A Genetic Programming Slant on the Way to Record De-Duplication in Repositories

Preethy.S
Department of Information Technology
N.P.R.College of Engineering and Technology, Dindigul, Tamilnadu, India

Daniel Das.A
Department of Mechanical Engineering
N.P.R.College of Engineering and Technology, Dindigul, Tamilnadu, India

Abstract-Several systems that rely on consistent data to offer high-quality services, such as digital libraries and e-commerce brokers, may be affected by the existence of duplicates, quasi replicas, or near-duplicate entries in their repositories. Because of that, there have been significant investments from private and government organizations for developing methods for removing replicas from its data repositories. This is due to the fact that clean and replica-free repositories not only allow the retrieval of higher quality information but also lead to more concise data and to potential savings in computational time and resources to process this data. In this paper, we propose a genetic programming approach to record de- duplication that combines several different pieces of evidence extracted from the data content to find a de-duplication function that is able to identify whether two entries in a repository are replicas or not. As shown by our experiments, our approach outperforms an existing state-of-the-art method found in the literature. Moreover, the suggested functions are computationally less demanding since they use fewer evidence. In addition, our genetic programming approach is capable of automatically adapting these functions to a given fixed replica identification boundary, freeing the user from the burden of having to choose and tune this parameter.

Keywords –Duplication, De-Duplication, Computation.

I. INTRODUCTION

This project describes the de-duplication of the records in databases. The input may be in any of the types. Genetic algorithms are well suited to solving production scheduling problems, because unlike heuristic methods genetic algorithms operate on a population of solutions rather than a single solution. In production scheduling this population of solutions consists of many answers that may have different sometimes conflicting objectives. As we increase the number of objectives we are trying to achieve we also increase the number of constraints on the problem and similarly increase the complexity. Genetic algorithms are ideal for these types of problems where the search space is large and the number of feasible solutions is small.

To apply a genetic algorithm to a scheduling problem we must first represent it as a genome. One way to represent a scheduling genome is to define a sequence of tasks and the start times of those tasks relative to one another. Each task and its corresponding start time represent a gene. A specific sequence of tasks and start times (genes) represents one genome in our population. To make sure that our genome is a feasible solution we must take care that it obeys our precedence constraints. We generate an initial population using random start times within the precedence constraints. With genetic algorithms we then take this initial population and cross it, combining genomes along with a small amount of randomness (mutation). We let this process continue either for a pre-allotted time or until we find a solution that fits our minimum criteria.

II.PROPOSED ALGORITHM

We present a genetic programming (GP) approach to record de-duplication. Our approach combines several different pieces of evidence extracted from the data content to produce a de-duplication function that is able to identify whether two or more entries in a repository are replicas or not. Since record de-duplication is a time consuming task even for small repositories, our aim is to foster a method that finds a proper combination of the best pieces of evidence, thus yielding a de-duplication function that maximizes performance using a small representative portion of the corresponding data for training purposes. Then, this function can be used on the remaining data or even applied to other repositories with similar characteristics. Moreover, new additional data can be treated similarly by the suggested function, as long as there are no abrupt changes in the data patterns, something that is very improbable in large data repositories.1 It is worth noticing that this (arithmetic) function, which can be thought as a combination of several effective de-duplication rules, is easy and fast to compute, allowing its efficient application to the de-duplication of large repositories.

A function used for record de-duplication must accomplish distinct but conflicting objectives: it should efficiently maximize the identification of record replicas while avoiding making mistakes during the process. The reason we have chosen GP as the basis of our approach is its known capability to find suitable answers to a given problem, without searching the entire search space for solutions, which is normally very large, and when there is more than one objective to be accomplished. In fact, we and other researchers have successfully applied GP to several information management related problems such as ranking function discovery document classification content-based image retrieval and content target advertising to cite a few, outperforming in many cases other state-of-the-art machine learning techniques. As for the record de-duplication problem, in our previous work we have successfully applied GP to provide solutions to it. In GP is applied to improve the Fellegi and Sunter's method by finding a better evidence combination than the simple linear summation used by that method. In, we use GP as a generic framework to address the record de-duplication problem independently of any other technique. As we showed in our experiments, our GP-based approach achieves better results than a state-of-the-art method based on Support Vector Machines (SVM), using less evidence to support it.

Thus, in this paper, we generalize our previous results in by showing that our GP-based approach is also able to automatically find effective de-duplication functions, even when the most suitable similarity function for each record attribute is not known in advance. This is extremely useful for the non-specialized user, who does not have to worry about selecting these functions for the de-duplication task. In sum, the main contribution of this paper is a GP-based approach to record de-duplication that outperforms an existing state-of-the-art machine learning based method found in the literature provides solutions less computationally intensive, since it suggests de-duplication functions that use the available evidence more efficiently 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 functions that better fit this de-duplication parameter.

III. MODULE DESCRIPTION

3.1 Cosine Similarity

Cosine similarity is a measure of similarity between two vectors of an innerproduct space that measures the cosine of the angle between them. The cosine of 0 is 1, and less than 1 for any other angle; the lowest value of the cosine is -1. The cosine of the angle between two vectors thus determines whether two vectors are pointing in roughly the same direction. Cosine similarity between strings was calculated using Levenstein distance and Soft-TFIDF similarity. Using cosine similarity function Genetic programming approach can automatically select populations. Cosine similarity function also used to in training phase to capture the characteristics of dataset. Based on the characteristic of data set populations are selected.

3.2 Feature Vector Extraction

Populations are the feature vectors are selected based on cosine similarity functions. Based on the population fitness functions are selected. If fitness function does not reach fitness value then populations are changed. e.g.: (att1, att2, att3). If fitness function does not reach fitness value then have to change the fitness function using genetic operations. Selected fitness function should have to reach the fitness value by machine learning approach. Genetic operations are crossover, mutation, reproduction. Selected fitness function had represented in tree format, for applying genetic operations easily.

3.3 Genetic Operation

If selected function does not reach fitness value have to apply genetic operations to change the fitness function.
Reproduction is the operation that copies individuals without modifying them. Usually, this operator is used to implement an elitist strategy that is adopted to keep the genetic code of the fittest individuals across the changes in

the generations. If a good individual is found in earlier generations, it will not be lost during the evolutionary process.

The crossover operation allows genetic content exchange between two parents, in a process that can generate two or more children. In a GP evolutionary process, two parent trees are selected according to a matching (or pairing) policy and, then, a random sub tree is selected in each parent.

The mutation operation has the role of keeping a minimum diversity level of individuals in the population, thus avoiding premature convergence.

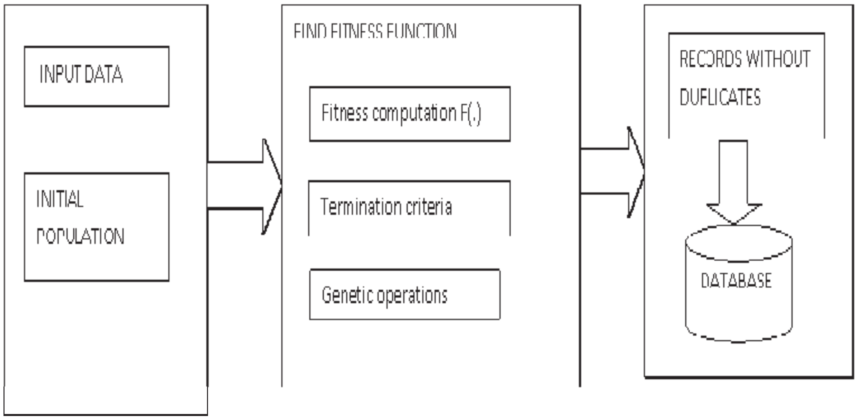

Figure 1 System Block Diagram

3.4 Removing Duplicate Records

After selecting fitness function have to calculate fitness value for all records using fitness function selected using machine learning approach. Selected fitness function can efficiently remove the records. Before storing records into database have to calculate fitness value for all records if two records meet same fitness value then have to remove one record.

IV. Experiment and Result

Table -1 Input data table

NAME	ADDRESS	CITY	PHONE NO	TYPE	CLASS
Arnie Morton of Chicago	435 s la cienegablv	Los Angeles	310/246-1501	American	0
Arnie Morton of Chicago	435 s la cienegablvd	Los Angeles	310-246-1501	Streakhouses	0
Arnie Morton of Chicago	435 s la cienegablvd	Los Angeles	310-246-1501	American	0
Arts Delicatessen	12224 venturablvd	Studio city	818/762-1221	American	1
Arts deli	12224 venturablvd	Studio city	818-762-1221	Delis	1
Hotel bel-air	701 stone canyon rd	Bel air	310/472-1211	Californian	2

Table -2Existing Records in DB

NAME	ADDRESS	CITY	PHONE NO	TYPE	CLASS
Felidia	243 e58th st	Newyork city	2127581479	Italian type	33

Four seasons grill room	**99e 52ndst**	**Newyork**	**2127549494**	**American**	**34**
Dawat	**210 e58thst**	**Newyork**	**2123557555**	**Asian**	**32**
Valentino	**3115 Pico blvd**	**Santa Monica**	**3108294313**	**Italian**	**21**
Spago	**1114 horn ave**	**Los Angeles**	**3106524025**	**California**	**20**
The palm	**9001 Santa Monica blvd**	**Los Angeles**	**3105508811**	**American**	**15**
Philippe’s the original	**1001 n alameda st**	**China Town**	**2136283781**	**Cafeterias**	**17**
River cafe	**1 water st**	**Brooklyn**	**7185225200**	**American new**	**56**
Coyote cafe	**3799 las Vegas blvd s**	**Las Vegas**	**7028917349**	**South western**	**68**

Table -3 After De-Duplication

NAME	ADDRESS	CITY	PHONE NO	TYPE	CLASS
Felidia	**243 e58th st**	**Newyork city**	**2127581479**	**Italian type**	**33**
Four seasons grill room	**99e 52ndst**	**Newyork**	**2127549494**	**American**	**34**
Dawat	**210 e58thst**	**Newyork**	**2123557555**	**Asian**	**32**
Valentino	**3115 Pico blvd**	**Santa Monica**	**3108294313**	**Italian**	**21**
Spago	**1114 horn ave**	**Los Angeles**	**3106524025**	**California**	**20**
The palm	**9001 Santa Monica blvd**	**Los Angeles**	**3105508811**	**American**	**15**
Philippe’s the original	**1001 n alameda st**	**China Town**	**2136283781**	**Cafeterias**	**17**
River cafe	**1 water st**	**Brooklyn**	**7185225200**	**American new**	**56**
Coyote cafe	**3799 las Vegas blvd s**	**Las Vegas**	**7028917349**	**South western**	**68**
Arnie Mortons of Chicago	**435 s la cienegablv**	**Los Angeles**	**3102461501**	**American**	**0**
Arnie Mortons of Chicago	**435 s la cienegablvd**	**Los Angeles**	**3102461501**	**Steakhouses**	**0**
Arts delicatessen	**12224 venturablvd**	**Studio City**	**8187621221**	**American**	**1**
Hotel bel-air	**701 stone canyon rd**	**Bel Air**	**3104721211**	**Californian**	**2**

V. CONCLUSION AND FUTURE ENHANCEMENT

5.1 Conclusion

There are various sources and process which will lead to the formation of duplicate records in the repository or DB. In our project we efficiently reduce the de-duplication of the records using Genetic Programming methodology. The reason for using GP is to powerfully eliminate the replicas produced by various users and the DBA by fault. GP will create individual values for each data. It uses the Genome concept for generating values. Thus the unique value generated does not match with each other and the elimination of replicas is drastically improved. This project will provide consistent data to digital libraries and e-commerce brokers need to remove the duplicate records. We propose genetic programming approach to identifying and handling duplicate records in the repositories. They automatically suggest de-duplication functions to handle duplicates based on the evidence present in the repositories. All are done in system dynamically and it is merit of our system. More type of data can be processed.

5.2 Future Enhancement

Reduce computational cost by using PSO (practical swarm optimization) algorithm. Time to find duplicate record is less while compared with genetic algorithm. Accuracy of PSO algorithm is high. This can also implements two or more kinds of DB's in a single system. Like, restaurant DB, voter identification DB and etc, is possible.

REFERENCES

[1] 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.

[2] H.M. de Almeida, M.A. Goncalves, M. Cristo, and P. Calado, "A Combined Component Approach for Finding Collection-Adapted Ranking Functions Based on Genetic Programming," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 399-406, 2007.

[3] P. Christen, "Probabilistic Data Generation for Deduplication and Data Linkage," Intelligent Data Eng. and Automated Learning, pp. 109-116, Springer, 2005.

[4] R.d.S. Torres, A.X. Falcao, M.A. Goncalves, 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.

[5] Zhang, Y. Chen, W. Fan, E.A. Fox, M. Goncalves, M. Cristo, and P. Calado, "Intelligent gp Fusion from Multiple Sources for Text Classification," Proc. 14th ACM Int'l Conf. Information and Knowledge Management, pp. 477-484, 2005.