ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2014

A Simplified Framework for Data Cleaning and Information Retrieval in Multiple Data Source Problems

Agusthiyar.R, [1], Dr. K. Narashiman [2]

Assistant Professor (Sr.G), Department of Computer Applications, Easwari Engineering College, Chennai, India [1]

Professor & Director, AUTVS Centre for Quality Management, Anna University, Chennai, India [2]

ABSTRACT: Nowadays, data cleaning solutions are very essential for the large amount of data handling users in an industry and others. Normally, data cleaning, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. There are number of frameworks to handle the noisy data and inconsistencies in the market. While traditional data integration problems can deal with single data sources at instance level. But the data cleaning is especially required when integrating heterogeneous data sources and should be addressed together with schema-related data transformations. This paper proposed a framework to handle errors in heterogeneous data sources at schema level and this framework detecting and removing errors and inconsistencies in a simplified manner and improve the quality of the data in multiple data source of the company having different sources of different locations.

KEYWORDS: Data cleaning, Data quality, Attribute selection, Data warehouse

.

I. INTRODUCTION

Data cleaning process deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data [1]. For data cleaning in single data source can dealt with attribute selection or feature selection method to detecting and removing errors significantly. This method gives the quality data for the end users or business community for homogeneous data sources. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems data cleaning process is very crucial. Data warehouses require and provide extensive support for data cleaning. They load and continuously refresh huge amounts of data from a variety of sources so the probability that some of the sources contain "dirty data" is high. Furthermore, data warehouses are used for decision making, so that the correctness of their data is vital to avoid wrong conclusions. For instance, duplicated or missing information will produce incorrect or misleading statistics ("garbage in, garbage out"). Due to the wide range of possible data inconsistencies and the sheer data volume, data cleaning is considered to be one of the biggest problems in data warehousing.

II. RELATED WORK

During data cleaning, multiple records representing the same real life object are identified, assigned only one unique database identification, and only one copy of exact duplicate records is retained. A token-based algorithm for cleaning a data warehouse is the notion of "token records" was introduced for record comparison and the smart tokens are more likely applicable to domain-independent data cleaning, and could be used as warehouse identifiers to enhance the process of incremental cleaning and refreshing of integrated data. [4] Data cleaning is a process of identifying or determining expected problem when integrating data from different sources or from a single source. There are so many problems can be occurred in the data warehouse while loading or integrating data. The main problem in data warehouse is noisy data. The attribute selection algorithm is used for the attribute selection before the token formation. An attribute selection algorithm and token formation algorithm is used for data cleaning to reduce a complexity of data cleaning process and to clean data flexibly and effortlessly without any confusion. [5] Every attribute value forms

DOI: 10.15680/IJIRSET.2014.0308046

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2014

either a special token like birth date or an ordinary token, which can be alphabetic, numeric, or alphanumeric. These tokens are sorted and used for record match. The tokens also form very good warehouse identifiers for future faster incremental warehouse cleaning. The idea of smart tokens is to define from two most important fields by applying simple rules for defining numeric, alphabetic, and alphanumeric tokens. Database records now consist of smart token records, composed from field tokens of the records. These smart token records are sorted using two separate most important field tokens. The result of this process is two sorted token tables, which are used to compare neighbouring records for a match. Duplicates are easily detected from these tables, and warehouse identifiers are generated for each set of duplicates using the concatenation of its first record's tokens. These warehouse identifiers are later used for quick incremental record identification and refreshing. [6]

III. MULTIPLE DATA SOURCE PROBLEMS

Each source may contain dirty data and the data in the sources may be represented differently, overlap or contradict. This is because the sources are typically developed, deployed and maintained independently to serve specific needs. This results in a large degree of heterogeneity with respect to data management systems, data models, schema designs and the actual data. The main problems with respect to schema design are naming and structural conflicts. Naming conflicts arise when the same name is used for different objects (homonyms) or different names are used for the same object (synonyms). Structural conflicts occur in many variations and refer to different representations of the same object in different sources, e.g., attribute vs. table representation, different component structure, different data types, different integrity constraints, etc.

A main problem for cleaning data from multiple sources is to identify overlapping data, in particular matching records referring to the same real-world entity (e.g., customer). This problem is also referred to as the object identity problem [3], duplicate elimination or the merge/purge problem [2]. Frequently, the information is only partially redundant and the sources may complement each other by providing additional information about an entity. Thus duplicate information should be purged out and complementing information should be consolidated and merged in order to achieve a consistent view of real world entities.

A. DATA CLEANING IN MULTIPLE DATA SOURCE WITH SOCIAL ASPECT

The data quality problem is raised in government sector databases such as census data, voter id information, driving license data bases and personal id information. These databases have the errors like missing information, invalid data, data entry mistakes, duplicated records and spelling errors. When the government would decides to implement or distribute any welfare scheme benefits to the peoples, the duplicate records or noisy data leads wrong decision or the welfare scheme would not be reached every one. This framework is used to detect and remove noisy data in single and multi source databases and it is used to clean the data and get the quality data for good decision making in government sectors and other business organizations.

DOI: 10.15680/IJIRSET.2014.0308046

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2014

Problem	Column name	Noisy data	Solution	Clean data
Data entry problem	Place	Ch/enn,ai Mu$m@bai Cal&cat/ta De<lhi	Remove the noisy	Chennai Mumbai Calcutta Delhi
Name conflicts	Customer Name	John Abraham S Suresh kumar B Hari babu M	Split the name	First Name: John Abraham / Suresh kumar / Hari babu Last Name: S / B / M
Same name with different value	Gender	Male Female M F 1 0	Use Common format	Male Female
Check inheritance	DOB	1/January/1980 1-jan-1960 01/01/1970	Use Common date format	01/01/1980 01/01/1960 01/01/1970
Abbreviation and acronym	Organization	Govt, Pvt	Abbreviate the terms	Government Private
Address conflicts	Address	Address 133-A, Krishna Street, MGR Nagar Chennai – 600 078	Split the address column	Street: 133-A Krishna Street Area: MGR Nagar City: Chennai Zip code: 600078
Column name with space and underscore	Person_detail First name	Connecting two different meanings	Better to avoid space and underscore	Persondetail Firstname

Table 1: An Example of Multiple Data Source Problems

B. AN EXAMPLE OF MULTIPLE DATA SOURCE PROBLEM

The above Table 1: shows that the multiple data source problems occurred in different situations. Most frequently, error is made while data entry, but this will occur only human unconscious, it will rectify easily by this framework. The name conflict of the customer name will leads the data quality errors at the time of decision making in business organization. So, this type of error can detect and remove by this framework efficiently. The same name with different value, date format and short terms would be replaced by common format and abbreviation appropriately. The address conflicts could be split as Street, Area, City and Zip code format. The last row of the table shows that the column or attribute header problem with space and underscore. This could be better to avoid and give meaningful attribute header to the table. This table shows that some of the data source problems, when two tables or data sources would be integrated.

DOI: 10.15680/IJIRSET.2014.0308046

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2014

IV. PROPOSED FRAMEWORK

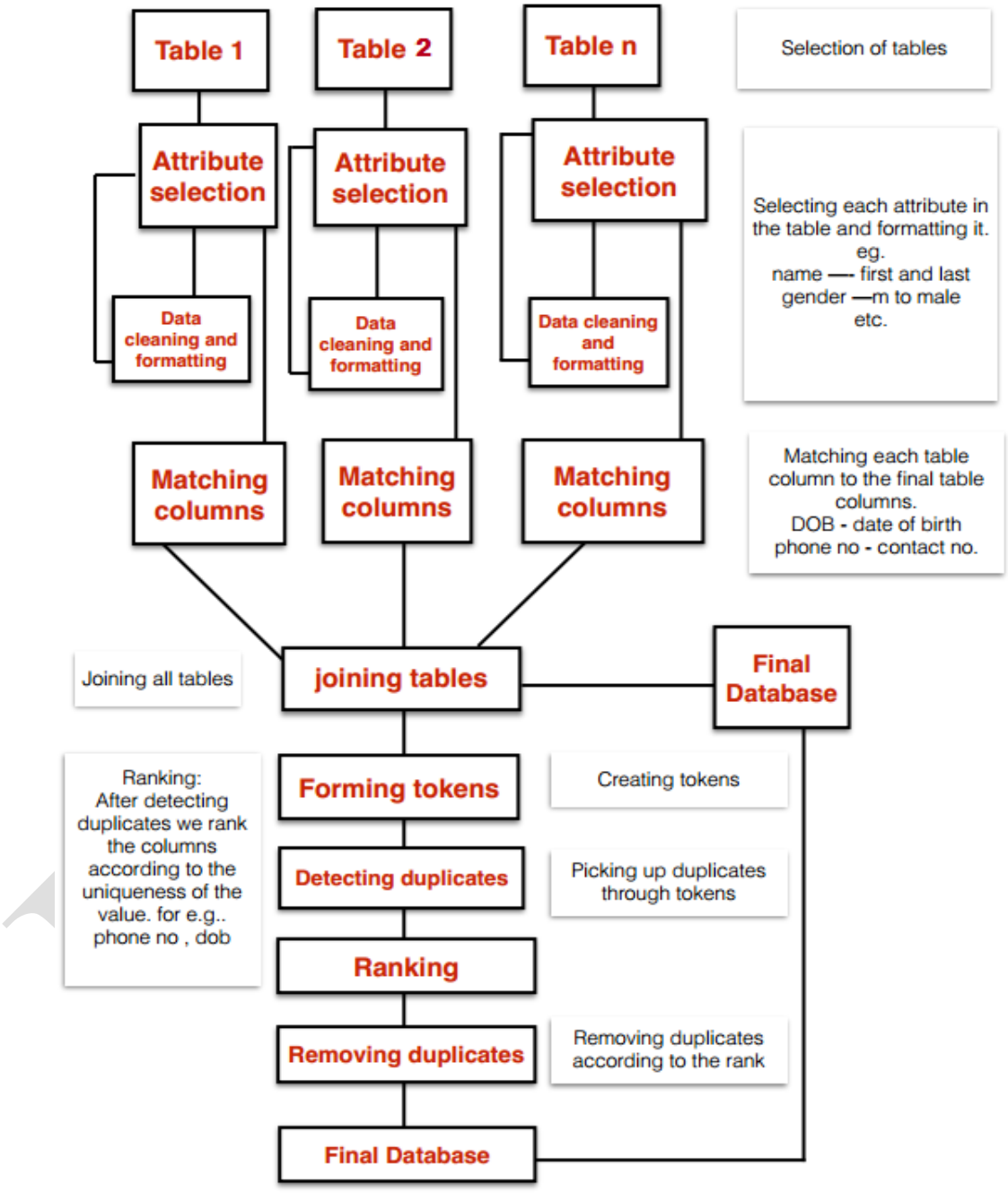

Figure 1: A Framework for Data cleaning in Multiple Data Sources

DOI: 10.15680/IJIRSET.2014.0308046

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2014

The Figure 1: shows that the framework implementation of the proposed research work. This step by step process of data cleaning methodology explains the user to detect the error and inconsistencies and then clean the noisy data efficiently.

1. Selection of tables from multiple data sources
2. Selecting each attribute in the table and formatting the attributes.
3. Match each table column to the final table column.
4. Joining all the tables.
5. Create Tokens and to find duplicates through tokens
6. Ranking: After finding duplicates, we rank the columns according to the uniqueness of the value
7. Removing duplicates according to the rank
8. Finally get the cleaned data then store in the database

1. SELECTION OF TABLES FROM MULTIPLE DATA SOURCES

In this step, data tables have been selected from various sources of single domain. Data coming from different origins and may have been created different times by different peoples and using different conventions. In this context, the question of deciding which data refers to the same real object becomes crucial. A company may have information about its clients stored in different tables, because each client buys different services that are managed by distinct departments. Selection of table from the multiple data source is the core research process.

2. SELECTING EACH ATTRIBUTE IN THE TABLE AND FORMATTING OF ATTRIBUTES

After selecting the table from the multiple data sources, each attribute of the table has been selected for data cleaning. In this step, attribute selection was done for the formation of attributes. For example, the Contact Name attribute may be split into two columns (firstname and lastname) and Gender attribute would be changed as m to male, f to female. Such kind of attribute formation is useful for the further research step.

3. MATCH EACH TABLE COLUMN TO THE FINAL TABLE COLUMN

When multiple tables are selected each table have different attributes in different name. A company may have information about its clients stored in different tables, because each client buys different services that are managed by distinct departments. Once it is decided to build a unified repository of all the company clients, the same customer may be referred to in different tables, by slightly different but correct names. This kind of mismatching is called the Object Identity problem. This problem would be solved by the common table and each table attribute should be matching to the common table attribute and it would be maintain a unified repository of all the tables which has been already selected.

4. JOINING ALL THE TABLES

After the formation of each table attributes to common table attributes, all the tables would be combined and then the common table format should be saved as final database. After creating unified repository of all the tables, the tokens will be created.

5. CREATE TOKENS AND TO FIND DUPLICATES THROUGH TOKENS

Creating and forming tokens is that the very important for data cleaning process. While forming tokens, the key value attributes acted as crucial role. This step makes use of the selected attribute field values to form a token. The tokens can be created for a single attribute field value or for combined attributes. For example, contact name attribute is selected to create a token for further cleaning process. The contact name attribute is split as first name, middle name and last name. The first name and last name is combined as contact name to form a token. Tokens are formed using numeric values, alphanumeric values and alphabetic values. The field values are split. Unimportant elements are removed [title tokens like Mr., Dr. and so on [6]. This step is eliminates the need to use the entire string records with multiple passes, for duplicate identification.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2014

6. RANKING: RANK THE COLUMNS ACCORDING TO THE UNIQUENESS OF THE VALUE

Select and ranks two or three fields that could be combined to most uniquely identify records. The condition for "field selection and ranking" is that the user is very familiar with the problem domain, and can select and rank fields according to their unique identifying power. We assume that the user in our banking domain Birth," "Name," and "Address "and ranked them in the order given [6]. After detecting the duplicates we rank the columns according to the uniqueness of the value. For example, Phone no, DOB, etc.,

7. REMOVING DUPLICATES ACCORDING TO THE RANK

Data Mining primarily works with large databases. Sorting the large datasets and data duplicate elimination process with this large database faces the scalability problems. This framework solves this problem by ranking the attributes according the dependency of the attribute.

8. GET THE CLEANED DATA THEN STORE IN THE DATABASE

The above step by step process of this framework gives the cleaned data of quality for further use. This quality data will store in to the data base and used for good decision making.

V. CONCLUSION

This framework is simplified the data cleaning process compare than other approaches previously used. These sequential steps are easy to handling data cleaning and information retrieval. This new framework consists of eight steps: Selection of tables, Selection of attribute, matching the column, joining all the tables, Forming tokens, Detecting duplicates, Ranking and removing the duplicates by ranking algorithm, and Merged to the database. This framework will be useful to develop a powerful data cleaning tool by using the existing data cleaning techniques in a sequential order.

ACKNOWLEDGEMENT

We would like to thank the reviewers for their comments and suggestions to develop the framework as data cleaning tool further.

REFERENCES

1. Erhard Rahm Hong Hai Do, Data Cleaning: Problems and Current Approaches, pp: 1-10
2. Hernandez, M.A.; Stolfo, S.J.: *Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem*. Data Mining and Knowledge Discovery 2(1):9-37, 1998.
3. Helena Galhardas - Daniela Florescu - Dennis Shasha - Eric Simon, An Extensible Framework for Data Cleaning.
4. Timothy E. Ohanekwu, C.I. Ezeife, A Token-Based Data Cleaning Technique for Data Warehouse Systems, PP:1-6
5. J. Jebamalar Tamilselvi , Dr. V. Saravanan, Handling Noisy Data using Attribute Selection and Smart Tokens, proceedings of International Conference on Computer Science and Information Technology 2008, PP: 771-774.
6. Christie I. Ezeife, Timothy E. Ohanekwu, Use of Smart Tokens in Cleaning Integrated Warehouse Data, International Journal of Data Warehousing & Mining (IJDW), Vol.1 No.2,PP: 1-22, Ideas Group Publishers, April-June 2005.
7. J. Jebamalar Tamilselvi , Dr. V. Saravanan, J. A Unified Framework and Sequential Data Cleaning Approach for a Datawarehouse, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.5,PP: 117-121 May 2008.

BIOGRAPHY

Agusthiyar. R. He received his Post Graduate MCA degree from Madurai Kamaraj University, Tamil Nadu, India in 2001, and M. Phil degree from Periyar University, Tamil Nadu, India in 2008 respectively. He is pursuing his Ph.D in Data mining at Anna University and he started his teaching profession from 2002 to till date and now he is working as Asst. Professor (Sr.G) in

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2014

Dept. of Computer Applications. His research interests include Data mining, and Artificial Intelligence.

Prof. Dr. K. Narashiman is the Director for AU TVS Center for Quality Management, Anna University Chennai. India. He is a well known Teacher, Trainer, Consultant and Researcher in the area of Quality Management. He has been recognized by the state of Bremen, Germany for implementing Quality Management System. Trained with scholarship in JAPAN, he is successfully propagating Quality Management to industrial and academic community.

DOI: 10.15680/IJIRSET.2014.0308046