Hybrid Approaches for Data Cleaning in Data Warehouse

Prerna S. Kulkarni
M.E I.T Student
Pillai's Institute of Information Technology, Panvel

J.W. Bakal
Principal
S.S. Jondhale College of Engineering
Sonarpada, Dombivli (E)

ABSTRACT
The quality of data can only be improved by cleaning data prior to loading into the data warehouse as correctness of data is essential for well-informed and reliable decision making. Data warehouse is the only viable solution that can bring that dream into a reality. The quality of the data can only be produced by cleaning data prior to loading into data warehouse. Data Cleaning is a very important process of the data warehouse. It is not a very easy process as many different types of unclean data can be present. So correctness of data is essential for well-informed and reliable decision making. Also, whether a data is clean or dirty is highly dependent on the nature and source of the raw data. Many attempts have been made till now to clean the data using different types of algorithms. In this paper an attempt has been made to provide a hybrid approach for cleaning data which combines modified versions of PNRS,

Transitive closure algorithms and Semantic Data Matching algorithm can be applied to the data to get better results in data corrections.

General Terms
Data Warehouse, Data Cleaning.

Keywords
PNRS, Transitive closure, Semantic Data matching

1. INTRODUCTION
Data cleaning is an essential step in populating and maintaining data warehouses. Data Warehouses needs extensive support for cleaning the data. It is used for decision making. So the correctness of the data is very important, due to which wrong decisions can be avoided. The ETL process [1] is used for cleaning the data, in which the data is collected from different sources, the extracted data is send to the next phase which is transformation, in transformation process series of rules are applied to the extracted data and then all the cleaning of data is performed in separate data staging area. Once the cleaned data is ready it is loaded into the data warehouse. The process is shown in Figure.1.

Managing information in an enterprise typically involves integrating data from across the enterprise and beyond, cleaning the data, matching the data to remove any duplicates, standardize the data, enrich the data, making the data in such a way that it is permissible and fulfilling the requirements as the data is needed, and then storing the data in a central location with all the necessary security settings. The data warehouse consists of huge amount of data whose quality keeps degrading with time and various operations performed on it. Operations such as insertion of new data, deletion of the data and updating causes changes in the data which is reflected in the data warehouse eventually leading to inconsistencies, incorrectness and thus inaccessibility of the quality data. Along with time the data becomes obsolete which also causes problem of inconsistency and inaccessibility. Thus the issue of dirty data is addressed using data cleaning strategies which is the first step in Business Intelligence. Also the quality of data in a data mart hugely affects the performance of an organization and their decisions, thus leading to importance of accuracy and correctness of data in a data warehouse.

Here an attempt has been made to provide approach to data cleaning using modified versions of two basic algorithms namely – PNRS,

Transitive Closure and Semantic Data Matching algorithm to get better results in data correction.

And also Near Miss Strategy will not be working for singular and plural words.

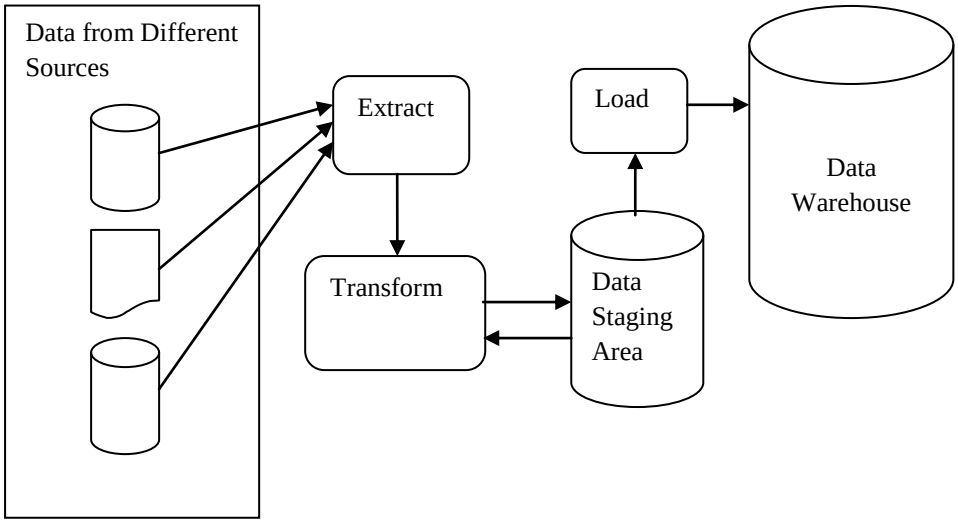

Figure 1 ETL Process [Source 1]

2. OVERVIEW

2.1 PNRS (Personal Name Recognizing Strategy)

This approach [9] is mainly working on the two algorithms; they are a) Near Miss Strategy and b) Phonetic Algorithm.

(a) Near Miss Strategy: For cleaning the data they will follow the following rules

- By inserting a blank space
- By interchanging 2 letters
- By changing/adding/deleting a letter

(b) Phonetic Algorithm: This approach will get the sounds of each Near Miss Strategy words of phonetic approximation sounds will be added and remaining words are eliminated.

The drawback of this approach is that, phonetic sounds are not available for all language

2.1.1 Modified version of PNRS

The spelling errors and ambiguity in terms is a common data entry error found in the database system, to serve this type of dirty data the concept of dictionary is used where the spellings errors are detected using the standard dictionary. Many organizations have different terms assigned to the posts of their employees which may not match with other organizations and serve as jargons. To address this issue many organizations make use organization specific dictionary.

The modified version of PNRS [9] (Near Miss Strategy) will accept the raw data as the input and clean the data based on some dictionary database. This process will get the data by applying the steps, those are adding, deleting or updating the data, removing some blank spaces and sometime inter changing the two letters. Here, in Modified version of PNRS along with standard dictionary, organizational specific dictionary is also used which is important because most of the verbal data present in data warehouses are official data and contain organizational jargons, sometimes even limited to a particular organization.

2.2 Transitive Closure technique:

This approach [9] is having lot of techniques for grouping the words. This is used as the priority of the approaches for cleaning the data. To clean the data, we consider multiple priorities but first priority is only considered. The drawback of grouping the priority approach we get some duplicate data or some miss entries which are available.

2.2.1 Modified version of Transitive Closure

In Transitive Closure, [9] matching of records is done on only one key (attribute) match. It also needs manual intervention to decide whether there is any duplication or correction in the records.

But in Modified version of Transitive Closure more than one key is used for matching the records into one group. The approach [9] is described as follows: prioritizing the keys in the Transitive Closure Algorithm at 2 levels.

At First level, keys are divided into 3 categories:

1) Primary: unique for a person (either one-to-one or one-to-many)
2) Secondary: relatively unique
3) Tertiary: not so unique

At second level, inside the categories, order of the keys is based on decreasing priority of uniqueness/importance.

For e.g. consider the Primary keys.

1) A person has one and only one Unique Id Number which will be given as topmost priority.

2) A person has one Driver's License at a time. But, if a person changes his state/branch as the case may be, he might have to change the driving license.

3) An email id or a mobile number. But a person can have many email ids as well as many mobile numbers so these keys will be kept at lesser order although a mobile no. /email id is unique for a person records.

Here, there is more than 1 key match to conclude that the records are related

- 2 matches are found if at least one of them is a primary key.
- 3 matches if at least two are secondary keys.
- 4 match if at least one key is a secondary key.

3. PROPOSED DATA CLEANING ALGORITHM

A hybrid algorithm is called as HADCLEAN [9] that includes usage of modified versions of PNRS and Transitive Closure algorithms.

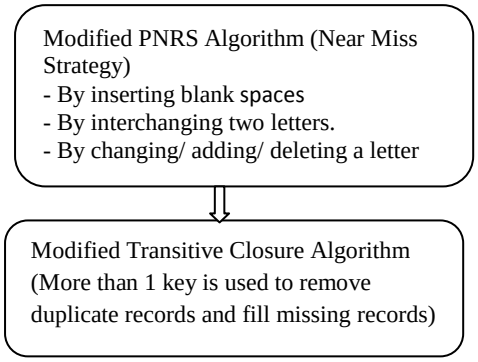

Figure 2 Flowchart of HADCLEAN [Source 9]

This approach work is based on the PNRS and Transitive Closure techniques. Its implementation is based on both the techniques which are slight modified, which will help to solve the draw backs of PNRS approaches and some of Transitive Closure techniques.

The Figure-3 represents the System architecture of the Semantic Data Matching approach along with the modified versions of PNRS and Transitive Closure. This architecture explains about various modules. In this approach we will be using three techniques. They are:

1) PNRS (Near Miss Strategy)
2) Transitive Closure
3) Semantic Approach based on Learning Systems

Here we will be using Semantic Data Cleaning algorithm [6] along with modified versions of PNRS and Transitive Closure. The Semantic Approach will be using the learning system to clean the data. The learning system will get some reference sets from the data based on the key values. Those keys will match that keys upgrading with the exact meaning key values. This approach getting some keys will relate to the some attributes will upgrade in the reference set of database, these keys will be helpful in getting the information cleaned effectively and providing the meaningful information while some grammar and spelling mistakes.

It has been found that in Table.1 [9] the name of the same city is represented by different name. But by using Semantic Approach with learning system we can take care by keeping a unique consistent name of the city based on the semantic similarity between the attribute values in different documents.

4. CONCLUSIONS

Semantic Data Matching algorithm can be applied to the data along with the HADCLEAN algorithms to get better results in data corrections. By using Semantic Data Matching, we can take care of this by keeping a unique consistent name based on the semantic similarity between the attribute values in different documents. This could be tested on huge Enterprise Data that can give us better knowledge of performance and efficiency of this algorithm.

Poor quality data costs businesses vast amounts of money every year. Defective data leads poor business decisions, and inferior customer relationship management. Data are the core business asset that needs to be managed if an organization is to generate a return from it. The above algorithms work as an important tool in the area of data warehouses to obtain quality data. Where alliance rules make use of mathematical calculations to calculate scores, it can be avoided by replacing it with PNRS algorithms.

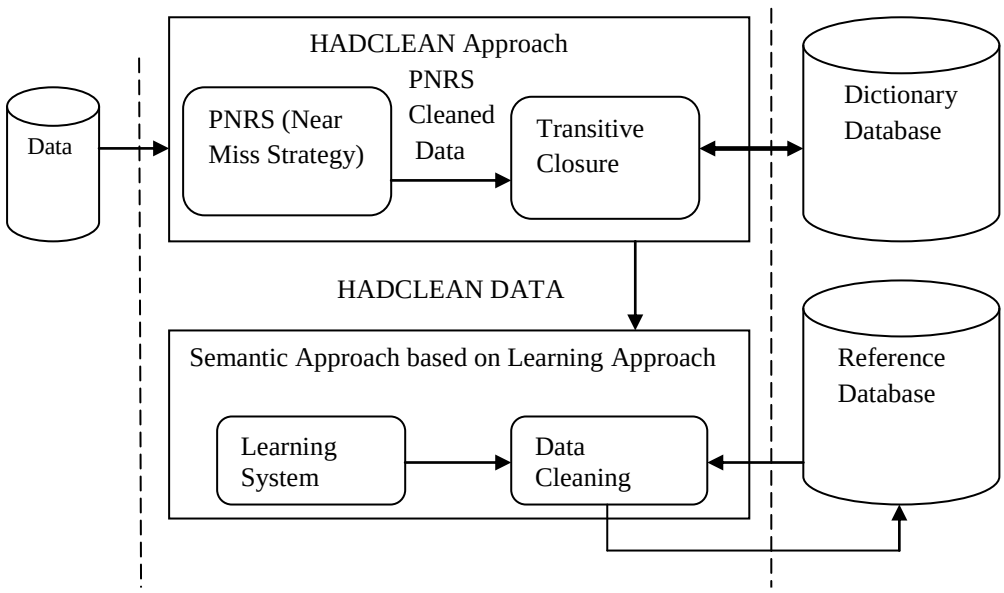

Figure 3: System Architecture for Semantic Data Matching Algorithm

The future work can involve a research on taking a judgment of replacing huge calculations with lesser involving the usage of hybrid approach, the advantage of dictionary strategy and semantic relation based on the reference lists can be extremely useful at several places. The alliance rules helps in error identification and detection, in future enhancement.

5. ACKNOWLEDGMENTS

My sincere thanks to all the members who have guided me throughout this work

6. REFERENCES

[1] E Rahm, Hong Hai Do, "Data Cleaning Problems and Current Approaches" IEEE Bulletin of the Technical Committee on Data Engineering, 2000, 24, 4.

[2] C. Varol, C. Bayrak, R. Wagner and D. Goff, "Application of the Near Miss Strategy and Edit Distance to Handle Dirty Data", Data Engineering - International Series in Operations Research & Management Science, vol. 132, pp. 91 -101, 2010.

[3] W. N. Li, R. Bheemavaram, X. Zhang, "Transitive Closure of Data Records: Application and Computation", Data Engineering – International Series in Operations Research & Management Science, Springer US, vol.132, pp. 39-75, 2010.

[4] M.A. Hernandez and S.J. Stolfo, "Real world Data is Dirty: Data Cleansing and The Merge/Purge Problem", Data Mining and Knowledge Discovery, Springer Netherlands, vol.2, no.1, pp.9-37, 1998.

[5] K. Kukich, "Techniques for Automatically Correcting Words in Text", ACM Computing Surveys, vol. 24, no. 4, pp.377-439, 1992.

[6] Deaton, Thao Doan, T. Schweiger, "SemanticData Matching Principles and Performance", DataEngineering - International Series in Operations Research & Management Science, Springer US, vol. 132, pp. 77-90, 2010.

[7] T. Redman,"The impact of poor data quality of typical enterprise", Communications of ACM, vol. 41, no. 3, pp.79-82, 1998

[8] Anders Haug, FrederikZachariassen, Dennis van Liempd; "The costs of poor data quality" JIEM, 2011 – 4(2): 168-193 – Online ISSN: 2013-0953.

[9] Arindam Paul, VaruniGanesan, JagatSeshChalla, Yashvardhan Sharma, "HADCLEAN: A Hybrid Approach to Data Cleaning in Data Warehouses", Information Retrieval & Knowledge Management (CAMP), International Conference on 2012, Page(s): 136-142.

[10] C.M. Strohmaier, C. Ringlstetter, K. U. Schulz and S.Mihov, "Lexical Post correctionof OCR-Results: The Web as a Dynamic Secondary Dictionary", Seventh International Conference on Document Analysis and Recognition (ICDAR'03),VOL. 2, pp.1133,2003

[11] S.M. Beitzel, E.C. Jensen and D.A. Grossman, "Retrieving OCR text: A Survey of Current Approaches", Symposium on Document Image Understanding Technologies (SDUIT), Greenbelt, MD, 2003.

[12] P. Jokinen, J. Tarhio and E. Ukkonen, "A Comparison of Approximate String Matching Algorithms", Journal of Software Practice and Experience, vol.1,no.1,pp.1-4,1988.Matching

Table 1 Sample Data after Modified PNRS and Modified Transitive Closure

Record no.	UID	Driving License	First Name	Middle Name	Last Name	City	State
6	12345	MH899878	Priya	Balwant	Patil	**Mumbai**	Maharashtra
7	23456	OR456557	Varuna	G	Mukharjee	Bhubaneshwar	Orissa
13	546521	KA474898	Narayan	Raghavendra	Murthy	Gulbarga	Karnataka
15	215253	WB89999	Suman	Hemant	Korawar	Kolkatta	Bengal
9	778974	MH699668	Sahil	Hanumant	Bhalerao	**Bombay**	Maharashtra
27	58757	PB336656	Shanti	Raju	Gaur	Ferozabad	Punjab
11	996851	TN885878	Krishna	Murthy	Nair	Chennai	Tamil Nadu
3	475626	KA745414	Ram	Vitthal	Hubli	Bijapur	Karnataka