A STUDY OF TRADITIONAL DATA ANALYSIS AND SENSOR DATA ANALYTICS

Kavita Ahuja[1] and N.N.Jani[2]

[1]Shree Madhav Institute of Computer & Information Technology, Surat, Gujarat, India
[2]Ex-Director, SKPIMCS - MCA, Gandhinagar, Gujarat, India

ABSTRACT

The growth of smart and intelligent devices known as sensors generate large amount of data. These generated data over a time span takes such a large volume which is designated as big data. The data structure of repository holds unstructured data. The traditional data analytics methods well developed and used widely to analyze structured data and to limit extend the semi-structured data which involves additional processing over heads. The similar methods used to analyze unstructured data are different because of distributed computing approach where as there is a possibility of centralized processing in case of structured and semi-structured data. The under taken work is confined to analysis of both verities of methods. The result of this study is targeted to introduce methods available to analyze big data.

KEYWORDS

Sensor data, big data, data analysis, data analytics

1. INTRODUCTION

The technological growth has rapidly influenced automation of devices but now this automated system is being made intelligent not only for purpose of operation but also for effective control. This initiative has opened up the window of smart devices, smart homes, smart hospitals, smart agriculture, smart citizen services, smart medical services and many more to improve their quality of life and services. The core at this initiative requires instant data acquisition, immediate processing and storage for analysis - dynamic as well as futuristic, requires deployment of wireless sensors of diversified types for data capture and high performance processing, communication and storage on cloud infrastructure. This will help to integrate all related areas and give analytics based decision support as well as intelligent control. Sensor data pose challenges with respect to data acquisition, storage and efficient and real-time processing of massive volumes [1] of possibly unstructured data resulting in big data. Therefore, analytics of sensor data raise a set of issue with respect to traditional methods of data analysis.

According to the estimation of Steve Lever [2], the growth of data is more than 50% a year, estimated by IDC, a technology research firm. There are now countless digital sensors worldwide in the devices used in areas such as agriculture, industries, automobiles, defense, and shipping transport etc. These devices can trap sensors data based on geographical location, movement, vibration, temperature, pressure, environmental changes, chemical changes and many more.

DOI : 10.5121/ijist.2016.6219

Although the advances of computer systems and internet technologies have witnessed the development of computing hardware following the Moore's law for several decades, the problems of handling the large-scale data still exist when we are entering the age of wireless sensor data which result in big data. That is why Fisher et al. [3] pointed out that big data means that the data is unable to be handled and processed by most current information systems or methods because data in the big data era will not only become too big to be loaded into a single machine, it also implies that most traditional data mining methods or data analytics developed for a centralized data analysis process may not be able to be applied directly to big data.

In response to the problems of analyzing large-scale data, quite a few efficient methods[4], such as sampling, data condensation, density-based approaches, grid-based approaches, divide and conquer, incremental learning, and distributed computing, have been presented. Of course, these methods are constantly used to improve the performance of the operators of data analytics process. The results of these methods illustrate that with the efficient methods at hand, the large

–scale data may be analyzed in a reasonable time.

To make the whole process of knowledge discovery in databases (KDD) more clear, Fayyad et al [5] summarized the KDD process by a few operations, which are selection, preprocessing, transformation, data mining, and interpretation/evaluation. These operators will be able to build a complete data analytics system to gather data first as Data Input, than Data Analytics and gives the result/output to the user.

Because the traditional data analysis methods are not designed for large-scale and complex data, they are almost impossible to be capable of analyzing the big data. Redesigning and changing the way the data analysis methods are designed are two critical trends for big data analysis. Several important concepts in the design of the big data analysis method will be given in the following sections.

Data streams are generated at multiple Distributed computing sensor nodes, which requires data mining approaches to optimized communication cost across different nodes, also the cost of computation, storage requirements at each node. Charu Aggrawal [6],draw attention to distributed computing network, which contains very large amount of sensor nodes, the aggregation of streams of data of every node become quite challenging task, also it pose challenges to mining problems, because the result of one node's mining algorithm must be integrate across the different nodes of the network.

Further, according to Russom [7], the data that need to be analyzed are not just large, but they are composed of various data types, and even including streaming data. Since big data has the unique features of "massive, high dimensional, heterogeneous, complex, unstructured, incomplete, noisy, and erroneous," which may change the statistical and data analysis approaches [8]. Although it seems that big data makes it possible for us to collect more data to find more useful information, the truth is that more data do not necessarily mean more useful information. It may contain more ambiguous or abnormal data. For instance, a user may have multiple accounts, or an account may be used by multiple users, which may degrade the accuracy of the mining results [9]. Sensor data brings numerous challenges with it in the context of data collection, storage and processing. This is because sensor data processing often requires

efficient and real-time processing from massive volumes of possibly uncertain data. Therefore, Charu Aggrawal [10] introduced several new issues of sensor data processing such as Data collection, Sensor Mining, Application-Specific issue, security, storage, and quality of data.

Different from traditional data analytics, Baraniuk [11] pointed out that the bottleneck of big data analytics will be shifted from sensor to processing, communications, storage of sensing data for its better management and maintenance in one single machine, as shown in Figure 1.

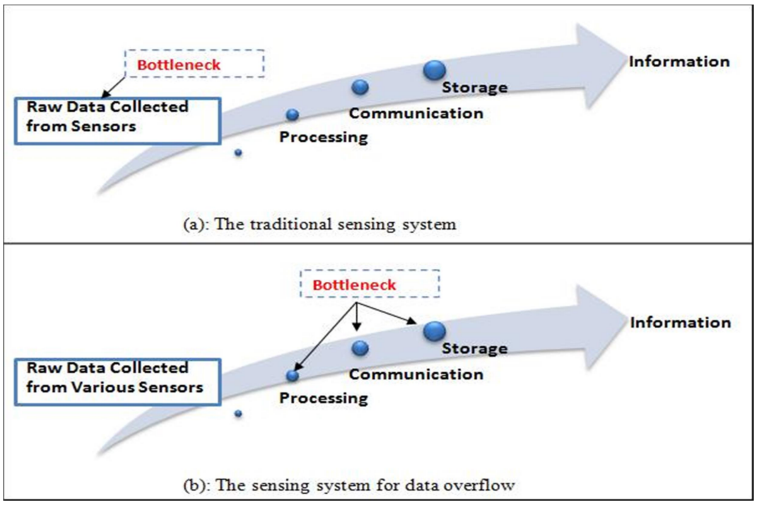

Figure 1. Difference of Traditional data and big data analysis on wireless sensor network

2. COMPARATIVE ANALYSIS

Many efficient analysis mining methods exist to analyze the data with respect to structure and semi-structure format, and those are as given in Table 1:

Table 1: Efficient data analytics methods for data mining

Problem	Method
Clustering	BIRCH
	DBSCAN
	RKM
	TKM
Classification	SLIQ
	TLAESA
	FastNN
	SFFS
Association Rules	CLOSET
	FP-tree
	CHARM
	MAFIA
	FAST
Sequential patterns	SPADE
	CloSpan
	PrefixSpan
	SPAM

Most data analysis methods have limitations for big data analysis with regards to Unscalability and centralization, Non-dynamic and they have uniform data structure. Because the traditional data analysis methods are not designed for large-scale and complex data, they are almost impossible to be capable of analyzing the big data.

Several studies attempted to present an efficient or effective solution from the perspective of system (e.g., framework and platform) or algorithm level. A simple comparison of these big data analysis technologies from different perspectives is described in Table 2. The "Perspective" column of this table explains that the study is focused on the framework or algorithm level; the

"Description" column gives the further goal of the study.

Table 2: big data analysis framework and methods

Perspective	Name	Description
Analysis framework	DOT	Add more computation resources via scale out solution
	GLADE	Multi-level tree-based system Architecture
	Starfish	Self-tuning analytics system
	MRAM	Mobile Agent Technology
	CBDMASP	Statistical computation and data mining approaches
	SODSS	Decision Support System Issue
	HACE	Data Mining Approaches
	HADOOP	Parallel computing platform
	CUDA	Parallel computing platform
	Storm	Parallel computing platform
	MLPACK	Scalable machine learning library
	Mahout	Machine-learning algorithms
	PIMRU	Machine-learning algorithms
	Radoop	Machine-learning algorithms
Mining Algorithm	CLOUDVISTA	Cloud computing for clustering
	MSFCUDA	GPU for clustering
	BDCAC	Ant on grid computing environment for clustering
	Corest	Use a tree construction for generating the corsets in parallel for clustering
	Cos	Parallel computing for classification
	Quantum SVM	Quantum computing for classification
	DPSP	Applied frequent pattern algorithm to cloud platform
	DHTRIE	Applied frequent pattern algorithm to cloud platform
	SPC, FPC, DPC	Map-reduce model for frequent pattern Mining

3. CONCLUSION AND FUTURE RESEARCH

In this article, we reviewed studies on the data analytics from the traditional data analysis to the recent big data analysis. From the system perspective, the KDD process is used as the framework for these studies and is summarized into three parts: input, analysis, and output.

From the analysis framework perspective, this study shows that big data framework, platform, and machine learning are the current research trends in big data analytics system. For the mining algorithm perspective, the clustering, classification, and frequent pattern mining issues play the vital role of these researches because several data analysis problems can be mapped to these essential issues.

REFERENCES

[1] Charu. C. Aggrawal, (2012), "Real-Time Data Analytics in Sensor Networks", Managing and Mining Sensor Data Journal, Springer publisher, ISBN 978-1-4614-6309-2, pp173-201.

[2] Steve Lohr, The age of big data.Rep. (2012). http://wolfweb.unr.edu/homepage/ania/NYTFeb12.pdf

[3] Fisher D, DeLine R, Czerwinski M, Drucker S., (2012), Interactions with big data analytics. Interactions Journal, ACM, Volume:19, pp50–9

[4] R, Wunsch D. (2009), Clustering Hoboken: Wiley-IEEE Press.

[5] Fayyad UM, Piatetsky-Shapiro G, Smyth P. (1996), "From data mining to knowledge discovery in databases". AI Mag, Vol. 17, pp37–54.

[6] Charu. C. Aggrawal, (2012), "Mining Sensor Data Streams", managing and Mining Sensor Data Journal, Springer publisher, ISBN 978-1-4614-6309-2, pp143-166.

[7] Russom P. (2011), big data analytics. TDWI: Tech. Rep.

[8] Ma C, Zhang HH, Wang X. (2014), "Machine learning for big data analytics in plants", Trends Plant Sci., Vol. 19, pp798–808.

[9] Boyd D, Crawford K. (2012), "Critical questions for big data". Inform Commun Soc., Vol. 15, pp662–79.

[10] Charu. C. Aggrawal, (2012) "An Introduction to sensor Data Analytics", Managing and Mining Sensor Data Journal, Springer Publisher, ISBN 978-1-4614-6309-2, pp1-8.

[11] Baraniuk RG., (2011)," More is less: signal processing and the data deluge" Science. Vol. 331, pp717-9.

Authors

Kavita Ahuja

(Assistant Professor) received the B.Sc. in mathematics and M.C.A. degree in 2005 and 2008 respectively from Veer Narmad South Gujarat University, Surat, India, and currently involved in researcher as a Ph.D. scholar in the area of sensor data analytics from Hemchandracharya North Gujarat University, Patan, India from 2014, under the guidance of Dr. N. N. Jani (Director, KSV).She is currently working as Assistant Professor in B.C.A college since 2008.

N. N. Jani

Currently contributing as Mentor to Institute Industry Interaction Cell of KSV University at Gandhinagar. He has served Saurashtra University Rajkot For a period of more than two decades as Prof and Head, Dept of Computer Science till July 2008 and thereafter as Director at SK Patel Institute of Management and Computer Studies, Gandhinagar up to Aug,2015with a rich teaching and research experience of 41 years. He successfully guided 36 scholars who completed their PhD in Computer Science. He published more than 82 research papers during this tenure. His research area is:

High Performance Computing, Big Data Analytics, Smart Embedded Systems, Nano Materials Characterization.