

## AN EXTENDED RANGE OF USE OF GENETIC PROGRAMMING APPROACH TO RECORD DEDUPLICATION

SUPRIYA THATAVARTHI<sup>1</sup>, GURU RAMANADHA BABU THOTA<sup>2</sup> & VIJAY SOWPATI<sup>3</sup>

<sup>1</sup>M.Tech Student, Department of PG (CSE), Loyola Institute of Technology and Management, Sathenapalli, Guntur  
Affiliated to Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India

<sup>2</sup>Assistant Professor, Department of CSE, Loyola Institute of Technology and Management,  
Sathenapalli, Guntur, Andhra Pradesh, India

<sup>3</sup>Assistant Professor, Department of CSE, Satyam Learning Campus Institute of Engg. & Tech.,  
Hyderabad, Telangana, India

### ABSTRACT

The task of recognizing, in a data warehouse, account that pass on to the matching real world entity regardless of misspelling words, kinds, special writing styles or even unusual schema versions or data types is called as the record deduplication. In presented research [1] they offered a genetic programming (GP) approach to record deduplication. [2] Their approach combines several different parts of substantiation extracted from the data content to generate a deduplication purpose that is capable to recognize whether two or more entries in a depository are duplications or not. Because record deduplication is a time intense task even for undersized repositories, their aspire is to promote a method that discovers a proper arrangement of the best pieces of confirmation, consequently compliant a deduplication function that maximizes performance using a small representative portion of the corresponding data for preparation purposes also the optimization of process is less. Our research deals these issues with a novel technique called modified bat algorithm for record duplication. The incentive behind is to generate a flexible and effective method that employs Data Mining algorithms. The structure distributes many similarities with evolutionary computation techniques such as Genetic programming approach [1]. this scheme is initialized with an inhabitant of random solutions and explores for optima by updating bat inventions. Nevertheless, disparate GP, modified bat has no development operators such as crossover and mutation. We also compare the proposed algorithm with other existing algorithms, together with GP from the experimental results.

**KEYWORDS:** Database Integration, Data Sets, Data Ware House, Evolutionary Computing and Genetic Algorithms

### INTRODUCTION

Several systems such as digital libraries and other database systems like organization databases are affected by the duplicates. We propose a genetic programming approach to find a deduplication function that is able to identify whether two entries in a repository are replicas or not. Deduplication is a task of identifying the duplicate data in a repository that refer to the same real world entity or object and systematically substitutes the reference pointers for the redundant blocks; also know as storage capacity optimization. Dirty data is defined in various categories (1) Performance degradation— a s a d d it ion n al u seless d at a d eman d more processing, more time is required to answer simple u ser queries; (2) Qua lit y loss—the presence of replicas and other inconsistencies leads to distortions in reports and misleading conclusions based on

the existing data; (3) Increasing operational costs—because of the additional volume of useless data, investments are required on ore storage media and extra computational processing power to keep the response time levels acceptable.

To avoid these problems[1], it is necessary to study the causes of “dirty” data in repositories. A major cause is the presence of duplicates, quasi replicas, or near- duplicates in these repositories, mainly those constructed by the aggregation or integration of distinct data sources. The problem of detecting and removing duplicate entries in a repository is generally known as record deduplication.

In our project we remove the dirty date[4] in the blood bank management system. As a part of genetic programming approach the gaining concepts and the entropy calculations are used to deduplicate the records.

## RELATED WORKS

Record deduplication is a growing research topic in database and many other fields as we mentioned above. The data collected from disparate sources having the redundant data[5]. Other replicas present because of the OCR documents. This leads to the inconsistent that may affect the originality of the database and the database management systems. This could be overcome by the Genetic programming approach an evolutionary algorithm- based methodology inspired by biological evolution to find computer programs that perform a user-defined task. It is a specialization of Genetic Algorithms (GA)[6] where each individual is a computer program. It is a machine learning technique used to optimize a population of computer programs according to a fitness determined by a program’s ability to perform a given computational task.

The main contribution of this paper is a GP- based approach to record deduplication that: Outperforms an existing state-of-the-art machine learning based method found in the literature; provides solutions less computationally intensive, since it suggests deduplication functions that use the available evidence more efficiently and frees the user from the burden of choosing how to combine similarity functions and repository attributes .This distinguishes our approach from all existing methods, since they require user- provided settings; frees the user from the burden of choosing the replica identification boundary value, since it is able to automatically select the de duplication f un ctions tha t be tter f it th is deduplication parameter.

## PROPOSED WORK

In Figure 1 overview of our project record deduplication detection is shown. The Blood group data set in which we are going to find duplicates is taken. The entropy value is calculated for the data set as a whole, i.e., for positive as well as negative. Based on the entropy value the donor records will be displayed in a Tree structure in which the blood groups are grouped together [7]. Entropy is the part of gain process. The entropy value is applied into the gain formula which is used to display the donor record with the highest priority.



**Figure 1: Overview of Our Project of Record Deduplication**

**DESCRIPTION**

- **Administrator**

The person who is responsible for setting up and maintaining the system is called as the system administrator. Administrator maintains the database in a secure manner. Responsible for installation, configuration, monitoring and administration and improving all the duplicates in the database performance and capacity. Admin are usually charged with installing, supporting and maintaining servers or other computer systems and planning for and responding to service outages and other problems. Administrators also responsible for creating backup and recover policy monitor network communication.

Our project is implemented for blood bank system. Here the administrator maintains the whole database which contains the details like registration of users and their blood donation details. Also the various blood bank branches for that particular blood bank Admin can look for all those details with that he can insert, update any information into the database as well as he can delete any unwanted information from the database. He calculate the entropy and gain values in order to group the user details with priority. So that he can display the blood groups by order [9]. From that the admin can come to know which blood group is required most. Finally, he merge/integrate all the blood bank branches to find out the duplicate entries in the database by means of gaining values. So that the duplicates will be displayed separately. After that he will send the mails to the blood banks which are having the duplicates in their database.

The admin create new branches with branch id's, they can change the user name and password of each branch admin's, lock/unlock their accounts, monitor the security over the truncations.

- **Creation of DB and Entropy Calculation**

Entropy is one kind of measurement procedure in information theory, details about Entropy is in here. In here, we will see how to calculate Entropy of given set of data.

$$Entropy(S) = \sum_{n=1}^n - p(I) \log_2 p(I)$$

p(I) refers to the category of the blood group.

S refers to the collection size we will set the example for entropy calculation from the following tables.

From Table 1 its known that there are two persons with positive type of AB blood group, one with negative type of AB and other with Positive type of B has been registered.

$$AB + = 2$$

$$AB - = 1$$

$$B += 1$$

**Table 1: Master Table**

Phone Number	Blood Group	Blood Type
123456789	AB	+
1234567891	AB	-
1234567892	AB	+
1234567893	B	+

The entropy value is calculated for each of the blood group by means of the above formulae.

- Entropy (AB+) =  $1 - (2/5) \log_2(2/5) = 0.960$
- Entropy (AB-) =  $1 - (1/5) \log_2(1/5) = 0.590$
- Entropy (B +) =  $1 - (1/5) \log_2(1/5) = 0.789$
- **Integration of Dataset and Detection**

Using Gaining Value

The gaining value is calculated for the records. Based on the gaining value the records which have the same key attribute values are grouped and they are displayed with their highest priority. Grouping records makes easier in identify the duplicate records and also this makes easy access of records. It improves the system performance in searching and retrieving the records. After finding entropy we next going to find gain value. Entropy is the part of gaining process. Information gain is  $G(S,A)$ , where S is the collection of the data in the data set and A is the attribute for which information gain will be calculated over the collection S.

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum(|S_v|/|S|) \times \text{Entropy}(S_v)$$

The entropy value is applied into the above formula in order to find the gain value for each blood group the gain value is calculated to the corresponding entropy value.

**Table 2: Transaction Table (Excluding Dirty Data)**

	Phone Number	Donation Date
B1	123456789	10/1/12
B2	1234567893	10/1/12
B1	1234567891	1/1/12
B1	1234567892	2/1/12
B2	123456789	1/7/12
B1	123456789	10/1/12
B2	1234567893	10/1/12
B1	1234567891	1/1/12
B1	1234567892	2/1/12
B2	123456789	1/7/12

Table 2 is the transaction table which shows that blood donors can donate blood at different branches with their personal details.

- **Display the Duplicates**

The administrator can merge the database to find out the duplicate entries for a whole-based on the gain value the admin can come to know the duplicate entries in the database[10]. For example if the gain is negative value means the admin can know that the corresponding blood group is duplicated, most he will get the overall duplicates from all the database. To detect the duplicates from each branch he can split the databases. From each branch the duplicate entries will be displayed and the mails have to be sent to those branches.

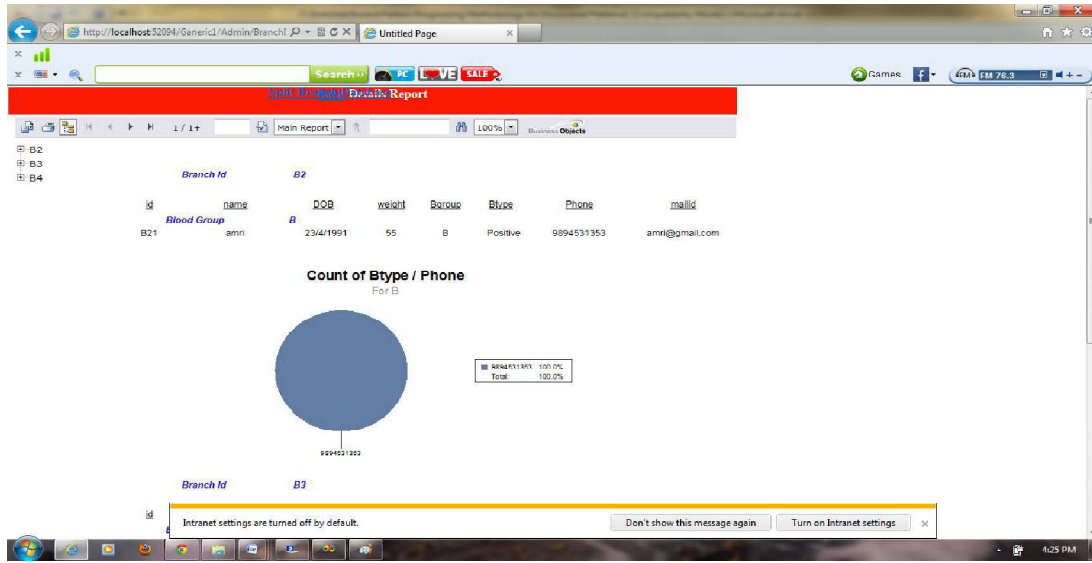


Figure 2: Duplicates at Each Branch



Figure 3: Duplicate Entries for All Branches

- **Genetic Programming Approach**

The problem of record duplication is solved by some of the evolutionary techniques [1]. Genetic programming is one of the best known evolutionary programming techniques. The main aspect that distinguishes GP from other evolutionary techniques is that it represents the concepts and the interpretation of a problem as a computer program and even the data are viewed and manipulated in this way. This special characteristic enables GP to model any other machine learning representation, another advantage of GP over other evolution ary techniques, its applicability to symbolic regression problems, since the representation structures are variable. Gp is able to discover the independent variables and their relationships with each other and with any dependent variable. Thus, GP can find the correct functional form that fits the data and discover the appropriate coefficients.

- **Experimental Results**

Figure 2 depicts about the duplicate records in each branch and figure 3 is the output of genetic programming approach which clearly shows the duplicate entries from entire branch.

## CONCLUSIONS

Identifying and handling replicas is important to guarantee the quality of the information made available by the data intensive systems such as digital libraries and e-commerce brokers. [1] 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. The system requires the assistance of an user when the committee members do not reach a decision whether the record pair represents a duplicate or not. At the end of the evolution process, the best committee is used identify replicas. The proposed method was tested in three datasets and compared to two other deduplication methods previously proposed in the literature, namely the supervised genetic programming method proposed in [1] and the semi-supervised active learning method proposed in [2]. Our approach is able to automatically suggest deduplication functions based on evidence.

## REFERENCES

1. Moise's G. de Carvalho, Alberto H.F. Laender, Marcos Andre' Gonc,alves, and Altigran S. da Silva(2012). *Genetic Programming Approach to record deduplication*
2. Bell R and Dravis F (2006), "Is You Data Dirty? and Does that Matter?," Accenture Whiter Paper, <http://www.accenture.com>.
3. Bhatt acharya I and Ge too r L (2004), "Iterative Record Linkage for Cleaning and Integration," *Proc. Ninth ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery*, pp. 11-18.
4. Ch aud huri S, Ga nja mK, Gan ti V a nd Motwani R (2003), "Robust and Efficient Fuzzy Match for Online Data Cleaning", *P r o c A C M S I G M O D I n t ' l C o n f. Management of Data*, pp. 313-324.
5. de Carvalho M G, Gonc,alves M A, Laender A H F and da Silva A S (2006), "Learning to Deduplicate", *Proc. Sixth ACM/IEEE CS Joint Conf. Digital Libraries*, pp. 41-50.
6. Fellegi I P and Sunter A B (1969), "A Theory for Record Linkage," *J. Am. Statistical Assoc.*, Vol. 66, No. 1, pp. 1183-1210.
7. Koudas N, Sarawagi S and Srivastava D (2 0 0 6), "Record Linkage : Similarity Measures and Algorithms", *Proc. ACM SIGMOD Int'l Conf. Management of Data*, pp. 802-803.
8. Koza J R (1992), *Gentic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press.
9. Verykios V S, Moustakides G V and Elfeky M G (2003), "A Bayesian Decision Model for Cost Optimal Record Matching," *The Very Large Databases J.*, Vol. 12, No. 1, pp. 28-40.