CHAPTER 1

INTRODUCTION

1. Introduction

1.1 Data Warehouse

In the 1990's as organizations of scale began to need more timely data for their business, they found that traditional information systems technology was simply too cumbersome to provide relevant data efficiently and quickly. Completing reporting requests could take days or weeks using antiquated reporting tools that were designed more or less to 'execute' the business rather than 'run' the business.

From this idea, the data warehouse was born as a place where relevant data could be held for completing strategic reports for management. The key here is the word 'strategic' as most executives were less concerned with the day-to-day operations than they were with a more overall look at the model and business functions.

The first data warehouses were developed in the 1980s. Many of the systems that existed in the 1980s were not powerful enough to store and manage large amounts of data. The term Data Warehouse was coined by Bill Inmon in 1990, who defined data warehouses as “subject-oriented, integrated, time-varying, non-volatile collections of data that is used primarily in organizational decision making” [Inm, 02]. Ralph Kimball provided a much simpler definition of a data warehouse. Kimball stated data warehouse as “copy of transaction data specifically structured for query and analysis”. This
definition provides less insight and depth than Mr. Inmon's, but is no less accurate [KM, 02].

A data warehouse is a process of transforming data into information and making it available to users in a timely enough manner to make a difference [PTS, 07]. The goal of using a data warehouse is to store and monitor information in a way that allows it to be easily analyzed.

Managing a business has never been easy, but it gets harder every year to maintain margins in the face of foreign competition, rising employee costs, and demanding customers. Cost cutting alone is no longer enough. The businesses that thrive in this challenging atmosphere are the ones with current, actionable information that allows them to make better decisions than competitors. The Data Warehousing Institute defines business intelligence as the process, technologies, and tools needed to turn data into information, information into knowledge, and knowledge into plans that drive profitable business action. Business intelligence encompasses data warehousing, business analytic tools, and content/knowledge management.

Finally, a data warehouse is basically a database and having unintentional duplication of records created from the millions of data from other sources can hardly be avoided. In the data warehousing community, the task of finding duplicated records within data warehouse has long been a persistent problem and has become an area of active research. There have been many research undertakings to address the problems of
data duplication caused by duplicate contamination of data. In this research work, a framework is designed to handle duplicate data effectively.

1.2 Data Mining

Clearly, a lot of data is being collected. However, what is being learned from all this data? What knowledge is gained from all this information? “The organizations are drowning in information but starved for knowledge” [Jay, 99]. The problem today is that there are not enough trained human analysts available who are skilled at translating all of this data into knowledge, and thence up the taxonomy tree into wisdom. Data mining is becoming more widespread every day, because it empowers companies to uncover profitable patterns and trends from their existing databases. Companies and institutions have spent millions of dollars to collect megabytes and terabytes of data but are not taking advantage of the valuable and actionable information hidden deep within their data repositories.

Just as with any new information technology, data mining is easy to do badly. A little knowledge is especially dangerous when it comes to applying powerful models based on large data sets. For example, data analyses carried out on un-preprocessed data can lead to erroneous conclusions. If deployed, these errors in analysis can lead to very expensive failures.

Furthermore, the data are growing which may cause the size of database become Terabyte or even Petabyte. Moreover, valuable information might be hidden in this huge
database and also due to the advanced technology recently, it has enabled the possibility of data mining in library. As a result, systematic efforts are needed in order to develop a data warehouse for easing the use of data mining techniques [TNC, 07].

Data Mining is the process of selection, exploitation and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results. It is the process of analyzing a huge amount of data intended to find useful information for decision making. It can be defined as the process of finding correlations or patterns among dozens of fields in large relational databases. However, if there is no useful information hidden in the data, it will obviously be impossible to obtain interesting results. Data in the real world can have duplicate and inconsistent data while integrating data which is collected from multiple data sources. The preprocessing of data is the initial and often crucial step of the data mining process.

To increase the accuracy of the mining result one has to perform data preprocessing because 80% of mining efforts often spend their time on data quality. So, data cleaning is very much important in data warehouse before the mining process.

1.3 Data Quality

Data quality is one part of a larger data management process, which is concerned not only with the quality but the accessibility of data. Quality data is, simply put, data that meets business needs. Quality data does not necessarily mean perfect data. It is essential to set quality expectations, especially in a warehouse setting where deliberate trade-offs must often be made among speed, convenience and accuracy.
Data quality is an important issue in decision environments due to large data volumes and complex, data-intensive decision-tasks that they support. It is estimated that as high as 75% of the effort spent on building a data warehouse can be attributed to backend issues, such as readying the data and transporting it into the data warehouse [Atr, 98]. Data warehousing is emerging as the cornerstone of an organization’s information infrastructure. It is imperative that the issue of data quality be addressed if the data warehouse is to prove beneficial to an organization [Pam, 98].

Data quality has been defined as "fitness for use". The nature of this definition directly implies that the concept of data quality is relative. As a decision support information system, a data warehouse must provide high level quality of data and quality of service. Coherency, freshness, accuracy, accessibility, availability and performance are among the quality features required by the end users of the data warehouse. A decision-maker is usually concerned with the quality of the data stored, their timeliness and the ease of querying them through the OLAP tools [Pan, 00]. Data mining enables to efficiently apply analytic techniques to discover and interpret patterns in the data warehouse. Bad data causes so much trouble in terms of lost time, money, resources and customer satisfaction in data mining process so we need to improve data quality before the mining process. Every organization has some dirty data and “defects” that can’t be prevented. The best place to clean data is in the source system so those defects are corrected in business operations and cannot spread to the data warehouse.
1.4 Data Cleaning

Data cleaning is often studied in association with data warehousing, data mining and database integration. These areas have received much attention from the database research community in recent years. One of the first and most important steps in data warehousing is data cleaning to verify or correct data values to improve quality of the data. Data cleaning is very important in the data warehousing. The quality of the data needs to be improved in the data warehouse before the mining process. The data cleaning is the process of identifying or detecting and removing the duplicate values and errors in the data warehouse. Most of the organizations are in need of quality data. Data Cleaning is also referred to as data scrubbing, the act of detecting and removing and/or correcting a database’s dirty data. The goal of data cleansing is not just to clean up the data in a database but also to bring consistency to different sets of data that have been merged from separate databases. Next, goal of data cleansing is to minimize these errors, and to make the data as useful and meaningful as possible. Furthermore, data warehouses are used for decision making, hence that the correctness of their data is vital to avoid wrong conclusions. For instance, duplicated or missing information will produce incorrect or misleading statistics. Hence data cleaning is very important in data warehouse before the mining process to produce good accurate result.

Data Cleaning is a process of avoiding the unnecessary information in the process of data maintenance. Data Cleaning can be done by using clustering. It is a process of identifying and changing the inconsistencies and inaccuracy. Data cleaning is
a self explanatory term. Most of the data warehouses in the world source data from multiple systems - systems that were created long before data warehousing was well understood, and hence without the vision to consolidate the same in a single repository of information. In such a scenario, the possibilities of the following are there:

i. Missing information for a column from one of the data sources
ii. Inconsistent information among different data sources
iii. Orphan records
iv. Outlier data points
v. Different data types for the same information among various data sources, leading to improper conversion
vi. Data breaching business rules

In order to ensure that the data warehouse is not infected by any of these discrepancies, it is important to cleanse the data using a set of business rules, before it makes its way into the data warehouse.

To solve these problems, many data cleaning techniques have been developed, to clean the data before loading it into the data warehouse. Existing data cleaning techniques are time consuming as a result of which data is not timely available for the mining process. The important concept in data warehouses is called, ‘Data Cleaning and Data Transformation’. As the name suggests, data transformation is a process in which information transferred from specific sources should be cleaned and loaded into a repository. Data cleaning process can either be a manual or automated process. Data
duplication is one of the major research areas in data cleaning. This research work handles duplicate data in order to reduce the time taken for data cleaning and improve the efficiency of the data.

1.5 Duplicate Data

Decision support analysis on data warehouses influences important business decisions; therefore, accuracy of such analysis is crucial. However, data received at the data warehouse from external sources usually contains errors, spelling mistakes, inconsistent conventions, etc. Hence, significant amount of time and money are spent on data cleaning, the task of detecting and correcting errors in data.

The problem of detecting and eliminating duplicated data is one of the major problems in the broad area of data cleaning and data quality. Many times, the same logical real world entity may have multiple representations in the data warehouse. Duplicate elimination is hard because it is caused by several types of errors like typographical errors, and equivalence errors—different (non-unique and nonstandard) representations of the same logical value. For instance, a user may enter “TN, INDIA” for “Tamil Nadu, IND”. Equivalence errors in product tables (“winxp pro” for “windows XP Professional”) are different from those encountered in bibliographic tables (“VLDB” for “very large databases”), etc., It is also important to detect and clean equivalence errors because an equivalence error may result in several duplicate tuples.
The problem of duplicate detection is long since known and several communities have worked on it using different terminology. Pioneering in this area has been the statistics community, which calls the problem “record linkage”, the database community calls it “de-duplication”, “the merge/purge problem” or “eliminating fuzzy duplicates”, other communities like machine learning or natural language processing investigate similar problems such as “identity uncertainty”, “object identification”, “object consolidation”, “co-reference resolution” or “entity resolution” [Pat, 06].

In any case, data duplication happens all the time. In large data warehouses, data duplication is an inevitable phenomenon as millions of data are gathered at very short intervals. Duplicate detection is an integral part of data cleansing. It consists of finding set of duplicate records, correctly identifying the original or more presentative record and removing the rest. A lot of this data is collected from, and entered by humans and this causes noise in the data from typing mistakes, spelling discrepancies, varying schemas, abbreviations and more. Because of this, data cleansing and approximate duplicate detection is now more important than ever [Rob, 07].

The number of factors that influence matter that is typed, and which can lead to inconsistencies in the data is written:

i. A different understanding of how a word or name is spelt (for example U.S. English compared to British English).

ii. A typographical error.

iii. An omission of one or several words in the field value.
iv. Differences in the formatting of data.

v. Incorrect or lacking knowledge of the spelling of a name.

vi. Lack of easy keyboard access to certain letters, like letters diacritical marks.

Recent research efforts have focused on the issue of duplicate elimination in databases. This entails trying to match inexact duplicate records, which are records that refer to the same real-world entity while not being syntactically equivalent. This research work mainly focuses on efficient detection and elimination of duplicate data. The main objective of this research work is to detect exact and inexact duplicates by using the token based blocking methods. This approach is used to improve the efficiency of the data.

1.6 Motivation

Data warehouse contains large amounts of data for data mining to analyze the data for decision making process. Data miners do not simply analyze data, they have to bring the data in a format and state that allows for this analysis. It has been estimated that the actual mining of data only makes up 10% of the time required for the complete knowledge discovery process [Dor, 99]. According to Jiawei, the precedent time consuming step of preprocessing is of essential importance for data mining. It is more than a tedious necessity: The techniques used in the preprocessing step can deeply influence the results of the following step, the actual application of a data mining algorithm [JM, 01]. Hans-peter stated that the role of the impact on the link of data preprocessing
to data mining will gain steadily more interest over the coming years. Preprocessing is one of the fourth future trend and major issues in data mining over the next years [HKP+, 07].

In data warehouse, data is integrated or collected from multiple sources. While integrating data from multiple sources, the amount of the data increases and as well as duplication of data. Data warehouse may have terabyte of data for the mining process. The result of the data mining process will not be accurate because of the data duplication and poor quality of data. The problem of detecting and eliminating duplicated data is one of the major problems in the broad area of data cleaning and data quality. The duplicate elimination problem of detecting multiple tuples, which describe the same real world entity, is an important data cleaning problem. Duplicate elimination is hard because it is caused by several types of errors like typographical errors, and equivalence errors, different (non-unique and nonstandard) representations of the same logical value. Also, it is important to detect and clean equivalence errors because an equivalence error may result in several duplicate tuples. There are many existing methods available for duplicate data detection and elimination. But the speed of the data cleaning process is very slow and the time taken for the cleaning process is high with large amount of data. So, there is a need to reduce time and increase speed of the data cleaning process as well as need to improve the quality of the data.

There are two issues to be considered for duplicate detection: Accuracy and Speed. The measure of accuracy in duplicate detection depends on the number of false
negatives (duplicates that were not classified as such) and false positives (non-duplicates which were classified as duplicates). The algorithm’s speed is mainly affected by the number of records compared, and how costly these comparisons are. Generally CPUs are not able to do duplicate detection on large databases within any reasonable time, so normally the number of record comparison needs to be cut down [Rob, 07].

In this research work, a framework is developed to handle any duplicate data in a data warehouse. The main objective of this research work is to improve data quality and increase speed of the data cleaning process. A high quality, scalable blocking algorithm, similarity computation algorithm and duplicate elimination algorithm are used and evaluated on real datasets from an operational data warehouse to achieve the objective.

1.7 Problem Definition

Investigate and propose a sequential approach for duplicate data detection and elimination with suitable methods for a data warehouse. Duplicate records are detected by comparing records which is integrated from multiple sources using token based approach and clustering/blocking method.

i. An attribute selection algorithm is proposed to identify important and appropriate attribute for the data cleaning process. The selected attribute contains sufficient information to identify the duplication of the record.

ii. A smart token-based and block-token key approach is used to identify exact and inexact duplicates which help in reducing the time taken for the
data cleaning process. It also eliminates the entire long string records with multiple passes for duplicate identification. It also finds what criteria are used to form tokens and which field is having the highest criteria.

iii. To handle duplicate data, smart tokens are used in the Clustering/Blocking method. Using blocking methods, the records are grouped which minimizes the number of record comparisons, and based on the similarities the duplicate records are eliminated.

iv. The rule based duplicate record detection and elimination techniques are used to reduce the probability of false mismatches, with a relatively small increase in the run time.

Software agents are used in the research work to carry out all the above task. The best attribute is selected with certain criteria used by the software agents.

a. Software agents identify, which criteria is used for forming tokens and which field has the highest criteria.

b. Eliminating duplicate data based on certain similarities are being carried out by the software agents.

### 1.8 Contribution

The goal of this research work is to improve the quality of the data and increase speed of the data cleaning process. Quality of the data is improved by eliminating duplicate records and correcting common errors. The speed of the data cleaning process
is increased by introducing token based cleaning. The contributions of the research work are outlined below:

i. A framework is developed to handle duplicate data and common errors in a sequential order. This framework leads to find similarity between records by blocking or clustering records. This research work accommodates a number of parameters to reduce false mismatches and true mismatches to identify exact duplicate record to improve the quality and accuracy of the data.

ii. An attribute selection algorithm is introduced to select subset of attributes from data warehouse to increase speed of the data cleaning process. This attribute selection is efficient to select good and best attribute which is well suited for the cleaning process.

iii. Automatic blocking key generation is developed by selecting best attribute for blocking records using certain criteria. Several blocking key generations are introduced for identifying which blocking key is best to block the records for duplicate elimination.

iv. Token-Based Similarity computation method is used to compare token value instead of comparing long string value. Token formation algorithm is developed to form smart token for the selected attribute field values. This token based method reduces time taken for comparison. This token formation algorithm uses external file for expanding abbreviations and removing common and special characters.
v. A New blocking function is developed to provide efficient and accurate selection of approximately similar object pairs for record linkage and clustering tasks. Previous work on blocking methods has relied on manually constructed blocking key with manually tuned parameters, while this research work automatically constructs blocking key and blocks the records to obtain good result using training data that can be naturally obtained within record linkage and clustering tasks.

vi. Rule-based elimination approach is implemented to eliminate correct and poor quality duplicate value. Duplicate elimination is very important to identify which duplicate to retain and which duplicate is to be removed.

1.9 Outline of the thesis

The thesis is further structured as follows:

**Chapter 2: Review of Literature** describes and assesses related works in the area of duplicate detection on an attribute selection for data cleaning, token formation, blocking records in record linkage, record matching algorithms, duplicate detection and elimination approaches and merging of records from several communities.

**Chapter 3: Framework Design and Dataset Analysis** presents sequential framework for duplicate detection and elimination of this research work and gives summary information on dataset used for result analysis.
Chapter 4: Attribute Selection and Token Formation provides attribute selection algorithm and token formation algorithm for duplicate data detection and elimination. This chapter also presents the results of the experiments that show the effectiveness of the two steps developed in the framework and compares it with other state of the art methods.

Chapter 5: Blocking and Similarity Computation presents blocking or clustering algorithm, blocking key generation with multiple combination and token based similarity computation which is adopted in this research work. At the end of this chapter, the results of the experiments are provided that show the effectiveness of block-token-key, blocking methods and similarity computations.

Chapter 6: Duplicate Elimination and Merge introduces the duplicate detection and elimination approach to identify exact and inexact duplicates and eliminate poor quality duplicate data by retaining only one copy of high quality duplicate data. This chapter also provides the results of the experiments and merging strategies to obtain high quality data.

Chapter 7: Conclusion and Future Work concludes and shows potential results of this research work. This chapter discusses several directions for future research based on the work presented in this thesis.