ijecs open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2 Issue 6 June, 2013 Page No. 1874-1878

AN OPTIMIZED APPROACH FOR RECORD DEDUPLICATION USING MBAT ALGORITHM

Subi S, Thangam P

[1]*PG-Scholar, M.E-CSE*
Coimbatore Institute of Engineering and Technology
subissuresh20@gmail.com
[2]*Assistant Professor,CSE Department*
Coimbatore Institute of Engineering and Technology
saithangam@gmail.com

Abstract: ***-Record deduplication[1] is the task of identifying, in a data storage, records that refer to the same real entity or any object in spite of spelling mistakes, typing errors, different writing styles or even different schema representations or data types. In the existing system aims at providing Unsupervised Duplication Detection method which can be used to identify and remove the duplicate records from different data storge. UDD, which for a given query, can effectively identify duplicates from the query result records of different web databases. After removing the same source duplicates, the supposed" non duplicate records from the same data storage can be used as training examples alleviating the trouble of users having to manually labeled training examples. Starting from the non duplicate reocord set, the two different classifiers, a Weighted Component Similarity Summing Classifier (WCSS) is used to knowing the duplicate records from the non duplicate record and presently a genetic programming (GP) approach to record deduplication. The approach joins several different pieces of attribute with similarity function extracted from the data content to produce a deduplication function that is able to identify whether two or more entries in a repository are replicas or not. Since record deduplication is a time taking task even for small repositories, the aim is to foster a method that finds a proper combination of the proper pieces of attribute with similarity function, thus yielding a deduplication function that maximizes performance using a small representative portion of the corresponding data for training purposes. But the optimization of result is less . The proposed system has to develop new method, modified bat algorithm for record duplication. The aim behind is to create a flexible and effective method that uses Data Mining algorithms. The system shares many similarities function with generational computation techniques such as Genetic programming approach.***

Keywords- Deduplication, Gp, Modified bat algorithm

1 INTRODUCTION

Record deduplication[1] is the task of identifying, in a data storage, records that refer the same real entity or object in spite of spelling mistakes words, typing errors, different writing styles or even different schema representations or data types. In this Research, the existing a genetic programming (GP)[5] approach to record deduplication[2] joines several different pieces of evidence extracted from the data content to produce a deduplication function that is able to identify whether two or more entries in a repository are replicas or not. Since record deduplication is a time taking task even for small repositories, our motive is to foster a method t0 finds a proper combination of the proper pieces of attribute with similarity function, thus yielding a deduplication function that maximizes performance using a small representative part of the corresponding data for guidance purposes. Then, the function can be used on the left over data or even applied to other repositories with similar characteristics. Moreover, new additional web cora data can be treated similarly by the suggested function, as long as there are no abrupt changes in the data structure, something that is very improbable in large data storage .The existing genetic programming approach consider all the data to found the duplicate records.

1.1 RECORD DEDUPLICATION

Deduplication[3] is a key operation in integrating data from multiple data 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. Record deduplication is the task of identifying, in a data storage, records that refer to the same real world entity or any object in spite of spelling mistakes, typing errorfs, different writing styles or even different schema representations or data types.

2 RELATED WORK

Record deduplication is a growing research topic,several existing methods are present for record deduplication

2.1 OVERVIEW OF THE GENETIC PROGRAMMING APPROACH IN RECORD DEDUPLICATION

GP[4] evolves a population of length-free data structures, also called records, each one representing a single solution to a given problem. During the generating process, the records are handled and modified by genetic operations such as reproduction, crossover, and mutation, in an iterative way that is expected to spawn better records (solutions to the proposed problem) in the subsequent generations.

In this work, the GP[6] generation wise process is guided by a generational evolutionary algorithm. This means that there are well defined and distinct generation cycles. It can adopted this approach since it captures the basic idea behind several generation wise algorithms. The algorithm steps are the following:

1. Initialize the population (with random or user provided records).
2. Evaluate all records in the present population, assigning a numeric rating or fitnessfunction to each record.
3. If the termination criterion is satisfied, then execute the last step. Otherwise continue.
4. Select 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 records selected. Their children will compose the next

Population. Replace the existing generation by the generated population and go back to Step 2.

7. Present the best record(s) in the population as the output of the evolutionary process.

2.2 EDIT DISTANCE APPROACH

The edit distance[10] between two strings 1 and 2 is the minimum number of edit operations of single characters needed to transform the string 1 into 2. There are three types of edit operations:

- insert a any word into the string.
- delete a word from the string, and
- modify one word with a different character.

To employ learnable text distance operations for each database field, and demonstrate that such measures are capable of adapting to the specific notion of similarity that is appropriate for the field's domain.Different edit operations have varying significance in different domains. For example, a digit substitution makes a major difference in a street address since it effectively changes the house number,while a single letter substitution is semantically insignificant because it is more likely to be caused by a typo or an abbreviation. Therefore, adapting string edit distance to a particular domain requires assigning different weights to different edit operations.Edit distance[9] metrics are widely used not only for text processing but also for biological equence alignment

2.3 ACTIVE LEARNING APPROACH

The success of this method criticallyhinges on being able to provide a covering and challenging set of training pairs that bring out the subtlety of the deduplication[9] function. This is non-trivial because it requires manually searching for various data inconsistencies between any two records spread apart in large lists An active learner[5] starts with a limited labeled and a large unlabeled pool of instances. The labeled set forms the training data for an initial preliminary classifier. The goal is to seek out from the unlabeled pool those instances which when labeled will help strengthen the classifier at the fastest possible rate. The initial classifier will be sure about its predictions on some unlabeled instances but unsure on most others. The unsure instances are those that fall in the classifier's confusion region. This confusion region is large when the training data is little. The classifier can perhaps reduce its confusion by seeking predictions on these uncertain instances. This intuition forms the basis for one major criteria of active learning, namely, selecting instances about which the classifier(s) built on the current training set is most uncertain.For example to show how selecting instances based on uncertainty can help reduce a classifier's confusion.

3 MODIFIED BAT ALGORITHM BASIC CONCEPTS

Metaheuristic algorithms such as particle swarm optimization, firefly algorithm and harmony search are now becoming powerful methods for solving many tough optimization problems. In this paper, propose a new metaheuristic method, the Bat Algorithm, based on the echosound behaviour of bats(basic attribute). It can also intend to join the advantages of existing algorithms into the new bat algorithm. The vast majority of heuristic and metaheuristic algorithms have been derived from the behaviour of biological systems and/or physical systems in nature. For example, particle swarm optimization was developed based on the swarm behaviour of birds and fish while simulated annealing was based on the annealing process of metals. New algorithms are also emerging recently, including harmony search and the firefly algorithm. The former was inspired by the improvising process of composing a piece of music, while the latter was formulated based on the flashing behaviour of fireflies. Each of these algorithms has certain advantages and disadvantages. For example, simulating annealing can almost assurance to find the best solution if the cooling process is slow enough and the simulation is running long enough; however, the fine adjustment in parameters does affect the union rate of the optimization process. A natural question is whether it is possible to join major advantages of these algorithms and try to build up a potentially improved algorithm.

3.1 BEHAVIOR OF THE MBATS

Most microbats(with basic attribute) are insectivores. Microbats use a produce of sonar, called, echolocation, to detect pre(record), avoid obstacles, and locate their roosting crevices in the dark. These bats emit a very loud sound pulse and listen for the echo that bounces back from the surrounding things. Their pulses change in properties and can be linked with their hunting strategies, depending on the type. Most bats use small, frequency-modulated signals to sweep through about an octave, while others more often use constant-frequency signals for echosound. Their signal bandwidth varies depends on the species, and often increased by using more harmonics.

By idealizing some of the echolsound characteristics of micro-bats(small keys), we can develop various bat-inspired algorithms or bat algorithms. Here developed Modified Bat Algorithm with Doppler Effect. For simplicity, here some of the approximate or idealized rules:

1. All bats(with abasic key record) use echosound to identify distance, and they also „know" the difference between food/prey(records) and background barriers in some magical way;

2. Bats(with a basic key) fly randomly with velocity vi at position xi with a fixed frequency fmin, varying wavelength λ

and loudness A0 to search for prey9original records). They can automatically adjust the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission r € [0,1], depending on the proximity of their target;

3. Doppler Effect is the within frequency of a wave for an observer moving relative to the source to destiny of the wave. The received frequency is higher (compared to the emitted frequency) during the approach, it is the same at the instant of passing by, and it is lower during the recession.

where *vs* is positive if the source is running away from the observer, and negative if the source is running towards the observer.

$$f = \left(\frac{c}{c+v_s}\right) f_0 \qquad (1)$$

(ii) where the similar convention applies: *vr* is positive if the observer is running towards the source, and negative if the

$$f = \left(\frac{c+v_r}{c}\right) f_0 \qquad (2)$$

(III) Single equation with both the source and receiver moving.

$$f = \left(\frac{c+v_r}{c+v_s}\right) f_0 \qquad (3)$$

C is the velocity of waves in the medium(air)
V_r is the velocity of the receiver relative to the medium; if the receiver is moving towards the source.

Vs positive is the velocity of the source relative to the medium; positive if the source is moving away from the receiver.

4. Although the loudness can vary in many ways, we assume that the loudness varies from a large (positive) A0 to a minimum constant value Amin

4 EXPERIMENTAL DATASET

For experiment evaluation used real data sets commonly employed for evaluating record deduplication approaches which are based on real data gathered from the web. In addition

The first real data set, the Cora data set, is a collection of 1,295 distinct citations to 59 computer science papers taken from the Cora research paper search engine.These citations were divided into different attributes (authornames, year, title, venue, and pages and other info) by an information extraction system.

In which for experimental evaluation u F1 metric,presion and recal mesurments are used. The F1metric harmonically combines the traditional precision (P)and recall (R) metrics commonly used for evaluating accuracy

F1- accuracy measured according to the precision and recall measurement.

$$P=\frac{Number\ Of\ Correctly\ Identified\ Duplicated\ Pairs}{aNumber\ Of\ Identified\ Duplicated\ Pairs} \qquad (4)$$

$$R=\frac{Number\ Of\ correctly\ identified\ Duplicated\ Pairs}{Number\ Of\ True\ Duplicated\ Pairs} \qquad (5)$$

$$F1=\frac{2\times P\times R}{P\cdot R} \qquad (6)$$

Equation for checking the fi metric.

CORADATASET	GENETIC	MBAT
10	0.71	0.78
20	0.8	0.83
30	0.91	0.96
40	1.1	1.5

Table 1 precision for 'cora' dataset

Table 1 lists the precision of the existing method and MBAT for cora data set.It shows mbat algorithm performs the good optimization result

CORADATASET	GENETIC	MBAT
10	0.7	0.8
20	0.75	0.83
30	0.8	0.95
40	1.2	1.6

Table 2 recall for 'cora' dataset

Table 2 lists the recall of the existing method and MBAT for cora data set.It shows mbat algorithm performsthe good optimization result

METHOD	F ACCURACY (%)
GENETIC	75
MBAT	80

Table 3 F-accuracy measures for 'cora' dataset

Table 3 lists the f-accuracy of the existing method and MBAT for cora data set.It shows mbat algorithm performsthe good optimization result

4.1 SYTEM FLOW DIAGRAM

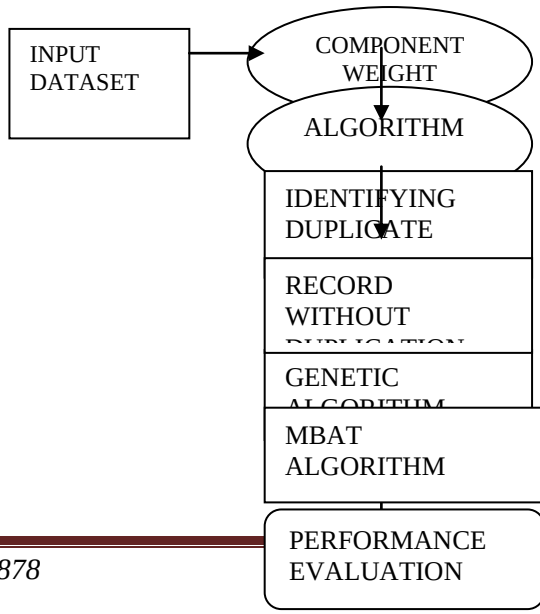

Fig.1 System flow diagram

4..2 SAMPLE SCREENSHOTS

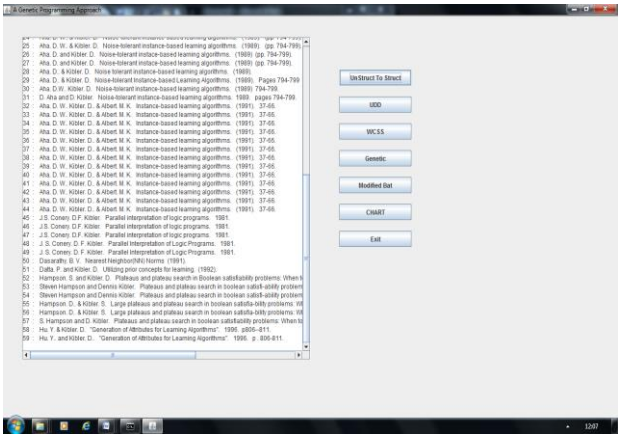

Fig.2 Home Page

Fig.2 shows the main home page of the output

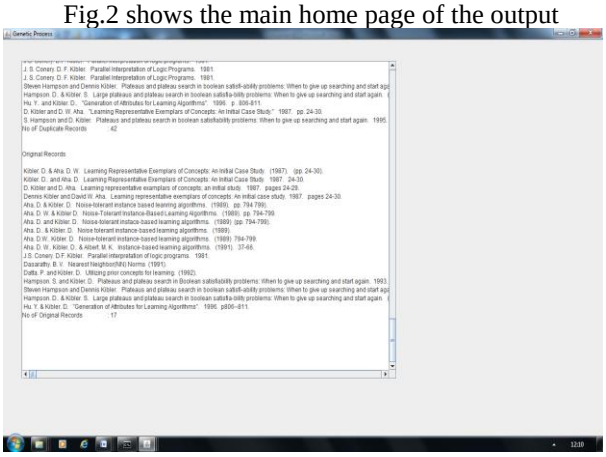

Fig.3 the existing system output using genetic programming approach

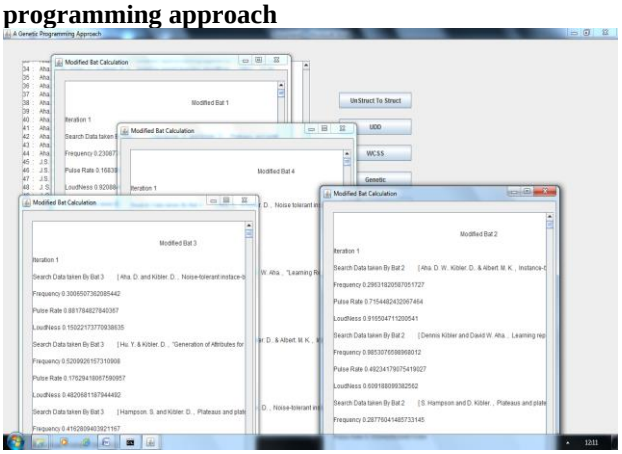

Fig.4 the proposed system output using MBAT algorithm

4 RESULTS AND DISCUSSION

In Fig 1 and Fig 2 it can notice that MBAT performs the best compared to th existing system result.

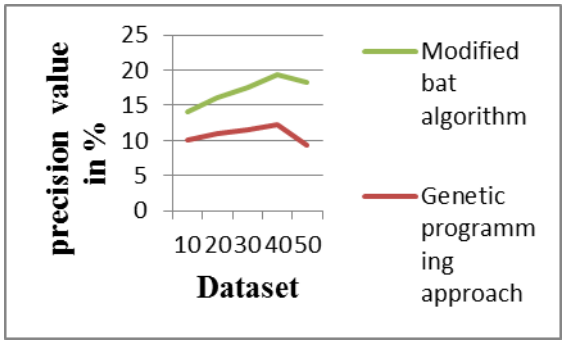

Fig.5 Comparing the precision of genetic and MBAT method in line graph

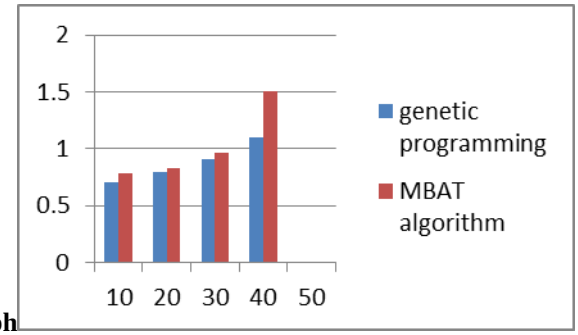

Fig.6 Comparing the recall of genetic and MBAT method in bar graph

In Fig 3 it can notice that MBAT performs the best optimization result according the evaluation f=accuracy metrics.

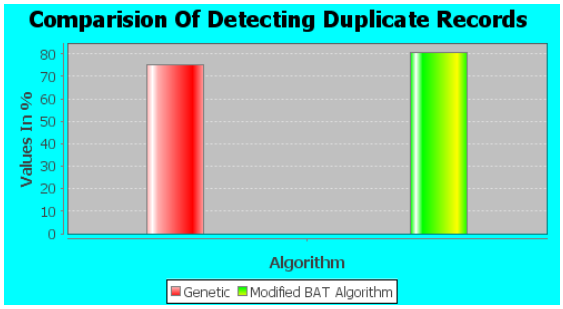

Fig.7 Comparing the f accuracy of genetic and MBAT method in bar graph

6 CONCLUSION

Duplicate detection is an important step in data integration and this method is based on offline learning techniques, which requires training data. In the Web cora database scenario, where records to match are greatly query dependent. The genetics programming approach combines several different pieces of attribute with similarity function extracted from the data content to produce a deduplication[10] function that is able to identify whether two or more entries in a repository are replicas or not. Their aim is to foster a method that finds a proper combination of the best pieces of attribute with similarity function, thus yielding a deduplication[11] function that maximizes performance using a small representative portion of the corresponding data for training purposes. In the proposed system increases the optimization of process and increases the most represented data samples are selected, it finds the best optimization solution to deduplication of the records. MBAT shares many similarities with evolutionary computation techniques such as Genetic

Algorithms[12]. The system is initialized with a population(set of records) of random solutions and searches for optima by updating generations. MBAT search the best optimal by updating generations. In MBAT takes and less error rate when comparing to the GP. It is a one -way information sharing mechanism.

References

[1] Bhattacharya I and L. Getoor, (2004) "Iterative Record Linkage forCleaning and Integration," Proc. Ninth ACM SIGMOD WorkshopResearch Issues in Data Mining and KnowledgeDiscovery, pp. 11-18.

[2] de Almeida H.M, M. Cristo M.A. Gonc¸alves, and P. Calado, "A Combined Component Approach for Finding Collection-AdaptedRanking Functions Based on GenetiProgramming," Proc. 30thAnn. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 399-406, 2007.

[3] Laender A.H.F , A.S. daSilva , M.A. Gonc¸alves, and, M.G. de Carvalho (2006) "Learning to Deduplicate" Proc. Sixth ACM/IEEE CS JointConf. Digital Libraries, pp. 41-50.

[4] Moustakides G.V, M.G. Elfeky, and, V.S. Verykios "Bayesian Decision Model for Cost Optimal Record Matching," The VeryLarge Databases J., vol. 12, no. 1, pp. 28-40, 2003.

[5] Sunter A.Band, I.P. Fellegi "A Theory for Record Linkage," J. Am.Statistical Assoc., vol. 66, no. 1, pp. 1183-1210, 1969.

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

[7] Falcao A.X, B. Zhang, , E.A. Fox, J.P. Papa, M.A. Goncalves, R.d.S. Torres, and, W.Fan "A Genetic Programming Framework for Content-Based Image Retrieval," PatternRecognition, vol. 42, no. 2,pp. 283-292, 2009.Citation Indexing," Computer, vol. 32, no. 6, pp. 67-71, June 1999.

[8] 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 transactions on knowledge and data engineering, vol. 24, no. 3, march 2012.

[9] Mikhail Bilenko and Raymond J. Mooney Department of Computer Sciences University of Texas at Austin" Adaptive Duplicate Detection Using Learnable String Similarity Measures" Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003), Washington DC, pp.39-48, August, 2003.

[10]. Elmagarmid A.K, P.G. Ipeirotis, and V.S. Verykios, "Duplicate Record Detection: A Survey," IEEE Trans. Knowledge and Data Eng.,vol. 19, no. 1, pp. 1-16, Jan. 2007.

Author Profile

Ms.Subi.S is currently pursuing M.E Computer Science and Engineering at Coimbatore Institute of Engineering and Technology, Coimbatore, Tamil Nadu, (Anna University, Chennai). She completed her B.E in Information Technology from Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, (Anna University, Coimbatore) in 2011. Her research interests include Data Mining.

Ms.P.Thangam received her B.E Degree in Computer Hardware and software Engineering from Avinashilingam University, Coimbatore in 2001. She has received her M.E degree in Computer Science and Engineering from Government College of Technology, Coimbatore in 2007. She is currently doing her PhD in the area of Medical Image Processing under Anna University, Chennai. Presently she is working as an Assistant Professor in the Department of Computer Science and Engineering at Coimbatore Institute of Engineering and Technology, Coimbatore. Her research interests are in Image Processing, Medical Image Analysis, Data Mining, Classification and Pattern Recognition.