Generalization of the Distributed Database with Specialization and Correctarisation on specific Query

Neelendra Badal, PhD
Associate Professor
Kamla Nehru Institute of Technology, Sultanpur

Ajay Kumar Trivedi
Research Scholar
Kamla Nehru Institute of Technology, Sultanpur

ABSTRACT
As the technology in the field of Computer Science evolved it facilitates the availability of information easily. To get any information, from heterogeneous or homogeneous databases, Distributed over different sites, through any authorised channel is now the matter of few key strokes and in few seconds information desired is before end user. But in this scenario it is seen that many other parts of the information made available to end user is not needed to him. This also generates traffic over the network. The user needed to do further analysis, on the retrieved information, to get some fruitful result that they actually needed. The work presented in this paper deals with this aspect of the information, in the form of datasets, provided to user. Now it is required to put that information in such a way that a Generalised representation of the information / dataset, with only required attributes, has been provide to the end user, which is worthy for him. The extraction & presentation of Generalised dataset followed by Specialisation and Correctarisation of datasets, saves the time of end user and space as well.

General Terms
Generalisation of distributed dataset

Keywords
Genaralisation, Correctarisation, Specialisation, Distributed Dataset

1. INTRODUCTION
A Distributed Database is scattered over different sites, communicated over computer network. The end users did not bother about the organisation and location of database distribution over the network while accessing the information. This incurred the cost in terms of time to the query which is used to extract the desired information and overhead on the network. In case of distributed database the end user is transparent to the database and is made available as if it is centralised.

The work proposed in this paper tends to analyse the time required to extract the Generalise [3,4,6,7,8] the datasets achieved from different databases, scattered over different sites, whether Heterogeneous or Homogeneous databases. The analysis is done by comparing with extraction of ungeneralised datasets and Generalised datasets. These Databases are not required to be logically inter-related, as each of them may have different sets of data. These databases are specialised with respect to query on the desired database or databases and generate output based on the correction made on the Specialised query with respect to their attributes which in turn Generalise the datasets thus achieved.

The Generalisation is aimed to reduce the communication cost, in terms of time to access and retrieved desired datasets, among the different databases, from which the dataset or datasets are to be extracted, by searching and putting them together for execution of specialised query. This also reduces the load over the network.

The database is Specialised and Correctarised, the extraction of Generalised dataset, is easy based on the specified qualifiers in the query.

2. GENERALISATION
Generalisation is the process of extracting the information from different tables, that resides at different sites and in different databases (heterogeneous or homogeneous), that extracts several tuples, and present them in the form where all desired attributes are collected and placed as information of single tuple.

In ***Generalisation***, all participating tables, which are temporary and generated at runtime, from different databases are collected at one place, at local site, and generated respective single, tuple based, tables on which the specialization [1,2] and Correctarisation of specific query is applied.

The Generalisation is achieved implementing Correctarisation on the specialised Datasets extracted from different distributed databases, whether homogeneous or heterogeneous.

The aim of Generalisation to collect attributes, from different tables, resides at different sites of distributed databases, to form a single tuple of information at on place through specific single query

3. GENERALISATION PROCESS

3.1 Steps of Generalisation
To achieve Generalisation following steps are taken :

1. Search for the location of the desired database through mining.
2. Search for the desired table in that database as per the query qualifiers.
3. Extract the desired tuple or tuples from that particular searched table to place them into local site.
4. Generate a local database (once) and generate a table, based on the schema of extracted tuples, runtime and place the data on it.
5. Repeat the above steps 1 to 3 till all database has been searched. For each extracted tuple there should be a table, at step 4.
6. Now, extract desired attributes from different tables, created runtime on local site, through specialised query, by specifying the attributes to be selected for next step.

7. By making Correctarisation, i.e. applying conditions for the desired result.
8. The above step makes a Generalised form of distributed database at local site which reduces communication cost between databases at different location and network load and dependency for the same for some extent as we have gathered all desired tuples at local site.

If a user requires dataset, a tuple, generated from attributes from different tables resides at heterogeneous or homogeneous databases, different queries has to be executed to extract datasets as per predicates of query. Then extract desired attributes from these output tuples and frame a resultant tuple that has desired result for the end user. This process requires queries to be executes several times and separately.

The Generalisation process proposed in this paper requires to execute query at once that gathers all the desired tuples from different distributed databases, which are dedicatedly connected to the local database, make them homogeneous at local database at local site and generate resultant dataset to the end user.

The specialisation phase from the above figures ensures that tuples from different database is extracted as per the predicates of the query and gathered at local database, in homogeneous form, with their original schema in different tables. Then correctarise these tables, extract desired attributes as the requirements of the end user, and generate a generalised tuple having attributes from different tables.

The above process gives user collective information at once, and end user did not to need to process this data further to generate its desired result from that output. This saves time and space as well to the user and system.

From the figure 1, it is apparent the Generalisation is achieved broadly in two phases

Phase 1

a) Searching the Databases where desired information or datasets are resides.

b) Indentify the tables where tuples or datasets resides.

c) Extract those tuples and bring them into one place, i.e. place them into tables generated at local database as per the schema of the extracted tuples.

Phase 2

a) Extract tables with specific tuples, as per the need of the query, from local database, which are participating in the process of Generalisation.

b) Extract only desired attributes, from the above extracted tables, to get resultant dataset or datasets, as per the conditions of the query.

c) This creates a Generalised form of Database, which has all necessary information, the desired attributes in a tuple, required from query.

Once the datasets are at local site, it is very easy to extract the desired dataset or datasets, without having any burden on the network traffic and also saves the time to extract the information in comparison to directly fetched from the source at another site..

3.2 Block Diagram of Generalisation Process

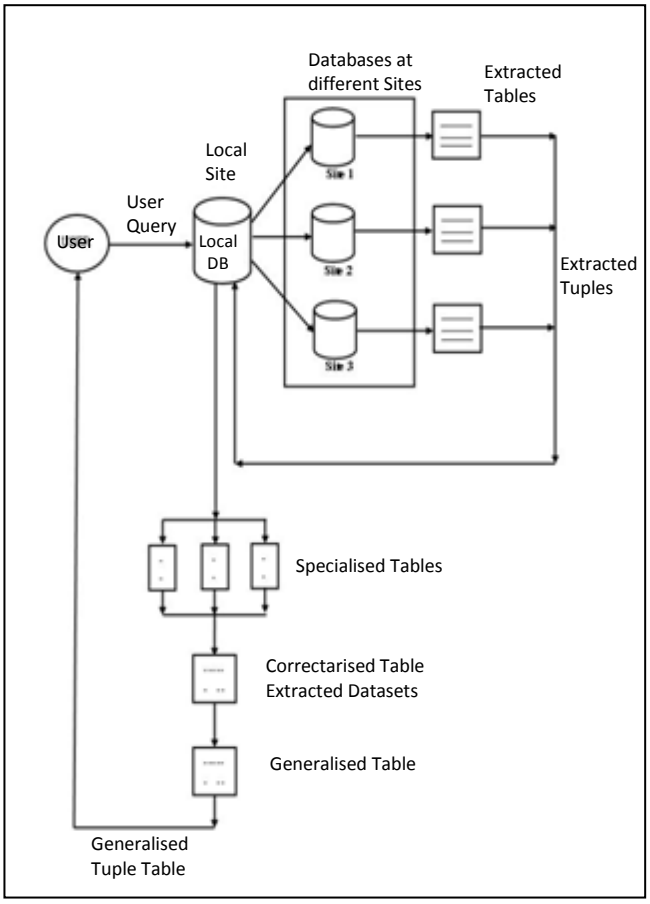

Fig 1. Block diagram of Generalisation

The above block diagram describes that when a user wants to extract information, the user query must be send to the database resides at local site. This database stores all the required datasets in different tables, which can be accessed by the user. If the information seek by the user is present then it may be extracted to the user as per the query of the user. If it is not present in the local database then, the query is further send to the dedicated distributed databases to search the information. The dedicated databases (heterogeneous / homogeneous), gives the permission to the user, to access the information at their site

Once information found in the tables, firstly target tables are extracted and then the target tuple(s)/dataset(s) from that table has been extracted and bring it into the local database, without altering the schema of the original dataset.

The specialization is performed by extracting the dataset as per the query and retaining the participating attributes only in the table. In case of Correctarisation only strictly participating attributes are selected and used for extraction of the information as per the query from the user.

3.3 Flow Chart of the Generalisation

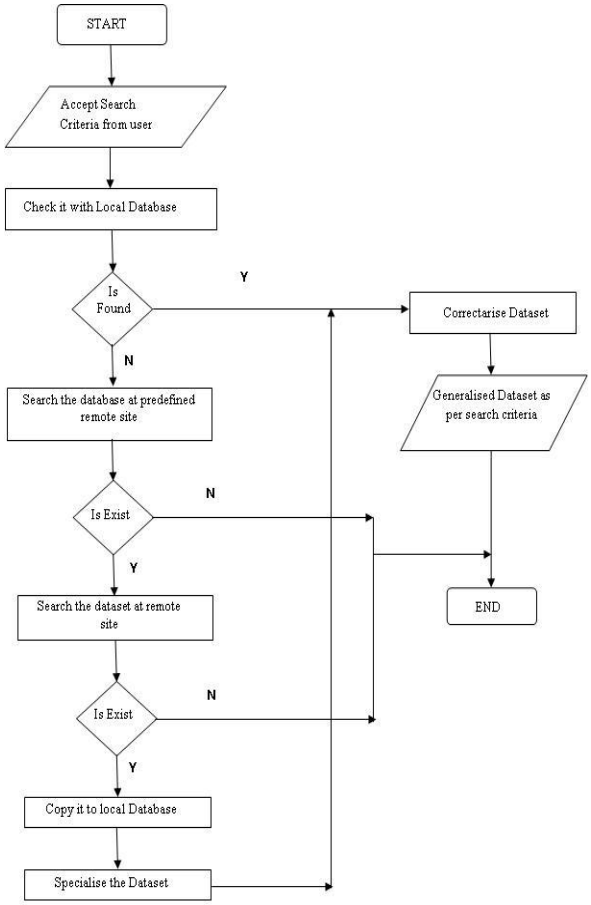

Fig 2: Flow Chart of Generalisation

4. EXPERIMENT IMPLEMENTATION

4.1 Extraction of Ungeneralised Dataset

For the extraction of ungeneralised dataset, following tables are used, which are assumed to be located at different sites and heterogeneous in nature. Also it is assumed that users are allowed and authorised to connect these locations to extract dataset(s).

The first database contains table **Student** having current information of the students, as per following schema

Table 1. Table shows the Schema of Student table

Attribute Name	Description
Stroll	Unique ID of Student
stEnroll	Enrollment id of the Student
stName	Name of Student
fName	Father's Name
DOB	Date of Birth
stCategory	Category of Student
pAddress	Permanent Address
pPhone	Phone Number
Eml	Email id
URank	Rank in Examination
adm Course	Admission taken in the Course
Stream	Stream of the Course
admYear	Year of Admission
Semester	Semester of the Admission
stType	Student Status Regular / Private
Sex	Sex of the Student

The second database contains table **Academic** having other Academic information of the students, as per following schema

Table 2. Table shows the Schema of Academic table

Attribute Name	Description
stName	Student Name
Roll	Roll Number
fName	Father Name
DOB	Date of Birth
stCatogary	Category
stClass	Class
passYr	Passing year
Board	Board/Univ of Study
Inst	Institute of Study
Subj	Subjects
perMarks	Marks obtained in %

It is not necessary that both tables has any direct relationship, as both are located at different database resides at different sites. Both may have an attributed, single or composite, in common to extract dataset. For example an student studying at any institute may have its previous academic credentials at different institutes databases which are not directly connected with current institution of the student. But current institute is allowed to extract students' dataset as per requirement from previous institute database.

Following is the Query that extracts datasets from other database(s) regarding students' previous academic credentials, as per the search criteria given to the query. Here this experiment uses MS Access as one database and MS EXCEL as another database, both are heterogeneous.

```
//Following code extracts specialised datasets based on the
   searching criteria, name and father name, given by the
   user
If txtnm.Text <> Empty And txtfnm.Text <> Empty Then
qry1 = "select * from STUDENT where STUDENT.stName
          LIKE '" & txtnm.Text & "%' and
          STUDENT.fName LIKE '" & txtfnm.Text & "%'"
//Following code extracts specialised datasets based on the
   partial searching criteria, name, given by the user
qry1 = "select STUDENT.stRoll, STUDENT.stName,
          STUDENT.fName, STUDENT.DOB,
          STUDENT.stCat from STUDENT where
          STUDENT.stName LIKE '" & txtnm.Text &
          "%'"
//Following code extracts specialised datasets based on the
   searching criteria, Unique ID, given by the user
qry1 = "select STUDENT.stRoll ,STUDENT.stName,
          STUDENT.fName, STUDENT.DOB,
          STUDENT.stCat from STUDENT where
          STUDENT.stRoll LIKE '" & txtRoll.Text & "%'"
//Following Code Executes the Query to Extract the Dataset
from other database
          rec.Open qry1, con, adOpenDynamic,
adLockOptimistic

Extraction of Dataset from another database, MS EXCEL,
   based on the search criteria input by the user
   If txtnm.Text = Empty And txtfnm.Text = Empty And
   txtRoll.Text = Empty Then
          sql = "select * from [studAcademic$] where
   stName LIKE '" & txtnm.Text & "%'"
//Code Checks if student name & father name as searching
   criteria for extraction of unGeneralised dataset
   If txtnm.Text <> Empty And txtfnm.Text <> Empty
   Then
          sql = "select * from [studAcademic$] where
          stName LIKE '" & txtnm.Text & "%' and fName
          LIKE '" & txtfnm.Text & "%'"
//Code Checks if student name is enter as searching criteria
   for extraction of unGeneralised dataset
   If txtnm.Text <> Empty Then
          sql = "select * from [studAcademic$] where
   stName LIKE '" & txtnm.Text & "%'"
//Code Checks if Unique ID is enter as searching criteria for
   extraction of unGeneralised dataset
   If txtRoll.Text <> Empty Then
          sql = "select * from [studAcademic$] where stRoll
LIKE '" & txtRoll.Text & "%'"
   End If

//Following Code Executes the Query to Extract the Dataset
from one database
res.Open sql, conx, 3, 3, 1 – adCmdText
```

Fig 2. Query Code for Extraction of ungeneralised Datasets from different databases

Below is the table that shows extraction of unGeneralised datasets with the time variation

Table3. Table shows the Output of the experimental setup of ungeneralised dataset

Searching / Extracting Records from tables in different Databases			
Extracted Record	**Total Available Records**	**Time (ms)**	**Search Criteria**
3	606	47	KAILASH YADAV
6	606	30	rahulsingh
3	606	15	0618731006 Unique ID
3	606	16	0618731046 Unique ID
3	606	15	0618731044 Unique ID
4	1851	63	2818710402 Unique ID
3	1851	31	AMIT NAIK
3	1851	31	AASTHA SINGH
3	1851	46	rubikanujiya , srikailashkanujiya
3	1851	47	0718740035 Unique ID
3	1851	31	0718710040 Unique ID
3	1851	64	PREETI , SRI RAMDAS

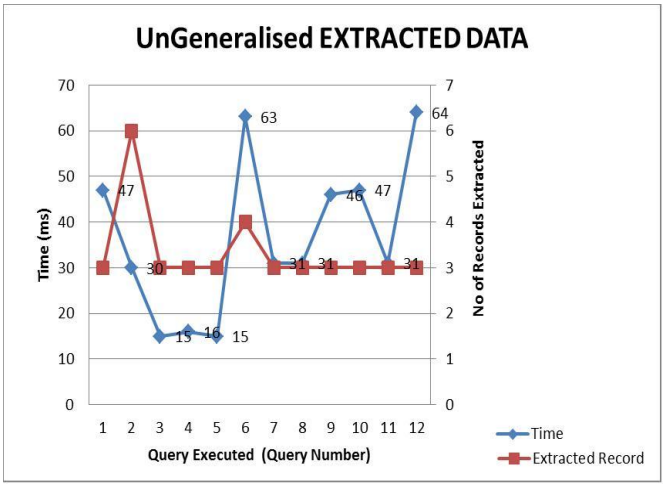

Fig 3: Graphical representation of Extraction of ungeneralised dataset

4.2 Extraction of Generalised Dataset

Following tables, ***Student*** and ***Academic***,are used for Generalisation that contains attributes extracted after specialization & Correctarisation. From table ***Student*** following attributes are selected

Table 4: Table shows the Extracted attributes for the Generalisation from Student table

Attribute Name	Description
Stroll	Roll Number of the Student
stName	Name of Student
fName	Father's Name of Student
DOB	Date of Birth
stCategory	Category of Student

From table ***Academic*** following attributes are selected

Table 5: Table shows the Extracted attributes for the Generalisation from Academic table

Attribute Name	Description
Stroll	Roll Number of the Student
stClass	Class
passYr	Passing year
Board	Board/Univ of Study
perMarks	Marks obtained in %

Following is the Query, implemented in experimental setup, used to extract the Generalised dataset from above tables

"select STUDENT.stRoll, STUDENT.stName, STUDENT.fName, STUDENT.DOB, STUDENT.stCat, a.stClass, a.board, a.passYr, a.perMarks, b.stClass, b.board, b.passYr, b.perMarks from STUDENT, (select * from Academic where stRoll LIKE ' " &txtRoll.Text& "%' and stClass='10') as a, (select * from Academic where stRoll LIKE '" &txtRoll.Text& "%' and stClass='12') as b where (STUDENT.stRoll LIKE '" &txtRoll.Text& "%' and (STUDENT.stRoll=a.stRoll) and (STUDENT.stRoll=b.stRoll))"

Fig 4. Query Code for Extraction of Generalised Datasets

Table 6: Table shows the Experimental output of Generalised Extraction of Dataset

GENERALISED EXTRACTION of datasets			
Record Extracted	**Total Records**	**Time (ms)**	**Search Criteria**
1	606	15	KAILASH YADAV
8	606	15	rahulsingh
1	606	15	0618731006 Unique ID
1	606	16	0618731046 Unique ID
1	606	15	0618731044 Unique ID
1	1851	31	2818710402 Unique ID
1	1851	31	AMIT NAIK
1	1851	16	AASTHA SINGH
1	1851	31	rubikanujiya , srikailashkanujiya
1	1851	31	0718740035 Unique ID
1	1851	31	0718710040 Unique ID
1	1851	31	PREETI , SRI RAMDAS

Below depicted graph shows the time variation of extracted dataset against no of dataset extracted

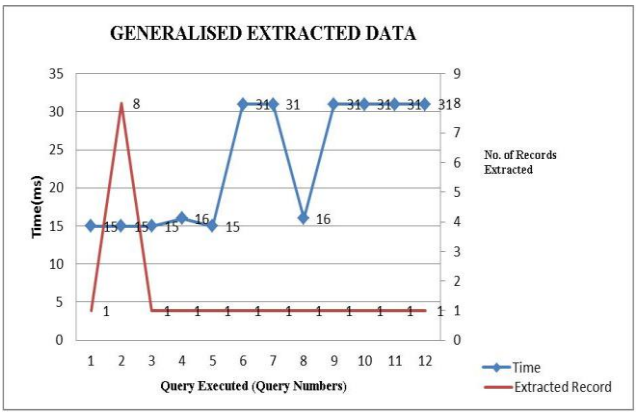

Fig 5. Graphical representation of Extraction of Generalised dataset

5. RESULT ANALYSIS

As per the output gathered from the experimental setup following results are analysed. For the comparison, following e graph represents the comparison between unGeneralised and Generalised dataset.

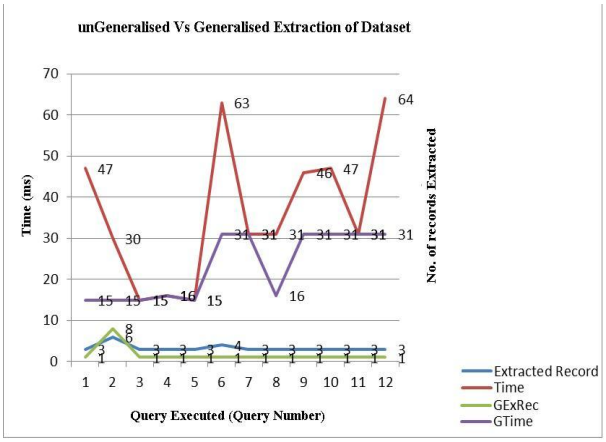

Fig 6. Graph representing of Extraction of ungeneralised and Generalised dataset

From the above graph it is apparent that it takes less time to extract the dataset in Generalised method in comparison to the ungeneralised method. Also it is also apparent that number of records extracted in Generalised method is less than in comparison to the ungeneralised method.

Following is the table that shows cross comparison of the data extracted as unGeneralised dataset and Generalised datasets.

Table 7: Table shows the Experimental output of Generalised Extraction of Dataset

UNGENERALISED EXTRACTION				GENERALISED EXTRACTION			
Record Extracted	**Total Records**	**Time**	**Search Criteria**	**Record Extracted**	**Total Records**	**Time**	**Search Criteria**
3	606	47	KAILASH YADAV	1	606	15	KAILASH YADAV
6	606	30	rahulsingh	8	606	15	rahulsingh
3	606	15	0618731006 Unique ID	1	606	15	0618731006 Unique ID
3	606	16	0618731046 Unique ID	1	606	16	0618731046 Unique ID
3	606	15	0618731044 Unique ID	1	606	15	0618731044 Unique ID
4	1851	63	2818710402 Unique ID	1	1851	31	2818710402 Unique ID
3	1851	31	AMIT NAIK	1	1851	31	AMIT NAIK
3	1851	31	AASTHA SINGH	1	1851	16	AASTHA SINGH
3	1851	46	rubikanujiya , srikailashkanujiya	1	1851	31	rubikanujiya , srikailashkanujiya
3	1851	47	0718740035 Unique ID	1	1851	31	0718740035 Unique ID
3	1851	31	0718710040 Unique ID	1	1851	31	0718710040 Unique ID
3	1851	64	PREETI , SRI RAMDAS	1	1851	31	PREETI , SRI RAMDAS

The above table compares the time taken to extract dataset as well as number of records extracted, between ungeneralised and Generalised method of extraction of datasets

Following is the average comparison table and graph of the same that the Generalisation takes less time to extract the whole record as a tuple in comparison to ungeneralised method where records are extracted separately from different distributed databases

Table 8: Table shows Average Comparison Search Result

Average Search Result						
UnGeneralised Search			**Generalised Search**			
Extracted Record	Total Records	Time	Extracted Record	Total Records	Time	Variation in Time (%)
4	606	25	3	606	16	36
4	1851	41	1	1851	30	27
8	3269	39	4	3269	47	15
4	4676	59	1	4676	47	21
3	5283	55	1	5283	51	8
10	6669	77	4	6669	68	12
3	6669	66	1	6669	60	10
			Average Variation (%)			**19**

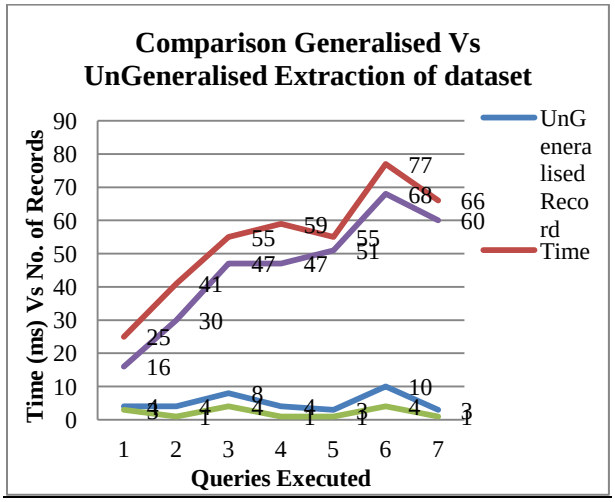

Figure 7. Graph showing Average Comparison

It is analysed from the average table and graph obtained, that an improvement in the performance of 19% has been observed while using the Generalised approach of the extraction of Dataset.

6. CONCLUSION

From the above result it is clear that extraction of generalised record takes less time and space as well in comparison to extraction of ungeneralised records from different databases. From the above graph and table it is apparent that an improvement of 19% has been observed for Generalisation.

The Generalisation result brings up all the desired attributes together as one tuple, and the end user did not need to do any further analysis on the dataset extracted. While in ungeneralised approach, the as shown in the above table and graph, shows that for extraction of a datasets, as per the search value, from different tables, it brings whole tuples from both tables. This will leave end user to further analysis on extracted datasets, which will also consumes time.

The other part of this work is the number of datasets extracted. In Generalisation a single tuple is extracted after Specialisation and Correctarisation has been applied on the single specific Query. While in case of extraction of ungeneralised dataset, separate query is needed to be

executed, for separate database and table. This will extract separate datasets/tuples with undesired attributes, for a specific search criteria, against a single tuple in Generalisation.

The dataset thus achieved after Correctarisation step formed as a Generalised dataset. In Generalisation only searched dataset or tuple with desired attributed has extracted from different tables. This also done with the less time and less space, as it extract only one dataset per search criteria as compared to ungeneralised dataset.

7. FUTURE SCOPE

This work has taken up the case study of student and its academic credentials Database, heterogeneous databases,there may be a good scope to work with other databases such as Transport Department, Passport Issuing Authority and many more where the verification of a person's credentials is needed rapidly.

This work may be extended in the direction of parallel processing /computing environment. In this case, Generalisation process may be taken up by several processors in parallel processing / computing environment. This may increase increases the reliability of the system.

8. ACKNOWLEDGEMENT

Authors of this paper extends our cordial thanks to Director KNIT & HOD Computer Science & Engg., Dr. A.K. Malviya, Professor, to support our efforts to complete this work.

9. REFERENCES

[1] Saxena N., Arora R., Sikarwar R. and Gupta A., "*An efficient approach of association rule mining on Distributed Database Algorithm",* in International Journal of Information and Computation Technology, ISSN 0974-2239 Volume 3, No 4 (2013), pp 225-234.

[2] SikhaBagui (University of West Florida, USA), "*Mapping Generalisations and Specialisations and Categories to Relational Database",* a Chapter in Handbook of Research on Innovations in Database Technologies and Applications: Current and Future Trends.

[3] S.M. Deen, R.R. Amin, G.O. Ofori-DWUMFUO and M.C. Taylor, "The Architecture of a *Generalised Distributed Database System-PRECI*",* PRECI Project, Department of Computing Science, University of Aberdeen AB9 2UB, Scotland.

[4] S.M. Deen et al., *Thedesign of canonical database system (PRECI)*, the Computer Journal , Vol. 24, No. 3 (1981).

[5] Park B. and Kargupta H., "*Distributed Data Mining: Algorithms, Systems and Applications",*Deptt of Computer Sc. & Electrical Engg, University of Maryland Baltimore County.

[6] M. Tamer Ozsu, Patrick Valduriez, a book titled "*Principles of Distributed Database Systems",* Springer, third Edition.

[7] Johnson E. & Kargupta H., (1999). *Collective hierarchical clustering from distributed, heterogeneous data*. In Lecture notes in computer science (Vol. 1759, p.221-224). Springer-Verlag.

[8] Kaundal G., Kaur S., Vashisht S., *"Review on Fragmentation in Distributed Database Environment"* in IOSR Journal of Engineering (IOSRJEN), Vol. 04, Issue 03 (March,2014), V6, PP28-32

10. AUTHOR'S PROFILE

Dr. Neelendra Badal, B.E., M.E., Ph.D. Working as Associate Professor at Kamla Nehru Institute of Technology, Sultanpur in the Department of Computer Science & Engineering. He has more than of 20 years of teaching experience. His area of specialization is Distrtibuted Computing, Web Technology, Data Warehouse & Mining, Communication, Control, Networking, Information Technology, GIS. He has published more than 36 National/International publications.

Mr. Ajay Kumar Trivedi, B.Sc. C-DAC, MCA, is an Research Scholar of M.Tech., at Kamla Nehru Institute of Technology, Sultanpur in the Department of Computer Science & Engineering. He has more than of 10 years of working experience in the field of Computer Science & Engineering, at FGIET, Raebareli. His area of interest is Distrtibuted Database, Web Technology, Networking, Information Technology.