

Review on Effective Data Mining Technique use with Structured and Unstructured data of Big Data

Mr. Dnyandeo Sopan Khemnar ¹, Prof. Priyanka Dhasal ²

¹ PG Student, Computer Science Engineering, Patel college of Science & Technology, Indore, Madhyapradesh, India

² Professor, Computer Science Engineering, Patel College of Science & Technology, Indore, Madhyapradesh, India

ABSTRACT

With the Invention of Big data. Big Data is collection of large and complex data. It consist of structured, semi-structured and unstructured types of data. Data get generated from various sources and from different fields. In today era data is been generated on huge amount. The Whole world is moving towards the digitalization. Social media sites, digital pictures and videos and many other. All this such type of data is known as big data. Data mining is a useful technique for extracting a patterns. Which is helpful from large scale data sets. Useful and meaningful data can be extracted from this big data with the help of data mining by processing on that data. In this paper We collect the healthcare data which consist of all the details of the patients their symptoms, disease etc. After the collection of data then there will be pre-processing on that all the details of the patients data as we need only filtered data for our analysis. The data will be stored in Hadoop. User can retrieve the data by symptoms, disease etc.

Keyword : Big data, Data mining, Hace theorem, Map Reducer, Privacy Preservation Mechanism etc.

1. INTRODUCTION

In current era as industrialization increases the pollution leves get increases. peoples are more careful about fitness and health and they want to be more protected, in case of their healthcare and health related problems. In healthcare surroundings it is normally seen that there is a huge amount of information available but the knowledge in its poor one. There is large data availabe with the healthcare systems records but they don't have effective analysis way to dis-cover 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. Quality service implies administering treatments that are effective according to diagnosing patients correctly. Poor clinical decisions can direct to terrible consequences. The proposed system aims at simplify the task of doctors and medical students as well as insurance company. When the doctor fires a query regarding symptoms or disease then the system provides the information regarding the diseases, Records about that inferred disease. 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 re-search areas. 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 sorting all this we use the HACE theorem. 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 important medical related information in compact representation. We acknowledge the actuality that are tools able to finding the relevant and reliable information in the medical domain standpoint as basic building blocks pillars for a healthcare record system that is upto date with the recent survey and discoveries in the medical fields. 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 system, In this patient has a medical care personalized according to its his requirement. Its 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 .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. We have to used new technologies to process such kind of data and discover the pattern by using the data mining. The paper content are as Section 2 discusses objectives , Section 3 discuss related work, Section 4 presents problem statement, Section 5 our system overview Section 6 Methodology to solve the task, Section 7 Contains Analytical analysis and Section 8 Contains conclusions

2. OBJECTIVE

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

3. LITERATURE SURVEY

The huge amount of data from various sources is heterogeneous type, and data having different characteristics of data content in big data. 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. So system make used of parallel computing, it's a correspondent programming support and software to capably analyse and mine the entire data in different format are the target focus of big data process to transform in quantity to quality. Currently processing of big data relay on parallel computing technic like Map Reducer supply cloud computing as a good platform big data for community as service.

In Paper [1] It deal with sustainability of health care system. It focus on changing approach of peoples to deal with health care system and make it personalized by using big data concept as mention in table.

Table -1: Addressed Table

Research Tracks	Addressed Challenges
Tele-monitoring and translational data for personalized VPH for diagnosis and prediction	Integration of medical knowledge Multi-modal and multi-scale Self-adaptive algorithms Personalized aid decision tools Interpretable Patterns
Merging data and knowledge driven ML for interpretable, personalized and incremental modeling and data mining	Integration of medical knowledge and clinical evidence Multi-modal and multi-scale Interpretable patterns Treat and correct data from uncontrolled conditions Holistic workflows & Decision Support System
Data fusion for robust bio-signal processing	Integration of medical knowledge and clinical evidence Treat and correct data from uncontrolled conditions Multi-modal and multi-scale
Patient empowerment	Behavior modelling User profiling Behavioral ontology
Organizational issues and translational research	Use in medical daily routine Institutional support Funds for deployment and scale up New organizational and communication schemes

In Paper [2] Finding associations between concepts have applications in different domains, ranging from Medicine to Material sciences. The associations between different types of diseases ,genes, symptoms, treatments, and drugs can be used to understand the functional relationships between medical concepts, and potentially come up with new preventive care,diagnostic methods, or cures for diseases. Extending the association analysis to 'indirect' association mining enables us to find more subtle relationships between concepts and propose new hypotheses for further research. This is of great benefit to researchers to focus on hypotheses with stronger evidences. In paper [3] 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. 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. 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 backpropagation neural networks 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. 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. Reducer node collect all the processed data and collect into summation. Proposed application of Map Reducer to sup-port parallel programming and multiprocessor system which include three different data mining algorithm K-means,linear regression principal component analysis. the Map Reducer mechanism in Hadoop execute the algorithm in single-pass, query based and iterative frame work of Map Reducer , dis-tributing 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 midsize clusters. Polarimetries and sun[4].In this they proposed a mutual distributed aggregation (DisCo) frame work for pre-processing of practically and collaborative technic. The performance in Hadoop it is and open source Map Reducer project show that DisCo have

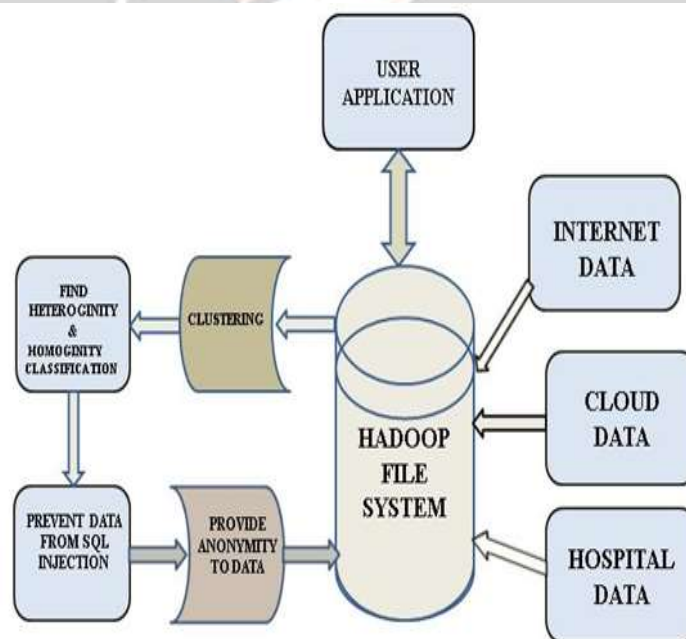


Fig -1: System Architecture

ideal which is accurate and can analyse and process enormous data. 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. 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. It is beyond their limits for processing it. The most important challenge in Hadoop is to process the Big Data and to get the valuable information from that large data sets. There valuable data obtained can be used for the future measure.

4. PROBLEM STATEMENT

User can independently and concurrently access data base or data from hadoop. User can perform SQL operations. An extraction of information about diseases and symptoms from hadoop data sets using data mining.

5. PROPOSED SYSTEM

We propose the HACE theorem of the Big Data. HACE theorem discovers the useful knowl-edge from the big 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. To perform the operation on Big Data the system should be well systemic design to make full use of Big Data. The following figure depicts the System Architecture of the System. Health related data is get collected from the website and it getting stored in text format for further processing. Once data get collected then data is classified as homogeneous and heterogeneous data. For this we apply the NLP algorithm SOM on the data set which removes the Stop words/Special Character/HTML tag. Then data is getting classified disease wise means similar disease data form cluster by K means algorithm. Means homogeneous data is formed. One disease may have different attributes then that data forms heterogeneous data cluster. 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. And anonymous data is provided to user.

5.1 K Mean

In Big Data huge data come from distributed control all the data is been collected and cluster of data are built up. Clustering of data is been done in the K-Means process. Data with same characteristic are built in the cluster. Pre-processing 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. Pre-processing is very essential in the case of multi variant data. 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. The informative sentence are collected as label informative which contain the information about diseases and there treatment. Other are noted not informative sentence.

5.2 K means algorithm

Data mining :

Data mining contain different classes of tasks:

Clustering : the process where data get collected into a group of data similar data is grouped together.

Anomaly detection: Outlier/change/deviation detection. It identifies the error data which need to be further processed. Association Rule Learning(Dependency modelling) Association relation between the entity. A common habit followed by the people generally in the market place. 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. Heterogeneous data one disease may have different attributes, different types of specific diseases may be present it stores such type of information. Summarization. It contain the Abstract information which is been summary of all the detail record providing a more com-compact view representation of data set.

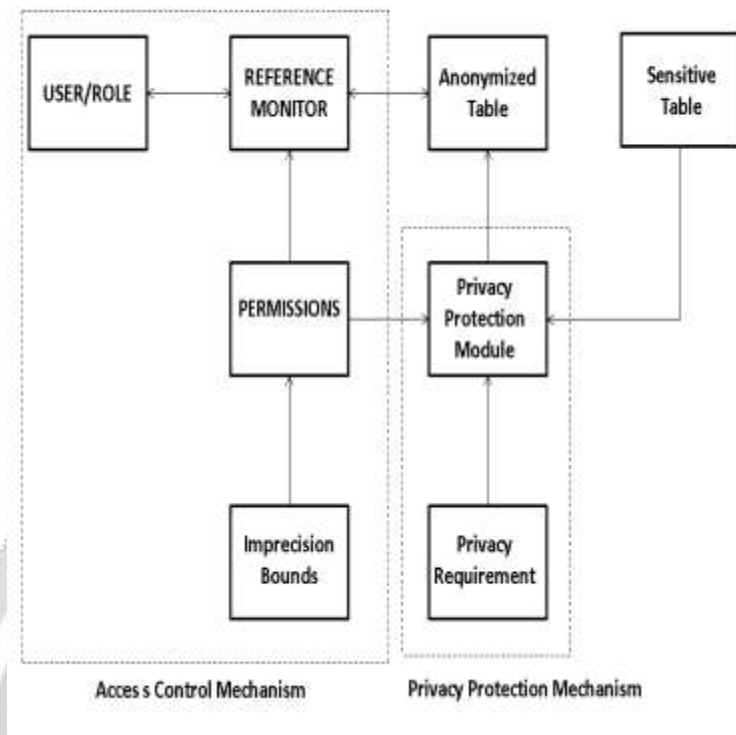


Fig -2: Privacy Preservation Mechanism

5.3 K-anonymity algorithm

In this module the anonymous data is been generated which contains only range values without containing actual values. Sensitive details are hidden from the user. Anonymous data converting algorithm provide provides only required information which is not sensitive. 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.

5.4 Authentication

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 is been prevented by the password and userid through which they can get access in to the system to get the required data. If anybody try to use the previous session password it get rejected does not allow the user to log in into the system. It provides the general security to the overall system. After entering the authentication details in the system the further action take pace. It prevents any illegal access to the system.

5.5 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. Such type of cluster is been presented to the user. Most frequently occurring term in the data set such data set is been selected in this module and presented to the user. 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. Huge Data with Heterogeneous and Diverse Dimensionality. The Big Data is the massive volume increasing velocity having diverse dimensionality and heterogeneous. This is because number of information sources collects their information in their own schema, and

the diverse data representations done in there format. For example, in biomedical each individual is been repres ented in different parameters age, gender, family disease history. For example X-ray checking and CT scan MRI report of each individual, images or videos are used to characterize the results because they provide visual information and details for doctors to carry in depth examinations on human being. In medical field the different doctors may have their own schema format to characterize patient and there details. The diverse dimensionality and data heterogeneity issues become major challenges to enable data aggregation bringing together data collected by different sources. Autonomous Sources with Distributed and Decentralized Control Information collected from different independent servers and distributed decentralized controls. Being autonomous, each data source collect information from autonomous source without relying on any centralizes control. For example www that is world wide web provide some quantity of information without relaying

ID	Age	Zip	Disease
1	5	15	Flu
2	15	25	Fever
3	28	28	Fever
4	22	28	Flu
5	35	25	Diarrhea

Table -2: Sensitive Table

ID	Age	Zip	Disease
1	0 to 20	0 to 20	Flu
2	0 to 20	20 to 30	Fever
3	20 to 30	20 to 30	Fever
4	20 to 30	20 to 30	Flu
5	30 to 40	20 to 30	Diarrhea

Table -3: Anonymous Table

on another server. There may be possibility of attack and inconsistency if we depend on centralize server. The large organization such as Google, WalMart and Face book a many number of server farms are deployed in all Corner of world to make sure quick responses and nonstop. Complex and Evolving Relationships Data information collected from different sources. Which evolve the complex data within it. By collecting the information regarding the personal details example age, gender, income, education. In this we consider each individual as an single entity. Without considering their social connection. We can link people with different entity not only by their Education. The relation can be evolved using their common habitat. Person having the hobby of panting can get linked with the person having the hobby of painting. There may be different in their age or education other personal details. People can come together through this and relationship can be evolved through this way. Many to Many relation can be built up by this property.

6.MATHEMATICAL MODULE

I= I1, I2, I3, I4 Where,
 I1=Username which is entered by the user.

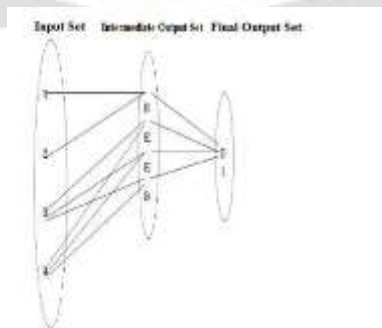


Fig -3: Venn Diagram

I2=Password of the session is given.

I3=Keywords which we are going to search in the data set

I4=Dataset Intermediate Output Set after pro-cessing of data.

E= E1, E2 Where,

E1=check out the validation of the user Authorized User

E2=cluster of data which is formed by clustering Homogeneous Data

E3=Entity having the different attribute are been placed in Heterogeneous Data

E4= 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

7.CONCLUSION AND FUTURE WORK

Big data is collection of complex data sets, Data mining is technic to explore data search of knowledge consistent patterns and then to validate the results by applying the patterns to new subsets of data. Through this system we get expected information when the user enters the disease name or disease symptoms. System processes all the data collected from different sources. All the data related to application users query accordingly is provided to the user realtime. 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. Priyanka Dhasal (Assistant Professor, Computer Science Engineering Department,) for her encouragement, guidance, and insight throughout the research and in the preparation of this dissertation. she truly exemplifies the merit of technical excellence and academic wisdom.

REFERENCES

- [1] "Challenges in personalized systems for Personal Health Care" C.Fernández-Llatas Member, IEEE, A. Martinez-Romero, A.M. Bianchi Member IEEE, J.Henriques, P. Carvalho Member, IEEE, and V. Traver Member, IEEE978-1-5090-2455-1/16/\$31.00 ©2016 IEEE.
- [2] "Mining and Visualizing Associations of Concepts on a Large-scale Unstructured Data" Reza Sadoddin,Osvaldo Driollet, 2016 IEEE Second International Conference on Big Data Computing Service and Applications, 978-1-5090-2251-9/16 \$31.00 © 2016 IEEE DOI 10.1109/BigDataService.2016.
- [3] "Research and Improve on K-means Algorithm Based on Hadoop",Kehe Wu,Wenjing Zeng, Tingting Wu, Yanwen An, 978-1-4799-8353-7/15/\$31.00 ©2015IEEE
- [4] 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.
- [5] Bo Liu, Member, IEEE, Keman Huang Jianqiang Li, and MengChu Zhou, "An Incremental and Distributed Infer-ence Method for Large-Scale Ontologies Based on MapReduce ParadigmKnowledge and Information Systems", vol. 45, no. 3, pp. 603-630, Jan.2015.
- [6] Muhammad MazharUllahRathore, Anand Paul "A Data Mining with Big Data" IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014.
- [7] 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.
- [8] 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.
- [9] S. Banerjee and N. Agarwal "Analyzing Collective Behavior from Blogs Using Swarm Intelligence,Knowledge and Informa-tion Systems", vol. 33, no.
- [10] J. Mervis, "Science Policy: Agencies Rally to Tackle Big Data,Science", vol. 336, no. 6077, p. 22, 2012.
- [11] D. Luo, C. Ding, and H. Huang "Parallelization with Mul-tiplicative Algorithms for Big Data Mining", IEEE 12th Intl-Conf. Data Mining, pp. 489-498, 2012.

[12] J. Bughin, M. Chui, and J. Manyika, "Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch.", McKinsey Quarterly, 2010.

BIOGRAPHIES

	<p>Mr. Dnyandeo Sopan Khemnar PG Student, Computer Science Engineering, Patel college of Science & Technology, Indore, Madhya Pradesh, India</p>
	<p>Prof. Priyanka Dhasal, Professor, Computer Science Engineering, Patel college of Science & Technology, Indore, Madhya Pradesh, India</p>

