Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Cleaning Heterogeneous Data & Removing Duplicates

Prajakta Naxane Pravin, Prof. Mangesh R. Wanjari
CSE, Ramdeobaba College of Engineering and Science,
Nagpur, Maharashtra, India

Abstract— Data, in everyday life, often times, we present with numbers, conclusions, studies or sometimes spelled out words. The purpose is to make data more deeply, make more rigorous conclusions derived from data. For example, the U. S. Census Bureau produces yearly statistical reports that track data between census year in 2011 they found that the real median household income was $ 50,054, a 1.5 percent decline from 2010. Data is everywhere. It is a collection of facts. There are different forms of data. Firstly, the qualitative data- the textual description. For example, the lotte has robust aroma, frothy appearance, and strong taste in a white cup. Second, the quantitative data- which represents a number. For example, the lotte contains 12 ounces of drink, 150 degrees Fahrenheit, and costs $ 4.95. There are two types of quantitative data. One is categorical data, that puts item into category. For example, we ve a set of shirts and among them we have categorized in different sizes like small, medium, large. The another is continuous data, which can take on any value in a certain range, for example, height. In this paper, we will discuss the data issues to be cleaned and removing duplicate methods with the help of new developed generic method using Jaccard Index method.

Keywords— Data Cleaning, De-Duplication, Jaccard Method of similarity.

I. INTRODUCTION

1.1 Data Mining:

The concept of data mining is drawing popularity in E- commerce, Business activities, etc. basically, we are an information economy where more and more data being generated in every aspect one can think of. For example, every time you swipe your grocery card, or when you try to get discount for your product, that is the data being downloaded from the database. On all transactions one does, there is some sort of data downloaded organizations or storing, processing and analyzing data.

Data Mining is incorporate of quality of messagesthat may include mathematical equations, algorithms, some problem methodologies or segmentation, classification, etc in industry sectors. If you have data, it is the application of powerful mathematical techniques, that extracts trends and patterns applies to Business Analytics, Healthcare, E-Commerce, Supply chain process, number of businesses, that are being mines with some techniques.

Any organization having data and has processes, can be analyzed with Data Mining, and results are extracting actual information from these data resources increases efficiency.

Data Mining topic concept is growing popularity because data make things to grow. Lets consider social networking sites like LinkedIn, Twitter, Facebook, etc, people are updating what they do, what they like, what hey are, etc. using services. All these data gathering and capturing i.e. extracting data sing data resources is nothing but Data Mining.

1.2 Data Cleaning:

Data cleaning is done in step by step manner. We iterate on first noticing and the correcting bad records. For example, numeric data can be denoted either in word format or in terms of set of numbers (like Two – 2, Three- 3, Four - 4,etc). It may happen that data items are designed in different manner which are different from our requirement. There might be some missing files or sometimes extra files can be the issue. For example, we search for document, files might not have the structure we expect at all. Let us consider one more example, in office for data entry, employee encoding various dates in U.S format i.e. 08/24/15 whereas system was expecting U.K format i.e. 24/08/15. For such data, we need data cleaning techniques.

Before cleaning, we need to collect data which can be done in two ways i.e. from primary as well as secondary sources. Data collected by researcher, through observation, case studies and questionnaire can be categorized under primary data. And the data that is already accessed, i.e. official statistics, weather information, government reports, previous research, etc. comes under secondary data collection.

What should you clean-up?

- Resolve duplicate information that you don't need.
- Correct spelling and punctuation error.
- Develop and enforce naming convention.
- Fill in data that is missing from your records.

Once you clean up your data, you are ready to prepare your import file.

There are various data issues we generally come across. These are as follows:

1. Sparse variables and cases : variables or cases with too many missing values. A variable will not proide beneficial information to the model if we do not have enough observations for it. A case with too many missing entries no longer adds tangible information to the model. These variables and cases should be removed.
2. Invariant variables : a variable has very little or no variance is nothing but an invariant variable. It does not add anything to the model building.
3. Duplicate records : data can artificially weight duplicated cases in the analysis. For example, a customer has applied for loan for multiple times and that is listed in dataset for multiple times. If this duplicates are not removed, this individual will have greater influence on the model. So, likely all but one record should be removed.

Why does it take time to Clean-up?

- Accuracy: cleaned data is accurate but dirty data is not. Let us consider an example, ABC and ABC-Labs is a name of same company written in two ways, similarly for Acme-NY and Acme-NYC. While selecting the company name, which one should be preferred by third person is an issue.
- Clean data is usable, dirty data ruins the efficiency and dulicate data makes existing data unusable.
- Clean data is creditable but dirty data is not trustworthy.
- Clean data is adoptable.

Steps of Data Cleaning

1. Plan

Our first and foremost step is to identify the set of data or information of a particular thing which is stable for making your working effort to be the best in all possible ways. When looking at data, it should focus on urgent data, and start small. The fields that are need to be recognized will be only one of its kind to the business and what information is exclusively looking for.

2. Analyze to cleanse

After some plans has been made of the main concern to be achieved, it is now very important to check for the data that exists already but with some missing values if any, the extraneous values that can be deleted, and the gaps in between them, if any. It needs some techniques or mechanisms to cleanse these exceptions manually. The amount of manual involvement is directly associated to the amount of satisfactory levels of data quality. For this, list of rules need to be built, then cleansing becomes much easier.

3. Implement Automation

Once the cleansing begins, it should be for standardization and cleansing the flow of new data as it enters the system by creating scripts or workflows. These can be run in real-time or in batch (daily, weekly, monthly) depending on how much data has been taken for working. These routines can be applied to new data, or to previously put-in data.

II. REVIEW OF LITERATURE

According to the data granularity, De-duplication strategies can be categorized into two main categories: file-level De-duplication and block-level De-duplication, which is nowadays the most common strategy. In block-based De-duplication, the block size can either be fixed or variable. Another categorization criterion is the location at which De-duplication is performed if data are de-duplicated at the client, and then it is called source-based De-duplication, otherwise target-based. In source-based De-duplication, the client first hashes each data segment he wishes to upload and sends these results to the storage provider to check whether such data are already stored: thus only "not de-duplicated" data segments will be uploaded by the user on the cloud. While De-duplication at the client side can achieve bandwidth savings, it unfortunately can make the system vulnerable to side-channel attacks whereby attackers can immediately discover whether a certain data is stored or not. On the other hand, by de-duplicating data at the storage provider.

Many people now store huge amount of personal and corporate data on laptops or home computers. These often have poor connectivity, and are susceptible to theft or hardware failure...Below there is brief about the papers which we referred for our project.

- ***Paper by Nidhi Choudhary titled " A Study over Problems and Approaches of Data Cleansing/Cleaning"***
  This paper defines what data cleansing is and the elements that data contains. Data cleaning process are briefly described i.e. planning, analyzing for cleaning, implementing automation, appending missing data, and monitoring data. It also includes significance of data quality and and the challenges arising while cleaning data. Various approaches for cleaning data are proposed. We came to a conclusion that Data cleaning is not only useful for data warehousing but also it is beneficial for query processing on heterogeneous data sources like in web-based information systems [Web-Designing].
- ***Paper by Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios on "Duplicate Record Detection: A Survey"***
  We cover similarity metrics that are commonly used to detect similar field entries, and we present an extensive set of duplicate detection algorithms that can detect approximately duplicate records in a database. We also cover multiple techniques for improving the efficiency and scalability of approximate duplicate detection algorithms. We conclude with coverage of existing tools and with a brief discussion of the big open problems in the area. In this survey, we have presented a comprehensive survey of the existing techniques used for detecting non-identical duplicate entries in database records.

- ***Paper by Srivatsa Maddodi, Girija V. Attigeri, Dr. Kaarunakar A. K. titled "Data Deduplication Techniques and Analysis"***
  It provided an architecture consisting of twin clouds for securely outsourcing of user private data and arbitrary computations to an untrusted commodity cloud. Privacy aware data intensive computing on hybrid clouds - Zhang et al also presented the hybrid cloud techniques to support privacy-aware data-intensive computing. We used public cloud of elastic infrastructure.
- ***Paper by Shital Gaikwad, Nagaraju Bogiri titled "A Survey Analysis on Duplicate Detection in Hierachical Data"***
  Different kind of errors adjust data superiority from the heterogeneous domains. As an alternative, evaluate all objective representations by means of a probably composite like method, to identify whether tht object is real world or not. This paper has given detailed survey analysis and groundwork on duplicate detection in hierarchical data.
  Pruning algorithm has been applied to check for the similarity proportion between objects. This survey paper is useful for doing research in Duplicate Detection in Hierarchical Data or XML Data. And the techniques in XML data for detecting duplicated records are being referred.
- ***Ranak Ghosh, Sujay Halder, Soumya Sen on "An Integrated Approach to Deploy Data Warehouse in Business Intelligence Environment"***
  This paper has given Proposed Architecture to Integrate Data Warehouse with Business Intelligence. The details of all the contents in architecture with advantages and shortcomings in them. It has been also proposed how to overcome those drawbacks for future use in brief.

III. COMPARISON

Sr. No.	*Paper*	*Pros*	*Future Scope*
1.	*Efficient and Effective Duplicate Detection in Hierarchical Data*	*Proposed algorithm is capable to accomplish elevated precision and remember scores in a variety of data sets.*	*To evaluate XML objects along with application of machine learning techniques, increase scope of the Bayesian Network model creation algorithm.*
2.	*An Effective Change Detection Algorithm for XML Documents*	*combine key XML structure characteristics and standard tree-to-tree correction techniques.*	*To enhance the time efficiency of the improved XDiff algorithm.*
3.	*Structure-aware XML Object Identification*	*If the diameters of the clusters is small as compared to distance between the clusters then proposed algorithm is robust.*	*To vigorously identifying q-grams of the cluster evaluate algorithm.*
4.	*XML Documents in Highly Dynamic Applications*	*Proposed algorithm in this paper having the capacity to formulate use of information from the structure and the content of XML documents.*	
5.	*Data Cleaning: Problems and Current Approaches*	*The process of data cleaning is used to identifying, take out errors and incompatibility from the available data to increase the data superiority.*	*Enhancing the design and implementation of the best language approach for supporting both schema and data transformations.*
6.	*Domain-Independent Data Cleaning via Analysis of Entity-Relationship Graph*	*Proposed algorithm is capable to recognize the distinctive entity to which every description corresponds*	*This confronts, make the approach a plug-and-play solution, and also improve accuracy and efficiency.*

IV. WORKING

Data cleaning is nothing but abolition of flawed data caused by disagreement, inconsistency, keying mistakes, missing bits, etc used mainly in databases. BI is a set of learning or developing skills to cram the past and come up with the strategies to improve the future. And the resulting information is helpful for commercial executives, business managers and other end users to make more well-versed business decisions.

We have considered employees data with various attributes. To remove duplicates in the data we have used Jaccard method to compare all the attributes in our data. Jaccard index is a statistic which is used for comparing similarity and diversity of samples sets. Jaccard coefficients measures similarity between finite sampled sets and is defined as size of the intersection divided by the union of samples sets. Jaccard index is also called as Jaccard similarity coefficient.

Earlier, we have compared the data attributes in a manner as an entire row was compared with all the other rows and using jaccard coefficient, we were detecting and removing the duplicate data from our dataset. But, to make it more efficient, we then estimated the generic method to compare duplicate data. The generic flow chart is being explained below:

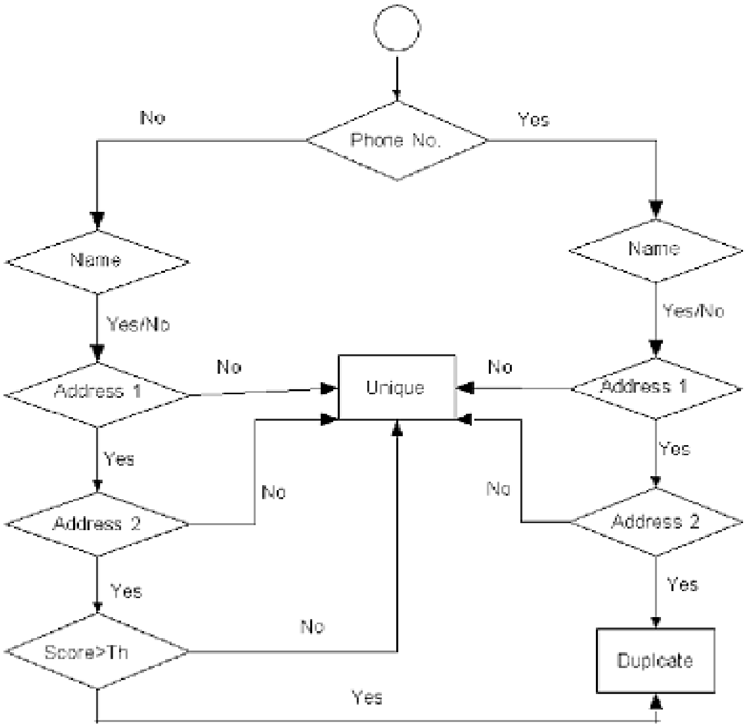

Figure 1: Process of comparing attributes

While entering name, it may happen that some may enter name with incorrect spelling or some can have name like A. B. or Amitabh Bachchan, both can either be same or different. So, to find out duplicates in them, we have worked on such type of data.

V. RESULT

The original and duplicate data are being differentiated using Jaccard coefficient method on the dataset. And the generalized method used to compare data is also efficient. This method takes less time. Data cleaning process helps in poor data accuracy, manual data entry error, CRM/ERP Integration effectiveness, Migration of legacy system & database, Limited customer insight, Acceleration of data dependant projects.

VI. CONCLUSION

This paper contains the details of data cleaning approaches, their advantages and shortcomings to be performed on heterogeneous data. On detailed study**,** we came to a conclusion that Data cleaning is not only useful for data warehousing but also it is favorable for query processing on heterogeneous data sources like in web-based information systems [Web-Designing].

REFERENCES

[1] Nidhi Choudhary, " A Study over Problems And Approaches of Data Cleansing/ Cleaning", *IJARCSSE, Volume 4, Issue 2, February 2014.*

[2] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios, "Duplicate Record Detection: A Survey", *IEEE Transaction on Knowledge and Data Engineering Vol 19 No.1, January 2017.*

[3] Srivatsa Maddodi, Girija V. Attigeri, Dr. Kaarunakar A. K. , "Data Deduplication Techniques and Analysis", *Third International Conference on Emerging Trends in Engineering and Technology.*

[4] Shital Gaikwad, Nagaraju Bogiri, "A Survey Analysis On Duplicate Detection in Hierarchical Data", *International Conference on Pervasive Computing (ICPC)*

[5] Ranak Ghosh, Sujay Halder, Soumya Sen, "An Integrated Approach to Deploy Data Warehouse in Business Intelligence Environment".

[6] Soumya Sen, Ranak Ghosh, Debanjali Pal, Nabendu Chaki, "Integrating Related XML Data into Multiple Data Warehouse Schemas". Int'l wrokshop on Software Engineering and Applications, ISSN: 2231-5403, 2012.

[7] Cohen, W.: *Integration of Heterogeneous Databases without Common Domains Using Queries Based Textual Similarity*. Proc. ACM SIGMOD Conf. on Data Management, 1998.

[8] 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.

[9] Lee, M.L.; Lu, H.; Ling, T.W.; Ko, Y.T.: *Cleansing Data for Mining and Warehousing*. Proc. 10th Intl. Conf.Database and Expert Systems Applications (DEXA), 1999.

[10] Milo, T.; Zohar, S.: *Using Schema Matching to Simplify Heterogeneous Data Translation*. Proc. 24th VLDB, 1998.

[11] Luı´s Leita˜ o, Pa´ vel Calado, Melanie Herschel "Efficient and Effective Duplicate Detection in Hierarchical Data", Knowledge and Data Engineering, IEEE Transactions, Volume: 25, Issue: 5, ISSN: 1041-434, May 2013.

[12] Yuan Wang David J. DeWitt Jin-Yi Cai "Local X-Diff: An Effective Change Detection Algorithm for XML Documents" Data Engineering, 2003. Proceedings. 19th International Conference, ISBN: 0-7803-7665, March 2013.

[13] Diego Milano, Monica Scannapieco, Tiziana Catarci, "Structure-aware XML Object Identification", in' CleanDB', 2006.

