A DE-DUPLICATION TOOL: IMPLEMENTATION OF DUPLICATE DETECTION ALGORITHMS

J.Shana*

Dr.T.Venkatachalam**

ABSTRACT

Data quality is an important issue in any data store especially when it is used for analysis. Poor quality of data would drastically affect the decision making process. Cleansing is an activity performed before data is loaded into the warehouse to enhance the quality and consistency of the data. This paper addresses one of the key data quality problem namely detection of duplicates. A duplicate detection tool is built that implements two algorithms namely Sorted Neighborhood method and Token based cleansing method. The tool was tested using student dataset of 118 records and showed an accuracy of 88.09%. This paper is a primitive analysis of how duplicates can be detected in a dataset.

Keywords- data quality, data cleansing, duplicate detection, data warehouse

* Department of MCA, Coimbatore Institute of Technology, Tamilnadu, India.

** Department of Physics, Coimbatore Institute of Technology, Tamilnadu, India.

INTRODUCTION

Explosion in the amount of data entered is leading to increased corporate attention on improving the quality and accuracy of business data. The data given as input for the data mining process should be of high quality in order for the results to be accurate and reliable. Before it is stored in warehouse or mined the data goes through a process called data cleansing [6][9]. Poor data quality can have a severe impact on the overall effectiveness of information systems. Data cleansing tools help in improving data consistency and accuracy. A TDWI survey published in a research report on data quality showed that customer and product data are the two main types of business entity susceptible to quality problems. Customer data involves variations in names, titles, phone numbers and addresses [10].Detecting and eliminating duplicated data is an important problem in the broader area of data cleansing and data quality. A duplicate value is the most serious kind of dirty data which leads to wastage of various resources. Many times the same logical real world entity has multiple representations in a relation due to data entry errors, varying conventions and other reasons. Hence a significant amount of time and money is spent on the task of detecting and eliminating duplicates.

1. EXPERIMENTAL EVALUATION

We developed a prototype of a de-duplication tool that implements the two algorithms, Sorted Neighborhood Method and Token Based cleansing algorithm.[5] The dataset used is a student dataset consisting of 118 records. The dataset consist of fields namely Roll Number, Name, Gender, Date of Birth, Father's name, Address, Phone Number.

2.1 Preprocessing

Preprocessing is one of the important steps in the design of data warehouse or prior to data mining. Preprocessing is done to standardize the data before implementing the algorithm to remove errors such as mistyped letters or abbreviations. For example, DOB formats such as 19-Dec-1978 or 19-12-1978 are converted into one standard format 19-12-1978. Abbreviations such as rd, st are converted into road and street respectively. We introduce two types of errors common in data warehouse namely equivalence errors and spelling errors [3]. Exact duplicate tuples are those for which all the fields are the same. If there is a mismatch in any one field then it is inexact duplicate. It is due to typographic errors or format mismatch in data.

2.2 Evaluation metrics

Precision and Recall metrics are used to evaluate the duplicate elimination algorithms [2]. Precision is ratio of identified duplicates to the original duplicates. Recall is the fraction of pairs of tuples an algorithm returns which are truly duplicates.

$$Recall = \frac{No.of\ duplicates\ identified}{Total\ no.of\ duplicates\ present} * 100$$

(1)

2. ALGORITHMS USED IN THIS STUDY

3.1 Sorted Neighborhood Method

This method involves the following phases.

Create Key: A key for each record in the list is computed by extracting relevant fields or portions of fields.

Sort Data: The records are sorted by using the key found previously. By sorting, duplicate records will come closer in the list. The effectiveness of this method highly depends upon the comparison key that is selected to sort the records.

Merge: A fixed size window is moved through the sequential list of records in order to limit the comparisons for matching records to those in the window. If the size of the window is 'w' records then every new record that enters that window is compared with the previous 'w-1' records to find matching records. The first record in the window slides out of it.

3.2 Token Based data cleansing method

This algorithm is implemented as follows:

Select Tokens: Select and rank 2 or 3 fields based on their record identifying abilities. Extract smart token for each selected field as follows. Form numeric, alphabetic or alphanumeric tokens after removing stop words and unimportant characters [4]. Table 1 gives the category of tokens used.

Table 1: Tokens for the dataset

Tokens
Numeric
Alphanumeric
Alphabetic

Sort Data: The records in the database are sorted by using the token found in the previous step. To detect the duplicates two ratios, Similarity Match Count (SMC) and Similarity Match Ratio (SMR) are calculated. They are calculated using the equation (2) and (3) as given below.

$$SMC = \frac{no\ of\ matching\ token\ fields}{no\ of\ token\ fields\ used} \quad (2)$$

$$SMR = 2 * \frac{no\ of\ common\ characrters\ in\ the\ 2\ tokens}{Total\ no\ of\ characters\ in\ the\ 2\ tokens} \quad (3)$$

SMC and SMR use the token fields and identify all pairs of records as duplicates using Table 2.

Table 2: SMC values for identifying the duplicates

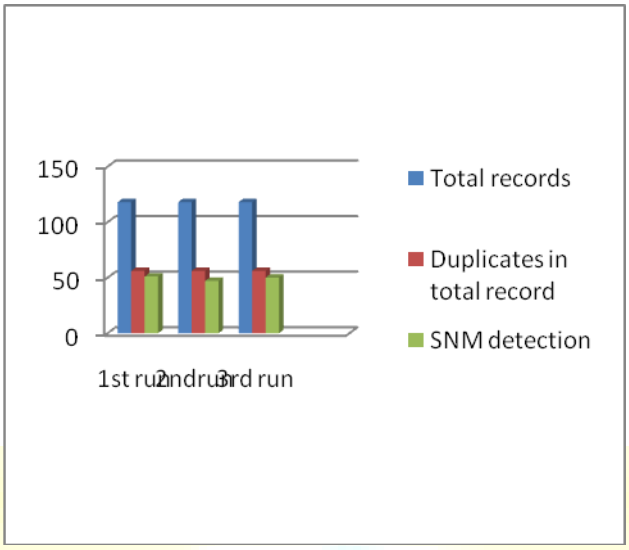
Match type	**SMC Values**
Perfect match	1.0
Near perfect match	Between 0.67 and 0.99
No Match	Less than 0.33
May be match	Between 0.33 and 0.66

Calculate SMR for may be match and the two tokens are a match if their SMR is greater or equal to 0.67.

3. EXPERIMENTAL RESULTS

4.1 SNM Method

Figure 1 shows the result of three runs made to detect duplicates using SNM Method.



Recall values for different runs of records in Sorted Neighborhood Method.

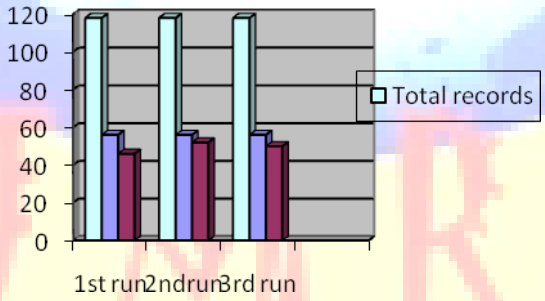
Recall value for 1st run = **91.07 %**

Recall value for 2nd run = **83.92 %**

False Positive error for 3rd run = 89.2**%**

Average recall value for SNM is **87.5%**

4.2 Token Based Data Cleansing Method



Recall values for different runs of records in Token Based data cleansing Method.

Recall value for 1st run= **82.14 %**

Recall value for 2nd run= **92.86 %**

Recall value for 3rd run= **89.28 %**

Average recall value for Token based data cleansing method is 88.09

Figure 3: Screen shot for SNM Method

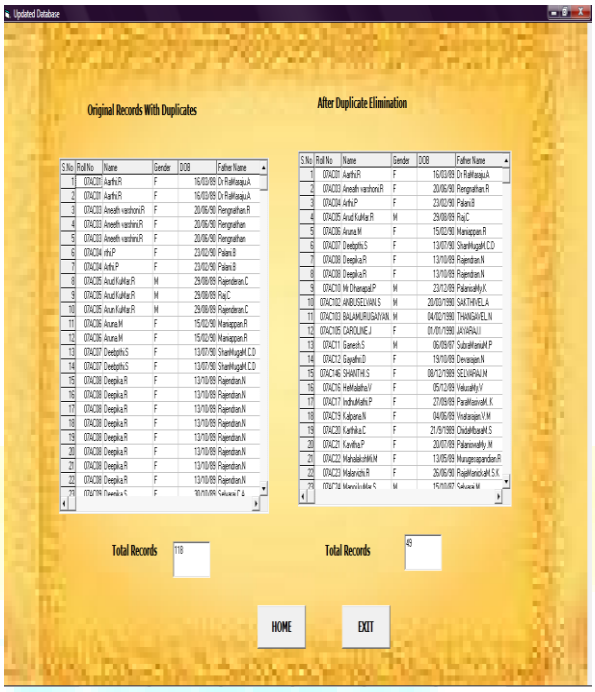
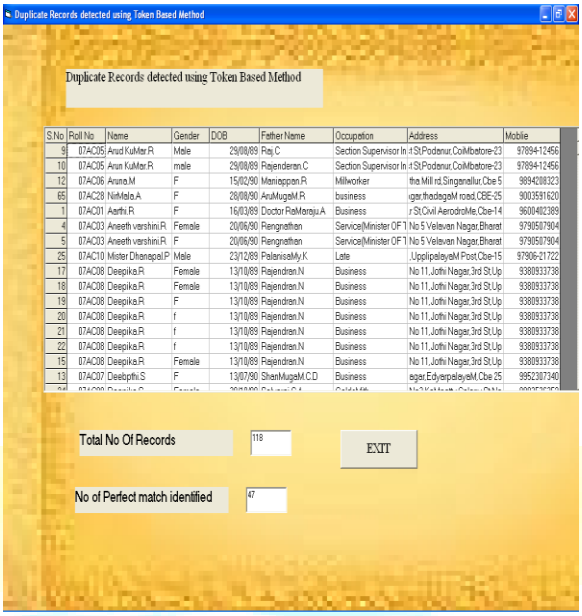


Figure 4:Screen shot for Token Based Method for single record.

Figure 5:Screen shot for Token Based Method



4. CONCLUSION

The data de-duplication tool is implemented successfully using Sorted Neighborhood Method and Token Based Data Cleansing Method. Both the algorithms were tested with the same dataset and the result analyzed. The tool was successfully tested for various runs. The analysis was done based on recall value and false positive error. SNM could detect and eliminate duplicates but it also identified false positives. Token based could detect duplicates with near match ad may be match also. Further processing for near perfect match records needs to be done on token based method to detect the duplicates very efficiently.

References

[1] Ramzi A. Haraty,Ralph Varjabedian Lebanese, ADD:Arabic Duplicate Detector- A Duplicate Detector ,IEEE June 2009

[2] Surajit Chaudhuri, Venkatesh Ganti, Rajeev Motwani, Robust Identification of Fuzzy Duplicates, IEEE, June 2008.

[3] M.Bilenko, R.Mooney ,On Evaluation and Training-set Construction for Duplicate Detection , Proceedings of the ACM SIGKDD workshop on Data Cleaning, Record Linkage and Object Identification, August 2003.

[4] Erhard Rahm and H. Hai Do. Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, December 2000

[5] Timothy Ohanekwu, C.I Ezeife, A Token Based Data Cleansing Technique for Data warehouse System, IJDWM, June 2003.

[6] Abdelkader Hameurlain, Rosine Cicchetti, Roland Traunmuller, Database and Expert System Applications, Proceedings of International Conference DEXA'97,1997

[7] Bing Chen,Beizhan Wang, Analysis and Solution of Data Quality in Data Warehouse of Chinese Materia Medica, Proceedings of International Conference on Computer Science, 2009.

[8] Vijayshankar Raman , Joseph M. Hellerstein, Potter's Wheel: An Interactive Data Cleaning System, Proceedings of the 27th International Conference on Very Large Data Bases, September , 2001

[9] R. Colin, Developing Universal Approach to Cleansing Customer and Product Data, BI research white paper from Business Object.

[10] Philip Russom, Taking Data Quality to the Enterprise through Data Governance, A White Paper.

[11] Ronald Forina, Data Equality –A Behind the Scene Perspective on Data Cleansing, http://www.dmreview.com/, March 2001