ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Text Stream Classification Techniques and Research Issues: A Review

Abhinandan Vishwakarma*
Research scholar
Technocrats Institute of Technology
Bhopal, Madhya Pradesh, India
abhinandantit@gmail.com

Prof. Sini Shibu
HOD Department of Computer Science
Technocrats Institute of Technology
Bhopal, Madhya Pradesh, India
sinijoseph@hotmail.com

Abstract: With the rapid growth of applications that generates massive text streams, text stream classification is the key technique for handling and organizing text data. Text classification is the process of sorting documents containing text streams into one or more predefined categories. These streams are generated continuously and in a very high fluctuating data rates, example includes social networks, news collections, chat and e-mail etc. These streams are transient open end rather than persistent on disk. As text streams are dynamic, infinite in length, can't reproduce for processing, arrives at very high speed, text stream classification is more challenging as compare to static text classification. However most of the reported work are concentrated on structural data and seldom focus on unstructured data such as textual document. In this paper we present a foundation on text stream analysis, text stream classification, challenges in classification and identify the direction of future research.

Keywords: Text stream, data stream, text stream classification, data stream classification, pre-processing.

I. INTRODUCTION

Text classification (also known as text categorization or topic spotting) is the task of assigning a given text document to one or more predefined categories. This problem has received a special and increased attention from the researchers in the past few decades due to the many reason like the massive amount of digital and online documents that are easily accessible and the increased demand to organize and retrieve these document efficiently. Efficient text categorization systems are beneficial for many applications, for example, information retrieval, classification of news stories, text filtering, and categorization of incoming e-mail messages and classification of web pages.

Automated text classification is attractive because it frees the organization from the need of manually organizing document base, which can be too expensive, or simple not feasible given the time constraint of the application or the number of documents involved. A large number of machine learning, knowledge engineering, and probabilistic-based methods have been proposed for text classification can be obtained in [6,7,10,12,16,18,20]. The most popular methods include Bayesian probabilistic methods, regression models, example-based classification, decision trees, decision rules, Rocchio method, Neural networks, support vector machines (SVM), and association rules mining.

Recently a new class of emerging application has become widely recognized: application in which data is generated at very high rates in the form of transient data streams. Text Streams have become ubiquitous because of a wide variety of application in social networks, news collections, and other form of activity which results in the continuous creation of massive stream. Some specific examples of applications which create text stream are as follows:

a) In chat, e-mail and social network such as twitter, facebook etc. continuous communication between the actors (users) can generate massive amount of text stream.
b) Several news article may be received by the various news aggregator agencies, generates the huge volume of text stream which are more structured than the messages generated in social media.
c) Over the network web crawlers may collect huge volumes of documents in a small time frame which results in stream of documents.

Other example of data stream includes financial applications, Network monitoring, security, telecommunication data management, sensor networks and others. In such data intensive application the data is modeled as transient open end stream rather than persistent tables on disk. As the data stream are open end in nature and huge in volume, it is impossible to hold the entire data stream in memory for analysis and store on the disk as fast processing is required. With the advent of advanced data streaming technologies [1], we are now able to continuously collect large amounts of data in various application domains, e.g., daily fluctuations of stock market, traces of dynamic processes, credit card transactions, web click stream, network monitoring, position updates of moving objects in location-based services and text streams from news etc [2]. It is important to manage data stream promptly and effectively. For example in a email system researchers want to be able to automatically classify vast amount of incoming email into different categories like spam, personal email, business email etc.

The paper is organized as follows section II represents the theoretical background of data stream analysis. Text stream Classification technique and system are viewed in section III and IV respectively. Research issues are discussed in section V. Finally section VI summarizes this paper.

II. FOUNDATION

Research problems and challenges that have arisen in data stream can find its solution in statistical and computational approaches [24]. The online nature of the text data streams and their potentially high arrival rates impose high resource requirement on data stream processing system. These solutions can be categorized into data based or summarization and task-based ones. In data based approaches we either examine only the subset of the whole data set or transform the data vertically or horizontally to an approximate smaller size data representation. On the other side on task-based approaches, techniques from the computational theory have been adopted for time and space efficient solution.

A. Data Based Or Summarization Technique:

Summarization technique refers to the process of transforming data to a suitable form for stream data analysis.

Data based technique refer to summarizing the whole data set or choosing a subset of incoming stream to be analyzed. Three techniques which summarize the whole data set are sampling, load shading and sketching techniques. Synopsis data structure and aggregation represents choosing a subset of incoming stream. An excellent review of data reduction technique is presented in [11]. Basic of these technique with examples of their applications are as follows:

a. Sampling:

Sampling refers to the process of statistically selecting a data item data item to be processed or not [15]. Researchers have been working on sampling since the end of nineteenth century. Boundaries of the error rate of the computation are given as a function of sampling rate. Very fast Machine learning technique [13] have used hoeffding bounds to measure the sample size according to some derived loss function. The problem with using the sampling in the context of data stream is the unknown dataset size. Therefore to find the error bounds special treatment of data streams are required. Sampling also does not address the problem of fluctuating data rates. When using sampling it would be worth investigating the relationship among the three parameters data rate, sampling rate and error bound. Designing sampling-based algorithms that can produce approximate answers that are provably close to the exact answer is an important and active area of research.

b. Sketching:

Sketching [8,23] involves building a summary of a data stream using a small amount of memory. It is the process of vertically sample the incoming stream. Alon [5] introduced the notion of *randomized sketching* which has been widely used ever since. Techniques based on sketching are very convenient to distributed computation over multiple streams. The major drawback of sketching is that of accuracy. Principal Component Analysis (PCA) would be a better solution if being applied in streaming applications [21].

c. Load shedding:

Load shedding refers [1,26] to the process of dropping a sequence of data streams. Load shedding has been used successfully in querying data streams. It has the same problems of sampling. Load shedding is not preferred in time series analysis because it drops chunks of data streams that could be used in the structuring of the generated models or it might represent a pattern of interest in time series analysis.

d. Synopsis Data Structures:

Synopsis data structures are the idea of small space, approximate solution to massive data set problems. Wavelet analysis [19], histograms, quantities and frequency moments [1] have been proposed as synopsis data structures.

Wavelets [33] are one of the often-used techniques for providing a summary representation of the data. Wavelets coefficients are projections of the given signal (set of data values) onto an orthogonal set of basis vector.

Histograms [34,35] approximate the data in one or more attributes of a relation by grouping attribute values into "buckets" (subsets) and approximating true attribute values and their frequencies in the data based on a summary statistics maintained in each bucket.

e. Aggregation:

Aggregation is the process of computing statistical measures such as means and variance that summarize the incoming stream. The problem with aggregation is that it does not perform well with highly fluctuating data distributions. Aggregation has been successfully used in distributed stream data environments and with continuous queries over data streams [9].

B. Task Based Approach:

Task-based techniques are those methods that modify existing techniques or invent new ones in order to address the computational challenges of data stream processing. Sliding window is one of the most popular approaches which represent this category.

a. Sliding window:

The idea behind sliding window is to perform detailed analysis over the most recent data items and over summarized versions of the old ones. Imposing sliding windows on data streams is a natural method for approximation that has several attractive properties. It is well defined and easily understood. Babcock, Datar and Motwani [36] studied sampling in sliding window model. It is deterministic, so there is no danger that unfortunate random choices will produce a bad approximation. Most importantly, it emphasizes recent data, which in the majority of real-world applications is more important and relevant than old data. This idea has been adopted in many techniques in the undergoing comprehensive data stream mining system *MAIDS* [14].

III. TEXT STREAM CLASSIFICATION

Text Stream classification is a kind of text classification with some unique characteristic as follows:

a) The order in which data elements arrived to be process is not fixed.
b) Because of the concept drift in the coming text stream it is require to continuously model maintenance in order to have the quality classification of stream.

c) It is difficult to collect the negative training data for classifier because once the data stream has been processed it is discarded.
d) Only limited memory space is provided to rebuild the classifier and operate the system. Updating and operating must be conducted very fast.

Classification of data stream mining is a challenging area of research. There are many problems to be solved, such as handling continuous attributes, Concept drift, Sample taken question, Classification accuracy problem, Data stream management and pretreatment of data stream.

A. Handling Continuous Attributes:

When classification text stream face the real time and memory limit, the reach of how to compute the evaluation function more quickly, how to more effectively compressed storage properties deserve further studies.

B. Concept Drifts:

Concept drift occur in the text data stream when the target concept in a classification problem changes over time. In the real world classification problem the concept being modeled is not static but rather changes over time – a situation known as concept drift. Mining concept drifts from text data stream is one of the most important field in data mining. The reach of how more rapidly and accurately judge concept drift, how to effectively use concept drift acquisition, save and heavy use concept, and the trend of concept drift need to serious study.

C. Sample Sampling:

Although there is Hoeffiding inequality of sampling method, how to get better with less precision of the sample, remains a problem worthy to study.

D. Classification Accuracy:

High classification accuracy is the goal of all classification algorithms. How to improve the classification accuracy is very important research.

E. Data Stream Management:

Traditional database technology has already promoted the development of information technology, but traditional technology look powerless to data stream.

F. Pretreatment of Data Stream:

Pretreatment of data stream also need to consider. The reach of how to design a lightweight preprocessing algorithm to guarantee the quality of mining result is very important. The pretreatment of Data Stream occupies most if the running time, and how to decrease the running time is also important.

G. Re- use of Traditional Classification Methods:

Traditional classification is Decision rule, Bayesian classification, Back propagation method, related classification, K nearest neighbor classifier, Example based reasoning, Evolutionary algorithm, Rough set method, and fuzzy set and so on. The current study was applied some of these method to data stream. How to use the characteristic of the data stream for the application of these methods will be very valuable. Continuous arrival of text stream makes it different from the static text data so it is difficult to use the existing technique of text categorization for the text stream categorization.

IV. METHOD OF CLASSIFICATION

Existing text classification techniques can be grouped into three categories as follows:

i) Supervised Learning
ii) Semi-Supervised Learning
iii) Unsupervised Learning

In supervised Learning a set of (often manually) labeled training documents of every class is used by a learning algorithm to build a classifier. Existing text classification technique includes the naïve Bayesian method [4], Support Vector Machine [3] and many others.

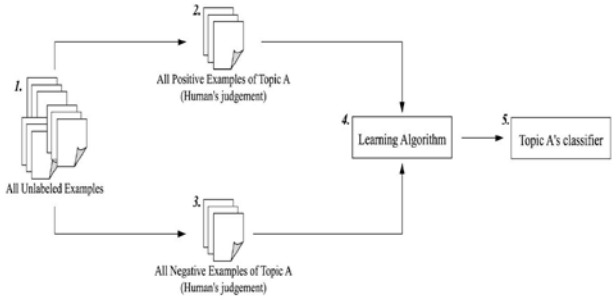

Figure 1:supervised learning technique

Due to the problem of manually labeling, semi-supervised learning is studied, which includes two main paradigms

a) Learning from a small set of labeled examples and a large set of unlabeled examples and
b) Learning from positive and unlabeled example (with no labeled negative examples)

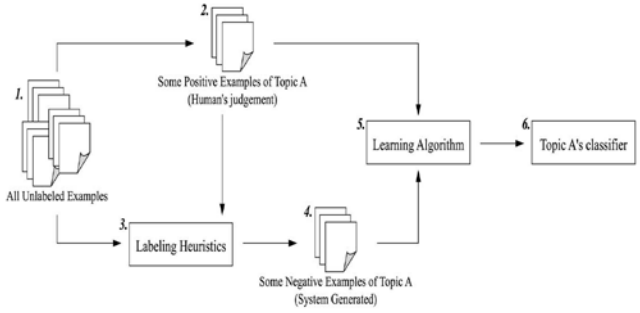

Figure 2:semi-supervised learning technique

Many researches have studied learning first paradigm can be obtained in [27-29]. In learning from positive and unlabeled example, some theoretical studies and practical algorithm are reported in [30,31].

Compared with the traditional data collection, the data stream is a real time, continuous, orderly, time varying and infinite tuple. A data stream has the following distinctive features a) Orderly, b) can't reproduce, c) high speed, d) infinite, e) high dimensional, f) dynamic.

So the traditional text classification methods are not applicable with text streams. Classification of the text stream is much more difficult as compare to other data stream because of the lack of the structure of the text data.

The concept drifting problem in text data stream classification has been addressed by several authors. Wang [25] have proposed a general framework for classification of concept drifting data stream. The proposed technique uses weighted classifier ensembles to mine data stream. The proposed algorithm combines multiple classifiers weighted by their expected prediction accuracy. Also the selection of number of classifiers instead of using all is an option in the proposed framework without loosing accuracy in the classification process. Ganti [17] have developed analytically two algorithms GEMM and FOCUS for model maintenance and change detection between two data sets in terms of the data mining results they induce. The algorithms have been applied to decision tree models and the frequent item set model. GEMM algorithm accepts a class of models and an incremental model maintenance algorithm for the unrestricted window option, and outputs a model maintenance algorithm for both window-independent and window dependent block selection sequence. FOCUS framework uses the difference between data mining models as the deviation in data sets. Last [22] has proposed an online classification system that can adapt to concept drift. The system rebuilds the classification model with the most recent examples. Using the error rate as a guide to concept drift, the frequency of model building and the window size are adjusted. The system uses info-fuzzy techniques for model building and information theory to calculate the window size. Aggarwal [5] have presented a different view on the data stream classification problem from the perspective of a dynamic approach, in which simultaneous training and testing streams are used for dynamic classification of data sets.

V. RESEARCH ISSUES

Data stream classification is the stimulating field of study that has raised many challenges and issues that need to be addressed by the machine learning and data mining communities. The characteristic of text data mining pointed out in section I indicates that when developing a technique of this kind, there are more issues need to be considered. The following is a brief discussion of some crucial open research issues in data stream classification:

A. Memory Management:

The first fundamental issue we need to consider is how to optimize the memory space consumed by the mining algorithm. Memory management is a particular challenge when processing streams because many real data streams are irregular in their rate of arrival, exhibiting burst ness and variation of data arrival rate over time. A stream mining algorithm with high memory cost will have difficulty being applied in many situations, such as sensor networks. More research needs to be done in developing new summarization techniques for collecting valuable information from data streams. Fully addressing this issue in the mining algorithm can greatly improve its performance [32].

B. Model Overfitting:

The overfitting problem in data stream has not been addressed so far in the literature. Using some techniques such as cross validation is very costly in the case of data streams. Novel techniques are required to avoid model overfitting.

C. Compact Data Structure:

Due to bounded memory size and the huge amount of data streams coming continuously, efficient and compact data structure is needed to store, update and retrieve the collected information. Failure in developing such a data structure will largely decrease the efficiency of the mining algorithm. Even if we store the information in disks, the additional I/O operations will increase the processing time. Incremental maintaining of the data structure is a necessity since it is not possible to rescan the entire input. Also, novel indexing, storage and querying techniques are required to handle the continual fluctuated flow of information streams.

VI. SUMMARY

The spreading of data stream phenomenon in real life application has influenced in great manner the development of mining algorithm. Classifying data stream has raised number of challenges for the data mining community. The proposed techniques have their roots in statistic and theoretical computer science. Data based and task based techniques are the two categories of algorithms. Based on these two categories number of classification techniques have been developed. Systems have been implemented to use these techniques in real applications.

Classification of data stream is still in immature, growing field of study. There are many open issues that need to be addressed. The development of systems that will fully address these issues is crucial for accelerating the science discovery in the fields of business and financial applications [21]. This would improve the real time decision making process in almost every area of life.

VII. REFERENCES

[1]. S. B. B. Babcock, M. Datar, R. Motwani, and J. Widom, *Models and issues in Data Stream Systems*, PODS, 2002.

[2]. H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, Querying and Mining of time Series Data: Experimental Comparison of Representations and Distance Measures, VLDB, 2008.

[3]. Vapnik, V. *The nature of statistical learning theory*, 1995.

[4]. Lewis, D., and Gale, W. (1994). A sequential algorithm for training text classifiers. *SIGIR-94*.

[5]. N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proc. of the 1996 Annual ACM Symp. on Theory of Computing*, pages 20–29, 1996.

[6]. I. Dhillon, S. Mallela, and R. Kumar, "A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification," J. Machine Learning Research, vol. 3, 2003.

[7]. S.T. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive Learning Algorithms and Representations for Text

Categorization," Proc. Seventh Int'l Conf. Information and Knowledge Management, 1998.

[8]. Jon Kleinberg. Bursty and hierarchical structure in streams. In KDD, pages 91–101, 2002.

[9]. S. Babu, and J. Widom Continuous queries over data streams. SIGMOD Record, 30:109-120, 2001.

[10]. G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," J. Machine Learning Research, vol. 3, 2003.

[11]. D. Barbara et al. The New Jersey data reduction report. Bull. Technical Committee on Data Engineering, 20:3-45, Dec. 1997.

[12]. T. Joachims, "A Statistical Learning Model of Text Classification with Support Vector Machines," Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, 2001.

[13]. P. Domingos and G. Hulten, A General Method for Scaling Up Machine Learning Algorithms and its Application to Clustering, Proceedings of the Eighteenth International Conference on Machine Learning, 2001, Williamstown, MA, Morgan Kaufmann

[14]. G. Dong, J. Han, L.V.S. Lakshmanan, J. Pei, H. Wang and P.S. Yu. Online mining of changes from data streams: Research problems and preliminary results, In Proceedings of the 2003 ACM SIGMOD Workshop on Management and Processing of Data Streams. In cooperation with the 2003 ACM-SIGMOD International Conference on Management of Data, San Diego, CA, June 8, 2003.

[15]. P. Domingos and G. Hulten, A General Method for Scaling Up Machine Learning Algorithms and its Application to Clustering, Proceedings of the Eighteenth International Conference on Machine Learning, 2001, Williamstown, MA,

[16]. F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[17]. V. Ganti, J. Gehrke, and R. Ramakrishnan: Mining Data Streams under Block Evolution. SIGKDD Explorations 3(2), 2002.

[18]. N. Tishby, F.C. Pereira, and W. Bialek, "The Information Bottleneck Method," Proc. 37th Ann. Allerton Conf. Comm., Control, and Computing, 1999.

[19]. A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, M. Strauss: One-Pass Wavelet Decompositions of Data Streams. TKDE 15(3), 2003

[20]. Y. Yang and J.O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proc. 14th Int'l Conf. Machine Learning (ICML '97), pp. 412-420, 1997.

[21]. H. Kargupta et al. VEDAS: A Mobile and Distributed Data Stream Mining System for Real-Time Vehicle Monitoring, Proceedings of SIAM International Conference on Data Mining, 2004.

[22]. M. Last, Online Classification of Nonstationary Data Streams, Intelligent Data Analysis, Vol. 6, No. 2, pp. 129-147, 2002.

[23]. G. S. Manku and R. Motwani. Approximate frequency counts over data streams. In Proceedings of the 28th International Conference on Very Large Data Bases, Hong Kong, China, August 2002.

[24]. S. Muthukrishnan, Data streams: algorithms and applications. Proceedings of the fourteenth annual ACM-SIAM symposium on discrete algorithms 2003.

[25]. H. Wang, W. Fan, P. Yu and J. Han; Mining Concept-Drifting Data Streams using Ensemble Classifiers, in the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Aug. 2003, Washington DC,USA.

[26]. N. Tatbul, U. Cetintemel, S. Zdonik, M. Cherniack, M. Stonebraker. Load Shedding on Data Streams, In Proceedings of the Workshop on Management and Processing of Data Streams, San Diego, CA, USA, June 8, 2003.

[27]. Nigam, K., McCallum, A., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39.

[28]. Ghani, R. (2002). Combining labeled and unlabeled data for multiclass text categorization. *ICML-02*.

[29]. Goldman, S. and Zhou, Y. (2000). Enhancing supervised learning with unlabeled data. *ICML-00*.

[30]. Denis, F. (1998). PAC learning from positive statistical queries. *ALT-1998*.

[31]. Liu, B., Lee, W. S., Yu, P., and Li, X. (2002). Partially supervised classification of text documents. *ICML-02*.

[32]. L. Golab and M. T. Ozsu. Issues in Data Stream Management. In SIGMOD Record, Volume 32, Number 2, June 2003.

[33]. Y. Matias, J. Vitter, and M. Wang. Wavelet-based histograms for selectivity estimation. In *Proc. of the 1998 ACM SIGMOD Intl. Conf. on Management of Data*, pages 448–459, June 1998.

[34]. Y. E. Ioannidis and V. Poosala. Histogram-based approximation of set-valued query-answers. In *Proc. of the 1999 Intl. Conf. on Very Large Data Bases*, pages 174–185, Sept. 1999.

[35]. V. Poosala and V. Ganti. Fast approximate answers to aggregate queries on a data cube. In *Proc. of the 1999 Intl. Conf. on Scientific and Statistical Database Management*, pages 24–33, July 1999.

[36]. B. Babcock, M. Datar, and R. Motwani. "Sampling from a Moving Window over Streaming Data." In Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2002, pages 633–634.

