A HEURISTIC APPROACH TO RECORD DEDUPLICATION

Lata[1], Shashikala.B[2]
[1]*Dept. of Computer Science Engineering,*
[2]*Assistant Prof. Dept. of Computer Science Engineering,*
BTL Institute of Technology & Management, Bangalore-562125, Karnataka, India

Abstract- **Databases and database related technologies are having a major impact on the growing use of computers. Many global data repositories collect data from various data sources, due to this the chances of duplicates in repositories are more. The duplicate present in database is the result of misleading words and different writing styles. The presence of duplicate records decreases the system performance as it will take more time to retrieve correct relevant data from database. The clean and replica free repositories allow retrieval of higher quality information. The record deduplication is process of identifying and removal of duplicates present in database. The different approaches used to design the deduplication function are domain knowledge approach, probabilistic approach, and machine learning approach. These approaches additionally require human judgment and large computation time. To resolve the above problem, this project proposes a model to design the deduplication function for identifying the duplicate records presents in data repository by using genetic programming approach. Genetic Programming (GP) approach is a heuristic approach which automatically suggests deduplication function based on the evidence present in the data repositories. The deduplication function will help to predict whether the records are duplicates or not. Its main policy is to avoid the problems that arise due to the existence of duplicate values in the database. The proposed model uses the jaro winkler similarity function to calculate similarity measure between the records.**

Index Terms- **Database Administration, Record Deduplication, Genetic programming.**

I. INTRODUCTION

A database is a collection of related data. A database administrator (DBA) is responsible for the performance, integrity and security of database so the increasing volume of information available in Digital media has become a challenging problem for data administrators.

A major cause of dirty data in repositories is the presence of duplicates, quasi replicas, or near duplicates, mainly those conducted by the aggregation or integration of distinct data sources. Records that refers to same real world entity in a spite of misspelling words, different writing styles or even different schema representation or the data types which causes "dirty" data in the repository. The effect of replicas in a data repository is as follows.

Performance Degradation: As additional useless data demands more processing, more time is required to answer the simple user queries.

Quality Loss: The presence of replicas and the other inconsistencies leads to distortions in the reports and misleading conclusions based on the existing data.

Increasing Operational Costs: Because of the additional volume of useless data, investments are required on more storage media and extra computational processing power to keep the response time levels acceptable.

The clean and replica-free repositories allow retrieval of higher quality information. It gives clear information about data representation .Data cleaning save the computational time of resources to process data.

This paper describes the previous work done on record deduplication based on Genetic programming approach which combines several different pieces of evidence from the data content to find the deduplicate function which is used to identify whether two entries in a repository are replicas or not.

II. LITERATURE SURVEY

A literature survey is mainly carried out in order to analyze the background of the current work, which helps to find out fault in the existing

and guides on which unsolved problems can work out. Following section explores different references that discuss about several topics related to collective behavior.

A .Fuzzy Clustering for Online Data Cleaning

Fuzzy matching is process of matching of input tuples against the reference table. To ensure high quality data, data repository must validate and cleanse incoming data type from external sources.

1. Architecture

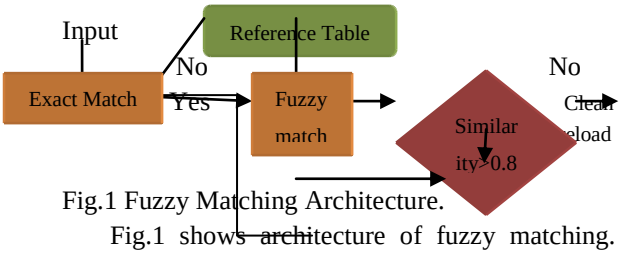


Fig.1 Fuzzy Matching Architecture.

Fig.1 shows architecture of fuzzy matching. Fuzzy match algorithm is the similarity function used for comparing tuples. Given the similarity function and an input tuple, the goal of fuzzy match operation is to return the reference tuple- a tuple in a reference relation, which is close to input tuple.

The fuzzy match similarity (fms) function is based on string matching, i.e tuples matching on high weight tokens are more similar than tuples matching on low weight tokens. The similarity between an input tuple and a reference tuple is cost of transforming. The transformation operations are token replacement, token insertion and token deletion [1].It is upto one attribute similarity.

B. Iterative Record Linkage

I. Bhattacharya and L. Getoor proposed, "Iterative Record Linkage for Cleaning and Integration," Record Linkage is the problem of determining when two records refer to the same entity [2].

1) *Operation:* By examining the context of the tuple i.e., the other tuples to which it is linked; there is a chance of more accurate linkage decision.

Example: - Two records (If we are comparing 2 census records for Jon Doe, Jonathan Doe) should be more likely to match them if they are both married to Jennette Doe.

ITERATIVE DEDUPLICATION

dup (ri,rj) = true if d(ri,rj) < t

dup (ri,rj) is the distance measure.

$d(ri,rj) = (1-\alpha)\times dattr(ri,rj)+\alpha\times dgroup(G(ri),G(rj))$.

G(r) is all the groups that r

d (ri,rj) = 1- sim (ri,rj)

sim(g1,g2) = |common(g1,g2)|/max(|g1|,|g2|)

Common (g1, g2) = {(r1,r2)|dup(r1,r2),r1 ∈ g1,r2 ∈ g2}.

C.Record Matching

V.S. Verykios, G.V. Moustakides, and M.G. Elfeky, proposed "A Bayesian Decision Model for Cost Optimal Record Matching". Record Matching or Linking is the process of identifying records, in db that refer to the same real world entity [3].

There are 2 types of record Matching.

i. The first one is called exact or deterministic and it is primarily used when these are unique identifiers for each record.
ii. The second one is called approximate record matching.

1) OPERATION: There are 2 principal steps in record matching process.

Searching step – This is used for searching of potentially linkable pair of records.

Matching step – It is a decision whether or not a given pair is correctly matched.

The aim of searching step is to reduce the no of failures to bring linkable records together for comparison. For matching step, the problem is how to enable the computer to decide whether or not a pair of records corresponds to same entity.

2) RULE: P (M/X) >= P (U/X) then x belongs to M else x belongs to U.

In the record matching process, there is a comparison vector x and matching was decide upon whether the comparison record corresponds to a Matched pair M of source records or whether the comparison record corresponds to an Unmatched pair U of source records.

D. Distance-Based Approach

It is based on rules to define whether 2 records are same or not. Rule based approaches can be considered as distance-based techniques, where the distance of 2 records is either 0 or 1.

Ex: - If age>22 then status = undergraduate
Else status = graduate

By using such a rules, unique keys are generated that can cluster multiple records that represent the same real-world entity.But the restriction is that the rules should always be correct. So rules should not be heuristically defined but should reflect absolute truths [4].

Ex: - FORALL (r1, r2) in Employee.
If r1, name is similar to r2.name
And r1.address = r2.address

Then r1 matches r2.

E.Genetic Programming Framework

*R.D.S.*Torres, A.X.Falcao, M.A. Gonc¸J.P. Papa, B. Zhang, W.Fan, and E.A. Fox proposed, "A Genetic Programming Framework for Content-Based Image Retrieval" [5].

1) *Genetic Programming*

Genetic programming (GP) is a model of programming which uses the ideas of biological evolution to handle a complex problem. It is a specialization of genetic algorithms (GA) where each individual is a computer program.

2) *GP Components*

In order to apply GP to a given problem several key components of a GP system need to be defined.

1. Reproduction: - A genetic operator that copies the individuals with the best fitness values directly into the population for the next generation.

2. Crossover: - A genetic operator that exchanges subtrees from 2 parents. Its aim is to improve the diversity and genetic fitness of population.

3. Mutation: - A genetic operator that replaces a selected individual's subtree, whose root is picked, with a randomly generated subtree.

In this project analysis of the literature on duplicate record detection.In this approach GP is used to design deduplication function.And this function is used to identification of two records are replica or not [6].

III. EXISTING SYSTEM

This section describes some methods to improve the quality of solution given by previous methods.

A.Domain Knowledge Approaches

A.K.Elmagarmid, P.G.Ipeirotis, S.Verykios proposed a matching algorithm, if a record given from a file or repository, then it looks for another record in a reference file. If it matches with the first record according to a given similarity function as threshold and if it returns more than one record that matches with that, then at that time the user is required to choose one record from that which is very close to the first one. Records matching on high-weight tokens (strings) are more similar than those matching on low-weight tokens. The authors used the vector space model for computing similarity among fields from different sources and combining the similarity scores of each field.

As a result of their experiment, they found that using evidence extracted from individual attributes improves the results of the replica identification task. Here it is necessary to define the matching threshold.

B. Probabilistic Approaches

Fellegi and Sunter (1969) proposed a more elaborated statistical approach to deal with the problem,that automatically handle duplicates. Their method relies on the definition of two boundary values that are used to classify a pair of records as being replicas or not. It is implemented with Bayes's rule and Naive based classification. This method is usually works with two boundaries as follows:

1. Positive identification boundary—if the similarity value lies above this boundary, the records are considered as replicas.
2. Negative identification boundary—if the similarity value lies below this boundary, the records are considered as not being replicas.In this case human judgement is necessary to identify the boundary values.

C. Machine Learning Approaches

S.Tejada.C.A.Knoblock and S.Minton proposed "Learning Object Identification Rules for Information Integration," Information Systems, according to this method use small portion of the available data for training. Training data set is assumed to have similar characteristics to those of the test data set, which makes feasible to the machine learning techniques to generalize their solutions to unseen data.

It requires large computation and high memory storage for mapping rules.

IV. PROPOSED SYSTEM

This section focuses on approaches to determine record deduplication using genetic programming approaches.

A. Genetic Programming (GP)

Genetic Programming is one of the evolutionary programming techniques which have the properties of natural selection. It is having mainly three operations such as selection, crossover and mutation. All the operation has been incorporated in the algorithm. At each point during the search space preserve a generation of individuals.

Features of Genetic Programming (GP)

1. GP works with multi-objective problems.
2. GP has good performance on searching over very large search spaces, where the optimal solution in many cases is not known, but it can provide near-optimal solution.

3. GP can be applied to symbolic regression problems.
4. GP represents the concepts and the interpretation of a problem as a computer program and even the data are viewed and manipulated.

B. Modeling the Record Deduplication Problem with GP

Modeling the record deduplication with GP to solve a problem, there are basic requirements that must be fulfilled, which are based on the data structure used to represent the solution. This section had chosen a tree-based GP representation for the deduplication function.
These requirements are the following:

1. All possible solutions to the problem must be represented by a tree, no matter its size.
2. The evolutionary operations applied over each individual tree must, at the end, result into a valid tree.

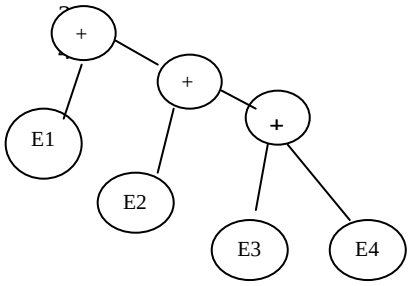


Fig.2. Tree based GP Representation.

3. Each individual tree must be automatically evaluated. For Requirement 1, it is necessary to take into consideration the kind of solution intends to find. In the record deduplication problem, look for a function that combines pieces of evidence.

Fig.2.shows the tree based GP representation, where internal node represents operators and leaf node represents evidence function.

In this approach, each piece of evidence E is a pair <attribute; similarity function> that represents the use of a specific similarity function over the values of a specific attribute found in the data being analyzed. For example,deduplicate a database table with four attributes (e.g., name, surname, address, and postal code) using a specific similarity function (e.g., the Jaro function), would have the following list of evidence: E1<name; Jaro>, E2<surname; Jaro>, E3<address; Jar>, and E4<postal code; Jar>. For this example, a very simple function would be a linear combination such as F (E1, E2, E3, E4) =E1+E2+E3+E4, which is represented as tree in fig.2.

After doing these comparisons for all candidate record pairs, the total number of correct and incorrect identified replicas is computed using fitness function.The fitness function is the GP component that is responsible for evaluating the generated individuals along the evolutionary process. If the fitness function is badly chosen or designed, it will surely fail in finding a good individual. In this approch F1 metric as our fitness function. The F1 metric harmonically combines the traditional precision (P) and recall (R) metrics commonly used for evaluating information retrieval systems as defined below:

P =Number of Correctly Identified Duplicated Pairs/ Number of Identified Duplicated Pairs (1)

R =Number of Correctly Identified Duplicated Pairs/ Number of True Duplicated Pairs (2)

F1= (2*P*R)/ (P+R) (3)

Tthis metric is used to express, as a single value, how well a specific individual performs in the task of identifying replicas. In summary, GP-based approach tries to maximize these fitness values by looking for individuals that can make more correct decisions with fewer errors.And best fitness value individuals are used for next genetication.

V. IMPLEMENTATION

This section describes the implementation of deduplication function based on GP.

A.Alogorithm

The steps of this algorithm are the following:
1. Initialize the population (with random or user provided individuals).
2. Evaluate all individuals in the present population, assigning a numeric rating or fitness value to each one.
3. If the termination criterion is fulfilled, then execute the last step. Otherwise continue.
4. Reproduce the best n individuals into the next generation population.
5. Select m individuals that will compose the next generation with the best parents.
6. Apply the genetic operations to all individuals selected. Their offspring will compose the next population. Replace the existing generation by the generated population and go back to Step 2.

7. Present the best individual in the population as the output of the evolutionary process.

The evaluation at Step 2 is done by assigning to an individual a value that measures how suitable that individual is to the proposed problem.

B. Modules

1. Similarity Function: In this approach the similarity function is used for comparing similarity between two records, output of preprocessing is input to genetic duplicate function finder. In this approach the Jaro–Winkler distance is used to measure of similarity between two strings.The Jaro distance d_j of two given strings s_1 and s_2 is

$$d_j = \begin{cases} 0, & if\ m = 0 \\ \frac{1}{3\left(\frac{m}{|s1|}+\frac{m}{|s2|}+\frac{(m-t)}{m}\right)}, & otherwise \end{cases} \quad (4)$$

Where, *m* is the number of matching characters,

t is half the number of transposition.

2.Evidence Function: In this approach, each piece of evidence is pair of <attribute, similarity function> that represent the use of a specific similarity function over the value of a specific attribute.This evidence function is used GP evolutionary process and which is given as input the genetic duplication finder.

TABLE 1

Data Set with Duplicate Records

Id	Fname	Lname	Street	City	Pincode
01	martha	math	vijaynagar	Bangalore	560000
02	marhta	math	rajajinagar	Bidar	585327
03	lata	math	vijaynagar	Banglore	560000
04	lat	mat	vijaynagar	Banglore	560000
05	latha	math	rajajinagar	Bangalore	560004
06	rama	sheelwant	shapure	Bidar	585327
07	smith	john	basavanagar	Gulbarga	560004
08	lata	math	vijaynagar	Bangalore	560000

The table 1 shows dataset, which contains six attributes with set of records.

The evidence function is calculated as <Fname(E1),Jaro>,<Lname(E2),Jaro>,<Street(E3), Jaro>,<City(E4),Jaro> ,<PinCode(E5),Jaro>.

3. Grouping Operator: Grouping operator is arithmetic operator (+,-, /,*), which are used with evidence function to generate number of evolution generations.

*4. Fitness evaluator:*The fitness function is the GP component that is responsible for evaluating the generated individuals along the evolutionary process. In this project, we have used the f1 metric fitness function.It is calculated by using equation (3).

5. Duplicate Finder: Duplicate Files Finder is used for finding and removing duplicate files by deleting; in this approach duplicate finder is used for finding duplication between the two records.

6. Performance Evaluator: It is used to analyze performance using equations (1), (2) and (3).

C. System Architecture

System architecture is the conceptual design that defines the structure and behavior of a system.

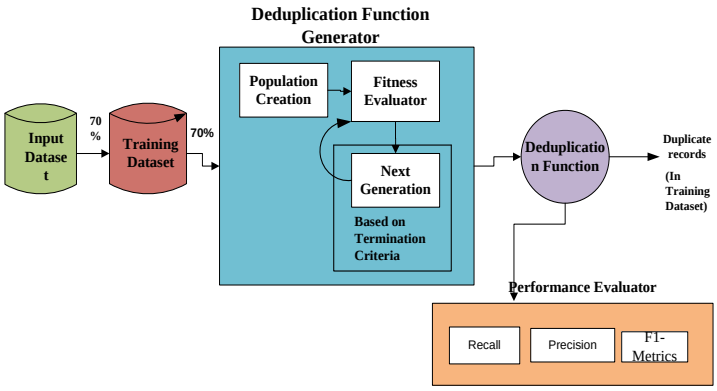


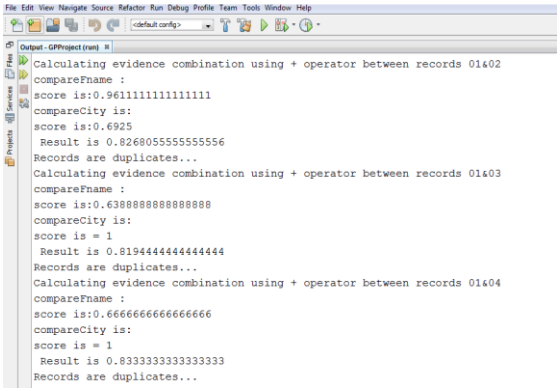
Fig.3. Architecture of Proposed Methodology.

D.Screen Shots

```
01#martha#math#856#vijaynagar#Bbb#Banglore#560000#24/07/1991#23#890462599#1bt13scs0
7#125545#4
02#marhta#math#856/29#rajajinagar#AAA#Bidar#585327#24/08/1993#19#8904625626#1bk1oec
02#123456#3
03#lata#math#856#vijaynagar#Bbb#banglore#560000#24/07/1991#23#8904625599#1bt13scs07
#1255454#4
04#lat#mat#856#vijaynagar#baa#banglore#560000#24/07/1991#23#8904625599#1bt13scs06#1
23562#4
05#latha#math#855#rajajinagar#CBB#banglore#560004#24/06/1990#23#890623669#1bt13scs0
7#123456#5
06#rama#sheelwant#658#shapure#B.K#Bidar#585327#12/07/1991#22#89032654321#3bk08cs029#
213564#5
07#smith#John#223#shwej#C.D#CCN#586456#12/05/1990#23#8954621321#3cb12cs025#2145632#
6
08#lata#math#856#vijaynagar#Bbb#banglore#560000#24/07/1991#23#8904625599#1bt13scs07
#1255454#4
```

Snapshot 1: Dataset used for training purpose

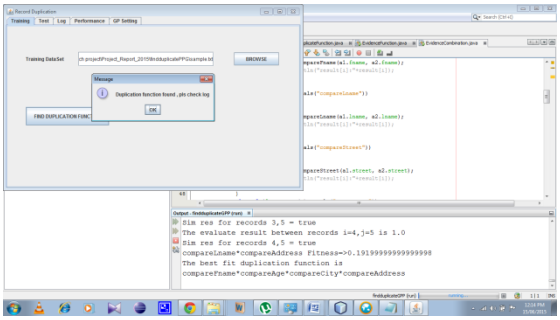
Snapshot 1 show the dataset used in project. Dataset contain fourteen attributes (Id, Fname, Lname, Door no, Steer, City, District, Pin-code, DOB, Age, Phone number, USN, Annualctc, duplicate info).



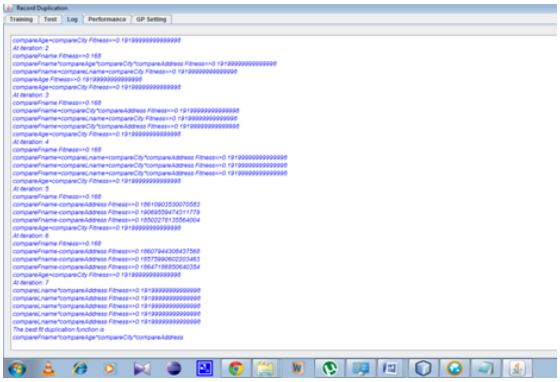
Snapshot 2: Combination of Evidence function and Grouping Operator using + operator.

Snapshot 2 show the calculation of evidence using jaro-winkler function and the evidence function (compareFname [E1] and compareCity[E2]) is combines with '+' operator i .e E1+E2 is linear combination.The threshold set here is 0.5. If the value is above the 0.5 are consider as duplicate and if below the threshold, the record are not duplicate.

Snapshot 3 is obtained when record deduplication function is generated using GP operations, which are crossover and mutation.

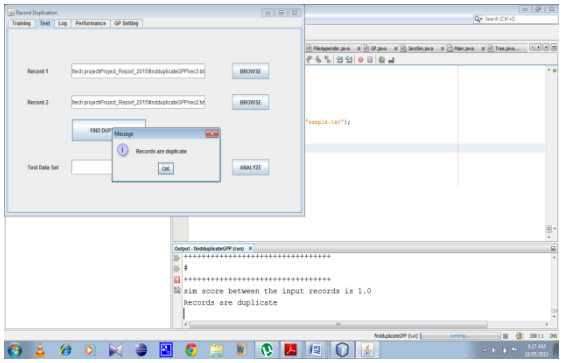


Snapshot 3.Front-End Design for Record Deduplication Using GP Approach.



Snapshot 4. Evaluation of All Individuals in Population..

Snapshot 4 shows evaluation of all individuals present in population which are based on the fitness value of each individual.



Snapshot 5. Checking Against Deduplication Function.

Snapshot 5 is obtained when two records are not duplicates and checked against the deduplicate function.

VI. CONCLUSION

The Genetic Programming (GP) based approach to record deduplication is able to automatically suggest deduplication functions based on evidence present in the data repositories.

Most approaches described in the literature used exactly the same similarity function for all available attributes, but GP is capable of combining distinct similarity functions that best fit each attribute which are considered.

GP-based approach is able to automatically find suitable deduplication functions, even when the best set of evidence is not previously known. This is extremely useful for the nonspecialized user, who does not have to worry about setting up the best set of evidence for the replica identification task.

As a future work, GP approach can be extended to implement record deduplication function based on different similarity function for different attribute (Example: For String attribute jaro_winkler function is used, for integer type one can use cosine similarity function). GP can also apply on non- linear equations to find deduplication function. Also time complexity required for execution using Jaro_Winkler can be reduced by using different parallel processing techniques.

REFERENCES

[1] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani, ―Robust and Efficient Fuzzy Match for Online Data Cleaning,‖Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 313-324, 2003.

[2] I. Bhattacharya and L. Getoor, ―Iterative Record Linkage for Cleaning and Integration,‖ Proc. Ninth ACM SIGMOD Workshop Research

Issues in Data Mining and Knowledge Discovery, pp. 11-18, 2004.

[3]V.S. Verykios, G.V. Moustakides, and M.G. Elfeky, ―A Bayesian Decision Model for Cost Optimal Record Matching,‖ The Very Large Databases J., vol. 12, no. 1, pp. 28-40, 2003.

[4]J.R. Koza, Gentic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, 1992.

[5]R.d.S. Torres, A.X. Falcao, M.A. Gonc¸alves, J.P. Papa, B. Zhang, W. Fan, and E.A. Fox, ―A Genetic Programming Framework for Content-Based Image Retrieval,‖ Pattern Recognition, vol. 42, no. 2, pp. 283-292, 2009.

[6]W. Banzhaf, P. Nordin, R.E. Keller, and F.D. Francone, Genetic Programming - An Introduction: On the Automatic Evolution of Computer Programs and Its Applications. Morgan Kaufmann Publishers, 1998.