# Learning to Group Web Text Incorporating Prior Information

Yu Cheng, Kunpeng Zhang, Yusheng Xie, Ankit Agrawal, Wei-keng Liao, Alok Choudhary
*Department of Electric Engineering & Computer Science*
*Northwestern University*
*Evanston, IL 60208*
*Email: {ych133,kzh980,yxi389,ankitag,wkliao,choudhar}@eecs.northwestern.edu*

***Abstract*—Clustering similar items for web text has become increasingly important in many Web and Information Retrieval applications. For several kinds of web text data, it is much easier to obtain some external information other than textual features which can be utilized to improve the performance of clustering analysis. This external information, called prior information, indicates label sign and pairwise constraints on sample points. We propose a unifying framework that can incorporate prior information of cluster membership for web text cluster analysis and develop a novel semi-supervised clustering model. The proposed framework offers several advantages over existing semi-supervised approaches. First, most previous work handles labeled data by converting it to pairwise constraints and thus leads to much more computation. The proposed approach can handle pairwise constraints together with labeled data simultaneously so that the computation is greatly reduced. Second, the framework allows us to obtain these prior information automatically or only with little human effort, thus, making it possible to boost the clustering learning performance relatively easily. We evaluated the proposed method on the real-world problems of automatically grouping online news feeds and web blog messages. Experimental results indicate the proposed framework incorporating prior information can indeed lead to statistically significant clustering improvements over the performance of approaches access only to textual features.**

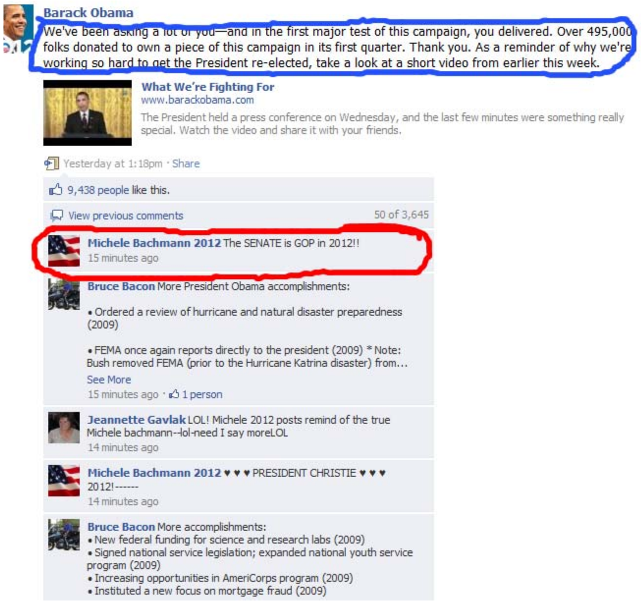
***Keywords*-web text; semi-supervised clustering; prior information; pairwise constraints**

## I. Introduction

As online communication becomes increasingly popular, texts become available in a variety of genres like blog & news feeds, forum, book & movie summaries, product descriptions, customer reviews, and so on. For example, popular online social media sites like Facebook, MySpace, Google Buzz or Twitter allow users to post short messages to their homepages. The message is called a status update and in particular, status updates from Twitter are more commonly called as tweets. Tweets are often related to different event-specific topics of interest like politics, sports or personal opinions. Twitter is an extremely popular blog service with more than 20 million unique monthly visitors and more than one million tweets each hour. Another case is the dynamic content of news and blog posts being delivered in the form of RSS or Atom. Many popular feed sources usually send a large number of items per day. For instance, Google News (http://news.google.com) publishes 500+ news items per day [1]. Visitors to Twitter.com and subscribers to the popular news items often face the problem of information overload due to the large number of messages published periodically. One way to deal with the information overload problem is to cluster similar items (i.e. by topic). Groupings similar messages or news feeds and filtering out duplicate or very similar items can make the information more manageable for a user. Systems that automatically clusters the documents belonging to the same topic and presents part of these items for a topic would enables the display of manageable amount of information for the users.

In real application domains, it is often the case that the experimenter possesses some prior information (about the domain or the data set) that could be useful in clustering the data. In this work, we propose to improve clustering performance by utilizing the priori information of cluster membership extracted from the web text. The prior information indicates label signs and more importantly, pairwise constraints on sample points, i.e., whether they are in the same class or not. The main reason we adopt such a framework is, for several kinds of web data, this prior information, at times, can be obtained automatically or only with little human effort and may be collected from different sources. Fig.1 illustrates an example of analysis on external information from Facebook.com. In Facebook.com, each user can have a personal page (like Barack Obama). Each personal page may publish its interesting status (we call it ”post” circled blue) and other users can reply to these posts as ”comments” (circled red). Any two comments corresponding to one post are in the same cluster (i.e. topic ”political”). Facebook provides APIs (http://developers.facebook.com/) for this kind of downloading data. The comments correspond to a specific post via an unique post ID. By analyzing this, one can get the pair information directly. Similar situation can also be concluded in Twitter from ”tweets” and ”replies”. In Fig.2 from Google News, first, the news feeds and the associated news (circled red) from other web sources (i.e. Yahoo Sports and Boston Herald) are linked, as the same class. Second, when clicking the associated link (circled yellow), it will show ”all related” news in one page. One can only use a coarse web analysis script to automatically detect the pair relation and label information of examples. Finally, the external information can also come from human

978-0-7695-4409-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDMW.2011.111



Barack Obama
We've been asking a lot of you—and in the first major test of this campaign, you delivered. Over 495,000 folks donated to own a piece of this campaign in its first quarter. Thank you. As a reminder of why we're working so hard to get the President re-elected, take a look at a short video from earlier this week.

What We're Fighting For
www.barackobama.com
The President held a press conference on Wednesday, and the last few minutes were something really special. Watch the video and share it with your friends.

Yesterday at 1:18pm · Share

9,438 people like this.

View previous comments 50 of 3,645

Michele Bachmann 2012 The SENATE is GOP in 2012!!
15 minutes ago

Bruce Bacon More President Obama accomplishments:

• Ordered a review of hurricane and natural disaster preparedness (2009)

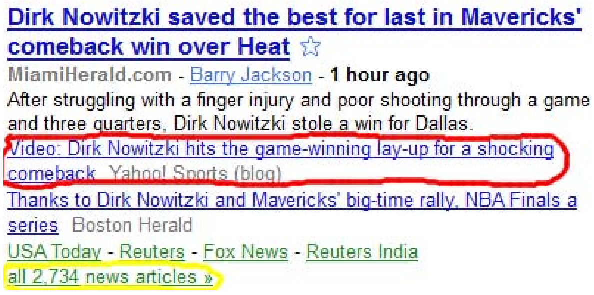
• FEMA once again reports directly to the president (2009) * Note: Bush removed FEMA (prior to the Hurricane Katrina disaster) from...
See More
15 minutes ago · 1 person

Jeannette Gavlak LOL! Michele 2012 posts remind of the true Michele bachmann--lol-need I say moreLOL
14 minutes ago

Michele Bachmann 2012 ♥ ♥ ♥ PRESIDENT CHRISTIE ♥ ♥ ♥ 2012!------
14 minutes ago

Bruce Bacon More accomplishments:
• New federal funding for science and research labs (2009)
• Signed national service legislation; expanded national youth service program (2009)
• Increasing opportunities in AmeriCorps program (2009)
• Instituted a new focus on mortgage fraud (2009)

Figure 1. An illustration of the form of data with prior information from Facebook.com



Dirk Nowitzki saved the best for last in Mavericks' comeback win over Heat ☆
MiamiHerald.com - Barry Jackson - 1 hour ago
After struggling with a finger injury and poor shooting through a game and three quarters, Dirk Nowitzki stole a win for Dallas.
Video: Dirk Nowitzki hits the game-winning lay-up for a shocking comeback Yahoo! Sports (blog)
Thanks to Dirk Nowitzki and Mavericks' big-time rally, NBA Finals a series Boston Herald
USA Today - Reuters - Fox News - Reuters India
all 2,734 news articles »

Figure 2. An illustration of the form of data with prior information from Google News

feedback, providing the exact label or pair information.

The key idea in this paper is to explore external information apart from textual features (hard and soft constraints) to boost the clustering performance. We work within a semi-supervised learning framework and generate a unifying model incorporating all information available, no matter whether it is about unlabeled data, data with constraints, or labeled data. Unlabeled and constrained data are integrated in a manner similar to the integration of labeled and unlabeled data so that we can treat pairwise constraints together with labeled data at the same time. The main contributions of our work are:
1. We analyze several web application cases (from Google News, Facebook.com and Twitter) and derive several types of prior information, especially pair-wise constraints information, to be incorporated into our framework to boost the learning performance.
2. A semi-supervised learning framework is developed which can handle pairwise constraints together with labeled data simultaneously so that the computation is greatly reduced.
3. The proposed method is generalized to automatically employ the pairwise constraints or other useful information of the web text to overcome the lack of large amount of labeled data.

The rest of the paper is organized as follows. In the next section, we first describe in detail how to integrate prior information into the model and present our modified algorithm, and then presents the general framework aimed at building text and web classifier. Experimental results are presented in Sections 3. Section 4 presents related work. Finally, Section 5 summarizes our contributions and discusses the future work.

## II. Proposed Methodology

In this section, we present a semi supervised clustering method using both the textual and the prior information extracted. The classifier is working with data and constraints so we call it Constraints Mixture Learning Model (CML). Our CML model encodes the prior information into to a model-based object function, and the clustering task is carried out by solving a convex optimization problem. We start with a natural formulation of the semi-supervised model-based classification that enables a smooth transition from the semi-supervised learning model to our proposed model. Finally we show the model framework and how to collect the constraints information.

### A. Problem Formulation

Since our technique builds upon traditional Model-based clustering/classification methods, we start with a brief review of that. The model-based clustering or classification assumes that the data were generated by a model and tries to recover the original model from the data. The model that we recover from the data then defines clusters and an assignment of data points to clusters. Usually we denote the model parameters by $\Theta$. In supervised model-based clustering, one has a labeled data set $\chi^l$ and the corresponding labels, denoted by $y_i$ for $x_i \in \chi^l$. Then best parameters $\Theta$ can be determined by minimizing the the negative data log-likelihood function $\Phi_l = -\Sigma_{x \in \chi^l} \log p(x_i, y_i|\Theta)$. Similarly to that, in unsupervised framework, the data density is $p(x|\Theta)$ for the data set $\chi^u$. Many machine learning algorithms attempt to choose the parameters $\Theta$ that minimize $\Phi_u = -\Sigma_{x \in \chi^u} \log p(x|\Theta)$. Usually, the Expectation-Maximization (EM) [2] algorithm is employed to obtain a local optimizer for this optimization problem.

For the semi-supervised learning task, the data set $\chi$ consists of both labeled data $\chi^l$ and unlabeled data $\chi^u$ and $\chi = \chi^l \cup \chi^u$. In order to use the information from both the labeled data and unlabeled data, the maximum likelihood criterion is to select the parameters $\Theta$ by minimizing the objective function:

$$\Theta = \arg\min_{\Theta}(\alpha\Phi_u + \beta\Phi_l) \tag{1}$$

here, $\alpha, \beta \in [0, 1]$ supervises the effect of the labeled/unlabelled data on the parameter estimation.

In our work, we represented the problem like this: for a given data set $\chi$, we can decompose it the into three parts: unlabeled data $\chi^u$, labeled data $\chi^l$ and data with pairwise assignment constraints $\chi^c$. Such that $\chi = \chi^l \cup \chi^u \cup \chi^c$. The unlabeled data, constrained data and the labeled data can be integrated in a manner similar to (1). We define the following joint object function, which is a convex combination of $\Phi_u$, $\Phi_l$ and $\Phi_c$:

$$\Theta = \arg\min_{\Theta}(\alpha\Phi_u + \beta\Phi_l + \gamma\Phi_c) \tag{2}$$

where $\alpha, \beta, \gamma \in [0, 1]$. Using the norm defined in (1), the optimal $\Theta$ can still be found by EM, while the result of the minimization is a parameter estimate that takes all the available prior information into account. Since we already have definition for $\chi^u$ and $\chi^l$, the goal of this work is to the integration of pairwise must-link and must-not-link constraints into the process of model fitting for data $\chi^c$. We want to achieve this in a way similar to the integration of partially labeled data and unlabeled data as described in semi-supervised learning framework in (1). Following the methods from [3], we provide a Maximum Entropy (ME) prior model defined on the hidden variables that captures the dependencies, and propose an efficient implementation of the model by means of a Mean-Field approximation. At first, we discuss constraint specification.

### B. The Constraints Mixture Learning Model

*1) integrating Constraints in the Inference:* The focus of the present work is the integration of pairwise must-link and must-not-link constraints into the process of model fitting. Since no exact labels are prescribed for the data in $\chi^c$, we consider a latent variable $y_i$ as the label for $x_i$. We introduce a binary indicator variable $a_{i,j}$ as the positive constraints (must-link), such that it is 1 if $x_i$ and $x_j$ should be in the same group, and 0 otherwise. The negative constraints (must-not-link) is defined similarly: $b_{i,j}$= 1 if $x_i$ and $x_j$ should not be linked, and 0 otherwise. Using the idea from [4, 5, 6], we penalize a constraint violation whenever the latent variables in a constraint are different (the same) while they are supposed to be the same (different). Hence, the penalty for violation of positive and negative constraints becomes $a_{i,j}\mathbf{1}(y_i \neq y_j)$, and $b_{i,j}\mathbf{1}(y_i = y_j)$, respectively, where $\mathbf{1}$ denotes the indicator function. We turn this pairwise information into a prior on the label assignment for the data in $\chi^c$ by applying the maximum entropy principle: find the prior distribution $p(\mathbf{y}) = p(y_1, ..., y_n)$ for the cluster labels of the data points $x_i \in \chi^c$ such that the entropy $H(p)$ is maximized while the expected number of constraint violations,

$$\sum_{y_1=1}^{n} ... \sum_{y_n=1}^{n} p(\mathbf{y})(a_{i,j}\mathbf{1}(y_i \neq y_j) + b_{i,j}\mathbf{1}(y_i = y_j)) \tag{3}$$

Note, we can convert the problem to the maximum entropy distribution as a Lagrangian functional with Lagrange parameters $\lambda_+$ and $\lambda_-$. The solution to this inference problem is the so-called Gibbs distribution, and in our case, it is

$$\frac{1}{Z}\exp(-\lambda^+ a_{i,j}\mathbf{1}(y_i \neq y_j) - \lambda^- b_{i,j}\mathbf{1}(y_i = y_j)) \tag{4}$$

where $Z$ is the normalization constant. In order to use more sophisticated optimization techniques such as EM, the problem of estimating marginalized posteriors can no longer be circumvented. In order to keep the optimization tractable, we approximate the posteriors in the E-step by the mean filed approximation.

*2) Mean-Field Approximation:* Assume that the data given in $\chi^c$ are independent. By Bayes rule, we have

$$p(\mathbf{y}|\chi^c) = \frac{1}{Z}\prod_i \exp(-h_i(y_i))p(\mathbf{y}) \tag{5}$$

where $h_i(y_i) = -\log p(x_i|y_i)$ for Gaussian class conditional densities. In the mean field approximation, one tries to find a factorial approximation, the mean field approximation, $q(\mathbf{y}) = \prod_i q_i(y_i)$ of the posterior $p(\mathbf{y}|\chi^c)$ such that the Kullback-Leibler divergence between the approximate and true posterior distributions is minimized, i.e.

$$\min_q \sum_y^n \frac{q(\mathbf{y})}{p(\mathbf{y}|\chi^c)} \tag{6}$$

such that $\sum_v q_i(v) = 1$, for all $i$. Because the approximation is factorial, the computation of the marginalized posterior probabilities becomes feasible, a prerequisite to optimize the model efficiently. Note that the above KL divergence can be decomposed as

$$-H_q - E_q p(\mathbf{y}|\chi^c) \tag{7}$$

where $H(q)$ denotes the entropy of the mean field approximation and $E_q$ denotes the expectation w.r.t. $q$. We seek to minimize the expression in (6) by looking for stationary points for the $q_i(v)$. Set $\gamma_{i,j} = \lambda_+ a_{i,j} - \lambda_- b_{i,j}$ and $\Delta_{\nu,\mu} = 1 - \delta_{\nu,\mu}$, where $\delta_{\nu,\mu}$ is the Kronecker delta function. Using this convention, we can summarize the exponents in (4) by $\gamma_{i,j}\Delta_{\nu,\mu}$ if $y_i = \nu v$ and $y_i = \mu$ . We want to emphasize that this approximation is only used for constrained data. Taking the derivative of (6) w.r.t $q_i(v)$ and setting it to zero leads to

$$q_i(v) = \frac{1}{Z_i}\exp(-h_i(v) - \sum_{j\neq i}\sum_{\mu} q_i(\mu)\gamma_{i,j}\Delta_{\nu,\mu}) \qquad (8)$$

where

$$Z_i = \sum_{v}\exp(-h_i(v) - \sum_{j\neq i}\sum_{\mu} q_i(\mu)\gamma_{i,j}\Delta_{\nu,\mu}). \qquad (9)$$

Since $\Delta_{\nu,\mu} = 1$ only if $\mu \neq \nu$, we can further simplify the expression for $q_i(v)$ to:

$$q_i(v) = \sum_{v}\exp(-h_i(v) - \sum_{j\neq i}\sum_{\mu} q_i(\mu)\gamma_{i,j}\Delta_{\nu,\mu}) \qquad (10)$$

Eventually, we have arrived at a factorial approximation of the marginal posterior probabilities. For the constrained data, these update equations can be used in the E-step for posterior probability estimation. So far, we have assumed that every data point in $\chi^c$ participates in a constraint and we minimize the data negative log-likelihood. The constraints part $\Phi_u$ in (2) is obtained via and the same convex optimization can be used,

*3) Optimization and Parameter Estimation:* Similarly, (labeled + constrained) and (constrained + unlabeled) data can be combined into a single objective function. In particular, the optimal $\Theta$ can still be found by EM, while allowing the inclusion of partially labeled as well as constrained data. The result of the minimization is a para-meter estimate that takes all the available prior information into account. For Gaussian class-conditional densities, we arrive at a similar formula as we did in the semi-supervised case. The $\mu_v$ are estimated in each EM iteration.

Clearly, the choice of $\alpha, \beta, \gamma$ is critical in this context since it might significantly determine the resulting model, in particular in the case of a model mismatch. Our framework assumes that the model parameters are actually modified and affected by the prior information, so a common strategy is to choose the $\gamma$ by $\gamma = |\chi^u|/|\chi^l \cup \chi^u \cup \chi^c|$.

### C. The Model Based Framework

In this section, we present the framework that aims at building web and text classifier with the Constraints Mixture Learning Model (CML). The framework is depicted in Figure 3 and consists of the following sub-problems: (1) analyzing and gathering pairwise constrains and label information, (2) integrating all the extracted priot information in CML, (3) building the classifier and doing the clustering analysis. For all the sub-problems, analyzing and gathering pairwise constrains and label information for the training data set is the most important. In general, this step can be implemented automatically, as we discussed in Section 1; or manually: querying the user and getting feedback. After integrating constraints and labels in the inference for training data, the final step of building the classifier is similar to any other training process to build a text classifier.

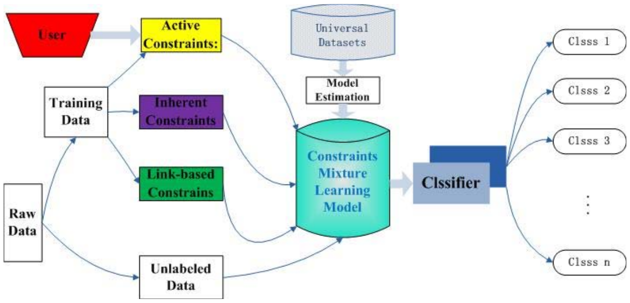


Figure 3. The general framework of learning to clustering web text with the CML model

As mentioned before, there are several types of pairwise constraints that can be extracted from the web text itself. In this paper, we pay particular attention to three types of pairwise constraints. The first two can be extracted automatically. The third one is based on the user's feedback in an active learning scheme.

**Link-based Constrains:** The web text or documents usually linked to other text with related topic. For some specific items, such as news, by analyzing their link status (both in and out), we can gather the constraints automatically. For example, in news, it is possible that two linked news items have the same topic.

**Inherent Constraints:** This type of constraints is are obtained by knowing the content of the text. As we mentioned before, in Facebook.com, two comments share the same post is from the same topic. In tweets, the sentence starting with "@" may be a reply sentence for topic-related tweet.

**Active Constraints:** In analogy to active learning paradigms, this type of constraints is obtained from users' feedback. Typically, the system gives users the most ambiguous pair of examples and users provide the constraint label as feedback.

## III. Experiments

In this section, we first present the details of our experimental procedure, including descriptions of the data sets and the evaluation methods. Then we analyze the performance results. Finally, we present a case study that validates our intuitions as to how the prior information boosts the classification performance.

### A. Data Collection and Preprocessing

The goals of the evaluation include (1) comparing the Constraints Mixture Learning Model (CML) performance against some of the existing classification methods; (2) empirically testing the performance of the overall framework on real world applications comparing it with the baseline approaches. K-means and two representative semi-supervised data clustering and classification methods: Semi-supervised K-means [7] and Transductive SVM [8] . We refer to these representation methods as Baseline. Evaluation metrics include accuracy and F-score, i.e. the harmonic mean of precision and recall. Note that an F-score of unity amounts to perfect classification. Two different data sets have been collected in our experiments. The first labeled data set was generated from the home page of Google News, where some of the news are clustered by some topics like ”business”, ”science” and ”health”. In our experience, we observed that the clustering strategy that Google uses is quite accurate. Thus it provides an easily available labeled data set for clustering. We collected 500 news documents from 4 different topics to create the News-500 . The second data set is fbs-500. A large corpus of comments from 20 public pages in Facebook.com have been collected and labeled manually based on the topic. The topics of these newsgroups are very diversified, ranging from music, automobiles, to religions, politics, etc. We selected 500 items from this data with 7 different categories into fbs-500. We pre-processed each text data set by tokenizing the text into bag-of-words. Then, we applied down casting, stop-words removal, and stemming to the words. Based on the processed words, a feature vector was constructed for each text sample.

### *B. Results and Analysis*

Here we present the performance results for the different methods we considered. The aim of this experiment is to test the performance of the CML model comparing with the other two semi-supervised clustering approaches: SK-means and Transductive SVM. In the experiment, 1%, 2.5%, 5%, 7.5% and 10% of the data, which are equal to about 5̃50 documents in most cases, were selected as training samples. The label for each news is known when collected from Google News, we can use it directly. Table I and Table II demonstrated that the results of the different semi-supervised learning algorithms. We can see that the proposed CML algorithm is very competitive and outperforms on most problem instances comparing with the baseline methods.

Our second experiment is on a comments classification task, where the goal is to grouping comments from Facebook.com into different clusters based on its topic. The experiment is presented to prove the effect of our overall framework. The fbs-500 data set with different percentage of the data in training sample are tested where the constraints are obtained using our proposed framework automatically. K-means is baseline method. The quality gap is particularly large in this case comparing with unsupervised algorithm in Table III. We also ran the experiments by selecting training samples from 1% to 30% and the results is showed in Fig.4. It should be pointed out that the accuracy and F-score did not increase significantly when the rate exceeded 20%.

Table III
THE ACCURACY AND F-SCORE OF K-MEANS AND CML ON FBS-500

| Evaluation Method | Accuracy | F-score |
|---|---|---|
| K-Means | 0.457 | 0.403 |
| CML(5% training) | 0.616 | 0.501 |
| CML(7.5% training) | 0.649 | 0.578 |
| CML(10% training) | 0.691 | 0.627 |

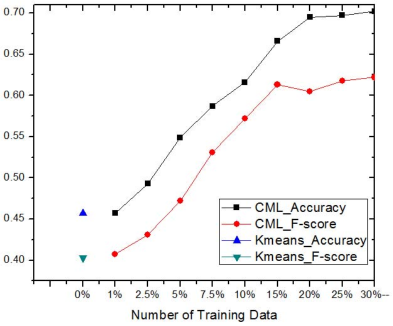


Figure 4. The accuracy and F-score of K-means and CML

All the constraints information in CML are generated automatically and it provides more information to overcome the paucity of textual information in short comment sentence. The proposed framework can easily be extended to the other applications in real world, like customer review analysis and topic-based blog classification.

### *C. Case Study*

We engage in a case study to show how our incorporated prior information can improve classification analysis. Table IV shows a selection of comments that only our algorithm clustered correctly from Facebook.com. We see that the text of these comments is often seeming hard for traditional to classify. Fig.5 shows such an example. The comment ”run down to big lots and get one for 30 bucks to”, coming from Amazon wall, is talking about the products which Amazon posted but the unsupervised method classified it to ”sports”. Instead, the proposed approach can first automatically obtain this comment as well as the other comments which share the same label with it (like ”sweet deal, thank you!”), then integrate these information into the model based framework and finally get the right class. In Fig.6, comments 3 is about the advertising of Pepsi. With the help of prior information from other sentences, we can classify it correctly.

## IV. RELATED WORK

In this section, we first review problems associated with web text & short text clustering methods, and then review related works on semi-supervised clustering with constraints.

With the popularity with web text, some interesting work has appeared to understand the web text. However, there are two main problems in clustering/classification of the web

Table I
THE ACCURACY OF NEWS-500 DATA SET WITH DIFFERENT PERCENTAGE OF TRAINING SAMPLES

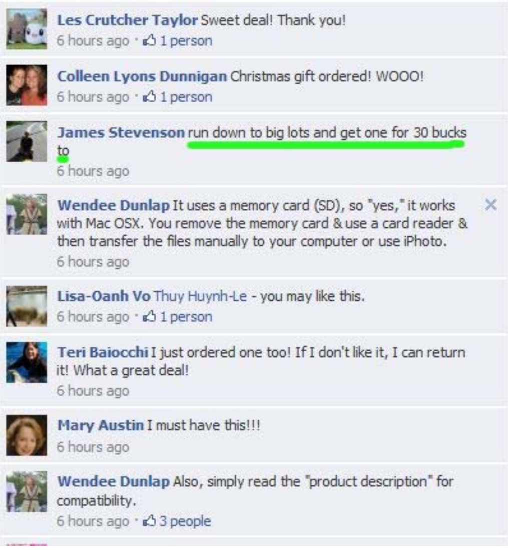
| Method | 1% | 2.5% | 5% | 7.5% | 10% |
|---|---|---|---|---|---|
| SK-means | 0.479 | 0.513 | 0.545 | 0.580 | 0.629 |
| Transductive SVM | 0.442 | 0.481 | 0.501 | 0.521 | 0.576 |
| CML Model | 0.459 | 0.492 | 0.549 | 0.632 | 0.682 |

Table II
THE F-SCORE OF NEWS-500 DATA SET WITH DIFFERENT PERCENTAGE OF TRAINING SAMPLES

| Method | 1% | 2.5% | 5% | 7.5% | 10% |
|---|---|---|---|---|---|
| SK-means | 0.486 | 0.502 | 0.526 | 0.540 | 0.575 |
| Transductive SVM | 0.403 | 0.431 | 0.458 | 0.485 | 0.524 |
| CML Model | 0.466 | 0.494 | 0.516 | 0.579 | 0.625 |

Table IV
SAMPLE OF COMMENTS CLUSTERED CORRECTLY ONLY IN CML FRAMEWORK

| Comments ID | True | CML | K-means | Comments |
|---|---|---|---|---|
| 1 | products | sports | products | run down to big lots and get one for 30 bucks to |
| 2 | products | movie | products | follow me @mad man bitch |
| 3 | advertising | politics | advertising | Watch for CCCS of Josephine County-Grants Pass,oregon to be here for waiting in August. Spread the word vote...PLEASE!! |
| 4 | movie | friend | movie | hate me I got in Pottermore ! :P |
| 5 | software | food | software | I like this Donut application:-) |
| 6 | computer | health | computer | Deviant art has the best themes, I recommend using the token icon set along with placebo. |



**Les Crutcher Taylor** Sweet deal! Thank you!
6 hours ago · 👍 1 person

**Colleen Lyons Dunnigan** Christmas gift ordered! WOOO!
6 hours ago · 👍 1 person

**James Stevenson** run down to big lots and get one for 30 bucks to
6 hours ago

**Wendee Dunlap** It uses a memory card (SD), so "yes," it works with Mac OSX. You remove the memory card & use a card reader & then transfer the files manually to your computer or use iPhoto.
6 hours ago

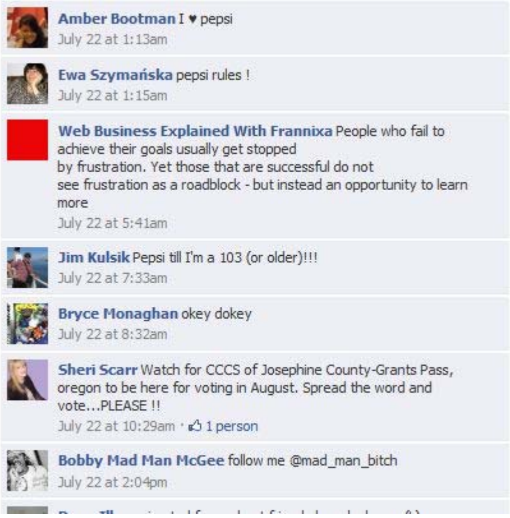
**Lisa-Oanh Vo** Thuy Huynh-Le - you may like this.
6 hours ago · 👍 1 person

**Teri Baiocchi** I just ordered one too! If I don't like it, I can return it! What a great deal!
6 hours ago

**Mary Austin** I must have this!!!
6 hours ago

**Wendee Dunlap** Also, simply read the "product description" for compatibility.
6 hours ago · 👍 3 people

Figure 5. The case study from Amazon wall (Facebook.com)



**Amber Bootman** I ♥ pepsi
July 22 at 1:13am

**Ewa Szymańska** pepsi rules !
July 22 at 1:15am

**Web Business Explained With Frannixa** People who fail to achieve their goals usually get stopped
by frustration. Yet those that are successful do not
see frustration as a roadblock - but instead an opportunity to learn more
July 22 at 5:41am

**Jim Kulsik** Pepsi till I'm a 103 (or older)!!!
July 22 at 7:33am

**Bryce Monaghan** okey dokey
July 22 at 8:32am

**Sheri Scarr** Watch for CCCS of Josephine County-Grants Pass, oregon to be here for voting in August. Spread the word and vote...PLEASE !!
July 22 at 10:29am · 👍 1 person

**Bobby Mad Man McGee** follow me @mad_man_bitch
July 22 at 2:04pm

Figure 6. The case study from Pepsi wall (Facebook.com)

data [9, 10]. First one is its dynamic property and the second one is each short text does not have enough content, or words specifically. Therefore, conventional machine learning and text mining algorithms cannot apply to these text directly. Existing work in the literature tries to address the aforementioned challenges from two directions. The former is to fetch external text to expand the text (e.g., [11]). Another direction is to discover a set of explicit or implicit topics and then connect the short text through these topics. For examples, in [12], the authors exploit the user-defined categories and concepts (e.g., Wikipedia1), while in [13], the authors derive a set of hidden topics through topic model LDA [14] from one large existing Web corpus.

Generally, text clustering algorithms are used in an unsupervised fashion. They are presented with a set of data instances that must be grouped according to some notion of similarity. However, in real application domains, it is often the case that the experimenter possesses some prior knowledge (about the domain or the data set) that could be useful in clustering the data. Based on this situation, recently there has been a growing interest in a hybrid setting, called semi-supervised, where the labels of only a portion of the data set are available for training [16, 17]. There have been research studies that modify traditional K-means clustering to incorporate labeled data [18]. [15, 19] introduced two types of constraints: the ”must-link” and the ”cannot-link” constraints, and their semi-supervised K-means produces data partitions by ensuring none of the user specified constraints are violated. The access to both similar and dissimilar pairs in the training data work with equivalence relations [20, 21, 22] These techniques discriminatively learn a distance metric for cluster-ing or classification when both similar and dissimilar pairs are given. In the absence of dissimilar pairs, Xing et al. [21] suggest treating the data pairs that are not similar as the dissimilar pairs. But, such heuristics have been shown to give lower accuracy in comparison to the techniques that use only similar pairs [20].

## V. Conclusion and Future Work

Learning with insufficient training data to classify or cluster objects has become an interesting topic in recent years. In this paper, the general idea is to explore external information except for textual features to help clustering or classification analysis. These external information, called prior information, sometimes can be obtained easily for web data. We demonstrated that the clustering performance can be significantly improved by incorporating information of labeled data as well as additional pairwise constraints for the web text. The proposed strategy allows us (i) to handle different information: constraint violations, soft constraints and labeled data, at the same time, (ii) to automatically gather the constraint or labeled information to boost the learning performance. Experiments on the real-world data showed that the proposed approach could achieve significantly improved performance, compared to the baseline classifiers.

There are several avenues for future work arising from this work. The proposed framework can be applied to other types of social media data (like customer reviews, blog messages) and incorporate more information in addition to partial labels and constraints. Regarding real world applications, it would be interesting to consider how to derive the pairwise constraints automatically from auxiliary information, and study how these different types of pairwise constraints can improve the performance of a discriminative classifier. Finally, we would like to further investigate the interplay between unlabeled, labeled and constraints information in both the theoretical and practical sense.

## Acknowledgment

This research was supported by the National Science Foundation under grants CCF-0621443, OCI-0724599, CCF-0833131, CNS-0830927, IIS-0905205, OCI-0956311, CCF-0938000, CCF-1043085, CCF-1029166, and OCI-1144061 and by the DOE grants DE-FC02-07ER25808, DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, and DE-SC0005340.

## References

[1] Jian Hu, Lujun Fang, Yang Cao, Hua-Ju n Zeng, Hua Li, Qiang Yang, and Zheng Ch en. Enhancing text clustering by leveraging wikipedia semantics. In Pro-ceedings of SIGIR, p ages 179-186, 2008.

[2] G. McLachlan and K. Basford. Mixture Models: Inference and Application to Clustering. Marcel Dekker, New York, 1988.

[3] E. T. Jaynes. Information theory and statistical mechanics i. Phys. Rev., 106:620-630, 1957.

[4] S. Basu, M. Bilenko, and R. Mooney. A probabilistic framework for semi-supervised clustering. In Proceedings of the 10th ACM SIGKDD, International Conference on Knowl-edge Discovery and Data Mining, pages 59-68, 2004.

[5] M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In Proceedings of the 21st International Conference on Ma-chine Learning, pages 81-88, 2004.

[6] Tilman Lange, Martin H. C. Law, Anil K. Jain, Joachim M. Buhmann. Learning with Constrained and Unlabelled Data. In Proceedings of CVPR (1)'2005. pp.731 738.

[7] K. Wagsta, S. Rogers, and S. Schroedl. Constrained K-means clustering with background knowledge. Proceedings of the 18th Internation Conference on Machine Lea rning, 577-584, 2001

[8] T. Joachims Transductive inference for text classification using support vector machines. Proc. 16th International Conf. on Machine Leaning, pp.200-209, Morgan Kaufmann, San Francisco, CA.

[9] Banerjee, S., Ramanathan, K., and Gupta, A. Clustering short texts using wikipedia. In Proceedings of SIGIR. 2007, 787-788.

[10] Mengen Chen, Xiaoming Jin and Dou Shen. Short Text Classification Improved by Learning Multi-Granularity Topics. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11). Barcelona, Spain. July 16-22, 201.

[11] Mehran Sahami and Timo-th y D. Heilm an. A web-based kernel functio n for measuring the similarity of short text snippets. In Proceedings o f WWW, p ages 377-386, 2006.

[12] Xia Hu, Nan Sun, Chao Zhang, and Tat-Seng Chua. Exploiting in ternal and external semantics forthe clustering of short texts using world knowledge. In Proceeding of CIKM, pages 919-928, 2009.

[13] Xuan Hieu Phan, LeMinh Nguyen, and Susumu Horiguchi. Learning to classify short and sp arse text & web with hidden topics from large-scale data collections. In Proceeding of WWW, pages 91-100, 2008.

[14] David M. Blei, Thomas L. Grifiths, Michael I. Jo rdan, and Jo shua B. Tenenbaum. Hierarchical topic m odels and the nested chinese r estaurant process. In Proceedings of NIPS , pages 17-24, 2003.

[15] M. Law, A. Topchy, and A. Jain. Model-based clustering with probabilistic constraints. In Proc. SIAM International Conference on Data Mining, 2005.

[16] S. Basu, A. Banerjee, and R. J. Mooney. Active semi-supervision for pairwise constrained clustering. In Proceedings of the SIAM International Conference on Data Mining, pages 333-344, 2004.

[17] K. Wagstaff and C. Cardie. Clustering with instance-level constraints. In Proc. the 17th International Conference on Machine Learning, pages 11031110, 2000.

[18] K. Wagstaff, C. Cardie, S. Rogers, and S. Scroedl. Constrained k-means clustering with background knowledge. In Proc. Int. Conf. on Machine Learning (ICML), 2001.

[19] K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidenite programming. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 988-995, 2004.

[20] S. X. Yu and J. Shi. Segmentation given partial grouping constraints. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(2):173-183, 2004.

[21] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning with application to clustering with side information. In Adv. NIPS, 2002.

[22] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In Proc. Int. Conf on Machine Learning (ICML), 2003.