UTILIZING ALLIANCE RULES FOR DATA WAREHOUSE CLEANSING

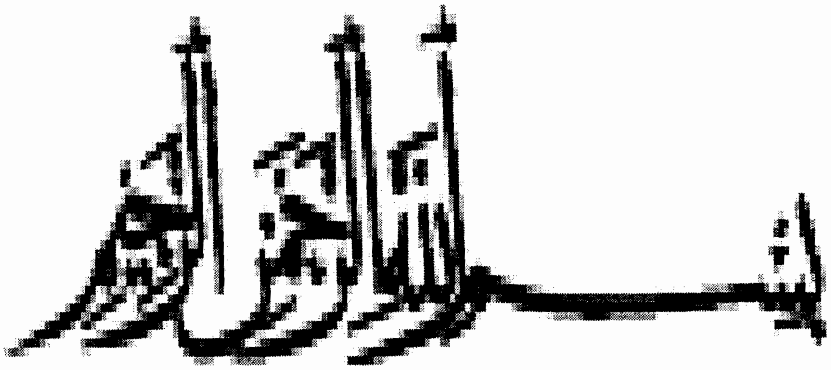
INTERNATIONAL ISLAMIC UNIVERSITY
ISLAMABAD

Ms. Masuma Abbas Rizvi
Registration # 515-FBAS/MSCS/F08

Supervised by:
Mr. Muhammad Imran Saeed Qureshi

Co-Supervised by:
Ms. Tehmina Amjad

Department of Computer Science
Faculty of Basic and Applied Sciences
International Islamic University Islamabad
2010



In the name of

ALLAH ALMIGHTY

The Most Merciful, The Most Beneficent

Department of Computer Science

International Islamic University Islamabad

Date: ____________

Final Approval

This is to certify that we have read the thesis submitted by **Masuma Abbas Rizvi**, registration # **515-FBAS/MSCS/F08**. It is our judgment that this thesis is of sufficient standard to warrant its acceptance by International Islamic University, Islamabad for the degree of **MS COMPUTER SCIENCE**

Committee:

External Examiner:

Dr. S. Afaq Husain
Professor,
Riphah International University,
Islamabad

____

Internal Examiner:

Ms. Zareen Sharf Khan
Assistant Professor,
Department of Computer Science,
Faculty of Basic and Applied Sciences,
International Islamic University,
H-10, Islamabad

____

Supervisor:

Mr. Muhammad Imran Saeed
Assistant Professor,
Department of Computer Science,
Faculty of Basic and Applied Sciences,
International Islamic University,
H-10, Islamabad

____

Co-Supervisor:

Ms. TehminaAmjad
Lecturer,
Department of Computer Science,
Faculty of Basic and Applied Sciences,
International Islamic University,
H-10, Islamabad

____

Dedicated to
my beloved parents,
sisters, brothers, friends
and my husband,
whose affection has always been the
source of encouragement for me,
and whose prayers have always been
a key to my success.

A dissertation Submitted To

Department of Computer Science,

Faculty of Basic and Applied Sciences,

International Islamic University, Islamabad

As a Partial Fulfillment of the Requirement for the Award of the

Degree of *MS Computer Science*

Declaration

I hereby declare that this Thesis ***"Utilizing Alliance Rules For Data Warehouse Cleansing"***neither as a whole nor as a part has been copied out from any source. It is further declared that I have done this research with the accompanied report entirely on the basis of my personal efforts, under the proficient guidance of my teachers especially my supervisor Mr. Muhammad Imran Saeed, and Co-Supervisor Ms. Tehmina Amjad. If any part of the system is proved to be copied out from any source or found to be reproduction of any project from any of the training institute or educational institutions, i shall stand by the consequences.

Masuma Abbas Rizvi
Registration # 515-FBAS/MSCS/F08

Acknowledgement

I bestow praises to Allah, The Most Merciful and Compassionate, whose bounteous blessings enabled me to recognize higher degrees of life, Who blessed me with good health, good circumstances, determination, and knowledge to accomplish the tasks.

I consider it a proud privilege to express my cordial gratitude and deep sense of obligation to my supervisor, **Mr. Muhammad Imran Saeed**, and co-supervisor **Ms. Tehmina Amjad**, who kept my morale high by their suggestions and appreciation. Their motivation led me to this success. Without their sincere and cooperative nature and precious guidance, I could never have been able to complete this task.

Finally I must mention that it was mainly due to my parents, sisters, brothers, friends and my husband's moral support and their prayers during my entire academic career that enabled me to complete my work dedicatedly.

Masuma Abbas Rizvi
Registration # 515-FBAS/MSCS/F08

Project In Brief

Project Title:	Utilizing Alliance Rules For Data Warehouse Cleansing
Undertaken By:	Masuma Abbas Rizvi Registration # 515-FBAS/MSCS/F08
Supervised By:	Mr. Muhammad Imran Saeed
Co-Supervised By:	Ms. Tehmina Amjad
Start Date:	July, 2009
Completion Date:	May, 2010
System Used:	Pentium IV
	Pentium IV

Chapter 1
INTRODUCTION

1. INTRODUCTION

1.1Motivation and Challenges

The completeness, correctness, and consistency of any large data set depend upon number of factors. Data entry and acquisition normally face errors both in nature simple and complex. Efforts have been given to this typical front-end process, but a chance of entry errors remains the same in large data sets. This failure to reduce errors in large data sets have number of impacts on organization, the higher operational costs can incurred, poor decision-making, the growing mistrust facts, within the organization and as a result, the diversion of focus of management from main issue, to something not that important. The solution to deal with this issue is to clean the data using some cleansing techniques.

Moreover to deal with data cleansing, a manual work is obviously not a good option as its laborious, time consuming and requires a huge amount to be spent during the entireoperation. The automation of the process of data cleansing for large data sets may be a good option with practical approach, which is also cost effective, time saving, and provides quality level of data in a data set. This problem has gained significant attention from researchers and these days work is being done to deal with data set issues and problems.

Data cleaning is particularly required for the integration of heterogeneous data sets, we specifically require data cleansing along with theschema transformations. Data cleaning isa major part of the so-called ETL processin data warehouses.

1.2Background

Data Cleansing is required to improve the data quality to make it fit for use by the user. Quality can be improved by removing inconsistent data, removing duplicates and reindexing existing data in order to achieve the most accurate and concise database.

Data cleaning is also required when the data is being merged from multiple parent databases. Data Cleansing can be performed manually or through specific software programs.

1.2.1 Data Warehouse

A **data warehouse** is a storage area of an organization's electronically stored data. Data warehouses are designed to assist reporting and analysis. According to Inmon, famous author for several data warehouse books, [14]

"*A data warehouse is a subject oriented, integrated, time variant, non volatile collection of data in support of management's decision making process*".

It not only stores the transaction data, but also the historical data. It distinguishes analysis workload from transaction workload and enables an organization to amalgamate data from multiple sources. Data warehouse gathers data from multiple sources in order to support organization's decision making, reporting, and analyzing. Data in a data warehouse should be coherent, accurate, and correct. Modifications have a strong impact on processes. The quality of data degrades due to constant updates.

Degraded quality of data produces incorrect results which lead to wastage of resources like time, money, human power, etc. It also affects the data mining process. Normally data mining is done to help out organization's decision making & planning. Data mining is a time consuming and costly process. Due to faulty data, inaccurate results will be produced.

1.2.2 Architecture of Data Warehouse

The architecture is the relationship between the parts of a system.The data warehouse architecture is shown in figure 1-1.

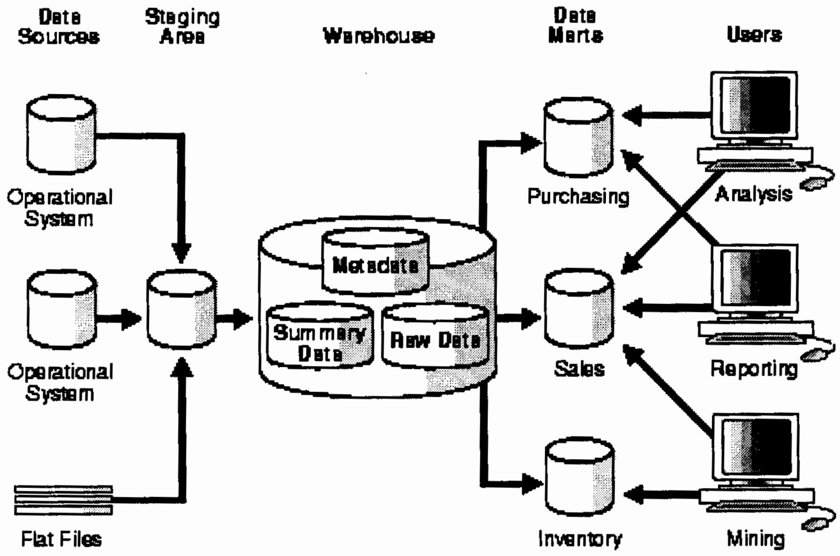


Figure 1-1: Architecture of Data warehouseadapted from [14]

The architecture of data warehouse includes tools for extracting data from multiple operational data sources (databases), as shown in figure 1-1. The main purpose of these tools is to first cleansethe data, then transform the data; and finally for loading the data into the data warehouse. It is also beneficial for periodically refreshing the warehouse to imitate updates at the sources and to flush out data from the warehouse, possibly onto slower backups.

There may be numerous departmental data marts in addition to the main warehouse. One or more warehouse servers are used to store and manage the data in the warehouse and data marts. A variety of front end tools like "*query tools*", "*report writers*", "*analysis tools*", and "*data mining tools*"are used to check the multidimensional views of data.

Finally, repository is used to store and manage metadata, and tools are used to monitor and administer the warehouse system.

1.2.3 Components of Data warehouse

According to [14], A Data Warehouse system comprises of three components: a database, a query tool and database interfaces.

Warehouse database extract data from existing operational database. To make it more useful and result driven, the warehouse system re-arranges the data.

The query tool helps in query generation, result displays, reports and data exports by allowing executives and other user's real-time access to the Data Warehouse database.

The database interfaces facilitates for updating the warehouse database on daily, weekly, monthly or yearly basis.To transfer the data from operational database to the data warehouse, these interfaces are used. It also re-packages the dataand carries out other operations such as pre-calculation of totals and ratios. These interfaces act like bridges to permit the query tool to access the data in its source environment.The construction of these bridges reduces the potential transfer problems as the data need not to be transported from the source system to the target system. It reduces the burden on IT employees, also provides the availability of data in real-time.

1.2.4 Data Mart

A data mart is a subset of a data warehouse for a single department or function.

> "*Data marts are analytical data stores designed to focus on specific business functions for a specific community within an organization. A data mart may have tens of gigabytes of data rather than hundreds of gigabytes for the entire enterprise.*"[14]

A data warehouse is a central repository (can be physically distributed); data mart is a data repository that may or may not be a result of a data warehouse.There can be one or more than one data marts of an organization;each data mart is related to one or more business functionsfor which it was designed. It is possible that the data marts are dependent on other data marts of the same organization.

Normally the related data marts are designed using some common facts and dimensions. The business unit is thought-out as the owner of the data mart including all the hardware, software and data.

1.2.5 Extract, Transform and Load (ETL)

Extract, Transform and Loaddefines a group of tools that helps in ensuring that the data has up-to-date records before being entered into the data warehouse.

ETL performs the following main tasks:-

1. Identification of important, appropriate and related information at the source side
2. Informationextraction
3. Transporting this information to the Data Staging Area (DSA)
4. Transforming of the information coming fromseveral sources into a common and standard format
5. Cleansing of the data set, onthe basis of database and business rules
6. Loadingthe data to the data warehouse and refreshing the data marts.

1.2.6Data Cleansing

According to [13],Datacleansing is defined as:

> *"Data cleansing is the process of ensuring that a set of*
> *Data is correct and accurate by detecting and removing*
> *and/or correcting a database's dirty data (that is incorrect,*
> *out-of-date,redundant, incomplete, or formatted incorrectly)"*

The target of data cleansing is to clean up the data in a database.It is also responsible for ensuring consistency among different sets of data that have been merged from multiple databases. Datacleansing can take place within a single set of records, or between multiple sets of data which need to be merged.

Data cleansing is also required, whenever two systems of data need to work together. If a company has two branches, sharing many of the same customers, then it is required for both the branches to have matching data and data must be complete and accurate. If a client updates his record in one branch, the system should automatically update the data at the other branchto ensure efficiency. This issue gives us another quality of data cleansing, that it is also responsible for ensuring accuracy and consistency among the different data sets

As data warehouse or data marts deal with millions or trillions of data records, so during the insertion,modification or deletion of the data, errors can be generated. The main target of data cleansing is to diminish these errors. The data act as useful and meaningful information only if it is free from errors. If the system does not go through regular cleansing process, it may produce inefficient results, leading to less competent work and complicated problems.

The general framework for data cleaning (after Maletic & Marcus 2000) is:

- *"Define and determine error types"*
- *"Search and identify error instances"*
- *"Correct the errors"*
- *"Document error instances and error types"*
- *"Modify data entry procedures to reduce future errors"*

It is a quite new field and the process is computationally expensive. The new faster computers provide data cleansing process within time framework on large amounts of data. However, there are numerous problems involve in data cleansing process i.e. missing data, determining record availability, erroneous data. Apart from the work done for data sets and there is not a single common definition of data cleansing, different definitions are being used during different types of processes. Data cleansing includes three major areas:

- Data warehouse
- Knowledge delivery in database

- Total quality management

a) The Need for Data Cleaning

Data cleansing is required to improve the data quality bydetecting and eliminating errors in the data. Identifying errors in data, correcting valuable records and eliminating bad records can be a tedious process but it is essential for effective decision making.

b) Data Quality

According to [14], High quality data needs to pass a set of quality criteria. Those include:

- ***Accuracy:*** *"The criteria of integrity, consistency and density"*
- ***Integrity:*** *'The criteria of completeness and validity"*
- ***Completeness:*** *"Achieved by correcting data containing anomalies"*
- ***Validity:*** *"Approximated by the amount of data satisfying integrity constraints"*
- ***Consistency:*** *"Concerns contradictions and syntactical anomalies"*
- ***Uniformity:*** *"Directly related to irregularities"*
- ***Density:*** *"The quotient of missing values in the data and the number of total values ought to be known"*
- ***Uniqueness:*** *"Related to the number of duplicates in the data"*

1.3 Research Domain

Use of a generalized algorithm for each data type present in a data warehouse is an area that needs further research work. Research done in [1] is in early phase as at present no single algorithm solves the error detection problem of different data types in standard automated manner.

Our research domain mainly focuses on investigation of Alliance Rules based on the principle of data mining association rules to test out its feasibility to use in a Data warehouse for data cleansing. Main goal of our research is to investigate all cleansing approaches reported in previous work.

We are not trying to re develop already established cleansing standards for detecting errors in the data setsbut try to provide a generalized algorithm for numeric data type present in the data warehouse.

1.4Proposed Approach

In order to provide a solution for all data types present in the data warehouse, all the requirements are thoroughly studied and deep analysis is carried out to developan automated generalized algorithm utilizing the Alliance Rules based on the principle of data mining association rules. Main focus is to target the accuracy of the existing algorithm by providing a generalized mathematical model with improved accuracy.

1.5Thesis Outline

The organization of this thesis has the following structure:

Chapter 2: Literature Survey is given to survey existing and previous Data cleansing approaches used in a data warehouse. Important findings are extracted and reassembled in the form of Finding Tableat the end ofthis chapter.

Chapter 3: Requirement Analysis gives an investigation of previousand existingdata cleansing requirements.

Chapter 4: Proposed Model explains in detail the proposedmodel utilizing the Alliance rules.

Chapter 5: Conclusion And Resultsprovides mathematical results, summarizes the contributions and suggest issues or directions for future research.

Chapter 2

LITERATURE SURVEY

2. LITERATURE SURVEY

2.1 Introduction

A Data warehouse is used for combining multiple data sources into a single data source for end user access. End users can use this information for performing queries and reporting of Data warehouse information. Data warehouse is used for analyzing and describing information for strategic decision making and other end user applications.

Metadata has to be referred for any kind of access or operation in the data warehouse. That is, whenever a user poses a query the DBMS or DW tool has to refer the metadata for verifying the correctness of query, for the existence of required data, to find the location of data in the repository, to set an efficient mechanism to access data, to know the format of data to read and present it in a proper format.

The main purpose of data cleansing is to cleanse the data as well as to bring consistency to different sets of data merged from multiple sources.Datacleansing can take place within a single set of records, or between multiple sets of recordsthat needs to be amalgamated.

In this chapter literature survey is being done. Main target is to provide a solid base for what we are trying to examine. In Section 2.2 research related todata warehouse cleansing is given. Section 2.3 organizesconcept matrix in tabular form. Section 2.4 provides the problems find in literature survey. Section 2.5 organizes research findings. Section 2.6 gives limitations of different Data warehouse cleansing strategies investigated in previous sections. Section2.7reviewswhole literature survey process.

2.2 Related Research

Within the data warehousing field, data cleansing becomes mandatory when multiple databases are amalgamated.Problem occurs when records belonging to the same entity are represented in different formats in the different data sets. Problems mentioned in previous papers are discussed here in detail.

2.2.1 Merge/Purge Problem

In [6][7] *merge/purge problem is discussed.* Duplicate records exist when different databases are merged. The main aim is to detect and to remove those duplicates. According to the merge/purge problem, it is not practical to evaluate each element of one database with the elements from the other; it also overloads the main memory. It collects all the records from the databases in one table and then assigns a key for each record. In next stage, the elements are sorted with respect to their key and lastly a fixed size window is moved through the list. Every time the window is passed over the list,a new element is compared with the ones already in the window. It merged thematching records. The process continues until all elements have been processed. The sorted neighborhood method is used in the merge/purge problem.

2.2.2 Inconsistencies in data

According to [5],"*data cleansing is the process of eliminating the errors and the inconsistencies in data, and solving the object identity problem*". Association rules are utilized for this purpose. "*Association rule mining are helpful in identifying not only interesting patterns for fields such as market basket analysis or census data, but also, by extension to ordinal association rules, patterns that uncover errors in other kind of data sets*" [5].

Although Ordinal association rules are similar to Boolean association rules to some extent, still they are better fitted to the problem of recognizing possible errors in the type of data sets being analyzed for the following reasons:

- Ordinal Association rules are easier and faster to compute than ratio-rules.
- It provides high-quality results in the case of finding (partial) ordering trends, though they are weaker than quantitative association rules.
- Distance-based association rules (over interval data) can be used for this problem, but it is essentiallydifficult to find the right intervals, while the specific domain knowledgelacks. This method isquite expensive.

2.2.3 Decomposing and Reassembling the data

Data cleaning is to decompose the data first and then to reassemble it. The main target is to update a record with good quality data.

Unfortunately, the problem of data cleansing is not clearlyreferred in the mentioned papers. Some papers have stressed on the process management issues while considering the data quality perspective, while others have focused on the definition of data quality.

2.2.4 Data cleansing : preprocessing step

According to [9], "*Datacleansing is regarded as a first step, or a preprocessing step, in the KDD process.*"Paper describes various KDD and Data Mining systems performing the data cleansing activities in aparticular fashion.

2.2.5 Garbage patterns

[10] Explores the informative patterns to perform one kind of data cleansing by discovering "*garbage patterns*"– meaningless or mislabeled patterns.

2.2.6 Computerized methods

In [9] data cleansing is defined as the process that implements computerized methods of examining the databases first, then finding the missing and incorrect data, and finally eliminating them.

2.2.7 Data quality problems

[12] Finds data quality problems when we switch from the old running database environment to a new database environment. There are some concepts which work correctly in old environment but give some problems in new environment.

2.2.8 Association Rules and Data Cleansing

To generate the automated tool for detecting the duplicity error in the name filed of data ware house, alliance rules are proposed [1]. Association rules with high confidence and support define a pattern. Records that do not follow these rules are thought-out as outliers. The meritof association rules is that they can deal with different data types. On the other hand, Boolean association rules do not offer enough quantitative and qualitative information. Ordinal association rules, defined by (Maletic and Marcus, 2000), are used to locate rules that provide information related to ordinal relationships between data elements.The ordinal association rules arerelated to the pattern-based method so it gives in special types of patterns.This method can be extensive to provide the statistical correlations and other kind of associations between different groups of data elements.

2.2.9 Problems in Data Sets

The data warehouse is updated as to store the present and new data, these updates may be intersections and deletions on the data sets. During the updating of data warehouse the probability and possibility of dirty data increases in data warehouse. [3]

2.2.10 Duplicity Error

The research [4, 5] specifically deals with the study of error types and there identification for only the date type data. It focuses on different date formats to cleanse the data. This research does not focus on any other data type. Another main issue is that it could not identify the duplicity error.

2.2.11 Updating Error

[12] Uses the concept of combining the databases to form data warehouse. It assumes that during merging of data from databases, data is not changed. There is a chance of induction of changed or wrong data entry while updating data in data warehouse which has not been dealt.

2.2.12 Semantic Heterogeneity Error

In [1], the errors are detected automatically. The duplicity in the names field of the data warehouse has been remarkably cleansed. Domain independency has been achieved using the concept of integer domain. But this algorithm restricts it only to “name” field.

The process to discover potential errors in data sets using ordinal association rules consisted of the following steps:

1. Apply ordinal rules with a minimum confidence c. This can be prepared with a variation of **apriori** algorithm

2. Recognize data items that bust the rules and can be thought-outas outliers or potential errors.

2.2.13 Data Matching

Number of different algorithms has been used for record matching. Some of them have been discussed in [12]. The following subsections present the algorithms.

a) A recursive record matching Algorithm

Paper [12] describes this domain independent algorithm for matching the strings. Thealgorithm works recursively. If records *X* and *Y*get a matching value equal to 1.0, it means they are alike but if the matching value is 0.0 that means no match. The paper accomplish the Matching process by dividing the record into *i* sub records, where it is thoughtthat *Xi* of *X* corresponds to *Yi* of *Y* with which it has the highest score. According to [12], The score ofthe match between *X* and *Y* equals

$$score\ (X,Y) = 1/|X| \sum_{i=1}^{|X|} \max{}^{|Y|}{}_{j=1}\ \mathrm{score}(X_i, Y_j)$$

This algorithm usesheuristics for acronym handling. The acronymmatching patterns are:

- the acronym is a prefix of its expansion, e.g. ”Univ” abbreviates ”University”
- the acronym combines a prefix and a suffix of its expansion, e.g. ”Dr”matches “Doctor”

- the acronym is an acronym for its expansion, e.g."IIUI" abbreviates "International Islamic University Islamabad"
- the acronym is a concatenation of prefixes from its expansion, e.g. "COBOL" matches "COmmon Business Oriented Language"

The algorithm is easy to implement, so special data structures are not used. The strings can be saved and accessed by theassociated records. First the associated records are determined by splitting the record X and Y. Then the record matching function is recursively called for each correspondingpair from X and Y. These comparisons determine which associated record of Y thatmatches every record in X. The recursion stops when Xor Y cannot be decomposed further.At this position the heuristics are used to match the strings.

b) The Hybrid Algorithm

Hybrid Algorithm is a mixture of Recursive algorithm and Smith-Waterfall Algorithm,sinceboth have some limitations.The hybrid algorithm is an extension of theSmith-Waterman algorithm. Combination of both the algorithms is used todeal without-of-order strings and errors.

c) Sorted Neighborhood Method

In [7], this method is discussed. The Merge/Purge problem is also utilized in this paper. According to themerge/purge problem, it is not practical to evaluate each element of one database with the elements from the other; it also overloads the main memory.

The main objective is to take full advantage of the number of matches while observance the number of falsepositives at least amount.This is accomplished by the congregationof all the records from the databases in one table, and allocating a key for every record. The key should have the appropriate information fromeach attribute. In the next stage, the elements are characterized according to their key and lastlya fixed size window is passed through the list.Every time the window is moved, thenew element is compared with the ones already in the window. It combines the matching records and the process continues to evaluate all the elements present in the data set.

This method is simple to use, the only complication is to construct akey. It is easy to sort the list hence less computing power is required. The key is to pay attentionto the problem especially the semantic errors, otherwise there will be problem. When the key is generated, the data is automatically correctedand spelling errors are removed.

d) Approximate string joins in databases

In [1], this technique is discussed. In this paper, small substrings of length q, called **q−grams**are compared to fulfill the matching process.Firstly, break the strings into q −grams of length q.
E.g. q − grams for ISLAMABAD are:

(1,##I) , (2,#IS) , (3, ISL) , (4, SLA) , (5, LAM) ,
(6, AMA),(7, MAB) , (8, ABA) , (9, BAD) , (10, AD$) , (11,D$$)

The strings are compared by checking how many q − grams are matched. Another important thing is the order of the q−grams,given that that two strings match if the order is the same.

e) Brute force method

The Brute force method accepts the metadata and loop through the elements present in the database to find the best match. If no match is found, it means data can be added or updated. The evaluationbetween the datawill be prolonged since every dataelement will haveto be compared with multiple entries in the database.

2.2.14 Personal name matching

[22]Provide the scenarios for personal name matching. When different databases are merged, there may exist number of different entries with the same name for different persons, or they could be the case of having number of entries in different formats for alike person. In this paper, test collections are used for the measurement of personal name matching algorithms. This paper provides two test methods based on real-world bibliographic data.

2.2.15 Data Quality Tools

[23] Survey the commercial and research data quality tools. The authors categorize tools according to a set of common functionalities. These functionalities include data types, data

sources supported at the targetand the different interfaces provided. Then the tools are grouped in six classes depending on the feature of data quality. Finally anassociation has been proposed between the data problems and the tools identified.

a) Generic Functionalities:

José Barateiro et al describe the generic functionalities as follows [23]:

- Data sources: Data sources may consist of relational databases, flat files, XML files, spreadsheets, legacy systems, application packages, and web based source
- Extraction Capabilities: to schedule extracts by time, set of rules, to select and to amalgamate records from multiple sources
- Loading Capabilities: to send data into various target systems in parallel.
- Incremental updates: to incrementally updates data targets
- Interface: to define data quality processes modeled as workflow
- Metadata repository: to store data schemasand information about the design of the data
- Performance Techniques: to accelerate data cleansing processes and to guarantee scalability
- Versioning: to look into different development versions of the source code.
- Function Library: to expand the function library with new and unique functions

2.3 Concept Matrix

S #	Title	Yr.	Author	Research Method	Technique	Ref
1	*"Alliance Rules for Data warehouse Cleansing"*	2009	Rajiv Arora,P. Pahwa, S.Bansal	Algorithm (association rules)	Devised an algorithm for the detection of errors and dirty data from data extracted from multiple data marts.	[1]
2	*"Managing Very*	2009	G.N.	Mining	Clustering is used to effectively organize	[2]

	Large Databases and Data Warehousing”		Wikramanayake, J.S. Goonetilla	Techniques	information by segmenting a heterogeneous population into a number of more homogeneous clusters	
3	*“Duplicate Record Detection for Database Cleansing”*	2009	Mariam Rehman, VatcharaponEsichaikul	Evaluation of different Data cleansing Algorithms	“*Adaptive duplicate detection algorithm*” is devisedfor the detection of duplicate records. *Recursive algorithm with word base* and *character base* are suitablefor approximate matching of data records.	[15]
4	*“A Unified Framework and Sequential Data Cleaning Approach for aDataWarehouse”*	2008	Jebamalar Tamilselvi, Dr. V. Saravanan	Framework , Solution	Thisresearch paper deals with the structure for data cleaning.A solution is presented by designing a new framework in a sequential order.	[21]
5	*“Automation of metadata updates in a time critical environment”*	2006	Johan Karlsteen	Algorithm	Data merging problem is solved by Brute Force algorithm	[12]
6	*“Problems, Methods, and Challenges in Comprehensive Data Cleansing”*	2003	Heiko Müller, Johann-Christoph Freytag	Survey is presented of data cleansing problems, approaches, and methods	Define a set of quality criteria that the cleansed data has to achieve.	[11]

S #	*Title*	Yr.	Author	Research Method	Technique	Ref
7	*"Data Cleansing: Beyond Integrity Analysis"*	2000	Maletic, Marcus	Framework	The results obtained from experimentation of applying the methods to a real world data set are also given	[5]
8	*"Data Cleaning: Problems and Current Approaches"*	2000	Erhard Rahm, Hong Hai Do	Survey	The paper conduct a survey by classifingthe *source(single or multiple)* and the *location (schema level or instance level)* of the error	[17]
9	*"Matching Algorithms within a Duplicate Detection System"*	2000	Alvaro Monge	Algorithm	To address the duplicate detection problem at the record level	[18]
10	*"Automatically Extracting Structure from Free Text Addresses"*	2000	V. R. Borkar. K. Deshmukh, SunitaSarawagi	Markov Model, Experimentation	The paper presents an automated approach to automatically convert postal addresses, seen as a plain text string, into unique structured elements like "City" and "Street name"	[19]
11	*"ARKTOS: A Tool For Data Cleaning and Transformation in Data Warehouse Environments"*	2000	Vassiliais, Vagena, S.Skiadopoulos, N.Karayannidis, T.Sellis	tool	A tool is presented for the modeling and executing several data cleansing activities	[20]
12	*"An Extensible Framework for Data Cleansing"*	1999	Galhardas H., Florescu D., Shasha D., Simon E.	Framework	Proposed SQL extension for specifying the macro operations & performance optimization like *mixed evaluation*, *neighborhood hash join*, *decision push down* and *short circuit computation*.	[6]

13	*“Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem”*	1998	Hernandez, S.Stolfo	Rule programming Module	An intelligent “*equational theory*” is developedthat identifies alikeelements by a complex,Domain-dependent matching process.	[7]
14	*“The impact of poor data quality on the typical enterprise”*	1998	Thomas C. Redman	Summary	Provides summary of the impacts of the poor data quality	[16]
15	*“Mining Association rules between Sets of Items in Large Databases”*	1993	Agrawal, Imielinski, SawamiArun	Algorithm	The paper provides “*Buffer management*” and “*Novel estimation and pruning techniques*”.	[9]

Table 2-1: Concept Matrix extracted from Literature survey

2.4 Problems found in Literature

S. #	Identified Problem	ProposedSolution	Ref
1	To detect the dirty and faulty data in data warehouse	Proposed Alliance rules based on the concept of mathematical Association rules.	[1]
2	To manage libraries electronically, large library databases has to be created	The data warehouse technology helps in discovering knowledge and improving services	[2]
3	The problem of *“duplicate record detection”* ariseswhen the databases are amalgamated from	Proposed a prototype to prove that the *“adaptive duplicate detection algorithm”* is	[15]

	multiple sources.	the feasible solutionfor the recognition of duplicate records	
4	Different Cleaning techniques provide different techniques. For instance, “*duplicate elimination cleansing tools*” can be used only for data elimination process while similarity “*cleaning tools*” can be used only for field similarity and record similarity.	Proposed and implemented a new framework to coverdifferentmethodologies as a single cleaning tool.	[21]
5	Minimize syntactical or semantic errors	To purge the old database	[12]
6	Problem is to automatically identifying and eliminating the potential errors in the data sets.	Statistical outlier detection, pattern matching, clustering, and data mining techniques.	[5]
7	Classification of data quality problems that can be debugged by data cleaning techniques	Provides asummary of the main approaches that provides solution to the defined problem	[17]
8	Errors occurred during Data entry, inappropriate abbreviations, or differences in the detailed schemas of records from multipledatabases	An algorithm is presented for minimizing the number of comparisons inpreviously presented window-based merge-purge algorithms.	[18]
9	Address Elementatization. For identical person, there can be different address records stored in different databases in different formats. Problem is tomerge all these addresses in a typical format where all the different fields are recognized and duplicates removed	“*Hidden Markov Modeling*” is a powerful statistical machine learning technique that is able to grip new data vigorously, is competent in computations and uncomplicated for humans to interpret. This technique is also appropriate for the “*address tagging problem*”.	[19]
10	How to handle the complications and competenceof the transformation and cleaning tasks. Software should be accountablefor the extraction of data from multiple sources, their clean-up, customization and inclusion into a datawarehouse.	A tool is presented, namely “*ARKTOS*”, skilledto do modeling and toperform practical scenarios. It also provides clear primitives for the capturing of common tasks.	[20]
11	1.The lack of universal keys across different	Data Transformation, Duplicate Elimination,	[6]

	databases(“*object identity problem*”) 2.Subsistence of keyboard errors in the data 3.Occurrence of faulty data extracted from multiple sources	Multi-table matching	
12	“*Merge/Purgeproblem*”	“*The basic sorted-neighborhood method*”	[7]
13	The problem of mining association rulesbetween different elements in a large database of customertransactions.	Used“*pruning techniques*” to shun measuring certainitem sets. It guarantees completeness.	[9]

Table 2-2: Identified Problems & Proposed Solution Table extracted from Literature survey

2.5Research Findings

S. #	Findings	Ref
1	Problems and errors related to the ‘name’ field are discussed. ‘Name’ field faces the problem of duplicity due to semantic heterogeneity among various sources of data.	[1]
2	Appropriate research is required to recognize the needs of urban and rural membership.	[2]
3	Takes more time for execution due to exact and approximate match activities.	[15]
4	Proposed framework will fulfill the cleaning requirements for the different kinds of data in the relational databases only. It can also be extended to other types of databases.	[21]
5	Research has to be done on intelligent string matching function	[12]
6	The management of multiple and alternative values has to be done.The organization anddocumentation of performed cleansing operations should be managed.Thespecification and development of a suitable framework supporting the data cleansingprocess should be done.	[11]
7	Different methods to detect errors should be integrated but it is not covered in this paper. It is essential to design and construct high quality and useful software tools to support the data cleansing process.	[5]
8	More efforts are required to develop and implement the language approach that maintain both schema and data transformations.	[17]

9	Domain-independent should be achieved. The algorithms should be able to designate the strengthof the match (or non-match) between the records.	[18]
10	The presented model can be updated byadding support for "*automated feature selection*", and "*incorporating a partial database*".Future work includes correcting thespelling mistakes in the data and automatically creating the segments of a largeheterogeneous dataset into smaller subsets to be checked on separate Hidden Markov Models	[19]
11	1. An impact analyzer should be designed to show how changes in the definition of a table or an activity influence other tables or activities in the data warehouse; 2. A metadata repository should be connected, to utilize its enhanced query facilities; 3. An optimizer should be created to achieveimproved efficiency	[20]
12	Research has to be done on significant functions and *n-gram.*Experimentation of the given framework is required	[6]
13	There is a need to seekin large windowsto achieve high accuracy	[7]
14	The database capability to classifyqueriesshould be enhancement	[9]

Table 2-3: Findings Table extracted from Literature survey

2.6 Limitations

2.6.1 "Alliance Rules for Data warehouse Cleansing"

Alliance rules for Data warehouse cleansing are not polishedthoroughly to satisfy the requirements of cleansing of all fields. Algorithm presented in the paper is restricted to only "name" field of string data type.Alliance rules show its applicability by considering a scenario but no implementation is provided.

2.6.2 "Duplicate Record Detection for Database Cleansing"

It is seen that the "*Adaptive duplicate detection algorithm*" produce better results in terms of accuracy. But if we consider efficiency, it takes extra time for execution due to exact and approximate match activities.

2.6.3 *"A Unified Framework and Sequential Data Cleaning Approach for a Data Warehouse"*

Proposed framework will fulfill the cleaning requirements for different kinds of data in the relational databases only. It can also be extended to other types of databases.Existing data cleaning techniques has covered some data cleaning problems, like data elimination and similarity problems. *"Duplicate elimination cleaning tools"* are appropriate for data elimination process and similarity cleaning problems are compatible for field similarity and record similarity. These approaches worked independently but this paper has proposed a new framework to encompass these approaches as a single data cleaning tool.

2.6.4 *"Automation of metadata updates in a time critical environment"*

Research has to be done on intelligent string matching function. The proposed approach is suitable for only one department. It can be extended to other departments as well.

2.6.5 *"Problems, Methods, and Challenges in Comprehensive Data Cleansing"*

Although this paper mostly focuses on the conversion and the eradication of duplicates, still the user has to be involved in the implementation details for the cleansing operation. Some data cleansing problems and challenges are not supported.

2.6.6 *"Data Cleansing: Beyond Integrity Analysis"*

Author gives an idea of an integrated approach for automatic error detection. The investigation of groups of correlated fields (e.g. based on statistical correlation) will be devised. Incorporation of various methods to address error detection is not covered in this paper. To support the data cleansing process, it is essential to design and construct high quality and useful software tools.

2.6.7 *"Data Cleaning: Problems and Current Approaches"*

Although the paper has covered *"schema- and instance-related data transformations"*in an integrated way, but programming has to be done manually. Besides that, the interoperability is also limited(in terms of proprietary APIs and metadata representations).To support bothschema and data transformations, special efforts are required on the design and implementation of the best language approach.

2.6.8 *"Matching Algorithms within a Duplicate Detection System"*

In this paper, the problem of *"approximate duplicate detection"*has been explored by the author. It provides that during amalgamation of data from discrete sources, common attributes must be obtained first. The paper has presented different *"record matching algorithms"* to determine the equivalence of records from source systems.Domain-independent should be achieved and the algorithms must point out the strengthof the match (or non-match) between the records.

2.6.9 *"Automatically Extracting Structure from Free Text Addresses"*

Support should be added for *"automated feature selection"* and *"incorporating a partial database"*. Further enhancements requires correcting the spelling mistakes in the data and automatically splitting a largeheterogeneous dataset into smaller subsets to be trained on separate Hidden Markov Modeling [HMM]

2.6.10*"ARKTOS: A Tool for Data Cleaning and Transformation in Data Warehouse Environments"*

More functionality can be added to ARKTOS, to provide user with richer transformation primitives. Some research issues are

1. An impact analyzer should be designed to show how changes in the definition of a table or an activity influence other tables or activities in the data warehouse;

2. A metadata repository should be connected, to utilize its enhanced query facilities;

3. An optimizer should be created to achieveimproved efficiency

2.6.11 *"An Extensible Framework for Data Cleansing"*

Research has to be done on significant functions and *n-gram.* Experimentation of the given framework is required

2.6.12 *"Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem"*

The "*sorted-neighborhood method*" is expensiveas it is used to search in large windows for high accuracy. The results reported in this paper form the basis of a "*DataBlade*Module".

2.6.13 *"The impact of poor data quality on the typical enterprise"*

Issues are discussed in this paper. While implementing data quality programs, practitioner must overcome these issues. The important issues are the dissatisfaction of the customer, amplifiedprice, unproductive decision making, the reduced ability to make and execute strategy, organizational suspect,difficulties in aligning the enterprise,and the ownershipissues.

2.6.14 *"Mining Association rules between Sets of Items in Large Databases"*

In this paper, the presented algorithm is tested by applying it on sales data captured from a retailing company. The efficiency of the algorithm is tested this way. The algorithm produces efficient and accurate results on the given data set.The procedure can further be improved due to the enhancement of the database capability with classification queries.

2.7 Summary

As we have seen the "Literature Survey", we have discussed different areas to highlight the problem. We have discussed the two main issue of Data warehouse i.e. Cleansing is mandatory for the effective performance of data warehouse and it is a complex task to merge data from multiple source systems into a Data ware house while maintaining the data consistency, coherency, efficiency, atomicity and reliability.

Rajiv Arora et al [1] discovered a problem related to the cleansing phase of the data warehouse. Cleansing of data is a procedure that determines and detects the *"unwanted", "corrupt", "inconsistent and faulty data"* and corrects it to improve the quality of data. The presence of faulty data leads to severe problems. Since the core purpose of designing a data warehouse is to do trend analysis, decision making, strategic planning and knowledge discovery for the enterprise, the presence of faulty data leads to inferior decision making and higher operational cost. The cleansing of data is the possible remedy to this problem. An automation of cleansing process will assist precise and accurate data cleansing operation. In the base paper, the authors provide an algorithm for an automated tool to detect and identify the corrupt data in the data sets of the data warehouse.Problems and errors related to the 'name' field are discussed. 'Name' field faces the problem of duplicity due to semantic heterogeneity among various sources of data.

Our proposed model differs from their approach. Our proposed mathematical model will utilize the Alliance rules based on the mathematical association rules, and wrap the uncovered errors in the fields of data sets. The reminder of this thesis presents a new methodology for the cleansing of a data warehouse.

Chapter 3
REQUIREMENT ANALYSIS

3. REQUIREMENT ANALYSIS

3.1 Introduction

In order to carry out requirement analysis, a deep understanding of data cleansing process is a requisite. The reason behind investing the requirements for data cleansing is to ensure that the data enter in data warehouse is free of errors, and satisfying the basicrequirements of data i.e.Accuracy, Integrity, Completeness, Validity, Consistency, Uniformity,Densityand Uniqueness.

Data Cleansing is an important issue. Data cleansing becomes obligatory whenever multiple operational data sources have merged in a single system. Databases distributed across an enterprise, and search engines are used to construct a data warehouse.The explosion of data from multiple sources makes the problem more complex hence highlights the relevance of data cleansing. Research is performed to provide suitable platforms for cleansing purposes, to decrease human dependencies and to develop efficient algorithms.

The core purpose of collecting data from multiple organizations is to make effective decision making in most of the times. Anomalies may exist in the data gathered from several sources. The situation becomes critical when the data has to be merged as a single database. The integrated database inherits the problems of the source systems. [15] The only solution is to clean the data. In our research work, emphasis is on one of the major issue of data cleansing i.e. “Duplicity”.

Due tothe awareness of importance of data, organizations are more emphasizing on the quality of data. The better the quality of data is, the high the quality of decision made. Therefore companies are putting their efforts to achieve utmostadvantages from their data to make better quality decisions. Improved quality of data can produce better results hence providing financial returns. Every organization needs accurate data. To ensurethe data quality, formal data quality assurance routines and procedures are required. The probability of containing dirty data increases, when data from heterogeneous sources is included. Therefore, Cleansing especially the removal of duplicates proves to be a major cleansing task.

Once the data is merged from multiple data marts into a single data ware house, it can be tested against the following criteria:

- *Completeness and Correctness:*

  Concepts exist in any component or data mart must be present uniquely in the resulting data warehouse after merging data from source systems. Therefore, the data warehouse can be thought out as a blend of all the attributes present in different data marts.
- *Minimality:*

  Concept must be represented once in the data warehouse, if it is presentedin more than one data marts.
- *Understandable:*

  The data should not be complicated to understand for the end user. The one that is the most understandable should be chosen among the several representations of the results of integration.

3.2 Requirement Analysis with critical Problem Scenarios

3.2.1 "*Alliance Rules for Data warehouse Cleansing*"

Algorithm presented in the paper is restricted to only "name" field of string data type.The paper provides its applicability on a scenario but no implementation is provided. Although the duplicity in the "names" field of the data warehouse has been cleansed and worked out, still there is a need to cover all other data fields. Domain independency has been achieved using the concept of integer domain.

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Completeness	Requirement 3 Automation	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
"Alliance Rules for Data warehouse Cleansing" [1]	Satisfied	Not Satisfied	Partially Satisfied	Satisfied	Satisfied	Partially Satisfied

3.2.2 *"Duplicate Record Detection for Database Cleansing"*

It is observed seen that the *"Adaptive duplicate detection algorithm"* produce better results in terms of accuracy. But in terms of efficiency, it takes extra time for execution due to exact and approximate match activities.

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Automation	Requirement 3 Efficiency	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
"Duplicate Record Detection for Database Cleansing"[15]	Satisfied	Satisfied	Not Satisfied	Satisfied	Not Satisfied	Partially Satisfied

3.2.3 *"A Unified Framework and Sequential Data Cleaning Approach for a Data Warehouse"*

The framework presented in this paperoffers the cleaning requirements for different kinds of data in the relational databases only.Services for data cleansing like selection of attribute, token formation,clustering algorithm selection, similarity function selection,elimination function selection and merge functions are provided.

It can also be extended to other types of databases. Existing data cleaning techniques has covered some data cleaning problems, like for data elimination process,duplicate elimination cleaning tools are appropriate andfor field similarity and record similarity, similarity cleaning tools are suitable. These approaches worked independently but this paper has proposed a new framework to encompass these techniques as a single tool for data cleaning.

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Automation	Requirement 3 Efficiency	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
"A Unified Framework and Sequential Data Cleaning Approach for a Data Warehouse"[21]	Satisfied	Not Satisfied	Not Satisfied	PartiallySatisfied	Not Satisfied	Satisfied

3.2.4 *"Automation of metadata updates in a time critical environment"*

Research has to be done on intelligent string matching function. The proposed approach is suitable for only one department. It can be extended to other departments as well.

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Automation	Requirement 3 Efficiency	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
" Automation of metadata updates in a time critical environment"[12]	Satisfied	Not Satisfied	Satisfied	Partially Satisfied	Not Satisfied	Not Satisfied

3.2.5 *"Problems, Methods, and Challenges in Comprehensive Data Cleansing"*

It has been concluded in the said paper that in the field of data cleansing, there are number of open problems and obstacles. It has mentioned the organization of several and different values, The administration and record keeping of performed cleansing operations, the cleansing extraction, and theobligation and development of an appropriate framework supporting the data cleansingprocess.

Although this paper mostly focuses on the resolution and the removal of duplicates, but the user has to participate in the implementation details of the cleansing procedure. Some data cleansing problems and challenges are not supported.

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Automation	Requirement 3 Efficiency	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
"Problems, Methods, and Challenges in Comprehensive Data Cleansing"[11]	Not Satisfied	Not Satisfied	Not Satisfied	Satisfied	Not Satisfied	Not Satisfied

3.2.6 *"Data Cleansing: Beyond Integrity Analysis"*

This paper has proposed the concept of an integrated approach for automatic error detection. The analysis of groups of correlated fields (e.g. based on statistical correlation) will be devised. The said paper does not cover the combination of various methods to deal with error detection.Tosupport the data cleansing process,it is obligatory to design and construct high quality and useful software tools.

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Automation	Requirement 3 Efficiency	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
"Data Cleansing: Beyond Integrity Analysis" [5]	PartiallySatisfied	Not Satisfied	Not Satisfied	Satisfied	Not Satisfied	Not Satisfied

3.2.7 *"Data Cleaning: Problems and Current Approaches"*

The author has provided an outline of commercial data cleaning tools. He has emphasizedon *"schema-and instance-related data transformations"* in an integrated way.More efforts are required to design and implement the best language approach for supporting bothschema and data transformations.

According to the research paper, data cleansing is also required for query processing on dissimilar data sources, e.g., in web based systems.

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Automation	Requirement 3 Efficiency	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
"Data Cleaning: Problems and Current Approaches" [17]	Not Satisfied	Not Satisfied	Satisfied	Satisfied	Not Satisfied	Not Satisfied

3.2.8 *"Matching Algorithms within a Duplicate Detection System"*

The paper has covered an important area of research i.e. the integration of information sources. To integrate data from multiple sources, the information common in these sources must be identified.The paper has presented number of record matching algorithms to find out the equivalence of records from several source systems. Domain-independent should be achieved and the algorithmsmust point out the potencyof the match (or non-match) between the records.

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Automation	Requirement 3 Efficiency	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
"Matching Algorithms within a Duplicate Detection System" [18]	Not Satisfied	Not Satisfied	Not Satisfied	Not Satisfied	Not Satisfied	Satisfied

3.2.9 *"Automatically Extracting Structure from Free Text Addresses"*

Support should be added for the automaticselection of the feature, incorporating a database and reducing amount of training data. Correction of spelling mistakes in the dataand automatically Splitting a heterogeneous dataset into smaller subsets to be trained on separate Hidden Markov Model [HMM]

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Automation	Requirement 3 Efficiency	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
"Automatically Extracting Structure from Free Text Addresses" [19]	Not Satisfied	Not Satisfied	Not Satisfied	Satisfied	Not Satisfied	Not Satisfied

3.2.10 *"ARKTOS: A Tool for Data Cleaning and Transformation in Data Warehouse Environments"*

More functionality can be added to ARKTOS, to provide user with richer transformation primitives. Some research issues are

1. An impact analyzer should be designed to show how changes in the definition of a table or an activity influence other tables or activities in the data warehouse
2. A metadata repository should be connected, to utilize its enhanced query facilities
3. An optimizer should be created to achieve improved efficiency

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Automation	Requirement 3 Efficiency	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
"ARKTOS: A Tool for Data Cleaning and transformation in Data Warehouse Environments" [20]	Not Satisfied	Not Satisfied	Not Satisfied	Satisfied	Not Satisfied	Not Satisfied

3.2.11 *"An Extensible Framework for Data Cleansing"*

Novel optimization techniques are presented in this paper to efficiently execute matching operations. Research has to be done on significant functions and *n-gram*. Experimentation of the given framework is required

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Automation	Requirement 3 Efficiency	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
"An Extensible Framework for Data Cleansing" [6]	Not Satisfied	Not Satisfied	Not Satisfied	Satisfied	Not Satisfied	Not Satisfied

3.2.12 *"Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem"*

The sorted-neighborhood method is expensive due to the need to search in large windows for high accuracy. The results reported in this paper form the basis of a "*DataBlade*Module".

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Automation	Requirement 3 Efficiency	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
"Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem"[7]	Not Satisfied	Not Satisfied	Not Satisfied	Satisfied	Not Satisfied	Not Satisfied

3.2.13 *"The impact of poor data quality on the typical enterprise"*

Issues are discussed in this paper. While implementing data quality programs, practitioner must overcome these issues. The important issues are the dissatisfaction of the customer, amplified price, unproductive decision making, the reduced ability to make and execute strategy, organizational suspect, difficulties in aligning the enterprise, and the ownershipissues.

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Automation	Requirement 3 Efficiency	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
"The impact of poor data quality on the typical enterprise" [16]	Not Satisfied	Not Satisfied	Not Satisfied	Satisfied	Not Satisfied	Not Satisfied

3.2.14 *"Mining Association rules between Sets of Items in Large Databases"*

In this paper, the presented algorithm is tested by applying it on sales data captured from a retailing company. The efficiency of the algorithm is tested this way. The algorithm produces efficient and accurate results on the given data set. The procedure can further be improved due to the enhancement of the database capability with classification queries.

Referenced Paper	Requirement 1 Accuracy	Requirement 2 Automation	Requirement 3 Efficiency	Requirement 4 Consistency	Requirement 5 Domain Independency	Requirement 6 Uniqueness
"Mining Association rules between Sets of Items in Large Databases" [9]	Satisfied	Not Satisfied	Not Satisfied	Satisfied	Not Satisfied	Not Satisfied

3.3 Summary

Requirement analysis is carried out on existing and previous Data Cleansing mechanisms of a data warehouse. Main aim of requirement analysis in our work is to find exact status of different data cleansing techniques. Evaluation table will help us if we want to know quickly what are the cleansing requirements still not satisfied by the existing models.

Some approaches are providing accuracy, while some algorithms are relying on efficiency only.

Some techniques support consistency, while others expect domain independency more important than consistency.

The requirement analyses will emphasis at requirement those need to be in data cleansing tool but most of the existing models are not emphasizing on them.

Chapter 4

PROPOSED MODEL

4. PROPOSED MODEL

4.1 Introduction

In this chapter, proposed model with its details are given. In order to devise algorithm for data warehouse cleansing, an investigative process is adopted. From extensive literature survey, gaps are identified. As a result of literature survey, findings table is developed. The main aim of construction is to support our model with valid issues that arise a s result of finding table. The biggest challenge for this study is to improve the results in terms of performance and accuracy of the algorithm to achieve better results as compared to the previous work done in the domain of duplicate record detection.

Chapter organization is given below.

Section 4.2 describes the detail of the design requirements obtained after Requirement analysis and are named as Accuracy, Completeness, Precision, Automation, Consistency, Domain Independence, and Uniqueness.

In Section 4.3 proposed architecture is given. Section 4.4 gives summary of whole design architecture.

4.2 Design Requirements

Design requirements are explained below in detail in order to increase the understandability of the proposed architecture.

4.2.1 Accuracy

Accuracy is one the main design requirements in designing algorithm for data warehouse cleansing. Accurate data is a fundamental requirement of a good information system. Proposed model should produce accurate results for different data types.

4.2.2 Completeness

Data completeness refers to an indication of whether or not all the data necessary to meet the current and future demands are presented in the data resource.

4.2.3 Automation

There is a need tomake the data cleansing tool fully automatic to avoid human intervention. The manual involvement in the proposed model is highly negligible resulting in high degree of automation and accuracy.

4.2.4 Efficiency

The proposed model should be efficient enough to produce the desired outputs within resource constraints.

4.2.5 Consistency

Proposed Model must transform the data from one consistent state to another consistent state. By eliminating or controlling redundancy, we can reduce the risk of inconsistencies occurring. If a data item is stored only once in a database, any update to its value has to be performed only once and the new value is available immediately to all users. If the data item is stored more than once and the system is aware of this, the system can ensure that all copies of the item are kept consistent.

4.2.6 Domain independence

The proposed model should be domain independent. Domain independency has been achieved using the concept of integer domain which even adds on to the memory saving capability of the algorithm.

4.2.7 Integrity

The proposed model should not violate the integrity of data. Integrity is usually expressed in terms of constraints, which are consistency rules that the database is not permitted to violate. Constraints may apply to data items within a single record or they may apply to relationships between records.

4.3 Proposed Model

We propose the mathematical model for the cleansing of data warehouse. The proposed modeldiscusses the problems in [1] and eliminates them by presenting a generalized algorithm for each data type present in the data warehouse.

4.3.1 Association Rules

The term "Association rules"was first introduced by Agrawal et. al. In this type of analysis, the data set is defined as the basket data $D=\{d_1,d_2,.....d_n\}$, where each basket d_i is subset of I, is the collection of items, where $I = \{i_1,i_2,........, i_k\}$is a set of k elements. An association rule is defined as follows.

i_1 --> $i_2$is an Associationrule if:

1. i_1 and i_2 occur together in at least s% of the n baskets, where s is the support of the rule
2. Of all baskets containing i_1, at least c% contains i_2, where c is the confidence of the rule.

This type of Association rules are referred as Classical or Boolean association rules in literature.

4.3.2 Ordinal Association Rules

According to Marcus and Maletic, the generalizations of the association rule concept can be used for the identification of possible erroneous data items with certain modifications. These considerations lead to a new extension of the association rule called ordinal association rules.

Let $D=\{d_1,d_2,.....d_n\}$ a data set, where each record $d_1$is a subset of I is a collectionof items, and $I = \{i_1,i_2,........, i_k\}$is a set of k items. Each item i_1 has the same numerical domain.

Then i_1,i_2->$i_1$op i_2 where op = $\{\leq , = , \geq\}$, is an ordinal association rule if:

1. i_1 and i_2 occur together in at least s% of the n records, where s is the support of the rule

2. In c% of the records where i_1 and i_2 occur together i_1 op i_2 is true, and where c is the confidence of the rule.

The process to identify potential errors in data sets using ordinal association rules is composed of the following steps:

1. Prepare the data
2. Get ordinal rules with a minimum confidence *c*
3. Identify data items that broke the rules and can be considered outliers(potential errors)

4.3.3 Data Alliance Rules

Based on the concept of mathematical association rules, Alliance rules [1] are defined as

Let A = $\{a_1, a_{2\ldots}\ a_n\}$ be a set of fields, each field is a subset of a data mart. There could be m Data marts.

Let S = $\{s_1, s_2 \ldots s_n\}$ be a set of scores, each score is a subset of fields.

Let D = $\{=, \neq, \leq, \geq\}$ be a set of relational operators, Set of scores belongs to D.

Then,

S_1, S_2=> S_1 op S_2, where op is a subset of D defines the Alliance rules as

1. S_1 and S_2 happen together in at least p% of the n records, where p is the support of the rule.
2. In C% of the rule, F => S_1 opS_2 op T where S_1,S_2 are reference table scoreand data warehouse table score respectively
3. Confidence of threshold value is 50%. If 50% letters match, then the letter is considered for further processing, otherwise it is considered as outlier value.
4. For q-gram matching, rule 2 is utilized. The q-grams[2] are the substrings of length q of a given string.

4.3.3 Proposed Algorithm

Different set of data marts and data sets are considered to form a complete data warehouse. The defined alliance rules are defined to identify and discover errors in the data warehouse. The main aim is to cleanse the data from duplicity errors. Duplicity is dangerous as wrong results will be produced due to faulty data.

An algorithm is produced here to remove the duplicity from the data sets in a data warehouse. The algorithm converts all the different data types into numbers. This way domain independency and advantages of memory concerns can be achieved. This also helps out in doing the mathematical analysis of the values of the data obtained by doing the calculations.

The Base paper only deals with the elimination of duplicity in the name field (string data type).It is extended to cover all the different data types.

The algorithm for the detection of different data types present in different data marts consist of the following three main stages:

1. Analysis phase
2. Implementation phase
3. Experimentation phase

In the analysis phase, data consisted of different data types are converted into a numerical value which is stored in a file for further processing. Alliance rules are utilized in theImplementation phase. The duplicity is detected and reported in the experimentation phase.

A. Analysis Phase

Twodifferent data mart sets, say DMS1 and DMS2, are considered in such a way that data record from DMS1 has to checked for duplicity in all the different data fields of DMS2. In the implementation phase, alliance rules are applied on the specific filed value to detect the duplicity and it will be reported in the last phase.

The main target of this phase is to get the numerical values of all the different data type values. The converted numerical values are stored in a file called *Attribute* and the values are called the *attributes*of the data.

Conversion can be done by first finding out the number of words in the specific field, if we are considering the name field, then for instance

Ahmed F. Khan

has 3 words. The number of attributes calculated for N words is N+1, as N+1 attributes comprises of N attributes for each word in the name and N+1th attribute for the initials of the name.

Similarly, if we are considering the address field, then for instance

H. # 35, St. # 7, F-10/2, Islamabad

has4 words. The number of attributes calculated for 4 words is 5, as N+1 attributes comprises of N attributes for each word in the address separated by commas and N+1th attribute for the initials of the address.

The values are converted into numbers using relation:

$$[(index)^{set\ value} * face\ value]mod\ m$$

Where *index* is defined as a set of 71 character (26 alphabets ,10 digits, 4 arithmetic operators (+, -, *, /), 6 relational operators(=,≠,<, >, ≤, ≥), 25 special characters (“.” , “,” , “#”, “@”, “$”, “%”, “^”, “&”, “_”, “/”, “?”, “ ; ”, “ :” , “~”, “\”, “|”, “`”,”(“, “)”, “{“, “}”, “[“, “]”, “!”, “’””)). The letters are considered case-insensitive.

The *face value* of each character can be obtained form table 4.1.The *set value* of characters are marked from right to left starting with zero. “*m*” is any large prime number.

Face Value	Characters
0	A
1	B
2	C
3	D
4	E
5	F

6	G
7	H
8	I
9	J
10	K
11	L
12	M
13	N
14	O
15	P
16	Q
17	R
18	S
19	T
20	U
21	V
22	W
23	X
24	Y
25	Z
26	0
27	1
28	2
29	3
30	4
31	5
32	6
33	7
34	8
35	9

36	+
37	-
38	*
39	/
40	=
41	≠
42	<
43	>
44	≤
45	≥
46	.
47	,
48	#
49	@
50	$
51	%
52	^
53	&
54	_
55	/
56	?
57	;
58	:
59	~
60	\
61	|
62	`
63	(
64	)
65	{

66	}
67	[
68	]
69	!
70	“

Table 4.1 Face values of characters

Using the above mentioned table, face values can be generated. All the calculated values for each data set are stored in tabular form in the Attribute file.

B. Implementation Phase

Alliance rules are utilized to detect the duplicity from the data sets in the data warehouse.

1. Take the primary key from DMS1.
2. Compare it with the matching value or the primary value of the data in the DMS2
3. Determine the number of words in the primary key. Let it be denoted by N.
4. Calculate N+1 attributes for the selected attribute. The $(N+1)^{th}$values provides the number of initials f the selected attribute.
5. Select the same attribute in DMS2. Calculate the Attribute values using the given formula.
6. Calculate the $(N+1)^{th}$ value of the attribute of DMS2
7. For instance, if we take the example of name, calculate the attributes of the last name in DMS1, and then cluster all those names that have the same attribute values in DMS2. Store in a file *Attribute 1.*
8. Now compare the middle name from DMS1 and cluster it with the matching fields of DMS2. Store it in a file *Attribute 2.*
9. Now compare it with different cases. This attribute matching helps in detecting the faulty errors.

In experimentation phase, different scenarios will be explained.

C. Experimentation Phase

The values obtained from implementation phase can be checked for different scenarios.

Scenario 1:- Best Match

a) *Single Entry Match:*

If a single entry is obtained, then no error is detected hence no duplicity.

b) *Multiple Entry Matches:*

In case of multiple entry matches, match the attributes with other values. If asingle entry is resulted while comparing all available attributes then there is no error. But if this resulted out with multiple attributes, then duplicate records exist in the data set.

c) *No Match:*

If S_1 attributes does not match, then compare it with S_{n+1}, it comes out in two conditions:

a) Same record

b) Different records

If initials match, then check for other records. If other record are same, then it is the same person, hence duplicity detected, otherwise it is a different record.

Scenario 2:- Worst Match

Worst case will happen if none of the match occurs from S_n to S_{n+1}. There could be many reasons behind this flaw. One of the reason is the absence of that entry in the data mart, or the entry exists but with some errors in the records. To remove this error, q-grams can be utilized. Consider q as any length smaller than the length of string. For instance, consider the q-gram with q=3.

q – grams for ISLAMABAD are:

(1,##I) , (2,#IS) , (3, ISL) , (4, SLA) , (5, LAM) ,
(6, AMA),(7, MAB), (8, ABA) , (9, BAD) , (10, AD#) , (11,D##)

Special symbol "_" can be used to representspace in the records. Evaluate q-grams for the remaining records of DMS1 and DMS2. If the values obtained from DMS1 matches with one obtained from DMS2, then compare it with the threshold value, then if it is greater than the threshold value then no error detected, but if it is less than the threshold value, then error is detected.

Process can be explained with the help of figure 4.1

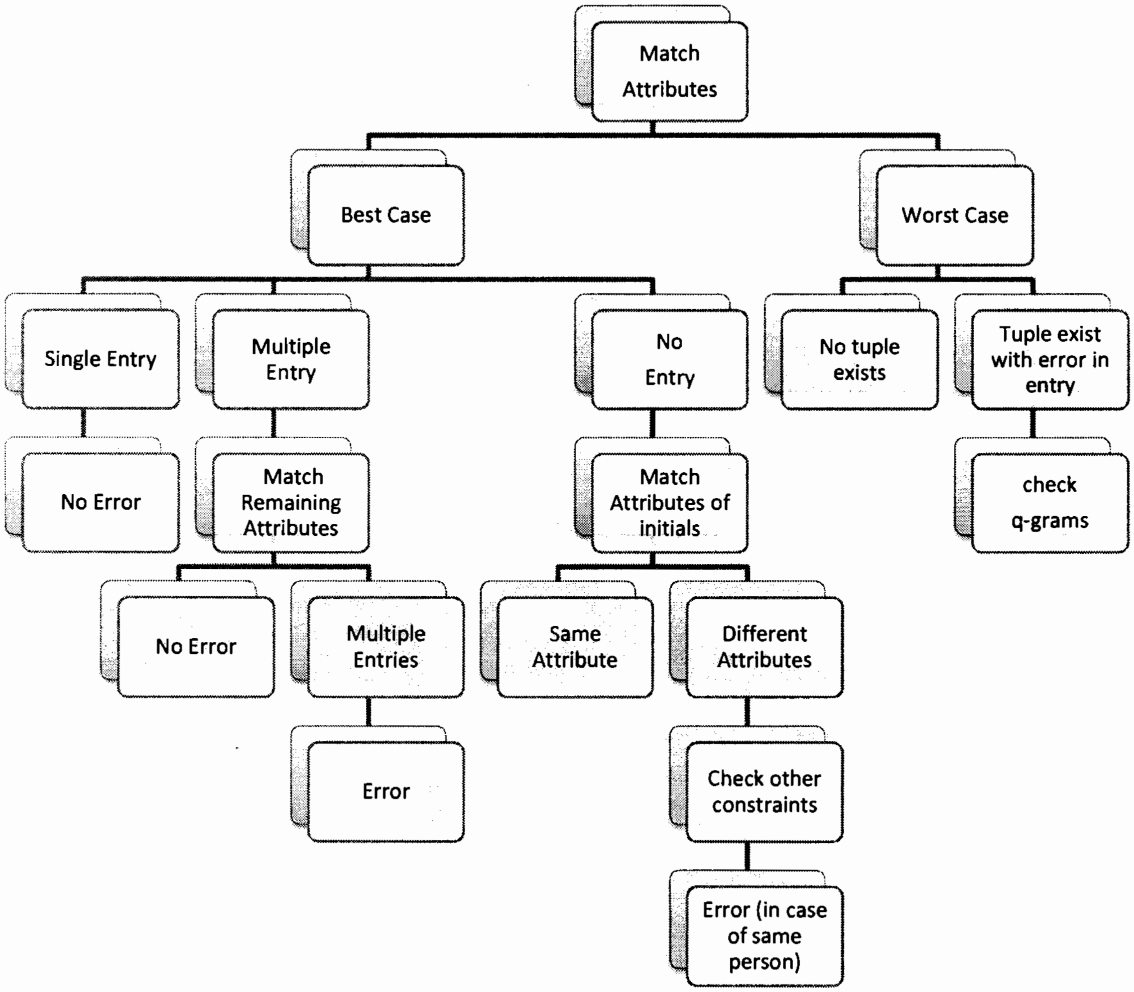


Figure 4-1: Attribute matching scenarios for error detection

4.4 Summary

In this chapter, proposed Algorithm based on the utilization of Alliance rules is given. It is basically a mathematical model. The whole process of model development follows a systematic process of investigation starting from literature survey. Important findings are obtained for providing valid base for our proposed model.

The design requirements obtained from requirement analysis phase were taken as input to our proposed model. Data Cleansing is essential in a data warehouse to enhance the overall quality of data collected from multiple sources. The proposed mathematical model has covered the hitches which were not discovered by the researchers in the previous research work.

Chapter 5
CONCLUSION AND RESULTS

5. CONCLUSION AND RESULTS

5.1 Introduction

In this chapter, results of the mathematical model and some concluding remarks are given. Achievements in form of what we really succeed in doing. Improvements and future work as what still need to be done as there is always some space for improvement in any research work.

5.2 Mathematical Results

Mathematical results are obtained by applying different data types on the algorithm presented in previous chapter. Consider the attributes that have the maximum probability of taking into account as a primary or composite key.

5.2.1 Name Attribute:

The values are converted into numbers using relation:

$$[(index)^{set\ value} * face\ value]\ mod\ m$$

Suppose, we are assuming that Name is the primary key in Data mart set 1(DMS1), then take any instance of name to find out the attribute value by using the above mentioned relation. Take "m" as any largest prime number. The relation considers value as case-insensitive.

For instance,

- Name = M. Adnan Raza
- Let N=3, as there are three words in the name
- Let N+1= MAR
- In Data mart set 2(DMS2), cluster the names that have same value of N. Now calculate N+1 attribute of DMS2.
- Now calculate the attribute for the last name $S_n$of name in DMS1 & $S_n$of each name in DMS2 and cluster all those names in DMS2 that have the same attribute as DMS1.Store the new value in Attribute file 1.

Let m=731

Last name = RAZA

Attribute Value = $[(71)^3 * 17 + (71)^2 * 0 + (71)^1 * 25 + (71)^0 * + 0]$ mod 731

= [6084487 + 0 + 1775 +0] mod 731

= [6086262]mod 731

= 687

- Now from Attribute file 1 of DMS2, pick the values and match it with the middle attribute value of DMS1. The Attribute value of middle name will be

Middle name = ADNAN

Attribute Value= $[(71)^4 * 0 + (71)^3 * 3 + (71)^2 * 13 + (71)^1 * 0 + (71)^0 * + 13]$ mod 731

= [0 + 1073733 + 65533 + 0 + 13] mod 731

= [1139279] mod 731

= 381

- Now match this attribute value obtained form DMS1 with the Attribute file 1 of DMS2. This comparison will further reduce the size of the file. Now place the matching values in a new file called Attribute File 2.
- Now consider the first word of the name attribute and compare the value with the Attribute file 2.

First name = M.

Attribute Value = $[(71)^1 * 12 + (71)^0 * + 46]$ mod 731

= [852 + 46] mod 731

= [898] mod 731

= 167

- Now compare this final value with the Attribute file 2.
- Now check the value for the best case scenario and the worst case scenarios.

- In case, there exist only a single entry in DMS2, then there is no duplicity and hence no error is detected.
- If multiple records are found in DMS2, then data set is infected with duplicate records.
- If the attribute value of first name do not match with Attribute file 2, this can result in two conditions:
    - a) *Same person with different initials*
    - b) *Name may enter in different formats*
- In this scenario, check the other attributes. Check for the composite keys first.
- Worst case occurs if no entry exists. This can be due to two reasons
    - a) *Entry does not exist*
    - b) *Entry exist with some syntax or semantic errors*
- For those entries in which syntax or semantic errors are found, apply the concept of q-grams and evaluate that.
- Assume any length of q. We are assuming q=3.
- Consider the previous example
- q-grams for ADNAN RAZA

((1, ##A), (2, #AD), (3, ADN), (4, DNA), (5, NAN) ,
(6, AN_),(7, N_R), (8,_RA) , (9, RAZ) , (10, AZA) , (11,ZA#), (12, A##))
Special symbol "_" is used to represent the space.

- The q-grams are evaluated for the names in DMS1, and DMS2. If the number of q-grams in DMS1 matches with the number of q-grams in DMS2, then match it with the threshold value. If the threshold value is less than the number of q-grams, then the match is perfect and the error is detected. But, if the DMS1 value is less than threshold, no error is detected.

5.2.2 Email Address Attribute:

Use the given relation:

$$\boldsymbol{[(index)^{set\ value} * face\ value]\ mod\ m}$$

Suppose, we are assuming that Email Address is the primary key in Data mart set 1(DMS1), then take any instance of email address to find out the attribute value by using the above mentioned relation. Take “m” as any largest prime number: The relation considers value as case-insensitive.

For instance,

- mail Address = adnan@yahoo.com
- It will assume special characters as separators, so N=3
- Let N+1= AYC
- In Data mart set 2(DMS2), cluster the email addresses that have same value of N. Now calculate N+1 attribute of DMS2.
- Now calculate the attribute for the last name S_n of email addresses in DMS1 & S_n of each email addresses in DMS2 and cluster all those email addresses in DMS2 that have the same attribute as DMS1. Store the new value in Attribute file 1.

Let m=731

First word = adnan

Attribute Value = $[(71)^4 * 0 + (71)^3 * 3 + (71)^2 * 13 + (71)^1 * 0 + (71)^0 * + 13\ ]\ \text{mod}\ 731$

$= [0 + 1073733 + 65533 + 0 + 13]\ \text{mod}\ 731$

$= [1139279]\ \text{mod}\ 731$

$= 381$

- Now from Attribute file 1 of DMS2, pick the values and match it with the middle attribute value of DMS1 including he special character “@”. The Attribute value of middle word will be

Middle word = @yahoo

Attribute Value = $[(71)^5 * 49 + (71)^4 * 24 + (71)^3 * 0 + (71)^2 * 7 + (71)^1 * 14 + (71)^0 * + 14] \bmod 731$

$= [88407238199 + 609880344 + 0 + 35287 + 994 + 14] \bmod 731$

$= [89017119551] \bmod 731$

$= 256$

- Now match this attribute value obtained form DMS1 with the Attribute file 1 of DMS2. This comparison will further reduce the size of the file. Now place the matching values in a new file called Attribute File 2.
- Now consider the first word of the name attribute and compare the value with the Attribute file 2.

Last word = .com

Attribute Value = $[(71)^3 * 46 + (71)^2 * 2 + (71)^1 * 14 + (71)^0 * + 12] \bmod 731$

$= [16463906 + 10082 + 994 + 12] \bmod 731$

$= [16474994] \bmod 731$

$= 447$

- Now compare this final value with the Attribute file 2.
- Now check the value for the best case scenario and the worst case scenarios.
- In case, there exist only a single entry in DMS2, then there is no duplicity and hence no error is detected.
- If multiple records are found in DMS2, then data set is infected with duplicate records.
- If the attribute value of first name do not match with Attribute file 2, this can result in two conditions:

  c) *Same person with different initials*

  d) *Name may enter in different formats*
- In this scenario, check the other attributes. Check for the composite keys first.
- Worst case occurs if no entry exists. This can be due to two reasons

  c) *Entry does not exist*

d) Entry exist with some syntax or semantic errors

- For those entries in which syntax or semantic errors are found, apply the concept of q-grams and evaluate that.
- Assume any length of q. We are assuming q=3.
- Consider the previous example
- q-grams for adnan@yahoo.com

((1, ##a), (2, #ad), (3, adn), (4, dna) , (5, nan) , (6, an@),(7, n@y) , (8,@ya) ,
(9, yah) , (10, aho) , (11,hoo), (12, oo.), (13,o.c), (14,com), (15,om#), (16,m##))

- The q-grams are evaluated for the names in DMS1, and DMS2. If the number of q-grams in DMS1 matches with the number of q-grams in DMS2, then match it with the threshold value. If the threshold value is less than the number of q-grams, then the match is perfect and the error is detected. But, if the DMS1 value is less than threshold, no error is detected.

5.2.2 Address Attribute:

We are assuming Address as the primary key in Data mart set 1(DMS1), then take any instance of email address to find out the attribute value by using the above mentioned relation. Take "m" as any largest prime number. The relation considers value as case-insensitive.

For instance,

- Address = H. # 567, St. 5, F-6/2, Islamabad
- It will assume special character"," as separators, so N=4
- Let N+1= HSFI
- In Data mart set 2(DMS2), cluster the addresses that have same value of N. Now calculate N+1 attribute of DMS2.

- Now calculate the attribute for the last name S_n of addresses in DMS1 & S_n of each email addresses in DMS2 and cluster all those email addresses in DMS2 that have the same attribute as DMS1. Store the new value in Attribute file 1.

First word = H. # 567

Attribute Value = $[(71)^5 * 7 + (71)^4 * 46 + (71)^3 * 48 + (71)^2 * 31 + (71)^1 * 32 + (71)^0 * + 33]$ mod 731

= [(1804229351*7) +(25411681*46) + (357911 *48) + (5041*31) + 2272+33] mod 731

= [12629605457 + 1168937326 + 17179728 + 156271 + 2305] mod 731

= 13815881087 mod 731

= 93

- Now from Attribute file 1 of DMS2, pick the values and match it with the middle attribute value of DMS1 including he special character "@". The Attribute value of middle word will be

Second word = St. 5

Attribute Value = $[(71)^3 * 18 + (71)^2 * 19 + (71)^1 * 46 + (71)^0 * + 31]$ mod 731

= [(357911 *18) + (5041*19) + 3266 + 31] mod 731

= [6442398+95779+3266+31]mod 731

= [194422973] mod 731

= 303

- Now match this attribute value obtained form DMS1 with the Attribute file 1 of DMS2. This comparison will further reduce the size of the file. Now place the matching values in a new file called Attribute File 2.
- Now consider the first word of the name attribute and compare the value with the Attribute file 2.

Third word = F-6/2

Attribute Value = $[(71)^4 * 5 + (71)^3 * 37 + (71)^2 * 32 + (71)^1 * 39 + (71)^0 * + 28]$ mod 731

= [(25411681*5) + (357911 *37) + (5041*32) + 2769 +28] mod 731

= [127058405+ 13242707 + 161312 + 2769 + 28] mod 731

= [140465221] mod 731

= 357

- Now compare this final value with the Attribute file 2.
- Now check the value for the best case scenario and the worst case scenarios.
- In case, there exist only a single entry in DMS2, then there is no duplicity and hence no error is detected.
- If multiple records are found in DMS2, then data set is infected with duplicate records.
- If the attribute value of first name do not match with Attribute file 2, this can result in two conditions:
    - *e) Same person with different initials*
    - *f) Name may enter in different formats*
- There is a possibility that different persons have the same address. In this case, check for other attributes.

5.3 Achievements

- In our Research work, first we did an informative deep analysis of existing Data Cleansing techniques. Evaluation table is constructed that provide an overview of existing and traditional data cleansing techniques.
- On the careful examination of Evaluation table, we are able to locate the exact situation as how many requirements of Data Cleansing are satisfying and what are those not satisfying.

- An algorithm has been designed utilizing the Alliance rules, based on the mathematical association rules. Our algorithm has the maximum support for the requirements of data cleansing as compared to the existing algorithms or techniques for cleansing purposes.
- The main target of this algorithm is to achieve higher accuracy by considering all the different aspects of faulty data. By doing this, although the complexity is also increased but we can not ignore the higher accuracy level.
- The calculated complexity of the algorithm is $O(n^2)$
- The proposed algorithm deals with different data types present in the data warehouse, while the previous paper deals only with the string data types.
- The proposed algorithm provides an automated and generalized solution for detecting the faulty data. The duplicity in different attributes of the data warehouse has been cleansed and worked out.

5.4 Future Recommendations and Improvements

- A mathematical model for the cleansing of data warehouse is developed in this research work. We plan to implement the proposed model in the next step and test it on working data marts.
- The future work can be scoped out to reduce the complexity of the algorithm. Considering all the different data types results in higher level of accuracy and efficiency while increasing the complexity.

REFERENCES

References:

[1] Rajiv Arora,PayalPahwa, ShubhaBansal, "Alliance Rules for Data Warehouse Cleansing", Department of IT GPMCE, Dehli, India.2009

[2] G.N. Wikramanayake, J.S. Goonetillake "Managing Very Large Databases and Data Warehousing",University of Colombo School of Computing, Sri Lanka,2009.

[3] P.Pooniah, "Data Warehousing Fundamentals- A comprehensive guide for IT professionals", New Dehli, India, 2006.

[4]Marcus, Maletic, Technical ReportCS-00-04, "Utilizing Association Rules for the Identification of Errors in Data", Division of Computer Science, The Department of Mathematical Science, the University of Memphis.

[5]Maletic, J.I. and Marcus, A. Data Cleansing: Beyond Integrity Analysis pp. 200-209 inProceedings of the Conference on Information Quality (IQ2000). Boston: Massachusetts Institute of Technology.

[6]Galhardas, H.; Florescu, D.; Shasha, D.; Simon, E.: "An Extensible Framework for Data Cleaning", Technical Report, Institute National de Recherche en Informatiqueét en Automatique, 1999

[7] Hernandez, M. A.; Stolfo, J. S.: "Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem", Journal of Data Mining and Knowledge Discovery, No. 2, 1998, pp. 9-37

[8] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.: "From Data Mining to Knowledge Discovery", in Advances in Knowledge Discovery and Data Mining, MIT Press/AAAI Press 1996,

[9] Agarwal, Rakesh, Imielinski, Tomasz, Sawami, Arun "Mining Association rules between Sets of Items in Large Databases" Proceedings of ACM SIGMOD International Conference on Management of Data, Washington May 1993, pp.207-216.

[10] EDD. Home page of DataCleanser tool: http://www.npsa.com/edd/

[11] Heiko Müller, Johann-Christoph Freytag , "Problems, Methods, and Challenges inComprehensive Data Cleansing", Humboldt-Universitätzu Berlin zu Berlin, 10099 Berlin, Germany, 2003

[12] J. Karlsteen "Automation of metadata updates in a time critical environment", 2006

[13] Oracle9*i* Data Warehousing Guide Release 2 (9.2) Part Number A96520-01

[14] Wikipedia

[15] Mariam Rehman,VatcharaponEsichaikul, "Duplicate Record Detection for Database Cleansing", Computer Science and Information Management Program, Asian Institute of Technology Pathumthani, 12120, Thailand mariam.rehman@ait.ac.th, vatchara@ait.ac.th , 2009

[16]Thomas C. Redman, "The impact of poor data quality on the typical enterprise", 1998

[17]Erhard Rahm,Hong Hai Do, "Data Cleaning: Problems and Current Approaches", University of Leipzig, Germany, http://dbs.uni-leipzig.de, 2000

[18] Alvaro E. Monge, "Matching Algorithms within a Duplicate Detection System" ,California State University Long Beach Computer Engineering and Computer Science Department, Long Beach, CA, 90840-8302 , 2000

[19] Vinayak R. BorkarKaustubhDeshmukhSunitaSarawagi, "Automatically Extracting Structure from Free Text Addresses", Indian Institute of Technology, Bombay, sunita@it.iitb.ernet.in, 2000

[20]PanosVassiliadis, ZografoulaVagena, SpirosSkiadopoulos, Nikos Karayannidis TimosSellis, "ARKTOS: A Tool For Data Cleaning and Transformation in Data Warehouse Environments", Knowledge and Database Systems Laboratory, Dept. of Electrical and Computer Engineering, National Technical University of Athens, 2000

[21] JebamalarTamilselvi, Dr. V. Saravanan, "A Unified Framework and Sequential Data Cleaning Approach for aData Warehouse", PhD Research Scholar Associate Professor & HOD Department of Computer Application, Karunya University, Coimbatore ,Tamilnadu, INDIA, 2008

[22] Patrick Reuther, Bernd Walter, "Survey on test collections and techniques for personal name matching", Department for Databases and Information Systems (DBIS), University of Trier, 54296 Trier, Germany, Department for Databases and Information Systems (DBIS), University of Trier, 54296 Trier, Germany

[23] José Barateiro, Helena Galhardas, "A survey of data quality tools ", Instituto Superior Técnico (IST), TechnicalUniversity of Lisbon,2005