Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 11, November 2017, pg.38 – 42

An Analysis on Removal of Duplicate Records using Different Types of Data Mining Techniques: A Survey

P. Selvi
Assistant Professor, KG College of Arts and Science, Coimbatore
pselvi99@yahoo.com

Abstract: ***In the current period rapid improvement of information technology provides to the need of large volume of storage to storing the dataset. From different data mart, most of the data warehouse access ability of data, by reason of this there is a prospect of latency of high record duplicates. Uncounted systems are mainly troubled by the habitation of duplication in the database which provides to the problem like slow performance, degradation of data quality, waste of data storage and high operating cost. In enlargement assurance of duplicates provides to the issue of misleading, the system reports as fails to recover the proper data for the entanglement of query and the time complication is big. The above said issues can be concluding by the process of record deduplication which is the one of the necessary task in data preprocessing. This process concluded in data cleaning and replica free repositories which allow recovering increased higher quality information. Record Deduplication is the process of analyzing and removing records in data storage which indicate to the same entity of different sources of data. Record Deduplication is necessary while linking entity based datasets that permit or not permit to share a frequent accessory. This paper discusses about the elaborate introduction to data deduplication. In this paper also granted the comprehensive study of different existing techniques for removal of data replication using deduplication.***

Keywords: Deduplication, Record, Mining, Replica, Repository.

1. INTRODUCTION

For every organization database is a most important origin which is collected from various types of origins. All different resource has diverse explanation for identical entity, which lead to replica in repository. Thus big investments are made by different companies to clear the replica from the repository. Data mining is the contemporary technology which clipped the useful instruction required by the company for getting an improved verdict. This is the step of KDD (Knowledge Discovery in Databases) method. Fayyad et.al. states that, "KDD is the method of identifying a accurate, potentially useful and finally comprehensible constitution in data" [1]. In the KDD procedure, Preprocessing is the data cleaning stage where the excessive information's to be isolated. The data cleaning is the method of identifying and rectifying the records from the database [2]. It includes parsing, renovation, and refusal of duplicates. One common approach to avert duplication, is the record duplication (also described as record linkage or data linkage) [3].

The record deduplication is the method of finding corresponding entity beyond various data sources. There are different techniques to record deduplication. They are:

1. Adhoc or domain information techniques – based on domain information and uses declarative languages.
2. Training based techniques – based on supervised or semi supervised learning.

2. RELATED WORK

A literature survey is cheerful carried out in sequence to examine the history of the present work, which assist to finding out the fault in the achievable and guides on which unresolved problems can work out. The succeeding sections analyze various references that explain about several topics related to collective act.

2.1 Divide and conquer method based deduplication:

Bilal khan et al. proposed an approach for duplicate record detection and removal. In this approach they transform the attributes of data into numerical form as first step. Next, the numeric form is declared to create clusters by using K-Means clustering algorithm. The main thing of clustering detects the number of comparisons. After using the divide and conquer technique is used to extending in same direction with these clusters for identification and removal of duplicated records. This divide and conquer technique determine all types of duplicated records like fully duplicated records, erroneous duplicated records and partially duplicated records. The preceding technique is only used for single table instead of multiple sorted tables. The measuring is used by the terms of underperformance like true positives, false positives, false negatives, precision, recall and F-Score [12].

2.2 An improvised Technique:

Manasa Veena Dokku et al. proposed deduplication is the key operation in data integration from multiple data sources. Duplicate record detection is important for data preprocessing and cleaning. Record linkage is the process of matching records from several databases that refer to the same entities. When applied on a single database, this process is known as deduplication. To achieve higher quality information and more simplified data representation, data preprocessing is required. Data cleaning is one among the data preprocessing steps. Removing duplicate records is crucial step in data cleaning process. Now-a-days in current databases, removing the duplicate records is more complex. This paper presents an analysis of record deduplication techniques and algorithms that detect and remove the duplicate records. In this paper, we proposed a new methodology which is divided two phases. They are: training phase and duplicate detection phase. These are again divided into four steps: (1) Similarity computation for all pair of records, (2) Computing feature vectors, (3) New similarity formulae generation and (4) Duplicate detection using the new similarity formulae. Our proposed system is very effective and efficient, when compared with remaining duplicate detection techniques [4].

2.3 Deduplication using Febrl system:

Peter Christen suggested record or data linkage is an important enabling technology in the health sector, as linked data is a cost effective resource that can help to improve research into health policies, detect adverse drug reactions, reduce costs, and uncover fraud within the health system. Significant advances, mostly originating from data mining and machine learning, have been made in recent years in many areas of record linkage techniques. Most of these new methods are not yet implemented in current record linkage systems, or are hidden within ‘black box’ commercial software. This makes it difficult

for users to learn about new record linkage techniques, as well as to compare existing linkage techniques with new ones. What is required are flexible tools that enable users to experiment with new record linkage techniques at low costs. This paper describes the Febrl (Freely Extensible Biomedical Record Linkage) system, which is available under an open source software licence. It contains many recently developed advanced techniques for data cleaning and standardization, indexing (blocking), field comparison, and record pair classification, and encapsulates them into a graphical user interface. Febrl can be seen as a training tool suitable for users to learn and experiment with both traditional and new record linkage techniques, as well as for practitioners to conduct linkages with data sets containing up to several hundred thousand records [5]. Record linkage is the problem of identifying similar records across different data sources. The similarity between two records is defined based on domain-specific similarity functions over several attributes. De-duplicating one data set or linking several data sets are increasingly important tasks in the data preparation steps of many data mining projects. The aim is to match all records relating to the same entity. Different measures have been used to characterize the quality and complexity of data linkage algorithms, and several new metrics have been proposed. An overview of the issues involved in measuring data linkage and de-duplication quality and complexity. A matching tree is used to overcome communication overhead and give matching decision as obtained using the conventional linkage technique. Developed new indexing techniques for scalable record linkage and de-duplication techniques into the febrl framework, as well as the investigation of learning techniques for efficient and accurate indexing[6].

2.4 Genetic Approach on Record Deduplication:

In this article L. Chitra Devi et al. going to discuss about how genetic programming can be used for record deduplication. Several systems that rely on the integrity of the data in order to offer high quality services, such as digital libraries and ecommerce brokers, may be affected by the existence of duplicates, quasi-replicas, or near-duplicates entries in their repositories. Because of that, there has been a huge effort from private and government organizations in developing effective methods for removing replicas from large data repositories. This is due to the fact that cleaned, replica-free repositories not only allow the retrieval of higher-quality information but also lead to a more concise data representation and to potential savings in computational time and resources to process this data. In this work, we extend the results of a GP-based approach we proposed to record deduplication by performing a comprehensive set of experiments regarding its parameterization setup. Our experiments show that some parameter choices can improve the results to up 30%. Thus, the obtained results can be used as guidelines to suggest the most effective way to set up the parameters of our GP-based approach to record deduplication [7].

Shital Gujar, Avinash Shrivash proposed, in the upcoming growing of technology the use of databases are very high. As the use of databases grows higher the dirty data on the other side is the biggest disadvantage with the databases. Dirty data can contain such mistakes as spelling or punctuation, incorrect data associated with a field, incomplete or outdated data or even data that is duplicated in the database. Various data cleaning software's are used to remove the dirty data. In our paper we are proposed a concept of Genetic programming approach to record Deduplication that combines several different pieces of evidence extracted from the data content to find a Deduplication function that is able to identify whether two entries in a repository are replicas or not. In addition, our genetic programming approach is capable of automatically adapting these functions to a given fixed replica identification boundary. We are applying this genetic programming approach for the blood bank database management to deduplicate the records [8].

J. R. Waykole, Prof. S. M. Shinde, suggested, in today's world, by increasing the volume of information available in digital libraries, most of the system may be affected by the existence of replicas in their warehouses. This is due to the fact that, clean and replica-free warehouse not only allow the retrieval of information which is of higher quality but also lead to more concise data and reduces computational time and resources to process this data. Here, we propose a genetic programming approach along with hash-based similarity i.e, with MD5 and SHA-1 algorithm. This approach removes the replicas data and finds the optimization solution to deduplication of records [9].

2.5 PSO Algorithm Based Deduplication:

K. Deepa et al. [10] proposed a heuristic global optimization method called Particle Swarm Optimization algorithm for record deduplication. The essential target behind the PSO algorithm is swarm. It includes two different phases namely training and duplicate record detection phase. It uses cosine similarity and levenshtein distance to obtain matches between the record pairs. The result data forms feature vectors is to represent the elements, which wish duplicate checking. Duplicate detection identifies the duplicates from the feature vectors means of PSO algorithm. It surpass genetic algorithm by providing high accuracy [10].

2.6 Artificial Bee Colony using Deduplication:

Lalitha.L et al. suggested deduplication is the key operation in data integration from multiple data sources. To achieve higher quality information and more simplified data representation, data preprocessing is required. Data cleaning is one among the data preprocessing steps. Data cleaning includes the process of parsing, data transformation, duplicate elimination and statistical methods. If two records represent the same real world entity then it is called duplicated records. The problem of detecting and eliminating duplicate records is called record deduplication. This paper presents a new combination algorithm called Tabu Artificial Bee Colony. It improves the optimization performance in detecting and removing the duplicate records [11].

3. CONCLUSION

A survey of the existing record deduplication techniques and frameworks is completed here. Deduplication and record linkage is a necessary step in data integration. By this survey, it is feasible to achieve that the existing algorithms wish much more memory for deduplication. This method also takes more time absorbing process. In forthcoming deduplication algorithm can be designed for detecting the lot of comparisons among the records such that it detects time absorbing and appropriate low memory space.

References

[1] Arun.K.Pujari, Data Mining Techniques.
[2] Lalitha. L, Maheswari. B, Dr. Karthick. S, "A Detailed Survey on various Record Deduplication Methods", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 8, October 2012, pp. 160-163.
[3] http//en.wikipedia .org/wiki/Record_linkage.
[4] Manasa Veena Dokku #1, A .Srinivasa Reddy#2 "An Improvised Technique for Record Deduplication", INTERNATIONAL JOURNAL FOR DEVELOPMENT IN COMPUTER SCIENCE & TECHNOLOGY ISSN-2320-7884, VOLUME-1, ISSUE-IIIIS.
[5] Peter Christen, "Febrl – A Freely Available Record Linkage System with a Graphical User Interface", Department of Computer Science, The Australian National University Canberra ACT 0200, Australia.
[6] K.Mala, Mr. S.Chinnadurai, "Efficient Record De-Duplication Identifying Using Febrl Framework", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 10, Issue 2 (Mar. - Apr. 2013), PP 22-27.

[7] L.Chitra Devi, S.M.Hansa, Dr.G.N.K.Suresh Babu, "A Genetic Programming Approach for Record Deduplication", International Journal of Innovative Research in Computer and Communication Engineering, ISSN (Print) : 2320 – 9798 1SSN (Online): 2320 – 9801 Vol. 1, Issue 4, June 2013.
[8] Shital Gujar, Avinash Shrivas, "Detection Of Duplicate Record Using Genetic Algorithm",SHITAL GUJAR et al. DATE OF PUBLICATION: DEC 20, 2014, ISSN: 2348-4098 Vol 2 Issue 8 Nov-Dec 2014.
[9] Miss. J. R. Waykole, Prof. S. M. Shinde, "A Survey Paper on Deduplication by Using Genetic Algorithm Alongwith Hash-Based Algorithm", Journal of Engineering Research and Applications www.ijera.com ISSN : 2248-9622, Vol. 4, Issue 1(Version 1), January 2014, pp.343-346.
[10] K. Deepa et al., "Record Deduplication using Particle swarm optimization Algorithm", European journal of scientific research ISSN 1450-216X., vol. 80,no. 3, pp. 366-378,2012.
[11] Lalitha.L, Sivaparthipan C.B, K.Sathyaseelan, Dr. Kalaikumaran.T, "Tabu Artificial Bee Colony Algorithm Based Record Deduplication in Data Mining Approach", International Journal of Computer Science and Technology Vol. 5, Issue 1, Jan - March 2014, ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print).
[12] Bilal Khan, Azhar Rauf, Sajid H. Shah and Shah Khusro, "Identification and Removal of Duplicated Records", World Applied Sciences Journal 13(5): ISSN 1818-4952, pp.1178-1184, 2011.
[13] Aswanandini.S, "Survey of a Multi-Agent System for Distributed Data Mining", International Journal of Innovative Computer Science & Engineering, Vol. 4., Issue 3, ISSN: 2393-8528.

