Subject Name -DDM /CS750

Name-Soubhik Baral
Reg No:1925504
CSE/02
Phone-76870357504
Email:soubhikbaral4@gmail.com

Title:
A Fragmentation Method Based on Data Semantics

Abstract

As nowadays there is a need for running organization from different geographical locations the need for distributed database system also rapidly increases day by day. The two main things that the distributed query processing depends for better performance and efficiency are fragmentation method and allocation method. The fragmentation solution is basically used for analyzing query patterns. This methods are not applicable during initial designing of databases ,we can make use of this method only to the existing distributed databases. The semantics or relationship of the fragmented data is not considered in most of the existing methods. While fragmenting if data relationship is considered we can have much better and time effective fragments that can be easily retrieved.
Clustering technique can used to obtain the related data from datasets. In this project clustering method for fragmentation is used. When we use clustered fragments instead of normal fragments ,we get a minimum query retrieval time.

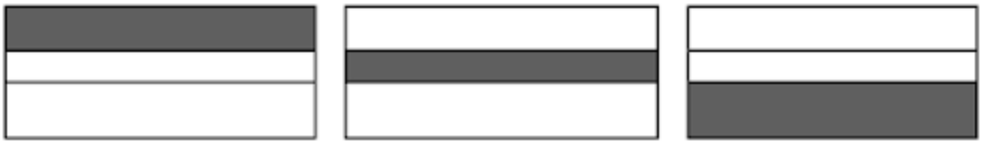
Keywords- Distributed database,Clustered fragments ,Fragmentation

INTRODUCTION

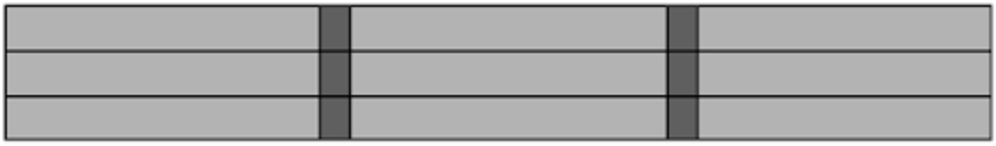
Nowadays,business are stretched globally and are not trapped to a centralized location,it basically spreads across all around the world,so keeping and using of data in a centralized place is a bad idea. A centralized storage area for data faces many issues , a node failure may result in overall failure and rise in traffic in network ,traffic in network leads to rise in access time of the query. This is why a system which basically stores the data in one or many number of nodes or sites is needed and any user can query to any sites for the retrieval of information. Even if one system fails the overall network does not fails and thus helps to maintain fault tolerant system. So which makes it a much better and effective distributed database system. As using a distributed system we can easily access the data much more efficiently and the overall access time also decreases as the no of nodes increases which contain the information in different sites rather than in a specific location. In order to increase the efficiency ,distributed databases system uses multiple technique which includes various types of fragmentation. Many researchers uses many unique allocation technique to enhance the performance. Many of time ,to reduce data loss in distributed system data are often copied to a certain location, basically the idea is to replicate the date and keep into different location so that no data is lost even if there exist some failure and it also helps to store the most fetched data to its local nodes for better performance.

There are different kinds of Fragmentation,which allows the data to divided in certain ways some technique divides the relation in horizontal way,some technique used to break the relation in a vertical way or we can combine both the strategy to find a mix of both the technique. Below we have *Figure 1* that represent each of this fragmentation technique. So we can combine this technique along with the process of copying data in different nodes so that we can get the required query without accessing all the nodes in worst case. This even makes the user access the data in a local nodes or sites rather than a particular centralized access. Along with fragmentation we allocate the data where the data is most commonly accessed. Another method to reduce the overall access time is data replication. It basically means copying important data from one database server to a different server ,this allows the data to be shared between the different servers hence user can share the data

without inconsistent. Another advantage is that user do not have to access the remotely located sites all the time.



(a) Horizontal Fragmentation



(b) Vertical Fragmentation

FIGURE 1:DIFFERENT TYPES OF FRAGMENTATION

There are different types of horizontal fragmentation technique,but those fragmentation technique can only be applied where query information is available in prior. Most of horizontal fragmentation methods, fails to establish a data relationship while making fragments. As the user query related information most of times establishing relationship make the user reduce the access time. In object oriented databases k-means clustering method has been proposed,for clustering based fragmentation. But this technique cannot be applied to all kind of databases. For flexible query answering clustering based fragmentation is also used. In order to apply flexible query answering there has been lot of methods used ,one of the way is to make the data of column relaxed when we are clustering the data.

Mostly the data being fragmented does not count the semantics,as most of the time user query for the data that are linked to one another in some way ,so if we do not make fragment keeping this criteria in mind the performance can be reduced and hence the overall effectiveness of the fragmentation reduces . This drawback can be solved if we keep in mind that the related data must be in a single fragment and which ,in turn will make the fragment resides in a same node,helping the user to access query faster.

Here in this project we solve this drawback using a multivariate clustering algorithm which allows the same kind of data to be in a same fragments .The main idea is to first cluster the data and that is to group the related data ,then use similarity index to join the similar clusters as clusters can be more than the numbers of nodes,similarity index basically joins the similar cluster in order to make the clusters that are relatable. These way we can make better clusters later which will be fragmented to different locations.

LITERATURE REVIEW

There has been lot of studies done extensively for single server systems for quite a few years but now, being in a distributed form of wave many technique related to query answering ,fragmentation has emerged in recent past.

CoBase [2] is expensive approach that are earlier used to answer the query,the method proposed a level of abstraction .This method use a centralized approach and used to face a lot of traffic issues. Halder and cortesi [3] is also a centralized based approach ,it is also expensive and this approach made use of the relaxation technique to address the query ,this is done basically at runtime.

As users query for related attributes together clustering methods have been proposed .A clustering method proposed [4] ,traditional clustering method which predefined the number of clusters earlier and hence the k-means is used. In this method object is represented in form of information. For clustering based fragmentation classical k-means clustering is used. But this technique can used where the overall data is present before hand.

Hadj Mahboubi and Jerome Darmount [5] proposed a logical condition which basically is used on the tuples to form a clusters .Logical tuples are extracted and this logical tuples are used for making clusters. The clustering technique used here is the tradition k-means algorithm.

Lena wiese [6] proposed a algorithm that also clusters the data and the after clustering it fragments the data for answering the query. Fragmentation approach based on Hybridized clustering [9] where we make primary partition and uses these partition to make level like clusters upon this primary clusters.

Various clustering algorithm is used to maintain a relation between the data [7].There are specific algorithm we use depending on the specific data,depending on need of the application. For predefined number of clusters we use k-means algorithm. Random shaped and oval shaped clusters cannot be found using Hierarchical algorithm. DBSCAN algorithm can be used to handle random shaped clusters.

Tian Zhana, Raghu Ramakrishnan and Miron Livny ,[8] BIRCH are used for large datasets it uses one scan for clustering and BIRCH can only be used in numeric database.

METHODOLOGY

In this work,the main motive is to make fragments that takes the idea of keeping the related data in same fragments. The idea is to connect all the data that are related and make use of this relationship so that we can get a better fragments which in turn can be used for faster query answering. There are many algorithm, that takes consideration of this facts and join the data that are common or related in some way. In this work k-prototype algorithm is used as it can handle multi dimensional data.
When we query for information,the information can be retrieved easily if it resides within one fragment. So to achieve this and accomplish fast access and data availability ,clustering technique is used. The cluster that are formed are needed to be distributed across various sites but as we have three nodes in this experiment we will use similarity index to reduce the number of clusters if we have more than three clusters. This process keeps on going until our cluster is equal to the number of nodes. Cluster prototypes are taken once the clusters are formed. This cluster prototype act as a representative for each clusters and are taken into fragmentation phase. The overall method can be divided into different modules ,in first module we will cluster the data using a clustering algorithm after that we fragments the cluster in horizontal way as we generally access the related data more.

1.*Clustering of data*

clustering is an unsupervised problem that is used to find the similarity of the data [10].There are different clustering algorithm available in data mining. Before the fragmentation the data set is clustered initially. In clustering we make groups based on the common criteria. Data with the common kind of values are grouped so that they are placed in the same geographical area which in turn lead to lesser nodes access. Traditionally k-means clustering is quite famous but as k-means cannot be used in categorical data as it cannot quantify a categorical data we have used a modified version of clustering algorithm that is k-prototype algorithm. This algorithm can be used to find the groups among data which will help to make the same kind data fall into the same group. This technique strengthen the idea of keeping the relatable data in the same location.

2.*Fragmenting the Clusters*

At first we used the Jaccard similarity coefficient [11]that is basically used to find the similarities in the data set,the similarity range is measured between the range of 0% to 100% ,0% indicates that the data is not similar and 100% means they are the same. Clusters having higher range are grouped in the same group and then this similar groups of data are placed in the same nodes. Not always similar data are merged that can also me made a different group and can be placed in different locations. This depends on application to application and also on the choice of keeping the data local or not.The formula used to calculate the Jaccard similarity coefficient :

$J(A, B) = A \cap B / A \cup B$

Jaccard similarity coefficient is defined by the size of intersection of the sample dataset divided by the union of the sample dataset. The following ***FIGURE 2*** gives the algorithm that is used to calculate Jaccard similarity coefficient ***FIGURE 3*** shows the merging algorithm that are used to merge the two similar clusters. After making the clusters we need to send the data to different sites but if the number of clusters are more than the number of nodes we need to merge the clusters so the number of cluster is equal to the number of nodes. Here in our experiment we have four clusters and we use the merge algorithm to make it to three clusters as the number of nodes in our experiment is three. ***FIGURE 4*** shows the overall clustering based fragmentation technique.

Algorithm 2:Similarity Algorithm

findSimilarity(Zi , Zj):
Input ← Cluster prototypes
Output ← Similar prototypes
for each representative r i and r j do
S(Z i , Z j) =Z i ∩ Z j / Z i U Z j
return S(Z i , Z j)
end

FIGURE 2

Algorithm 3: Merging Algorithm

merge(C i , C j):
Input ← Clusters
Output ← Merged prototypes
P ij = M i U M j ;
return P ij ;

FIGURE 3

Algorithm 1: Clustering based Fragmentation Algorithm

ClusterFragments(D,n):
Input ← Dataset
Output ← Fragments of the dataset f i where i = n
Let C 1 ,C 2 ,...C i be the clusters obtained from K-prototype
algorithm for each pair of clusters C i and C j do
Result ij = findSimilarity(Z i , Z j)
end
while i ≤ n do
P ij = merge(C i ,C j)
i=i-1;
end

FIGURE 4

Notations	Meanings
D	Dataset
A1,A2…..An	Attribute
n	Number of sites
i	Number of clusters
C_i	ith cluster of the dataset D
Zi	Representative of C_i
S(S i , S j)	Jaccard Coefficient between Si and S j
Result ij	Similarity between the cluster si and sj
M_k	Set of elements in the cluster k

TABLE 1:NOTATIONS USED

Experimentation and Results:

In this experiment we have used PostgreSql to make a distributed system and observe the behavior of both the fragments with respect to the different query. The datasets we used for this experiment are taken from UCI KDD database repository. In ***Table 3*** the details of the datasets is mentioned. In this work,we have three nodes that are used to store the fragments. Two types of fragments were made one was random fragments and the other one is the fragment created using the above clustered method for this experiment. Request from the users can be from any node. In this work we maintain a meta data table ***Table 2*** which shows the max range and min range the nodes have and the IP address of each nodes. ***Table 2*** show the overall structure of the meta data. So the query for a specific information will be send to the nodes that have the range within the max and min value, finally we obtain the combined output from the different nodes for the desired query.

TABLE 3:DATASET:

Dataset	**No of columns**	**No of instances**
Echo cardiogram	13	132
Bank Marketing	17	45211

TABLE 4:METADATA:

Attribute	**Data stored in each node**
IPaddress	IP address of different nodes
Min value	Min value of the column
Max value	Max value of the column

As discussed above we have used the k-prototype algorithm that is used to cluster the datasets so that we can find the relation between the data. The cluster that is formed using the k-prototype algorithm is passed to the merging algorithm ,the merging algorithm checks if the number of cluster formed is greater than the nodes that are there in the distributed system ,as the number of clusters formed is more we then uses the Jaccard Similarity coefficient that finds the similarities between the clusters. The similar clustered are merged ,this process continues until we find the number of cluster

less or equal to that of the nodes or sites. once the similar clusters are found we allocate them to the same nodes for quick access of related data.

A number of queries are given to observe the time required for both the type for fragments. Then the time for each type fragment is calculated for a particular query and are observed. There has been quite a difference in access time for normal query and clustered query. The normal query performance is not so effective whereas clustered query performs much better than the normal query, the access time reduces as we do not have to search for more no of nodes as related data are present in the same node. This is quite time saving when user query for the related query ,and for most of the time user query for the related query hence leads to lesser access time and faster answering of query.

The results shows that if the relationship between the data is maintained and if we keep this related data in single fragment then the query execution time will be minimum. The execution time will be reduced as users query related data most of the times. Figure 1 shows the comparison between Normal query and Clustered query in the Echo Cardiogram dataset. Figure 2 shows that clustered query performs better than normal query in the Bank Marketing dataset. The query excess time for clustered query is less than that of Normal query,hence making the clusters with related data saves time and improve performance.

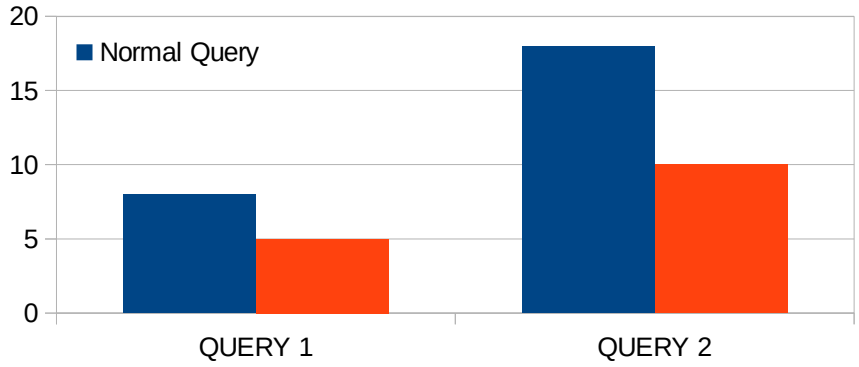


Figure 1:Comparison Chart for Echo cardiogram Dataset.

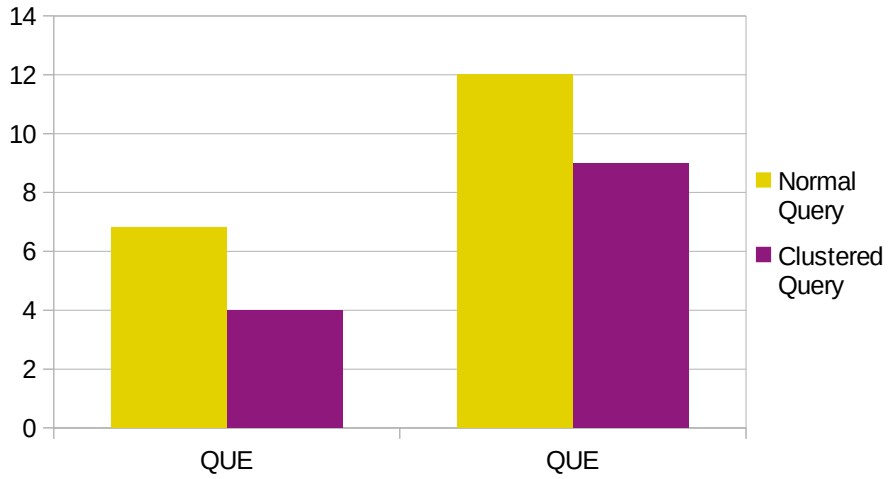


Figure 2:Comparison Chart for Bank Marketing Dataset.

CONCLUSION

The fragmentation based technique that uses k-prototype clustering was implemented to keep the relationship between the data. A fragmentation method is effective only when it can answer the query quickly,hence the success of such method is decided based on how quick it can answer the query. So faster query means that it can answer with minimum node access,by grouping a related data using clustering technique enhances the overall performance. So this experimentation shows that clustered fragments perform better than normal fragment and query answering time is reduced. In short clustering technique lead to better fragments which in turn makes the user retrieve information easily. Future work can be comparing this algorithm with the existing fragmentation technique.

REFERENCES

[1] Ceri S., Negri M., and Pelagatti, "Distributed Database Principles and System", 2008.

[2] Chu WW, Yang H, Chiang K, Minock M, Chow G, Larson ,"CoBase: a Scalable and extensible cooperative information system".JIIS 6(2/3):223–259,1996

[3] Halder R, Cortesi ,"Cooperative query answering by abstract interpretation." In: SOFSEM2011. LNCS, vol. 6543. Springer,Berlin/Heidelberg. Pp 284–296,2011

[4] Zhexue Huang, "Clustering large datasets with mixed numeric and categorical values," Journal Data Mining and Knowledge Discovery,Volume 2, 283 – 304, 1997.

[5] Hadj Mahboubi and Jerome Darmount, "Data Mining-Based fragmentation of XML data warehouses," ACM International Workshop on Data Warehousing and OLAP (CIKM/DOLAP 08), 2008.

[6] Lena Wiese, "Clustering-based fragmentation and data replication for flexible query answering in distributed databases," Journal of Cloud Computing Advances, Systems and Applications, 2014.

[7] Chintan Shah and Anjali Jivani, "Comparison of data mining clustering algorithms," Nirma University International Conference on Engineering (NUiCONE), 2013.

[8] Tian Zhana, Raghu Ramakrishnan and Miron Livny, "BIRCH: An Efficient Data Clustering Method for very Large Databases," SIGMOD'96 Proceedings of the 1996 ACM SIGMOD International Conference on Management of data, 103-114, 1996 .

[9] Sandhya Harikumar and Raji Ramachandran, "Hybridized fragmentation of very large databases using clustering," Signal Processing,

[10] http://home.deib.polimi.it/matteucc/Clustering/tutorial_html.

[11] https://en.wikipedia.org/wiki/Jaccard_index.