International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

A NOVEL APPROACH TOWARDS MULTIPLE ATTRIBUTES BASED CLUSTERING FOR DATASET

Sonam Rani, M.Tech Student ,CSED,KNIT Sultanpur; Dr.Neelendra Badal, Associate Prof, CSED, KNIT Sultanpur;

Abstract

Clustering, on the basis of attributes of datasets, is the main concern of this paper. One type of clustering related to attributes is Single-value Attribute based Clustering. Single attribute value is taken into consideration for the purpose of clustering. With the limitation of using only single value, this approach is common to use for the clustering purpose of datasets. To overcome this limitation, a new approach named as Multiple Attributes based Clustering is being presented in this paper. The proposed Multiple Attributes based Clustering uses the more than single attribute of datasets for clustering in an efficient way and improved manner. With the help of experimental results, it is shown that the proposed Multiple Attributes based Clustering approach performs better than the existing Single-value Attribute based Clustering approach with respect to attribute usage and memory space for the multiple[7] attributes datasets.

Introduction

Organizing data into sensible groupings is one of the most fundamental modes of understanding and learning. The technique used to organize data is known as Clustering[3]. The aim of clustering is to find structure in data and is therefore exploratory in nature. Clustering has a long and rich history in a variety of scientific fields. It is a main task of exploratory data mining and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics, data compression. Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It is an unsupervised learning means no information is provided to the algorithm on which data points belong to which cluster. So, The Clustering basically focuses on organization of objects into groups whose members are similar in some way.

For clustering, many techniques have been introduced. One of the techniques used for clustering is Single-value Attribute based Clustering (SAC) which performs clustering on the basis of single value attribute. It is used to cluster datasets of single value attribute. It is a good technique used for clustering on the basis of attribute but still there is a limitation that existing SAC approach does not support the multiple attributes datasets[4]. Therefore, there is a requirement to introduce a new approach that incorporates the problems in existing SAC approach. This paper proposed the new approach named as Multiple Attributes based Clustering (MAC) to overcome the drawback of existing SAC approach that is efficient clustering of the multiple attributes datasets.

Literature Review

Cluster analysis or Clustering was originated in Anthropology by Driver and Kroeber in 1932 and introduced to psychology by Zubin in 1938 and Robert Tryon[1][9] in 1939 and famously used by Cattell[2] beginning in 1943 for trait theory classification in personality psychology.

Cluster Analysis or Clustering is the process of organizing objects into groups in such a way that objects in one group are more similar to each other than to those in other groups.

Hartigan[6] provides a treatment of modern clustering theory from the statistical point of view. His book contains detailed discussions of a variety of algorithms and their application to real data sets ranging from medical and biological data to political data. Anderberg's book[5] entitled *Cluster Analysis for Applications* deserves a special mention.

Proposed Solution including more than one Attribute:

This Venn diagram is used here to show the different-different clusters and their combinations on the basis of need. Three cases are discussed below to show the best possible combination of clusters on the basis of raised queries.

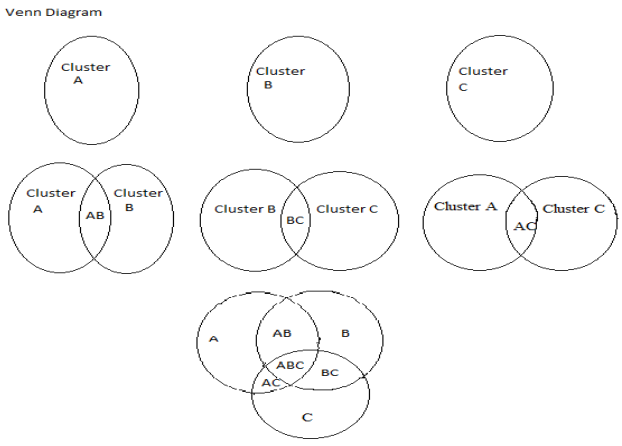

Figure 1.1(a) for single cluster, 1.1(b) for two cluster, 1.1(c) for more than two cluster.

ISSN:2319-7900

Case 1: Suppose if there are queries that are related to single cluster then this type of clusters are shown in fig.1.1 (a).In this fig1.1(a) there are three independent clusters named as Cluster A, Cluster B and Cluster C. Queries if raised on any of three cluster (cluster A, cluster B and cluster C) then it can be solved easily or it is easy to answer all the queries by independent clusters.

Case 2: Suppose if there are some queries that are related to cluster A and cluster B or related to cluster B and cluster C or related to cluster A and cluster C. Now, if these joint queries are tried to solve by independent clusters as shown in fig 1.1(b) then it will take more time and more memory space to store the fetched data from independent clusters. Overhead also increased as in this case the fetched data have to b saved again and again till the query finish.

So, we tried to overcome from these problems and made a joint cluster of two independent clusters on the basis of raised queries on respective two clusters. Suppose if queries are on cluster A and cluster B then we joined these two cluster together and a new cluster discovered named as cluster AB shown in fig 1.1(b).Same if queries are on cluster B and cluster C or on cluster A and cluster C then same we can join the respective clusters to make a new joined cluster named as cluster BC or cluster AC. The benefit of this joint cluster is that it reduces the access time as the data is now fetched from a single joint cluster and also it takes less memory space as there is no need to save data again and again which we have to do if we are solving queries by independent clusters, Overhead also minimized by this joint cluster as it requires only one time access from the joint cluster.

Case 3: Suppose if there are some queries that are raised on all of three clusters means on cluster A, cluster B and cluster C. Then if we solved raised queries by case 1 then the problems of access time, memory space and overhead remains same or even more as data increases in queries. To fetch data from independent clusters, access time increases and to save this fetched data more memory space needed. Therefore first case is not a good method to answer the raised queries on more than one cluster. Now, if we solved raised queries by case 2 then access time reduces as there is joint cluster of two independent clusters and also less memory space is needed as compare to case 1 because here we only three times save the data as an example suppose queries are on cluster A, cluster B and cluster C then in case 2 we have cluster AB, cluster BC and cluster AC so queries on cluster A and cluster B are solved by single joint cluster AB, same queries on cluster B and cluster C are solved by cluster BC, same queries on cluster A and cluster C are solved by cluster AC. So only three times we save the data to get final answer of the raised queries. But we want to overcome from all problems of three times saving data and different-different accesses from joint clusters.

Therefore, we joined all three clusters together in order to make a single joint cluster named as cluster ABC. All the queries raised on more than two clusters can be solved easily and efficiently by this single cluster. This joint cluster reduces the access time even more better as compare to case 1 and case 2 as only one time access is done in this case, It also decreases the memory space to save the data as in this case only one time saving of data is required. So this single joint cluster of all three clusters reduces all the problems occurs in using case 1 and case 2 for multiple clusters dependent queries. In the other hand, the fig 1.1(c) shows the combination of different- different clusters. All possible clusters are available in one single cluster; we can easily solve queries raised on independent cluster or also queries raised on more than one cluster and final cluster is very useful in case of queries raised on more than two clusters.

Single-Value Attributes based Clustering

This type of clustering is based on single value attribute. Clustering is done on the basis of single value attribute of dataset. All the Queries that are related to single value attribute can be solved easily by this clustering .Based on the Queries of datasets, the data are fetched from the clustered data and desired output in find out. Single-value attribute based clustering is useful for the queries that are based on single value attribute of dataset. But if there are some queries that are based on multiple attributes of datasets [8] then single value attribute based clustering doesn't play a good role to satisfy the queries related to multiple attributes of datasets.

Algorithm for SAC (Single-Value Attributes based Clustering) Approach

An Algorithm for SAC Approach is presented below.

Step 1: Input Record or Table (T=1).* Enter the table as an input*\

Step 2: Input Queries (Q=1 to k).* Enter the queries can be 1 or more*\

Step 3: Create Attribute Usage Matrix.* Matrix is maintained on the basis of attributes*\

Step 3.a. Put '1' in attribute usage matrix if attribute used.

Step 3.b. Put '0' in attribute usage matrix if attribute doesn't use.

Step 4: Find out Result.

Step 5: End.

An illustration of SAC Algorithm step by step is presented here:

In Step 1, Tables are used as a input for this approach. Here one table is used as a input that's why the value of t is 1.

ISSN:2319-7900

In Step 2, Queries are entered. These queries can be ranged between 1 to k means k no. of queries can be used or entered as a input.
In Step 3, A Attribute Usage Matrix is created on the basis of attributes used in the respective queries of step 2.Step 3.a is related to the usage of attribute. If attribute used then '1' is entered in the attribute usage matrix. Otherwise '0' is entered if attribute is not used shows in step 3.b.
In Step 4, Final result is carried out means how much attributes are used by using SAC Approach.
In Step 5, SAC Algorithm ends.
Based on the above SAC Algorithm, An Example is solved below:
Here, with the help of an example we show the clustering based on single value attribute of dataset. In this example let there is a company A that have three main components or attributes named as COM-A'S COMPUTERS represented by A1,COM-A'S LAPTOPS represented by A2 and COM-A'S PRINTERS represented by A3.Now these A1,A2 and A3 have their own attributes that are shown by three tables respectively.
First table is regarding with COM-A'S COMPUTERS (A1) which have three attributes or components with their range named as COM-A'S CRT COMPUTERS represented by (A1') which are of range 10,000/-; COM-A'S LCD COMPUTERS represented by (A1") which are of range 15,000/-; COM-A'S LED COMPUTERS represented by (A1'") which belongs to range 20,000/-.Now second table is regarding with COM-A'S LAPTOPS (A2) which have three attributes or components with their range named as COM-A'S HD LAPTOPS represented by (A2') which are of range 40,000/-; COM-A'S INTRA DISPLAY LAPTOPS represented by (A2") which are of range 30,000/-; COM-A'S INSPIRON LAPTOPS represented by (A2'") which belongs to range 35,000/-.Third table is regarding with COM-A'S PRINTERS (A3) which have three attributes or components with their range named as COM-A'S DOT MATRIX PRINTER represented by (A3') which are of range 10,000/-; COM-A'S LASER PRINTER represented by (A3") which are of range 24,000/-; COM-A'S SIMPLE PRINTER represented by (A3'") which belongs to range 16,000/-.

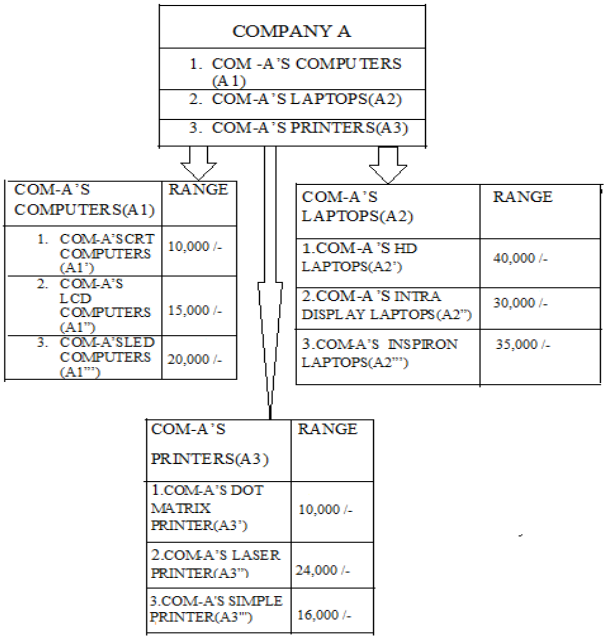

Figure1.2 Service provider and their services

Now, Some Queries that are raised on the above table are shown below. These queries are as follows:
Q.1. SELECT A2 FROM A AND A2" FROM A2 WHERE RANGE=30,000.
Q.2. SELECT A1 FROM A AND A1'" FROM A1 WHERE RANGE=20,000.
Q.3. SELECT A2 FROM A AND A2' FROM A2 WHERE RANGE=40,000.
Q.4. SELECT A3 FROM A AND A3'" FROM A3 WHERE RANGE=16,000.
In the Query Q.1. Attribute A2 is selected from COMPANY A Table and then attribute A2" of table COM-A'S LAPTOPS is selected from the table. So only one attribute A2" is used in this query.
In the Query Q.2. Attribute A1 is selected from COMPANY A Table and then attribute A1'" of table COM-A'S COMPUTERS is selected from the table. So only one attribute A1'" is used in this query.
In the Query Q.3. Attribute A2 is selected from COMPANY A Table and then attribute A2' of table COM-A'S LAPTOPS is selected from the table. So only one attribute A2' is used in this query.
In the Query Q.4. Attribute A3 is selected from COMPANY A Table and then attribute A3'" of table COM-A'S PRINTERS is selected from the table. So only one attribute A3'" is used in this query.
On the basis of above Queries we calculated a Attribute usage matrix, In this matrix, The attributes that are used in queries are shown by 1 and the attributes that are not used in queries are shown by 0.Attribute usage matrix by using the attributes of above queries is shown below:

ISSN:2319-7900

Table(i) Attribute Usage Matrix

	A1	A2	A3
Q1	0	1	0
Q2	1	0	0
Q3	0	1	0
Q4	0	0	1

Result For SAC Approach:

In this Attribute Usage Matrix, Out of three attributes only one attribute is used that is shown by 1, other attributes are unused that are shown by 0.These unused attributes simply takes memory to save. So here is a wastage of available memory space in SAC (Single value attribute based Clustering).
The above queries are related to one type of service provider or one type of cluster. These queries can be easily carried out by using the cluster data of one type.
But if there are suppose more than one service provider or more than one data sets are available then, Questions Arises,

- Is Single value attribute based clustering solved the queries effectively and efficiently in minimum time that are based on multiple attributes of dataset?
- Is the Single value attribute based clustering will take less storage space in order to save multiple attributes related query data ?
- Is the Single value attribute based clustering is good for the multiple attribute based queries in order to minimize the overhead to provide solution for the query?

The Answer is NO,

- Because single value attribute based clustering (SAC) will take more time in order to find out the desired data or desired output, it is because this approach fetch the data from multiple clusters instead of one cluster so fetching will take more time in case of multiple attribute of dataset.
- Overhead increases in this type of clustering when applied for multiple attribute attribute based queries because data are fetched and collected through different different clusters in order to save the fetched data, more storage space is needed that causes overhead.

Therefore, in order to solve all these problems with single value attribute based clustering in case of multiple attributes; we proposed a new Approach that is MULTIPLE ATTRIBUTE BASED CLUSTERING (MAC) APPROACH that resolved all the problems of SAC Approach and plays an efficient and effective role in order to provide result for multiple attribute based Queries.

Multiple Attribute based Clustering

Multiple attributes based clustering (MAC) is an approach that is based on multiple attributes of dataset. It clusters the data on the basis of need. The queries that focuses more on the datasets (more than one datasets), according to the dependency on more than one attributes of dataset, clustering is performed. The attributes that have dependency of more than one query are clustered together in order to satisfy the query in minimum time and with less overhead. The cluster of more than one dataset gives the more efficient result in comparison to single dataset cluster.
The result of multiple attributes based query is easy to find-out with the multiple attributes based clustering than to single value attribute based clustering because in MAC Approach the desired data are fetched from the single cluster of more than one dataset but this is not possible in the case of SAC approach.
In SAC Approach, Different tables of datasets[10] are maintained that takes the memory space.
In order to find-out the desired data-
-If query is related to single cluster, then it is easy and time saving approach for result.
-But if query is related to more than one cluster of dataset then it doesn't play an effective role of clustering over there. Because, for every needed attribute of query, a different table is used to fetch the desired data, for another attribute of query, another table is used to fetch data.
If this process continues to find out the result of query (based on multiple attribute) then it will take more memory space that causes overhead and also it will take more time as compare to MAC approach.

Therefore, Our proposed MAC Approach is a better solution to all these problems.
In MAC Approach, the tables are maintained on the basis of need of clustering. The datasets that have dependency more are clustered together and by clustering a new single cluster is formed that contains both the dependent attributes of query. As compare to SAC approach, now a single new cluster table is maintained in the system that reduces the memory space because a single table plays role of both different tables.
This single new table is used for query and in the less time with less overhead, the result of queries (based on multiple attribute) are carried out. This MAC approach is useful for 2,3,4,5,6-------n no. of datasets. The n no of datasets can be clustered together on the basis of need.

ISSN:2319-7900

Proposed Algorithm for MAC (Multiple Attributes based Clustering) Approach-

A proposed algorithm for our new approach Multiple Attributes based Clustering (MAC) is presented below:
Step 1: Input: Record or Table (T=1 to n).*Tables are entered*\
Step 2: Input: Queries (Q=1 to k).*k no. of queries is entered*\
Step 3: Join Tables on the basis of respective Queries of step 2.
Step 4: Final joint table or cluster of independent table or cluster is carried out.
Step 5: Create Attribute Usage Matrix for final joint Tables or clusters (from step 4).
Step 5.a. Put '1' in attribute usage matrix if attribute used.
Step 5.b. Put '0' in attribute usage matrix if attribute doesn't use.
Step 6: Compare Attribute Usage Matrix of SAC and MAC Algo.
Step 7: Find out final result.
Step 8: End.

An illustration of SAC Algorithm step by step is presented here:
In Step 1, n no. of tables is entered as an input. Value of T can be 1 to n.
In Step 2, k no. of queries is entered means queries can be of range 1 to k.
In Step 3, Tables are joined on the basis of queries entered in second step.
In Step 4, A Single joint table of independent tables entered in first step is created. Final joint table is carried out in this step.
In Step 5, An Attribute Usage Matrix is created for the final joint table developed in step fourth. Step 5.a is related to the usage of attribute. If attribute is used then show by '1' in the attribute usage matrix otherwise '0' is used to show the unused attribute in attribute usage matrix as step 5.b says.
In Step 6, both attribute usage matrix of SAC approach and MAC approach is compared with each other.
In Step 7, Final result is carried out on the basis of comparison done in sixth step means how much attribute usage increases in MAC approach? Or which one approach (SAC or MAC) is better to utilize the attribute maximum? The answers of these questions are find out in this step.
In Step 8, the algorithm for MAC approach gets ends here.
The working of the MAC Algorithm is presented next by a Flow Chart :

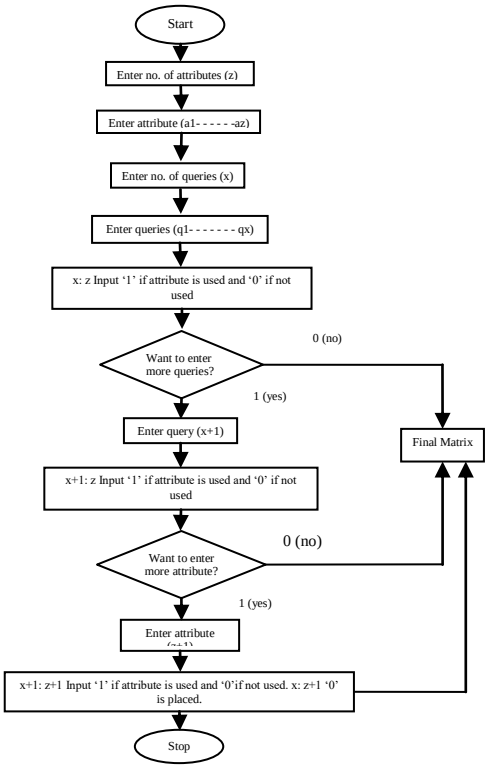

Figure1.3 Flow Chart for MAC methodology

On the basis of above proposed MAC algorithm the following example is solved. This example is solved by using MAC algorithm. The following example is as follows:
Let there are n types of service providers(SP)-Here we are taking the value of n=4 means n=1,2,3,4.So service providers are represented by SP1,SP2,SP3 and SP4.SP1 provides the Hotel services,SP2 provides the food services,SP3 provides the Vehicle services and SP4 provides the Tours services.SP1 Hotel table is shown by table A and it contains 4 type of hotels that are named as A1,A2,A3 and A4,these A1,A2,A3,A4 have their own tables and attributes are represented such as A1',A1",A1"' of table A1,A2',A2",A2"' of table A2,A3',A3",A3"' of table A3 and A4',A4",A4"' of table A4.Same like table A, table B have 4 types of food services that are named as B1,B2,B3 and B4,these B1,B2,B3,B4 have their own tables and attributes are represented such as B1',B1",B1"' of table B1,B2',B2",B2"' of table B2,B3',B3",B3"' of table B3 and B4',B4",B4"' of table B4. Same like table B, table C have 4 types of Vehicle services that are named as C1,C2,C3 and C4,these C1,C2,C3,C4 have their own tables and attributes are represented such as C1',C1",C1"' of table C1,C2',C2",C2"' of table C2,C3',C3",C3"' of table C3 and C4',C4",C4"' of table C4. Same like table C, table D have 4 types of Tours services that are named as D1,D2,D3 and D4,these D1,D2,D3,D4 have their own tables and attributes are represented such as D1',D1",D1"' of table D1,D2',D2",D2"' of

table D2,D3',D3",D3"' of table D3 and D4',D4",D4"' of table D4.

Suppose First table (A) is regarding with HOTELS that have four different hotels means attributes that are as follows: Himalayan Regency represented by attribute A1,Hill View Hotel represented by attribute A2,Start Point Hotel represented by attribute A3 and Mountain Hill Hotel represented by attribute A4.Now these attributes A1,A2,A3 and A4 have their own tables respectively. First table HIMALAYAN REGENCY (A1) have three attributes named as Delux Room represented by A1' belongs to range 800 /-, Super Delux Room represented by A1" belongs to range 1000 /- and Normal Room represented by A1"' belongs to range 600 /-.Now, Second table Hill View Hotel (A2) have three attributes named as Delux Room represented by A2' belongs to range 700 /-, Super Delux Room represented by A2" belongs to range 900 /- and Normal Room represented by A2"' belongs to range 400 /-.Third table Start Point Hotel (A3) have three attributes named as Delux Room represented by A3' belongs to range 600 /-, Super Delux Room represented by A3" belongs to range 850 /- and Normal Room represented by A3"' belongs to range 350 /- and last table Mountain Hill Hotel (A4) have three attributes named as Delux Room represented by A4' belongs to range 800 /-, Super Delux Room represented by A4" belongs to range 1200 /- and Normal Room represented by A4"' belongs to range 700 /- .First table is shown below:

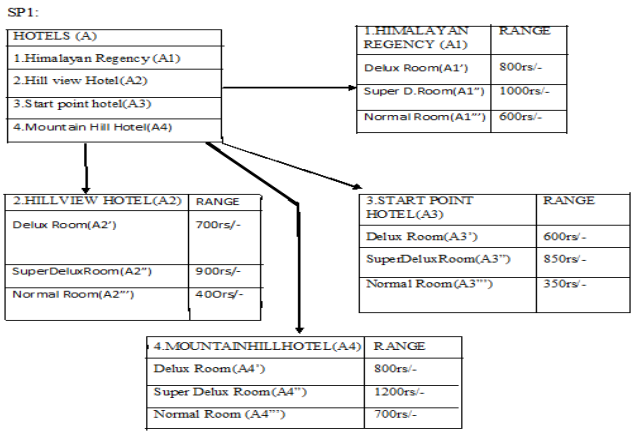

Figure1.4 SP1 and thier services

Another Second table (B) is suppose regarding with MEAL that have four different Meals means attributes that are as follows: Chinese Meal represented by attribute B1,Continental Meal represented by attribute B2,Indian Meal represented by attribute B3 and Punjabi Meal represented by attribute B4.Now these attributes B1,B2,B3 and B4 have their own tables respectively.First table Chinese Meal (B1) have three types of food or can say attributes named as Chowmin represented by B1' belongs to range 150 /-, Momoz represented by B1" belongs to range 40 /- and Hakka Noodles represented by B1"' belongs to range 250 /- .Now, Second table Continental Meal (B2) have three attributes named as Snacks represented by B2' belongs to range 70 /-, Spring Roll represented by B2" belongs to range 80 /- and Burger represented by B2"' belongs to range 60 /-.Third table Indian Meal (B3) have three attributes named as Indian Thali represented by B3' belongs to range 300 /-, Paneer Pasanda represented by B3" belongs to range 200 /- and Malai Kofta represented by B3"' belongs to range 150 /- and last table Punjabi Meal (B4) have three attributes named as Punjabi Thali represented by B4' belongs to range 350 /-, Saag and Roti represented by B4" belongs to range 450 /- and Paneer Punjabi represented by B4"' belongs to range 250 /-.Second table is shown below:

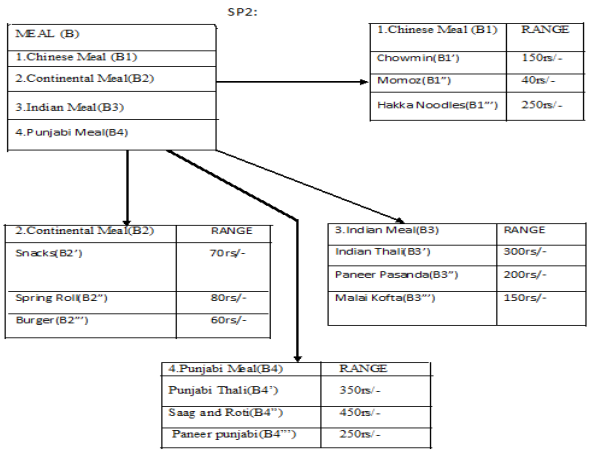

Figure1.5 SP2 and thier services

Now, Suppose Third table (C) is regarding with Vehicle Service that have four different Vehicle Services means attributes that are as follows: Volvo Bus Service represented by attribute C1, Himalaya Bus Service represented by attribute C2, Toranto Bus Service represented by attribute C3 and Hill View Bus Service represented by attribute C4.Now these attributes C1,C2,C3 and C4 have their own tables respectively. First table Volvo Bus Service (C1) have three types of Buses or can say attributes named as Normal Bus represented by C1' belongs to range 150 /-, AC Bus represented by C1" belongs to range 200 /- and Delux Bus represented by C1"' belongs to range 250 /-.Now, Second table Himalaya Bus Service (C2) have three attributes named as Normal Bus represented by C2' belongs to range 250 /-, AC Bus represented by C2" belongs to range 560 /- and Delux Bus represented by C2"' belongs to range 780 /-. Third table Toranto Bus Service (C3) have three attributes named as Normal Bus represented by C3' belongs to range 200 /-, AC Bus represented by C3" belongs to range 300 /- and Delux Bus represented by C3"' belongs to range 450 /-. and last table Hill View Bus Service (C4) have three attributes

named as Normal Bus represented by C4' belongs to range 350 /-, AC Bus represented by C4" belongs to range 450 /- and Delux Bus represented by C4"' belongs to range 250 /-. The respective third table is shown below:

SP3:

Vehicle Service(C)
1.Volvo Bus Service(C1)
2.Himalaya Bus Service(C2)
3.Toranto Bus Service(C3)
4.Hill View Bus Service(C4)

1.Volvo Bus Service (C1)	RANGE
Normal Bus(C1')	150rs/-
AC Bus(C1")	200rs/-
Delux Bus(C1"')	250rs/-

2.HimalayaBusService(C2)	RANGE
Normal Bus(C2')	250rs/-
AC Bus(C2")	560rs/-
Delux Bus(C2"')	780rs/-

3.TorantoBusService (C3)	RANGE
Normal Bus(C3')	200rs/-
AC Bus(C3")	300rs/-
Delux Bus(C3"')	450rs/-

4.HillViewBusService(C4)	RANGE
Normal Bus(C4')	350rs/-
AC Bus(C4")	450rs/-
Delux Bus(C4"')	250rs/-

Figure1.6 SP3 and thier services

Last table (D) of this example is suppose concerned with TOURS that have four different types of tours means attributes that are as follows: Hill View Tour represented by attribute D1, Rafting Tour represented by attribute D2, Boating Tour represented by attribute D3 and Sketting represented by attribute D4.Now these Tours or can say attributes D1,D2,D3 and D4 have their own tables respectively. First table Hill View Tour (D1) have three types of tours or attributes named as Top Point represented by D1' belongs to range 150 /-, Himalaya View represented by D1" belongs to range 300 /- and Lake View represented by D1"' belongs to range 250 /-.Now, Second table Rafting Tour (D2) have three attributes named as 15 min Rafting represented by D2' belongs to range 250 /-, 30 min Rafting represented by D2" belongs to range 400 /- and 1 hour Rafting represented by D2"' belongs to range 800 /-.Third table Boating Tour (D3) have three attributes named as Simple Boat represented by D3' belongs to range 250 /-, Double Sitter Boat represented by D3" belongs to range 350 /- and Peddle Boat represented by D3"' belongs to range 400 /- and last table Sketting Tour (D4) have three attributes named as 1 km sketting represented by D4' belongs to range 250 /-, 5 km sketting represented by D4" belongs to range 350 /- and 8 km sketting represented by D4"' belongs to range 550 /-.Last or fourth table is shown below:

SP4:

TOURS (D)
1. Hill View Tour(D1)
2.Rafting Tour(D2)
3.Boating Tour(D3)
4.Sketting Tour(D4)

1. Hill View Tour(D1)	RANGE
Top point(D1')	150rs/-
Himalaya View(D1")	300rs/-
Lake View(D1"')	250rs/-

2.Rafting Tour(D2)	RANGE
15 min Rafting(D2')	250rs/-
30 min Rafting(D2")	400rs/-
1 hour Rafting(D2"')	800rs/-

3.Boating Tour(D3)	RANGE
Simple boat(D3')	250rs/-
Double sitter boat(D3")	350rs/-
Peddle boat(D3"')	400rs/-

4.Sketting Tour(D4)	RANGE
1 km sketting (D4')	250rs/-
5 km sketting (D4")	350rs/-
8 km sketting (D4"')	550rs/-

Figure1.7 SP4 and thier services

Let there are some queries that are raised on the above four tables. These queries are as follows:

Q.1. Select A4 from A where A4"=1200rs/- and B2 from B where B2"'= 60rs/- and D4 from D where D4'=250rs/-.

Q.2. Select A3 from A where A3'=600rs/- and C2 from C where C2"= 560rs/- and D3 from D where D3"'=400rs/-.

Q.3. Select B4 from B where B4'=350rs/- and D2 from D where D2"=400rs/-and A1 from A where A1"'=600rs/-.

Q.4. Select C3 from C where C3'=200rs/- and A2 from A where A2"=900rs/- and D4 from D where D4"'=550rs/-.

Q.5. Select B1 from B where B1'=150rs/- and D2 from D where D2"= 400rs/-and A3 from A where A3"'=350rs/-.

Q.6. Select D3 from D where D3"=350rs/- and A1 from A where A1'=800rs/- and B2 from B where B2"=80rs/-.

Q.7. Select A4 from A where A4'=800rs/- and D3 from D where D3"=350rs/- and C2 from C where C2"'=780rs/-.

Q.8. Select C4 from C where C4"=400rs/- and A2 from A where A2'=700rs/- and D1 from D where D1"'=250rs/-.

In the Query.1. A4 is selected from table A and A4" is selected from table A4 where A4" belongs to range 1200 /- and B2 is selected from table B and B2"' is selected from table B2 where B2"' belongs to the range 60 /- and D4 is selected from table D, D4' is selected from table D4 where D4' belongs to the range 250 /-. So in this query three attributes A4", B2"' and D4' is used.

In the Query.2. A3 is selected from table A and A3' is selected from table A3 where A3' belongs to range 600 /- and C2 is selected from table C and C2" is selected from table C2 where C2" belongs to the range 560 /- and D3 is selected from table D, D3"' is selected from table D3 where D3"' belongs to the range 400 /-. So in this query three attributes A3', C2" and D3"' is used.

In the Query.3. B4 is selected from table B and B4' is selected from table B4 where B4' belongs to range 350 /- and D2 is selected from table D and D2" is selected from table D2 where D2" belongs to the range 400 /- and A1 is selected

ISSN:2319-7900

from table A, A1’” is selected from table A1 where A1’” belongs to the range 600 /-. So in this query three attributes B4’, D2” and A1’” is used.

In the Query.4. C3 is selected from table C and C3’ is selected from table C3 where C3’ belongs to range 200 /- and A2 is selected from table A and A2” is selected from table A2 where A2” belongs to the range 900 /- and D4 is selected from table D, D4”’ is selected from table D4 where D4”’ belongs to the range 550 /-. So in this query three attributes C3’, A2” and D4”’ is used.

In the Query.5. B1 is selected from table B and B1’ is selected from table B1 where B1’ belongs to range 150 /- and D2 is selected from table D and D2” is selected from table D2 where D2” belongs to the range 400 /- and A3 is selected from table A, A3’” is selected from table A3 where A3’” belongs to the range 350 /-. So in this query three attributes B1’, D2” and A3’” is used.

In the Query.6. D3 is selected from table D and D3” is selected from table D3 where D3” belongs to range 350 /- and A1 is selected from table A and A1’ is selected from table A1 where A1’ belongs to the range 800 /- and B2 is selected from table B, B2” is selected from table B2 where B2” belongs to the range 80 /-. So in this query three attributes D3”, A1’ and B2” is used.

In the Query.7. A4 is selected from table A and A4’ is selected from table A4 where A4’ belongs to range 800 /- and D3 is selected from table D and D3” is selected from table D3 where D3” belongs to the range 350 /- and C2 is selected from table C, C2’” is selected from table C2 where C2’” belongs to the range 780 /-. So in this query three attributes A4’, D3” and C2’” is used.

In the Query.8. A4 is selected from table A and A4’ is selected from table A4 where A4’ belongs to range 800 /- and D3 is selected from table D and D3” is selected from table D3 where D3” belongs to the range 350 /- and C2 is selected from table C, C2’” is selected from table C2 where C2’” belongs to the range 780 /-. So in this query three attributes A4’, D3” and C2’” is used.

On the basis of dependency upon more than one attribute of datasets, we form cluster of dependent attribute of datasets .The dependent attributes are clustered together in order to provide the fast access of data from a single cluster of multiple dependent attribute. After reviewing all queries, on the basis of need, we make two clusters that are useful n significant in all aspects such as time, memory overhead. The two clusters are ACD and ABD. The ACD cluster is now represented by D cluster and ABD cluster is now represented by E cluster and attributes are now represented by D1, D2,D3 respectively of D cluster and E1,E2,E3 respectively of E cluster.

CLUSTER D:

	D1		D2		D3
1.A1	A1’=800rs/- A1”=1000rs/- A1”’=600rs/-	C1	C1’=150rs/- C1”=200rs/- C1”’=250rs/-	D1	D1’=150rs/- D1”=300rs/- D1”’=250rs/-
2.A2	A2’=700rs/- A2”=900rs/- A2”’=400rs/-	C2	C2’=250rs/- C2”=560rs/- C2”’=780rs/-	D2	D2’=250rs/- D2”=400rs/- D2”’=800rs/-
3.A3	A3’=600rs/- A3”=850rs/- A3”’=350rs/-	C3	C3’=200rs/ C3”=300rs/- C3”’=450rs/-	D3	D3’=250rs/- D3”=350rs/- D3”’=400rs/-
4.A4	A4’=800rs/- A4”=1200rs/- A4”’=700rs/-	C4	C4’=350rs/- C4”=450rs/- C4”’=250rs/-	D4	D4’=250rs/- D4”=350rs/- D4”’=550rs/-

Figure 1.8 New generated cluster using MAC Approach

In this table, A1 represented Himalayan Regency where,
A1’=Delux-Room,A1”=Super-D.Room and
A1’”=NormalRoom
A2 represented Hill View Hotel where,
A2’=Delux-Room,A2”=SuperDeluxRoom and
A2’”=Normal Room
A3 represented Start Point Hotel where,
A3’=Delux-Room,A3”=SuperDeluxRoom and
A3’”=NormalRoom
A4 represented Mountain Hill Hotel where,
A4’=Delux Room,A4”=Super Delux Room and
A4’”=Normal Room
C1 represented Volvo Bus Service where,
C1’=Normal Bus,C1”=AC Bus and C1’”=Delux Bus
C2 represented Himalaya Bus Service where,
C2’=Normal Bus,C2”=AC Bus and C2”’=Delux Bus
C3 represented Toranto Bus Service where,
C3’= Normal Bus,C3”=AC Bus and C3’”=Delux Bus
C4 represented Hill View Bus Service where,
C4’= Normal Bus,C4”= AC Bus and C4’”=Delux Bus
D1 represented Hill View Tour where,
D1’= Top point,D1”=Himalaya View and D1’”=Lake View
D2 represented Rafting Tour where,
D2’=15 min Rafting,D2”=30 min Rafting,D2’”=1 hour Rafting
D3 represented Boating Tour where,
D3’=Simple boat,D3”=Double sitter boat and D3’”=Peddle boat
D4 represented Sketting Tour where,
D4’=1 km sketting,D4”=5 km sketting and D4’”=8 km sketting.

CLUSTER E:

	E1		E2		E3
1.A1	A1'=800rs/- A1"=1000rs/- A1"'=600rs/-	B1	B1'=150rs/- B1"=40rs/- B1"'=250rs/-	D1	D1'=150rs/- D1"=300rs/- D1"'=250rs/-
2.A2	A2'=700rs/- A2"=900rs/- A2"'=400rs/-	B2	B2'=70rs/- B2"=80rs/- B2"'=60rs/-	D2	D2'=250rs/- D2"=400rs/- D2"'=800rs/-
3.A3	A3'=600rs/- A3"=850rs/- A3"'=350rs/-	B3	B3'=300rs/- B3"=200rs/- B3"'=150rs/-	D3	D3'=250rs/- D3"=350rs/- D3"'=400rs/-
4.A4	A4'=800rs/- A4"=1200rs/- A4"'=700rs/-	B4	B4'=350rs/- B4"=450rs/- B4"'=250rs/-	D4	D4'=250rs/- D4"=350rs/- D4"'=550rs/-

Figure1.9 New generated cluster using MAC Approach

In this table,A1 represented Himalayan Regency where,
A1'=Delux-Room,A1"=SuperD.Roomand
A1"'=NormalRoom
A2 represented Hill View Hotel where,
A2'=Delux Room,A2"=SuperDeluxRoom and A2"'=Normal Room
A3 represented Start Point Hotel where,
A3'=Delux-Room,A3"=SuperDeluxRoom and A3"'=NormalRoom
A4 represented Mountain Hill Hotel where,
A4'=Delux Room,A4"=Super Delux Room and A4"'=Normal Room
B1 represented Chinese Meal where,
B1'=Chowmin,B1"=Momoz and B1"'=Hakka Noodles
B2 represented Continental Meal where,
B2'=Snacks,B2"=Spring Roll and B2"'=Burger
B3 represented Indian Meal where,
B3'= Indian Thali, B3"=Paneer Pasanda,B3"'= Malai Kofta
B4 represented Punjabi Meal where,
B4'=Punjabi Thali, B4"=Saag and Roti and B4"'=Paneer Punjabi
D1 represented Hill View Tour where,
D1'= Top point, D1"=Himalaya View and D1"'=Lake View
D2 represented Rafting Tour where,
D2'=15 min Rafting, D2"=30 min Rafting,D2"'=1 hour Rafting
D3 represented Boating Tour where,
D3'=Simple boat, D3"=Double sitter boat and D3"'=Peddle boat
D4 represented Sketting Tour where,
D4'=1 km sketting, D4"=5 km sketting and D4"'=8 km sketting.

Table(ii) Attributes usage matrix for cluster D:

	D1	D2	D3
Q1	1	1	1
Q2	1	1	1
Q3	1	1	1
Q4	1	1	1

Table(iii) Attribute Usage Matrix of Cluster E:

	E1	E2	E3
Q1	1	1	1
Q2	1	1	1
Q3	1	1	1
Q4	1	1	1

In the both table for cluster D and cluster E, all the attributes are used efficiently. There is no wastage of memory space in this approach as compare to previous approach (SAC).

Result & Analysis:

After comparison of table (i) from SAC approach and table (iii) from MAC approach; it is clear that attribute usage increases in table (iii) that is produced by using our new idea new approach that is MULTIPLE ATTRIBUTE BASED CLUSTERING.
As in SAC approach, we have the table based on single value attribute, if we want to solve query based on multiple attributes then also we have to fetch data from different-different table of datasets that takes more steps and also time in order to provide the solution to the query (related to multiple attributes of dataset).Thus, the usage of attributes in table by using SAC approach is not so efficient as some attributes remains unused. Due to the unused attributes, the memory space is also wasted.

ISSN:2319-7900

Therefore, to overcome from these problems of unused attributes in table and wastage of memory space, we used our proposed approach, MULTIPLE ATTRIBUTES BASED APPROACH, that increases the usage of attributes by maintaining one table for the dependent multiple attributes.
All the attributes that are needed in order to provide solution or on the basis of need are maintained in one table that reduce the access time and also increases the use of memory space.

Experimental Result:

On the basis of queries and related attributes usage, the result is calculated. Different-different no. of queries and attributes usage are taken into consideration in order to calculate the performance of SAC (Single-value Attribute based Clustering) and MAC(Multiple Attribute based Clustering) Approach-
To calculate the performance of Clustering on available attributes by using SAC Approach, different number of queries and attributes are taken...these are as follows-
Suppose,
If Query=4 and Attribute=3 then Attribute Usage=4/12=0.333
If Query=8 and Attribute=3 then Attribute Usage=8/24=0.333
If Query=10 and Attribute=4 then Attribute Usage=10/40=0.25
If Query=20 and Attribute=4 then Attribute Usage=20/80=0.25
If Query=30 and Attribute=5 then Attribute Usage=30/150=0.2
If Query=40 and Attribute=5 then Attribute Usage=50/200=0.25
If Query=50 and Attribute=4 then Attribute Usage=50/200=0.25
If Query=100 and Attribute=4 then Attribute Usage=100/400=0.25
If Query=200 and Attribute=4 then Attribute Usage=200/800=0.25
If Query=300 and Attribute=4 then Attribute Usage=300/1200=0.25
If Query=400 and Attribute=4 then Attribute Usage=400/1600=0.25
If Query=500 and Attribute=5 then Attribute Usage=500/2500=0.2
If Query=600 and Attribute=5 then Attribute Usage=600/3000=0.2
If Query=700 and Attribute=4 then Attribute Usage=700/2800=0.25
If Query=800 and Attribute=4 then Attribute Usage=800/3200=0.25
If Query=900 and Attribute=5 then Attribute Usage=900/4500=0.2
If Query=1000 and Attribute=5 then Attribute Usage=1000/5000=0.2
If Query=1500 and Attribute=6 then Attribute Usage=1500/9000=0.167
If Query=2000 and Attribute=7 then Attribute Usage=2000/14000=0.1428
If Query=1500 and Attribute=6 then Attribute Usage=3000/21000=0.1428
To calculate the performance of Clustering on available attributes of datasets by using MAC Approach, different number of queries and attributes are taken...these are as follows-
Suppose,
If Query=4 and Attribute=3 then Attribute Usage=12/12=1
If Query=8 and Attribute=3 then Attribute Usage=24/24=1
If Query=10 and Attribute=4 and Suppose,
Case: 5 Query used 3 attributes and 5 Query utilizes 4 attribute then,
Attribute Usage=35/40=0.875
If Query=20 and Attribute=4 and Suppose,
Case: 10 Query uses 3 attributes and 10 Query utilizes 4 attribute then,
Attribute Usage=70/80=0.875
If Query=30 and Attribute=5 and Suppose,
Case: 20 Query uses 4 attributes and 10 Query utilizes 5 attribute then,
Attribute Usage=130/150=0.867
If Query=40 and Attribute=5 and Suppose,
Case: 20 Query uses 3 attributes, 10 Query utilizes 4 attribute and 10 Query uses 5 attribute then,
Attribute Usage=150/200=0.75
If Query=50 and Attribute=4 and Suppose,
Case: 20 Query uses 3 attribute and 30 Query uses 4 attribute then,
Attribute Usage=180/200=0.9
If Query=100 and Attribute=4 and Suppose,
Case: 30 Query uses 3 attribute and 70 Query uses 4 attribute then,
Attribute Usage=370/400=0.925
If Query=200 and Attribute=4 and Suppose,
Case: 100 Query uses 3 attribute and 100 Query uses 4 attribute then,
Attribute Usage=700/800=0.875
If Query=300 and Attribute=4 and Suppose,
Case: 100 Query uses 3 attribute and 200 Query uses 4 attribute then,
Attribute Usage=1100/1200=0.9167
If Query=400 and Attribute=4 and Suppose,
Case: 200 Query uses 3 attribute and 200 Query uses 4 attribute then,
Attribute Usage=1400/1600=0.875
If Query=500 and Attribute=5 and Suppose,
Case: 250 Query uses 5 attribute, 100 Query uses 3 attribute and 150 Query uses 4 attribute then,
Attribute Usage=2150/2500=0.86

ISSN:2319-7900

If Query=600 and Attribute=5 and Suppose,
Case: 150 Query uses 3 attribute, 200 Query uses 4 attribute and 250 Query uses 5 attribute then,
Attribute Usage=2500/3000=0.833
If Query=700 and Attribute=4 and Suppose,
Case: 300 Query uses 3 attribute and 400 Query uses 4 attribute then,
Attribute Usage=2500/2800=0.8928
If Query=800 and Attribute=4 and Suppose,
Case: 400 Query uses 3 attribute and 400 Query uses 4 attribute then,
Attribute Usage=2800/3200=0.875
If Query=900 and Attribute=5 and Suppose,
Case: 400 Query uses 4 attribute and 500 Query uses 5 attribute then,
Attribute Usage=4100/4500=0.9111
If Query=1000 and Attribute=5 and Suppose,
Case: 500 Query uses 4 attribute and 500 Query uses 5 attribute then,
Attribute Usage=4500/5000=0.9
If Query=1500 and Attribute=6 and Suppose,
Case: 1000 Query uses 6 attribute and 500 Query uses 5 attribute then,
Attribute Usage=8500/9000=0.944
If Query=2000 and Attribute=7 and Suppose,
Case: 1500 Query uses 7 attribute and 500 Query uses 6 attribute then,
Attribute Usage=13500/14000=0.96428
If Query=3000 and Attribute=7 and Suppose,
Case: 2000 Query uses 7 attribute and 1000 Query uses 6 attribute then,
Attribute Usage=20000/21000=0.9523

A Table is maintained for the above calculated data by using SAC and MAC Approach..The following table shows the above calculated data-

Table(iv)Calculated data by SAC and MAC Approach

Data utilization in SAC Approach	Data utilization in MAC Approach
0.333	1
0.33	1
0.25	0.875
0.25	0.875
0.2	0.867
0.2	0.75
0.25	0.9
0.25	0.925
0.25	0.875
0.25	0.9167
0.25	0.875
0.2	0.86
0.2	0.833
0.25	0.8928
0.25	0.875
0.2	0.9111
0.2	0.9
0.167	0.944
0.1428	0.96428
0.1428	0.9523

The following graph on the basis of data maintained in above table, shows the attribute utilization of datasets by using SAC and proposed MAC Approach-

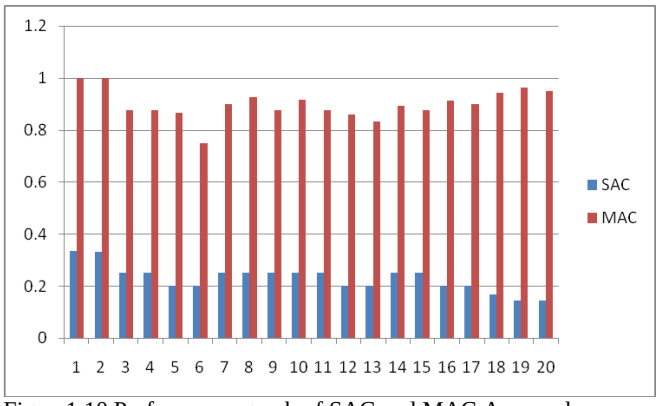

Figure1.10.Performance graph of SAC and MAC Approach

The x-axis of graph represents the no. of iterations and y-axis of graph represents the range of data.
The calculated result by using different no of queries and attribute of datasets by applying MAC Approach is represented by RED LINES in graph and The calculated result by using different no of queries and attribute of data-sets by applying SAC Approach is represented by BLUE LINES in graph.
The graph clearly shows the attribute utilization in MAC Approach is more in comparison to SAC Approach..Although all possible cases are also applied in MAC Approach but still it performs better than SAC Approach. Attribute utilization increased in MAC Approach so overall experimental result shows that MAC Approach is better than SAC Approach. If proposed MAC Approach is used then with attribute utilization, time and memory space also utilized in much better way in comparison to SAC Approach.

Benefits of the Proposed MAC (Multiple Attributes based Clustering) Approach:

1. **Improved Availability:** As all the needed data are available in one cluster so availability of data improves in this approach in compare to SAC approach.
2. **Reduce Ambiguity:** As data is clustered together at one place. So there is no ambiguity in this approach.
3. **Improve Ability:** As clustering is performed on the basis of need so this approach improves our ability to make clusters of multiple attribute datasets more efficiently in order to provide the result effectively n with minimum overhead.
4. **Reduce Access time:** In SAC approach, the data is collected through different different tables that increases the overall access time but in MAC approach the needed data is fetched from a single table or cluster so this new approach reduces the access time.
5. **Improve usage of available memory space**: Clustering is done in a order to use the available memory space more efficiently so MAC approach reduces the wastage of memory space that happens in SAC Approach because different different tables of clusters are maintained that takes memory space.

CONCLUSION AND FUTURE SCOPE:

The new presented MAC (Multiple Attributes based Clustering) approach improves the efficiency of existing SAC (Single Value Attribute based Clustering) approach for multiple attributes based datasets. It is found that MAC approach is best suitable for the datasets which are related to multiple attributes. This approach reduces the access time and wastage of available memory space and is more reliable than existing SAC approach in case of multiple attributes datasets.

A proposed MAC approach is thus a very useful and efficient approach to use for the multiple attributes of dataset. It plays a very important role in the presence of multiple attributes datasets in order to provide the clustered dataset that helps to give solutions to the queries in less time or in minimum time with better usage of memory space.

Further, work may be extended to introduce the new value of n means in this paper we take the value of n=4 means four multiple attributes datasets, So this value of n can be increased and also work can be done on this increased value of n in future.

References

[1] Tryon, Robert C. (1939). Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers.

[2] Cattell, R. B. (1943). "The description of personality:Basic traits resolved into clusters".Journal of Abnormal and Social Psychology 38: 476–506. doi:10.1037/h0054116.

[3] Everitt, B.S., Landau, S. and Leese, M. (2001), Cluster Analysis, Fourth edition,Arnold.

[4] ZUPAN, J. 1982. Clustering of Large Data Sets. Re search Studies Press Ltd., Taunton ,UK .

[5] MR. Anderberg, *Cluster Analysis for Applications*, Academic Press, London (1973).

[6] J. A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, New York (1975).

[7] Peter Hannappel, Reinhold Klapsing, and Gustaf Neumann. MSEEC—a multi search engine with multiple clustering. In Proceedings of the 99 Information Resources Management Association Conference, May 1999.

[8] Tian Zhang, Raghu Ramakrishnan, Miron Livny. "An Efficient Data Clustering Method for Very Large Databases." In: Proc. Int'l Conf. on Management of Data, ACM SIGMOD, pp. 103–114.

[9] Bailey, Ken (1994). "Numerical Taxonomy and Cluster Analysis". Typologies and Taxonomies. p. 34. *ISBN 9780803952591*.

[10] M. Hernandez and S. Stolfo. The merge/purge problem for large databases. In *Proceedings of the 1995 ACM SIGMOD*, May 1995.

Biographies

KM SONAM RANI received the B.Tech degree in Computer Science and Engineering from Moradabad Institute of Technology, Moradabad, UPTU, (U.P), and Pursuing M.Tech in Computer Science & Engineering from KNIT, Sultanpur (U.P), India.

N. Badal is an Associate Professor in the Department of Computer Science & Engineering at KNIT, Sultanpur (U.P.), INDIA. He received B.E. (1997) from Bundelkhand Institute of Technology (BIET), Jhansi in Computer Science & Engineering, M.E. (2001) in Communication, Control and Networking from Madhav Institute of

ISSN:2319-7900

Technology and Science (MITS), Gwalior and PhD (2009) in Computer Science & Engineering from Motilal Nehru National Institute of Technology (MNNIT), Allahabad. He is Chartered Engineer (CE) from Institution of Engineers (IE), India. He is a Life Member of IE, IETE, ISTE and CSI-India. He has published more than 50 papers in International/National Journals, conferences and seminars. His research interests are evinced at Distributed System, Parallel Processing, GIS, Data Warehouse & Data mining, Software engineering and Networking.