Scientific Journal of Impact Factor (SJIF): 5.71
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 10, October -2018

A DUPLICATE DETECTION MECHANISM FOR DATA WAREHOUSING

Ms. Kanchan Chande[1], Prof. Vijay Shelake[2]

[1]PG Scholar, Department of Computer Engineering, ARMIET, Mumbai, India
[2] Assistant Professor, Department of Computer Engineering, YTCEM, 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. So for accurate data, improving the data quality is compulsory. Sometimes data may include text errors, quantitative errors or duplication of the data. Linking of data is an important component in the process of finding duplicates. The proposed algorithm deal with linkage mechanism for detecting duplicates in data warehousing. It performs better and gives higher level of results as compared to existing techniques of duplicate detection for data warehousing. Hence, the improved technique will helps the users to get better quality of data.*

Keywords— *Data Cleaning, Duplicate Detection*

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.

II. LITERATURE SURVEY

The various approaches for duplicate detection in data warehousing are as follows:

- **HADCLEAN: A Hybrid Approach to Data Cleaning in Data Warehouses**

Data Cleaning is an important part of the data warehouse management process. It is difficult process as many different types of unclean data (bad data, incomplete data, typos, etc) can be present. Generally data is clean or dirty is highly dependent on the nature and source of the raw data. Thus, it is clean the data using blocking algorithms, phonetic algorithms, etc. Thus the analysis establishes the superiority of the proposed algorithm in automating the data cleaning process[2].

- **Transitive Closure of Data Records**

Data quality improvement has become a critical issue for many companies and organizations because poor data quality degrades organizational performance whereas improved data quality results in cost saving and customer satisfaction. Activities such as identify and removing “duplicate” database records from a single database, and correlating records from different databases that identify the same real world “entity” are used routinely to improve data quality[3].

- **A Parallel and Distributed Approach for Finding Transitive Closures of Data Records**

The objective of finding transitive closures is to reduce the number of records to be considered in the second step, from a whole data source having hundreds of millions to billions of records to the range of hundreds to thousands. To process hundreds of millions to billions of records, an efficient approach is essential that works in distributed environment. As a part of this approach, this paper presents an efficient distributed algorithm for solving distributed transitive closure problem on large data sets[4].

- **Personalized Spell Checking using Neural Networks**

In this paper they present a proof of concept spell checking system that is able to intrinsically avoid many of these problems. The typist performed the actual corrections that provide the basis for error detection. These corrections are used to train a feed-forward neural network therefore if the same error is remade, the network can flag the offending word as a possible error[5].

- **Big Data Cleaning**

The existing data cleansing algorithms through improved sorting algorithm or using better data to quickly calculate the equivalence class, the time complexity is reduced, but can only be dealt with in the small data sets of main memory.

This dissertation focuses on Cloud Computing data centre, policy of replica, and scheduling mechanism, and makes deep research around such issues[6].

- **Adaptive Windows for Duplicate Detection**

The Sorted Neighbourhood method is a standard algorithm, but due to the fixed window size, it cannot efficiently respond to different cluster sizes within a data set. In this paper we have examined different strategies to adapt the window size, with the Duplicate Count Strategy as the best performing. We have proven that with a proper (domain- and data-independent!) threshold, DCS++ is more efficient than SNM without loss of effectiveness[7].

- **Name Matching in Record Linkage**

The authors use the largest collection of genealogy person records in the world together with user search query logs to build name matching models. The procedure for building a crowd-sourced training set is outlined together with the presentation of this method. They cast the problem of learning alternative spellings as a machine translation problem at the character level[10].

III. PROPOSED SYSTEM ARCHITECTURE

The block diagram of proposed system describes the three modules. The first part is training, then testing module and final part is producing output.

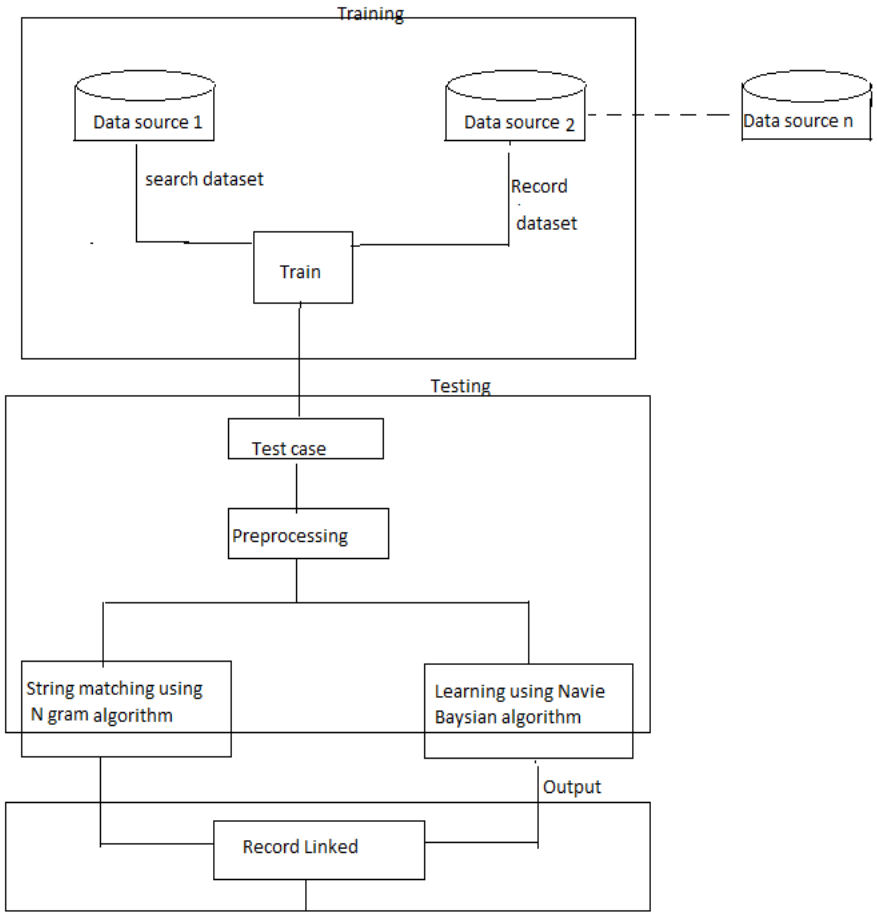

Fig. 3.1 Block diagram of Duplicate Detection for Data Warehousing

The datasets containing records and search datasets having some data which will test before training of those data and then matching record are created. After training using given dataset(Record and Search) a set of matching record are created and then it shown using N-gram. Suppose we take testing 100 and testing 200 dataset, then it will checked first record of testing 100 with first record of testing 200 , then again check first record of testing 100 with second record of testing 200. It process continues up to first record of first dataset totally check all of records of second dataset. In pre-processing, all letters of given name is converted into lower case and it removes any special characters or numerical. The given string is matching using N-gram and learning using Naive Bayesian algorithm.

In analysis part, actual records and system find records are analysed using N-gram and Naive Bayesian method. In this, no. of record tested, record merging, how many duplicate records are tested and then precent of record found.

IV. CONCLUSION AND FUTURE WORK

Data pre-processing is a necessary step for various data warehousing tasks. The proposed algorithm deal with linking mechanism for detecting duplicates in data warehousing. It performs better and gives higher level of results as compared to existing techniques for duplicate detection. Thus, the improved technique will helps the users to get better quality of data.

The task of finding duplicates has been effectively handled by the proposed system. The proposed system can be extended for big data pre-processing. The data pre-processing along with duplicate detection mechanism can be utilized for real world data from multiple data sources for improving big data quality.

REFERENCES

[1] Prerna S. Kulkarni and J W Bakal, "Hybrid Approaches for Data Cleaning in Data Warehouse", International Journal of Computer Applications, Vol. 88 – No.18, February 2014

[2] Arindam Paul, VaruniGanesan, JagatSeshChalla, Yashvardhan Sharma, HADCLEAN: A Hybrid Approach to Data Cleaning in Data Warehouses, Department of Computer Science & Information Systems, 978-1-4673-1090, 2012.

[3] Johnson Zhang, RoopaBheemavaram, Wing NingLi , "Transitive Closure of Data Records: Application and Computation", University of Central Arkansas, March 3, 2006.

[4] Johnson Zhang, RoopaBheemavaram, Wing Ning Li , "A parallel and Distributed Approach for finding Transitive Closure of Data Records ", University of Central Arkansas, March 3, 2006.

[5] Tyler Garaas, Mei Xiao, and Marc Pomplun," Personalized Spell Checking using Neural Networks".University of Massachusetts Boston, 100 Morrissey Blvd., Boston, MA 02125-3393, USA.

[6] Zhang Feng , XueHui-Feng, Xu Dong-Sheng , Zhang Yong-Heng,YouFei, "Big Data Cleaning Algorithms in Cloud Computing", Yulin University, Yulin, China, Vol. 9, Issue 3, July 2013.

[7] UweDraisbach , Felix Naumann , SaschaSzott , Oliver Wonneberg , "Adaptive Windows for Duplicate Detection" ,International Conference on Data Engineering, 1084-4627/12, 2012.

[8] Erhard Rahm and Hong Hai Do, "Data Cleaning: Problems and Current Approaches", University of Leipzig, Germany, 2000.

[9] 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.

[10]Jeffreysukharev, Leonid Zhukov, and Alexandrin propescul,"parallel corpus approach for name matching in record linkage",2014 IEEE International conference on Data Mining,1550-4786/14,2014.

