Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 12 (2014) 255 – 263

Procedia Technology

The 7th International Conference Interdisciplinarity in Engineering (INTER-ENG 2013)

Data mining – past, present and future – a typical survey on data streams

M.S.B. PhridviRaj[a,]*, C.V. GuruRao[b]

[a]Department of CSE, Kakatiya Institute of Technology and Science, Warangal, INDIA
[b] Department of CSE, S.R. Engineering College (Autonomous), Hasanparthy, Warangal,INDIA

Abstract

Data Stream Mining is one of the area gaining lot of practical significance and is progressing at a brisk pace with new methods, methodologies and findings in various applications related to medicine, computer science, bioinformatics and stock market prediction, weather forecast, text, audio and video processing to name a few. Data happens to be the key concern in data mining. With the huge online data generated from several sensors, Internet Relay Chats, Twitter, Face book, Online Bank or ATM Transactions, the concept of dynamically changing data is becoming a key challenge, what we call as data streams. In this paper, we give the algorithm for finding frequent patterns from data streams with a case study and identify the research issues in handling data streams.

Keywords: Clustering; Streams; Mining; Dimensionality reduction; Text stream; Data streams

1. Introduction

Data mining is a process of discovering hidden patterns and information from the existing data. The difference between data in the databases and a data warehouse is in a database the data is in the structured form where as in the

* Corresponding author. Tel.: +9030076521.
E-mail address: prudviraj.kits@gmail.com

doi:10.1016/j.protcy.2013.12.483

data warehouse the data may or may not be present in the structured format. The structure of the data may be defined to make it compatible for processing. Hence in data mining; we also need to primarily concentrate on cleansing the data so as to make it feasible for further processing. The process of cleansing the data is also called as noise elimination or noise reduction or feature elimination. The process of cleansing data can be either made by using tools such as ETL, tools available in the market or may be done by using various suitable techniques available. The important aspect for consideration in data mining is whether the data considered is static or dynamic. Handling static data is comparatively much easier to handling dynamically varying data. In the case of a static dataset, the entire data is available for analysis purpose in hand before processing and is generally not a time varying data. However dynamic data refers to high voluminous continuously varying information which is not a stand still data and also is not at the hand for processing or analyzing.

Data mining requires an algorithm or method to analyze the data of interest. Data may be a sequence data, sequential data, time series, temporal, spatio- temporal, audio signal, video signal to name a few. The concept of data streams has gained a lot of practical interest in the field of data mining. A data stream is an infinite sequence of data points defined usually either using time stamps or an index. We may also view data in the data streams as equivalent to a multidimensional vector containing integer, categorical, graphical with the data in structured or unstructured format. If the data is not structured we may have to transform in to a suitable format for processing by the algorithm being used. With the very high voluminous structured or unstructured continuous data being generated from various applications and devices, the concept of data is no more static but is turning out to be dynamic. This brings a lot of challenges in analyzing the data. Traditional data mining algorithms are not suitable for handling data streams because the algorithms designed perform multiple scans over the data which is not possible when handling the data streams. This brings actual challenge before the data mining researchers working in the area of data streams.

Further, Many of the existing data mining algorithms available for clustering, classification and finding frequent pattern in the literature are suitable for only static data sets and are no more practically suitable for handling data streams or for mining the stream data. Data streams may be time series or temporal or spatio temporal. The concept of clustering and classification is widely used and turned out as a choice of typical interest among the current data mining researchers. Section 2 discusses various related works in detail. In Section 3, we discuss various research issues in data mining and problems in handling data streams. We conclude the survey in section 4 finally.

2. Related works

In case of data streams, the number of distinct features or items that exist would be so large which makes even the amount of on cache memory or system memory available not suitable for storing the entire stream data. The main problem with data streams is the speed at which the data streams arrive is comparatively much faster than the rate at which the data can be stored and processed.

In the ACM KDD International conference held in 2010, the authors discuss the problem of finding the top-k frequent items in a data stream with flexible sliding widows [3]. The idea is to mine only the top-k frequent items instead of reporting all the frequent items. But the crucial factor or limitation that evolves here is the amount of memory that is required still for mining w.r.t to finding of top-k frequent items is still a bounding factor. The authors finally discusses that there exists however a memory efficient algorithms by making some assumptions.

In [2] the authors focus on developing a framework for classifying dynamically evolving data streams by considering the training and test streams for dynamic classification of datasets. The objective is to develop a classification system in which a training system can adapt to quick changes of the underlying data stream.

The amount of memory available for mining stream data using one pass algorithms is very less and hence there is chance for data loss. Also it is not possible to mine the data online as and when it appears because of mismatch in speed and several other significant factors.

In [4] the authors discuss the method of finding most frequent items by using a hash based approach. The idea is to use say ‘h’ hash functions and build the hash table by using linear congruencies. Data streams can be classified into two types as 1. Offline data streams and 2. Online data streams.

In [6] the method of singular valued decomposition is used to find the correlation between multiple streams. The concept of SVD was particularly used to find offline data streams. Clustering text data streams is one of the topics which have evolved as important challenge for data mining researchers. The problem of spam detection, email