# A survey of Record Deduplication Techniques

Pallavi P. Gujar
*Department of Computer Engineering*
*St. Francis Institute of Technology, Borivli, Maharashtra, India*

Priyanka Desai
*Department of Computer Engineering*
*Thakur College of Engineering and Technology, Kandivali, Maharashra, India*

**Abstract- Record Deduplication refers to the task of finding and deleting records in a data set that refer to the same entity across different data sources. Record Deduplication is necessary when joining data sets based on entities that may or may not share a common identifier, as may be the case due to differences in record shape, storage location, and/or curator style or preference. In this paper, we present the detailed analysis of deduplicating the data. We have discussed the different deduplication techniques and analysed the same. For non duplicate set, the two cooperating classifiers, a Weighted Component Similarity Summing Classifier (WCSS) and Support Vector Machine (SVM) are used to iteratively identify the duplicate records from the non duplicate record and a genetic programming (GP) approach to record deduplication. The GP-based approach is also able to automatically find effective deduplication functions.**

**Keywords – Data Deduplication, Duplicate Detection, Genetic Programming, Data Identification**

## I. INTRODUCTION

Data linkage and deduplication can be used to improve data quality and integrity, which helps to re-use of existing data sources for new studies, and to reduce costs and efforts in obtaining data. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. In the deduplication process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. Deduplication is a key operation in integrating data from multiple sources. The main challenge in this task is designing a function that can resolve when a pair of records refers to the same entity in spite of various data inconsistencies.
Deduplication of data has two big advantages over a normal file system:

- Reduced Storage Allocation - Deduplication can reduce storage needs by up to 90%-95% for files and backups. Basically situations where you are storing a lot of redundant data can see huge benefits.
- Efficient Volume Replication - Since only unique data is written disk, only those blocks need to be replicated. This can reduce traffic for replicating data by 90%-95% depending on the application.

In the proposed paper, we are comparing different data deduplication techniques such as Active Learning Technique, Distance Based Technique, Unsupervised Duplicate Detection Technique and Genetic Programming Approach for Deduplication [4].

## II. RELATED THEORY

Record duplication mainly arises when data are collected from disparate sources using different information The designed technique is a description styles and metadata standards. Other common place for replicas is found in data repositories created from OCR documents [4]. These situations can lead to inconsistencies that may affect many systems such as those that depend on searching and mining tasks.
The common problems are:
1) The existing structured databases of entities are organized very differently from labeled unstructured text.
2) There is significant format variation in the names of entities in the database and the unstructured text.
3) In most cases the database will be large whereas labeled text data will be small. Features designed from the databases should be efficient to apply and should not dominate features that capture contextual words and positional information from the limited labeled data.

To solve these inconsistencies it is necessary to design a deduplication function that combines the information available in the data repositories in order to identify whether pair of record entries refers to the same real-world entity.

In the following section different deduplication techniques are discussed. Deduplication of data is mainly done to extract valuable information in spite of misspelling, typos, different writing styles or even different schema representation or data types.

## III. DATA DEDUPLICATION TECHNIQUES

Duplicate record detection is the process of identifying different or multiple records that refer to one unique real-world entity or object. Typically, the process of duplicate detection is preceded by a data preparation stage during which data entries are stored in a uniform manner in the database, resolving (at least partially) the structural heterogeneity problem. Deduplication is a specialized data compression technique for eliminating redundant data. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent across links. . There are multiple techniques for improving the efficiency and scalability of approximate duplicate detection algorithms.

- *Active-Learning Technique*

Active learning [5] is an umbrella term that refers to several models of instruction that focus the responsibility of learning on learners. Learning-based deduplication system was introduced which discovered challenging training pairs using method called Active learning. learning based deduplication system that allow automatic construction of the deduplication function by using a novel method of interactively discovering challenging training pairs. In this method the learner is automated to do the difficult task of bringing together the potentially confusing record pairs. So the user has to only perform the easy task of labeling the selected pairs of records as duplicate or not.

The system for deduplication consists of three primary inputs they are:

a) Database of records (D): The original set D of records in which duplicates need to be detected.

b) Initial training pairs (L): An optional small(less than ten) seed L of training records n1, n2 arranged in pairs of duplicates or non-duplicates.

c) Similarity functions (F): A set F of n1 functions each of which computes a similarity match between two records based on any subset of d attributes

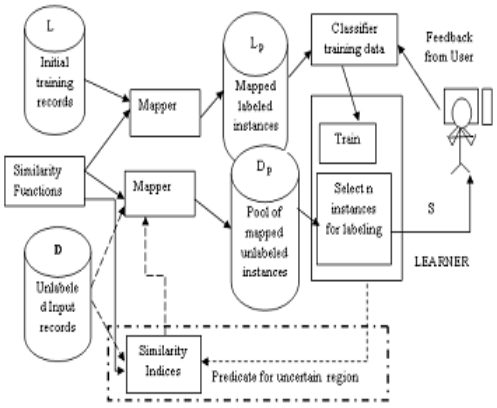


Figure 1: Overall Design and Working of Active Learning based Technique [5].

The system starts with small subsets of pairs of records designed for training which have been characterized as either matched or unique. This initial set of labeled data forms the training data for a preliminary classifier. In the sequel, the initial classifier is used for predicting the status of unlabeled pairs of records. The goal is to seek out from the unlabeled data pool those instances which, when labeled, will improve the accuracy of the classifier at the fastest possible rate. Active-learning-based system is not appropriate in some places because it always requires some training data or some human effort to create the matching models.

- *Distance-Based Techniques*

To avoid training data is to introduce a distance metric for records which does not need tuning through training data. Without using training data sets with help of distance metric and an appropriate matching

threshold, it is possible to match similar records without the need for training. Each record is considered as a field and the similarity between the records is calculated with the help of different field matching techniques such as Character based Similarity Metrics, Token based Similarity Metrics, Phonetic Similarity Metrics, and Numeric Similarity Metrics [1]. One of the problems of the distance-based techniques is the need to define the appropriate value for the matching threshold. In the presence of training data, it is possible to find the appropriate threshold value. However, this would nullify the major advantage of distance-based techniques, which is the ability to operate without training data.

- *Unsupervised Duplicate Detection Technique*

One way to avoid manual labeling of the comparison vectors is to use clustering algorithms and group together similar comparison vectors. The idea of unsupervised learning for duplicate detection has its roots in the probabilistic model proposed by Fellegi and Sunter. The WCSS classifier act as the weak classifier which is used to identify "strong" positive examples and an SVM [3] and classifier acts as the second classifier. The first classifier utilizes the weights set to match records from different data sources. Then, with the matched records being a positive set and the nonduplicate records in the negative set, the second classifier further identifies new duplicates. Finally, all the identified duplicates and nonduplicates are used to adjust the field weights set in the first step and a new iteration begins by again employing the first classifier to identify new duplicates. The iteration stops when no new duplicates can be identified.

Algorithm for Unsupervised Duplicate Detection [6]:

**Input**: Potential duplicate vector set P
Non-duplicate vector set N

**Output**: Duplicate Vector set D

C1: a classification algorithm with adjustable parameters
W that identifies duplicate vector pairs from P

C2: a supervised classifier

1. D=Ø
2. Set the parameters W of C1 according to N
3. Use C1 to get a set of duplicate vector pairs d1 from P
4. Use C1 to get a set duplicate vector pairs f from N
5. P=P – d1
6. While $|d1| \neq 0$
7. N'=N-f
8. D=D+d1+f
9. Train C2 using D and N'
10. Classify p using C2 and get a set of newly identified duplicate vector pairs d2
11. P=P - d2
12. D=D+d2
13. Adjust the parameters W of C1 according to N' and D
14. Use C1 to get a new set of duplicate vector pairs d1 from P
15. Use C1 to get a new set of duplicate vector pairs f from N
16. N=N'
17. Return D

- *Genetic Programming Approach for Deduplication*

The data is gathered from various resources. Thus it contains "dirty data". The data without any standard representation and presence of replicas are said to b dirty data. To deal with this problem approach based on Genetic programming is used. Evolutionary programming is based on ideas inspired on the naturally observed process that influence virtually all living beings, the natural selection. Genetic Programming is one of the best known evolutionary programming techniques. It is a direct evolution of programs or algorithms used for the purpose of inductive learning (supervised learning), initially applied to optimization problems.

During the evolutionary process, the individuals are handled and modified by genetic operations such as reproduction, crossover, and mutation [2], in an iterative way that is expected to spawn better individuals.

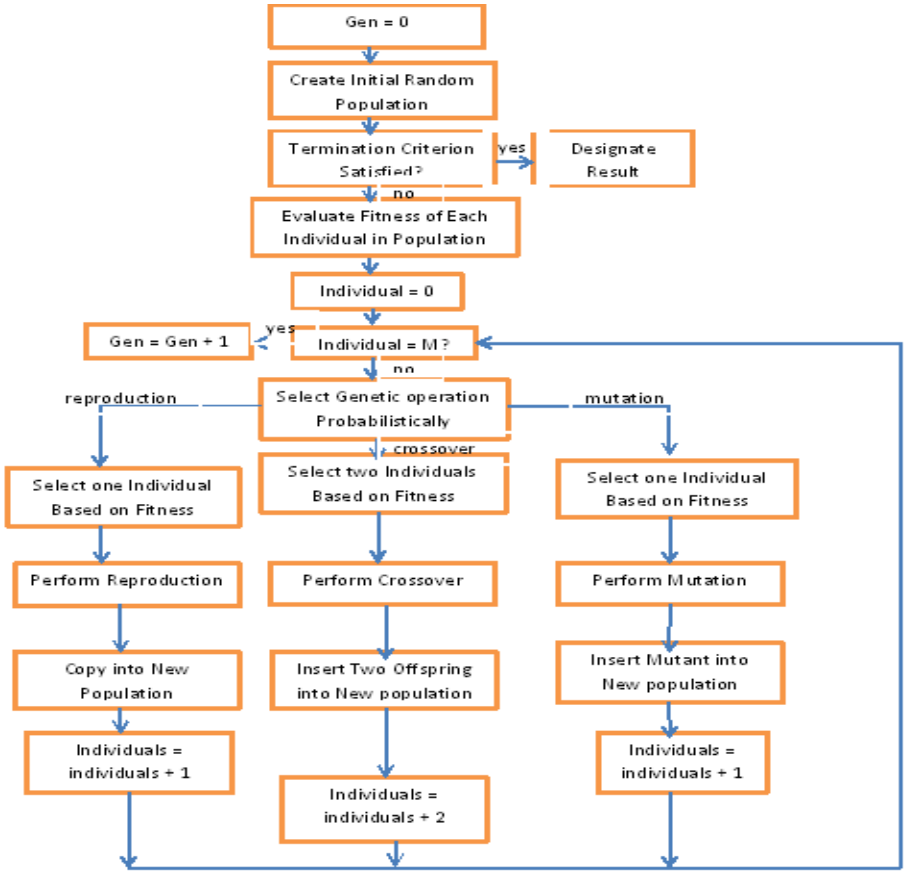


Figure 1: Flowchart of Genetic Programming

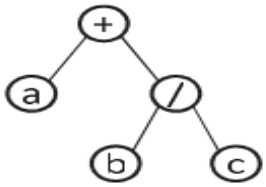


Figure 2: graphical representation of GP

## IV.CONCLUSION

The problem of identifying and handling replicas is considered important since it guarantees the quality of the information made available by data intensive systems. These systems rely on consistent data to offer high-quality services, and may be affected by the existence of duplicates, quasi replicas, or near-duplicate entries in their repositories. Deduplication, a key operation in integrating data from multiple sources, is a time-consuming, labor-intensive and domain-specific operation. In this survey, we have presented a comprehensive survey of the existing techniques used for detecting non identical duplicate entries in database records.

REFERENCES

[1] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios,"Duplicate Record Detection: A Survey", IEEE transactions on knowledge and data engineering, vol. 19, no. 1, January 2007.

[2] Moise´s G. de Carvalho, Alberto H.F. Laender, Marcos Andre Gonc¸alves, and Altigran S. da Silva "A Genetic Programming Approach to Record Deduplication" IEEE Transaction on knowledge and data engineering,vol.24, No.3, March 2012.

[3] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg, "Adaptive Name Matching in Information Integration," IEEE Intelligent Systems, vol. 18, no. 5, pp. 16-23, Sept./Oct. 2003.

[4] P.Shanmugavadivu, N.Baskar, "An Improving Genetic Programming Approach Based Deduplication Using KFINDMR", International Journal of Computer Trends and Technology- volume3Issue5- 2012

[5] Sarawagi and A. Bhamidipaty, "Interactive Deduplication Using Active Learning," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 269-278, 2002.

[6] Weifeng Su, Jiying Wang, and Frederick H. Lochovsky," Record Matching over Query Results from Multiple Webatabases",Knowledge Discovery and Data Mining, VOL. 22, NO. 4, APRIL2010