ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 5, May 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Survey of Data Duplication Detection and Elimination in Domain Dependent and Domain-Independent Databases

Akshata Anil Dagade[1]
Computer Engineering
Vishwakarma Institute of Information Technology
Pune, Maharashtra – India

Prof. Manisha P. Mali[2]
Computer Engineering
Vishwakarma Institute of Information Technology
Pune, Maharashtra – India

Prof. Narendra P. Pathak[3]
Computer Engineering
Vishwakarma Institute of Information Technology
Pune, Maharashtra – India

Abstract: ***A major problem of today is the record linkage. The record linkage problem occurs when there are multiple representations of the same real-world entity. In record linkage problem, record matching is done because of that we can find out duplicate records from the system, and the duplication detection is the process of finding duplicate records which having different representations. Data integration is the important step for data duplication detection. To eliminate duplicate records from the database data cleaning is needed which the important phase in data is warehousing. It improves the data quality to provide better decisions support system. This paper provides the thorough survey on data duplication detection and elimination using several methods which reduced the record linkage problem, record comparison and elimination time in single and multiple the databases. Section V shows the implementation of securing the database in encrypted format.***

Keywords: Record Matching, Data Integration, Data Cleaning, Duplicate Record Detection, Duplicate Record Elimination, Data Quality, Record Linkage.

I. INTRODUCTION

The detection of similar duplicate records is a difficult task, especially when the records are domain-independent. there are several steps in data mining preprocessing one of the sub-step is data cleaning (scrubbing).the intent of the process is to detect and remove errors and inconsistencies from data and improve their quality[4]. Record linkage is an important initial step in many research and data mining projects in the biomedical and other sectors, where it is used to improve data quality [6]. Duplicate detection is the problem of determining that two different the database entries, In fact represent the same real-world object, and performing this detection for all objects represented in the the database. Duplicate detection" is also known as record linkage, object identification, record matching, and many other terms [7]. Data mining algorithms assume that data will be clean and consistent. However, in practice, this is not always the case, and for this reason, the detection and elimination of duplicate records is an important part of data cleaning. The presence of similar-duplicate records causes over-representation of data. If the the database contains different representations of the same data, the results obtained from the data mining algorithm will be erroneous. Records can be similar duplicates because of missing values, typing errors, abbreviations, extra words, word transposition, illegal values, inconsistent values, multiple values in a single free-form field, misfielded values, synonyms and nicknames, initials are all forms of dirty data [2]. A major consequence of dirty data is the existence of duplicates (i.e. multiple entries in the the database – stand-alone or integrated, referring to the same real-world entity). The removal of duplicates constitutes a major cleaning task. There are many existing methods available from which we can find the duplicate records in less time to solve record linkage problem, which we discuss in part II.

The rest of the paper is organized as follows: Section II gives the information about existing methods for duplication detection in domain dependent and domain-independent the databases. Section III presents the overview of the literature survey; Section IV shows the challenges and future scope for the topic. Section IV shows the discussion of one the challenge implemented and Section VI presents the conclusion of this paper.

II. EXISTING METHODS FOR DUPLICATION DETECTION

Duplication detection is done in two ways either detecting duplicate record in the single the database (same the database) or detecting duplicate record in multiple other the databases. The methods/algorithms which are used to detect the duplicate record are like: Blocking, Windowing, chunk estimation, Domain-independent duplication detection algorithm, weight assignment, and Domain- independent transitivity rule (DIT) and so on. These methods are categorized in domain dependent and domain-independent data duplication detection.

A. Domain Dependent Duplication Detection Methods:

a) Blocking:

Blocking method divide the set of tuples into disjoint partitions (blocks). It compares all pairs of tuples only within each block so the number of comparisons gets reduced. In blocking method, the important decision is the choice of good partitioning predicate which determines number and size of partitions and in sorted neighborhood algorithm all records are gets sorted on the particular key. The list of records is compared with the currently processed record [6-7].

b) Windowing :

In windowing method, window size is provided on the dataset to compare duplicate records on that window size. The windowing method divided into three phases which is also called as a sorted neighborhood method (SNM):

- Sorting key is assigned to each record
- All records are sorted according to key

  The first two phases are comparable to the selection of a partitioning predicate in blocking method.

- Window of fixed size across the sorted list of records. All pairs of records that appear in the same window are compared.

To reduce the number of comparisons and overall execution time, records can be clustered first which means that as for blocking records are assigned to disjoint clusters. If the window size is too small, some duplicates might be missed [6-7].

c) Clustering:

More specifically, many de-duplication approaches in the literature essentially attempt to match or cluster duplicated records [11] and do not guarantee an adequate level of scalability. On the other hand, the usage of traditional clustering algorithms is made unviable by the high number of object clusters expected in a typical de-duplication scenario, which can be of the same order as the size of the the database. However, this approach does not cope with incremental issues. Recently, some approaches have been proposed [5, 8] which exploit efficient indexing schemes based on the extraction of relevant features from the tuples under consideration. Such approaches can be adapted to deal with the de-duplication problem, even though they are not specifically designed to approach the problem from an incremental clustering perspective. The solution we propose essentially relies on an efficient and incremental clustering technique that allows discovering all clusters containing duplicate tuples.

d) Canopy Clustering:

Canopy Clustering with TFIDF (Term Frequency/Inverse Document Frequency) forms blocks of records based on those records placed in the same canopy cluster. A canopy cluster is formed by choosing a record at random from a candidate set of

records (initially, all records) and then putting in its cluster all the records within a certain loose threshold distance of it. The record is chosen at random and any records within a certain tight threshold distance of it are then removed from the candidate set of records [6].Canopy clustering provides better accuracy than other blocking methods but with loose threshold value of 1.5.

B. Domain-Independent Duplication Detection Methods:

a) Domain Independent Duplicate Detection (DIDD):

DIDD algorithm is detecting the duplicate records in multi-the databases. It attempts to sort the records once initial input dataset is getting, it sorts the records to bring the similar records nearer. After that, it utilizes window based system to limit the number of comparisons. The purpose of this step is to determine the neighbor records with which each record will be directly compared. In this algorithm, The first record of the first the database compared with whole records from the second the database linearly. Likewise, data duplication is done by comparing records linearly [4, 12].

b) Domain Independent Transitivity Rule (DIT):

Transitivity rule states that if A is equivalent to B and B is equivalent to C then A is equivalent to C. Transitivity rule is used for similar-duplicate detection. The proposed Domain-Independent Transitivity (DIT) algorithm for similar-duplicate detection is totally domain-independent, and it does not require any rule, *c.f.*, or user involvement. In DIT algorithm, SIM stands for similarity and PR stands for the power of a record. We call the latest record (the highest numbered record) in a window as the window creator. An Input to the algorithm is two records: window creator and record to which the window creator declares similarity (i.e. passes a threshold value). It should be noted that if the window creator does not declare similarity with some records, then there is no need to invoke DIT [4].

c) Token Based Similarity:

In token based similarity sequence of the record will be check and rearrangements of the words are done [12]. e.g. dagade akshata vs akshata dagade

d) Phonetic Similarity:

In phonetic similarity, the phonetic rules are provided to convert the record into specific generic name [12]. e.g. replace “i” to “ee” i.e. sujit to sujeet

e) *Character Based Similarity*:

In character based similarity typographical errors are prevented [12]. e.g. akshata to aksta

III. Literature Study

The existing system has done searching of duplicate records either in single the database or in multi-the databases. The data cleaning deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. In 1996 A.E. Monge and C. P. Elkan has proposed the field matching problem algorithm with WEBFIND application. They have worked on the abbreviations types. They had proposed the field matching algorithm, recursive field matching algorithm and smith waterman algorithm. Their study addresses the problem of reconciling information from heterogeneous sources. Such sources may represent entities differently, so identifying equivalent information is difficult [1]. Paper [2] presents a field matching algorithm that incorporates the positional ordering of characters and words and positional algorithm for field matching in place of edit distance and smith waterman algorithm. The two levels of domain-independence are identified in this work, domain-independence at attribute level and domain- independence at the record level to provide more accurate results. Vandana Dixit Kaushik et al. has proposed the edit distance algorithm which is used commonly for approximate string matching of demographic data .In this, they have used Levenshtein distance to measure the amount of difference between two sequences. The Levenshtein distance between two strings is the minimum number of edits that are required to transform one string into the

other, with the help of edit operations like insertion, deletion or substitution of a single character. The proposed de-duplication algorithm consists of two major components where first is enrolment of demographic data of an individual in the the database and second is searching the the database for a query demographic data to find potential duplicates. They have defined some phonetic rules to reduce the search space in the the database. Each name is reduced to its generic name [11]. Kazi shah Nawaz Ripon et al. has presented the domain independent data cleaning algorithm for detecting the similar duplicates. They have modified the positional algorithm [2] slightly to search the similar duplicates by calculating the match score. Similarly, they presented a domain independent technique for sorting similar records and a modification to the transitivity rule to apply into domain-independent cases. They have also proposed the DIDD algorithm which is fully domain-independent algorithm with the threshold value to clean the selected duplicates [7]. The de-duplication of data is done for reducing the record linkage problem. Peter Christen et al. presented a comparison of fast blocking methods for record linkage. Bigram indexing and canopy clustering with TFIDF are the two new blocking methods which provide better accuracy than standard blocking and sorted neighborhood blocking.[9]. Later on in 2007 Peter christen developed blocking techniques that do not require extensive parameter tuning, The proposed work replacing the global thresholds with nearest neighbor based parameters, which results in much-reduced variance in the quality of the candidate record pairs generated, and thus improves the robustness of these blocking techniques to changes in parameter settings, thus making blocking more applicable in practice [10]. The IBM had research on the estimation of de-duplication ratios in large data sets. In which they estimate the de-duplication ratio with respect to time, CPU memory, Disk access and compression ratio. They used chunk estimation algorithm for de-duplication of data. They have done work by using sampling technique [3]. In duplication detection, some challenges are not yet addressed like during entry level registration of the system duplication is not detected. There are too many representations of same data which leads to the bad quality of data. Because of that record linkage problems are occur. Also, the duplication detection is done separately in the single the database and in multi-the databases, our proposed system, combining these two domains to detect the duplicate record in the the database, which reduce the memory storage. In duplicate detection system only some of the techniques are considered from which we will get the duplicate records like token based similarity, phonetic similarity, character based similarity, word transposition, missing data, illegal values, synonyms, abbreviations and initials, extra words, misfielded values. In our proposed system all possible techniques are going to involve getting good quality of data by reducing record linkage problem. Our proposed system will also provide the security to the customer by providing the OTP (one-time password) on mobile number. We will also provide security to the database by encrypted data. There is no chance to hack the original data from the database because they are in encrypted format.

IV. Challenges and Future Scope

Challenges in data duplication detection and elimination will be like detecting and eliminating the duplicate records at entry level while entering a new record into the the database, so there will be no any duplicate record in the the database. In data warehousing, it will help to integrate data from desperate sites and cleaned data is provided for decision support system or for any other analysis of data. Also, when we are going to searching for duplicate records on domain-independent the databases that time consider the various formats of the database files so the problem of format of the database will get reduce. In existing system of data duplication detection and elimination some of the constraints are considered while detecting the duplicate record but when we are going to detect the duplicate record, consider all the constraints of records together so there will be a cleaned data into the system. Similarly, when the new record entering into the system that time search for the similar duplicate record in other the databases linearly and fetch that record if it exists in the system. While fetching the record details hide the sensitive information or details from the the database as security concerned. After the new record gets inserted into the the database update the information into the the database and send the message to the user's mobile number that their information is get updated. Hence, these will help to indicate unauthorized user trying to access the information from the database.

V. DISCUSSION

This paper presents the encryption of data in the database. The purpose of this technique is for security concern. Because of this, no one can able to hack the information from the database. Either the database is opened or it is copied to other location. In this paper, AES algorithm for encryption is used, which provides greater security on data. In Fig.1.the data stored in MySQL is in encrypted format so, no one can able to decrypt it easily.

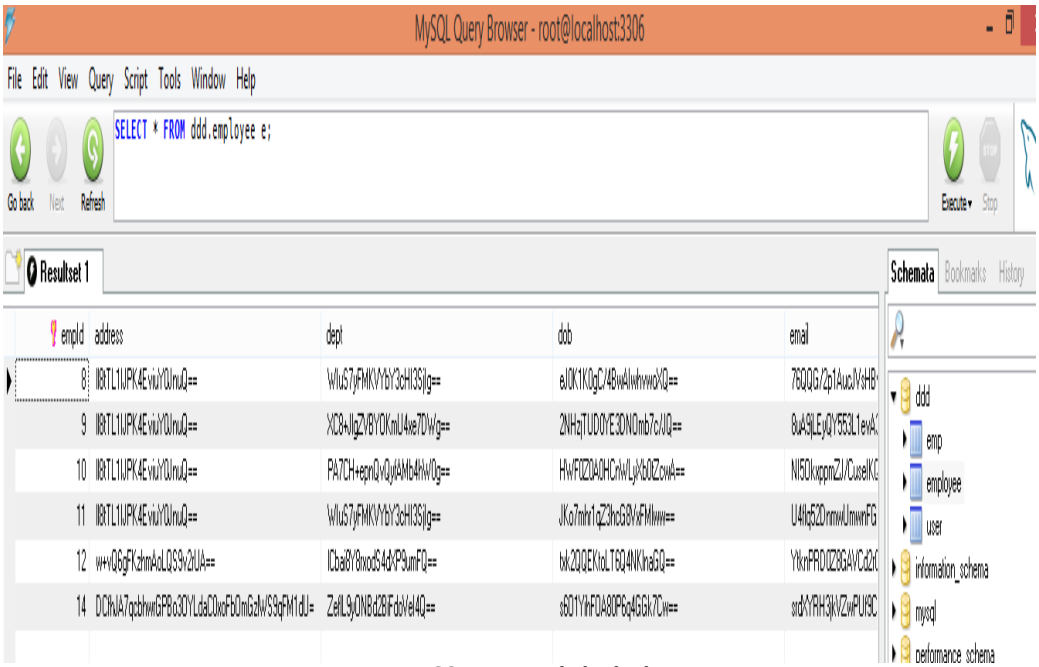
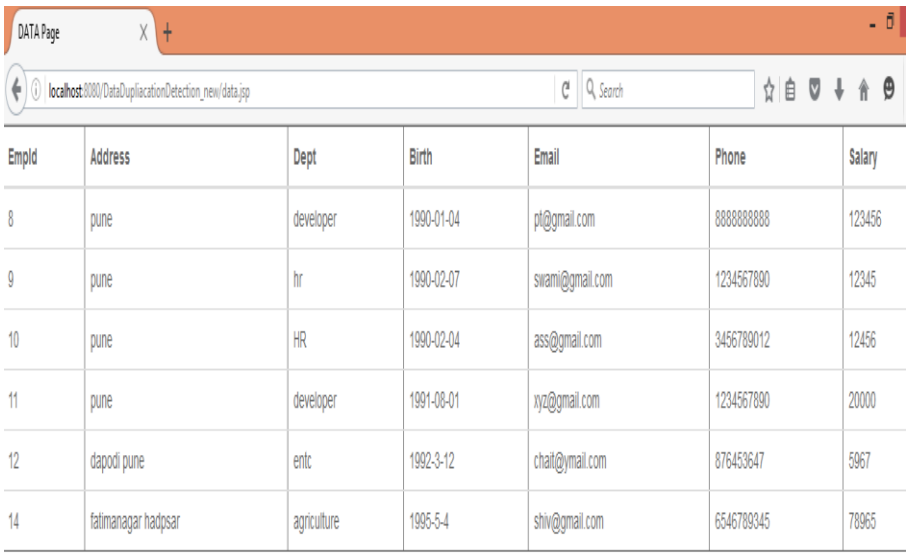


Fig.1 MySQL Encrypted The database

In Fig.2. shows that the original data on localhost site. Only admin will be able to see this original data on server site.



EmpId	Address	Dept	Birth	Email	Phone	Salary
8	pune	developer	1990-01-04	pt@gmail.com	8888888888	123456
9	pune	hr	1990-02-07	swami@gmail.com	1234567890	12345
10	pune	HR	1990-02-04	ass@gmail.com	3456789012	12456
11	pune	developer	1991-08-01	xyz@gmail.com	1234567890	20000
12	dapodi pune	entc	1992-3-12	chait@ymail.com	87643647	5967
14	fatimanagar hadpsar	agriculture	1995-5-4	shiv@gmail.com	6546789345	78965

Fig.2 Server the database

VI. CONCLUSION

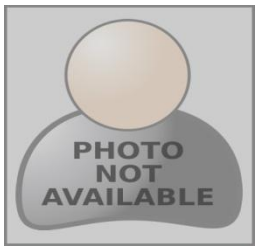
The paper focuses on the different methods to find the duplicate records in domain dependent and domain-independent the databases. The current problem of today is record linkage, which is also called as data duplication problem. The record linkage problem occurs when there are multiple representations of the same real-world entity. To reduce this problem, blocking, SNM, DIDD, DIT, canopy clustering, phonetic similarity, token based similarity, character based similarity etc. methods are used. The basic idea of these methods is to reduce the number of duplicate records from the data warehouse and minimize the time to execute. It will provide good quality of data by eliminating the duplicate record from the data warehouse.

References

1. A. E. Monge and C. P. Elkan, The field matching problem: Algorithms and applications, In Proc. SIGKDD Conference, 1996. Page no 267270.
2. Ajumobi Udechukwu, Christie Ezeife, Ken Barker, " Independent De-duplication in Data Cleaning" Dept. of Computer Science, University of Calgary, Canada, Original scientific paper.
3. Danny Harnik, Oded Margalit , Dalit Naor, Dmitry Sotnikov, Gil Vernik, "Estimation of De-duplication Ratios in Large Data Sets" IBM Research, Haifa, Israel. IEEE, 2013.
4. Kazi Shah, Nawaz Ripon, Ashiqur Rahman and G.M. Atiqur Rahaman, "A Domain-Independent Data Cleaning Algorithm for Detecting Similar-Duplicates", JOURNAL OF COMPUTERS, VOL. 5, NO. 12, DECEMBER 2010 Page No 1800-1809.
5. Peter Christen, "Towards parameter-freeblocking for scalable record linkage". Technical Report TR-CS-07-03, The Australian National University, August 2007
6. Rohan Baxter, Peter Christen, and Tim Churches, "A comparison of fast blocking methods for record linkage" In SIGKDD Workshop on Data Cleaning, Record Linkage and Object Consolidation, 2003.
7. Uwe Draisbach Hasso "A Generalization of Blocking and Windowing Algorithms for Duplicate Detection", Plattner Institute 14482 Potsdam, Germany. 978-1-4577-0866-4/2011 IEEE Page No.18-24.
8. Vandana Dixit Kaushik. "An Efficient Algorithm for De-duplication of Demographic Data", Department of Computer Science & Engineering, Hartcourt Butler Technological Institute, Kanpur 208002, India.
9. Wang, Yufeng , Tan, Chiu C. , Mi Ningfang, "Using Elasticity to Improve Inline Data De-duplication Storage Systems", IEEE 7th International Conference. 2014. Page No. 785-792 .
10. Feng Li, Hui Liu, "A High-Performance Online De-duplication Cluster", IEEE CONFERENCE PUBLICATIONS. 2012, Page No. 1062-1065
11. Mikhail Bilenko, Beena Kamath, and Raymond J. Mooney, "Adaptive blocking: Learning to scale up record linkage". In Industrial Conference on Data Mining (ICDM), 2006.
12. V.Subramaniya swamy and chentur pandiyan ,"A complete survey of duplicate record detection using data mining techniques", Information technology journal 11,copyright 2012 asian network for scientific information .
13. Akshata Dagade,Manisha mali, "Data duplication detection and elimination in domain dependent and domain independent the databases",FIFTH POST GRADUATE CONFERENCE OF COMPUTER ENGINEERING, CPGCON 2016.

AUTHOR(S) PROFILE

Akshata A. Dagade, is a Masters student at Vishwakarma Institute of Information Technology, Pune(India) in Department of Computer Engineering. She has completed her B.Tech. Degree from Vishwakarma Institute of Technology, Pune(India) during 2012 to 2014. Her research interests include Data mining and Soft computing.



Mrs. Manisha P. Mali, is currently working with VIIT, Pune at Assistant Professor in Department of Computer Engineering. She has completed her Post Graduation in Computer Engineering in 2008 from Bharati Vidyapeeth. Her current research interests include text mining, data mining, sentiment mining and soft computing.

