A Technique towards Retrieving Data Set from Clustered Big Data

Mr.Kanse Satyawan Laxman[1], Prof. Kulkarni P.R.[2], Prof Kulkarni V.B.[3]

[1] *ME Student, Software Engineering, Aditya College of Engineering, Beed, Maharashtra.*
[2] *Assistant Professor, Software Engineering, Aditya College of Engineering, Beed, Maharashtra.*
[3] *Assistant Professor, Software Engineering, Aditya College of Engineering, Beed, Maharashtra.*

ABSTRACT

In recent days the information technology facing problems due to occurrence of large amount of data called as Big data. A buzzword "Big Data" has basically three attributes like Volume, Velocity and Varity. A mining data from large data set is very tedious and complex task, to overcome this here defined approach shows a method to mine big data efficiently with the help of clusters.

Keyword: - *Big data, Clusters, Classification, Complexity.*

1. Introduction

In current century everyone creating and deploying data in his own format as related to or independently to others. The era of petabyte has come and almost gone, leaving us to confront the exabytes era now. Technology revolution has been facilitating millions of people by generating tremendous data via ever-increased use of a variety of digital devices and especially remote sensors that generate continuous streams of digital data, resulting in what has been called as "big data". It has been a confirmed phenomenon that enormous amounts of data have been being continually generated at unprecedented and ever increasing scales. Big data frequently comes in the form of streams of a variety of types. Time is an integral dimension of data streams, which often implies that the data must be processed or mined in a timely or (nearly) real-time manner. Besides, the current major consumers of big data, corporate businesses, are especially interested in "a big data environment that can accelerate the time-to-answer critical business questions that demonstrate business values".

We are sure living in an interesting era – the era of big data and cloud computing, full of challenges and opportunities. Organizations have already started to deal with petabyte-scale collections of data; and they are about to face the exabyte scale of big data and the accompanying benefits and challenges. Technology revolution has equipped millions of people the ability to quickly generate tremendous stream data at any time and from anywhere using their digital devices (either for business profits or individual leisure); besides, remote sensors have been ubiquitously installed and utilized to produce continuous streams of digital data. Massive amounts of heterogeneous, dynamic, semi-structured and unstructured data are now being generated from great diverse sources and applications such as mobile-banking transactions, calling-detail records, online user-generated contents (e.g., tweets, blog posts, keeks videos), online-search log records, emails, sensor networks, satellite images and others.

2. Literature Survey

The literature survey defines a past working details of some author related to same topic. By identifying the methodologies and techniques of them we are going to construct an efficient one method to retrieve big data. Chi Yang et all explains a technique on A Time Efficient Approach for Detecting Errors in Big Sensor Data on Cloud, introduces as Big sensor data is prevalent in both industry and scientific research applications where the data is generated with high volume and velocity it is difficult to process using on-hand database management tools or traditional data processing applications. Cloud computing provides a promising platform to support the addressing of this challenge as it provides a flexible stack of massive computing, storage, and software services in a scalable

manner at low cost. Some techniques have been developed in recent years for processing sensor data on cloud, such as sensor-cloud. However, these techniques do not provide efficient support on fast detection and locating of errors in big sensor data sets.

Xuyun Zhang et all introduces the topic of Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud explains, cloud computing provides promising scalable IT infrastructure to support various processing of a variety of big data applications in sectors such as healthcare and business. Data sets like electronic health records in such applications often contain privacy-sensitive information, which brings about privacy concerns potentially if the information is released or shared to third-parties in cloud. A practical and widely-adopted technique for data privacy preservation is to anonymize data via generalization to satisfy a given privacy model. However, most existing privacy preserving approaches tailored to small-scale data sets often fall short when encountering big data, due to their insufficiency or poor scalability.

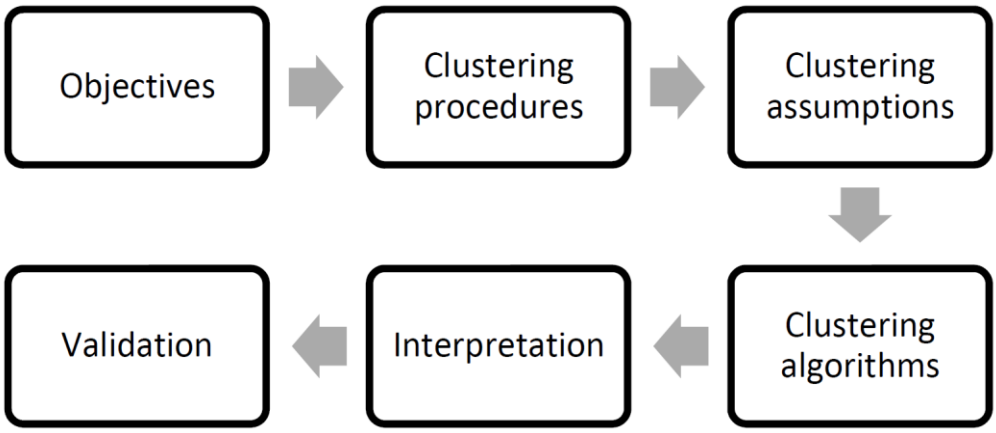
Huan Ke et all defines The MapReduce programming model simplifies large-scale data processing on commodity cluster by exploiting parallel map tasks and reduce tasks. Although many efforts have been made to improve the performance of MapReduce jobs, they ignore the network traffic generated in the shuffle phase, which plays a critical role in performance enhancement. Traditionally, a hash function is used to partition intermediate data among reduce tasks, which, however, is not traffic-efficient because network topology and data size associated with each key are not taken into consideration.

Bo Liao et all put the concept on Efficient Feature Ranking Methods for High-throughput Data Analysis, here they had defined Efficient mining of high-throughput data has become one of the popular themes in the big data era. Existing biology related feature ranking methods mainly focus on statistical and annotation information. In this study, two efficient feature ranking methods are presented. Multi-target regression and graph embedding are incorporated in an optimization framework, and feature ranking is achieved by introducing structured sparsity norm. Unlike existing methods, the presented methods have two advantages: (1) the feature subset simultaneously account for global margin information as well as locality manifold information. Consequently, both global and locality information are considered. (2) Features are selected by batch rather than individually in the algorithm framework. Thus, the interactions between features are considered and the optimal feature subset can be guaranteed. In addition, this study presents a theoretical justification. Empirical experiments demonstrate the effectiveness and efficiency of the two algorithms in comparison with some state-of-the-art feature ranking methods through a set of real-world gene expression data sets.

3. Clustering Approach

Clustering is the process of grouping a set of data elements into multiple groups or clusters so that objects/elements within a cluster have high similarity, but are very dissimilar to objects in others clusters. Dissimilarities and similarities are accessed based on the attribute values describing the objects and often involve distance measures. Clustering of objects is as ancient as the human need for describing the salient characteristics of men and objects and identifying them with a type. Therefore, it embraces various scientific disciplines: from mathematics and statistics to biology and genetics, each of which uses different terms to describe the topologies formed using this analysis. From biological “taxonomies”, to medical “syndromes” and genetic “genotypes” to manufacturing “group technology” The problem is identical: forming categories of entities and assigning individuals to the proper groups within it.

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. When dealing with larger datasets, organizations face difficulties in being able to create, manipulate, and manage big data. Big data is particularly a problem in business analytics because standard tools and procedures are not designed to search and analyze massive datasets. Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Although big data doesn’t refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data.The defined system architecture mainly focus on the following attributes of big data.



Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

Clustering methods can be divided into two basic types: hierarchical and partitional clustering. Within each of the types there exists a wealth of subtypes and different algorithms for finding the clusters. Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters. The clustering methods differ in the rule by which it is decided which two small clusters are merged or which large cluster is split. The end result of the algorithm is a tree of clusters called a dendrogram, which shows how the clusters are related. By cutting the dendrogram at a desired level a clustering of the data items into disjoint groups is obtained. Partitional clustering, on the other hand, attempts to directly decompose the data set into a set of disjoint clusters. The criterion function that the clustering algorithm tries to minimize may emphasize the local structure of the data, as by assigning clusters to peaks in the probability density function, or the global structure. Typically the global criteria involve minimizing some measure of dissimilarity in the samples within each cluster, while maximizing the dissimilarity of different clusters.

The hierarchical agglomerative clustering methods are most commonly used. The construction of an hierarchical agglomerative classification can be achieved by the following general algorithm.

1. Find the 2 closest objects and merge them into a cluster
2. Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.
3. If more than one cluster remains , return to step 2

4. Applications

Data clustering has immense number of applications in every field of life. One has to cluster a lot of thing on the basis of similarity either consciously or unconsciously. So the history of data clustering is old as the history of mankind. cluster analysis is used to describe and to make spatial and temporal comparisons of communities (assemblages) of organisms in heterogeneous environments; it is also used in plant systematics to generate artificial phylogenies or clusters of organisms (individuals) at the species, genus or higher level that share a number of attributes. clustering is used to build groups of genes with related expression patterns (also known as coexpressed genes) as in HCS clustering algorithm . Often such groups contain functionally related proteins, such as enzymes for

a specific pathway, or genes that are co-regulated. High throughput experiments using expressed sequence tags (ESTs) or DNA microarrays can be a powerful tool for genome annotation, a general aspect of genomics.

On PET scans, cluster analysis can be used to differentiate between different types of tissue and blood in a three-dimensional image. In this application, actual position does not matter, but the voxel intensity is considered as a vector, with a dimension for each image that was taken over time. This technique allows, for example, accurate measurement of the rate a radioactive tracer is delivered to the area of interest, without a separate sampling of arterial blood, an intrusive technique that is most common today. Cluster analysis can be used to analyse patterns of antibiotic resistance, to classify antimicrobial compounds according to their mechanism of action, to classify antibiotics according to their antibacterial activity. Cluster analysis is widely used in market research when working with multivariate data from surveys and test panels. Market researchers use cluster analysis to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers, and for use in market segmentation, Product positioning, New product development and Selecting test markets.

5. Conclusion

As per above defined concepts the handling of big data can be possible with the help of clustering method which distribute large data amongst different clusters that are used further to classify the attributes. To mine data according to its characteristics the procedure has to be followed described in the figure. With the help of clustering anyone can able to divide data into different groups which are independent to others. By firing quires on such clusters effective data get mined.

ACKNOWLEDGEMENT

Here with I am giving my heart full thanks to Prof. Kulkarni P. R. and Kulkarni V.B for their valuable guidance and support. I am also thankful to friends and all departmental staff members of college for their help.

REFERENCES

[1]. Zhixu Li, Mohamed A. Sharaf, Laurianne Sitbon, Xiaoyong Du, and Xiaofang Zhou, "CoRE: A Context-Aware Relation Extraction Method for Relation Completion ", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 4, April 2014.
[2]. Evan Wei Xiang, Bin Cao, Derek Hao Hu, and Qiang Yang, "Bridging Domains Using World Wide Knowledge for Transfer Learning ", IEEE Transactions On Knowledge And Data Engineering,Vol. 22, No. 6, June 2010.
[3]. Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios, " Duplicate Record Detection: A Survey", IEEE Transactions On Knowledge And Data Engineering, Vol. 19, No. 1, January 2007.
[4] Yanfeng Zhang, Shimin Chen, Qiang Wang, and Ge Yu,"i2 MapReduce: Incremental MapReduce for Mining Evolving Big Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 7, July 2015.
[5]. Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen, Susumu Horiguchi,"A Hidden Topic-Based Framework toward Building Applications with Short Web Documents ", IEEE Transactions On Knowledge And Data Engineering, Vol. 23, No. 7, July 2011.
[6]. Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana, "Relevance Feature Discovery for Text Mining ", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 6, June 2015.
[7]. Mirjana Mazuran, Elisa Quintarelli, and Letizia Tanca, "Data Mining for XML Query-Answering Support ", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 8, August 2012.
[8]. Oznur Kirmemis Alkan, Pinar Karagoz , "CRoM and HuspExt: Improving Efficiency of High Utility Sequential Pattern Extraction ", IEEE Transactions On Knowledge And Data Engineering,May. 2012.
[9]. Marc Sole and Josep Carmona, "Region-Based Foldings in Process Discovery ", IEEE Transactions On Knowledge And Data Engineering,Vol. 25, No. 1, January 2013.
[10]. Saima Aman, Yogesh Simmhan, and Viktor K. Prasanna, "Holistic Measures for Evaluating Prediction Models in Smart Grids ", IEEE Transactions On Knowledge And Data Engineering,Vol. 27, No. 2, February 2015.

[11] Huan Ke, , Peng Li, Song Guo, Senior and Minyi Guo, ", On Traffic-Aware Partition and Aggregation in MapReduce for Big Data Applications ", IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 3, March 2014.
[12] Fan Zhang, Junwei Cao Wei tan, Samee U. Khan,Keqin Li, And Albert Y. Zomaya, ",Evolutionary Scheduling of Dynamic Multitasking Workloads for Big- Data Analytics in Elastic Cloud ", IEEE Transactions On Emerging Topics In Computing , Volume 2, No. 3, September 2014.
[13] Chamikara Jayalath, Julian Stephen, and Patrick Eugster, ", From the Cloud to the Atmosphere: Running MapReduce across Data Centers ", IEEE Transactions On Computers, Vol. 63, No. 1, January 2014.
[14] Yijie Wang, Xingkong Ma, ",A General Scalable and Elastic Content-based Publish/ Subscribe Service ", IEEE Transaction On Parallel And Distributed Systems, , Vol. 6, No. 1, January 2013.
[15]. Daisuke Takaishi, Hiroki Nishiyama, Nei Kato, And Ryu Miura "Toward Energy Efficient Big Data Gathering in Densely Distributed Sensor Networks ", IEEE Transactions On Emerging Topics In Computing , Volume 2, No. 3, September 2014.