Stream Short Text Document Clustering

Chanon Jenakom and Methawee Apinainarong

March 28, 2018

Abstract

Short text documents, such as instant messages, SMS, or news headlines, have been increasingly useful for data analysis in recent times. Furthermore, these text documents are presented in real time, requiring a form of stream clustering technique. A data stream is a continuously generated sequence of data for which the characteristics of the data evolve over time. In this paper, we propose a short text document clustering technique which supports continuous data streams in real time using E-Stream algorithm as a stream clustering technique, Distributed Word Representation for representing each document, and Word Mover's Distance as the distance metric. It is expected that this proposed algorithm will offer a new way to effectively represent short text documents in real-time and offer meaningful patterns for future analysis.

1 Introduction

In social media, stream short text documents are text documents that contain very few words, such as instant messages, SMS, or news headlines that are ordered sequences of documents that arrive in timely order. Different from data in traditional static databases, data streams are continuous, unbounded, usually come with high speed and have a data distribution that often changes with time. Therefore, developing data mining techniques to handle the large volume of short text documents from data stream has become an important goal. Short text clustering is already a challenging task; due to the sparsity and noise, they provide very few contextual clues for applying traditional data mining techniques; therefore, short documents require different or more adapted approaches. The representation of short-text segments needs to get enriched by incorporating information about correlation between terms. Data streams, because of their unique features, have further posed many new challenges to short text document clustering. There are three main challenges: single access of data, unbounded data, and real-time response. In addition to the aforementioned challenges, applying stream clustering to short text documents requires an efficient method to represent and store documents for computation of clusters. In this paper, the focus is on developing a new clustering algorithm that is suitable for clustering short text documents from differing sources of data streams. Some previously proposed algorithms are chosen as a basis for developing this stream short text document clustering algorithm. Then, the result and performance of the proposed algorithm will be shown on a web-based application.

2 Literature Summary

Table 1: Six papers on stream clustering and Short text document clustering, with their scopes and goals.

Paper	Scope	Goal
E-Stream [1]	Propose stream clustering that supports five evolutions	Stream Clustering (SC)
Similarity Measures [2]	Compare and analyze document distance measures	Document Distance (DD)
SE-Stream [3]	Propose stream clustering that supports high dimensional data streams	Stream Clustering (SC)
Distributed Representations of Words [4]	Present several extensions that improve both the quality of the vectors and the training speed	Document Representation (DR)
Supervised Word Mover's Distance [5]	Propose an efficient technique to learn a supervised metric	Document Distance (DD)
Short Text Document Clustering [6]	Presents a method for clustering short text documents	Document Distance (DD) and Document Representation (DR)

Based on the research papers' goals, we can divide these papers into three categories: stream clustering (SC), document distance (DD), and document representation (DR). E-Stream [1] and SE-Stream [3] discuss about two stream clustering algorithms where SE-Stream is an extension of E-Stream. Similarity Measures for Text Document Clustering [2], Supervised Word Mover's Distance [5], and Short Text Document Clustering [6] papers then discuss about commonly used document distance metrics. Similarity Measures for Text Document Clustering paper compares five metrics and their performance, while Supervised Word Mover's Distance and Short Text Document Clustering papers focus on one specific metric. Short Text Document Clustering and Distributed Representations of Words [4] discuss two Distributed Word Representation algorithms, where one is a supervised extension of the former.

Table 2: Six papers on stream clustering and Short text document clustering, with their algorithms and performances.

Paper	Algorithm	Performance
E-Stream [1]	E-Stream	Polynomial with respect to the number clusters

Paper	Algorithm	Performance
Similarity Measures [2]	Euclidean Distance, Cosine Similarity, Jaccard Coefficient, Pearson Correlation Coefficient, Averaged Kullback-Leiber Divergence	The averaged KL divergence and Pearson coefficient tend to outperform the cosine similarity the Jaccard coefficient, except for the classic dataset
SE-Stream [3]	SE-Stream	Quadratic with respect to the number of dimensions
Distributed Representations of Words [4]	Distributed Word and Phrase Representation	This results in a great improvement in the quality of the learned word and phrase representations
Supervised Word Mover's Distance [5]	Supervised Word Mover's Distance	S-WMD manages to capture difference in words based on the context of the article
Short Text Document Clustering [6]	Distributed Word Representation and Word Mover's Distance	The combination between the two algorithms outperforms others significantly

From the performance comparison between stream clustering algorithms, SE-Stream has better performance since it tried to reduce the number of dimensions in the incoming data before using them to compute the clusters. In addition to that, while there are many document distance metrics that are used, Word Mover's Distance is generally the most popular.

References

[1] Komkrit Udommanetanakit, T. R., and Kitsana Waiyamai (2007), E-Stream: Evolution-Based Technique for Stream Clustering. *LNCS, 4632*, pp. 605-615.

[2] Huang, A. (2008). Similarity Measures for Text Document Clustering. *NZCSRSC*, pp. 49-56.

[3] Rattanapong Chairukwattana, T. K., Thanawin Rakthanmanon, Kitsana Waiyamai. (2013). Evolution-Based Clustering of High Dimensional Data Streams with Dimension Projection. *ICSEC*, pp. 190-195.

[4] Tomas Mikolov, I. S., Kai Chen, Greg Corrado, Jeffrey Dean. (2015). Distributed Representations of Words and Phrases and their Compositionality. pp. 1-9.

[5] Gao Huang, C. G., Matt J. Kusner, Yu Sun, Kilian Q. Weinberger, Fei Sha. (2016). Supervised Word Mover's Distance. *NIPS, 30*, pp. 1-9.

[6] Supavit KONGWUDHIKUNAKORN, K. W. (2017). Short Text Document Clustering using Distributed Word Representation and Document Distance. *Walailak J Sci and Tech, 14*, pp. 1-13.