RESEARCH ARTICLE OPEN ACCESS

An Efficient For Selective Data Mining Algorithm for Big Data Analytics Hadoop

Ms.Jennifa Angelin Robinson[1], Mr. S. Rajeshkumar, M.E[2].,
[1]*M.E., Student, *Department of Computer Science and Engineering, Annai Vailankanni college of Engineering, Kanyakumari.*
[2]*Asst.Prof *Department of Computer Science and Engineering, Annai Vailankanni college of Engineering, Kanyakumari.*

ABSTRACT

Big Data is a large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seek to explore complex and evolving relationships among data. Big data processing challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible. In this system, we propose a series of novel data mining mechanism that can be used for efficient data extraction from the data set. HACE theorem is used to reveal the characteristics based on which proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations.

***Keywords:* -** Data mining; frequent pattern mining; graph streams, Big data.

I. INTRODUCTION

Recent years have witnessed a dramatic increase in our ability to collect data from various sensors, devices, in different formats, from independent or connected applications. This data flood has outpaced our capability to process, analyze, store and understand these datasets. Consider the Internet data. The web pages indexed by Google were around one million in 1998, but quickly reached 1 billion in 2000 and have already exceeded 1 trillion in 2008. This rapid expansion is accelerated by the dramatic increase in acceptance of social networking applications, such as Face book, Twitter, etc., that allow users to create contents freely and amplify the already huge Web volume. Furthermore, with mobile phones becoming the sensory gateway to get real time data on people from different aspects, the vast amount of data that mobile carrier can potentially process to improve our daily life has significantly outpaced our past CDR (call data record)-based processing for billing purposes only. It can be foreseen that Internet of things (IoT) applications will raise the scale of data to an unprecedented level. People and devices are all loosely connected.

Big data is defined as large amount of data which requires new technologies and architectures to make possible to extract value from it by capturing and analysis process. New sources of big data include location specific data arising from traffic management, and from the tracking of personal devices such as Smart phones. Big Data has emerged because we are living in a society which makes increasing use of data intensive technologies. Due to such large size of data it becomes very difficult to perform effective analysis using the existing traditional techniques. Since Big data is a recent upcoming technology in the market

which can bring huge benefits to the business organizations, it becomes necessary that various challenges and issues associated in bringing and adapting to this technology are need to be understood. Big Data concept means a datasets which continues to grow so much that it becomes difficult to manage it using existing database management concepts & tools. The difficulties can be related to data capture, storage, search, sharing, analytics and visualization etc.

II. BIG DATA CHARACTERISTICS

Data Volume: The Big word in Big data itself defines the volume. At present the data existing is in peta bytes (10,s) and is supposed to increase lo zetta bytes (I021) in **nearby** future. Data volume measures the amount of data available to an organization, which does not necessarily have to own all
of it as long as it can access it.

Data Velocity: Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data hut also speed at which the data flows and aggregated.

Data Variety: Data variety is a measure of the richness of the data representation - text, images video, audio, etc. Data being produced is not of single category as it not only
includes the traditional data but also the
semi structured data from various resources like web Pages. Web Log Files, social media sites, e-mail, documents.

Data Value: Data value measures the
usefulness of data in making decisions. Data science is exploratory and useful in getting to know the data, but "analytic science" encompasses the predictive power of big data. User can run certain queries against the data stored and thus can deduct important
results from the filtered data obtained and can also rank it according to the dimensions they require. These reports help these people to find the business trends according to which they can change their strategies.

Complexity: Complexity measures the
degree of interconnections (possibly very large) and interdependence in big data structures such that a small change (or combination of small changes) in one or a few elements can yield very large changes or a small change that ripple across or cascade through the system and substantially affect its behavior, or no change at all.

III. ISSUES IN BIG DATA

The issues in Big Data arc some of the conceptual points that should be understood by the organization to implement the technology effectively. Big data Issues ate need not be confused with problems but they are important to know and crucial to handle.

Issues related to the Characteristics

Data Volume As data volume increases, the value of different data records will decrease in proportion to age. Type, richness, and quantity among other factors. The social networking sites existing are themselves producing data in order of terabytes every day and this amount of data is definitely difficult to be handled using the existing traditional systems.

Data Velocity our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion. Ecommerce has rapidly increased the speed and richness of data used for different business transactions (for example, web-site clicks). Data velocity management is much more than a bandwidth issue. Data Variety All this data is totally different consisting of raw, structured, semi structured and even unstructured data which is difficult to be handled by the existing traditional analytic systems, from an analytic perspective; it is probably the biggest obstacle to analytic sprawl. Data Value As the data stored by different organizations is being used by them for data analytics. It will produce a kind of gap in between the Business leaders and the IT professionals the main concern of business leaders would be to just adding value to their business and getting more and more profit unlike the IT leaders who would have to concern with the technicalities of the storage and processing. Data Complexity One current difficulty of big data is working with it using relational databases and desktop statistics packages, requiring massively parallel software running on tens, hundreds, or even thousands of servers. It is quite an undertaking to link, match, cleanse and transform data across systems coming from various sources. It is also necessary to connect and correlate relationships, hierarchies and multiple data linkages or data can quickly spiral out of control.

Storage and Transport Issues

The quantity of data has exploded each time we have invented a new storage medium. The difference about the most recent data explosion, mainly due to social media, is that there has been no new storage medium. Moreover, data is being created by everyone and everything. Current disk technology limits arc about 4 terabytes (10i:) per disk. So, I Exabyte (I0IK) would require 25.000 disks. Even if an Exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks. Access to that data would overwhelm current communication networks. Assuming that a I gigabyte per second network has an effective sustainable transfer rate of 80% the sustainable bandwidth is about 100 megabytes.

Thus, transferring an Exabyte would take about 2800 hours, if we assume that a sustained transfer could be maintained. It would take longer time to transmit the data from a collection or storage point to a processing point than the time required to actually process it.

IV. CHALLENGES IN BIG DATA

Privacy and Security

It is the most important challenges with big data which is sensitive and includes conceptual technical as well as legal significance. Information regarding the people is collected and used in order to add value to the business of the organization. Another important consequence arising would be Social tratification where a literate person would be taking advantages of the Big data predictive analysis and on the other hand underprivileged will be easily identified and treated worse. The scalability issue of Big data has lead towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters. This requires a high level of sharing of resources which is expensive and also brings with it various challenges like how to run and execute various jobs so that we can meet the goal of each workload cost effectively. It also requires dealing with the system failures in an efficient manner which occurs more frequently if operating on large clusters. Collection of huge amount of data and its storage comes at a cost.

V. DENSE GRAPH STREAMS

Streams of data can be produced in many applications such as social networks, sensor networks, bioinformatics, and chemical informatics. These kinds of streaming data share a property in common—namely, they can be modeled in terms of graph-structured data. Here, the data streams generated by graph data sources in these applications are graph streams. To extract implicit, previously unknown, and potentially useful frequent patterns from these streams, efficient data mining algorithms are in demand. Many existing algorithms capture important streaming data and assume that the captured data can fit into main memory. However, problems arise when such as vertical mining algorithm makes good use of the information captured in the DSMatrix
for mining.

VI. CONCLUSION

As technology advances, streams of data (including graph streams) are produced in many applications. Key contributions of this paper include a simple yet powerful alternative disk-based structure—called *DSMatrix*—for efficient frequent pattern mining from streams (e.g., dense graph streams) with limited memory, tree-based frequent pattern mining algorithms, and an effective frequency counting technique, which avoids keeping too many FP-trees in memory when the space is limited. Such a technique requires only one FP-tree for a projected database to be kept in the limited memory. Analytical and experimental results show the benefits of our DSMatrix structure and its corresponding mining algorithms.

REFERENCE

[1]. Aggarwal CC. On classification of graph streams. In: Proceedings of the SDM 2011. SIAM; 2011, p. 652–663.

[2]. Aggarwal CC, Li Y, Yu PS, Jin R. On dense pattern mining in graph streams.PVLDB2010;3 (1–2):975–984.

[3]. Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. In: Proceedings of the VLDB 1994. Morgan Kaufmann; 1994,p. 487–499.

[4]. Bifet A, Holmes G, Pfahringer B, Gavald` a R. Mining frequent closed graphs on evolving data streams. In: Proceedings of the ACM KDD 2011. ACM; 2011, p. 591–599.

[5]. Buehrer G, Parthasarathy S, Ghoting A. Out-of-core frequent pattern mining on a commodity. In: Proceedings of the ACM KDD 2006. ACM; 2006,p. 86–95.

[6]. Cameron JJ, Cuzzocrea A, Jiang F, Leung CK. Frequent pattern mining from dense graph streams. In: Proceedings of the EDBT/ICDT Workshops 2014.CEUR-WS.org; 2014, p. 240–247.

[7]. Cameron JJ, Cuzzocrea A, Leung CK. Stream mining of frequent sets with limited memory. In: Proceedings of the ACM SAC 2013. ACM; 2013,p. 173–175.

[8]. Cameron JJ, Leung CK, Tanbeer SK. Finding strong groups of friends among friends in social networks. In: Proceedings of the IEEE DASC (SCA)2011. IEEE; 2011, p. 824–831.

[9]. CaoL, Yang D, Wang Q, Yu Y, Wang J, Rundensteiner EA. scalable distance-based outlier detection over high-volume data streams. In: Proceedings of the IEEE ICDE 2014. IEEE; 2014, p. 76–87.

[10]. ChiL, Li B, Zhu X. Fast graph stream classification using discriminative clique hashing. In: Proceedings of the PAKDD 2013, Part I. Springer;2013,p. 225–236.