Enhancement Clustering of Cloud Datasets using Improved Agglomerative Technique

Prof. Madhuri h Parekh
Smt. J.J.Kundaliya Commerce College, Rajkot, Gujarat, India.
Email: madhurisuchak@yahoo.com

----------**ABSTRACT**----------

Cloud computing is the latest technology that delivers computing resources as a service such as infrastructure, storage, application development platforms, software etc. Huge amount of data is stored in the cloud which needs to be retrieved efficiently. In Cloud Computing using of Clustering Process from Heterogeneous Network fetch the data find out the row data. Clustering is consists of many of the same or similar type of Machine. Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram.The retrieval of information from cloud takes a lot of time as the data is not stored in an organized way. Data mining is thus important in cloud computing. We can integrate data mining and cloud computing which will provide agility and quick access to the technology. The integration should be so strong that it will be able to deal with increasing production of data and will help in efficient mining of massive amount of data. In this paper, we provide brief description about cloud computing and clustering techniques. Then, it also describes about cloud data mining. This paper proposes a model that applies traditional hierarchical improved agglomerative clustering algorithm and distributed on heterogeneous network.

Keywords - **Cloud Computing, Clustering, Agglomerative algorithm, Distributed Algorithm**

I. INTRODUCTION TO CLOUD COMPUTING

Cloud computing relies on sharing of resources to achieve coherence and economies of scale, similar to a utility over a network. At the foundation of cloud computing is the broader concept of converged infrastructure and shared services. In Cloud computing using of clustering provide datasets in well organization and fetch data from heterogonous network and distributed over Network.[1] The tree is not a single set of cluster but rather a multilevel hierarchy where cluster at one level are joined as cluster at next level this decide the level or scale of clustering that is most appropriate for application. we have argued that multimode transfer operation have significant impact on performance of data intensive cluster application using of agglomerative enables global control both across and within transfer to optimize the performance. We provide brief description about cloud computing and clustering techniques. it also describes about cloud data mining.

In Cloud Computing using of Clustering Process from Heterogeneous Network fetching the data find out the row data. Using of agglomerative approach Find the similarity or dissimilarity between every pair of objects in data set. Group the objects into binary, hierarchical cluster tree. Determine where to cut the hierarchical tree into cluster.[2]

Major advantages & disadvantage of using Cloud Computing.

Fig-2. Advantage and disadvantage of Cloud Computing

1.2 Cloud Computing Architecture

Cloud computing system can be divided into 2 sections, one is Front end and another one is Backend. The front end is the interface for the user example client and the back end is the cloud section for the whole system. Front end and Backend connected with each other via network like internet.

There are 3 types of layers related to Cloud services

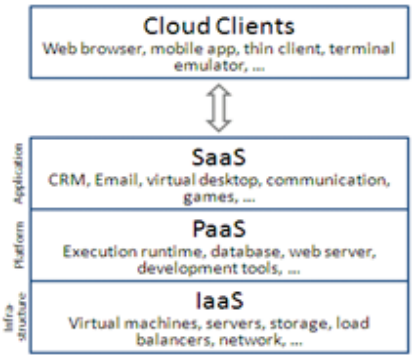

Fig-3 Cloud Architecture

Cloud Infrastructure Services (Infrastructure as a Service "IaaS"):This service provider bears all the cost of servers, networking equipment, storage, and back-ups.Rather than purchasing servers, software, data-center space or network equipment, clients instead buy those resources as a fully outsourced service.[3]

Cloud Application Services (Software as a Service "SaaS"):This service provider will give your users the service of using their software, especially any type of applications software. Google Apps., Salesforce.com,and various other online applications are use cloud computing as Software-As-Service (SAAS) model.[3]

Cloud Platform Services (Platform as a Service "PaaS"): Platform cloud services are used by software developers to build new applications and by operations managers to manage their application, compute and storage cloud services.[3]

II. BACKGROUND WORK

Agglomerative hierarchical clustering
Each object initially represents a cluster of its own. Then clusters are successively merged until the desired cluster structure is obtained.

The result of the hierarchical methods is a dendrogram, representing the nested grouping of objects and similarity levels at which groupings change. A clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level.

Single-link clustering (also called the connectedness, the minimum method or the nearest neighbor method) — methods that consider the distance between two clusters to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the data consist of similarities, the similarity between a pair of clusters is considered to be equal to the greatest similarity from any member of one cluster to any member of the other cluster.

Complete-link clustering (also called the diameter, the maximum method or the furthest neighbor method) - methods that consider the distance between two clusters to be equal to the longest distance from any member of one cluster to any member of the other cluster.

Complete-link clustering (also called the diameter, the maximum method or the furthest neighbor method) - methods that consider the distance between two clusters to be equal to the longest distance from any member of one cluster to any member of the other cluster.

Average-link clustering (also called minimum variance method) – methods that consider the distance between two clusters to be equal to the average distance from any member of one cluster to any member of the other cluster. Such clustering algorithms may be found.[3]

The disadvantages of the single-link clustering and the average-link clustering can be summarized as follows

Single-link clustering has a drawback known as the "chaining effect": A few points that form a bridge between two clusters cause the single-link clustering to unify these two clusters into one.

Average-link clustering may cause elongated clusters to split and for portions of neighboring elongated clusters to merge.

The complete-link clustering methods usually produce more compact clusters and more useful hierarchies than the single-link clustering methods, single-link methods are more versatile.

Advantage of hierarchical Method
Versatility — The single-link methods, for example, maintain good performance on data sets containing non-isotropic clusters, including well separated, chain-like and concentric clusters.

Multiple partitions — hierarchical methods produce not one partition, but multiple nested partitions, which allow different users to choose different partitions, according to the desired similarity level. The hierarchical partition is presented using the dendrogram.

Limitation of hierarchical Methods
Inability to scale well—The time complexity of hierarchical algorithms is at least O(m2) (where m is the total number of instances), which is non-linear with the number of objects. Clustering a large number of objects using a hierarchical algorithm is also characterized by huge I/O costs. Hierarchical methods can never undo what was done previously. Namely there is no back-tracking capability.

III. PROPOSED WORK

Agglomerative Clustering Algorithm

The algorithm forms clusters in a bottom-up manner, as follows:

1. Initially, put each article in its own cluster.
2. Among all current clusters, pick the two clusters with the smallest distance.
3. Replace these two clusters with a new cluster, formed by merging the two original ones.
4. Repeat the above two steps until there is only one remaining cluster in the pool.

Thus, the agglomerative clustering algorithm will result in a **binary** cluster tree with single article clusters as its leaf nodes and a root node containing all the articles.

In the clustering algorithm, we use a distance measure based on log likelihood. For articles A and B, the distance is defined as

The log likelihood LL(X) of an article or cluster X is given by a unigram model:

$$LL(X) = \log \prod_{w \in X} p_X(w)^{c_X(w)} = \sum_{w \in X} c_X(w) \log c_X(w) - N_X \log N_X$$

Here, $c_X(w)$ and $p_X(w)$ are the count and probability, respectively, of word *w* in cluster *X*, and N_X is the total number of words occurring in cluster *X*.

Notice that this definition is equivalent to the weighted information loss after merging two articles:

$$d'(A, B) = (N_A + N_B)H(A \cup B) - (N_A H(A) + N_B H(B)) \quad (2)$$

where

$$H(X) = -\sum_{w \in X} P_X(w) \log P_X(w) \ .$$

To avoid expensive log likelihood recompilation after each cluster merging **step**, we define the distance between two clusters with multiple articles as the maximum pair wise distance of the articles from the two clusters:

$$d(C_1, C_2) = \max_{A \in C_1, B \in C_2} d(A, B) \quad (3)$$

where C_1 and C_2 are two clusters, and *A*, *B* are articles from C_1 and C_2, respectively.

Once a cluster tree is created, we must decide where to slice the tree to obtain disjoint partitions for building cluster-specific LMs. This is equivalent to choosing the total number of clusters. There is a tradeoff involved in this choice. Clusters close to the leaves can maintain **more** specifics of the word distributions. However, clusters close to the root of the tree yield LMs with more reliable estimates, because of the larger amount of data.

We roughly optimized the number of clusters by evaluating the perplexity of the Hub4 development **test** set. We created sets of 1, 5, 10, 15, and 20 article clusters, by slicing the cluster tree at different points. A back off trigram model was built for each cluster, and interpolated with a trigram model derived from all articles for smoothing, to compensate for the different amounts of training data per cluster. Then, the set of LMs that maximizes the log likelihood of the Hub4 development data was selected. Given a cluster model set $LM = \{LM_i\}$, the test set log likelihood was obtained as an approximation to the mixture-of-clusters model:

$$P(w \mid LM) = \sum_i P(LM_i) * P(w \mid LM_i) \approx P(LM_{i^*}) * P(w \mid LM_{i^*}) \propto P(w \mid LM_{i^*})$$

Where

$$i^* = \arg\max_i P(LM_i \mid \Lambda) \ ,$$

and $P(LM_i)$ and $P(LM_i \mid \Lambda)$ are the prior and posterior cluster probabilities, respectively. In training, *A* is the reference transcript for one story from the Hub4 development data. During testing, *A* is the 1-best hypothesis for the story, as determined using the standard LM. Note that $P(w \mid LM)$ depends on the smoothing **weights** used to compute $P(w \mid LM_i)$, which in turn determine which cluster a story is assigned to, which in turn determines the best smoothing weights. Therefore, we jointly optimize smoothing and cluster assignment in an iterative procedure. First, the posterior probabilities of the smoothed cluster LMs given reference transcripts for a story were calculated. Then, stories with the highest posterior probability of a *same* cluster LM were merged. The interpolation weight for the cluster LM and the general LM was tuned by maximizing the likelihood of the segments in the story cluster corresponding to the cluster LM. These steps were iterated until all cluster assignments became stable and the interpolation weights converged. [3]

IV. CONCLUSION

In Cloud Computing using of Clustering Process from Heterogeneous Network fetching the data find out the row data. Using of agglomerative approach Find the similarity or dissimilarity between every pair of objects in data set, Group the objects into binary, hierarchical cluster tree. Determine where to cut the hierarchical tree into cluster. Data mining is used for extracting potentially useful information from the row data. Data mining techniques are very much needed in cloud computing. Future work on improve the agglomerative algorithm and integrated with hadoop and map reduce technique.[4]

REFERENCES

Journal Papers:

[1] Cloud computing adoption and usage in community colleges. Behaviour & Information Technology , 30 (2), 231–240.

[2] Vouk, M. A. (2008). Cloud Computing – Issues, Research and

[3] Katzan, H. (2010). The Education Value Of Cloud Computing.

[4] Yang Chen-zhu. The Reseach of Data Mining Based on HADOOP[D].ChongQing. Chongqing University 2010.11:pp42~43

[5] Bhupendra Panchal and R. K. Kapoor, "Performance Enhancement of Cloud Computing with Clustering", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-5, June 2013

Books:

[1] Algorithms of clustering data by Anil k jain,Richard c.dubes Michigan state university(by Prentice-Hall,Inc A Division of simon & Schuster Englewood cliffs, New Jersey 07632)