HADCLEAN: A Hybrid Approach to Data Cleaning in Data Warehouses

Arindam Paul, Varuni Ganesan, Jagat Sesh Challa, Yashvardhan Sharma
Department of Computer Science & Information Systems
Birla Institute of Technology & Science, Pilani
Pilani, Rajasthan, India - 333031
e-mail: arindampaul.bits@gmail.com, varuni.ganesan@gmail.com, jagatsesh@gmail.com, yash@bits-pilani.ac.in

Abstract—**Data Cleaning is a very important part of the data warehouse management process. It is not a very easy process as many different types of unclean data (bad data, incomplete data, typos, etc) can be present. 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 to clean the data using blocking algorithms, phonetic algorithms, etc. In this paper an attempt has been made to provide a hybrid approach HADCLEAN for cleaning data which combines modified versions of PNRS and Transitive closure algorithms.**

Keywords- PNRS; HADCLEAN; transitive closure; near miss; phonetic algorithm; data warehouse

I. Introduction

Data cleaning is an essential step in populating and maintaining data warehouses. Owing to likely differences in conventions between the external sources and the target data warehouse, as well as due to a variety of errors, data from external sources may not conform to the standards and requirements at the data warehouse. Therefore, data has to be transformed and cleaned before it is loaded into the warehouse so that downstream data analysis is reliable and accurate. This is usually accomplished through an Extract-Transform-Load (ETL) process. Typical data cleaning tasks include record matching, de-duplication, and column segmentation which often go beyond traditional relational operators. This has led to the development of a broad range of methods intending to enhance the accuracy and thereby the usability of existing data. Data cleansing is the first step, and most critical, in a Business Intelligence (BI) or Data Warehousing (DW) project, yet easily the most underestimated. T. Redman [1] suggests that the cost associated with poor quality data is about 8-12% of the revenue of a typical organization. Thus, it is very significant to perform data cleaning process for building any enterprise data warehouse.

An attempt has been made in this paper to provide a hybrid approach to data cleaning using modified versions of two basic algorithms namely – PNRS and Transitive Closure. These have been explained in further sections.

The remainder of the paper is as follows. Section II describes the related work in the field of data cleaning. Section III briefly elucidates the two basic algorithms that were used as main reference to this paper. Section IV & V describe the algorithm proposed and analysis respectively. Section VI gives conclusions and recommendations for future work.

II. Related Work

Researchers have proposed various approaches to data cleaning. Dictionary based data cleaning is very commonly used. Christian M. Strohmaier, et al. [2] has proposed post correction of OCR-results for text documents by using fixed, static large scale dictionaries, dynamic dictionaries (retrieved via an automated analysis of the vocabulary of web pages from a given domain) and mixed dictionaries. Beitzel, S.M, et al. [3] provided an overview of work done to improve the effectiveness of retrieval of OCR text. Various mechanisms in consideration include IR Models for OCR text, processing OCR text for categorization, auto-correction of OCR errors and improved string matching on noisy Data. JM Trenkle, et al. [4] presented the design of a high-performance recognition system for recognizing low-quality characters extracted from postal address blocks. They employed disambiguation and spell-correction methods in order to substantially improve the performance. K. Kukich [5] aimed at correcting words in the text focusing mainly on 3 problems – i) non word error detection, ii) isolated word error correction, iii) context dependant word correction. C. Varol et al. [6] have proposed PNRS algorithm which stands for Personal Name Recognizing Strategy. It has two algorithms Near Miss Strategy and Phonetic Algorithm which correct words using standard Verbal and Vocal Dictionaries.

Various researchers have attempted to clean the data on the basis of transitive closure technique. M. A. Hernández, et al. [7] proposed an approach that helps in finding the duplicates in the data using transitive closure technique. R. Bheemavaram et al. [8] attempted to group related data records together using the transitive closure. R. Bheemavaram et al. [9] have proposed a suitable algorithm for computation of transitive closure algorithm in a distributed and parallel way while dealing with data cleaning to huge data. P. Jokinen et al. [10] gave a comparative study of string matching algorithms which include Dynamic Programming, Galil Park algorithm, Ukonen Wood Algorithm, etc. W.N. Li et al. [11] used transitive closure in

978-1-4673-1090-1/12/$31.00 ©2012 IEEE

filling of missing records, removing data redundancies and grouping of similar records together.

Next section gives the background of the basic algorithms that have been extended to design our algorithm for data cleaning.

III. Background

In this section, a brief description PNRS and Transitive Closure is given.

A. PNRS

The PNRS algorithm, proposed by C. Varol et al. [6], corrects the phonetic and typographical errors present in the raw data, using standard dictionaries. It mainly employs two algorithms which are explained below.

- **Near Miss Strategy** – Two words are considered "near" if they can be made identical –
  - By inserting a blank space
  - By interchanging 2 letters
  - By changing/adding/deleting a letter

  If a valid word is generated using this technique, it is added to temporary suggestion list, which can be reviewed and corrected in the original data automatically or with some manual intervention.
- **Phonetic Algorithm** - Phonetic Algorithm uses a rough approximation of how each word sounds. This is important as "near miss" doesn't provide us with the best list of suggestions when a word is truly mis-spelled. This compares the *phonetic code* of the mis-spelled word to all the words in the word list. If the *phonetic code* matches, then the word is added to the temporary suggestion list, which can be reviewed and corrected in the original data automatically or with some manual intervention.

In this way, PNRS corrects the errors in the data. Next section explains the second algorithm – Transitive Closure.

B. Transitive Closure

Transitive Closure algorithm for data cleaning has been proposed by W.N. Li, et al [11]. This algorithm preprocesses the data to categorize millions and billions of records into groups of related data.

The ETL tool using following algorithms processes the individual groups for data cleaning which involves

- Identifying and removal of redundancies - This is especially valuable when we are migrating data from different source systems where we might store same data in different formats
- Filling the blank cells.
- Establishment of "group" relationship between different records leading to faster querying.

In this technique the records are matched on the basis of matching of the keys (keys are selected attributes of the data). Each key is matched one after the other, so as to obtain related group of records. These groups can further be analyzed and corrected. Blanks can be filled, and redundancies can be removed.

Next section explains the proposed data cleaning algorithm along with a case study.

IV. Proposed Data Cleaning Algorithm With A Case Study

A hybrid algorithm called HADCLEAN is being proposed in this paper that includes usage of modified versions of PNRS and Transitive Closure algorithms. The proposed approach is explained using a sample data as shown in Table 1. Each algorithm is applied one after the other to obtain the cleaned data.

First the modification proposed to PNRS has been explained in the subsequent section.

A. Modified version of PNRS algorithm

The contemporary PNRS algorithm [6] was correcting the spelling mistakes in the data on the basis of any Standard English dictionary or a standard dictionary available for that particular language. It takes care of phonetic and typographical errors present in the raw data. It has a limitation that it works effectively only for English language as the phonetic algorithm is available at present only for English language. It only removes errors in the words where two words can be made equal by inserting, interchanging or deleting a letter. But sometimes, same word can exist in slightly different formats which cannot be corrected by contemporary PNRS.

A modification that can be proposed here is to use an organization specific dictionary, along with a standard dictionary, for checking the spelling mistakes. This 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. For example, in Table 1, the field "Identification Marks" can be corrected using a Standard Dictionary, but the field "Occupation" cannot be corrected as most of the Occupational titles could be in regional languages as well which don't exist in the standard dictionaries. Also there could be organizational specific titles to the designation or the occupation. We can observe that in records 1 and 25, PNRS has corrected "Occupation" and the "Identification marks".

The modified (corrected) data after applying modified PNRS algorithm is shown in Table 2.

In our data PNRS can be applied for the fields "City", "State", "Identification Marks" and "Occupation" but not to the fields like "Name" and "Address" as they are not found in any dictionaries with clarity. So here comes the role of transitive closure which is explained in section that follows.

B. Modified version of Transitive Closure algorithm

Transitive Closure algorithm matches two or more records into one group when one of the key (attribute) matches between the two records. But, sometimes by matching only one key to group records would result in mistakes as we cannot rely only on matching one key. Also this runs completely in semi automatic mode, where the records are grouped together and then leaves room for manual intervention to study these groups and then declare that duplication or correction of data in the records.

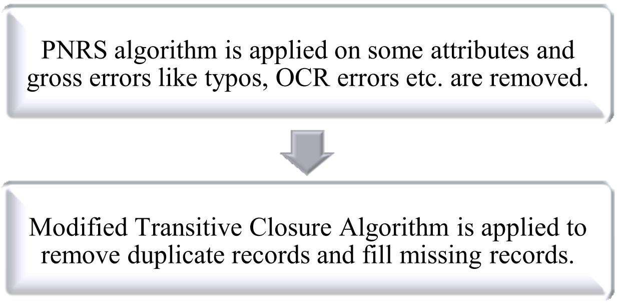

PNRS algorithm is applied on some attributes and gross errors like typos, OCR errors etc. are removed.
Modified Transitive Closure Algorithm is applied to remove duplicate records and fill missing records.

Figure 1. Flowchart of HADCLEAN

In our paper we are modifying this transitive closure in a way to fully automate it without any manual intervention. This is primarily based on using more than one key to match the records into one group or rather saying that these records are the same. The approach is described as follows:

We are suggesting prioritizing the keys in the Transitive Closure Algorithm at 2 levels. At First level, we divide our keys into 3 categories:

a) Primary: unique for a person (either one-to-one or one-to-many)
b) Secondary: relatively unique
c) Tertiary: not so unique

In our case study we categorize the keys into primary, secondary and tertiary as follows:

Primary – UID, Driving License, Mobile Number, Email-ID
Secondary – name, street address
Tertiary – pin-code, DOB, Identification mark

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

For e.g. consider the Primary keys.

1) A person has one and only one UID. So this presumes the 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 is mapped to only one person but not vice-versa. So, a person can have many mobile numbers and email ids. So, we keep these keys at lesser order although a mobile no. /email id is unique for a person. Again, the probability of having multiple mobile nos. for a person is lesser than that of having multiple email ids, so that comes at a later priority.

So, in the case of primary key, we keep **UID>Driver’s License>mobile number>email id** (For simplicity, we have removed mobile number in the sample data). For secondary key, we have **name>street address**. We consider a combination of first, middle and last name for the name key. For tertiary keys, we keep **pin-code >DOB>Identification mark**.

Now after categorizing the attributes, we apply following rules on the records to find out the related records.

Number of key matches to conclude them as related records -

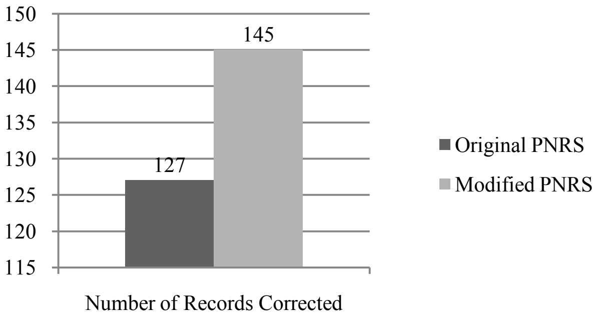


Figure 2. Contrast of No. of Records Corrected

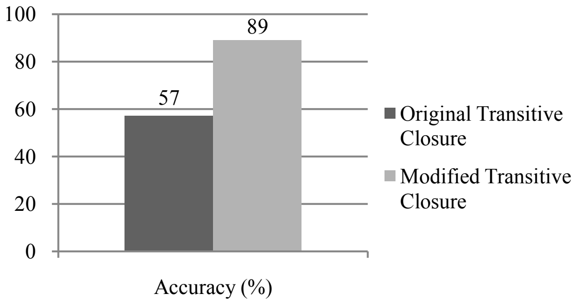


Figure 3. Contrast of Accuracy of Both Algorithms

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

Record 1 is removed according to rule 1 as there are two primary key matches (UID and Driving License) with Record 25. Record 5 is removed by rule 2 as there are two secondary key matches and one tertiary key match (Name, Street address, pincode). Record 8 is removed by rule 3 as there is one secondary key match and three tertiary key matches (name, pincode, DOB and identification mark).

The modified (corrected) data after applying the modified transitive closure algorithm appears in Table3.

The flow chart of the algorithm **HADCLEAN** is shown in Figure 1.

V. Analysis

Experiment has been performed comparing the modified algorithms with the originally proposed algorithms. Public data consisting of 1200 records has been generated for experimental purposes. The sample data presented in Tables 1, 2 and 3 is a subset of the whole data. Both original algorithms and the modified versions of the algorithms have been coded and applied on the data generated. The results obtained have been contrasted with as explained below.

PNRS and the modified version of the PNRS have been run on the data generated. It was found the modified PNRS was correcting 18 records more than what original PNRS was able to correct. Figure.2 shows the comparison of the number of records corrected.

TABLE I.

Sample Data before Data Cleaning Algorithm

Record No.	UID	Driving License No.	email id	First Name	Middle Name	Last Name	Street Address	City	State	Pincode	DOB	Occupation	Marital Status	Identification Mark
1	2932922	KA273782	nandu@fb.com	Nandamanohar		Kulkarni	34/2, Netaji Nagar	Bidar	Karnataka	232332	15.01.1965	Software Enginear	Unmarried	Red sopt on left ear
2	7226377	TN263238	ganeshansubr@redif	Ganesan		Subramanium	56/3, Anna Mudai Street	Chennai	TamilNad	873237	21.08.1990	Teachee	Married	Black spot on left cheek
3	8349326	PB278323	shipramnga@gmail.	Shipra	K	Mongia	23/5, Sector-23, Defence Colony	Firozabad	Punjab	266327	21.1.1991	Student	Unmarried	Mole on left ear
4	8349324	PB278323	shipramnga@gmail.	Shipra	Kamath	Monga	23/5, Sector-23, Defence Colony	Ferozabad	Punjab	266327	21.1.1991	Student	Unmarried	Mole on left ear
5	8349324	KA234983	shanti_hasaan@fac	Shanti	Padma	Bhushan	29-A, Main Street	Hassan	Karnataka	234534	23.04.1962	State Govt Employee	married	
6	3455656			Shanti	Padma	Bhushan	29-A, Main Street	Hassan	Karnataka	234534	23.04.1962	State Govt Employee	married	six fingers on the left hand
7	3475656			Shanti	Padma	Bhushan	29-A, Main Street	Hassan	Karnataka	234534	23.04.1972	State Govt Employee	married	
8	98908			Mary	Anbu	Mani	658-H, MuthuMari Street	Mumbai	Maharashtra	546456	07.12.1966	HouseWife	married	scar on cheen
9	9890809	TN873249	MaryMani@gmail.c	Mary		Mani	658-H, MuthuMari Street	Bombay	Maharashtra	546467	07.12.1956	HouseWife	married	scaar on chin
10	2362989	WB723273	aksahagrwl@gmail.c	Akash		Agarwal	Block-F, Plot-35,Ground Floor, New-Alipore	Kolkata	Paschim Banga	700053	27.8.1983	Researcher	Unmarried	Mol on nose
11	2362989	WB723273	aksahagrwl@gmail.c	Akash		Agarwal	Block-F, Plot-35,Ground Floor, New-Alipore	Calcutta	Paschim Bangal	700053	27.8.1983	Researcher	Unmarried	Mol on nose
12	6223723			Samrat	Kanti	Bose	14/2, James Long Sarani	Kolkata	West Bengal	772837	23.04.1987	Consultant	Divorced	Red spot on left cheek
13	6223723	WB723234	samratbose@ymail.	Samrat	Kant	Bose		Kolkata	West Bengal	772837		Consultant	Married	Black spot on right chick
14	6223723	WB723234	samratbose@ymail.	Samrat	Kanti	Bose	21B, S.N. Bose Road	Kolkatta	Bengal	772837		Consultant	Divorced	
15	2328728	OR232893	sumanmohanty@gr	Suman	K	Mohanty	21/3, M.G. Road	Bhubaneshwar	Orissa	283782	30.11.1988	IT Professional	Unmarried	Mole near left eyebrow
16	2328728	UK726822		Sarthak	Kumar	Mohanty	21/2, Vasant Nagar	Haridwar	Uttarkhand	282929	31.8.1965	Teacher	Married	Scar on right elbow
25	2932922	KA273782	nandu@fb.com	Nandmanohar		Kulkarni	34/2, Netaji Nagar	Bidar	Karnataka	232332	15.01.1965	Software Engineer	Unmarried	Red spot on left ear
26	7226377	TN263238	ganeshansubr@redif	Ganesan		Subramanium	56/3, Anna Mudai Street	Chennai	Tamil Nadu	873237	21.08.1990	Teacher	Married	Black spot on left cheek
27	98908			Mary	Anbu	Mani		Mumbai	Maharashtra	546456	07.12.1966	HouseWife	married	scar on cheen

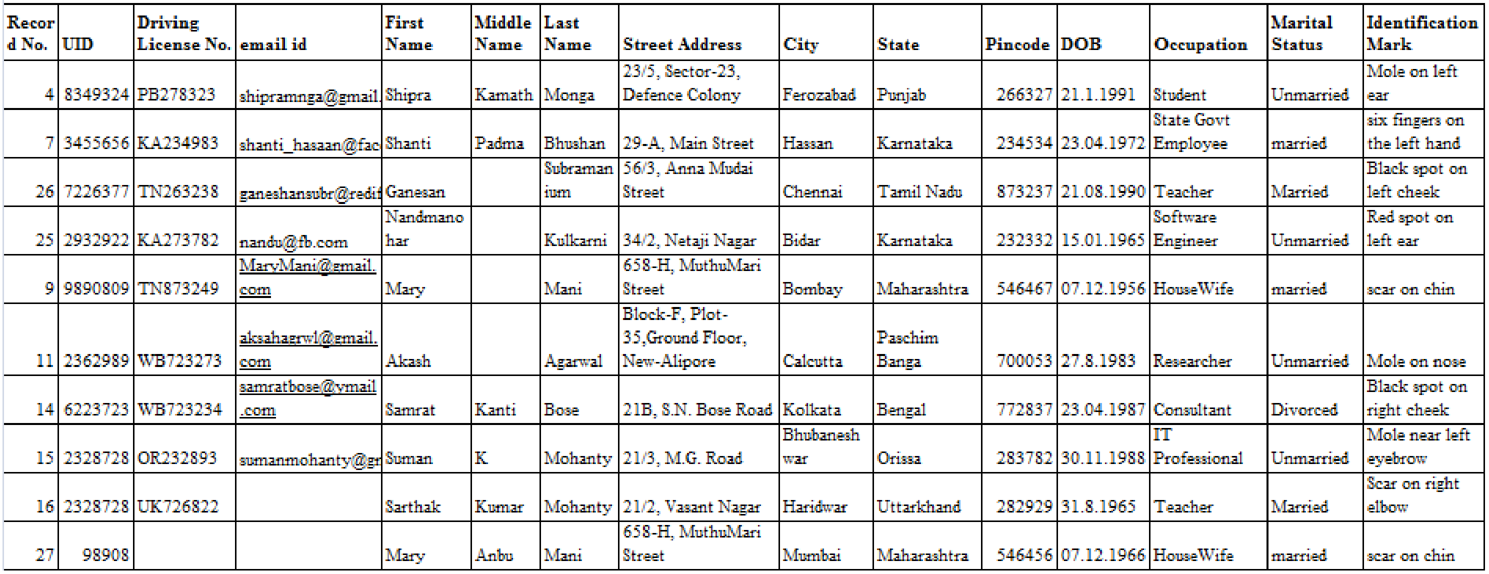
TABLE II.

Sample Data after Modified PNRS

Record No.	UID	Driving License No.	email id	First Name	Middle Name	Last Name	Street Address	City	State	Pincode	DOB	Occupation	Marital Status	Identification Mark
1	2932922	KA273782	nandu@fb.com	Nandamanohar		Kulkarni	34/2, Netaji Nagar	Bidar	Karnataka	232332	15.01.1965	Software Engineer	Unmarried	Red spot on left ear
2	7226377	TN263238	ganeshansubr@redif	Ganesan		Subramanium	56/3, Anna Mudai Street	Chennai	TamilNadu	873237	21.08.1990	Teacher	Married	Black spot on left cheek
3	8349326	PB278323	shipramnga@gmail.	Shipra	K	Mongia	23/5, Sector-23, Defence Colony	Ferozabad	Punjab	266327	21.1.1991	Student	Unmarried	Mole on left ear
4	8349324	PB278323	shipramnga@gmail.	Shipra	Kamath	Monga	23/5, Sector-23, Defence Colony	Ferozabad	Punjab	266327	21.1.1991	Student	Unmarried	Mole on left ear
5	8349324	KA234983	shanti_hasaan@fac	Shanti	Padma	Bhushan	29-A, Main Street	Hassan	Karnataka	234534	23.04.1962	State Govt Employee	married	
6	3455656			Shanti	Padma	Bhushan	29-A, Main Street	Hassan	Karnataka	234534	23.04.1962	State Govt Employee	married	six fingers on the left hand
7	3455656			Shanti	Padma	Bhushan	29-A, Main Street	Hassan	Karnataka	234534	23.04.1972	State Govt Employee	married	
8	98908			Mary	Anbu	Mani	658-H, MuthuMari Street	Mumbai	Maharashtra	546456	07.12.1966	HouseWife	married	scar on chin
9	9890809	TN873249	MaryMani@gmail.c	Mary		Mani	658-H, MuthuMari Street	Bombay	Maharashtra	546467	07.12.1956	HouseWife	married	scar on chin
10	2362989	WB723273	aksahagrwl@gmail.c	Akash		Agarwal	Block-F, Plot-35,Ground Floor, New-Alipore	Kolkata	Paschim Banga	700053	27.8.1983	Researcher	Unmarried	Mole on nose
11	2362989	WB723273	aksahagrwl@gmail.c	Akash		Agarwal	Block-F, Plot-35,Ground Floor, New-Alipore	Calcutta	Paschim Banga	700053	27.8.1983	Researcher	Unmarried	Mole on nose
12	6223723			Samrat	Kanti	Bose	14/2, James Long Sarani	Kolkata	West Bengal	772837	23.04.1987	Consultant	Divorced	Red spot on left cheek
13	6223723	WB723234	samratbose@ymail.	Samrat	Kant	Bose		Kolkata	West Bengal	772837		Consultant	Married	Black spot on right cheek
14	6223723	WB723234	samratbose@ymail.	Samrat	Kanti	Bose	21B, S.N. Bose Road	Kolkata	Bengal	772837		Consultant	Divorced	
15	2328728	OR232893	sumanmohanty@gr	Suman	K	Mohanty	21/3, M.G. Road	Bhubaneshwar	Orissa	283782	30.11.1988	IT Professional	Unmarried	Mole near left eyebrow
16	2328728	UK726822		Sarthak	Kumar	Mohanty	21/2, Vasant Nagar	Haridwar	Uttarkhand	282929	31.8.1965	Teacher	Married	Scar on right elbow
25	2932922	KA273782	nandu@fb.com	Nandmanohar		Kulkarni	34/2, Netaji Nagar	Bidar	Karnataka	232332	15.01.1965	Software Engineer	Unmarried	Red spot on left ear
26	7226377	TN263238	ganeshansubr@redif	Ganesan		Subramanium	56/3, Anna Mudai Street	Chennai	Tamil Nadu	873237	21.08.1990	Teacher	Married	Black spot on left cheek
27	98908			Mary	Anbu	Mani		Mumbai	Maharashtra	546456	07.12.1966	HouseWife	married	scar on chin

TABLE III.

Sample Data after Modified PNRS and Modified Transitive Closure



Record No.	UID	Driving License No.	email id	First Name	Middle Name	Last Name	Street Address	City	State	Pincode	DOB	Occupation	Marital Status	Identification Mark
4	8349324	PB278323	shipramnga@gmail.	Shipra	Kamath	Monga	23/5, Sector-23, Defence Colony	Ferozabad	Punjab	266327	21.1.1991	Student	Unmarried	Mole on left ear
7	3455656	KA234983	shanti_hasaan@fac	Shanti	Padma	Bhushan	29-A, Main Street	Hassan	Karnataka	234534	23.04.1972	State Govt Employee	married	six fingers on the left hand
26	7226377	TN263238	ganeshansubr@redif	Ganesan		Subramanium	56/3, Anna Mudai Street	Chennai	Tamil Nadu	873237	21.08.1990	Teacher	Married	Black spot on left cheek
25	2932922	KA273782	nandu@fb.com	Nandmanohar		Kulkarni	34/2, Netaji Nagar	Bidar	Karnataka	232332	15.01.1965	Software Engineer	Unmarried	Red spot on left ear
9	9890809	TN873249	MaryMani@gmail.com	Mary		Mani	658-H, MuthuMari Street	Bombay	Maharashtra	546467	07.12.1956	HouseWife	married	scar on chin
11	2362989	WB723273	aksahagrwl@gmail.com	Akash		Agarwal	Block-F, Plot-35,Ground Floor, New-Alipore	Calcutta	Paschim Banga	700053	27.8.1983	Researcher	Unmarried	Mole on nose
14	6223723	WB723234	samratbose@ymail.com	Samrat	Kanti	Bose	21B, S.N. Bose Road	Kolkata	Bengal	772837	23.04.1987	Consultant	Divorced	Black spot on right cheek
15	2328728	OR232893	sumanmohanty@gr	Suman	K	Mohanty	21/3, M.G. Road	Bhubaneshwar	Orissa	283782	30.11.1988	IT Professional	Unmarried	Mole near left eyebrow
16	2328728	UK726822		Sarthak	Kumar	Mohanty	21/2, Vasant Nagar	Haridwar	Uttarkhand	282929	31.8.1965	Teacher	Married	Scar on right elbow
27	98908			Mary	Anbu	Mani	658-H, MuthuMari Street	Mumbai	Maharashtra	546456	07.12.1966	HouseWife	married	scar on chin

Transitive Closure algorithm and the modified version of the Transitive Closure algorithm have been run on the data generated. It was found the modified transitive closure was able to group the records more accurately than the original algorithm. The contrast of the average accuracy of grouping of records has been shown in Figure.3.

The original transitive closure algorithm was only grouping related records together. It was unable to identify identical records or redundancies. It is required to manually study related groups of records to find out redundancies. But in the modified version, the redundancies are removed automatically which establishes its superiority over the older ones.

VI. Conclusions, Limitations & Future Work

Thus the analysis establishes the superiority of the proposed algorithm in automating the data cleaning process. The following section states a few limitations and future work.

A. Limitations

- The modified version of the transitive closure algorithm has prioritization of the attribute keys which is data specific. This cannot be automated and needs manual intervention.
- Apart from these three primary algorithms, there can always be other data specific algorithms to clean the data. For example the errors in the field “date of birth” can be corrected. For e.g. year of birth 1850 may not be correct for an employee of a particular firm. It can be auto corrected to 1950. There is always a room for such data specific corrections.
- We are not able to combine record 9 and 27 as there is only one secondary key and two tertiary key matches. They can be merged by manual correction.

B. Future Work

- Semantic Data Matching algorithm can be applied to the data along with the above algorithms to get better results in data corrections. This algorithm has been clearly explained by R. Deaton, et al. in [12]. For e.g. in our final sample data set after applying Transitive Closure in the records 9 and 27, Bombay and Mumbai both refer to the same city but two different representations are present. By using Semantic Data Matching, we can take care of this by keeping a unique consistent name for city based on the semantic similarity between the attribute values in different documents.
- One area that has been identified for future work is the usage of Principal Component Analysis (PCA) while modifying the transitive closure algorithm. Based on the input data we can modify the attributes so as to prioritize them appropriately using PCA. These prioritized attributes can be used as keys in Transitive Closure algorithm.
- As of now, our algorithm has been tested on data with only 1200 records. This could be tested on huge Enterprise Data that can give us better knowledge of performance and efficiency of this algorithm.

Acknowledgment

We express benign thanks to Mr. Nikhil Agarwal and Mr. Naveen Goyal (BITS-Pilani, India) for their support.

REFERENCES

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

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

[3] 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.

[4] J.M. Trenkle and R. C. Vogt, "Disambiguation and spelling correction for a neural network based character recognition system", Proceedings of SPIE, vol. 2181, pp. 322-333. 1994.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.