An Optimized Approach of Modified BAT Algorithm to Record Deduplication

Faritha Banu,A
Research Scholar in Computer Science
Sree Narayana Guru College
Coimbatore – 641 105 . TamilNadu, India.

Chandrasekar C
Asst Prof, Department of Computer Applications
Sree Narayana Guru College
Coimbatore – 641 105 . TamilNadu, India.

ABSTRACT
The task of recognizing, in a data warehouse, records that pass on to the identical real world entity despite misspelling words, kinds, special writing styles or even unusual schema versions or data types is called as the record deduplication. In existing research they offered a genetic programming (GP) approach to record deduplication. 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. 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, including GP from the experimental results.

Keywords: Genetic Programming, Deduplication Function, Modified Bat Algorithm, Data Mining algorithms.

1. INTRODUCTION

Deduplication is a key operation in integrating data from multiple 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. Most existing systems use hand-coded functions. One way to overcome the tedium of hand-coding is to train a classier to distinguish between duplicates and non-duplicates. The success of this method critically hinges on being able to provide a covering and challenging set of training pairs that bring out the subtlety of the deduplication function. This is non-trivial because it requires manually searching for various data inconsistencies between any two records spread apart in large lists. Then to overcome this kind of disadvantage, a learning-based deduplication system that uses a novel method of interactively discovering challenging training pairs using a method called Active Learning came into existence [1]. The Active Learning is done on real-life datasets which shows significantly reduced number of instances needed to be achieved for high accuracy. Even active learning techniques require some training data or some human effort to create the matching models. In the absence of such training data or the ability to get human input, supervised and active learning techniques are not appropriate. One way of avoiding the need for training data is to define a distance metric [1] for records which does not need tuning through training data. Using the distance metric and an appropriate matching threshold, it is possible to match similar records without the need for training.

Deduplication is a mission of identifying record reproductions in a data repository that refer to the same real world entity or object and systematically alternates the orientation indicators for the unnecessary blocks that as well recognized as storage capacity optimization. The task of merging database records that refer to the same underlying entity is moreover referred as Record deduplication. In relational databases, accurate deduplication for records of one type is often dependent on the merge decisions made for records of other types. While all previous approaches almost have combined records of different types autonomously, this effort replicas these inter-dependencies clearly too collectively deduplication evidences of multiple types. We make a provisional random field replica of deduplication that confines these relational dependencies, and then make use of a novel relational paneling algorithm to jointly deduplication records. The rising amount of information on hand in digital media has turned out to be a challenging problem for information administrators. Regularly built on records assembled from different foundations may nearby records with different arrangement [5]. Nowadays, it is probable to say that the ability of an institute to provide useful repairs to its users is comparative to the excellence of the data obtainable in its systems. In this setting, the choice of observance repositories with “unclean” data goes distant away from technical difficulties, such as the overall speed or concert of data executive systems. The resolutions obtainable for dealing with this trouble needs more than technical endeavors, they require managing and intellectual alterations as well [5] [6].

Organizer systems frequently contain surplus copies of information: the same files or sub-file provinces, perhaps accumulated on a single multitude, on an allocated storage cluster, or reversed-up to secondary storage. Reduplicating storage systems acquire benefit of this redundancy to decrease the underlying space needed to include the file systems. Deduplication can work at also the sub-file [7] [8] [9] or whole-file level. Data deduplication strategies can be cataloged according to the essential data components they handle. In this high estimation there are two main data deduplication strategies: File-level deduplication, in which only a single copy of each file is stored. Two or more files are recognized as matching if they have the same hash value. In

terms of the architecture of the deduplication solution, there are two basic approaches. In the target-based approach deduplication is handled by the target data-storage device or service, while the client is unaware of any deduplication that might occur. Resource based deduplication proceeds on the data at the user ahead of it is transmitted. Purposely, the client software converses with the support server to ensure for the continuation of files or blocks. The topical studies [10] disclosed that, suitable to the out-of-memory to enormous backed up data; large piece-level de-duplication has an intrinsic latency and throughput difficulty that importantly influences the backup performance. There are a number of other techniques that have been used for the reason deduplication with efficiency and accuracy. The methods are deduplication using genetic algorithm, semantic methods, cloud services etc. The methods that uses GA are overcame some difficulties plotted above. This research is to find the optimization techniques that can perform better among the recent methods.

In the directed methods are not applicable for the web database scenario, anywhere the evidences to contest are uncertainty results enthusiastically produced on-the fly. Such evidences are Query- Dependent, which are configured information and have a more rapidly growth rate. To deal with the difficulty of record matching, an unsupervised record matching method, UDD, for which a given query, can efficiently are employed. They presented a genetic programming (GP) approach to record deduplication. Their recognize replicas from the inquiry results of manifold web databases. To identify duplicates two types of classifiers approach combines a number of different portions of confirmation extracted from the data content to construct a deduplication function that is able to identify whether two or more entries in a repository are replicas or not using modified bat algorithm to find the best or optimization solution to improve the accuracy of the classifier.

The main contributions are as follows:

1) First they consider the set of documents and accumulate the all information and keywords from the document. Then extract the keywords from the collection of documents.
2) In the proposed system presents an approach new modified bat algorithm to conquer the difficulty and difficulty of the genetic programming approach.
3) An optimization problem consists of maximizing or minimizing a genuine function by methodically choosing input significances from within a permitted set and computing the value of the function.
4) In the proposed system, we implement the IBAT (Modified Bat) which is a metaheuristic algorithm that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality.
5) The Modified Bat Algorithm is based on the echolocation activities of micro-bats with changeable pulse rates of emission and loudness with Doppler Effect.

2. DUPLICATE DETECTION TECHNIQUES

Data Cleaning is a time consuming process since of its lengthy activities. Since data preparation is done from multiple sources, there precedes data redundancies which brings problem in data storage capacity, processing capacity and also manual vagueness to maintainability. Deduplication is a specialized data compression technique for eliminating coarse-grained redundant data. The technique is used to get better storage utilization and can also be applied to network data transports to decrease the number of bytes that should be sent across a link. There are multiple methods for improving the effectiveness and scalability of estimated duplicate detection algorithms. The main idea after this system is that most duplicate and nonduplicate pairs are obviously separate. The system starts with minute subsets of pairs of records planed for training which have been characterized as either matched or unique. This initial set of labeled data forms the training data for a preliminary classifier. In the sequel, the initial classifier is used for predicting the status of unlabeled pairs of records.

The initial classifier will make clear determinations on some unlabeled instances but lack determination on most. The goal is to seek out from the unlabeled data collection those instances which, when labeled, will advance the accuracy of the classifier at the fastest possible rate. Pairs whose status is difficult to decide serve to strengthen the integrity of the learner. Conversely, instances in which the learner can easily predict the status of the pairs do not have much effect on the learner. Using this technique, Active-learning-based system can quickly learn the peculiarities of a data set and rapidly detect duplicates using only a small number of training data. Active-learning-based system is not appropriate in some places because it always requires some training data or some human attempt to generate the matching models. But the calculation part of weighted reserve shifts bit probabilistic and complex.

An alternative approach of creating the distance metric that is based on ranked list merging. Here the idea is to evaluate only one field using matching algorithm and discover out the best contests and rank them according to the similarities, where the best match catches the top location in rank. At last, one of the problems of the distance-based techniques is the need to define the appropriate value for the matching threshold. In the presence of training data, it is possible to find the suitable threshold value. However, this would nullify the major advantage of distance-based techniques, which is the ability to operate without training data. Unsupervised duplicate detection **(**UDD) can effectively identify duplicates from the query result records of multiple web databases for a given query it employs two classifiers. The WCSS classifier act as the frail classifier which is used to identify “strong” positive examples and an SVM classifier acts as the second classifier. Foremost, each field’s weight is set according to its “relative distance,” i.e., dissimilarity, in the middle of records from the approximated negative training set. Then, the first classifier exploits the weights set to match records from different data sources. Next, with the matched records being a positive set and the nonduplicate records in the negative set, the second classifier supplementary identifies fresh duplicates. To end with, all the identified duplicates and nonduplicate are used to adjust the field weights set in the first step and a new iteration begins by again employing the first classifier to identify new duplicates.

The iteration ends when no new duplicates can be identified. This method is well costumed for only web based data but still it requires an initial approximated training set to assign weight. Compared to the existing work, UDD (Unsupervised

Deduplication Detection) is specifically designed for the Web database scenario. Moreover, UDD focuses on studying and addressing the field weight assignment issue rather than on the similarity measure. UDD identifies duplicates as follows: WCSS classifier weight is assigned to each set of the field according to the relative distance that is the dissimilarity, among records from the approximated negative training set and the WCSS classifier, which utilizes the weights set in the first step of the UDD algorithm. It is used to match records from different data sources. then matched records being a positive set and the non duplicate records in the negative set, the SVM classifier further identifies new duplicates. Finally, all the identified duplicates and non duplicates are used to adjust the field weights set in the first step and a new iteration begins by again employing the first classifier to identify new duplication.

3. GENETIC PROGRAMMING APPROACH FOR DEDUPLICATION

The data gathering is done from multiple sources to make data repository. Data repository at that stage is said to contain "dirty data". The data with no standard representation and presents of replicas is said to be dirty data. Due to this kind of contamination usage of such repository faces few problems. They are 1) performance degradation—as additional ineffective data demand more processing, more time is required to answer simple user queries; 2) quality 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 wanted on more storage space media and extra computational processing power to keep the response time levels acceptable. The problem of detecting and removing duplicate entries in a repository is generally known as record deduplication

To deal with the above problem approach based on Genetic programming is used. This approach combines 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. Record deduplication is a kind of time consuming process so the aim is to make out duplication function for small repository and resulting function is applied to other areas. The resulting function should be able to efficiently maximize the identification of record replicas while avoiding making mistakes during the process. Genetic Programming is one of the best known evolutionary programming techniques. During the evolutionary process, the individuals are handled and modified by genetic operations such as reproduction, crossover, and mutation, in an iterative way that is expected to spawn better individuals (solutions to the proposed problem) in the subsequent generations. The steps of Genetic 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
4. The last step. Otherwise continue.
5. Reproduce the best n individuals into the next generation population.
6. Select m individuals that will compose the next generation with the best parents.
7. 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.
8. Present the best individual(s) 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. In our GP experimental environment, individuals are evaluated on how well they learn to predict good answers to a given problem, using the set of functions and terminals available. The resulting value is also called raw fitness and the evaluation functions are called fitness functions. The results are represented in tree format in this case, the rule is that each possible solution found is placed in the tree and evolutionary operation is applied for each tree. 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.

Using GP approach three set of researches are done with different conditions (a) GP was used to find the best combination function for previously user-selected evidence (b) GP was used to find the best combination function with automatically selected evidence (c) GP was tested with different replica identification boundaries. The boundary decides whether the pair is replica or not. This method is able to automatically suggest deduplication functions 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. As the result of GP approach following criteria must be satisfied: outperforms an existing state-of-the-art machine learning based method, provides solutions less computationally intensive, frees the user from the burden of choosing how to combine similarity functions and repository attributes, frees the user from the burden of choosing the replica identification boundary value, since it is able to automatically select the deduplication functions that better fit this deduplication parameter.

4. RECORD DEDUPLICATION WITH MBAT

By idealizing some of the echolocation characteristics of micro-bats, we can extend different bat-inspired algorithms or bat algorithms. Here we developed Modified Bat Algorithm with Doppler Effect. For effortlessness, here some of the approximate or idealized rules:

1. All bats use echolocation to sense distance, and they also know the difference between food/prey and background barriers in some magical way;

2. Bats fly randomly with velocity v_i at position xi with a fixed frequencyf_{min}, varying wavelength λ and loudness A_0 to search for prey. They can automatically adjust the wavelength

(or frequency) of their emitted pulses and adjust the rate of pulse emission $r \in [0,1]$, depending on the proximity of their target;

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

(i) Where v_s is positive if the source is moving away from the observer, and negative if the source is moving towards the observer.

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

(ii) (or) where the similar convention applies: v_r is positive if the observer is moving towards the source, and negative if the

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

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

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

Where,

- c the velocity of waves in the medium
- v_r is the velocity of the receiver relative to the medium; positive if the receiver is moving towards the source.
- v_s 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) A_0 to a minimum constant value A_{min}

Another obvious simplification is that no ray tracing is used in estimating the time delay and the three dimensional topography. Though this might be a good feature for the application in computational geometry, however, we will not use this as it is more computationally extensive in multidimensional cases. In addition to these simplified assumptions, we also use the following approximations, for simplicity. In general the frequency f in a range $[f_{min}, f_{max}]$ corresponds to a range of wavelengths $[\lambda_{min}, \lambda_{max}]$. In the actual implementation, we can adjust the range by adjusting the wavelengths (or frequencies), and the detectable range (or the largest wavelength) should be chosen such that it is comparable to the size of the domain of interest and then matching down to smaller ranges. Furthermore, we do not necessarily have to use the wavelengths themselves; instead, we can also vary the frequency while fixing the wavelength λ. This is because λ and f are related due to the fact λ f is constant. We will use this later approach in our implementation.

In the proposed system presents an approach new modified bat algorithm to overcome the difficulty and complexity of the genetic programming Approach .The new algorithm finds the best optimization solution for random selection of the input values and removes the duplicate records in the system. The algorithm reduces the number of the steps in the Genetic programming approach. Optimization is nothing but selection of a best element from some set of available deternatives. An optimization problem consists of maximizing or minimizing a real function by systematically choosing input values from within an allowed set and computing the value of the function. In the proposed system, we implement the IBAT (Modified Bat) which is a metaheuristic algorithm that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. The Modified Bat Algorithm is based on the echolocation behavior of micro-bats with varying pulse rates of emission and loudness with Doppler Effect.

Objective function $\boldsymbol{f(x)}, \boldsymbol{x} = (\mathbf{x_1} \ldots . \boldsymbol{x_d})^T$

Initialize the bat population $(\mathbf{x_i} = \mathbf{1}, \mathbf{2} \ldots \boldsymbol{n})$ and $\boldsymbol{V_i}$

Define Pulse frequency $\boldsymbol{f_i}$ at $\boldsymbol{x_i}$

Initialize the rates $\boldsymbol{r_i}$ and the loudness $\boldsymbol{A_i}$

While (t < Max number of iterations)

Generate new solutions by adjusting frequency,

Apply equation (1)

And updating velocities and locations /solutions [Equations (2) and (4)]

If $(\boldsymbol{rand} > \boldsymbol{r_i})$

Select a solution among the best solutions

Generate a local solution around the selected best solution

End if

Generate a new solution by flying randomly

If $(\boldsymbol{rand} < \boldsymbol{A_i} \ \& \boldsymbol{f(x_i)} < \boldsymbol{f(x_*)})$

Accept the new solutions

Increase $\boldsymbol{r_i}$ and reduce $\boldsymbol{A_i}$

End if

Rank the bats and find the current best $\boldsymbol{x_*}$

End while

The overall Block diagram is in Fig 1.

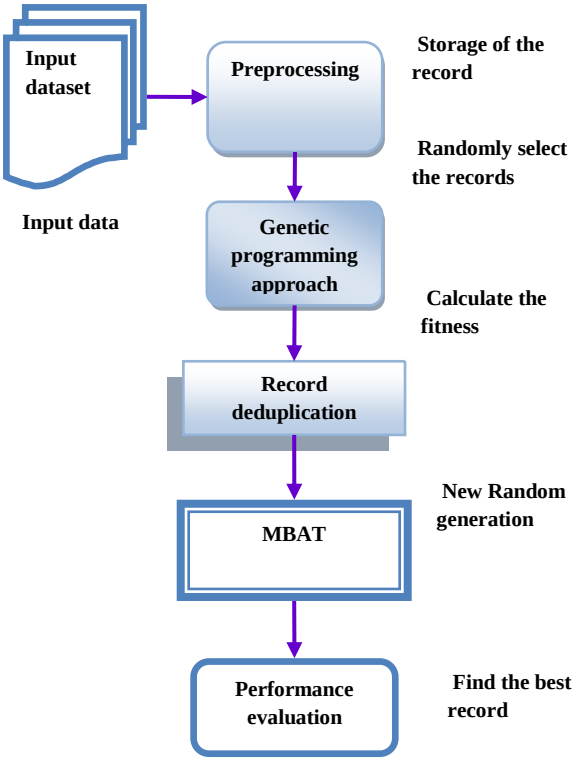

Fig 1: System Flow Diagram

5. EXPERIMENTAL RESULTS

5.1 Precision Comparison

In this section, we present and discuss the results of the Experiments performed to evaluate our proposed algorithm to record deduplication. In our experiments, we used Cora dataset to found the duplicate records. The first real data set, the Cora Bibliographic data set, is a collection of 1,295 distinct citations to 122 computer science papers taken from the Cora research paper search engine. These citations were divided into multiple attributes (author names, year, title, venue, and pages and other info) by an information extraction system.

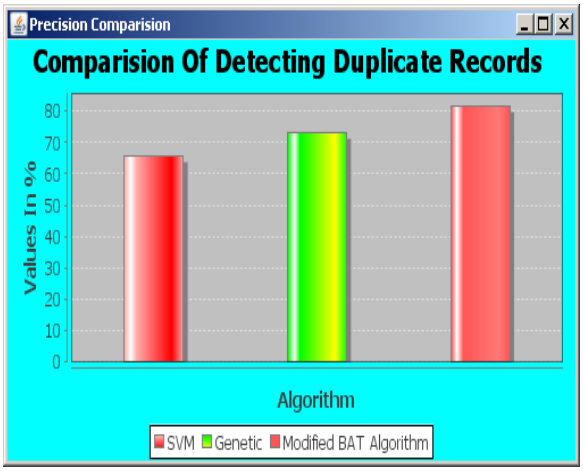

Fig 2: Precision comparison

In this Figure 2, the Precision comparison of the system between the SVM, Genetic programming approach, Modified bat algorithm. We measure the precision value in % at Y-axis as algorithm and consider the Cora dataset in the X-axis. The precision value of the MBAT (Modified Bat Algorithm) is higher than the GP and the precision value of the GP is higher than the SVM. Finally our proposed algorithm achieves the higher level of the precision value rather than the other algorithm.

5.2 F-Measure Comparison

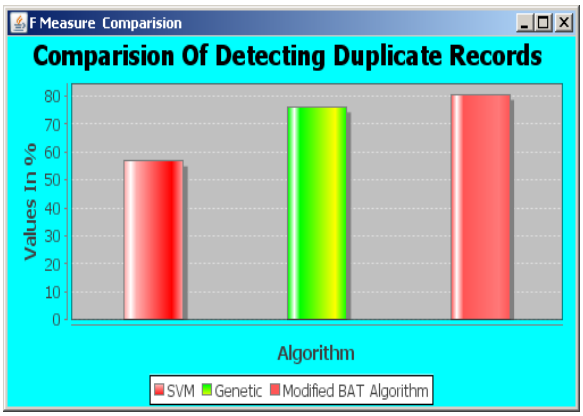

Fig 3: F measure comparison

In this Figure 3 shows that the F measure Comparison of the system between the SVM , Genetic programming approach, Modified Bat Algorithm find most relevant sample selection. We measure the F measure value in % at Y-axis as algorithm and consider the Cora dataset in the X-axis. The F measure value of the Modified Bat Algorithm is higher than the GP and the F measure value of the GP is higher than the SVM. Finally our proposed algorithm achieves the higher level of the F measure value than the other algorithm.

5.3 Recall Comparison

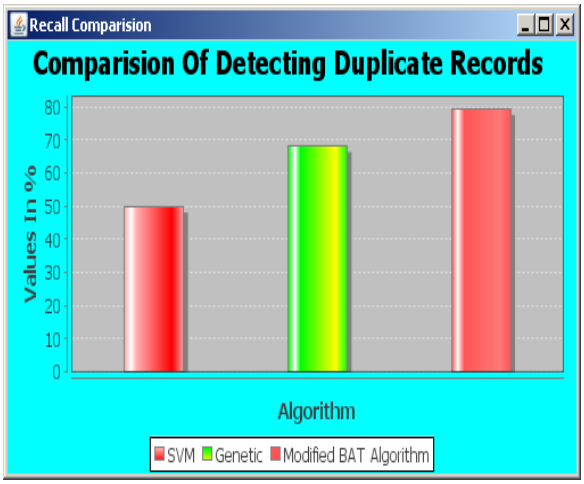

Fig 4: Recall comparison

In this Figure 4, the recall comparison of the system between the SVM, Genetic programming approach, Modified Bat Algorithm find most relevant sample selection. We measure the recall value in % at Y-axis as algorithm and consider the Cora dataset in the X-axis. The recall value of the Modified Bat Algorithm is higher than the GP and the recall value of the GP is higher than the SVM. Finally our proposed

algorithm achieves the higher level of the Recall value than the other algorithm.

5.4 Time Comparison

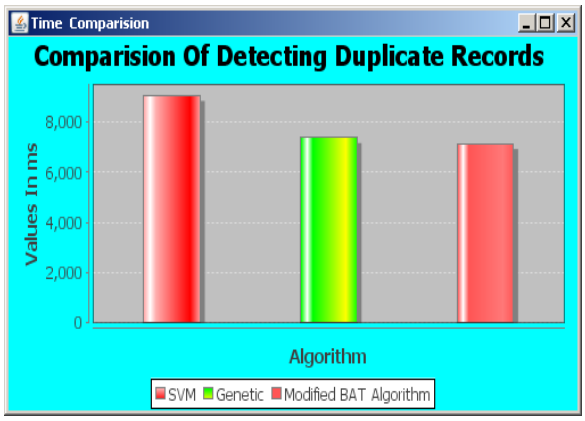

Fig 5: Time comparison

In this Figure 5, the recall comparison of the system between the SVM, Genetic programming approach, Modified Bat Algorithm find most relevant sample selection. We measure the time value in ms at Y-axis as algorithm and consider the Cora dataset in the X-axis. The time value of the Modified Bat Algorithm is higher than the GP and the time value of the GP is higher than the SVM. Finally our proposed algorithm achieves the higher level of the time value than the other algorithm.

6. CONCLUSION

Duplicate detection is a vital step in data integration and this technique is based on offline learning techniques, which requires training data. The inherited programming approach unites several different pieces of evidence mined as of the data content to generate a deduplication function that is able to identify whether two or more entries in a storehouse are duplications or not. Their intend is to promote a method that discovers a suitable grouping of the most excellent pieces of evidence, consequently yielding a deduplication function that exploits performance using a undersized representative portion of the corresponding data for training purposes. Our research work enlarges the optimization of procedure and increases the most represented data samples are selected, it discovers the best optimization solution to deduplication of the records. MBAT divides a number of similarities with evolutionary computation procedures such as Genetic Algorithms. The system is initialized with an inhabitant chance of solutions and explores for optima by updating creations. MBAT seek the best optima by updating generations. In MBAT attains and less error rate once comparing to the GP. It is a one -way information sharing method. The development only looks for the best solution. Compared with GP, the complete bat finds the best optima solution for each one of the random selection input data.

REFERENCES

[1] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios, "Duplicate Record Detection: A Survey", IEEE transactions on knowledge and data engineering, vol. 19, no. 1, January 2007.

[2] Gengxin Miao1 Junichi Tatemura2 Wang-Pin Hsiung2 Arsany Sawires2 Louise E. Moser11 ECE Dept., University of California, Santa Barbara, Santa Barbara, CA, 93106 2 NEC Laboratories America, 10080 N. Wolfe Rd SW3-350, Cupertino, CA, 95014, "Extracting Data Records from the Web Using Tag Path Clustering".

[3] Imran R. Mansuriimran@it.iitb.ac.in IIT Bombay, Sunita Sarawagi sunita@it.iitb.ac.in IIT Bombay, "Integrating unstructured data into relational databases".

[4] Jaehong Min, Daeyoung Yoon, and Youjip Won,"Efficient Deduplication Techniques for Modern Backup Operation", IEEE transactions on computers, vol. 60, no. 6, June 2011.

[5] Moises G. de Carvalho, Alberto H. F. Laender, Marcos Andre Goncalves, Altigran S. da Silva,"A Genetic Programming Approach to Record Deduplication", *IEEE Transaction on Knowledge and Data Engineering*, pp 399-412, 2011.

[6] N. Koudas, S. Sarawagi, and D. Srivastava, "Record linkage: similarity measures and algorithms," in Proceedings of the2006 *ACM SIGMOD International Conference on Management of Data*, pp. 802–803, 2006.

[7] Dutch T. Meyer and William J. Bolosky, "A Study of Practical Deduplication", *Computer and Information Science*,pp: 1-13,2011.

[8] C. Dubnicki, L. Gryz, L. Heldt, M. Kaczmarczyk, W. Kilian, P. Strzelczak, J. Szczepkowski, C.Ungureanu, and M. Welnicki.Hydrastor: a scalable secondary storage. In Proc. 7th USENIX Conference on File and Storage Technologies, 2009.

[9] C. Ungureanu, B. Atkin, A. Aranya, S. Gokhale, S. Rago, G. Cakowski, C. Dubnicki, and A.Bohra. Hydrafs: A high-throughputfile system for the Hydrastor content-addressable storage system. In Proc. *8th USENIX Conference on File and StorageTechnologies*, 2010.

[10] D. Bhagwat, K. Eshghi, D. D. Long, and M. Lillibridge, "Extreme Binning: Scalable, Parallel Deduplication for Chunkbased File Backup," *HP Laboratories, Tech. Rep. HPL-2009- 10R2*, Sep. 2009.

[11] A New Metaheuristic Bat-Inspired Algorithm, Xin-She Yang, Department of Engineering, University of Cambridge.

[12] "Bats behaviour", www. Swam inteLligence.org.