CluChunk: Clustering Large Scale User-generated Content Incorporating Chunklet Information

Yu Cheng, Yusheng Xie, Kunpeng Zhang, Ankit Agrawal, Alok Choudhary
EECS Department, Northwestern University
2145 Sheridan Road
Evanston,IL,USA
{ych133,yxi389,kpz980,ankitag,choudhar}@eecs.northwestern.edu

ABSTRACT

The exponential rise of online content in the form of blogs, microblogs, forums, and multimedia sharing sites has raised an urgent demand for efficient and high-quality text clustering algorithms for fast navigation and browsing of users based on better document organization. For several kinds of these user-generated content, it is much easier to obtain the input in small sets, where the data in each set belongs to the same class but with unknown class labels. Such data is viewed as weakly-labeled data and the inherent chunklet information is very useful for improving clustering performance. In this paper, we propose a system - CluChunk (clustering chunklet data) to cluster unlabeled web data which incorporates chunklet information. We try to transfer the original feature space by a discriminatively learning linear transformation such that simple unsupervised learning techniques (such as K-Means) in the transformed space can achieve good clustering accuracy. Using larger scale data from some web applications (social media and online forums), we demonstrate that the clustering performance can get significantly improved by: 1)incorporating the inherent weakly-labeled information into the clustering framework; 2)enriching the representation of short text with additional features extracted from the chunklet subset. The proposed approach can be applied to other mining tasks with large scale user-generated content, like product review summarizing and blog content clustering/classification task.

Categories and Subject Descriptors

H.1 [**Models and Principles**]: General; H.3.3 [**Information Storage and Retrieval**]: Clustering

General Terms

Algorithms,Experimentation

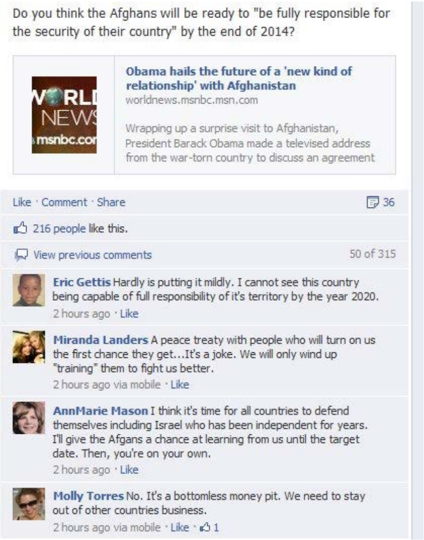
BigMine'12, August 12, 2012 Beijing, China.

Keywords

User-generated Content, Text Clustering, Data Transformation, Chunklet

1. INTRODUCTION

In this era of information explosion, there is a large amount of web content being generated by users every moment in the form of forums, blogs, microblogs, customer reviews, and so on. Twitter is an online microblogs which allows users to publish content (called "tweets") Users in Twitter generate more than 200 millions tweets per day [1]. In Facebook.com, there are about 100 millions comments posted by users in one day [6]. This huge amount of information invariably makes its manual comprehension infeasible for a person, and urges the development of automated methods geared towards helping the user better understand this information. For example, in a online shopping website, automatically clustering the customer reviews into categories and filtering out duplicate or very similar items can make the information more manageable for a user to view. However, web content is very different from traditional documents, and automatically clustering/classifying such data usually faces two main problems. The primary problem is that traditional supervised and semi-supervised approaches for text classification often require labeled data for learning classifiers. When applied to a large amount of web data, creating such labeled data, even with a few documents per category, can be a time-consuming and error-prone process. However it is difficult to get good performances using traditional unsupervised clustering methods such as K-Means and normalized cut. How can we find an efficient and high-quality clustering algorithm for the large scale content data? In this paper, we aim at addressing the problem of grouping user-generated content data by exploring the inherent weakly labeled information, which is very helpful for improving the clustering performance. We find that, for several kinds of web user-generated content, it is much easier to obtain the input in subsets, where the data in each subset comes from the same but unknown class. For example, in Facebook.com, a person can create and manage pages (also called "walls") to represent their organizations or companies. Facebook provides some functions to allow the page managers to create content (called "posts") publicly in their pages to which their audiences can make comments to. These posts and comments can be classified based on their topics, such as political, education, games/sports, health etc. Figure 1 shows a post published by msnbc.com(a news/media company) referring to the election of president Obama, and other users made



Do you think the Afghans will be ready to "be fully responsible for the security of their country" by the end of 2014?

Obama hails the future of a 'new kind of relationship' with Afghanistan
worldnews.msnbc.msn.com

Wrapping up a surprise visit to Afghanistan, President Barack Obama made a televised address from the war-torn country to discuss an agreement

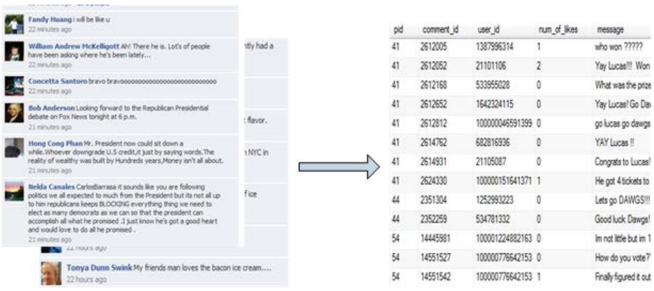
Eric Gettis Hardly is putting it mildly. I cannot see this country being capable of full responsibility of it's territory by the year 2020.
2 hours ago · Like

Miranda Landers A peace treaty with people who will turn on us the first chance they get...It's a joke. We will only wind up "training" them to fight us better.
2 hours ago via mobile · Like

AnnMarie Mason I think it's time for all countries to defend themselves including Israel who has been independent for years. I'll give the Afgans a chance at learning from us until the target date. Then, you're on your own.
2 hours ago · Like

Molly Torres No. It's a bottomless money pit. We need to stay out of other countries business.
2 hours ago via mobile · Like · 1

Figure 1: An illustration of a post and its comments in msnbc.com wall from Facebook.com



Fandy Huang i will be like u
22 minutes ago

William Andrew McKelligott Ah! There he is. Lot's of people have been asking where he's been lately...
22 minutes ago

Concetta Santoro bravo bravoooooooooooooooooooooooooo
22 minutes ago

Bob Anderson Looking forward to the Republican Presidential debate on Fox News tonight at 6 p.m.
21 minutes ago

Hong Cong Phan Mr. President now could sit down a while.Whoever downgrade U.S credit,it just by saying words.The reality of wealthy was built by Hundreds years,Money isn't all about.
21 minutes ago

Nelda Canales CarlosBarrasa it sounds like you are following politics we all expected to much from the President but its not all up to him republicans keeps BLOCKING everything thing we need to elect as many democrats as we can so that the president can accomplish all what he promised .I just know he's got a good heart and would love to do all he promised .
21 minutes ago

Tonya Dunn Swink My friends man loves the bacon ice cream....
22 hours ago

pid	comment_id	user_id	num_of_likes	message
41	2612005	1387996314	1	who won ?????
41	2612052	21101106	2	Yay Lucas!!! Won
41	2612168	533955028	0	What was the prize
41	2612652	1642324115	0	Yay Lucas! Go Da
41	2612812	100000046591399	0	go lucas go dawgs
41	2614762	682816936	0	YAY Lucas !!
41	2614931	21105087	0	Congrats to Lucas!
41	2624330	100000151641371	1	He got 4 tickets to
44	2351304	1252993223	0	Lets go DAWGS!!!
44	2352259	534781332	0	Good luck Dawgs!
54	14445981	100001224882163	0	Im not little but im 1
54	14551527	100000776642153	0	How do you vote?
54	14551542	100000776642153	1	Finally figured it out

Figure 2: An illustration of how the data with chunklet information is extracted from web sites and stored in database

their comments to this post. Nearly all of the comments talk about the election or opinions to Obama and belong to the same topic (i.e. election or political). Such comments be obtained automatically and stored in a database structurally as shown in Figure 2. Facebook provides APIs (http://developers.facebook.com/) for the downloading of posts and comments. When downloading a post we can also get all the comments within that post, with each comment corresponding to a unique post ID. In our database, the comments as well as the corresponding post are stored. Comments related to the same post are collected in one subset with the same (although unknown) label, and there are many small subsets in the whole dataset. This assumption would be true in a majority of cases in the real-world data. Similar paradigm can be found in forums data. In a forum like *Flyertalk* (http://www.flyertalk.com/forum/), there are a lot of user-generated content posted per day with different topics varied from flight, hotel, traveling etc. One user can post a message(usually called "thread") related to a specific topic and others can reply the thread. All the replies made to a specific thread are viewed in a subset and have the same topic label.

This kind of data is actually weakly labeled data and all the subsets are known as *chunklets*. The *chunklet* is defined as a subset of points that are known to belong to the same although unknown class [4]. It is important to notice that such a scenario yields positive pairwise relation, *i.e.*, 'X is similar to Y' if X and Y belong to the same chunklet but never gives negative relation 'X is dissimilar to Z' if X and Y belong to different chunklets as even two items from different chunklets may have the same (unknown) label. There are some previous works about data with chunklet information [4, 19, 17, 10] using Relevant Componet Analysis(RCA). RCA is an effective linear transformed algorithm, which can be used as a preprocessing step for unsupervised clustering or nearest neighbor classification of the data. Through the transformation based on a group of chunklets, RCA can assign larger weights to relevant features and low weights to irrelevant features[17]. In this paper, we address the problem by proposing a novel method for discriminatively learning linear transformations using the chunklet information. The key advantage of our technique is that it results in transformation space that are better for class discrimination than RCA. We focus on discriminatively learning linear transformations in order to improve the subsequent performance of unsupervised learning techniques in the transformed feature space. The second problem, when dealing with web text, it may raise challenges because of its shortness and sparseness, which makes it extremely hard to build a feature space directly for clustering and classification. In this paper, we propose a method for improving the clustering accuracy by enriching the representation of the short texts to be clustered. In our method, the conventional bag of words representation of one text item is augmented with the features extracted from other items within the same chunklet.

The main contributions of this work are as follows:
(1)We propose an algorithm to discriminatively learn a linear transformation using the inherent chunklet information extracted from web content;
(2)We improve the performance of clustering by enriching the representation of short web text using external feature source from the same chunklet;
(3)Building a complete system (Cluchunk) to extract the data with chunklet information from web, structurally store them into database and cluster them based on topic.
The proposed approach demonstrates superior performance two real-world datasets for clustering task. The rest of this paper is organized as follows: Section 2 describes the related work. Section 3 presents the architecture of our proposed framework as well as the method for for enriching the feature space. The details of the proposed transformation for clustering is given in Section 4. Experimental results are presented in Sections 5. Section 6 concludes the paper with directions for future work.

2. RELATED WORK

In this section, we first review some related work on web text and short text classification/clustering methods, and then review some problems associated with learning on the chunklet data.

2.1 Web Text Classification/Clustering

The exponential growth of online text has raised an urgent demand to understand the web text. As mentioned earlier, web text is different from traditional documents in its shortness and do not provide enough word co-occurrence or shared context for a good similarity measure. Two major approaches have been exploited to enrich the representation of short text. One is to derive a set of explicit or implicit topics from existing large corpus and to enrich the representation of short text by using these topics. In [12, 11], the authors use Wikipedia concept and category information to enrich document representation to address semantic information loss caused by bag-of-words representation. Wikipedia is used in [9] to build a concept thesaurus to enhance traditional content similarity measurement. [7] proposed a method for improving the modeling for short text by leveraging topics at multiple granularity. Another direction for addressing the problem of short text is to directly fetch external text to expand the original text. In [3, 14], semantic similarity between words was obtained by leveraging page counts and text snippets returned by the search engine. Similarly, titles and snippets we combined to enrich the original short text, which leads to a significant improvement in similarity measurement in [2, 18, 16]. In this paper, we will follow the second approach by directly enriching the original short text with the external text. The original feature is expanded using other text within the same chunklet. In other words, the new features are a combination of the original features and the features extracted from the chunklet data. This method is efficient and suitable for large scale web data.

2.2 Learning with Chunklet Data

In many unsupervised learning tasks, it is much easier to obtain the data in chunklets, without the need for labels. Each chunklet is the set in which the data comes from the same class but the actual class labels are not known. Such a scenario yields partial equivalence relations. There are some existing approaches in the literature about the learning with partial equivalence relations [20, 5]. One of the algorithms for this purpose is Relevant Component Analysis (RCA) [17]. RCA is an effective linear transformed algorithm used for data representation, which finds a linear transformation of the data such that irrelevant variability in the data is reduced. This "irrelevant variability" is estimated using chunklets. A nonlinear extension of RCA called kernel RCA was proposed by Tsang et al.[19]. However, the major drawback of RCA, similar to Principal Component Analysis (PCA), is that the transformations of RCA are optimized for representation or compression of data in a group, but may not be good enough for class discrimination [4]. In this paper, we address this problem by proposing a method for discriminatively learning linear transformations using the chunklet data. The key advantage of our technique is that it results in transformations that are better for class discrimination than RCA.

3. SYSTEM ARCHITECTURE

3.1 The General Framework

In this section we introduce the framework of the proposed system Cluchunk, which aims at building a web text clustering system with unlabeled data which incorporates inherent chunklet information. The workflow is depicted in Figure 3 and consists of three main consecutive phrases. For ease of illustration, we present an example to show how the comments from Facebook.com are grouped in the framework.

Feature Extraction: Suppose the input is comments with short text from a Facebook wall. As we discussed in Section 1, when using bag-of-words model to represent short text, it often leads to the problem that the short text is insufficient to provide enough term frequency for classification and clustering. In this paper, we consider the closeness of data in the same chunklet and explore the chunklet feature to enrich the representation of the short texts to be clustered. We propose a 3-step-framework of feature extraction: the extraction of features from the original text, the generation of features from chunklet data and the combination of the original and generated features.

Preprocessing for unsupervised clustering: In this phase, the goal is to get the discriminatively learnt linear transformation matrix using chunklets, in order to improve the subsequent performance of unsupervised clustering techniques. Then the original data is transferred into a new representation by the transformation matrix. This transformation can be viewed as a preprocessing step for unsupervised clustering. The details will be presented in next section.

Building a Cluster: After the features are obtained and the input features have been transformed into the new space, the task is to group the comments into a specific topic. In this work, we analyze a simple paradigm: given some predefined topics, the system can cluster all the comments into groups based on the topic.

3.2 Feature Extraction Using Chunklet

Similar to the approach in [14], we expand the original feature by directly fetching the external text. The feature representation is augmented with related chunklet text. The motivation is the data in one chunklet belong to the same class and we can enrich the feature representation with the other data in the same chunklet. The workflow consists of three consecutive steps, including feature extraction, feature generation and feature combination, as shown in Figure 4. The first step, feature extraction uses a simple bag of words representation. That is, each article is represented by a vector of terms appearing in the article [15]. The weight of each term is the frequency of the term in the text. The feature generated in this step is denoted as $f_{original}$. In the second step, the term frequency vector of the above method is augmented with selected chunklet data. For a given comment we use its related post and all other comments in the same chunklet as additional sources. Then we combine the two and apply bag-of-words to the combined text to get the feature f_{chunk}. Finally, the new representation of the given text is generated by a linear combination of the two: $f = w_1 \times f_{original} + w_2 \times f_{chunk}$, where w_1 and w_2 are the weight of the two features, and are determined empirically in this paper. We refer to this feature representation method as the *Chunk* method.

4. PREPROCESSING FOR CLUSTERING

In this section, we present how to find a good representation of the original data for clustering. We assume that the data is represented as vectors of features, and that the Euclidean distance in the feature space is used. The key point is that we transform the original data by a transformation matrix W which is learnt using the chunklet information, such that the Euclidean distance in the new feature space is so discriminative for clustering. Our work is inspired by the traditional RCA and the work from [13], in which the authors find the best Fisher's Linear Discriminant Analy-

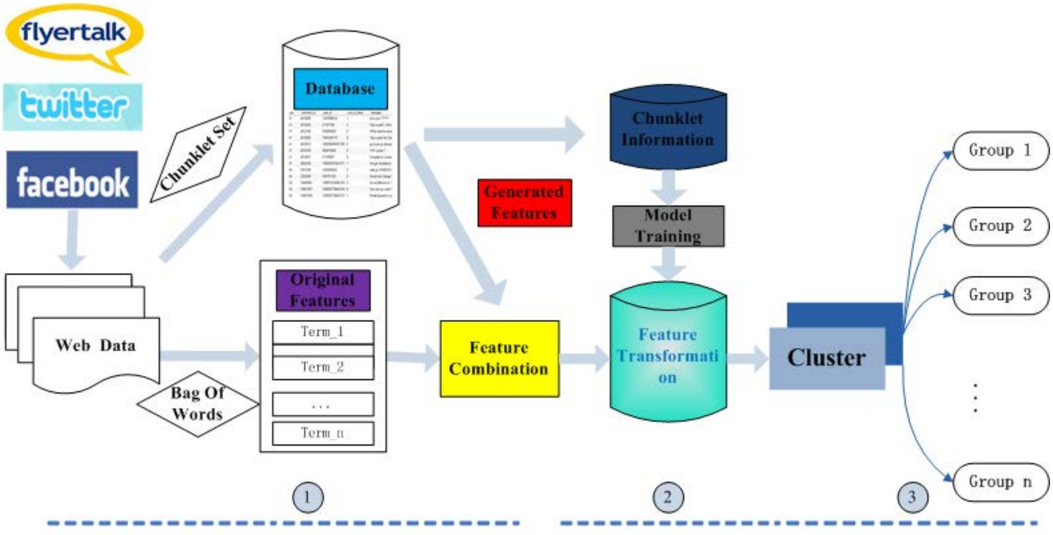


Figure 3: The overall framework of our system: Cluchunk

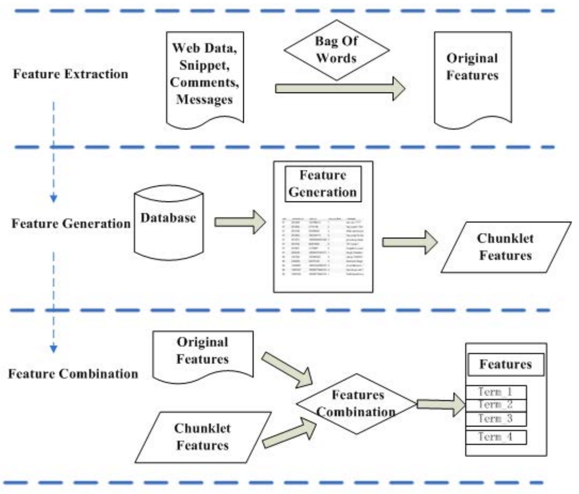


Figure 4: The workflow of feature extraction

sis (LDA) [8] projections using fully label data. Since our technique builds upon traditional RCA and LDA, we start with a brief review of RCA and LDA. Then we show that the method in [13] for fully supervised classification problem can be naturally extended to address the problem of clustering with chunklet information.

4.1 Problem Formulation

Suppose the sample data is given by $X = \{x_i\}_{i=1}^{|X|}$ where $|X|$ denotes the size of the sample. Let X_m denote the set of all the data with class label m, and M denote the number of classes where $\bigcup_M X_m = X$. We are provided with grouped data such that data in a chunklet share the same class label but without the exact label. Let H_n denote the sample of the nth chunklet where $\bigcup_N H_n = X$. We assume that the number of chunks N is larger than M, and $\forall H_n \subseteq X_i$ with some unknown class label i. In traditional RCA, it assumes that the distribution of the mth class is normal, i.e $X_m \sim N(\mu_m, \Sigma_m)$ and approximates the within-class scatter S_W, defined as

$$S_W = \sum_{m=1}^{M} \sum_{x \in X_m} (x - \mu_{X_m})(x - \mu_{X_m})^T \tag{1}$$

with the within-chunk scatter matrix, S_g, defined using the chunklet sets $\{H_n\}_{n=1}^{N}$ as,

$$S_g = \sum_{n=1}^{N} \sum_{x \in H_n} (x - \mu_{H_n})(x - \mu_{H_n})^T \tag{2}$$

where μ_{X_m} is the mean of data in class X_m, μ_{H_n} is the mean of the data in set H_n. Then the transformation matrix W^T is defined as $W^T = V\Lambda^{-\frac{1}{2}}$, V is the orthogonal matrix of eigenvectors of S_g and Λ is its corresponding matrix of singular values. V and Λ may be obtained from the singular value decomposition of $S_g = V\Lambda V^T$. Then apply W^T to the original data points x: $x_{new} = W^T x$. The key technique in RCA is it approximates S_W with S_g under the assumption that the chunklet is chosen randomly and is large enough. However, the projections of RCA are optimized for representation or compression of data in a chunklet, but not good for class discrimination [4].

Let us look at the supervised learning method LDA. In LDA, all the class labels are available during training and the LDA utilizes the label information to find informative projection(transformation) matrix W by maximizing the following objective function $J(W)$:

$$J(W) = \frac{|W^T S_b W|}{|W^T S_W W|} \tag{3}$$

where $|.|$ denotes the matrix determinant and S_W is defined in (1). S_b is between-class scatter and defined as:

$$S_b = \sum_{m=1}^{M} N_{X_m}(\mu_{X_m} - \mu)(\mu_{X_m} - \mu)^T \tag{4}$$

Here μ is the mean of the total input X and N_{X_m} is the number of points in class X_m. In our case, since we do not have access to the class labels, LDA cannot be applied directly to learn the desired transformation. So the main problem is, given the grouped data in the form of $X = \{H_n\}_{n=1}^{|N|}$ (instead of $X = \{X_m\}_{m=1}^{|M|}$), how can we learn a optimal transformation W as in using the chunklet information. This problem will be discussed in the following section.

4.2 Discriminatively Learning Linear Transformation

The work of Huang et al. [13] aimed at finding the best LDA projections given fully labeled data. Following its idea, we would like to find the transformations that minimizes the within-class scatter and maximizes the between-class scatter. In other words, we want to define an objection function like (3). Since we do not have access to the class labels, we make the same assumption as made in RCA and approximate the within-class scatter, S_W, with S_g defined in (2). However the "between-chunk" scatter $S_{\bar{g}}$, defined as,

$$S_{\bar{g}} = \sum_{n=1}^{N} N_{H_n}(\mu - \mu_{H_n})(\mu - \mu_{H_n})^T \tag{5}$$

is not a good approximation of the between-class scatter S_b (N_{H_n} is the number of points in chunklet H_n). As mentioned in Section 1, two input chunklets may contain data from the same class. In such a case, it is not reasonable to define a function like (3) and no discrimination would be possible between different classes. In this work, we can just minimize the within-chunk scatter S_g while keep the between-chunk scatter $S_{\bar{g}}$ not collapse to zero. We look for the transformation W that optimize the following objective functions:

$$\begin{aligned} &\min_{w} \quad |W^T S_g W| \\ s.t. \quad &|W^T S_{\bar{g}} W| > 0, \quad and \\ &||\omega_i||^2 = 1 \quad for \quad i = 1, ..., m \end{aligned} \tag{6}$$

where ω_i are the columns of the transformation matrix W and this objective function can be solved via generalized Eigenvalue decomposition. Similar to RCA, when W is obtained, the original data is projected by W (in the transformed space). This transformation is named as chunklet learning transformation (ChunkLT in this paper). In the transformed space, the unsupervised clustering methods are applied to the data. It is easy to prove that the ChunkLT and RCA have the equivalent computational complexities.

5. EXPERIMENTS ON BENCHMARK DATA

In this section, we demonstrate the effectiveness of our proposed system CluChunk. To evaluate the proposed framework on a real-word data, we applied it to the problem of grouping text from Facebook.com and online forums into different clusters based on its topic. For this, two large corpus have been collected and labeled manually. The first dataset fbs-5000 contains 5000 comments and posts from Facebook.com with 10 topics varied from music, automobiles, to religion, politics, etc. It is important to note that the 5000 items are from 228 chunklets. The second one is fly-3000, which contains 3000 forum threads and replies from Flytertalk. The 3000 items are divided into 6 categories and originally came from 145 small chunklets.

5.1 Performance Comparison

In this section, the first experiment is to study how the proposed ChunkLT method helps in the clustering of data, and compare it with RCA. The K-means method is used to cluster the data after it has been processed by the ChunkLT and RCA. We run the K-means algorithm by fixing the number of clusters, K. We refer the method using only K-means as the *Baseline*, and compare it with RCA and the proposed ChunkLT. The features in this experiment are all generated using *Chunk* method before applied to the three algorithm. Cluster quality is evaluated by two metrics: Purity and F-score. Purity measures the frequency whereby data points that belong to the same cluster share the same class label. F-score combines the information of precision and recall which is extensively applied in information retrieval. Both the two metrics range from 0 to 1, and the higher their value, the better the clustering quality is. Tables 1 and 2 show the results of the three algorithms with different group sizes ($M = 4, 6, 8, 10$) on fbs-5000 and fly-3000. ChunkLT achieves higher F-score and accuracy than RCA and *Baseline* for all group sizes. However, the performance on fly-3000 is better than fbs-5000. The reason is the two datasets have different structure (different chunklet size).

The goal of the second experiment is to show the effectiveness of the proposed method *Chunk* for the feature representation and the performance of the system CluChunk. We compare the results of four frameworks (feature generation + clustering) below:
Km (baseline): Traditional bag-of-words model, only using K-means, no RCA or ChunkLT before clustering
Km+*Chunk*: The *Chunk* method proposed in Section 3, only using K-means, no RCA or ChunkLT before clustering
ChunkLT: Traditional bag-of-words model, using ChunkLT before clustering
ChunkLT+*Chunk*: The *Chunk* method propose in Section 3, using ChunkLT before clustering
Table 3 and 4 shows the results of the four approaches with different class sizes on fbs-5000 and fly-3000. From the results, it can be concluded that the *Chunk* method for the feature generation is very powerful. Our problem system with ChunkLT+*Chunk* can get great improvement than the method only using Km.

5.2 The Effect of Chunklet Size

The variance of the RCA estimator is effected by the number of chunklets in the dataset. In [4] Hillel et al. point that the variance of the RCA estimator using small chunklets rapidly converges to the variance of the best estimator. To better understand how the chunklet size affects the clustering performance based on ChunkLT, we control the structural characteristics in the two datasets and perform the experiments using different chunklet sizes. The results are shown in Figures 5 and 6 (with the class size 5 and 10 respectively). It is concluded that as the chunklet size increases, the accuracy decreases.

6. CONCLUSIONS

Automatical clustering of large scale web text data usually encounters two main problems. Firstly, as the acquisition of a large quantity of labeled data is time-consuming and expensive, it is difficult to get accurate performance using traditional unsupervised clustering method. Secondly, web text is short and does not have enough content to di-

Table 1: Results of K-means clustering applied to fbs-5000 dataset with three different preprocessing and four different class sizes

Algorithm	*Baseline*		RCA		ChunkLT	
M	F-Score	Purity	F-Score	Purity	F-Score	Purity
4	0.598	0.652	0.654	0.683	0.713	0.744
6	0.522	0.543	0.575	0.602	0.622	0.653
8	0.462	0.489	0.546	0.525	0.564	0.581
10	0.406	0.422	0.472	0.503	0.532	0.545

Table 2: Results of K-means clustering applied to fly-3000 dataset with three different preprocessing algorithms and four different class sizes

Algorithm	*Baseline*		RCA		ChunkLT	
M	F-Score	Purity	F-Score	Purity	F-Score	Purity
4	0.622	0.684	0.704	0.732	0.752	0.805
6	0.573	0.626	0.601	0.640	0.655	0.677
8	0.536	0.579	0.569	0.595	0.608	0.625
10	0.487	0.518	0.516	0.539	0.561	0.595

Table 3: Results of different algorithms applied to fbs-5000 with four class sizes

RCA	Km		Km+*Chunk*		ChunkLT		ChunkLT+*Chunk*	
M	F-Score	Purity	F-Score	Purity	F-Score	Purity	F-Score	Purity
4	0.537	0.622	0.598	0.652	0.644	0.667	0.713	0.744
6	0.487	0.508	0.522	0.543	0.569	0.614	0.622	0.653
8	0.437	0.450	0.462	0.489	0.512	0.535	0.564	0.581
10	0.389	0.411	0.487	0.518	0.482	0.504	0.532	0.545

Table 4: Results of different algorithms applied to fly-3000 with four class sizes

RCA	Km		Km+*Chunk*		ChunkLT		ChunkLT+*Chunk*	
M	F-Score	Purity	F-Score	Purity	F-Score	Purity	F-Score	Purity
4	0.600	0.575	0.622	0.684	0.635	0.703	0.752	0.805
6	0.521	0.533	0.573	0.626	0.595	0.638	0.655	0.677
8	0.485	0.488	0.536	0.579	0.555	0.601	0.608	0.625
10	0.404	0.457	0.482	0.504	0.567	0.592	0.561	0.595

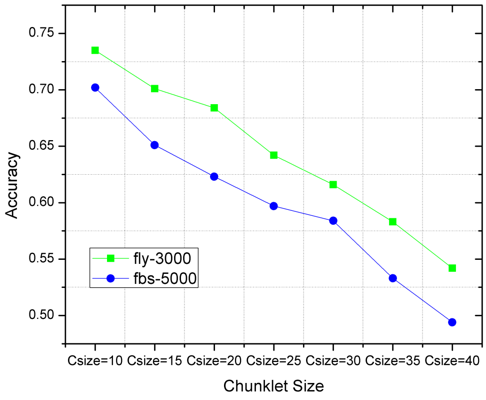


Figure 5: Accuracy on test set with different chunk size when the class size is 5

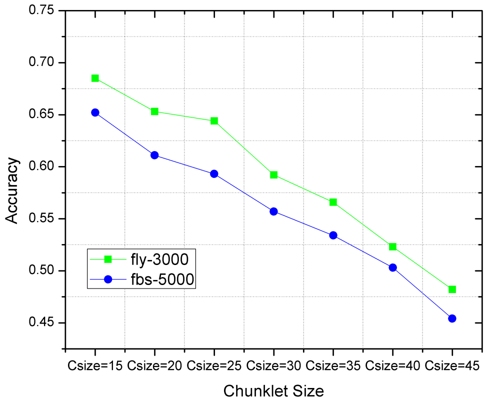


Figure 6: Accuracy on test set with different chunk size when the class size is 10

rectly build a feature space for clustering. In this paper, we have presented a general framework which learns discriminative linear transformations using only unlabeled chunklet data. The chunk information can be obtained easily or even automatically for several kinds of web application. We demonstrated that the classification/clustering performance for web text can be significantly improved by: 1)incorporating the inherent chunklet information into the clustering framework; 2)enriching the representation of short text with additional features from related chunklets. Interestingly, the ChunkLT method got large improvement, even though it only needs slightly more prior information than the unsupervised method.

The key idea in this paper, which is to explore inherent or automatically obtained chunklet information to help clustering analysis, represents an interesting research direction in web text mining. There are many potential future directions from this work. A straightforward task would be applying it to the other type of social media data, like topic-based blog summarizing, and feature-based customer review analysis. In the theoretical sense, the current version is a linear machine and we plan to extend it to a non-linear machine with the kernel technique in future.

7. ACKNOWLEDGMENT

This work is supported in part by NSF award numbers CCF-0621443, OCI-0724599, CCF-0833131, CNS-0830927, IIS-0905205, OCI-0956311, CCF-0938000, CCF-1043085, CCF-1029166, and OCI-1144061, and in part by DOE grants DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, DE-SC0005340, and DE-SC0007456. We also would like to thank Facebook.com, Voxsupinc.com for providing access to their data.

8. REFERENCES

[1] 200 million Tweets per day. http://blog.twitter.com/2011/06/200-million-tweets-per-day.html.

[2] Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 757–766, New York, NY, USA, 2007. ACM.

[3] S. Banerjee, K. Ramanathan, and A. Gupta. Clustering short texts using wikipedia. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788, New York, NY, USA, 2007. ACM.

[4] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *In Proceedings of the Twentieth International Conference on Machine Learning*, pages 11–18, 2003.

[5] S. Becker, S. Thrun, and K. Obermayer, editors. *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*. MIT Press, 2003.

[6] B. Carter. How To Get More Likes And Comments On Facebook. http://allfacebook.com/how-to-get-more-likes-and-comments-on-facebook-book-excerpt.

[7] M. Chen, X. Jin, and D. Shen. Short text classification improved by learning multi-granularity topics. In T. Walsh, editor, *IJCAI*, pages 1776–1781. IJCAI/AAAI, 2011.

[8] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, New York, 2. edition, 2001.

[9] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, 2007.

[10] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press, 2004.

[11] J. Hu, L. Fang, Y. Cao, H.-J. Zeng, H. Li, Q. Yang, and Z. Chen. Enhancing text clustering by leveraging wikipedia semantics. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 179–186, New York, NY, USA, 2008. ACM.

[12] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou. Exploiting wikipedia as external knowledge for document clustering. In *In Proc. of Int. Conf. on Knowledge Discovery and Data Mining (KDD*, 2009.

[13] R. Huang, Q. Liu, H. Lu, and S. Ma. Solving the small sample size problem of lda. In *In: Proceedings of the 16 th International Conference on Pattern Recognition (ICPRąŕ02*, pages 29–32, 2002.

[14] M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 377–386, New York, NY, USA, 2006. ACM.

[15] G. Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[16] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang. Query enrichment for web-query classification. *ACM Trans. Inf. Syst.*, 24(3):320–352, July 2006.

[17] N. Shental, T. Hertz, D. Weinshall, and M. Pavel. Adjustment learning and relevant component analysis. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, pages 776–792, London, UK, UK, 2002. Springer-Verlag.

[18] W. tau Yih and C. Meek. Improving similarity measures for short segments of text. In *AAAI*, pages 1489–1494. AAAI Press, 2007.

[19] I. W. Tsang, P. ming Cheung, and J. T. Kwok. Kernel relevant component analysis for distance metric learning. In *In IEEE International Joint Conference on Neural Networks (IJCNN*, pages 954–959. IJCNN, 2005.

[20] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 577–584, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.