Volume 3, Issue 6, June 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

De-Duplication Tool For A Data Repository In E-Shopping Using Evolutionary Computing

Manvizhi.N[1]
[1]PG Scholar
SNS College of Technology
Coimbatore, India

Suguna.M[2]
[2]Assistant Professor
SNS College of Technology
Coimbatore, India

Abstract ***The any system for data repository is expected to produce quality service. The quality service in the sense to remove duplicate data in the data repository. As the duplicates increases the computational time for processing to increases the searching time in the repository. Lot of efforts is taken to prevent this duplicate and almost most of the efforts gone in vain. Hence it is considered to apply the evolutionary algorithm based on the natural behavior. A new genetic algorithm is given which prevents this duplication and de-duplication is carried out not merely by comparing the data but by combining several different features extracted from data content. In this novel de-duplication scheme the replicas are minimized to near perfection. Applied genetic programming approach which is capable of automatically adapting the functions to a limit mentioned by us for fixed replica identification boundary. Due to this enhanced feature the user is free from the burden of choosing the tuning parameter for identifying the duplicate. The Genetic Programming approach is very expensive and computationally demanding task, so the PSO (Particle Swarm Optimization) algorithm and ABC (Artificial Bee Colony) can be used. The PSO and ABC algorithm is implemented in e-shopping application and thus provides better performance and accuracy than Genetic algorithm based technique.***

Keywords: Database administration, evolutionary computing, genetic algorithms, Particle Swarm Optimization and Artificial Bee Colony.

1. Introduction

1.1 Genetic Programming

The genetic programming has been used, to remove the duplicate data from the data repository. So the original data set only can store into the repository. Genetic programming based approach is able to automatically find suitable de-duplication functions, even when the best set of evidence is not previously known. The proposed PSO (Particle Swarm Optimization) Algorithm can be used it can provide better performance and accuracy than the Genetic algorithm based technique. The preprocessing method is using for updating process into data set then the data can be viewed as this updating process. Shopping cart is a very important feature used in e-commerce to assist people making purchases online, similar to the US English term 'shopping cart'. The Business-to-Customer aspect of electronic commerce (e-commerce) is the most visible business use of the World Wide Web. The Primary goal of an e-commerce site is to sell goods and services online. E-commerce is fast gaining ground as an accepted and used business paradigm. More and more business houses are implementing web site providing functionality for performing commercial transactions over the web. It is reasonable to say that the process of shopping on the web is becoming commonplace.

1.2 Particle Swarm Optimization

"A population based stochastic optimization technique" (Hu) It provides a population-based search procedure Getting the best solution from the problem by taking particles and moving them around in the search space. The system is initialized with a population of random solutions and searches for optima by updating generations. There is no crossover and mutation. The particles fly through the problem space by following the current optimum particles. Easy to perform. Few parameters to adjust Efficient in global search.

1.3 Artificial Bee Colony

Artificial Bee Colony (ABC) is one of the most recently defined algorithms.Motivated by the intelligent behavior of honey bees. It is as simple as Particle Swarm Optimization (PSO) and Differential Evolution (DE) algorithms, and uses only common control parameters such as colony size and maximum cycle number. ABC as an optimization tool provides a population-based search procedure. Individuals called foods positions are modified by the artificial bees with time and the bee's aim is to discover the places of food sources with high nectar amount and finally the one with the highest nectar. Artificial bees fly around in a multidimensional search space and some (employed and onlooker bees) choose food sources depending on the experience of themselves and their nest mates, and adjust their positions. Some (scouts) fly and choose the food sources randomly without using experience. If the nectar amount of a new source is higher than that of the previous one in their memory, they memorize the new position and forget the previous one. Thus, ABC system combines local search methods, carried out by employed and onlooker bees, with global search methods, managed by onlookers and scouts, attempting to balance exploration and exploitation process.

2. Literature Survey

2.1 Role And Applications Of Genetic Algorithm In Data Mining

Data mining has as goal to extract knowledge from large databases. To extract this knowledge, a database may be considered as a large search space, and a mining algorithm as a search strategy. In general, a search space consists of an enormous number of elements, making an exhaustive search infeasible. Therefore, efficient search strategies are of vital importance. Search strategies based on genetic-based algorithms have been applied successfully in a wide range of applications. The suitability of genetic-based algorithms for data mining. To discuss the various application areas where genetic Algorithm plays evolutionary role with data mining technique. A genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. Genetic algorithms find application in bioinformatics, phylogenetics, computational science, engineering, economics, chemistry, manufacturing, mathematics, physics and other fields. [6]

2.2 Genetic Algorithm Based Data Mining Approach To E- Shopping

Shopping cart is a very important feature used in e-commerce to assist people making purchases online, similar to the US English term 'shopping cart'. The Business-to-Customer aspect of electronic commerce (e-commerce) is the most visible business use of the World Wide Web. The primary goal of an e-commerce site is to sell goods and services online. E-commerce is fast gaining ground as an accepted and used business paradigm. More and more business houses are implementing web site providing functionality for performing commercial transactions over the web. It is reasonable to say that the process of shopping on the web is becoming common place. Shopping Cart feature allows online shopping customers to "place" items in the cart. Upon "checkout" the software calculates as total for the order including shipping and handling postage, packing and taxes, if applicable. The Shopping Cart is very important feature used in e-commerce to assist people making purchases products online. It provides the user a catalog of different products available in the system. In order to purchase a shopping cart is provided to the user. The shopping cart application has been developed to allow business grows larger and faster. This site will let customers to view and order products online from any part of the world. Under this website many products and services can be ordered. The shopping cart is expanded permanently through new products and services in order to offer a product portfolio corresponding to the market.[7]

2.3 Particle Swarm Optimization

Particle Swarm Optimization is a new branch in evolutionary algorithms, which were inspired in group dynamics and its synergy and were originated from computer simulations of the coordinated motions. Particles moving in an n-dimensional space to search for solutions for n-variable function optimization problem. In PSO individuals are called particles and the population is called a swarm. In Particle Swarm Optimization, a set of randomly generated solutions propagates in the design space towards the optimal solution over a number of iterations based on large amount of information about the design space that is assimilated and shared by all members of the swarm. Particle Swarm Optimization is relatively recent heuristic method that is based on the idea of collaborative behavior and swarming in biological populations. PSO is similar to the genetic algorithm in the sense that they are both population based search approaches and that they both depend on information sharing among their population members to enhance their
Search process using a combination of deterministic and probabilistic rules. Getting the best solution from the problem by taking particles and moving them around in the search space. [8].

2.4 Artificial Bee Colony

Artificial Bee Colony algorithm is an optimization algorithm based on the intelligent behavior of honey bee foraging. The specialized bees try to maximize the nectar amount stored in the hive by performing efficient division of labor and self-organization. The three agents in Artificial Bee Colony are:

1. The Employed Bee
2. The Onlooker Bee
3. The Scout

The employed bees are associated with the specific food sources, onlooker bees watching the dance of employed bees within the hive to choose a food source, and scout bees searching for food sources randomly. The onlooker bees and scout bees are unemployed bees. Initially the scout bees discover the position of all food sources, thereafter, the job of employed bee starts. An artificial employed bee probabilistically obtains some modifications on the position in its memory to target a new food source and find the nectar amount or the fitness value of the new source. Later, the onlooker bee evaluates the information taken from all artificial employed bees and them chooses a final food source with the highest probability related to its nectar number. If the fitness value of new one is higher than that of the previous one, the bee forgets the old one and memorizes the new position. This is called as greedy selection. Then the employed bee whose food source has been exhausted once again. In ABC, the solutions represent the food sources and the nectar quantity of the food sources corresponds to the fitness of the associated solution. The number of the employed and the onlooker bee is same, and the number is equal to the number of food sources.[9]

2.5 Conclusion of the literature survey

The problem of detecting and removing duplicates in data repositories is known as record de-duplication but it is also denominated in the literature data cleaning, record linkage, and record matching More specifically, record de-duplication consists of identifying and removing, from data repositories, records referring to the same object or entity in the real world, even if writing styles, data types or schemas are different. Record de-duplication is a complex task whose

treatment requires a lot of time and processing power due to the large amount of record comparisons necessary. Recently have proposed an innovative method for record de-duplication that is based on a machine learning technique known as Genetic Programming (GP) Through this method, records are de-duplicated using evidence extracted from the data content to create similarity functions, generically called de-duplication functions, capable of pointing out which records of a repository are replicas. However, despite superior results when compared to other approaches found in the literature, machine learning techniques generally rely on a training phase in which examples for learning duplication patterns are usually generated manually. As such, the cost and time necessary for creating the training sets make it difficult to use such techniques in practice. The drawback of the GA is its expensive computational cost. Then next attempts to examine the claim that PSO has the same effectiveness (finding the true global optimal solution) as the GA but with significantly better computational efficiency (less
function evaluations) by implementing statistical analysis and formal hypothesis testing.

3. Existing System

The Genetic Programming is used for identify the duplicate data from the data repository. Here the Genetic Programming using for Reproduction, Crossover, Reproduction these three methods are using for finding the Duplicate data from the repository. The main disadvantage is Genetic Programming is not repeatedly used. Then Genetic Programming is very expensive can't use for small amount of data repository, only used for large type of data repository.

3.1 Advantage

Genetic programming is one of the best simple algorithms for Evolutionary computing. It can be basic ideas come from the properties of the genetic operations and natural selection system.

3.2 Disadvantage

Genetic programming approach would not be the most adequate to use. To identify the duplicate data is not suitable for all type of dataset.

3.3 Problem Definition

Performance is low – Additional useless data demand more process timing time to answer simple question.

Data quality loss – Presence of dirty data leads to distortions in reports & misleading conclusions based on the existing data.

Increased total cost – additional volume of useless data, investments are required on more storage media, increasing computation processing power to keep response time levels acceptable.

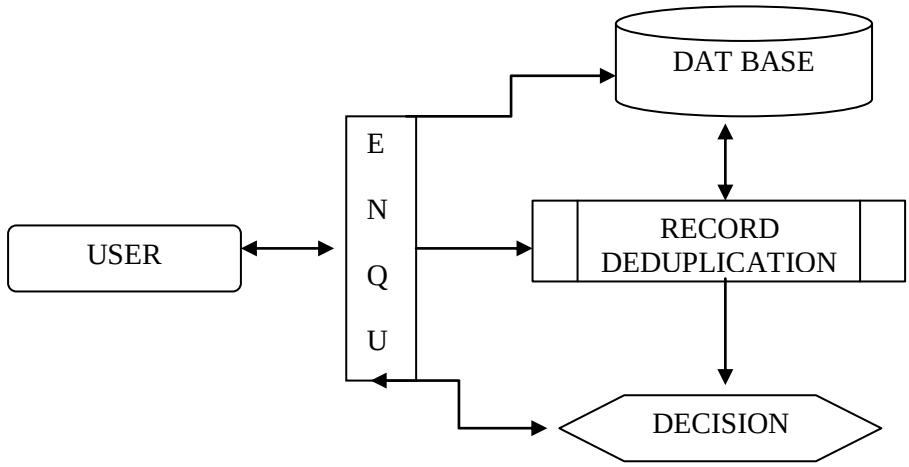

Figure 1: Block Diagram of Overall Process

4. Proposed System

The algorithm is Particle Swarm Optimization for using record de-duplication from the data repository. The system is initialized with a population of random solutions and searches for optima by updating generations. There is no crossover and mutation; the particles fly through the problem space the current optimum particles. The drawbacks of the existing system can be overcome in proposed system that is features implemented. In PSO (Particle Swarm Optimization) algorithm can be easy to perform, few parameters to adjust, Using both global and local search, efficient in global search. The next algorithm is Artificial Bee Colony is one of newly introduced optimization algorithm. The main features of ABC algorithm are:

1. Employed bee
2. Onlooker bee
3. Scout bee

The main objective of the employed bee is to generate the best solution, generate new set of bees. Onlooker bee is to create one new population generation and calculated fitness value of the new solution is better than old solution, the new solution is replaced the old solution. The problem with the abandon solution is solved with the scout bee. When an abandon solution is discovered, then that solution is replaced with a random generated solution. The newly introduced solution is called scout bee. The PSO and ABC algorithm is implemented in e-shopping application and thus provides

better performance and accuracy than Genetic algorithm based technique. There is no duplicate means it will store into the database otherwise reproduce the best individuals into the next generation population. Then select individuals will compose the next generation with best parents. Apply the genetic operations to all individuals selected replace the existing generation by the generated population and then assign the fitness value. Present the best individuals in the population as the output of the evolutionary process.

5. Evolutionary Algorithm

The overall concept for the duplicate detection method it will first start to select the data set from the data base. Mainly three methods are used crossover, selection, mutation. Then we have to calculate the fitness value for the data set, if the dataset fitness value will be high means the data is consider as original data otherwise the data will be consider as the duplicate data. If the duplicate data will be find out means the genetic programming method will apply to the dataset, the genetic programming method mainly using for three method the crossover method will just swap the record then the reproduction method will re produce the new data in data set and the mutation will be just adding their convenient for fitness value. At last the highest fitness value will be stored into the data base

Step 1: Initialize the population (with random or user provided individuals).
Step 2: Evaluate all individuals in the present population, assigning a number of rating or fitness value to each one.
Step 3: If the termination criterion is fulfilled, then execute the last step. Otherwise continue.
Step 4: Reproduce the best n individuals into the next generation population.
Step 5: Select the m individuals that will compose the next generation with the best parents.
Step 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.
Step 7: Present the best individuals in the population as the output of the evolutionary process.

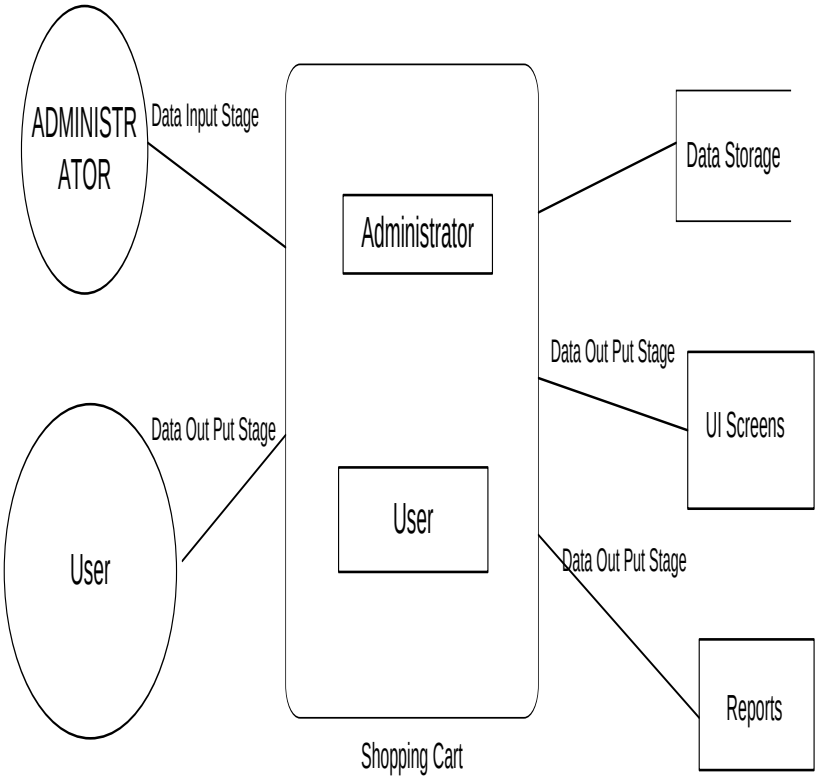

Figure: 2 Data Flow 0th level

The first administrator only stored all the details from the user. The user can enter into the data storage and take details. Suppose if any user need particular user details means enter into the administrator and register then get the details. If user details is not match into the data storage can't enter into data storage and does not get the data.

5.1 First Level Data Flow For Admin

The first level of the method is first open the login form to enter the username and password if the username and password is correct then enters into the administrator. The data storage is open to check the user detail is which country and which state. Then the products should be buying an online method. The user can decide which category the product needs should buy.

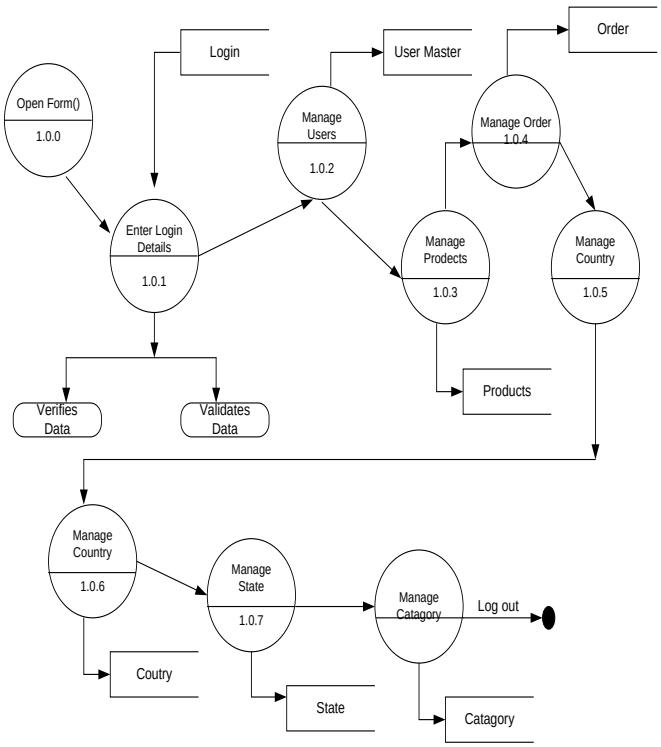

Figure: 3 1st Level of Data Flow for Admin

6. Results And Discussion

Identifying and handling replicas is important to guarantee the quality of the information made available by data intensive systems such as digital libraries and e-commerce brokers. 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. Thus, for this reason, there have been significant investments from private and government organizations for developing methods for removing replicas from large data repositories. The mainly using for Genetic programming algorithm in Genetic process automatically identifies the duplicate data.

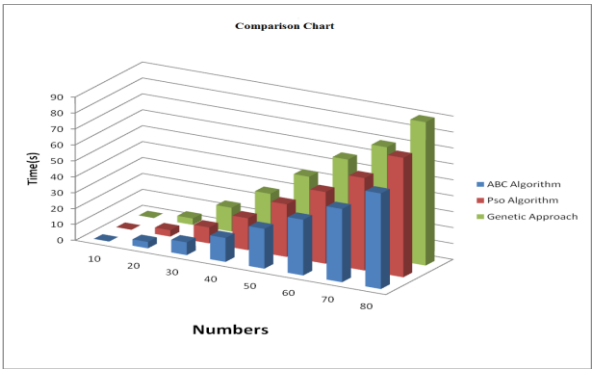

Figure: 4 Comparisons for De-Duplication In GA, PSO, ABC

Then calculate the fitness value for the data set, if the dataset fitness value will be high means the data is consider as original data otherwise the data will be consider as the duplicate data. If the duplicate data will be find out means the genetic programming method will apply to the dataset, the genetic programming method mainly using for three method the crossover method will just swap the record then the reproduction method will reproduce the new data in data set and the mutation will be just adding their convenient for fitness value. At last the highest fitness value will be stored into the data base. The next algorithm also can be using PSO (Particle Swarm Optimization) Compared to Genetic algorithm the PSO algorithm is best. The algorithm is Particle Swarm Optimization for using record de-duplication from the data repository. The system is initialized with a population of random solutions and searches for optima by updating generations. There is no crossover and mutation; the particles fly through the problem space the current optimum particles. The Artificial Bee Colony is one of the newly introduced optimization algorithm. The main features of ABC algorithm are Employed bee, Onlooker bee, and Scout bee. The employed bee phase is to generate the best solution. The Onlooker bee phase is to create one new population generation. The problem with the abandon solution is solved with the scout bee phase. The algorithms are used to remove the duplicate data from the data repository. The PSO and ABC algorithm is implemented in e-shopping application and thus provides better performance and accuracy than Genetic algorithm based technique.

7. Conclusion And Future Enhancement

A genetic approach to record de-duplication, our approach is able to automatically suggest de-duplication function based on evidence present in the data repositories. The suggested functions properly combine the best evidence available in order to identify whether two or more distinct record entries are replicas or not. Each particle keeps track of its coordinates in the solution space which are associated with the best solution (fitness) that has achieved so far by that particle. This value is called personal best,(**pbest*)***. Another best value that is tracked by the PSO is the best value obtained so far by any particle in the neighborhood of that particle. This value is called (**gbest*)***. The basic concept of PSO lies in accelerating each particle toward its pbest and the gbest locations, with a random weighted acceleration at each time. This can be implemented from using PSO (Particle Swarm Optimization) and the next algorithm for ABC (Artificial Bee Colony) these two algorithms can be going to be used. Then that three algorithm has been implemented for E-Shopping. The Artificial Bee Colony is given for best results to compare the Genetic Algorithm and Particle Swarm Optimization. The duplicate data also very low and the take very less time to remove the duplicate data in the data repository. In future the evolutionary computing algorithm for Ant Colony can be use for de-duplication method.

References

[1] A.Lacerda, M. Cristo,M.A.Goncalves,W.Fan,N.Ziviani, and B.Ribeiro-Neto(2006),"Learning to Advertise," proc.29th Ann.Int'l ACM SIGIR Conf.Research and department in information retrieval,pp.549-556.

[2] Aron Culotta Andrew McCallum(2004), Joint deduplication of multiple record types in relational data 56:89–113.

[3] B.Zhang,Y.Chen,W.Fan,E.A.Fox,andM.Crist,(2005),"Intelligent gp fusion from multiple sources for text classification," Proc.14th ACM Int'l Conf.information and knowledge management,pp.477-484.

[4] "Freely Extensible Biomedical Record Linkage(2007),"http// sourceforge.net/projects/febrl.

[5] Gabriel S.Goncalves Moises G.de Carvalho, Alberto H.F.Laender, Marcos A.Goncalves(2010), Automatic selection of training Examples for a record deduplication method based on genetic programming Journal of information and data management,vol 1.no.2.june.

[10] H.B. Newcombe, J.M.Kennedy,and S.Axford(1959) "Automatic Linkage of Vital Records,Science, vol.130,no.3381,pp.954-959,oct.

[11] H.M.de Almedia, M.A.Goncalves,and M. Cristo(2007),"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.

[12] I.Bhattacharya and L.Getoor(2004),"Iterative Record Linkage for Cleaning and Integration,"proc.Ninth ACM SIGMOD workshop Research issues in data mining and knowledge discovery,pp.11-18.

[13] I.P. Fellegi and A.B. Sunter(1969), "A Theory for Record Linkage," J. Am. Statistical Assoc., vol.66, no. 1, pp. 1183-1210.

[14] J.C.P. Carvalho and A.S. da Silva(1993),"Finding similar Identities among Objects from Multiple Web sources," proc.fifth ACM Int'l Workshop Web Information and data management,pp.90-93.

[6] Ashita, Bhagade, Parag (2012) "Artificial bee colony algorithm for optimization problem," G.H. Raisoni college of Engineering Nagpur.

[7] Gunjan verma, Vineetha Verma (2012) "Role and Application of Genetic Algorithm in Data Mining,"vol 48-No.17.

[8] Hokey Min, Tomasz, G.Smolinski(2000),"Geenetic Algorithm based Data Mining Approach For E-Purchasing,"Logistics and Distribution Institute university of Louisville.

[9] Rania Hassan* Babak Cohanim† Olivier de Weck(2005) Particle swarm optimization and genetic algoriyhm American Institute of Aeronautics and Astronautics.

