data. In *SIGIR'03*, pages 127–134, 2003.

[Blei and McAuliffe, 2007] David M. Blei and Jon D. McAuliffe. Supervised topic models. In *NIPS'07*, pages 121–128, 2007.

[Buckley *et al.*, 1994] Chris Buckley, Gerard Salton, and James Allan. The effect of adding relevance information in a relevance feedback environment. In *SIGIR'94*, pages 292–300, 1994.

[Hong *et al.*, 2011] Liangjie Hong, Byron Dom, Siva Gurumurthy, and Kostas Tsioutsiouliklis. A time-dependent topic model for multiple text streams. In *KDD'11*, pages 832–840, 2011.

[Lee and Liu, 2003] Wee Sun Lee and Bing Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *ICML'03*, pages 448–455, 2003.

[Li and Liu, 2003] Xiaoli Li and Bing Liu. Learning to classify texts using positive and unlabeled data. In *IJCAI'03*, pages 587–594, 2003.

[Li *et al.*, 2007] Xiaoli Li, Bing Liu, and See-Kiong Ng. Learning to identify unexpected instances in the test set. In *IJCAI'07*, pages 2802–2807, 2007.

[Li *et al.*, 2009] Xiaoli Li, Philip S. Yu, Bing Liu, and See-Kiong Ng. Positive unlabeled learning for data stream classification. In *SDM'09*, pages 257–268, 2009.

[Li *et al.*, 2010] Xiaoli Li, Bing Liu, and See-Kiong Ng. Negative training data can be harmful to text classification. In *EMNLP'10*, pages 218–228, 2010.

[Liu *et al.*, 2002] Bing Liu, Wee Sun Lee, Philip S. Yu, and Xiaoli Li. Partially supervised classification of text documents. In *ICML'02*, pages 387–394, 2002.

[Liu *et al.*, 2003] Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S. Yu. Building text classifiers using positive and unlabeled examples. In *ICDM'03*, pages 179–188, 2003.

[Lu and Zhai, 2008] Yue Lu and Chengxiang Zhai. Opinion integration through semi-supervised topic modeling. In *WWW'08*, pages 121–130, 2008.

[Nguyen *et al.*, 2011] Minh Nhut Nguyen, Xiaoli Li, and See-Kiong Ng. Positive unlabeled leaning for time series classification. In *IJCAI'11*, pages 1421–1426, 2011.

[Phan *et al.*, 2008] Xuan Hieu Phan, Minh Le Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *WWW'08*, pages 91–100, 2008.

[Rosen-Zvi *et al.*, 2004] Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *UAI'04*, pages 487–494, 2004.

[Sil *et al.*, 2011a] Dyut Kumar Sil, Srinivasan H. Sengamedu, and Chiranjib Bhattacharyya. Readalong: reading articles and comments together. In *WWW (Companion Volume)'11*, pages 125–126, 2011.

[Sil *et al.*, 2011b] Dyut Kumar Sil, Srinivasan H. Sengamedu, and Chiranjib Bhattacharyya. Supervised matching of comments with news article segments. In *CIKM'11*, pages 2125–2128, 2011.

[Tang *et al.*, 2009] Jie Tang, Limin Yao, and Dewei Chen. Multi-topic based query-oriented summarization. In *SDM'09*, pages 1147–1158, 2009.

[Tang *et al.*, 2012] Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain collaboration recommendation. In *KDD'12*, pages 1285–1293, 2012.

[Wang *et al.*, 2009] Chong Wang, David M. Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *CVPR'09*, pages 1903–1910, 2009.

[Yang *et al.*, 2011] Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li. Social context summarization. In *SIGIR'11*, pages 255–264, 2011.