A Comparative Analysis of Data Cleaning Approaches to Dirty Data

Sonal Porwal
M.E Computer Student
Vidyalankar Institute of Technology
Mumbai, India

Deepali Vora
Asst. Professor and Head of the Information and Technology Department
Vidyalankar Institute of Technology
Mumbai, India

ABSTRACT

Data Cleansing or (data scrubbing) is an activity involving a process of detecting and correcting the errors and inconsistencies in data warehouse. Thus poor quality data i.e.; dirty data present in a data mart can be avoided using various data cleaning strategies, and thus leading to more accurate and hence reliable decision making. The quality data can only be produced by cleaning the data and pre-processing it prior to loading it in the data warehouse.

As not all the algorithms address the problems related to every type of dirty data, one has to prioritize the need of its organization and use the algorithm according to their requirements and occurrence of dirty data.

This paper focuses on the two data cleaning algorithms: Alliance Rules and HADCLEAN and their approaches towards the data quality. It also includes a comparison of the various factors and aspects common to both.

General Terms

Data Mart, Data Warehouse, Dirty data.

Keywords

HADCLEAN, PNRS, phonetic algorithm, alliance rules, transitive closure, near miss strategy, scores

1. INTRODUCTION

The data warehouse consists of huge amount of data whose quality keeps degrading with time and various operations performed on it. Operations like insertion, deletion 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.

The alliance rules algorithm [1] identifies the errors in the data by using a mathematical association rule concept. The mathematical concept helps reduce the comparison time of the strings for errors as it makes use of numerical integers for analysis.The token based approach is also proposed in [6] to reduce the time for the comparison process and to increase the speed of the data cleaning process.

The HADCLEAN algorithm [2], uses a hybrid approach of the concept PNRS of and the transitive closure. The PNRS strategy makes use of standard dictionary to detect errors like misspelled words, typo errors etc. The transitive closure approach makes use of related data grouping concept.

An attempt has been made in this paper to provide with various approaches in data cleaning and also highlight the importance of a quality data. The basic algorithms will be explained in the further sections.

The remainder of the paper is as follows. Section II describes the overview of the data cleaning concept and its importance in the business world. Section III briefly explains the two algorithms that were used as a main reference to this paper. Section IV describes the comparative study obtained in the form of a table. Section V includes the conclusions and recommendations of future work.

2. OVERVIEW OF DATA CLEANING

The most important step in any database organization is to confirm the quality of data. The data processing is an essential pre –requisite to verify the data and confirm their values in accordance to some set of records. For example a customer id field should include a unique number and not an age of the customer instead, here although the age of the customer is correct but still the data is misplaced and thence is an error or (dirty data).

2.1 Importance of Data Cleaning

Any organization which uses its database for knowledge discovery and decision making will be required to keep its database updated and error free. The failure leads to loss of quality data and increase in operational costs. The data warehouse users use the features of data like coherency, correctness and accuracy of the data, which degrades with time and regular updates which in turn has an effect on the integrity of the data residing in a data warehouse.

These errors thus lead to poor decision making and errors in the trend analysis. Clean data is an essential requirement to any sales, marketing and distribution strategy. The avoidance of dirty data will help to decrease operational costs and time, thus leading to an improving brand image of an organization.

Data cleaning is thus a process of maintaining data quality by identifying incorrect or invalid or may be duplicate entries in the information systems. Data quality is the degree to which data meets the specific needs of specific customers, which contains several dimensions [3].Also the data quality is determined by the quality of the data source. The measures of data quality are validity, completeness, accuracy, non-duplication, and precision, timeliness [3], [7], and [9]

It is often performed as a step in data profiling [4] activity. Various sources of errors like:

a) Data entry errors

b) Measurement errors

c) Distillation errors

d) Data integration errors lead to a poor quality data called as dirty data.[3,5]

2.2 The Concept of Dirty Data

The concept of dirty data can be said as any data which is not consistent with the already residing data in a data warehouse. The types of dirty data could be misspellings like "green" replaced by "rgeen","1" replaced by "l",typographical or phonetic errors. Some fields also have numerical constraints like weight cannot be "negative", people cannot have "more than two parents", a human cannot be of "more than 100-120 years". The outlier errors like a "20 feet man", inconsistencies like incorrect zip code for a city also lead to dirty data. Some fields require a default value to be entered, failure of which leads to an erroneous data. Misfielded values are the ones that are actually correct but wrongly placed.eg: country="Mumbai" also called as (column shift error). Another critical type of dirty data is an obsolete data which loses its significance with time, such data leads to wastage of resources hence an update or deletion process should be followed. Also some organizations have different formats of field like a date field is of the format DD/MM/YY or MM/DD/YY which leads to a data entry error, while entering the data manually, a different unit of measurement e.g.:" meters" v/s "inches", different modes of payment e.g.: "daily" v/s "weekly" or "monthly" v/s "annually". Another major form of dirty data is duplicity which leads to wastage of resources. Duplicity may be due to spelling variations, naming conventions etc. [8].It can be eliminated using character based similarity metrics, token-based and empirical [10].

Some critical applications however require a conformance check on the data set, where a dirty data is eliminated followed by a verification process of data entry. This however increases the operational costs and time but leads to performance upgrade and organization's appraisal. Hence data cleansing forms an integral part of the data processing and increases the reliability of the results obtained.

2.3 Steps involved in data cleansing

Data cleansing is usually a two-step process including detection and then correction of errors in a data set.

The steps involved in Data Cleansing are:

a) Identification of errors-records could have incomplete or corrupted data.

b) Perform error verification-whether it is truly an error or not. This situation occurs in organizations where there exists a usage of organizational jargons [2].

c) Extract the data to be cleaned-the data is extracted and stored in a temporary table, operations are performed and the data is repaired and verified, then it is replaced in the target table.

d) Perform data cleaning-which can be done automatically or manually.

Manual process is however avoided as it is highly time consuming and tedious in nature. It is limited by human capabilities like speed, accuracy in error detection and correction. Thus leading to more error prone performances and degrading the quality of data, which in turn leads to increase in operational costs and hence poor decision making.

It is extremely important to categorize the data according to the rate of its criticality.

a) Critical errors-needs to be immediately addressed i.e.; error reporting ,verification and cleansing

b) Non-critical errors-can be temporarily ignored.

Further section will involve a discussion about the algorithms that can be followed for data cleansing.

3. OVERVIEW OF ALGORITHMS

3.1 Alliance Rules Algorithm

3.1.1 The need

The need of this algorithm raised in the data mart system of an organization which involved a customer bill generation. The storage of large amount of information about the customers suffers from the problem of dirty data. A data warehouse is formed by merging data from different sources which can have different field formats. A data mart of a telecommunication system may involve sections like bill generation, account section, personal information section etc. [1] which on merging can lead to several inconsistences and hence dirty data.

This algorithm addresses the errors and issues occurred in the 'name' field of the data warehouse. The name field being an important aspect in a customer based organization forming an integral part of the bill generation and their strategies, any duplicity or field mismatch can lead to organization mistrust and wastage of time.

3.1.2 Steps involved

This algorithm address the duplicity error of string data type (name filed) and uses the algorithm of *de-duplicity in the name field of the data warehouse.*The steps involved are

I. Preprocessing

II. Alliance rules application

III. Detection of error

I. Pre processing

Here the strings in the name field are converted into a numerical value which is stored in another file called *Score* for reference. The integer values are called *scores* of the name. The string is converted into numbers using relation

[((radix) ^place value)* face value] mod m

Figure 1. Calculation of Scores

The formula described in Figure 1[1] shows the calculation of *scores.* The paper refers the total number of words in a name defined as N. e.g.: Sonal S.Porwal has N=3. The total number of scores would hence be N+1.

The elements in [1] describing the formula are:

a) Radix is 27 characters (26 alphabets and '.'),

b) Face value is the sequence of occurrence of characters in the world of alphabets starting with 0-a---25-z and 26-(.)

c) The place value is marked from right to left starting from 0.

d) M is any large prime number

e) letters are case-insensitive

II. Alliance rules application

Here the 2 data marts are considered such that a name from DM1 is to be checked and matched for duplicity with all the names in another data mart DM2.The steps involved are briefly introduced in paper [1]as*alliance rules application and duplicity detection.*

III. Detection of errors

The errors in the name are evaluated using the concept of q-grams testing. The q-grams are the substring of a given name string. The length of the substring can be of any value smaller than the length of the name string itself.

E.g.: SON POR

Q=3

The q-grams are as follows:

(1,##S),(2,#SO),(3,SON),(4,ON_),(5,N_P),(6,_PO),(7,POR),(8,OR#),(9,R##).

Here the initial '##' determines that the name is started, and the later determines the end. This method also considers the space between the words as well.

3.1.3 Limitations

a) This approach specifically deals with the study of error types and their detection related to string data types.

b) It mainly focuses on the 'name' string format and does not focus on any other format.

c) Another drawback was that it could not detect the duplicity error efficiently in some cases; it needed the date of birth of that person as a reference.

d) In cases the DOB field is incorrect or blank field then the cleaning process could suffer.

e) A lot of manual work is required in the pre-processing phase i.e.; the calculation of the *scores*, and hence error –prone.

3.2 HADCLEAN Algorithm

3.2.1 The need

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 f their employees which may not match with other organizations and serve as jargons. To address this issue many organizations make use organization specific dictionary.

Also some records have blank fields that can be filled using transitive closure algorithm.

3.2.2 Steps involved

I. PNRS

The Personal Name Recognition Strategy corrects the phonetic and typo errors using standard dictionaries.it employs two strategies:

a) Near miss strategy: It addresses the errors in the words which are nearly missed and shows errors.it is done by inserting a blank space e.g.: co-education, by interchanging two letters e.g.:'rgreen'with 'green', by changing/adding/deleting a letter. All these actions are taken with the help of reference to standard dictionaries.

b) Phonetic algorithm: It uses the concept of phonetic codes which is calculated for every word and then it is matched with the phonetic code of the standard dictionary.it helps to detect errors for words like 'seen' and 'scene' which sound same (phonetic) but have different meaning and different phonetic codes respectively.

c) The modified PNRS-some organizations involve the usage of jargons and often have their organizationsdesignation in regional languages, for such situations an organization specific dictionary can be used as a reference.

II. Transitive Closure

Here the records are matched using the attribute keys. Using key the records are grouped and matched as a set of related records and then errors are detected and corrected after detailed analysis. This helps to fill in the blank cells (fields) and also remove the duplicity errors and redundancies.

The modified transitive closure makes use of more than one key to match and group the related records. Primary secondary and tertiary key concept is used to group the related records, and when the records are matched blanks are filled and redundancies are removed.

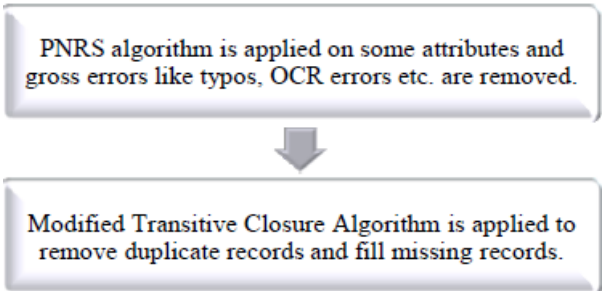

Figure 2. Flowchart of HADCLEAN

As shown in the Figure 2 [2], the flowchart above summarizes the working of the HADCLEAN approach. The PNRS strategy is applied followed by modified transitive closure algorithm [2].

3.2.3 Limitations

a) The prioritization to the attribute keys [2] in modified transitive closure makes the algorithm data specific, hence needs manual intervention.

b) The modified transitive closure algorithm has some specifications defined in order to combine the records as related. The rules being strict, sometimes we are not able to combine the records even if they are related because it has only one secondary key and two tertiary key matches. This can be observed from the work done in the[2] in the form of table which has the list of records before and after applying the algorithm.

c) Another situation occurs in case of records which have values 'Mumbai' and some 'Bombay'. Semantic Data Matching Principles [11] can be applied to the data along with the above to get better results. The work in [5] clearly explains how this problem can be avoided by keeping a unique consistent name for the city based on the semantic similarity between the attribute values e.g.:street address, pin code etc.

4. COMPARATIVE ASPECTS OF THE ALGORITHMS

The alliance rules algorithm deals with error identification and detection of string data types. The HADCLEAN also stresses on the detection of errors in the string data types but with a little different approach. While alliance rules algorithm makes uses of the mathematical concept to detect errors,HADCLEAN on the other side uses dictionary based concept to detect spelling errors.

The alliance rules algorithm focuses majorly on the 'name' field which is of string data type. Whereas the HADCLEAN approach covers the spelling errors, typographicalerrors (string data types) etc. along with blank fields errors i.e. Missing data [1], [4] too using (transitive closure) [2].Where alliance rules algorithm involves a lot of mathematical calculations which is tedious and error prone,HADCELAN on the other hand uses dictionary based approach and keys of the database which is easily accessible and less error prone and less complex.

But one major drawback of the HADCLEAN strategy is that it is only applicable to English language. On the other hand, one major advantage of the alliance rules is that it converts the string data types into integer numbers (*Scores*), thereby addressing the memory concerns.

The duplicity in the name field in alliance rules algorithm is addressed with the help of *Score*matching and the error is detected using *Q-grams* [1].Transitive closure helps to convert *n* no of records into related group and not only identification but removal of redundancies as well.

Modified transitive closure improves the result with the help of primary, secondary and tertiary key concept. The modified PNRS however corrects more number of errors than PNRS in many situations, as shown by the work done in [2] in the form of collection of the records for experimental purpose and analysis obtained in the form of an output table with cleaner records. Similarly the analysis was obtained for the transitive closure algorithm resulting in better result for modified version.

The table below highlights the essential points of comparison between the two algorithms:

Table 1. Comparative analysis of Alliance rules and HADCLEAN algorithms

Factors	Alliance Rules	HADCLEAN
Approach	Inter-related	Hybrid
Steps required	Pre-processing, Alliance rules Detection and Q-gram	PNRS and Transitive Closure
Strategy	Scores calculation and comparison	Uses dictionary based approach and comparison using phonetic codes
Dependency	Pre-processing forms the base of the other steps, alliance rules is applied on the basis of the *Scores* obtained.in case the above two steps fail-grams is applied.	The Steps are independent of each other, yet they are used as an input to another with the intention of hybrid approach
Complexity	Since the algorithm involves dependencies in the steps performed, complexity is bound to be greater as compared to HADCLEAN. Also mathematical calculation in the pre-processing stage increases the complexity.	The independency of the steps performed in the algorithm makes it relatively less complex. However the hybrid approach makes the algorithm slightly more complex.
Types of dirty data addressed	Misspelled data Duplicate data	Nearly Misspelled data, phonetic errors, and typographical errors.
Accuracy	Far better because it uses absolute matching(q-grams)	Approximate because transitive closure uses the rules of keys matching to group records)
Ease of implementation	Difficult as defining *scores* involves huge mathematical calculation	Easy as less calculations are involved
Application	Customer oriented data warehouse	Organization specific data warehouse
Output	Error is identified and detected	Error is detected and corrected
Drawback	Many calculations involved	The dictionary based approach for PNRS works only for English language.

5. Conclusions and Future work

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 [3]

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.

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 can be extremely useful at several places. The alliance rules helps in error identification and detection, thus only partial work is done. This can be completed with the HADCLEAN approach where the error is corrected as well.

6. ACKNOWLEDGMENTS

My sincere thanks to the members who have guided me throughout the research.

7. REFERENCES

[1] Rajiv Arora,PayalPahwa and ShubhaBansal,"Alliance Rules for Data Warehouse Cleansing", 2009.IEEE Press, Pages 743-747.

[2] ArindamPaul,VaruniGanesan,"HADCLEAN:A Hybrid Approach to Data Cleaning in Data Warehouses",2012.IEEE Press,Pages 136-142.

[3] Dr.MortadhaM.Hamad,AlaaAbdulkarJihad,"An Enhanced Technique to Clean Data in the Data Warehouse",2011,IEEE.

[4] Kamran Ali,MubeenAhmed,"A framework to implement Data Cleaning in Enterprise Data Warehouse for Robust Data Quality",2010,IEEE Press,Pages 1-6.

[5] W. Kim, B. Choi, E. Hong, S. Kim and D. Lee, "A taxonomy of dirtydata," Data Mining and Knowledge Discovery, 7, 81–99, 2003.

[6] J.Jebamalar Tamilselvi,Dr.V.Saravanan,"Handling Noisy Data using Attribute Selection and Smart Tokens",2008.IEEE Press,Pages 770-774.

[7] Yan Hao,"Research on Information Quality Driven Data Cleaning Framework",2008.IEEE ,Pages 537-539

[8] WaiLupLow,Mong Li Lee, "A Knowledge based Approach for Duplicate Elimination in Data Cleaning", School of Computing, National University Singapore.

[9] Lukasz Ciszak,"Application of Clustering and Association Methods in Data Cleaning,2008,IEEE,proceedings of the International Multiconference on Computer Science, Pages 97-103.

[10] Mariam Rehman,"Duplicate Record Detection for Database Cleaning", 2009.IEEEconference.,Pages 333-338.

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