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 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.

Fig 2. The workflow of feature extraction

Clustering Using Chunklet Information: we propose an algorithm: ChunkLT for data pre-processing, 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.

Fig 2. The overall framework of the CluChunk system

Transformation Matrix: W, learned from the chunklet data

\[
\begin{align*}
\min_{W} & \quad |W^T S_W W| \\
\text{s.t.} & \quad |W^T S_W W| \geq 0, \text{ and} \\
\quad & \quad \|w_i\|^2 = 1 \text{ for } i = 1, 2, \ldots, m
\end{align*}
\]

\[
S_W = \sum_{i=1}^{m} \sum_{j=1}^{m} (x-u_{ij}(x-u_{ij})^T)
\]

\[
S_T = \sum_{i=1}^{m} |N_{ij}\nu_i - \mu_i|(|N_{ij}\nu_i - \mu_i|)^T
\]

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.