A Novel Text Stream Clustering Technique for Web Pages using Sliding Window

V.Kumuthavalli
Associate Professor, Department of Computer Science,
Sri Parasakthi College for Women, Courtallam, Tirunelveli, Tamil Nadu, India.
Email: saikumuthavalli@gmail.com
Dr.V.Vallimayil
Associate Professor & Head, Department of Computer Science & Applications,
Periyar Maniyammai University, Vallam, Thanjavur, Tamil Nadu, India.
Email: vallimayilv@gmail.com

Abstract- The text mining gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. In the current scenario, text data streams gains lot of significance in processing. Due to rapid development of the information technology, large numbers of electronic documents are available on the internet instead of hard copies. It provides beginning advice to information in social network for making decision, a clustering for text stream algorithm is proposed to cluster the text stream, which is formed by web crawler to continuously grab the web pages. The time sliding window able to split the text stream into continuous segments of web page news associated to velocity of stream and size of sliding window. Here, multilevel cluster method is used to merge the cluster in each sliding window. The results of experiments, used 2750 web page news simulate text stream by web crawler using the algorithm with executing efficiency and the higher clustering quality in terms of precision and recall rate. The experimentation and results with various documents and compared with existing methods and it provides better results.

Keywords- Text Categorization, sliding window, data stream, text mining, clustering.

1. INTRODUCTION

Today, information has a great value and the amount of information has been expansively growing during last years. Especially, text databases are rapidly growing due to the increasing amount of information available in electronic forms, such as electronic publications e-mail and the World Wide Web. The information around us, that becomes a problem to find related for our necessary. Because of this, there are many databases and catalogues of information classified into many categories, helping the viewer to easily navigate to the information. Most information in world is texts and here the text streaming comes to the scene.

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Generally, data mining is process of analyzing data from different perspectives and summarizing it into useful information. Text mining has been defined as discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. Text mining, which is sometimes referred to text analytics, is one way to make qualitative or unstructured data usable by computer. Text mining can help an organization derive potentially valuable business insights from text-based content.

1.1 Text Categorization

Text categorization is one of the well studied problem in data mining and information retrieval. Categorization is the process in which ideas and objects are recognized, differentiated and understood. Categorization implies that objects are grouped into categories, usually for some specific purpose. A category illuminates a relationship between the subjects and objects of knowledge. The data categorization includes the categorization of text, image, object, voice etc. With the rapid development of the web, large numbers of electronic documents are available on the Internet. Text categorization becomes a key technology to deal with and organize large numbers of documents. Text categorization is the assignment of natural language documents to one or more predefined categories based on their semantic content is an important component in many information organization and management tasks [1][2]. The techniques of text categorization are necessary for improving the quality of any information system dealing with textual information, although they cover only a fraction of document management.

2. REVIEW OF LITERATURE

In this research work, discussed about the methodology for feature extraction and document classification. In order to propose this works have analyzed various literatures which are very much relevant and helpful to do this work. The literature where are have retrieved and analyzed are presented in the following section.

With the development of mobile cloud computing and social network application, the value of data is changing the users consume and decision habit in the big data era. Especially, the rapidly increasing data with high velocity to form a data stream, which is continuous, unlimited and dynamic variable data set [3].

In order to help web users making decision in real-time, how to achieve and extract the valuable information from the data stream and to form different topic clusters in time are new challenges, especially, how to extract text feature vector with multi-dimension and to design a cluster algorithm for text stream with lower time and space complexity. Therefore, the evaluation about the quality of cluster algorithm, such as precision rate, recall rate, efficiency and robustness, are becoming the key points, when a single pass scanning method [4] is used in this process. To solve above problems, this paper proposes a Topic-Based Dynamic Clustering Algorithm for Text Stream (TBDC4TS), which uses a sliding time window to split the text stream into continuous segments and to transform the text stream cluster from flow data to continuous batch data processing.

There are three kinds of models to process the data stream, such as the time-limited model, the sliding window model and the snapshot model [4, 5]. The data scale of all these three models depend on the selection of time interval, which are defined by the time interval from an initial time to current time, a certain time widow size and a certain time interval between each snapshot operation, respectively.

Moreover, some researchers focus on the structure of data stream clustering, which includes some algorithms based on the Single-pass and Clu-stream algorithm[6]. Single-Pass is a classic incremental clustering algorithm with single scanning the whole data set. The upcoming data in stream, which is captured by system, should be compared to existed clusters one by one, if there is a cluster which has the highest similarity degree with the new data and larger than the threshold, then merge the new data into this cluster and recalculate the new average feature of cluster, else a new cluster can be created by this new data point. This algorithm is suitable to the large data with certain number of clusters, but not suitable to the situation that the number of cluster is varying when the data volume is constant increasing with data flow.

Based on the Single-Pass strategy, Zhu[7] analyze the influence factors on the efficiency and quality of clustering by feature's weighted coefficient for the dimension of a feature vector.

Yi[8] proposed a method of periodicity incremental clustering to obtain a new centre point of cluster. Yin[8] also put forward a method to split the data stream into a serials of chunk of data to optimize and decrease the effect of the sequence of data stream.

Clu-Stream algorithm is a clustering framework for data stream with two phases, including a real-time online clustering and off-line clustering. In the process of online clustering, the micro-cluster can cluster the data sets in different segments of data stream. Then, in the process of offline cluster, the macro-cluster can put these new clusters created in online into the whole cluster sets and merge them into existed clusters by similarity measuring. In addition, a Pyramid model of time framework is designed to store the clustering results in different granularities and phases.

Li[9] proposed a sliding time window, based on the Clu-Stream algorithm, to increase the efficiency of mirco-cluster in online. However, because the Clu-Stream adopts the hierarchical clustering method with BIRCH algorithm, which just is suitable for the data set with same number of feature's dimension, and not suitable to cluster the text data sets with variable dimensions of feature vector. But, all these methods have lower efficiency with larger computation to index the high frequency words in text.

3. METHODOLOGY

3.1 Single Pass Clustering

The single pass clustering is incremental clustering algorithm. It requires only pass the input dataset. The specialization of single pass clustering algorithm using centriod list and cosine similarity[11], the inputs of single pass clustering are dataset and a threshold. The proposed algorithm starts with setting first document vector as an initial centroid, then iterates all document vectors of datasets and finally computes the cosine similarity of each document vector and centroid list to find the nearest centriod. It compares distance of the nearest centroid with given threshold it less than specified threshold the document nearest centroid will be recalculated after assigning the document vector. The output of single pass clustering algorithm is centroid list.

3.2 Proposed Methodology

A major problem of traditional approach is high dimensionality of the feature vector. The feature vector with a large number of key terms is not only unsuitable but also easily to cause the over fitting problem. The goal of a classifier is to assign a category to given unseen documents. In general, the processing of automatic text streams, it first is the extraction of feature terms that become effective keywords and the second is classification of the document using these features.

The sampling of data stream is chosen at random. It can use the sliding window model to analyze stream data. The sliding-window model computation is motivated by running computations on all of the data seen. It makes decisions based only on recent data. At every time t, a new data element arrives. In this element expires at time t +w, where w is the window size. This model useful for complex technique for producing approximate answers to a data stream query for only recent events may be important uses this model. To reduce the memory requirements only a small window of data is stored. In present scenario, lot of on-line analysis software tools uses this model for generating summaries of data in stream. There are two types of windows called count based and time based can be developed. In past, the n items are stored and later can store only those items which have been generated. Micro-clustering used be perform construction of data streams. Single-Pass clustering algorithm has lower efficiency it not only calculate the similarity by distance between each new data and existed cluster in time serial processing, can be affected by sequence of data input time. Sliding Time Window can be used in this work provided batch processing in continuous window. It can improve the clustering efficiency and manipulate the data sequence added into stream. The clustering framework for text stream in the whole framework and analysis. The mechanism of clustering includes the process of clustering text stream are sliding window scheduler, text clustering. Web Crawler can capture the web pages text data source to form continuously text stream. In order to improve the efficiency on extraction of key words needs a pre-processing for text filtering like reduplication URL address, embed image, video, advertisement etc., keep away from make the negative influence for text feature extracting. The

Clustering of sliding window scheduler to control the basic time window (BWt)k with unique procedure and change the data stream into batch data stream. Text feature in document sets in (BWt)k can be extracted and indexed by segmenting the word tool and bag of words. The text feature vector can be represented by <Person, Address, Time and Frequent Terms>. After the text feature extracted and a feature vector of the document formed, the Text Clustering is triggered to cluster density algorithm the different feature vectors in k-th BWt, then a sets of clusters can be output, including {Clusterk1, Clusterk2…, Clusterkn}. In each cluster, same as a data point with its own feature is average of features documents in the cluster. Hence, each micro-cluster in one BWt are calculated the distance between each of present macro-clusters and then merged into macro-cluster which has shortest distance and less than the threshold. In addition to the process and result of clustering can be visualized to provide decision support. The framework of sliding time window and clustering process together to provide a multilevel strategy for text stream.

Where

i. Person sets of person name emerged in the document
person1, person2…, personn}

ii. Address sets of geographical place feature emerged in document
{addr1, addr2… , addrm}

iii. Time sets of time related feature information emerged in document
{time1, time2…, timek}

iv. Frequent terms sets of high-frequency words in document
{term1, term2…, termj}

Basic Time Window (BW_t):

Let t and p represent the time and time interval respectively. The document set is obtained from system as a time series in the time interval of (*t*, *t*+*p*). Then, Basic Time Window (BW_t) is document set in certain time interval.

$$BWt = \{Doc_{i,j},\ 0< i \leq n,\ t \leq j \leq (t+p)\} \text{…… (3.1)}$$

Where, the length of BW_t is the time interval *p*.

Sliding Time Window (SW_t):

The special BW_t that can forward and operate in *p* cycle and form a control flow for series of BW_t.

$$SWt = \{(BWt)_k,\ 0< k < m\} \qquad \text{…… (3.2)}$$

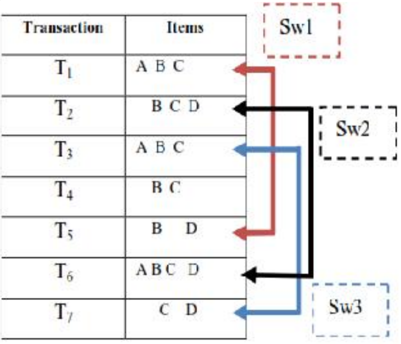
Where, k means the t moment, then *k+1* means the *t + p* moment. The size of SW_t is volume of data, which is captured in time interval of p, the length of BW_t and included a set of documents. In additional, the process of SW_t can become a continuous and infinite window stream.

Size of Basic Time Window (SBW_t):

Let *v* and *p* mean the velocity of data stream and the length of Basic Time Window, respectively. The SBW_t means that the total amount of data in one BW_t, which can be calculated as

$$SBW_t = p * v \qquad \text{…… (3.3)}$$

According to the above definition, let p=1(unit time) and v=1 (doc per unit time), then SBWt=1, which means that just one document flowed into the window in unit time interval. This process is become a real time continuous process classical Single-Pass process. Moreover, if p is fixed as a constant, the SWt also is a stable Sliding Time Window, but if p is variable, the SWt is sliding window with more complexity and flexibility.



Transaction	Items
T_1	A B C
T_2	B C D
T_3	A B C
T_4	B C
T_5	B D
T_6	A B C D
T_7	C D

Table.3.1 Transaction view Data Stream Sliding Window

3.3 ALGORITHM

Input : *Text Stream*
Output : *Sets of Cluster*

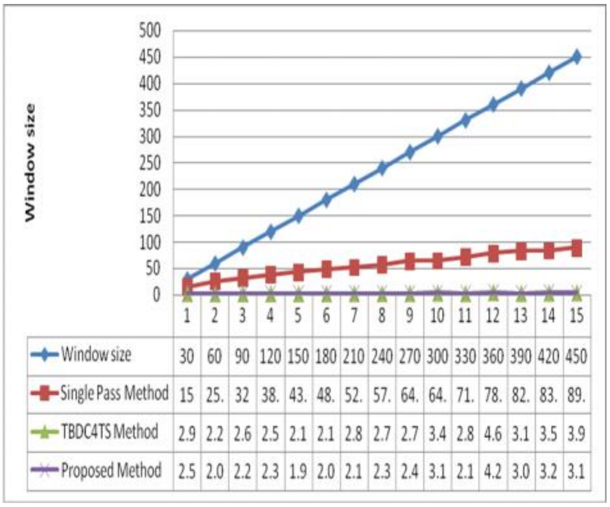
```
Procedure_ Sliding Time Window ( )
{
 Step1: Startup the Web Crawler to form a text stream.
Step2: To slide the k-th ($0<k\leq m$) time window (SWt)k with the
         length p by the module of SWS.
           For ($k=1, k\leq m, k++$)
           {
           }
Step3: Procedure_ Micro clustering ( )
Step4: Procedure_ Macro-clustering processing ( )
Step5: Repeat Next (SWt) k+1
Step6: Stop the process
}
```

```
Procedure_ Micro clustering ( )
{
Step1: Start a process with text stream.
Step2:  For ($i=1, i\leq n, i++$)
{
         doc Feature ik(Person, Address, Time, FT);
           //Mining i-th document and extracted feature of text,
            such as person's name, address, time//
Step3: build Index By Feature (Person, Address, Time);
Step4: DBSCAN({doc Feature ik} );
Step5: Return( {Mirco-clusters fk ,$0<f \leq n$} )
}
}
```

```
Procedure_ Macro clustering ( )
{
Step1: Start a process with micro cluster.
Step2:  FOR (f=1, f≤n, f ++)
{
  Step3:    Single-pass ({Mirco-clusters fk });
 Step4:    Return ({ClusterTn, 0<Tg≤k*n});
Step5:    Update the database with new clusters feature;
}
}
```

4. EXPRERIMENTATION & RESULTS

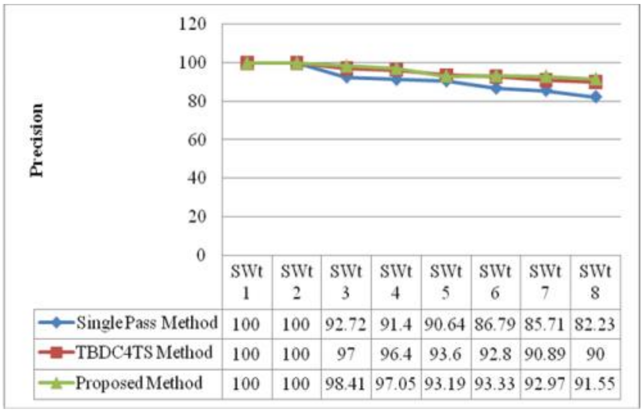
The proposed methodology is experimented with manually copied content to the text file from multiple websites, which is an unstructured data in the form of text represent the number of document for each category. In order to test the efficiency and precision the web page news can be taken from the website by web crawler to simulate the continuous text stream, which includes 7 topics and 2750 web pages as testing documents with a certain velocity. In the beginning phase, the parameters of system show the length, velocity and size of SWt are p, v and SBWt, respectively. The execution time are analysis with proposed algorithms to compare the execution time with variation of SBWt from 30 to450 stable interval of 30. From the experiments and results it shows Fig.4.1 proposed algorithm has better time executing efficiency than existing methods.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Window size	30	60	90	120	150	180	210	240	270	300	330	360	390	420	450
Single Pass Method	15	25.	32	38.	43.	48.	52.	57.	64.	64.	71.	78.	82.	83.	89.
TBDC4TS Method	2.9	2.2	2.6	2.5	2.1	2.1	2.8	2.7	2.7	3.4	2.8	4.6	3.1	3.5	3.9
Proposed Method	2.5	2.0	2.2	2.3	1.9	2.0	2.1	2.3	2.4	3.1	2.1	4.2	3.0	3.2	3.1

Fig.4.1 Comparison of execution time

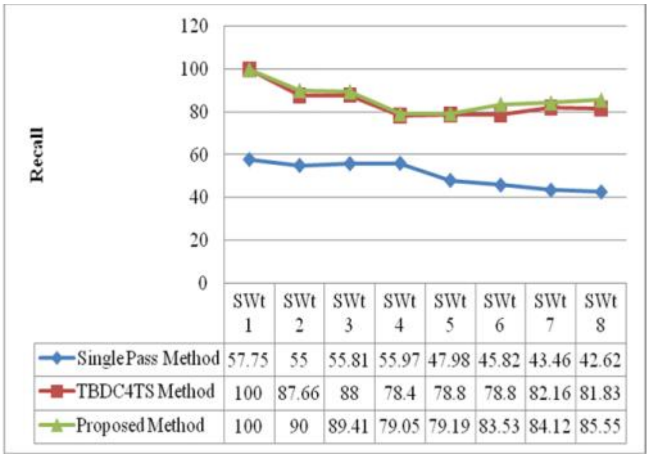
The experiment and results precision rate with existing methods are fig.4.2 with sliding of Time Window, the precision rate of both of two algorithms, which adopt four text feature vectors to describe the document, also decrease in a certain degree, but the precision of proposed method is still is higher than TBDC4TS and Single-pass. Especially, owning to batch processing in each window and multi-phase clustering, the influence by the sequence of text stream is avoided in TBDC4TS algorithm, but is still disturbed in Single-pass algorithm.



	SWt 1	SWt 2	SWt 3	SWt 4	SWt 5	SWt 6	SWt 7	SWt 8
Single Pass Method	100	100	92.72	91.4	90.64	86.79	85.71	82.23
TBDC4TS Method	100	100	97	96.4	93.6	92.8	90.89	90
Proposed Method	100	100	98.41	97.05	93.19	93.33	92.97	91.55

Fig.4.2 Comparison of Precision

The experiment and results of recall rate with existing methods are fig.4.3 proposed method has higher than TBDC4TS and Single-pass. The increasing of text documents, the new clusters in each of sliding time window (SWt) should be merged into existed clusters. If the density of existed clusters is sparser, which means the topics of text is lower interrelated and scattered distribution and then the recall rate in clustering also will be decreased. The experiments results in above mentioned shown that the proposed method higher efficiency and performance than TBDC4TS and Single-pass with multilevel clustering method.



	SWt 1	SWt 2	SWt 3	SWt 4	SWt 5	SWt 6	SWt 7	SWt 8
Single Pass Method	57.75	55	55.81	55.97	47.98	45.82	43.46	42.62
TBDC4TS Method	100	87.66	88	78.4	78.8	78.8	82.16	81.83
Proposed Method	100	90	89.41	79.05	79.19	83.53	84.12	85.55

Fig.4.3 Comparison of Recall

5. CONCLUSION

In this paper, a novel algorithm is proposed to modify the processing of text stream by sliding time window. When the web page news are captured continuously by web crawler and formed a text stream. The proposed sliding window algorithm to split the text stream into continuous segments, which merged the text data stream clustering. The every sliding time window, there be a set of web page news related to velocity of stream and size of sliding window. In addition, framework algorithm are designed to adopt a multilevel clustering method with sliding time window. The result shows 2750 web page news in websites to simulate text stream by web crawler. The proposed method provides better executing time efficiency, cluster quality, precision and recall than existing methods.

REFERENCES

[1] M.-L.Antonie and O.R.Zaiane, “Text document categorization by term association”, In Proc. of the IEEE 2002 “International Conference on Data Mining”, pp.19–26, Maebashi City, Japan, 2002.

[2] K.Androutsopoulos, K.V.Koutsias, Chandrinos, C.D. Spyropoulos, “An Experimental Comparison of Naïve Bayes and Keyword-based Anti-spam Filtering with Personal Email Message”, Proceedings of 23rd ACM SIGIR, pp.160-167, 2000.

[3] Huang Lei, Mining Stream Data: A Survey, Journal of Software (in Chinese), Vol. 15(1):Pp. 1-7, 2004.

[4] A. Forestiero, C. Pizzuti, G. Spezzano, A single pass algorithm for clustering evolving data streams based on swarm intelligence Data Mining Knowledge Discovery, Vol. 26:Pp.1-26, 2013.

[4] A. Arasu and G. Manku, Approximate counts and quartiles over sliding windows [C], the Processing of the 2004 ACM Symp. Principles of Database Systems, Pp.286-296, 2004.

[5] M. Oyamada, H. Kawashima, H. Kitagawa, Data Stream Processing with Concurrency Control, SIGAPP Applied Computing Review, Vol.13(2): p.54-64, 2013.

[6] C. Junghans, M. Karnstedt, M. Gertz, Quality-driven Resource-adaptive Data Stream Mining], SIGKDD Explorations Newsletter,Vol.13 (1): P.72-82, 2011.

[7] Zhu Hengmin, Zhu Weiwei, Study on Web Topic Online Clustering Approach Based on Single-pass Algorithm, New Technology of Library and Information Service(in Chinese), Vol.12, Pp.52-57, 2011.

[8] Yin Fengjing, Xiao Weidong, Ge Bin, etc., Incremental Algorithm for Clustering Texts in Internet-oriented Topic Detection, Application Research of Computers, Vol. 28(1): 249-252, 2011.

[9] Li Na, Xing Changzhen, Density-based Data Stream Clustering Algorithm over Time-based Sliding Window, Journal of Computer Applications, Vol. 31(5): 1363-1366, 2011.

[10] Shui Yidong, Qu Youli, Huang Houkuan, A New Topic Detection and Tracking Approach Combining Perodic Classification and Single-pass Clustering,, Journal of Beijing Jiaotong University (in Chinese), Vol.33 (5): Pp.85-89, 2009.

[11] E.Rasmussen, “Chapter 16:Clustering Algorithms”,in Frakes,W.B.Baeza-yates,R.(Eds),Information Retrieval:data structures & algorithms,(pp.419-442),Prentice Hall,1992.