Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Approach to Data Reduction in Data Warehouse

Md. Ishtiyaque Alam
Department of Computer Science & Engineering
Jamia Hamdard (Hamdard University)
New Delhi-110062

Jawed Ahmed
Assistant Professor
Department of Computer Science & Engineering
Jamia Hamdard (Hamdard University)
New Delhi-110062

Abstract: Data reduction is a process to convert polluted, unmanageable or duplicate data into final meaningful data for any organization. Duplication of record is a serious challenge in data warehouse. It makes our performance slow and reduces our storage capacity. To rid of this problem, we use data reduction techniques, algorithms and tools. There may be various reasons for this duplication issue which can arise due to integration of data at various steps of our end to end cycle of data migration. When we collect data from different sources and integrate it then we find some data may be identical or seems identical. Duplicate detection is very difficult to apply in real world. In continuation of past work purposing an algorithm which uses the cluster for comparison of record. A cluster contains sets of records. Each set contains one or more data from a detected cluster. Entire clusters should not always be saved due to efficiency reasons.

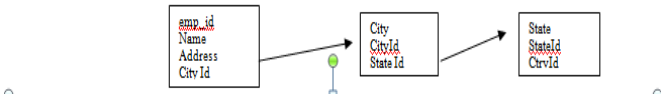
Keywords: data reduction, duplicate detection, data cleaning, partial duplication, clustering.

1 INTRODUCTION

A data warehouse is constructed by integrating data from multiple heterogeneous sources which support analytical reporting, structured and/or ad hoc queries and decision making. Data is integrated from multiple operational databases, which has different formats. Data integration is very important to decision making and analysis.

Now a day's databases play an important role in IT and economy based industries. Many companies depend on the efficiency of databases to perform all operations. Therefore, the qualities of data must be good so that company take appropriate decision to conduct business. If the quality of data is not good, the strategic decisions taken on the basis of that data may not be good which is not good or we can say disaster for any company.

Data has to be in integrity, and if exceeds the criteria then it is duplicate. Data is considered as an important asset of a company but due to data changes and sloppy data entries duplication arises[1]. Problem in data quality happens during data entry like missing integrity constraints, missing record, spelling mistake, and duplicated records. Data cleaning is process to remove from this complication. The act of detecting and removing or correcting a database's dirty data (i.e. redundant, incomplete)[2].



emp_ID	Name	Address	City Id
1	Ram Kumar Asso.	#1, Sea View Frq	C1
2	Chander pandey	#22, Main Road	C2
3	Chander pande	#22, Main Road	C3
4	Ram Kumar Associates	#1, Sea View	C4
5	Irfan Corp.	#10 kali ghati	C5
6	Vikas films	#5 vikas nagar	C6
7	Irfan Corporation	#10 K.G	C7
8	kailash Consultants Ltd.	#8 Chandan Road	C8
9	kailash Consultants	#8 Chand Rd.	C9

City Id	City Name	State Id
C1	Jabalpur	S1
C2	Jabalpur	S2
C3	Jabalpur	S4
C4	Jabalpur	S3
C5	Vijaynagar	S5
C6	Vijaynagar	S6
C7	Varanpur	S6
C8	Alipur	S7
C9	Alipur	S8

State Id	statename
S1	rajesthan
S2	rajesthan
S3	rajesthan
S4	ajmer
S5	up
S6	utter pardesh
S7	andhra pradesh
S8	andhra pradesh

Table: 1 Duplicate Record

Data reduction is process to transformation of numerical or alphabetical digital information into a corrected, ordered, and simplified form. The basic concept is the reduction of large amounts of data into the meaningful parts.

Data reduction is the way of keeping down the number of data that needs to be stored in a data warehouse. Data reduction will help to increase storage efficiency and reduce costs.

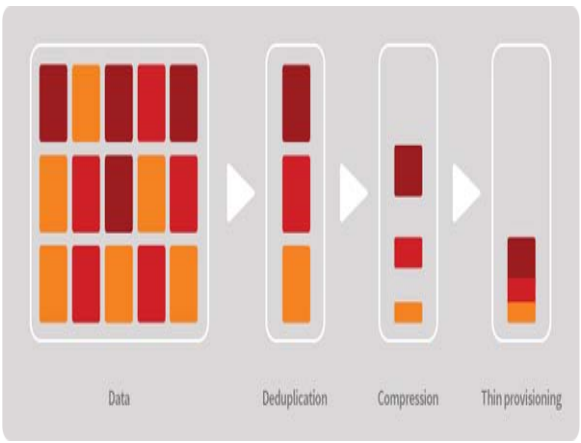


Fig. 1 Data Reduction

The proposed technique identifies and removes data quality problems occur because of data entry operator errors such as spellings mistakes, missing integrity constraints, missing, noise or contradicting entry, null values, misuse of abbreviations, and duplicated records not only fully but also partially duplicated records. The clustering algorithm is utilized to decrease the quantity of examinations by framing clusters and the divide and conquer approach is utilized to match records inside the clusters.

The algorithm scans the sorted database with a priority queue of record subsets belonging to the last few clusters detected. The priority queue contains a number of sets of records. My proposed technique deal with all such kinds of data quality problem in a record.

In this paper, we organized as follow. We describe the related work in Section 2. We represent the core of our approach in Section 3. Conclusions are introduced in Section 4.Finally, references in Section 5.

2 RELATED WORKS

Several earlier proposals exist for the problem of duplicate elimination.

Willetty et al [3] proposed Data Reduction Techniques with the objective of contrasting how an expanding level of pressure influences the execution of SVM-sort classifiers

Bitton et al. examined the end of copied records in substantial information documents by sorting which unites indistinguishable records [4].In the event that sorting depends on dirty fields, duplicate records can never come together. Sorting method is inefficient for large data files having typographical errors.

Hernandez et al. discuss the problem of merge/purge in a large database [5]. They form token keys of selected fields of the database table. The effectiveness of merge/purge approach relies on upon the nature of the picked keys which may flop in bringing conceivable copied records close to each other for subsequent comparison. Token based information purging strategy characterizes smart tokens that are utilized to recognize and get rid of similar data [6].

For example, token made for the name column containing values 'Bandana Persad' and 'Bhawna Persad' will be 'BP'. Even if real records are different but token made for these records is same. Thus non duplicated records are considered as duplicated records.

3 PROPOSED SYSTEM

In data warehousing, experts decide, such redundancy (duplicate information or deficient data) can make the investigation create the wrong outcome that misdirect the examiner along these lines the business will endure. There is a need to check and expel copied records from the data warehouse. Duplication impacts the general execution of data stockroom so data diminish confine the repetition. Data reduction procedure is utilized to recognize and evacuate copied records.

This issue can be clarified for instance in the hospital business. On the off chance that a client's record is put away more than one times, the organization will send him sends more than once as he is viewed as another individual, however, in reality, he is a similar individual. For the success of any company, we should collect correct data for decision making.

So we need to free from this problem by removing duplicated records from the data warehouse. As I mentioned that copied record affects the overall performance of data warehouse and also slows down the storage efficiency and capacity optimization.

The proposed algorithm is fundamentally used to recognize and remove copied records in data warehousing. This algorithm enhances the data quality as well as the execution of the data warehouse.

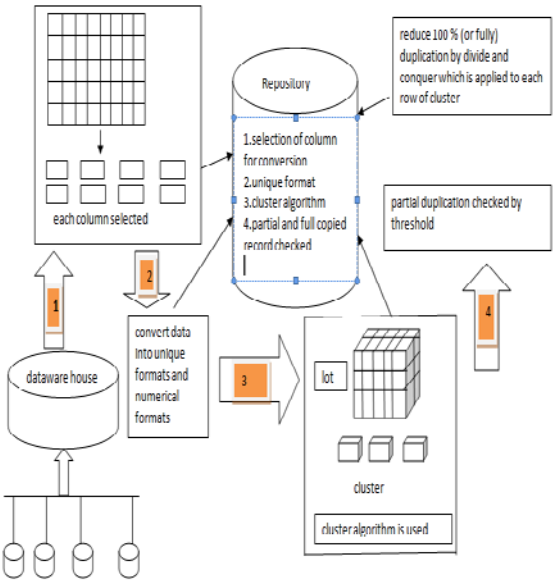


Fig. 2: Duplicate Data Reduction Process.

I. Convert data into uniform format

Algorithm initially brings the data into a uniform configuration. As the data encouraged to the information distribution center originates from various operational frameworks, there could be various designing issues in information. One of them is information sort arrange jumble. For instance, a date might be in the organizations of dd-mm-yyyy, mm-dd-yyyy, or yyyy-mm-dd. So also, a telephone number having nation code and city code in one record yet in another record same telephone number without city and nation code. Such arranging and missing qualities issues are settled and information is conveyed to a uniform arrangement.

Essentially, shortened forms are extended. To institutionalize and evacuate irregularity in the information, our approach brings the information into a uniform configuration and afterward changes over all field values (regardless of whether string, numeric or date) into numeric frame on information. After transformation of the field values into numeric shape, an additional section is annexed putting away all the figured qualities into that segment comparing to applicable column isolated with comma(,)[9].

Input: Table created with various data format and shortened forms

Output: Uniform format table with additional attached property.

Algorithm for Unique Format

Begin
Loop column j= 1 to last column, n
Loop row *i*= 1 to last row, m
1. Put column values into uniform format
2. The special character (like punctuation) should be ignoring.
3. The variation of column values should be removing.
4. Short format are converted to full format
End loop
End loop
End

II. Clustering

Clustering is whose the way toward sorting out articles into gatherings individuals are comparative somehow. In other words Clustering is the task of grouping a set of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups (or clusters). 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[8].Clustering is partition data set into clusters, and one can store cluster representation only. It can be very effective if data is clustered but not if data is "smeared" [7].

III. Matching

After clustering step completed, the divide and conquer method is apply on each row of the cluster. This approach divides the values recursively into small pieces and continues the process until certain smallest size is reached. Then compares the single value of one record is with the single value of other record. If match is found between values of records then the percentage duplication of records is calculated.

The difference between these records is due to data entry operator error.

For example, data entry operator types 'Avid' instead of typing 'Abid' in the name field value in second row. Such a difference is called erroneous difference and corrected by the domain expert.

Validating the duplicates:

```
Select count (1), Name, Fname, Job, Salary
From <table_name>
Group by Name, Fname, Job, Salary
Having count(1)>1;
```

Table 2: Partially duplicated records

Name	Fname	Job	Salary
Abid	Mohammad	manager	10500
Avid	Mohammad	manager	10500

Table 3: Partially duplicated records

Name	Fname	Job	Salary
Ayan	Irfan	Prof	78000
Aman	Irfan	Prof	78000

In Table 3, two records are 75% duplicated (3/4×100 = 75%). There is an original difference between these records. For example, difference occurs in the name field of row 1 and 2. These two records are for two different individuals. Only name field values are different i.e. 'Ayan' and 'Aman' and other attributes value of both the records are same. When domain expert analyzes that the difference is original then he/she keeps the records [9].

If both records are fully duplicated as in Table 4, then the duplicate records are discarded and the original row is kept for example, in Table 4 both the records are 100%duplicated, when the divide and conquer approach is applied on these records, the system identifies that these records are fully duplicated .

TABLE 4: Fully Duplicated Records

Name	Fname	Job	Salary
Aliya	khan	Prof	78000
Aliya	khan	Prof	78000

If records are partially duplicated then the threshold value is checked. If the percentage of duplication crosses the threshold, the program displays those records and mentions clearly those attribute values having difference among them as an output to analyze whether this was an actual difference or erroneous difference among those records. For example, the difference in Table 2 is an erroneous sentry which is corrected and stored. In this manner, both records become identical or fully duplicated. The duplicated records are then discarded and single row is kept. But in case of actual difference between the records as in Table 4, both rows are kept.

Sometimes column values don't look fully matched but they are actually matched. For example: column name having two values one is "Asim Ali Asghar" and other is "Asim Asghar" are actually matched but it does not seem that these values are matched.

To identify such matching records, domain expert defines threshold for column value and when threshold value is crossed, the algorithm considers that the values of column are matched. Our algorithm not only specifies the threshold for the single column value matched but it also specifies the threshold for all columns to identify the partial duplicated records. For example, we have 10 columns and two record shaving 8 columns matched values and values of two columns are not matched. In such a situation domain expert needs to specify the threshold. If the matching values of records cross the threshold then it display those records and mention then on-matching values. The domain expert can correct if there is any erroneous difference, discard the record and keep the original entity otherwise leaves the records [9].

There are some symbols which will use in Algorithm.

Symbol	Description
d	Percentage duplications between records.
t	Threshold value specified by the domain expert
p	First position of record i.e. 1
q	Last position of record i.e. n
v	Number of values after dividing the row into two port

Input: Table with appended column and duplicated records,
Output: Cleaned table
Algorithm
Begin
Loop row i = 1 exit when last row, n

1. loop v, p =1 to q
2. if v>1 then go to step 9
3. else compare all the value with the corresponding value of other row
4. if match found b/w values then
5. Calculate the d and go to step 9
6. else go to step 13
7. divide (p+q)/2 and go to step 2

8. end loop
9. if d = 100% then discard the duplicated record and go to step 15
10. else if d >= t then
11. display the records and mention the attributes values having difference b/w values
12. if difference is due to data quality then correct the entities and go to step 8
13. else go to 15
14. end loop
15. exit

4. CONCLUSION

In this paper, I explained how to identify and get rid of duplicate data through algorithm. This algorithm used for conversion of record in a unique format and compare duplicate data. I also mentioned clusters which are helpful in reducing the number of comparisons. On the basis of these clusters, divide and conquer technique is used in each small clusters to identify and remove the duplicated records. Due to this, storage capacity and performance efficiency is increases and cost decreases in data warehouse. Algorithms reduce large data-set to a manageable size without any loss of important information represented by the original record.

5. REFERENCES

[1] Progressive Algorithms for Efficient Duplicate Detection Anusha Kenno1, Bindhu J S2.
[2] J. J. Tamilselvia and V. Saravanan, "Handling noisy data using attribute selection and smart tokens,"
[3] R. Georgescu, C. R. Berger, P. Willett, M. Azam_, and S.Ghoshal,―Comparison of Data ReductionTechniquesBasedon the Performance of SVM-type Classifiers"
[4] "Duplicate Record Elimination in Large Data Files" DINA BITTON and DAVID J. DeWlTT.
[5] M. A. Hernandez and S. J. Stolfo, "The merge/purge problem for large databases,"
[6] E. Rahm and H. I. Do, "Data cleaning: Problem and current approaches,"
[7] "DATA REDUCTION TECHNIQUES TO ANALYZE NSL-KDD DATASET" Shailesh Singh Panwar, Dr. Y. P. Raiwani.
[8] https://en.wikipedia.org/wiki/Cluster_analysis.
[9] " Removing Fully and Partially Duplicated Records through K-Means Clustering " by Bilal Khan, Azhar Rauf, Huma Javed, Shah Khusro, and Huma Javed.