Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 3, March 2015, pg.360 – 370

RESEARCH ARTICLE

Study of Data Cleaning & Comparison of Data Cleaning Tools

Sapna Devi[1], Dr. Arvind Kalia[2]

[1]Himachal Pradesh University Shimla, India

[2]Himachal Pradesh University Shimla, India

[1] sapna.bahri106@gmail.com; [2] arvkalia@gmail.com

Abstract— Data Cleaning is a major issue. Data mining requires clean, consistent and noise free data. Incorrect or inconsistent data can lead to false conclusion and misdirect investment on both public and private scale. Data comes from various systems and in many different forms. It may be incomplete, yet it is a raw material for data mining. This research paper provides an overview of data cleaning problems, data quality, cleaning approaches and comparison of data cleaning tool.

Keywords: Data cleaning, Data Quality, Data Preprocessing

I. INTRODUCTION

Data cleaning is part of data preprocessing before data mining, prior to process of mining information in a data warehouse, data cleaning is crucial because of garbage in and garbage out principle[1].Data cleaning is also called data cleansing or scrubbing deals with detecting and removing errors and inconsistencies from data in order to improve quality of data. The main objective of data cleaning is to reduce the time and complexity of mining process and increase the quality of datum in data warehouse. Data Quality problems can be single source problems or multisource problems as shown in Figure1.[2]

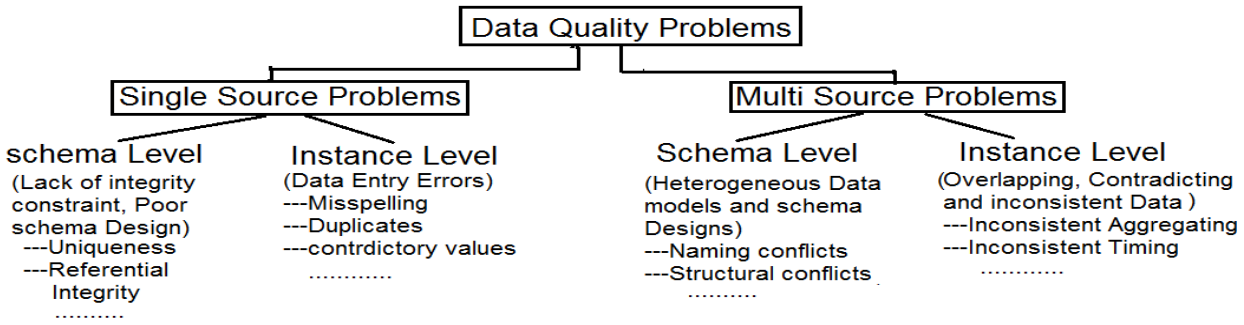

Figure 1. Classification of data quality problem in data sources

A. SINGLE SOURCE PROBLEMS

Schema- related data quality problems occur because of the lack of appropriate model-specific or application specific integrity constraints. E.g. Due to data model limitations or poor schema design, or because only a few integrity constraints were defined to limit the overhead for integrity control. Instance-specific or application specific problems relate to errors and inconsistencies that cannot be prevented at the schema level. Table1 and Table2 shows the single source problem at schema level and single source problems at instance level.

TABLE I

Examples for Single-source problems at schema level (violated integrity constraints)[2]

Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	**Illegal values**	**Bdate=30.13.70**	**Values outside the domain range**
Record	**Violated attribute dependencies**	**Age=22, bdate=12.02.70**	**Age = current year-birth year should hold**
Record type	**Uniqueness violation**	**Emp1=(name="John smith", SSN ="123456") Emp2=(name="Peter Miller",SSN="123456")**	**Uniqueness for SSN(social security number) violated**
Source	**Referential integrity violation**	**Emp=(name="John Smith",deptno=127**	**Referenced department (127) not defined**

TABLE II

Examples for Single-source problems at instance level [2]

Scope/Problem		Dirty data	Reasons/Remarks
Attribute	**Missing values**	**Phone=9999-999999**	**Unavailable values during data entry(dummy values or null)**
	Misspelling	**City="Liipzig"**	**Usually typos , phonetic errors**
	Cryptic values, Abbreviations	**Experience="B": Occupation="DB Prog"**	
	Embedded values	**Name="J. Smith 12.02.70 New York"**	**Multiple values entered in one attribute(eg. In a free-form field)**
	Misfielded values	**City="Germany"**	
Record	**Violated attribute dependencies**	**City="Redmond", zip=77777**	**City and zip code should correspond**
Record type	**Word transpositions**	**Name1="J. smith", name2 =" Miller P."**	**Usually in free –form field**
	Duplicated	**Emp1=(name="John smith "**	**Same employee represented twice due**

	records	**……) Emp2=(name="J. smith"….)**	**to some data entry errors**
	Contradicting records	**Emp1=(name="jhon smith", bdate=12.02.70)**	**The same real world entity is described by different values**
Source	**Wrong References**	**Emp=(name="John smith " , deptno =17)**	**Referenced department(17) is defind but wrong**

B. MULTISOURCE PROBLEMS

When multiple sources are integrated the problems present in single source are aggregated. Each source may contain dirty and inconsistent data and the data in the sources may be represented differently, overlap or contradict. This is because the sources are typically developed, deployed and maintained independently to serve specific needs. This results in a large degree of heterogeneity w.r.t. data management systems, data models, schema designs and the actual data.Table III shows the multisource problems.

TABLE III
Examples of multisource problems at Schema & Instance Level [2]

Customer(Source1)

CID	Name	Street	City	Sex
11	Kristen Smith	2 Hurley PL	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Client(Source 2)

Cno	LastName	FirstName	Gender	Address	Phone/Fax
24	Smith	Christoph	M	23 Harley St. Chicago IL,60633-2394	333-222-6542 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place,South Fork MN, 48503-5998	444-555-6666

Customers(Integrated target with cleaned data)

No.	LName	FName	Gender	Street	City	State	ZIP	Phone	FAX	CID	Cno
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

C. DATA QUALITY

The quality of data can be increased by using data cleaning techniques. To be processable and interpretable in a effective and efficient manner, data has to satisfy a set of quality criteria. Data satisfying those quality criteria is said to be high quality .In general data quality is defined as an aggregated value over a set of quality criteria.

Data must conform to the set of quality criteria.[3] Figure 2 describes the set of criteria of data quality and data must fulfill each criterion to ensure its quality .

Figure 2: Set of criteria of data quality.

Accuracy: It depend upon how the data is collected. The condition of quality of being correct or exact, freedom from errors and defects.

Integrity: It refers to maintaining and assuring the accuracy and consistency of data over its entire life cycle.

Completeness: Completeness of data is the extent to which the expected attributes of data are provided.

Validity: data validity is the correctness and reasonableness of data**.**

Consistency: Data consistency is the process of keeping data uniform through all its location.

Schema Conformance: Conformance of data value to schema design and requirement**.**

Uniformity: state or quality of being uniform overall sameness**.**

Density: Density refers to compactness and Distribution of data.

Uniqueness: State or quality of being unique without any kind of duplication.

D. DATA CLEANING PROCESS

Data Auditing: Auditing the data is done to find the types of anomalies contained within it. Statistical methods are used for auditing. Syntactical anomalies are detected using parsing. The results of auditing the data support the specification of integrity constraints and domain formats. Integrity constraints are depending on the application domain and are specified by domain expert. Each constraint is checked to identify possible violating tuples. For one-time data cleansing only those constraints that are violated within the given data collection has to be further regarded within the cleansing process.

Workflow Specification: Multiple operations over the data are applied for Detection and elimination of common order problems. This is called the data cleansing workflow. It is specified after auditing the data to gain information about the existing anomalies in the data collection at hand. One of the main challenges in data cleansing insists in the specification of a cleansing workflow that is to be applied to the dirty data automatically eliminating all anomalies in the data.

Workflow Execution: The data cleaning workflow is executed after specification and verification of its correctness.

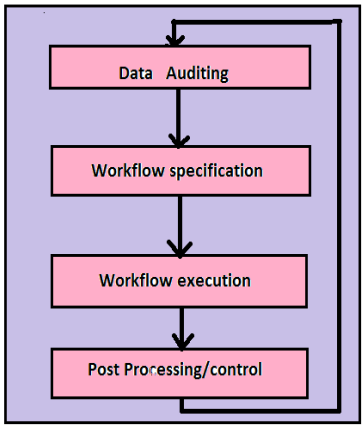

Figure 3: Data cleaning process[4]

Post-Processing/Control: After executing the cleansing workflow the results are checked to again verify the correctness of specified operations. Within the controlling step the tuples that could not be corrected initially are inspected intending to correct them manually.[4]

II. LITERATURE SURVEY

Mong Li Lee [1] in his paper examined the problem of detecting and removing duplicates records. Several different techniques to pre-process the records before sorting them so that potentially matching records will be brought to close neighborhood subsequently.
Rahm, Hong Hai Do [2] in their paper defined the various data cleaning problems and current approaches like Single source problems and Multisource problems and Data quality problems.
Hamid Ibrahim Housien et al[6] in their paper study the data scrubbing algorithms and frameworks in data warehouse .
Nidhi Choudhary[8] in his paper study the various problems and approaches in Data cleaning.
Joseph M. Hellerstein[9] in his paper discuss the quantitative cleaning of large databases**,** and defines the approaches to improve data quality.
Rajashree Y.Patil et al [10] have discussed various data cleaning algorithms for data warehouse.
Heiko Müller et al[4] in their paper discussed the various data cleaning process and compare the data cleaning frameworks.
Mong Li Lee et al[5] in their research paper they proposed a generic knowledge based framework for effective data cleaning that implements existing cleaning strategies and more.
Kofi Adu-Manu Sarpong et. al[11] in their paper conceptualized the data cleansing process from data acquisition to data maintenance. Data Cleansing is an activity involving a process of detecting and correcting the errors and inconsistencies in data warehouse.
Taoxin Peng [12] in his paper presented a framework for How to improve the efficiency while performing data cleaning and How to improve the degree of automation when performing data cleaning, which provides an approach to managing data cleaning in data warehouses by focusing on the use of data quality dimensions, and decoupling a cleaning process into several sub-processes.

III. NEED AND SCOPE OF STUDY

Every business and organization require the clean and noise free data. Data warehouse load and continuously refresh huge amount of data from variety of sources so the probability that some of sources contain dirty data is high. Data cleaning is used so that the correctness of their data is vital to avoid wrong conclusion. Data cleaning is necessary step in any data- related project. The need of data cleaning is for the improvement the data mining result. Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results. There is huge scope of data cleaning, since every organization is

using data and data can be from more than one sources , so to extract quality and efficient result data cleaning is very necessary.

IV. OBJECTIVE

A. To have the general understanding of data cleaning.
B. To study the data cleaning tools.
C. To clean the data using various data cleaning tools like Ms Excel Data cleaner, RapidMinor and Winpure Clean & Match.

V. RESEARCH METHDOLOGY

The theoretical study has been done from various sources like journals, research papers, books and internet. Data Cleaning tools has been used for the cleaning of various kind of excel data . On the basis of results obtained comparison of tools has been done to find the best tool for data cleaning.

VI. COMPARISON OF DATA CLEANING TOOLS

A. MS EXCEL DATA CLEANER

Excel Files Data Cleaning Utility is an useful addin for Excel to Clean and Organize Data. It is fast & Reliable and you can save your precious time & Money.[14]
Text cleaner: A Collection of Tools to Clean Text in Selected Cells.
Duplicate Cleaner: A Collection of Tools to eliminate duplicate entries.
Data Organizer: A Collection of Tools to Organize Data**.**
Text Organizer: A collection of tool to organize the text.

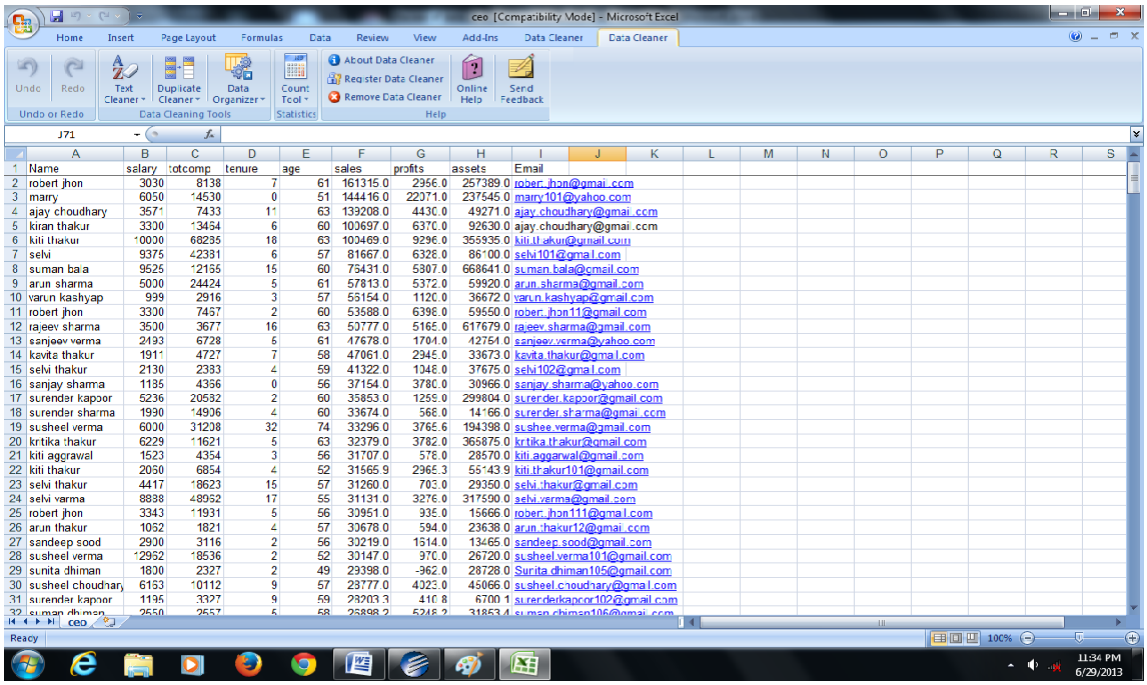

Figure 4: MS Excel with data cleaner

B. RAPIDMINOR:

Rapid minor is a software platform that is used for data mining. We can also clean our data using RapidMinor. This tool contain various operators for data cleaning or data cleansing. It released on 2006, latest version available is Rapid Minor 6. It can be installed on any operating system that is cross platform, Language independent, Licensed by AGPL proprietary. Rapid minor support about twenty two file format [7]. It easily reads and writes Excel files and different databases. Using RapidMinor we can clean ,transform our data[13] . The operators for data cleaning are:-

- Attribute name and Role Modification Operator
- Type conversion
- Attribute set reduction & transformation
- Value Modification
- Outlier Detection
- Filtering
- Sorting
- Rotation
- Aggregation
- Set operation

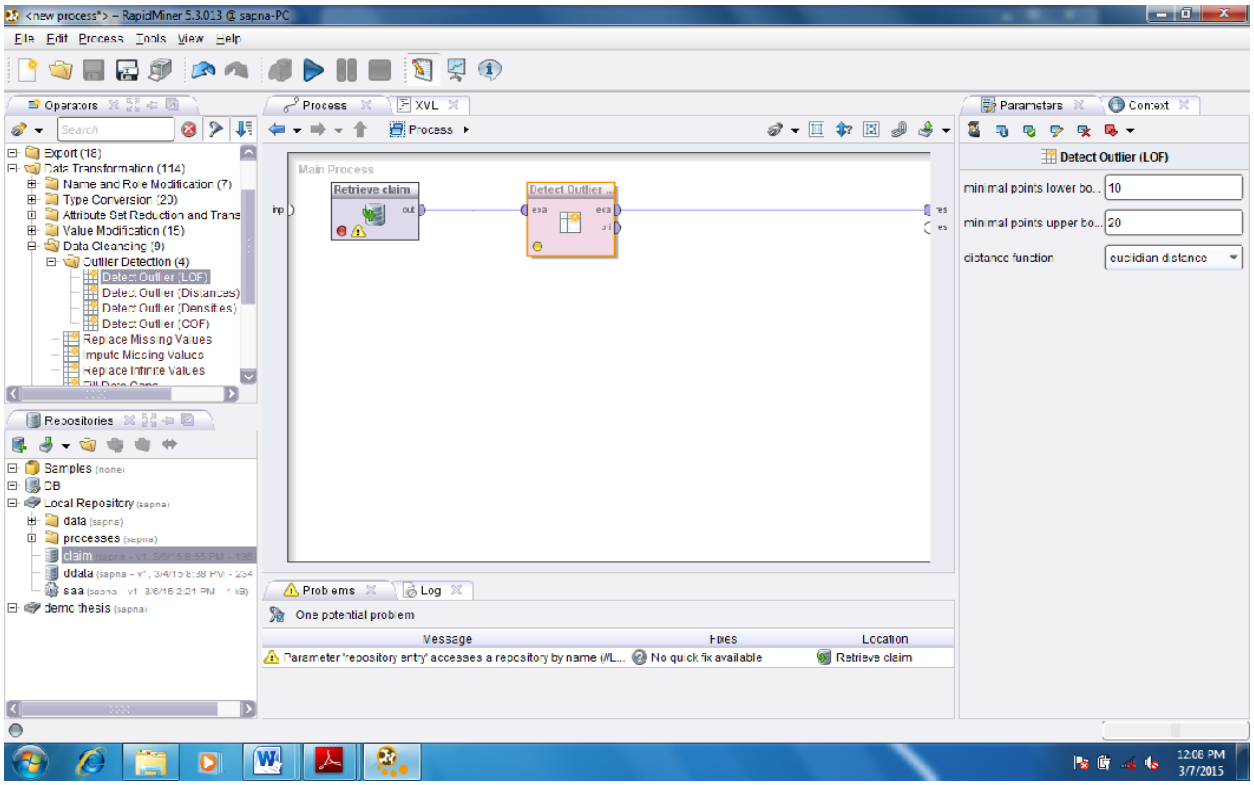

Figure 5: Screen shot of RapidMinor process

C. WINPURE CLEAN & MATCH

Its a comprehensive data cleansing, data deduplication software and a data listing software solution. It helps to clean mailing lists, spreadsheets, marketing databases and electronic mails. The software carries out data deduplication and offers the option of basic as well as advanced search. It helps to merge duplicate records on one or two lists which help to make the merging process effective and easier. "WinPure Clean & Match " offers some interesting features such as 'Safe Merge' options that make sure that no data is lost while merging the records. The following are some properties for "WinPure Clean & Match 2012" Software:

1. It determines a data duplication on the list to achieve the results set.

2. It helps in selecting the master record from each duplicate group, Each group of duplicates provides the ability to select a Master record. This Master record will be the record from each duplicate that will be kept. However this master record has missing values in columns.

So, for each of these duplicate groups we want to remove the duplicate AND also want to populate all the missing values in the master record, to give us a more accurate and populate record for master record. [14]

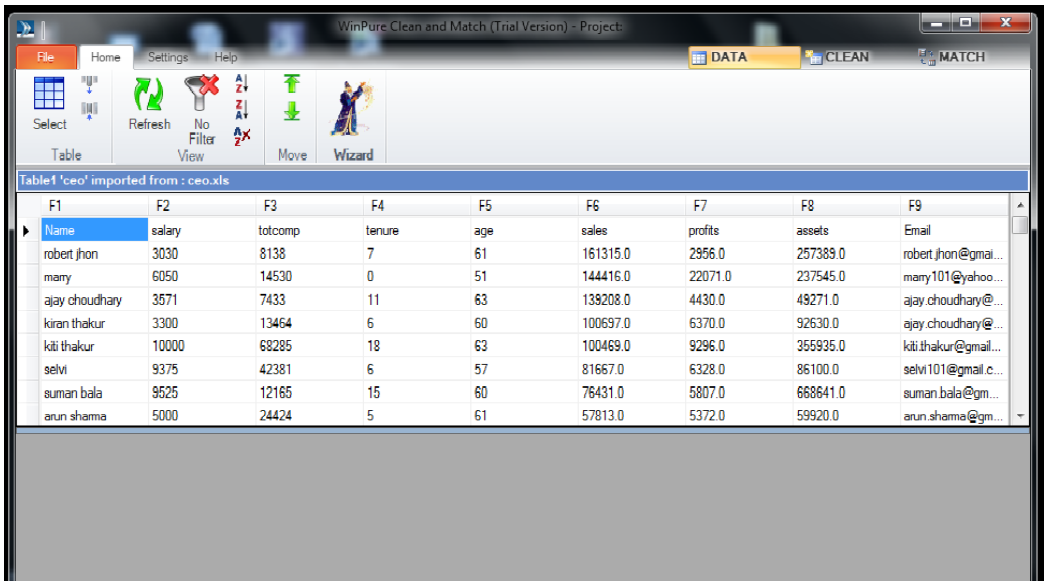

Figure 6: Winpure clean and Match

VII. RESULTS AND ANALYSIS

Table IV represent the three excel files ,cleaned using data cleaning tools , Student file contain 20 records .This file is cleaned using "MS Excel Data cleaner". It contain 10% Missing values and 4 duplicate records. CEO Data file is cleaned using "Winpure clean & Match", it contain 25% missing values and 11 duplicate records. Mis_claim Data file is cleaned using "RapidMinor", It contain 7 % missing values and 20 duplicate records.

TABLE IV

Files used and Properties

File Name	No. of records	No. of Fields	Missing Value	Duplicates Records
Student	**20**	**11**	**10%**	**4**
CEO data	**400**	**9**	**25%**	**11**
Mis_claim Data	**841**	**13**	**7%**	**20**

The Table V shows the result of analysis of above three tools. Missing values represent the how much efficient are these tools in finding missing values of a file, Availability tells us whether these tools are available on desktop or we have to use internet to access the tools, Other features are whether these tools are able to find out duplication, illegal values , Misspelling and merging the records, File format supported by these records and ease of use. Rapid minor provides various other extra feature like outlier detection and set operation , which are not available in other two tools .But Winpure clean & Match have more ease of use.

TABLE V
Comparison of Data Cleaning tools

Tools / Problems	MS Excel with data cleaner	RapidMinor	Winpure Clean & Match
Missing Values	**No**	**Yes**	**Yes**
Availability	**Desktop**	**Desktop**	**Desktop**
Duplication	**Yes**	**Yes**	**Yes uses the matching**
Illegal Values Elimination	**No**	**No**	**Yes**
Misspelling	**No**	**No**	**No**
Merge	**No**	**Yes**	**Yes**
File Format	**Excel**	**CSV, Database, Excel, Access, binary, XML**	**Text files, Excel , commercial DBMS,**
Ease of use	**Moderate**	**Moderate**	**High**

VIII. CONCLUSIONS

Data cleaning is very necessary part of data mining. From the above study we can see that there are different types of problems in data cleaning .Data cleaning methods and approaches depend upon the type of data which we want to clean and according to that we apply particular methods. This paper also present a comparison of data cleaning tools and determines the best tool. Each tool has its own specific features and depending upon the data we can use the tool to clean data. In future work we can check other functionality of these tools and suggest own.

REFERENCES

[1] Li Lee Mong , *Cleansing Data for Mining and Datawarehousing*, school of computing National University of Singapore, 1999 .

[2] Rahm E. & Hai Do Hong, *Data Cleaning: Problems and current approaches*, IEEE Bulletin of the Technical Committee on Data Engineering, 2000

[3] http://www.ukessays.co.uk accessed on 7-4-2014 at 4:18pm

[4] Müller Heiko & Christoph Freytag Johann , *Problems, Methods, and Challenges in Comprehensive Data Cleansing* ,Humboldt-Universität zu Berlin zu Berlin,10099 Berlin, Germany.

[5] Li Lee Mong, Wang Ling Tok & Lup Low Wai, *IntelliClean: A knowledge-based intelligent data cleaner*, Proceedings of the ACM SIGKDD, Boston, USA, 2000

[6] Ibrahim Housien Hamed, Zuping Zhang & Qays Abdulhadi Zainab,*A comparison study Of Data Scrubbing algorithm and framework in Data Warehousing*,,*International Journal of Computer Applications (0975 – 8887) April 2013*

[7] Mikut Ralf & Reischl Wiley Markus Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, pages 431–443, September/October 2011

[8] Choudhary Nidhi **,**A Study over Problems and Approaches of Data Cleansing/Cleaning, Volume 4, Issue 2, February 2014

[9] M. Hellerstein Joseph, Quantitative cleaning of large databases February 27,2008.

[10] Y.Patil Rajashree, Dr. Kulkarni R.V. , A Review of Data Cleaning Algorithms for Data Warehouse Systems. IJCSIT , Vol. 3 (5) , 2012.

[11] Sarpong Kofi Adu-Manu, Davis Joseph George, Panford Joseph Kobina , " A Conceptual Framework for Data Cleansing – A Novel Approach to Support the Cleansing Process" International Journal of Computer Applications , Volume 77– No.12, September 2013.

[12] Peng Taoxin ," A FRAMEWORK FOR DATA CLEANINGS IN DATA WAREHOUSES", School of Computing, Napier University,10 Colinton Road, Edinburgh, *EH10 5DT, UK.*

[13] www.rapidminor.com accessed on 25-2-2015 at 5:18pm .

[14] www.winpure.com accessed on 07-01-2015 at 4:13 pm.

[15] www.exceladin.net accessed on 25-12-2015 at 12:13pm.

