ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

A Novel Deduplication Technique Using an Evolutionary Approach

R. Gayathri[1], A. Malathi[2]

Assistant Professor[1], School of IT and Science, Dr. G.R.D. College of Science, Coimbatore, India[1]

Assistant Professor[2,] PG and Research Department of Computer Science, Government Arts College, Coimbatore, India[2]

Abstract: The process which identifies the records that refers to the same entity in data storage is known as record deduplication. UDD can effectively identify duplicates of different web databases. Initially from the non duplicate record set, the two different classifiers, a Weighted WCSS is used for deduplication. The approach joins different pieces of attribute with similarity function extracted from the data content to produce a deduplication function that able to identify presence of replicas. The main goal of this paper is to promote a method that finds a proper combination of the proper pieces of attribute with likeness function. In the existing approaches optimization is less. The proposed system has adapted the firefly algorithm for record duplication. The structure distributes many similarities with evolutionary computation techniques such as GP and SVM based approaches. The experimental result was done by comparing the proposed algorithm with other two existing approaches machine algorithm.

Keywords: Deduplication, genetic algorithm, optimization, firefly, data mining

I. RELATED WORK

An imperative feature of duplicate detection is to condense the number of record pair comparisons. There are so many existing methods available for detecting and reducing duplicated records. These techniques mostly fall under the one of the two major categories as mentioned in [1] Identifying Similarity function to determine the similar records using this metric the matching records are identified.

In [2] present two learnable text similarity measures suitable for this task: an extended variant of learnable string edit distance, and a novel vector-space based measure that employs a Support Vector Machine (SVM) for training. Experimental results on a range of datasets show that in this paper their proposed framework can improve duplicate detection accuracy over traditional technique.

To improve computational procedures in applications, of the Fellegi-Sunter model of record linkage. In [3] describes a method for estimating weights using the EM Algorithm under less restrictive assumptions. The weight computation automatically incorporates a Bayesian adjustment based on file characteristics.

An important aspect of duplicate detection is to reduce the number of record pair comparisons. Some existing methods [9, 10, 11] use different partitioning method to considerably reduce the comparison of records instead of comparing the same blocks.

Jaro [4] in his article presents the theoretical background to understand the statistical basis of record linkage and the methodology was developed for the estimation of parameters required by any record linkage activity.

II. PROBLEM DEFINITION

In existing system authors present a genetic programming [7, 8] (GP) approach to record deduplication. Their GP-based approach is also able to automatically find effective deduplication 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 deduplication task. In addition, authors show that their approach is also able to adapt the suggested deduplication function to changes on the replica identification boundaries used to classify a pair of records as a match or not. This releases the user from the burden of having to choose and tune these parameter values. . Their 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. Since record deduplication is a time consuming task even for small repositories, their aim is to foster a method that finds a proper combination of the best pieces of evidence, thus yielding a deduplication function that maximizes performance using a small representative portion of the corresponding data for training purposes.

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:

- Initialize the population of records.

- Evaluate all individuals in the present population record, using fitness value to each one.
- If the termination criterion was reached, then go to last step.
- Otherwise
- Reproduce the best n individuals (records) into the next generation population.
- Select m individuals (records) that will compose the next generation with the best parents.
- Apply the genetic operations like mutation or cross over to all individuals (records) selected. Their offspring will compose the next population. Replace the existing generation by the generated population and go back to Step 2.
- Present the best individual records in the population as the output of the evolutionary process.

Disadvantages

- The optimization of this process is less.
- Certain optimization problems cannot be solved by means of genetic algorithms. This occurs due to poorly known fitness functions which generate bad chromosome blocks in spite of the fact that only good chromosome blocks cross-over. There is no absolute assurance that a genetic algorithm will find a global optimum. It happens very often when the populations have a lot of subjects.

III. PROPOSED FRAMEWORK

This proposed framework consist of several phases after collecting the dataset the raw dataset has to be preprocessed using unstructured to structured conversion for precise processing. The records are converted to vectors using UDD process. The next step is to identify the Weight of the components using WCSS. After that the firefly algorithm was adopted for identifying and eliminating duplicate records. The comparison is done on the existing duplication removal approaches with the proposed algorithm.

The firefly algorithm (FA) is a meta heuristic algorithm, inspired by the flashing behaviour of fireflies [13]. The brightness of a firefly is affected or determined by the landscape of the objective function to be optimized [14], [15]. The primary purpose for a firefly's flash is to act as a signal system to attract other fireflies and find the duplicate records based on the flashing behaviour of the each fireflies. Xin-She Yang [12] formulated this firefly algorithm by assuming:

- All fireflies are unisexual, so that one firefly will be attracted to all other fireflies;
- Attractiveness is proportional to their brightness, and for any two fireflies, the less bright one will be attracted by (and thus move to) the brighter one; however, the brightness can decrease as their distance increases;
- If there are no fireflies brighter than a given firefly, it will move randomly

Advantages

- It is easy to implement and there are few parameters to adjust.
- Compared with GA, all the fireflies tend to converge to the best solution quickly even in the local version in most cases

Steps for Firefly algorithm (FA)

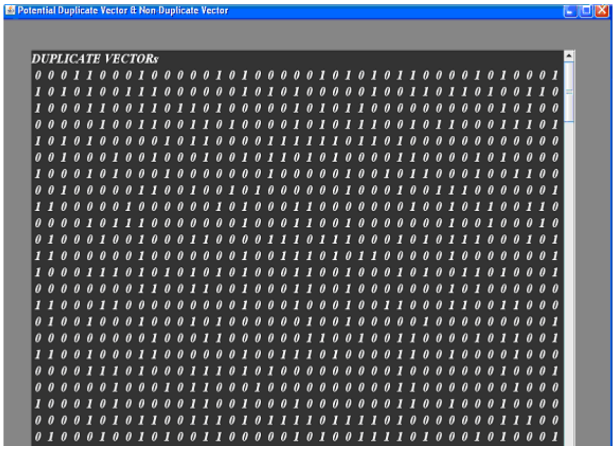
- Generate the objective function and initial population of fireflies based on number of records
- Each records matching is considered like light intensity
- Start comparing with each and every records i.e., fireflies
- Based on the attractiveness the records got matched.
- Follow the steps till all the records got compared.

Output: Number of duplicate and non duplicate records.

IV. EXPERIMENTAL RESULT

In this paper for conducting experiment Cora data set [17] is used. The first real data set, the Cora data set, is a collection of 1,295 distinct citations to 62 computer science papers taken from the Cora research paper search engine. These citations were divided into different attributes (author names, year, title, venue, and pages and other info) by an information extraction system.

The proposed system was implemented using java. The performance study of our proposed method for record deduplication using the optimized approach based on firefly algorithm is compared with two different existing techniques namely Support Vector Machine, and genetic algorithm.



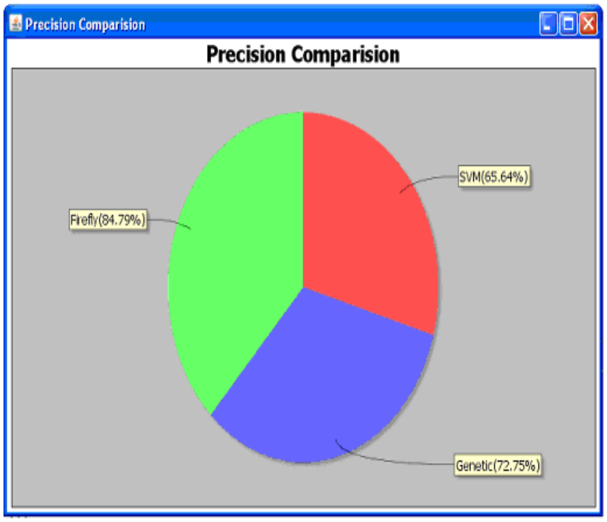
```
DUPLICATE VECTORs
0 0 0 1 1 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 1 1 0 0 0 0 1 0 1 0 0 0 1
1 0 1 0 1 0 0 1 1 1 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 1 0 0 1 1 0 1 1 0 1 0 0 1 1 0
1 0 0 0 1 1 0 0 1 1 0 1 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0
0 0 0 0 0 1 0 0 1 1 0 0 1 1 0 1 0 0 0 0 1 0 1 0 1 1 1 0 0 1 0 1 1 0 0 0 1 1 1 0 1
1 0 1 0 1 0 0 0 0 0 1 0 1 1 0 0 0 0 1 1 1 1 1 1 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 1 0 1 0 1 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0
1 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 1 0 1 1 0 0 0 1 0 0 1 1 0 0
0 0 1 0 0 0 0 0 1 1 0 0 1 0 0 1 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 1 1 1 0 0 0 0 0 0 1
1 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 1 0 1 1 0 0 1 1 0
0 0 0 0 1 0 1 1 1 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0
0 1 0 0 0 1 0 0 1 0 0 0 1 1 0 0 0 0 1 1 1 0 1 1 1 0 0 0 1 0 1 0 1 1 1 0 0 0 1 0 1
1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 1 1 0 1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1
1 0 0 0 1 1 1 0 1 0 1 0 1 0 1 0 1 0 0 0 1 1 0 0 1 0 0 0 1 0 1 0 0 1 1 0 1 0 0 0 1
0 0 0 0 0 0 0 0 1 1 0 0 1 1 0 0 1 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0
1 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 1 1 0 0 0 1 1 0 0 1 1 0 0 0
0 1 0 0 1 0 0 0 1 0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 1 1 0 0 1 0 0 1 1 0 0 0 0 1 0 1 1 0 0 1
1 1 0 0 1 0 0 0 1 1 0 0 0 0 0 0 1 0 0 1 1 1 0 1 0 0 0 0 1 1 0 0 1 0 0 0 0 1 0 0 0
0 0 0 0 1 1 1 0 1 0 0 0 1 1 1 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1
0 0 0 0 0 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0
1 0 0 0 1 0 1 0 0 0 0 0 1 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0 1 0 0 0 0
0 0 0 0 1 0 1 0 1 1 0 0 1 1 1 0 1 0 1 1 1 1 0 1 1 1 1 0 1 0 0 0 0 0 0 0 1 1 1 0 0
0 1 0 0 0 1 0 0 1 0 1 0 0 1 1 0 0 0 0 0 1 0 1 0 0 1 1 1 1 0 1 0 0 0 1 0 1 0 0 0 1
```

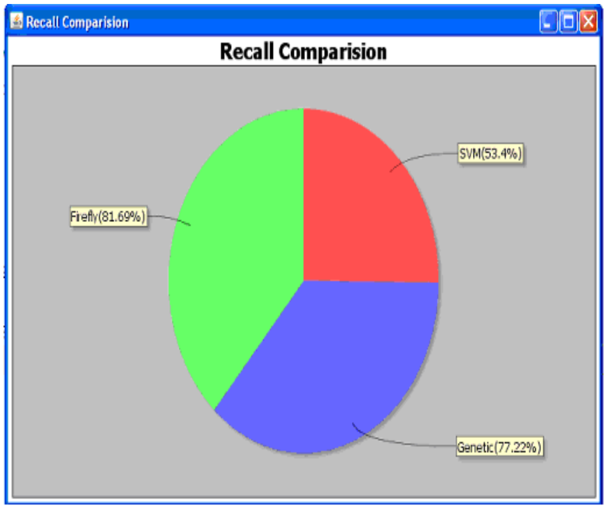
The above screen shows how the given dataset got converted to the vectors.

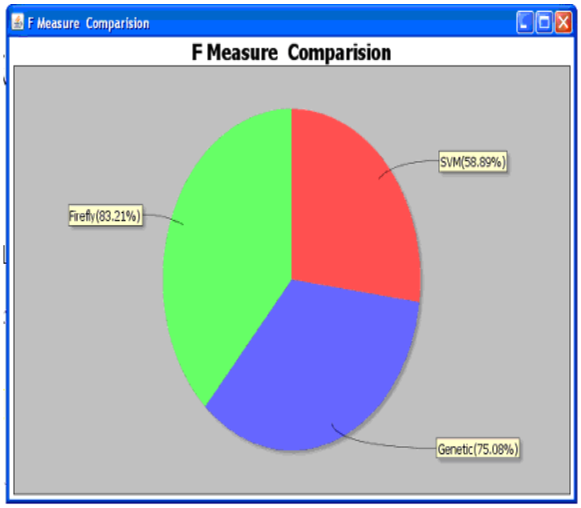
TABLE 1:
PERFORMANCE COMPARISON OF PROPOSED APPROACH WITH OTHER METHODS

Algorithm	Precision	Recall	F-Measure
Firefly	84.79 %	81.69%	83.21%
Genetic	72.75%	77.22%	75.08%
SVM	65.64%	53.4%	58.89%

The metrics used for comparison are precision, recall, f-measure and time taken. The table 1 shows the performance of firefly algorithm, genetic and the Support Vector Machine algorithm.







From the above result, number of duplicate records found in firefly algorithm is higher when compared to other two algorithms based on fmeasure, precision and recall.

V. CONCLUSION

Record deduplication task is important to identify the original and duplicate records in the dataset. To identify duplicate record this paper proposed a Firefly based record deduplication task which improves the duplicate detection performance by comparing the results with parameters like precision, recall, F-measure and time. In which for experimental evaluation F-measure, precision and recall measurements are used. The F-measure harmonically combines the traditional precision (P)and recall (R) metrics commonly used for evaluating accuracy F measure accuracy measured according to the precision and recall measurement. The result shows that the proposed optimization approach using firefly algorithm has 84.79% precision, 81.69% recall and F-measure 83.2%. Next to the firefly, genetic algorithm gives better performance and the worst case is support vector machine. The time complexity is also considerably lessen by our method.

REFERENCES

[1] A.K. Elmagarmid, 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.

[2] M. Bilenko and R.J. Mooney, “Adaptive Duplicate Detection Using Learnable String Similarity Measures,” Proc. ACM SIGKDD, pp. 39-48, 2003.

[3] W.E. Winkler, “Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage,” Proc. Section Survey Research Methods, pp. 667-671, 1988

[4] M.A. Jaro, “Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida,” J. Am. Statistical Assoc., vol. 89, no. 406, pp. 414-420, 1989

[5] 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

[6] 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

[7] Subi S, Thangam P , An Optimized approach for record deduplication using mbat algorithm, International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 6 June, 2013 Page No. 1874-1878

[8] 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

[9] 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

[10] Bhattacharya I 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.

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

[12] X. S. Yang, "Firefly algorithms for multimodal optimization", Stochastic Algorithms: Foundations and Appplications (Eds O.Watanabe and T. eugmann), SAGA 2009, Lecture Notes in Computer Science, 5792, Springer-Verlag, Berlin, pp. 169-178, 2009

[13] X. S. Yang, (2010). "Firefly Algorithm Stochastic Test Functions and Design Optimization". Int. J. Bio-Inspired Computation, vol.2, No. 2, pp.78-84, 2010

[14] X.-S. Yang, "Firefly Algorithm, Lévy Flights and Global Optimization", Research and Development in Intelligent Systems XXVI (Eds M. Bramer, R. Ellis, M. Petridis), Springer, pp. 209-218, 2010.

[15] X. S. Yang, "Engineering Optimization: An Introduction with Metaheuristic Applications". Wiley & Sons, New Jersey, 2010.

[16] L.M. Manevitz and M. Yousef, "One-Class SVMs for Document Classification," J. Machine Learning Research, vol. 2, pp. 139-154, 2001

[17] A. McCallum, "Cora Citation Matching," http://www.cs.umass.edu/~mccallum/data/cora-refs.tar.gz, 2004.

[18] V.P.Archana Linnet Hailey, N.Sudha, "An Optimization Approach of Firefly Algorithm to Record Deduplication" International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 2 Issue 9, September – 2013.

BIOGRAPHIES

R. Gayathri received her MCA and MPhil degree in Computer Science from Bharathiar University in 2007 and 2011 respectively. Presently she is pursuing Phd., degree in Computer science. She worked as assistant professor in V.L.B. Janakiammal College of Arts and Science, Coimbatore from 2007 to 2009. Presently she is working at Dr. G.R.D. College of Science, Coimbatore. She has published seventeen papers in various national, international conferences and journals. Her area of interest is data mining and networks.

A. Malathi received her Phd., degree in computer science. She worked in various colleges and presently she worked as Assistant Professor in PG and Research Department of Computer Science, Government Arts College, Coimbatore. She also published many papers in international journals. Her area of interest is data mining.