International Journal of Research in Advent Technology (IJRAT) Special Issue
E-ISSN: 2321-9637
Sharadchandra Pawar college of Engineering, Dumbarwadi, Pune 410504,
Organizes
National Conference "MOMENTUM-17", 14th & 15th February 2017
Available online at www.ijrat.org

Use of Data Mining Technique In Unstructured Data of Big Data

Mr. Sadashiv P. Shinde [1], Prof. Purbey Suniti [2]
[1]ME Student G. H. Raisoni College of Engineering & Management, Chas, Ahmednagar, Maharashtra, India
E-mail:sada.shinde83@gmail.com
[2]Asst. Prof. G. H. Raisoni College of Engineering & Management, Chas, Ahmednagar, Maharashtra, India
E-mail:sunitynu@gmail.com

Abstract- Big data is collection of data which is wide range and complex data.. Data get generated from every way, from different fields. These big data has structured semi-structured and unstructured types of data. In today era data is been generated on large scale. Social media sites, digital pictures and videos and many other. Whole world is moving towards the digitalization. All this type of data is known as big data. Data mining is a way for discovering a pattern which is useful from large scale data sets. We collect the healthcare data which include all the details of the patients their symptoms, disease etc. Once we collect the data then there will be pre-processing on that data as we need only filtered data for our analysis. Useful and meaningful data can be extracted from this big data with the help of data mining by processing on that data. The data will be stored in Hadoop. User can access the data by symptoms, disease etc.

Index Terms— Big data, Data mining,Privacy, Hace theorem,Hadoop.

1. INTRODUCTION

In healthcare surroundings it is usually seen that there is information rich but the knowledge in its poor one. People care extremely about fitness and health and they want to be more protected, in case of their healthcare and health related issues. Quality service implies administering treatments that are effective according to diagnosing patients correctly. There is large data present with the healthcare systems records but they not have effective analysis way to discover important data and hidden relationships in complex data or patterns in that data. A main challenge posed to the healthcare decision makers is to offer quality services. The proposed system aims at simplify the task of doctors and medical students as well as insurance company. Poor clinical decisions can direct to terrible consequences.When the doctor fires a query regarding symptoms or disease then the system provides the information regarding the diseases, Records about that inferred disease. The tools that are capable to recognize relevant information in the medical science domain stand as construction blocks for this healthcare system. In this system, we see diseases and there records facts, and the relation which is present between both. that exists. The method used to sort out all this we use the HACE theorem. Basically our paper aims to benefits of the two today very fast developing research areas which are data Pre-processing techniques and Data Mining by discovering a framework which incorporates both the research areas. Our objective aims for this work is to Data mining done on huge amount of big data techniques which illustration of information and which grouping algorithms are proper for classifying and identifying significant medical related information in short representation. In this research, we focus on relation between the diseases and recorded information. That is present between diseases and recorded. Our interests are in order to a personalized medicine, In this patient has a medical care personalized according to its his requirement. We acknowledge the actuality that are tools capable of finding the relevant and reliable information in the medical domain standpoint as basic building blocks pillars for a healthcare record system that is up-to date with the latest survey and discoveries in the medical fields. It's not adequate to know and read the information only necessary for treatment is help for disease healthcare should provide all the information and new invention discoveries about assured treatment and record to specify it may also have certain side effect to specific type of patient . We have to used new technologies to process such kind of data and discover the pattern by using the data mining. The good practice guide initially as educative and introductory sources of agencies seeing to bring in big data capability and opportunities that accomplish the different challenges of implementation. Even the element using big data and implementing smaller or greater in the government agency this will also highlight different challenges come under practical in main stream of performing and operation. The paper content are as Section 2 discusses objectives, Section3 discuss related work, Section 4 presents problem statement, Section 7Contains Analytical analysis and Section 8 Contains conclusions.

I. OBJECTIVE

Enables Data Mining in hadoop data sets and provide anonymous data with respect to privacy confidentiality preservation and authentication of the user.

2. LITERATURE SURVEY

The one of the important characteristic of big data is to perform computation on data present in GB and PB (petabyte) and even on exa-byte (EB) with the computational process. The different sources heterogeneous, huge and data having different characteristics of data content in big data. So system make used of parallel computing, it's a correspondent programming support and software to capably analyze and mine the entire data in different format are the target focus of big data process to transform in quantity to quality. Map Reducer is batch orientated parallel processing of data. There are some short come and performance gap with relational data base. To increase the performance and increase the nature of large data Map Reducer has used data mining algorithm and machine learning. Currently processing of big data relay on parallel computing technique like Map Reducer supply cloud computing as a good platform big data for community as service. The mining algorithm used in this are , including locally weighted linear regression, k-Means, linear support vector machines, logistic regression, Gaussian discriminant analysis, the independent variable analysis, expectation maximization, naive Bayes, and back-propagation neural networks . Data mining algorithm obtain the optimizes result it perform computing on large data. By increasing performance and appropriate algorithm are process in parallel programming which is applied to number of machine learning algorithm which is based on Map Reducer frame work .With the machine learning we can state that the process can be change to summation operation. Summation operation can be perform on subset of data separately and accomplish simply on Map Reducer programming. Reducer node collect all the processed data and collect into summation. Ranger et al. Proposed application of Map Reducer to support parallel programming and multiprocessor system which include three different data mining algorithm K-means ,linear regression principal component analysis. In paper [3] the Map Reducer mechanism in Hadoop execute the algorithm in single-pass, query based and iterative frame work of Map Reducer, distributing the data between number of nodes in parallel processing algorithm that the Map Reducer approach for huge data mining by checking standard data mining task on mid size clusters. Polarimetries and sun[4].In this they proposed a mutual distributed aggregation (DisCo) frame work for pre-processing of practically and collaborative technique. The performance in Hadoop it is and open source Map Reducer project show that DisCo have ideal which is accurate and can analyse and process enormous data. Therefore the large data set are can be divided into small subset and that subset can be assign to various number of machine in Mapper the data is process by the mapper it perform operation on it. To took up the poor analysed capabilities and the week analysed software which are traditional Hadoop system. In detail integration which give the data for the computation in parallel processing model that make use of full Hadoop. It is beyond their limits for processing it. Increase of big data application has increase in the areas where the data is generated more and more which can't be handled by the normal software. The most important challenge in Hadoop is to process the Big Data and to get the valuable information from that large data sets. An extraction of information about diseases and symptoms from hadoop data sets using data mining. There valuable data obtained can be used for the future measure.

3. PROPOSED SYSTEM

We propose the HACE theorem of the Big Data of Unstructured data. HACE suggest the characteristic of Big Data Heterogeneous, Huge and diverse data sources, Autonomous with distributed control complex and evolving in data and knowledge associations. HACE theorem discovers the useful knowledge from the big data. To perform the operation on Big Data the system should be well systemic design to make full use of Big Data. Health related data is get collected from the website and it getting stored in text format for further processing. The following figure depicts the System Architecture of the System. Once data get collected then data is classified Units

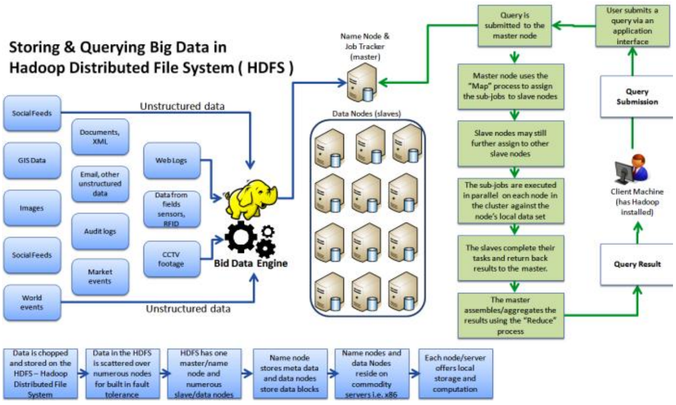

Storing & Querying Big Data in Hadoop Distributed File System (HDFS)
Social Feeds
GIS Data
Images
Social Feeds
World events
Documents, XML
Email, other unstructured data
Audit logs
Market events
Web Logs
Data from fields sensors, RFID
CCTV footage
Unstructured data
Unstructured data
Bid Data Engine
Name Node & Job Tracker (master)
Data Nodes (slaves)
Query is submitted to the master node
User submits a query via an application interface
Master node uses the "Map" process to assign the sub-jobs to slave nodes
Slave nodes may still further assign to other slave nodes
The sub-jobs are executed in parallel on each node in the cluster against the node's local data set
The slaves complete their tasks and return back results to the master.
The master assembles/aggregates the results using the "Reduce" process
Query Submission
Client Machine (has Hadoop installed)
Query Result
Data is chopped and stored on the HDFS – Hadoop Distributed File System
Data in the HDFS is scattered over numerous nodes for built in fault tolerance
HDFS has one master/name node and numerous slave/data nodes
Name node stores meta data and data nodes store data blocks
Name nodes and data Nodes reside on commodity servers i.e. x86
Each node/server offers local storage and computation

Fig. 1. System Architecture

as homogeneous and heterogeneous data. Then data is getting classified disease wise means similar disease data form cluster by K means algorithm. All the data cluster is get stored in Hadoop. User enters the symptoms or disease names then the all the details of the input are get fetched from Hadoop. For this we apply the NLP algorithm SOM on the data set which removes the Stop words/Special Character/HTML tag. Means homogeneous data is formed. One disease may have different attributes then that data forms heterogeneous data cluster. And anonymous data is provided to users.

3.1 K means

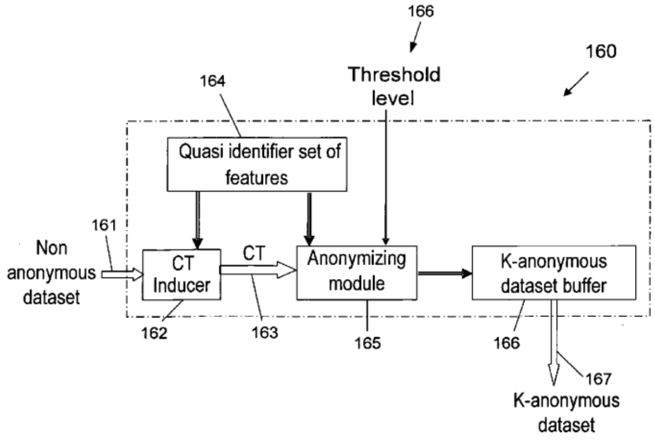
k-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem.In Big Data huge data come from distributed control all the data is been collected and cluster of data are built up. Data with same characteristic are built in the cluster. Clustering of data is been done in the K-Means process. Preprocessing Self optimize mapping (SOM): Data is collected from different sources. Before applying data mining algorithms to the Data Sets pre-processing is needed. As the data mining discover the pattern it should contain large data related to that field. The common sources of data are data warehouse. In pre-processing the stop words, HTML tags, special character are removed. The noise missing data identifying the diseases and treatments sentences published in medical information. Pre-processing is very essential in the case of multi variant data. The informative sentence are collected as label informative which contain the information about diseases and there treatment. Other are noted not informative sentence.

3.2 Data Mining

Data mining contain different classes of tasks: Clustering the process where data get collected into clusters a group of data similar data is grouped together. Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. Anomaly detection: Outlier/change/deviation detection. A common habit followed by the people generally in the market place. It identifies the error data which need to be further processed. Association rule learning (Dependency modeling) Association relation between the entity. Which product material is been purchased frequently by the customer. Homogeneous data to find out the information related to diseases, building the cluster of data of same diseases. Specific diseases information is been collected in one block. Summarization. It contain the Abstract information which is been summary of all the detail record providing a more compact view representation of data set. Heterogeneous data one disease may have different attributes, different types of specific diseases may be present it stores such type of information.

3.3 K annonymity algorithm

K-Anonymity is a privacy preserving method for limiting disclosure of private in-formation in data mining. In this module the anonymous data is been generated which contains only range values without containing actual values. For example identity of the person can be hidden from the user for security purpose. To avoid the misuse of personal details the anonymous data is generated and provided to the user. Algorithms use the data representation to create anonymous data of that captures data regarding diseases, feature values and labels in order to given query by the user.



3.4 Authontication.

Authentication is a process in which the credentials provided are compared to those on file in a database of authorized users' information on a local operating system or within an authentication server. It is been prevented by the password and user id through which they can get access in to the system to get the required data. In this module the SQL injection attacks are prevented. In which fake user could not get the access of the data base of the system. It provides the general security to the overall system. If the credentials match, the process is completed and the user is granted authorization for access. After entering the authentication details in the system the further action take pace. Authentication is any process by which a system verifies the identity of a User who wishes to access it. It prevents any illegal access to the system.

3.5 Apriory

In this module the frequent item set mining is been done. When user enters any query to the system then in which data set cluster that term is occurring number of times. Association rules analysis is a technique to uncover how items are associated to each other. There are three common ways to measure association. In this module the user only give the query to the system and then that query is processed by the system after processing the most relevant information is provided to the user.

II. Mathematical Module

S= S1, S2, S3, S4
Where,
S1=Username which is entered by the user.
S2=Password of the session is given.
S3=Keywords which we are going to search in the data set
S4=Dataset Intermediate Output Set after processing of data.
C= C1, C2
Where,
C1=check out the validation of the user Authorized User
C2=cluster of data which is formed by clustering Homogeneous Data
C3=Entity having the different attribute are been placed in Heterogeneous Data
C4= The output which is secure hiding the sensitive detail record Anonymous Data is generated. Final Output Set result from the data set is displayed which has anonymous data.
D= D1

4. Perfofmance Analysis

Consider the raw patient data in Table 1 (the attribute is just for the purposes of illustration). Each row in the table represents the information from a patient. Then, record can be uniquely identified, since she is the only with diagnostic codes in the raw data. Make the privacy attack easier for an adversary. Suppose that the adversary knows that the target patient is female and her diagnostic codes contain . Thus, identifying her record results in disclosure that she also has [10]. Note that, we do not make any assumption about the adversary's background knowledge. A differentially-private mechanism ensures that the probability of generating any output (released data) is almost equally likely from all nearly identical input data sets, and

ID	*AGE*	*ZIP*	*DISEASES*
1	6	3643	Fever
2	16	74532	Diarrhea
3	29	3241	Flu
4	23	3521	Fever
5	36	3451	fever

Table1.Sensitive Table

An adversary may have partial or full information about the set-valued data and she can try to use any background knowledge to identify the victim. Therefore guarantees that outputs are insensitive to any single individual's record. In other words, an individual's privacy is not at risk because of inclusion in the disclosed data set.

ID	*AGE*	*ZIP*	*DISEASES*
1	0 to 22	****	Fever
2	0 to 22	****	Diarrhea
3	22 to 32	****	Flu
4	22 to 32	****	Fever
5	32 to 42	****	fever

Table 2. Table Anynomous.

Differential privacy makes no assumption about an adversary's background knowledge. A differentially-private mechanism ensures that the probability of any output (released data) is almost equally likely from all nearly identical input data sets and thus guarantees that all outputs are insensitive to any single individual's data. In other words, an individual's privacy is not at risk because of inclusion in the disclosed data set. To prevent such linking attacks, a number of partition-based privacy models have been proposed [12, 13, 14, 15, 16]. However, recent research has indicated that these models are vulnerable to various privacy attacks [18, 23, 24, 25] and provide insufficient privacy protection. In this article, we employ differential privacy26, a privacy model that provides provable privacy guarantees and that is, by definition, immune against all aforementioned attacks.

5. Conclusion

Big data is collection of complex data sets, An Data mining and privacy preservation framework for big data has been proposed. Data mining allow to explore important knowledge and privacy preservation allow to provide the anonymous data to the user. The framework is combination of accessing mined data and Privacy preservation mechanism. System processes all the data collected from different sources. Through this system we get expected information when the user enters the disease name or disease symptoms. All the data related to application users query accordingly is provided to the user real-time. User enters the keyword to the system and system provide the related information regarding to the keyword.

Acknowledgment

We would like to take this opportunity to express my sincere gratitude to my Project Guide Prof. Purbey Sunita (Assistant Professor, Computer Engineering Department,) for his encouragement, guidance, and insight throughout the research and in the preparation of this dissertation. He truly exemplifies the merit of technical excellence and academic wisdom.

References

[1] Novel Metaknowledge-based Processing for multimedia Big Data clustering challenges, 2015 IEEE International Conference on Multimedia Big Data.

[2] Bo Liu, Member, IEEE, Keman Huang Jianqiang Li, and MengChu Zhou, "An Incremental and Distributed Inference Method for Large-Scale Ontologies Based on MapReduce ParadigmKnowledge and Information Systems", vol. 45, no. 3, pp. 603-630, Jan.2015.

[3] Muhammad MazharUllahRathore, Anand Paul "A Data Mining with Big Data" IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014.

[4] Xindong Wu, Fellow, IEEE, Xingquan Zhu "Real-Time Big Data Analytical Architecturefor Remote Sensing Application- Knowledge and Information Systems", vol. 33, no. 3, pp 707-734, Dec. 2015.

[5] Yanfeng Zhang, Shimin Chen, Qiang Wang, and Ge Yu "MapReduce:Incremental MapReduce for Mining Evolving Big Data ACM Crossroads", vol. 27, no. 2, pp. July 2015.

[6] S. Banerjee and N. Agarwal "Analyzing Collective Behavior from Blogs Using Swarm Intelligence,Knowledge and Information Systems", vol. 33

[7] J. Mervis, "Science Policy: Agencies Rally to Tackle Big Data,Science", vol. 336, no. 6077, p. 22, 2012.

[8] D. Luo, C. Ding, and H. Huang "Parallelization with Multiplicative Algorithms for Big Data Mining", IEEE 12th Intl- Conf. Data Mining, pp. 489-498, 2012.

[9] Xindong Wu, Fellow, IEEE, Xingquan Zhu "A Data Mining with Big Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014.

[10] Zahid Pervaiz, Walid G. Aref, Senior Member, IEEE, Arif Ghafoor, Fellow, IEEE, and Nagabhushana Prabhu Accuracy-Constrained Privacy-Preserving Access Control Mechanism for Relational Data IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 4, April 2014