CLUSTERING ALGORITHMS AND THEIR APPLICATIONS IN CLOUD COMPUTING ENVIRONMENT

U.S. Patki
Dept of Computer Science & IT, Science College, Nanded (M.S.), India
patkiulhas@gmail.com

Manuscript History
Number: IRJCS/RS/Vol.04/Issue04/APCS10081
Received: 12, March 2017
Final Correction: 20, April 2017
Final Accepted: 27, April 2017
Published: April 2017

Abstract—Cloud computing is Internet-based computing that provides shared computer processing resources and data to computers and other devices on demand. Cloud computing is the hottest purpose built architecture created to support computer users. The cloud addresses three main areas of operation SaaS (software-as-a-service), PaaS (platform-as-a-service) and IaaS (infrastructure as a service). The large amount of data can be stored into cloud Data centers with low cost. We Integrate Data Mining and Cloud Computing to provide a quick access to this large volume amount of data on cloud. This paper aimed to study Clustering algorithm which can be applicable in cloud computing. This paper also describes the role of soft clustering in Cloud computing environment.

Keywords—Cloud Computing, Clustering, Soft Computing, Hard Computing, K-means, FCM

I. INTRODUCTION
Cloud computing is the delivery of computing services—servers, storage, databases, networking, software, analytics and more—over the Internet (“the cloud”). Cloud has three major components.

i) Client Computers
ii) Distributed Servers
iii) Data Centers

Clients are the device that the end user interacts with cloud. Distributed servers are in geographically different places, but servers acts as if they are working next to each other. Data center is a collection of servers where application is placed and is accessed via internet.
SERVICE MODELS:
Service of Cloud Consists of three main models.
SaaS (Software as Service): Required software, operating system and network in the form of software is provided.
PaaS (Platform as Service): Operating System and Network is Provided.
IaaS (Infrastructure as a Service): Just Network is provided.

II CLUSTERING
Clustering is a most popular data mining technique used to find useful unknown pattern from data in large repository. Clustering is Grouping of data into different clusters such that elements belongs to same cluster are most similar while elements belongs to different cluster are dissimilar. Basically Clustering methods are divided into two broad categories.
   i) Hard Clustering
   ii) Soft Clustering
In Hard Clustering, each document can belongs to only one Cluster. Hard Clustering is also known as exclusive clustering. In Soft Clustering Same document can belong to more than one group. It is also known as Overlapping Cluster technique.

III. HARD CLUSTERING USED IN CLOUD COMPUTING.
A storage cloud consists of data coming from different sources and is of different types. The traditional data mining technique used to handle the data has many limitations. As the data available on cloud storage is heterogeneous, the traditional technique fails to handle such data[1] Clustered storage is the use of two or more storage servers working together to increase performance, capacity, or reliability. Clustering distributes work loads to each server, manages the transfer of workloads between servers, and provides access to all files from any server regardless of the physical location of the file. Two basic clustered storage architectures exist, known as tightly coupled and loosely coupled.
Mrs. Dhanamma Jagl[2] proposed a model for clustering a cloud data. The proposed model contains following steps
1. Extract SaaS data
2. Transform SaaS data
3. Load in RDBMS/MDBMS format
4. Apply clustering algorithm
5. Obtain appropriate cluster
In step 4 most popular K-means Clustering and Hierarchical Clustering algorithms are used. These algorithms are Hard Clustering algorithms in which data elements will belongs to one and only one cluster. Due to Heterogeneous nature of data in Cloud Computing environment, Soft Clustering algorithm like Fuzzy C-means are more effective.

**IV. PROPOSED SOFT CLUSTERING ALGORITHMS**

In Soft Clustering Algorithm, Fuzzy C-means, data can belong to more than one clustering. The most effective feature that makes fuzzy clustering most popular is “overlapping”. In fuzzy Clustering Overlapping is permitted. Each document has assigned a fuzzy membership value. This value represents the presence the degree belongings of an element into a cluster.

**ALGORITHMIC STEPS FOR FUZZY C-MEANS CLUSTERING**

Let \( X = \{x_1, x_2, x_3, \ldots, x_n\} \) be the set of data points and \( V = \{v_1, v_2, v_3, \ldots, v_c\} \) be the set of centers.

1) Randomly select ‘c’ cluster centers.
2) Calculate the fuzzy membership \( \mu_{ij} \) using:

\[
\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}}{d_{ik}} \right)^{2/(m-1)}}
\]

\[
v_j = \frac{\sum_{i=1}^{n} \left( \mu_{ij} \right)^m x_i}{\sum_{i=1}^{n} \left( \mu_{ij} \right)^m}, \forall j = 1, 2, \ldots, c
\]

3) Compute the fuzzy centers \( v_j \) using:

4) Repeat step 2) and 3) until the minimum ‘J’ value is achieved or \( ||U^{(k+1)} - U^{(k)}|| < \beta \).

where,
- \( k' \) is the iteration step.
- \( \beta' \) is the termination criterion between \([0, 1]\).
- \( U = (\mu_{ij})_{n \times c} \) is the fuzzy membership matrix.
- \( J \) is the objective function[7].

In Cloud Storage environment as the data is of heterogeneous type, it is very difficult to classify them. The data may contain similarities, overlapping, distinguishes etc. The limitation of hard clustering is that the data will belongs to exactly one group. The hard clustering algorithms are suitable in the situation where overlapping does not exist and data items are exactly dissimilar to each other. But in cloud environment Soft computational techniques are more suitable as data may have relations with more than one group.

**V. CONCLUSIONS**

Cloud storage is a challenging technology which helps large organizations to store and manage their enormous data. Different algorithms are proposed for fast and efficient retrieval of this cloud data. Clustering algorithms such as K-means Partition, Hierarchical Clustering, are used to group similar data on cloud. In the Distributed Environment, very large datasets need to be reduced. Fuzzy C-means, modified version of K-means algorithm, is appropriate in cloud clustering. As data is allowed to be fall in more than one cluster, the data retrieval becomes fast and more accurate.

**REFERENCES**

[7] Dibya Jyoti Bora, D r. Anil Kumar, A Comparative Study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm, Apr 2014.