K-Mean Clustering in Health Care using Map Reduce

Nandini Mehta[1], Preeti Sethi[2]
[1] Deptt. of Computer Engineering
YMCA University Of Science And Technology, Faridabad, Haryana, INDIA
[2] Deptt. Of Computer Engineering
YMCA University Of Science And Technology, Faridabad, Haryana, INDIA
[1]nandinimehta1993@gmail.com, [2]Preetisethi22@gmail.com

Abstract

In the present scenario, the use of internet enabled devices like smartphones, laptops, tablets etc. has increased to a big extent. As a result, a huge amount of data is generated on a daily basis in every work field. To manage this data, a suitable model is needed which can organize the database and keep it safe for future references. One such model is Data Clustering with machine learning and it is becoming an interesting research area. The present work imbibes the concept of clustering of vast data i.e. Big data in the field of Health sector. The work basically stores data in multiple clusters that would be made available at the particular health care center in isolated form & would be easily accessible. The Proposed model Algorithm consists of the map-reduce function which involves a certain procedure and it is discussed in the latter part of the paper.

Keywords: Big Data, Clustering, Data mining, Health care center, K-mean clustering, Map Reduce.

1. Introduction

Big Data can be defined as a huge amount of database which is very complex and difficult to manage. In the past several years, there had been a rapid growth and increase in the use of electronic devices and information technology for a number of purposes and in various fields like science, defense, medical etc. This has in turn resulted in generation of huge amount of data which is very complex and difficult to manage. Big Data consists of several components as V's which are described in the following diagram-

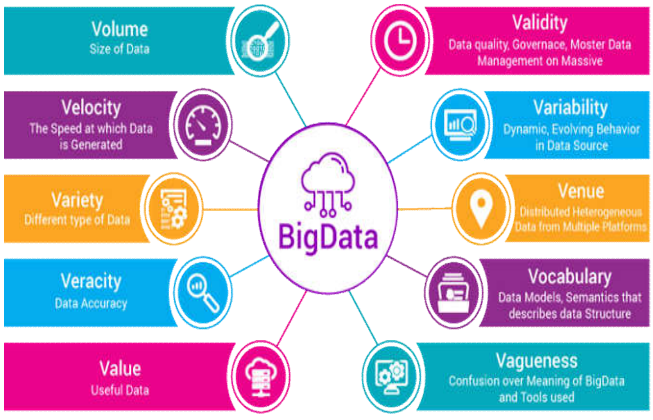

Figure 1: V's of Big Data

To organize the data sets, many techniques are being developed. One such algorithm is Map Reduce based on K-mean clustering model in which large data sets are managed and

an entity is introduced to determine the similarity between objects. Map Reduce, is a potentially labeled calculating paradigm used for large-scale execution of data in cloud computing. The main target behind handling the big data is to lower its complexity and achieve transparency to the fullest. Different aspects regarding the data are surveyed and examined so that authenticity of the system is maintained.

1.1 Healthcare Center

A regular healthcare center consists of large unstructured data related to a number of patients, doctors, nursing staff and other ward members. From personal details to the reports of a disease, a pile of data e.g. medical reports, medication, medical bills, discharge records, MRI and CT scan images, specific associations, specialized hospitals, descriptions, facilities provided and so on gets collected which needs to be stored by the system in an efficient way.
The various challenges for the healthcare sector Related to Important V's of Big Data were-

- **Volume of data-** According to a report provided by the health institute, any healthcare firm with around 1000 members has nearly 400 terabytes of data and this applied to other institutes as well. This data gets accumulated from various sources such as medical labs and their readings, 3D images etc. and it has contributed in making healthcare sector a huge data volume industry. Medical machines perform and collect a series of physiological information regarding a patient, like wise blood pressure, heartbeat rate, bio potential which simply adds on the database.

- **Variation in data-** The data can be characterized and classified as unstructured, structured and semi-structured due to its high complexity. Most of the unstructured data came from the handwritten work of the doctors and nurses. On the other hand, structured and semi structured data was referred to electronic billing and accounting, laboratory readings etc.

- **Velocity-** The static data consisting of x-ray films, paper files, scripts etc. The important aspect apart from collecting this data again and again is to process a real time data stream so that proper treatment gets available to the patient by analyzing the data accurately. Performing a real time analysis of the life threatening pathological changes could help the healthcare team to save a number of lives.

- **Veracity-** The authenticity of the data is very crucial as various sources adds to the proxy nature of the data and its quality can also be questioned. There are certain abnormalities or errors in the healthcare data which is a potential threat to the decision making process and treatment to the patient. The biggest problem is establishing a balance between patient information and maintaining the integrity of data. In the coming years, first aid will be integrated with the efficient and useful data that was processed and maintained earlier.

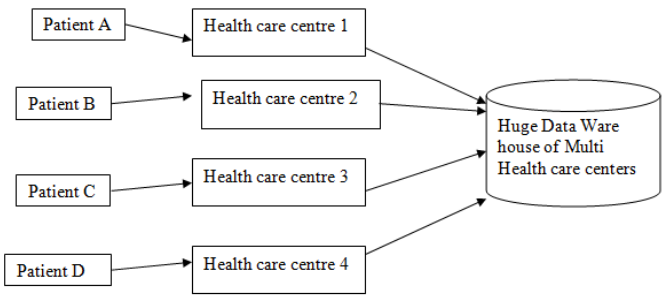

Figure 2: Existing Health Care

1.2 Clustering

Cluster can be defined as a group of data with similar characteristics and then establishing a relation or extracting information from the raw data by performing certain functions. Gigantic amount of data was generated in the form of data streams. Now it is extremely hard for the commonly used software tools to extract value, process and organize such a complicated database so its storage and processing was a difficult task. Technical implementation, analysis, sharing, storage, querying, visualization, updating and information privacy of the database were the main challenges. So there was a need for an accurate model to normalize and manage the big data. The main advantage of accuracy in big data leads to confident decision making and this enhances the operational efficiency of the system.

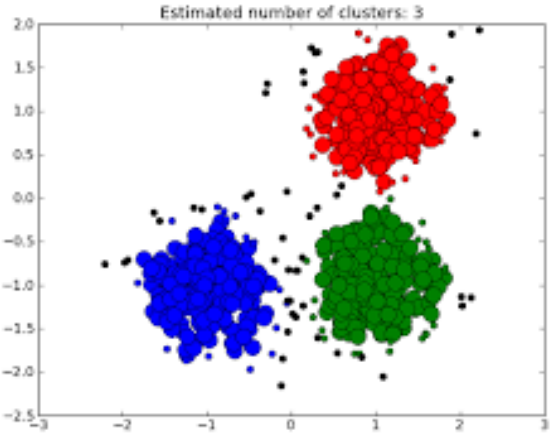

Figure: 3 Clusters

Applications of data clustering includes-

1**. Interpret the manufactured domain**- Appropriate information regarding the application and targeted goal is gathered.
2. **Accumulating the data streams**- In this part, raw data is collected and the required data sets are selected and if any variable set that is affecting the system is detected.
3. **Data cleaning, pre- processing and transformation**- The data is evaluated and if any defect like missing value or error is detected, it gets replaced with appropriate value. Data is grouped into forms essential for mining.
4. **Data integration**- The data collected from different sources is combined in a suitable form.
5. **Selection of function**- An advanced data mining function is selected for any of the problems described in stage 1.
6. **Data clustering algorithm**- To find the pattern in data, a mode of technique is accepted so that the desired action is performed.
7**. Data mining-** Here the pattern of data is converted into a specific representational form.
8. **Visualize and decode the data-** It includes tasks which decode the data patterns and make it easy to understand.
9**. Implementation**- The discovered knowledge is put into action in the manufactured performance system. The feedback is received and the knowledge can be modified further based on the statistics recorded.
10. **Storage, reuse and integration**- This includes the reuse of data in future if any similar case comes up and its possible integration into the manufacturing system.

1.2.1 K-Means Clustering

It is Unsupervised learning Algorithm which solve the well-known Clustering Problem. K-Means Clustering is fast, robust and easier to understand. This Clustering Technique is Relatively efficient: O(tknd), where n is objects, k is clusters, d is dimension of each object, and t is iterations. Normally, k, t, d << n. This Gives best result when data set are distinct or well separated from each other.

Formula Used for K-Means described as $J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$ where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j, is an indicator of the distance of the n data points from their respective cluster centers. The algorithm used for K-Means Clustering is described in later section.(Section 3 in Methodology).

1.3 Map Reduce

Map Reduce is a programming model that works on parallel and distributed algorithms in a cluster. Basically, it performs filtering and sorting functions and it follows a split-apply-combine strategy for data analysis. We can further understand this method using the given figure-

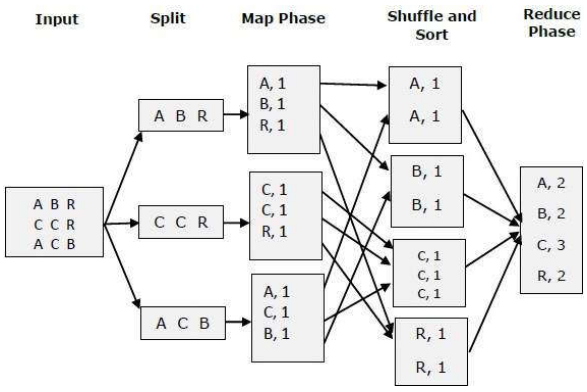

Figure 4: Map Reduce

2. Literature Review

A lot of work has been done in area of Big Data, health care & various research paper have studied for various clustering mechanisms & this Literature Portion gives idea of that existing work.
In this Section we summarizes and synthesizes on our topic.

Chen Long et.al [1] proposed the use of Map Reduce based clustering for large data sets which followed divide and conquer approach to split large amounts of data into chunks and this model reduces phases of work, independent of the other running map. Clustering quality metrics were not considered and evaluations based on different workloads was also a shortcoming.

Navjot K. [2] discussed Experimental results of K-means clustering & its performance in case of execution time. Some limitations in K-means clustering algorithm like as it took more time for execution. But this algorithm to decrease execution time. Authors has been

used Ranking Method & also shown that how clustering is performed in less execution time as compared to traditional method. This work makes an attempt at studying feasibility of K-means clustering algorithm in data mining using Ranking Method.

Arun Pushpan, Ali Akbar N .[3]considered data mining applications for health care sector which provides a variety of methods for data analysis and discover useful knowledge. The survey features various data mining techniques such as classification, clustering, association, regression in health domain. Human analyst may take time to discover useful information and since medical files are related to human subjects, privacy concern is taken more seriously in this research.

K. Chitra et.al [4] converted data and gathered useful information for supplementary purposes. The Clustering algorithms could be classified into partition-based, hierarchical based, density-based and grid-based algorithms. This paper focuses on a deep study of different clustering algorithms in data mining.

M.Umamaheswari et .al[5]studied about the applications of data mining in different fields like traffic control, weather forecasting, fraud detection, security, education enhancement and health care. In their research, drawbacks of existing machine learning algorithms were summarized and then a solution was formed using grouping mechanism.

Zeba K . et al[6] surveys the large scale data processing using Map Reduce and its various implementations to facilitate the governing bodies and other communities in developing the technical understanding of the Map Reduce framework. Different Map Reduce implementations are explored and their inherent features are compared on different parameters.

3. Methodology

3.1 Problem Formulation

To manage and utilize the data is a major task in the healthcare sector as it consists of patient queries, doctor's diagnose, medicinal description, treatment of diseases according to a particular age group and gender. They have certain **limitations** too such as
1. Fixed size of centroids in order to cluster the data using K-mean.
2. The generation of empty cluster during the procedure.

3.1.1 Health care system itself is affected by several issues such as
1. Management and storage of data sets
2. Applying the proper mechanism
3. Performance issue in cloud computing
4. Security and Authenticity
5. Availability of data in real time
6. Duplicate information increases workload

Health care [8] system is an organization set up in order to meet the medical needs of the country's population. Their extent vary among different territories and nations but with the common goal to help and improve the health of the citizens. In some countries, the healthcare project is undertaken by the private firms so that an advanced procedure can be conducted while in some developing nations, the healthcare system is the prime responsibility of the government organization and the cost of the process is economical. World Health Organization (WHO) is a centralized firm of the world which regulates the big projects related to diseases and medicines. For this, a robust financing mechanism,

well trained workforce, reliable information is essential based on which proper decisions and policies are carried out.
The major issue is that if patient gets medical assistance from one health care Centre that is connected to huge data ware house of multi health care Centre then performance is degraded in case when there are multiple patients according to different age group at different terminals. This research is step to improve performance of health care Centre services by making multiple clusters.
Our **main target** in the proposed model is to use the Map Reduce mechanism along with K-mean clustering technique that could benefit the society.

3.2 Proposed Model of Health Care Center

The huge data set that is stored on a centralized server & complex in nature would be distributed among different clients using enhanced clustering mechanisms. By storing data in multiple clusters data would be available at particular health care center in isolated form & would be quickly available to Patient.

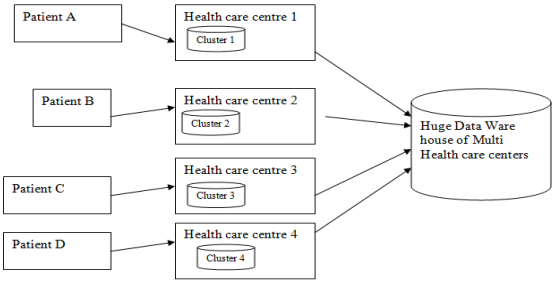

Figure 5: Proposed Health care model

The sub system of health care center is following artificial intelligence (AI) based mechanism in order to provide medical facilities & provide medical prescription according to Symptoms of disease. Enhanced clustering model would make cluster according to symptoms & diagnoses.

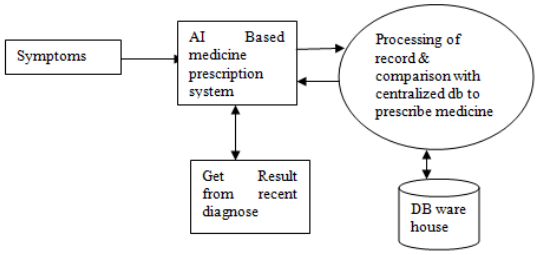

Figure 6: Subsystem in Health care center

3.3 Proposed Algorithm

A. Designing:

- At first, database should be designed based on categories such as symptoms and diseases. Also, the system must have three different users like Admin, Doctor and Analyst.
- Next is the project design based on the types /attributes on which analysis takes place.

- The data is uploaded and Pre-processed in the system.
- Tokenization, Mapper, Reducer and the analysis will be done for the said attributes.

B. Description of the Proposed Algorithm:

Here, the K-mean clustering algorithm is implemented on Map Reduce framework. The system provides the text file as an input and mapping is done on the data set. The output is stored in (key, value) pair form. Then mapper function store output in intermediate file. This file provides Reducer function as an input. The Reducer's job is to process the data that comes from the mapper and a new set of output is obtained. Now the mechanism takes place on this output.

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

Step 1: Randomly select 'c' cluster centers.
Step 2: Calculate the distance between each data point and cluster centers.
Step 3: Assign the data point to the cluster center whose distance from the cluster center is minimum of all the Cluster centers.
Step 4: Recalculate the new cluster center using eq. (1)
Step 5: Recalculate the distance between each data point and new obtained cluster centers.
Step 6: If no data point was reassigned then stop, otherwise repeat from step 3.
Step 7: End

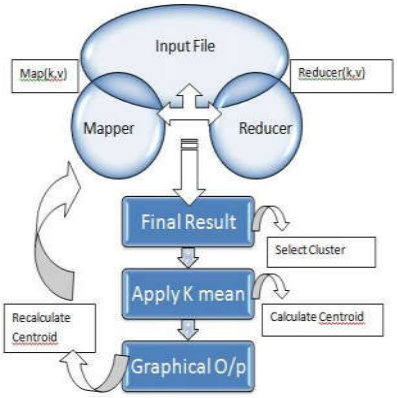

Figure 7: Implementation Steps

Map Reduce programming is designed for computer clusters. Map Reduce applications can process large datasets using several low-cost machines referred to as clusters. Individual computers in a cluster are often referred to as nodes in that cluster. Map Reduce involves two main broad computational stages (a map phase and reduce phase) that are applied in sequence on large volumes of data. The map function applies to each line of data and breaks data into chunks to form key-value pairs. A reducer function is then applied to all key-value pairs sharing the same key.

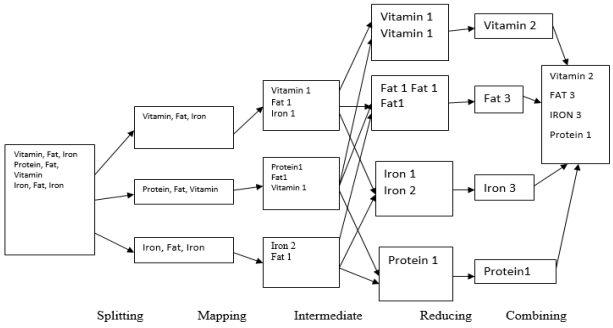

Figure 8: Health Care Map Reduce

Collection of the data related to healthcare center is made that consists of Nutrients like Protein, vitamin, iron, fat, diseases as keywords. This collection of data consists of query of patients as well as diagnosis from experts. Map reduction mechanism has been applied to the dataset in order to find the frequency of relevant nutrients as well as diseases. The outcome would be frequency. It would be clustered using K-mean clustering mechanism further the problem during K-mean clustering of relevant data would be considered and the issues would be solved.

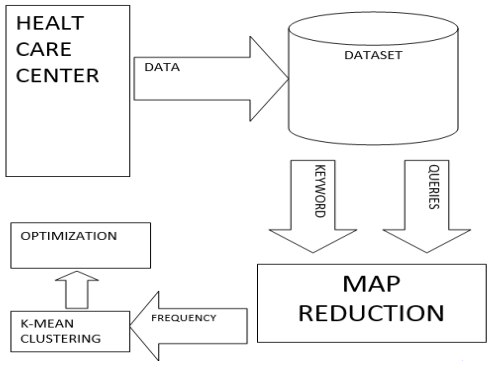

Figure 9: Flow Chart of Map Reduce

4. Scope of Research

Health care centers consists of huge data sets & it needs to be classified with best clustering mechanisms. In our proposed work, huge data set that is stored on a centralized server and then distributed among different clients. Map Reduce Mechanism for Faster Data Access would help patients from different locations to log in to a specific and specialist health care center for services. Such system would provide secure access to a patient with backup. Map Reduce is a programming model for processing large data sets that works on a parallel and distributed algorithm in a cluster. It has an extensive capability to handle the unstructured data as well. The proposed work could help in managing big data as well as make it easy to filter data set according to our requirement.

References

1. Chen M. Soft clustering for very large data sets. Comput Sci Netw Secur J. 2017;17(11):102–8.
2. Navjot K. (2012) "Efficient K-Means Clustering Algorithm Using Ranking Method In Data Mining"
3. Arun Pushpan, Ali Akbar N (2017) "Data Mining Applications in Healthcare", IOSR Journal of Computer Engineering
4. K. Chitra et al (2017) "Comparative Study of Various Clustering Algorithms in Data Mining", IJCSMC, Vol. 6, Issue. 8, August 2017, pg.109 – 115
5. M.Umamaheswari et al. [4] (2017) A Survey of Big Data Analytics in Healthcare, International Journal of Advanced Computer Science & Applications, Vol. 8, No. 6.
6. Zeba Khanam , Shafali Agarwal (2015),Map Reduce: A survey Paper on recent Expansion, (IJACSA) Vol. 6, No. 8.
7. Preeti Sethi, Naresh Chauhan(2014),Design of Communication strategy for Wireless Sensors in Non Deterministic environment using Mobile Agents,(IJCAR) Vol. 3, pg.104-117.
8. K. Srinivas , B. Kavitha Rani & Dr. A. Govrdhan, ―Applications of Data Mining Techniques in Healthcare & Prediction of Heart Attacks‖ International Journal on Computer Science & Engineering (2010).
9. ShwetaKharya, ―Using Data Mining Techniques For Diagnosis & Prognosis Of Cancer Disease‖, International Journal of Computer Science, Engineering & Information Technology (IJCSEIT), Vol.2, No.2, April 2012.
10. Arvind Sharma & P.C. Gupta ―Predicting Number of Blood Donors through their Age & Blood Groupby using Data Mining Tool‖ International Journal of Communication & Computer Technologies Volume 01 – No.6, Issue: 02 September 2012.
11. Nikita Jain & Vishal Srivastava, "Data mining techniques" international journal of research in engineering & technology.