A distributed k-mean clustering algorithm for cloud data mining

Renu Asnani

Computer Science department,
Rajiv Gandhi Proudyogiki Vishwavidyalaya
Address-E-7/54 Ashoka Society, Arera Colony, Bhopal(M.P.)

Abstract—cloud computing is a new generation computational manner. In this technique the way of computing is transformed into distributed computing, therefore the concept of cloud is used where the efficient computational experience and scalable computing is required. Not only has the cloud provided the scalable computing it also provides the scalable storage units. Therefore a significant amount of data is arrived in these units to handle them, among various kinds of storage the unstructured data storage is also a part of entire cloud data storage i.e. social networking text data, images and other electronic form of documents. Thus in this presented work a survey is introduced for cloud data storage, and their cluster analysis for utilizing the data into various business intelligence applications. in addition of that a new model of cluster analysis of data is proposed which provides the clustering as service.

Keywords— *cloud computing, cluster analysis, data mining, social networking, text clustering.*

I. INTRODUCTION

Data mining [1] is a technique of analysing data and extraction of meaningful patterns from the raw sets of data. The meaningful is termed here to indicate the patterns or knowledge recovered from the training samples which is further used to identify the similar pattern which belongs to the learned pattern. In data mining two main kinds of learning techniques are observed namely supervised learning technique and unsupervised learning technique. These learning models are used to evaluate data and create a mathematical model for utilizing to identify the similar data patterns arrived for classifying them in some pre-fined groups .

In supervised learning technique the data is processed with their class labels and here the class labels are working as teacher for learning algorithm. On the other hand in unsupervised learning technique the data not contains the class labels to utilize as the teacher. Therefore using the similarity and dissimilarity of the input training samples the data is categorized. Therefore the supervised learning processes are known as the classification of data and the unsupervised learning techniques are supporting the cluster analysis of data.

In this presented work the unlabelled data is used for analysis therefore the data analysis technique is used as the cluster analysis. Clustering is the unsupervised classification of patterns or input samples. That can used classify observations, data items, or feature vectors into groups. [3] These groups are in data mining is known as the cluster analysis of data. In the case of clustering, the problem is to group a given collection of unlabelled patterns into meaningful clusters. In a sense, labels are associated with clusters also, but these category labels are data driven; that is, they are obtained solely from the data.

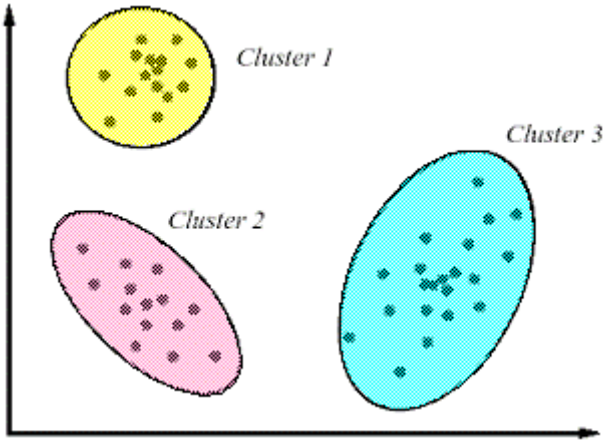

Figure 1 clustering example

The clustering example of the data is given using the figure 1. In this diagram the three different groups of points are given which is termed as the clusters and points are the objects available for classify in unsupervised manner. By nature the in different applications are found in both the manners structured and unstructured. The given work is intended to obtain a unique technique which helps to find similar patterns in unstructured data. This section provides the overview of the introduction of data mining and the selected domain for study in data mining. In the next section the different kinds of clustering algorithms are learned for understanding the technique behind the cluster analysis.

II. CLUSTERING BACKGROUND

This section provides the understanding of the clustering and their different available techniques. in addition of that it also includes the applications and need of cluster analysis.

A. Clustering technique

There are a significant amount of clustering algorithms and methods are available some essential techniques are described [4]:

Partitioning Method: in this clustering approach the n numbers of data or objects are provided, and k number of partitions are required from the data but the number of partition is such that k≤n. This means the partitioning algorithm will generate k partitions satisfying below condition:

a. Each group have minimum one object.
b. Each object should be a member of exactly one group.

Hierarchical Methods: Hierarchical method generates hierarchically manner of clusters organization. That can be achieved using the following manner:

a. **Agglomerative Approach:** It follows the bottom-up approach. Firstly, it generates separate group for each object of data. Next, it merges these groups on the basis of closer similarities. This process is repeated till the entire crowd of groups are not combined in a single or until the termination condition holds.
b. **Divisive Approach:** It follows the top-down approach. Process starts with a single cluster having all data objects. Then, it continues splitting the bigger clusters into smaller ones. This process continues until the termination condition holds. This method is inflexible that is after merge or split is finished, It can never be negated.

Density-Based Methods: This technique uses the perception of density. The main design is to keep expanding the cluster until the density of neighbourhood reaches certain threshold i.e. within a given cluster, the radial span of a cluster must possess certain number of points for each data points.

Grid-Based Method: This method quantizes the object space into a large no. of cells which together nurture a grid. The method having the flowing advantages:

a) Primary benefit the method provides is its fast processing.
b) The only dependability is relying upon the no. of cells in object space.

Model-Based Methods: In Model-based scheme, a model can be conjectured for every cluster along with that; it then identifies data fitting best into that model. This method supplies a means to automatically reveal number of clusters derived from the standard statistics, considering outlier or noise. As a result, it creates robust clustering methods.

Constraint-Based Method: It performs clustering on the basis of constraints either application oriented or user oriented. These constraints are actually the prospect or properties of the desired clustering results. These constraints make communication with the clustering process easy.

B. Applications of Cluster Analysis

There are a number of clustering applications are available in literature some of them are reported in this section [3][4].

a. Cluster Analysis is used in a number of applications for instance marketplace investigations, pattern evaluation, data scrutiny, and image processing.
b. It provides an aid to marketers to group their consumers on different basis. They can also group them on the basis of their purchase patterns.
c. It is also beneficial in bio field, to classify and categorize plants and animal taxonomy on the basis of their genes and functionalities.
d. In an earth observation database, it can be much helpful in categorizing similar lands. Also, it can be beneficial in dividing, houses, plots, and flats on the basis of geographical areas.
e. It can be helpful for information discovery in web data.
f. "Cluster Analysis" can work as a data mining function, to have glances of data to examine every cluster.

C. Requirements of Clustering

Following are some fundamental need of clustering in data mining [5]:

1. **Scalability** – To handle large set of database, scalability must be on high in a clustering technique, to deal with changes.
2. **Capability of handling diverse kind of attributes** – Ideal clustering algorithm must have ability to handle a diverse variety of attribute.
3. **Cluster Discovery with attribute shape** – The algorithm must be able to trace clusters of any shape and dimension.
4. **High dimensionality** – The algorithm must competent towards supervising data of squat dimensions as well as high dimensional space.
5. **Ability to deal with noisy data** – It is always possible to have noisy, incomplete entries into the database. The algorithm must be less sensitive towards these.
6. **Interpretability** - Outcomes of clustering must be easily interpretable, understandable and extensive.

This section provides the overview of the clustering analysis and the usages.

III. LITERATURE SURVEY

This section provides the recent studies over the cluster analysis techniques and their use in cloud computing domain.

With the rapid growth of emerging applications like social network analysis, semantic Web analysis and bioinformatics network analysis, a variety of data to be processed continues to witness a quick increase. Effective management and analysis of large-scale data poses an interesting but critical challenge. Recently, big data has attracted a lot of attention from academia, industry as well as government. ***Changqing Ji et al [6]***

introduce several big data processing techniques' from system and application aspects. First, from the view of cloud data management and big data processing mechanisms, we present the key issues of big data processing, including cloud computing platform, cloud architecture, cloud database and data storage scheme. Following the MapReduce parallel processing framework, and then introduce MapReduce optimization strategies and applications reported in the literature. Finally, we discuss the open issues and challenges, and deeply explore the research directions in the future on big data processing in cloud computing environments.

Yi Zhuang et al [7] present an efficient and robust content-based large medical image retrieval method in mobile Cloud computing environment, called the MIRC. The whole query process of the MIRC is composed of three steps. First, when a clinical user submits a query image Iq, a parallel image set reduction process is conducted at a master node. Then the candidate images are transferred to the slave nodes for a refinement process to obtain the answer set. The answer set is finally transferred to the query node. The proposed method including a priority-based robust image block transmission scheme is specifically designed for solving the instability and the heterogeneity of the mobile cloud environment, and an index support image set reduction algorithm is introduced for reducing the data transfer cost involved. Authors also propose a content-aware and bandwidth-conscious multi-resolution based image data replica selection method and a correlated data caching algorithm to further improve the query performance. The experimental results show that the performance of given approach is both efficient and effective, minimizing the response time by decreasing the network transfer cost while increasing the parallelism of I/O and CPU.

Data mining is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. It is the extraction of information from huge volume of data or set through the use of various data mining techniques. The data mining techniques like clustering, classification, neural network, genetic algorithms help in finding the hidden and previously unknown information from the database. Cloud Computing is a web-based technology whereby the resources are provided as shared services. The large volume of business data can be stored in Cloud Data centres with low cost. Both Data Mining techniques and Cloud Computing helps the business organizations to achieve maximized profit and cut costs in different possible ways. The main aim of ***Astha Pareek et al [8]*** is to implement data mining technique in cloud computing using Google App Engine and Cloud SQL.

Cloud computing is a powerful technology to perform massive-scale and complex computing. It eliminates the need to maintain expensive computing hardware, dedicated space, and software. Massive growth in the scale of data or big data generated through cloud computing has been observed. Addressing big data is a challenging and time demanding task that requires a large computational infrastructure to ensure successful data processing and analysis. The rise of big data in cloud computing is reviewed by ***Ibrahim Abaker et al [9]***. The definition, characteristics, and classification of big data along with some discussions on cloud computing are introduced. The relationship between big data and cloud computing, big data storage systems, and Hadoop technology are also discussed. Furthermore, research challenges are investigated, with focus on scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy, legal and regulatory issues, and governance. Lastly, open research issues that require substantial research efforts are summarized.

While high-level data parallel frameworks, like MapReduce, simplify the design and implementation of large-scale data processing systems, they do not naturally or efficiently support many important data mining and machine learning algorithms and can lead to inefficient learning systems. To help fill this critical void, ***Yucheng Low et al [10]*** introduced the GraphLab abstraction which naturally expresses asynchronous, dynamic, graph-parallel computation while ensuring data consistency and achieving a high degree of parallel performance in the shared-memory setting. In this paper, they extend the GraphLab framework to the substantially more challenging distributed setting while preserving strong data consistency guarantees. We develop graph based extensions to pipelined locking and data versioning to reduce network congestion and mitigate the effect of network latency. And also introduce fault tolerance to the GraphLab abstraction using the classic Chandy-Lamport snapshot algorithm and demonstrate how it can be easily implemented by exploiting the GraphLab abstraction itself. Finally, evaluate given distributed implementation of the GraphLab abstraction on a large Amazon EC2 deployment and show 1-2 orders of magnitude performance gains over Hadoop-based implementations.

IV. PROPOSED WORK

The cloud computing is providing ease in computation and large scale data storage. This manner of computing provides the efficient and scalable storage and computing experience for the remote users with any installation and maintenance. Therefore a huge amount of recently developed applications are getting benefits from the cloud computing style. Due to scalable storage the organizations and individuals are outsource their data to these units for long time preservation. Thus the proposed work is dedicated to perform the cluster analysis on the stored on the cloud.

Due to interaction of different kinds of the organizational data the storage contains a significant amount of unstructured data. Thus the proposed method is dedicated to find the method of text

clustering. The proposed text clustering includes the classification of text according to their orientation in terms of the social networking data. The term orientation is used to find the user mood in the available text communications. Therefore the twitter dataset is used for the text data analysis in the cloud computing environment.

Therefore the following objectives are involved to find the text clustering technique.

1. Study of clustering techniques: in this phase the different kinds of clustering technique is studied for finding the efficient and accurate scheme. Among a number of clustering schemes here the K-means algorithm is chosen for cluster analysis because the k-means clustering provide the efficient clustering as compared to other similar clustering scheme.

2. Study of text clustering techniques: in this phase different text analysis techniques are studied. Therefore the utilization of the k-mean algorithm is performed on the text data for classifying the data according to the hidden sentiments in text. Therefore some additional modifications are placed on the k-means algorithm for achieving the text analysis according to hidden sentiments using the unsupervised manner.

3. Design and implementation of the text clustering algorithm: after improving the design of the traditional k-means algorithm for the sentiments based text analysis in cloud environment the implementation of the given technique is performed using JAVA technology and for deploying the clustering as a service the openshift environment is utilized.

4. Performance analysis of the proposed technique: in this phase the implemented technique is evaluated over different performance parameters i.e. accuracy, error rate, memory and time consumption during analysis. According to the obtained performance the results are provided in near future.

In order to provide the sentiments based text clustering using k-means algorithm an system architecture is proposed in this paper. The proposed system architecture is given using the figure 2 and their different components are reported as follows. The given data model provides the overview of the proposed text clustering technique for social network text clustering technique. Therefore in first the twitter dataset is used to produce as input for the system. That is stored in the Hadoop data storage. The stored data in Hadoop is further pre-processed for reducing the noise in data additionally to remove the unwanted contents and special symbols from the data. After pre-processing of data the data is temporarily stored in row manner.

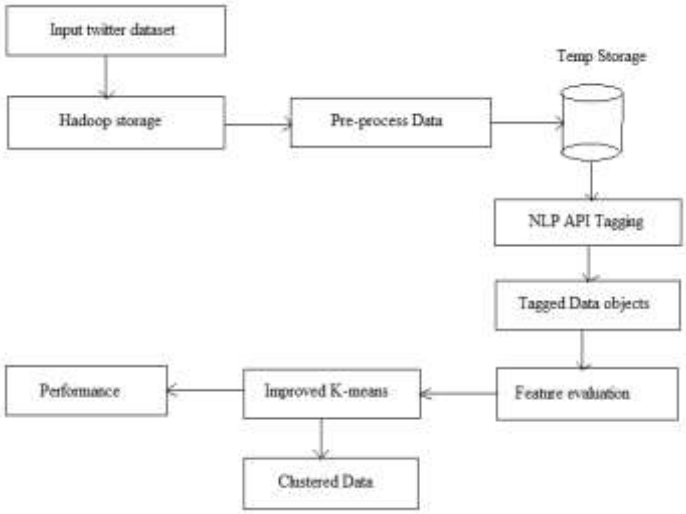

Figure 2 system architecture

Further the data is used with the NLP tool API which is used to produce the tag of data. These tags are provided in terms of their identifiers such as noun, pro-noun and others. The outcome of the NLP tool is rows with the tagging. The tagged data is used for feature extraction thus the data features are recovered and used with the improved k-means clustering. The improved k-means clustering is optimized in order to accept the text data features and provides the cluster analysis of text data and their performance of clustering.

This section provides the understanding of the work involved in the proposed work and the model by which the proposed text analysis is performed.

V. CONCLUSIONS

In this paper an overview of data mining and their techniques are provided first. Further the data mining technique is concentrated on the cluster analysis of data. Thus the different available techniques of clustering and their applications are learned. Additionally the recently developed techniques for the clustering of different kinds of data are also learned in this paper. Finally the key objectives are established and a new model for sentiments based text clustering data model is proposed. The proposed model is in near future implemented and their design and implementation with the obtained experimental results are provided.

REFERENCES

[1] Data mining Concepts and Techniques, Second Edition, Jiawei Han and Micheline Kamber, http://akademik.maltepe.edu.tr/~kadirerdem/772s_Data.Mining.Concepts. and .Techniques.2nd.Ed.pdf.

[2] "Data Mining - Classification & Prediction Introduction", http://www.idc-online.com/technical_references/pdfs/data_communications/Data_Mining_Classification_Prediction.pdf

[3] Data Mining - Cluster Analysis, http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

[4] Pavel Berkhin, "Survey of Clustering Data Mining Techniques", Accrue Software, 1045 Forest Knoll Dr., San Jose, CA, 95129

[5] Marina Meil˘a, "The stability of a good clustering", Journal of Artificial Intelligence Research 1 (1993) 1-15 Submitted 6/91; published 9/91

[6] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li, "Big Data Processing in Cloud Computing Environments", 2012 International Symposium on Pervasive Systems, Algorithms and Networks

[7] Yi Zhuang, Nan Jiang, Zhiang Wu, Qing Li, Dickson K.W. Chiu, Hua Hu, "Efficient and robust large medical image retrieval in mobile cloud computing environment", 2013 Elsevier Inc. All rights reserved.

[8] Astha Pareek, Manish Gupta, "Review of Data Mining Techniques in Cloud Computing Database", International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970), Volume-2 Number-2 Issue-4 June-2012

[9] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, Samee Ullah Khan, "The rise of "big data" on cloud computing: Review and open research issues", & 2014 Elsevier Ltd. All rights reserved.

[10] Yucheng Low, Joseph Gonzalez, Aapo Kyrola, Danny Bickson, Carlos Guestrin, Joseph M. Hellerstein, "Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud", Proceedings of the VLDB Endowment, Vol. 5, No. 8 Copyright 2012