Hybrid Technique for Data Cleaning

Ashwini M. Save
P.G. Student,
Department of Computer Engineering,
Thadomal Shahani Engineering College,
Bandra, Mumbai, India

Seema Kolkur
Assistant Professor,
Department of Computer Engineering,
Thadomal Shahani Engineering College,
Bandra, Mumbai, India

ABSTRACT

Data warehouse contains large volume of data. Data quality is an important issue in data warehousing projects. Many business decision processes are based on the data entered in the data warehouse. Hence for accurate data, improving the data quality is necessary. Data may include text errors, quantitative errors or even duplication of the data. There are several ways to remove such errors and inconsistencies from the data. Data cleaning is a process of detecting and correcting inaccurate data. Different types of algorithms such as Improved PNRS algorithm, Quantitative algorithm and Transitive algorithm are used for the data cleaning process. In this paper an attempt has been made to clean the data in the data warehouse by combining different approaches of data cleaning. Text data will be cleaned by Improved PNRS algorithm, Quantitative data will be cleaned by special rules i.e. Enhanced technique. And lastly duplication of the data will be removed by Transitive closure algorithm. By applying these algorithms one after other on data sets, the accuracy level of the dataset will get increased.

General Terms

Data Warehouse, Data cleaning

Keywords

Data cleaning, PNRS, Improved PNRS, Enhance Technique, Transitive.

1. INTRODUCTION

Data warehouse is important for storing the large amount of data which plays important role in management's decision support system. This data is stored in the data warehouse should be correctly entered, accurate and relevant. Because incorrect or inaccurate data i.e. dirty data may cause to create various problems like taking incorrect decisions or actions based on that dirty data. Many times incorrect data can be costly. For instance many companies may get problems such as sending mails repeatedly because of incorrectly entered address of the customer. Company may lose customers due to such incorrectly entered data [1][3].

Data quality is an important factor in the data warehousing projects. Dirty data must be detected and corrected to improve the data quality. Data cleaning or scrubbing is the process of correcting or removing corrupt or inaccurate records from the database [3].

Data cleansing is the first step and most critical, in a Business Intelligence (BI) or Data Warehousing (DW) projects Thus, it is very significant to perform data cleaning process for building any enterprise data warehouse [1].

This paper presents a hybrid (integrated) technique to perform data cleaning process for building any enterprise data warehouse by using algorithms that detect and correct most of the error types and expected problems.

- Text data will be cleaned by improved PNRS algorithm
- Quantitative data will be cleaned by special rules i.e. Enhanced Technique.
- Duplication of the data will be removed by Transitive closure algorithm.

So that by applying these algorithms one after other on data sets, the accuracy level of the dataset will be increased.

2. RELATED WORK

Many methods are proposed by researchers for data cleaning. Dictionary based data cleaning is widely used technique. In this the dictionary has been maintained for the mapping incorrect word and correcting it according to the dictionary word. Dictionary used for the data cleansing process can be both real world dictionary as well as organization level dictionary.

C.varol et al. [4] have proposed PNRS algorithm i.e. Personal Name Recognizing strategy. This algorithm works on text based entries. PNRS include Near-Miss strategy and Phonetic algorithm. This algorithm detects and corrects textual words using standard verbal vocal dictionaries. Arindam Paul et al. [1] have given approach for PNRS by using organization specific dictionary along with the real world dictionary. M.M. Hamad [2] has proposed data cleaning technique for quantitative data. Invalid quantitative data can be detected and corrected by applying some special rules. This enhanced technique can be used for quantitative data that has limited values.

Many researchers have worked on the transitive closer algorithm for cleaning the data. M.A. Hernandez, et al. [5] has worked on transitive closure algorithm that helps in finding the duplicates in the data. R. Bheemavarm et al. [7] have proposed approach to group related data records together using the transitive closure. W.N. Li et al. [8] have used transitive closure algorithm in filling of missing records, removing data redundancies and grouping of similar records together. Finding duplication of records and removing the duplications is one of the important tasks of the data cleaning system. [10][11]. Data ware house projects are highly dependant on the data quality; hence data cleaning system plays an important role in such data warehouse and business intelligent projects [12].

3. BACKGROUND

This paper propose the Hybrid technique for the data cleaning system. Hybrid technique is grouping of data cleaning algorithms like PNRS algorithm, Enhanced Technique, Transitive algorithm. In this process, first Improved PNRS algorithm will be applied which corrects text Enhanced Technique will be applied to correct quantitative data. And after correcting Text and quantitative data, lastly duplication of records will be removed

and missing values will be filled by applying Transitive algorithm.

3.1 PNRS Algorithm:

C.Varol et al. [4] Have proposed the PNRS Algorithm for Data Cleaning. It corrects the phonetic and typo-graphical errors present in the data set using standard dictionaries. PNRS algorithm mainly includes two Algorithms-

i) Near-Miss Strategy –

This approach works on the technique where two words are found identical by interchanging, inserting, or by deleting two letters. If valid word is generated by applying this technique; it is added to the temporary suggestion list. This can be reviewed and corrected in the original data by automatic or some manual intervention.

ii) Phonetic Algorithm-

When the word is truly miss-spelled, Near-Miss doesn't work efficiently as it's unable to give best list of suggestions. In Phonetic Algorithm phonetic code is calculated for Miss-Spelled word which has to be compared with the phonetic codes of the word list in dictionary. When it gets matched the word is added to the temporary suggestion list and which can be reviewed and corrected by automatic or some manual intervention.

As per A. Paul et al. [1], in the PNRS Algorithm, an organization specific dictionary is being used along with the standard dictionary for checking the spelling mistakes.

3.2 Enhanced Technique:

M. Hamad at el. [2] has attempted to solve all errors and problems that are expected in the quantitative data. An enhanced technique to clean data in the data warehouse uses a new algorithm to detect and correct most of the error types and expected problems, such as lexical errors, domain format errors, irregularities, integrity constraint violation. Here is presented a solution to handle data cleaning process by using an enhanced technique for data cleaning [2][9].

3.3 Transitive Closure Algorithm:

A. Paul et al. [1], have proposed the Transitive Closure Algorithm which works in fully automated way. This approach is based on using more than one key to match the records into same group.

This technique works at two levels.

i) At the first level, it divides keys in three categories i.e. into primary, secondary, and territory.
ii) Then at second level inside the categories order the keys based on decreasing priority of Uniqueness/importance. Then proper rules on records are applied depending upon number of keys matches which will find related records. And accordingly duplicate records get deleted from the dataset automatically and also missing values gets filled by this technique.

4. PROPOSED DATACLEANING TECHNIQUE

4.1 Improved PNRS:

The modification done in the PNRS Algorithm in this paper is, Using some predefined characteristics on some attributes; it will give more accurate and less number of suggested words in the suggestion list. That means it will avoid some non-relevant suggestions so that it will help user to choose the correct word easily from the suggestion list.

For example, if we apply this modified version of Near-Miss Strategy in PNRS algorithm on the 'NAME' attribute. if we have specified characteristic of 'AMIT' as 'MALE' and 'AMITA' AS 'FEMALE' in the dictionary itself; then while correcting the incorrectly entered word say 'AMI' whose Gender attribute is having value as MALE; will get 'AMIT' as a suggestion in the list. It will simply avoid 'AMITA' suggestion in the list even though it is present in the dictionary.

Taking another example, if we apply this modified version of Phonetic Algorithm in PNRS algorithm on the 'CITY' attribute. If we have specified characteristic of 'AHMEDABAD' as 'GUJRAT' and 'AHMEDNAGAR' AS 'MAHARASHTRA' in the dictionary itself; then while correcting the incorrectly entered word say 'AHAMED' whose State attribute is having value as GUJRAT; will get 'AHMEDABAD' as a suggestion in the list. It will simply avoid 'AHMEDNAGAR' suggestion in the list even though it is present in the dictionary.

4.2 Enhanced Technique:

The PNRS Algorithm which includes Near-Miss Strategy and Phonetic algorithm works efficiently on the 'Text Data'. i.e. this algorithm detects and corrects all textual errors effectively. These algorithms when works on the data set for data cleaning, the Quantitative data is completely ignored in these approaches. Hence in this paper we have proposed an enhanced technique for data cleaning; which will work on the quantitative data so that incorrectly entered numerical values will get identified and corrected.

M. Hamad at el. [2] has attempted to solve all errors and problems that are expected in the quantitative data. They have presented some rules needed in the data cleaning system.

Enhanced technique algorithm is applied on the quantitative attributes for example Date of birth and age. Which corrects wrongly corrected data.

If the entry of attribute mobile no. is incorrectly entered as '98964327', Then applying special rule -

{Mobile no. should have 10 digits}

It will give incorrect entry with the suggestion 'Mobile No. should be of 10 digits'

4.3 Hybrid Approach:

Proposed data cleaning technique used here is Hybrid system for the data scrubbing. The improved PNRS algorithm is used to correct the textual data then modified enhanced technique is used to detect and correct quantitative data. So that, data cleaning system can handles 'Text' as well as 'Quantitative' data. After applying these algorithms i.e. PNRS and Enhanced technique; The Transitive closure algorithm can be applied on the data. So that it will remove redundant data and fill missing values. Applying transitive algorithm at the end of the hybrid system will give more accuracy than applying it directly on the unclean dataset. And also, while filling missing values after removal of the duplication; it will take newly corrected data (Text/Numerical).

Hence rather than using these algorithms separately i.e. only to correct text data or quantitative data or to avoid duplicate records, this proposed hybrid technique covers all areas of the data i.e. text fields, quantitative fields and then removal of duplication and filling missing values. So this Hybrid approach gives best data

cleaning system to scrub the data in the dataset. Fig. 1 shows overall system flowchart.

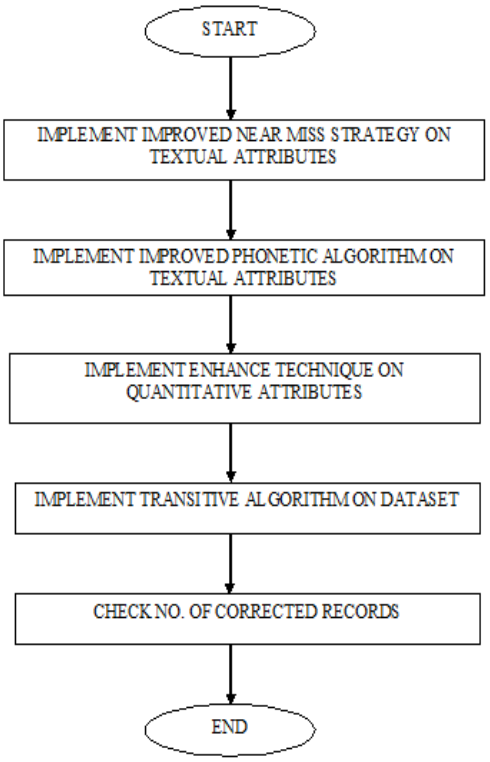

Fig1: System Flowchart

5. RESULTS AND ANALYSIS

Here results are discussed with help of examples in tables with some attributes as follows:

Table1. Unclean Dataset

Stu_ID	Religion	State	DOB	Age
1000	Hind	Maharash	11/5/1992	21
1001	Jian	Gujarat	18/10/1991	22
1002	Christian	Kerala	24/10/1991	22
1003	Sikh	Lunjab	13/3/1988	25
1004	Islam	Maharashtra	17/7/19992	21
1005	Hindu	Madras	10/8/1990	23
1006	Hindu	Karnataka	21/4/1990	21
1007		Maharashtra	28/3/1992	22
1008	Christian	Madras	15/15/1791	22
1001	Jain	Gujrat	18/10/1991	23
1007	Hindu	Maharashtra	28/3/1992	22

Table 1 show some attributes with wrongly entered data which is highlighted here with an underline. Also some records are repeated which means duplication of records is present in the table. This is considered as unclean dataset required as input for this system. Here hybrid approach is applied to clean this data in which algorithms like near-miss strategy, phonetic algorithm, enhanced technique with special rules and Transitive algorithm is applied on the dataset.

Table 2. Result by applying Improved PNRS Algorithm

Stu_ID	Religion	State	DOB	Age
1000	Hindu	Maharashtra	11/5/1992	21
1001	Jain	Gujarat	18/10/1991	22
1002	Christian	Kerala	24/10/1991	22
1003	Sikh	Punjab	13/3/1988	25
1004	Islam	Maharashtra	17/7/19992	21
1005	Hindu	Madras	10/8/1990	23
1006	Hindu	Karnataka	21/4/1990	21
1007		Maharashtra	28/3/1992	22
1008	Christian	Madras	15/15/1791	22
1001	Jain	Gujarat	18/10/1991	23
1007	Hindu	Maharashtra	28/3/1992	22

Table 2 shows result by applying only Improved PNRS Algorithm on the text data. Here it shows in the table that text data of the attributes Religion and state has been corrected by using Improved PNRS algorithm i.e. near miss strategy and phonetic algorithm. But attributes with quantitative data such as DOB and age has been remain unchanged.

Table 3. Result by applying Improved PNRS algorithm and Enhance technique.

Stu_ID	Religion	State	DOB	Age
1000	Hindu	Maharashtra	11/5/1992	21
1001	Jain	Gujarat	18/10/1991	22
1002	Christian	Kerala	24/10/1991	22
1003	Sikh	Punjab	13/3/1988	25
1004	Islam	Maharashtra	17/7/1992	21
1005	Hindu	Madras	10/8/1990	23
1006	Hindu	Karnataka	21/4/1990	23
1007		Maharashtra	28/3/1992	22
1008	Christian	Madras	15/15/1991	22
1001	Jain	Gujarat	18/10/1991	23
1007	Hindu	Maharashtra	28/3/1992	22

Table3 shows result by applying Improved PNRS Algorithm on the text data and enhanced technique on the quantitative data. Here it shows in the table that text data of the attributes Religion and state has been corrected by using Improved PNRS algorithm and also attributes with quantitative data such as DOB and age has been corrected by Enhanced Technique. But still we can observe here that although spelling mistakes and incorrect numerical values has been corrected but duplication of the record is present in the data set.

Table 4. Result by applying Hybrid approach (Improved PNRS, Enhanced technique, Transitive)

Stu_ID	Religion	State	DOB	Age
1000	Hindu	Maharashtra	11/5/1992	21
1001	Jain	Gujarat	18/10/1991	22
1002	Christian	Kerala	24/10/1991	22
1003	Sikh	Punjab	13/3/1988	25
1004	Islam	Maharashtra	17/7/1992	21
1005	Hindu	Madras	10/8/1990	23
1006	Hindu	Karnataka	21/4/1990	23
1007	Hindu	Maharashtra	28/3/1992	22
1008	Christian	Madras	15/12/1991	22

Table 4 shows result by applying Improved PNRS Algorithm on the text data, enhanced technique on the quantitative data and lastly Transitive algorithm to remove duplication of records and also fill missing values. After applying this hybrid approach; as an outcome it will give the dataset with all correct entries and with no duplication of records.

Taking some examples from above tables we can show how each algorithm works on the data:

5.1 Near-miss strategy:

Near-miss strategy is applied on the attribute Religion.

For example: If the entry of attribute Religion is incorrectly entered as 'Jian' instead of 'Jain' then applying near miss strategy two nearer letters 'i' and 'a' will be interchanged and the word will be corrected as 'Jain'.

5.2 Phonetic algorithm:

Phonetic algorithm is applied on the attribute State.

For example : If the entry of attribute State is incorrectly entered as 'Maharatr' whose phonetic code is M-636 then it will be matched with dictionary phonetic code M-623 as shown in fig. 2. Then the word will be corrected as 'Maharashtra'.

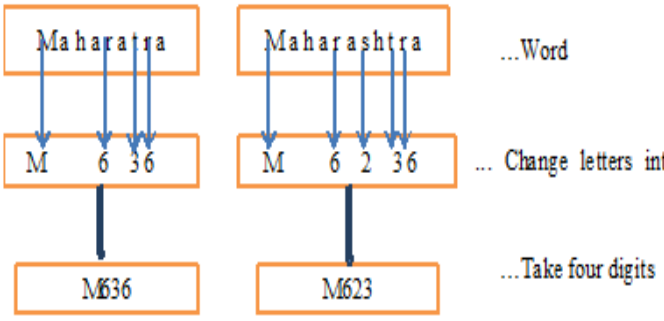

Fig 2: working of phonetic Technique

5.3 Enhanced technique:

Enhanced technique algorithm is applied on the quantitative attributes Date of birth and age.

For example: If the entry of attribute Birthday is incorrectly entered as '19992', then applying special rule

{Age = Current Date –Birthday; 0< age <120; no negative}

The entry will be corrected as '1992' and age will be calculated as 21.

5.4 Transitive algorithm:

Transitive algorithm will be implemented and applied on the dataset to remove duplicate records and fill missing values.

For example: Stud_

Id 1001 which is entered two times in dataset will be merged as a one record by feeling missing values if any [8].

6. CONCLUSION

The Improved PNRS algorithm proposed in this paper will avoid some non-relevant suggestions so that it will help user to choose the correct word easily from the suggestion list. It makes our system more accurate.

Also the Hybrid System for data cleaning has been proposed which includes Improved PNRS algorithm, Enhanced Technique and Transitive Closure Algorithm. Improved PNRS algorithm detects and corrects unclean Textual data. An Enhanced technique detects and corrects unclean quantitative data. and Transitive Closure removes duplication of data and fills missing values.

Experiment has been performed on sample dataset. Comparison between each algorithm applied separately and applying this hybrid data cleaning system shows that, applying hybrid system on the unclean data, system gives more accuracy than applying each algorithm separately. PNRS algorithm first clears all the Text data then in the next step Enhance Technique clears all quantitative values and once all the misspelled and incorrect values gets corrected, Transitive algorithm has to apply on the clean values of dataset which detects and removes the duplication and fill the missing values.

In future scope this data cleaning system can be implemented by making all these algorithms automatic instead of manual intervention. Algorithms can automatically select correct value from the dictionary and correct errors followed by removing the duplication and filling missing values.

7. REFERENCES

[1] Arindam Paul, V.Ganesan, and J.Challa, "HADCLEAN: A Hybrid Approach to Data Cleaning in Data Warehouses" IEEE, 2012.

[2] Mortadha M. Hamad and AlaaAbdulkhar Jihad, "An Enhanced Technique to Clean Data in the Data Warehouse"IEEE,2011.

[3] K. Ali and M. Warraich, "A framework to implement data cleaning in enterprise data warehouse for robust data quality" IEEE, 978-1-4244-8003-6/10, 2010.

[4] 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.

[5] M. A. Hernández 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.

[6] R. Bheemavaram, J. Zhang and W. N. Li, "Efficient Algorithms for Grouping Data to Improve Data Quality", Proceedings of the 2006 International Conference on Information & Knowledge Engineering (IKE 2006), CSREA Press, Las Vegas, Nevada, USA, pp. 149-154, 2006.

[7] R. Bheemavaram, J. Zhang, W. N. Li, "A Parallel and Distributed Approach for Finding Transitive Closures of Data Records: A Proposal", Proceedings of the Acxiom Laboratory for Applied Research (ALAR), pp. 71-81, 2006.

[8] 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.

[9] Ballou, D. (1999) "Enhancing data quality in Data Warehousing Environment," *Comm. ACM* (42:1), pp. 73-78.

[10] M. Bilenko and R. J. Mooney. "Adaptive duplicate detection using learnable string similarity measures" ACM SIGKDD, 2003, pp 39-48

[11] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. "Duplicate Record Detection": A Survey. IEEE TKDE, 19(1), 2007, pp 1-16

[12] S. Reddy, A. Lavanya, V. Khanna, L.S.S. Reddy, "Research Issues onData Warehouse Maintenance", IEEE, ICACC '09. InternationalConference Advanced Computer Control, Singapore, Jan 2009, Page(s): 623 – 627