# Stream Short Text Document Clustering

Chanon Jenakom and Methawee Apinainarong

***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.**

## I. Introduction

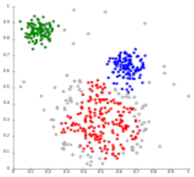

Fig. 1. An example of a result of clustering algorithm into 3 clusters

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, similar to the example seen in figure 1 on page 1, and performance of the proposed algorithm will be shown on a web-based application.

## II. Literature Summary

TABLE I

Six papers on stream clustering and Short text document clustering, with their scopes, goals, algorithm(s), and performance.

| Paper | Scope | Goal | Algorithm(s) | Performance |
|---|---|---|---|---|
| E-Stream [1] | Propose stream clustering that supports five evolutions | Stream Clustering (SC) | E-Stream | Polynomial with respect to the number clusters |
| Similarity Measures [2] | Compare and analyze document distance measures | Document Distance (DD) | 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] | Propose stream clustering that supports high dimensional data streams | Stream Clustering (SC) | SE-Stream | Quadratic with respect to the number of dimensions |
| Distributed Representations of Words [4] | Present several extensions that improve both the quality of the vectors and the training speed | Document Representation (DR) | 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] | Propose an efficient technique to learn a supervised metric | Document Distance (DD) | 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] | Presents a method for clustering short text documents | Document Distance (DD) and Document Representation (DR) | Distributed Word Representation and Word Mover's Distance | The combination between the two algorithms outperforms others significantly |

Based on the research papers' goals, we can divide these papers into three categories: stream clustering (SC), document distance (DD), and document representation (DR). According to column Goal in Table I, E-Stream [1] and SE-Stream [3] discuss about two stream clustering algorithms. SE-Stream is an extension of E-Stream, as seen in column Scope of Table I. In column Goal of Table I, 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. According to column Scope in Table I, 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, according to both column Scope and Goal in Table I.

As seen in column Performance in Table I, the performance comparison between stream clustering algorithms, SE-Stream has better performance since it tries to reduce the number of dimensions in the incoming data before using them to compute the clusters (Column Algorithm(s), Table I).

## III. Relevant Theory

### A. Stream Clustering

Stream clustering is a data mining technique that clusters data from data streams, producing results in real time. It is very important for stream clustering to be able to compute the incoming data within a single pass and using limited memory. Examples of such techniques are E-Stream [1] and SE-Stream [3].

*1) E-stream:* The main idea of this stream clustering technique is that data stream's behavior can evolve over time. There are five categories: appearance, disappearance, self-evolution, merge, and split.

*2) SE-Stream:* This stream clustering technique is an extension of the previous technique, E-Stream, to support high dimensional data streams. The cluster quality and execution time of SE-Stream is improved when compared to E-stream.

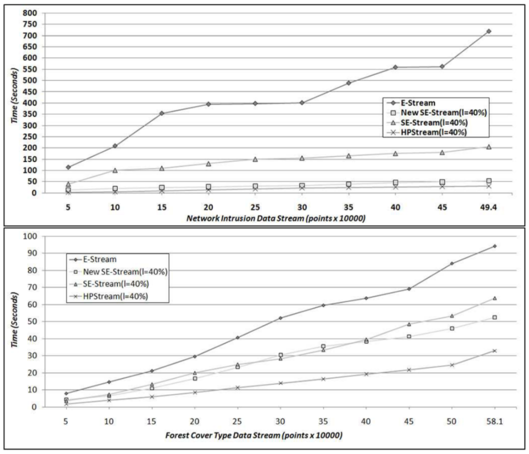

Fig. 2. Performance comparison in term of execution time between E-Stream and SE-Stream

In figure 2 on page 2 shows performance of SE-Stream and E-Stream in term of execution time. The execution time of SE-Stream is significantly lower than E-stream.

### B. Short Text Document Clustering

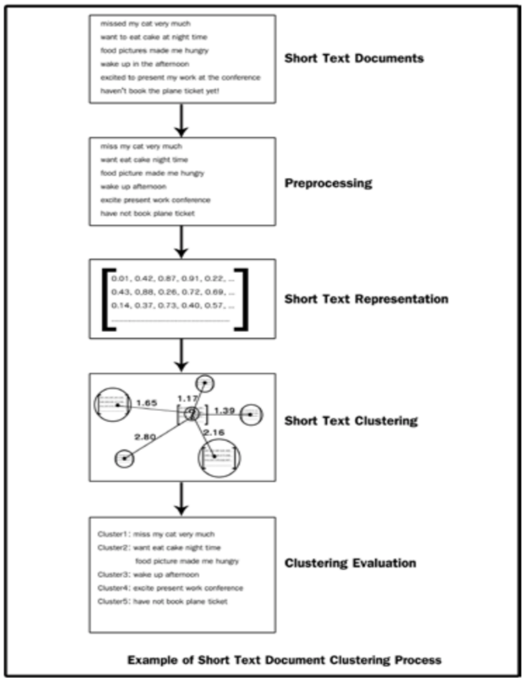
missed my cat very much
want to eat cake at night time
food pictures made me hungry
wake up in the afternoon
excited to present my work at the conference
haven't book the plane ticket yet!
Short Text Documents
miss my cat very much
want eat cake night time
food picture made me hungry
wake up afternoon
excite present work conference
have not book plane ticket
Preprocessing
0.01, 0.42, 0.87, 0.91, 0.22, ...
0.43, 0.88, 0.26, 0.72, 0.69, ...
0.14, 0.37, 0.73, 0.40, 0.57, ...
Short Text Representation
1.65
1.17
1.39
2.80
2.16
?
Short Text Clustering
Cluster1: miss my cat very much
Cluster2: want eat cake night time
food picture made me hungry
Cluster3: wake up afternoon
Cluster4: excite present work conference
Cluster5: have not book plane ticket
Clustering Evaluation
Example of Short Text Document Clustering Process

Fig. 3. An example of Short Text Document Clustering Process

Short text documents have become increasingly important for data analysis. In order to analyze the documents, an effective representation of the documents and a proper distance metric are required, such as Distributed Word Representation and Word Mover's Distance [6], respectively.

As seen in figure 3 on page 2, short text documents are firstly preprocessed by standard text preprocessing techniques, resulting in a vector space of unique words. Then, the trained neural network is applied to the preprocessed data to create the word representations. Lastly, K-Means algorithm is used to cluster these representations of the documents using WMD as the distance metric.

*1) Distributed Word Representation:* This algorithm learns and represents the words in vector space using a neural network model. It is able to capture semantic similarity between words. The main idea of this algorithm is to represent each word by a vector of certain dimension.

*2) Word Mover's Distance:* This distance metric is based on the idea of Earth Mover's Distance, measuring how far the words of one document must be "moved" to match another document. In order to use WMD, the documents must be represented in vector space of certain dimension, containing vocabularies, or unique words, from the documents.

## 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.