

THE 18TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING August 12-16, 2012.

Beijing, China

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

Yu Cheng, Yusheng Xie, Kunpeng Zhang, Ankit Agrawal, Alok Choudhary



EECS Department, Northwestern University Evanston IL, 60208, USA {ych133,yxi389,kpz980,ankittag,choudhar}@eecs.northwestern.edu

Introduction

Bigdata from web: a large amount of web content being generated by users in the form of forums, blogs, microblogs, customer reviews and so on.



Necessity: the huge amount of information invariably makes its manual comprehension infeasible for a person and urges the development of automated methods geared to help the user better understand this information. Eg. clustering/classification

Challenge from clustering/classification: 1) web text is short and sparse; 2) labeling such data is expensive.

Motivation

Chunklet Data: 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. This kind of data is called **Chunklet** data.



Fig 2. The overall framework of the ClChunk system

Transformation Matrix: W, learned from the chunklet data

$$\begin{split} \min_{w} & |W^{T}S_{g}W| \\ s.t. & |W^{T}S_{\overline{g}}W| > 0, \text{ and} \\ & ||\omega_{i}||^{2} = 1 \text{ for } i = 1, 2, ..., m \end{split}$$

-	An example with			94	Annor A	AAO 10	inention acces	Internege						
-	William Andrew McKelligott Ah! There he is. Lot's of people have been asking where he's been lately	ntly had a		41	2612005	1387996314	1	who won ?????						
	22 minutes ago			41	2612052	21101106	2	Yay Lucas!!! Won						
Å	Concetta Santoro bravo bravososososososososososososososososososos									41	2612168	533955028	0	What was the prize
105	Bob Anderson Looking forward to the Republican Presidential debate on Fox News tonight at 6 p.m. 21 minutes ago			41	2612652	1642324115	0	Yay Lucas! Go Dav						
7		: flavor.		41	2612812	10000046591399	0	go lucas go dawgs						
	Hong Cong Phan Nr. President now could sit down a	n NYC in		41	2614762	682816936	0	YAY Lucas !!						
	while. Whoever dowingrade U.S credit, it just by saying words. The reality of wealthy was built by Hundreds years, Money isn't all about. 21 minutes aco			41	2614931	21105087	0	Congrats to Lucas!						
-	Neida Canales Carlosliamasa it sounds like you are following	fice	V	41	2624330	100000151641371	1	He got 4 tickets to						
5	politics we all expected to much from the President but its not all up to him republicans keeps BLOCICING 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			4	2351304	1252993223	0	Lets go DAWGS!!!						
				44	2352259	534781332	0	Good luck Dawgs!						
	and would love to do all he promised . 21 minutes ago			54	14445981	100001224882163	0	im not little but im 1						
				54	14551527	100000776642153	0	How do you vote?						
	Tonya Dunn Swink My triends man loves the bacon 22 hours ago		54	14551542	100000776642153	1	Finally figured it out							

Fig 1. An illustration of a post and its comments from Facebook.com

Methodology

Feature Extraction Using Chunklet: 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 2.

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

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

RCA	Km		Km+Chunk		Chun	kLT	ChunkLT+Chunk	
М	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 3: Results of different algorithms applied to fbs-5000 with four class sizes

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

RCA	Km		Km+Chunk		ChunkLT		ChunkLT+Chunk	
М	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
e	0 491	0 535	0 579	0.696	0 505	0.600	0.655	0.677



Fig 2. The workflow of feature extraction

Clustering Using Chunklet Information: we propose an algorithm: ChunkLT for data pre-pocessing, which aims to discriminatively learn a linear transformation matrix using the inherent chunklet information, such that the Euclidean distance in the new feature space is so discriminative for clustering.

0.0000.0110.0210.000 0.073 0.0200.595 0.0300.5360.5550.6080.6250.4850.4880.5790.601ð 0.5950.5670.4570.4820.5040.5920.561100.404

Fig 3. Results of different algorithms on two datasets

Experiments

Data Sets: (1)fbs-5000 Facebook data; (2) fly-3000: forum data.

Comparison Algorithms: (1)Km+bow; (2)Km+*Chunk*; (3)ChunkLT+bow; (4)ChunkLT+Chunk.

Results: the *Chunk* method for the feature generation is very powerful. Our proposed system with ChunkLT+Chunk can get great improvement than the baseline methods.

Paper ID: BigMine-2