International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 9, September - 2013

An Optimization Approach of Firefly Algorithm to Record Deduplication

V.P.Archana Linnet Hailey,
M.Phil student, Department of Computer science,
Bishop Appasamy College of arts & Science,
Coimbatore -18, TamilNadu, India.

N. Sudha,
Asst. Professor, Department of Computer Science,
Bishop Appasamy College of Arts & Science,
Coimbatore-18, TamilNadu, India.

Abstract

In the field of data mining, the task of identifying the data records in the data warehouse to facilitate to equivalent real world entity in spite of misspelling language, special letter styles or even curious schema representations or data type is called as record deduplication. The main task of record deduplication is the task of identifying the replica in the records and to find the original data from its data repositories. In Genetic programming approach record Deduplication, works to find the replica records only in local repository and not in all records, when compared to other optimization it becomes less efficient. This new system introduces a Firefly algorithm (FA) based record deduplication that discovers or identifies more replica records in data warehouse than the GP approach .Firefly algorithm is one of the optimization algorithms and is inspired by fireflies' behaviour in scenery.

1. INTRODUCTION

Finding duplicate records in individuals records from data collected at various sources are more and more important task. Data linkage and deduplication can be second-hand to improve data quality and integrity, which helps to re-use of obtainable of existing data sources from new studies and to reduce costs and efforts in obtaining data. Traditional methods for collecting duplicate records are time consuming and expensive survey methods. To find the duplicate records in the dataset the genetic programming algorithm helps to identify the records in the dataset and solves the deal with the classification problem following a supervised approach [1], i.e., they believe to facilitate all fitness cases (examples) obtainable to evaluate their models are labelled. It solves the data deduplication problem in several areas such as spam detection and text and protein categorization, a lot of human effort is required to label the training data [2]. It reduces considerably the time required for data labeling while maintaining acceptable accuracy rates.

Semi-supervised methods work with a mixture of labeled and unlabeled data, and can be second-hand together in the contexts of classification and clustering [3]. Here we focus on semi-supervised method for classification. Numerous methods subsequent to this approach have been previously proposed, together with self-training [4] and co-training [5]. Nevertheless, we are not aware of any categorization method based on genetic programming next a semi-supervised approach, even though genetic semi-supervised clustering methods have already been proposed [6].

The data mining techniques can be practical when the information available is appropriate in a suitable format. To obtain this, information from various sources and repositories is to be prearranged. Numerous of the obtainable web data are in unstructured form. This unstructured information cannot be omitted because it contains valuable information. Thus this data is to be integrated to structured database to enable mining activities. This can be done using highly performing models, Conditional Random Fields, semi-markov models and matching [3].

This project describes the de-duplication of the records in databases. Genetic algorithms are well suited to solving production scheduling problems, because unlike heuristic methods genetic algorithms operate on a population of solutions rather than a single solution. Genetic algorithms are ideal for these types of problems where the search space is large and the numbers of feasible solutions are small. A specific sequence of tasks and start times (genes) represents one genome in our population. It solves the duplicate records in the dataset. To improve the record deduplication task in the dataset we proposed a firefly algorithm based deduplication task, it identifies the more records in the result than the existing genetic programming based approach.

2. RELATED WORK

There are several problem occurs by collection of the original data from different data sources ,before studying the proposed system of the FA with record deduplication first need is to identify the problems of the existing system and what are the steps followed by earlier work to solve the deduplication. The representation of the replica data by the way only system identify the deduplication task .Previous work of the replica of documents is made for OCR documents. This leads to inconsistency amongst the data stored in repositories. It becomes more difficult when a user need to obtain user-specified information from huge amount of data stored in large databases like repositories. First the information or data in the individual record is converted into some structured data and stored in databases with ideal structure this increases performance and accuracy. Data integration based semi supervised learning methods was proposed by the Imran R. Mansuriimran et .al [7], they proposed a Semi-Markov model for extracting information from structured data and labelled unstructured data bases of their format, structure and size variations. Yihong Ding [8] proposed an enhanced semi-automatic extraction method using DEG. It can solve the unstructured data to structure data formulation in three ways: first together the essential knowledge is collected and then transformed them into useable form. It can be obtained from any source such as encyclopaedia, a traditional relational database, a general ontology like Mik. Then the collected data are automatically generating the initial data-extraction ontology based on the acquired information and example target documents. After that finally the Gathered information is transformed into XML format and various XML documents are combined to produce a high level schema. Finally user validates the early data extraction ontology which is generated using set of justification documents with OntologEditor.

Gengxin Miao [9] proposed a tag path clustering to extract the web information from the web database using pair wise similarity match. It mainly focuses on how a different tag path repeated in the document, occurrence of the resulting tag path is called visual signals which is compared with estimation how likely these two tag paths represent the same list of objects. But still pair wise similarity match did not speak to the nested data structures or further complicated structure. To represent the data in the correct format and extracted data from data sources study the deduplication task.

Adaptive Duplicate Detection technique was proposed by Bilenko et al [10] with MARLIN (Multiply Adaptive Record Linkage with Induction), it employ a two-level knowledge approach. Primary string similarity measures are qualified for every database ground so that they can provide correct estimate of string distance among values for that ground. Next, a final predicate for detecting duplicate records is well-read from similarity metrics of the individual fields. They again make use of Support Vector Machines and demonstrate that they do better than decision trees. MARLIN can lead to improved duplicate detection accuracy over traditional techniques.

There are several duplicate and non-duplicate pairs which are generated by system and starts with small subsets of pairs of records considered for training the data to characterize the data unique. First the initial classifier is used to predict the status of unlabelled pairs of records. The goal is to search for the unlabelled data pool instances and will improve the accuracy of the classifier at the fastest possible rate. Equally instances in which the learner can straight forwardly calculate the status of the pairs which do not have great effect on the learner by Active-learning-based system. In this method the system can rapidly study the peculiarity of a data set and quickly detect duplicates by only a small number of training data [11].But still it is not appropriate in some places since it always require some training data or various human efforts to generate the matching models.

Weifeng Su [12] proposed an unsupervised learning based duplicate detection UDD (Unsupervised Deduplication Detection) to identify duplicates from the query result records of numerous Web databases for a known query it uses two classifiers. WCSS classifier and SVM classifier .From that WCSS classifier act like weak classifier to identify the similar or positive pairs in the dataset and SVM classifier acts like the learning based classifier to produce the deduplicate record results from the WCSS step. UDD (Unsupervised Deduplication Detection) is purposely premeditated for the Web database scenario. Furthermore UDD focus on study and address the field weight assignment subject rather than on the similarity measure.

3. RECORD DEDUPLICATION WITH GENETIC PROGRAMMING

In general the data collection of the user or individual personal information have been collected from multiple data sources, For example if a person is working somewhere based on their environment he/she maintains more than one address details. Database storage system maintain large database to store individual datum, but all the data or records belong to individual person only, to avoid these problems and

remove the replica data or several records of individual person, all the previous work follows a record deduplication task with data mining methods, but evolutionary based identification of the deduplication records make best result than the general data mining methods. These methods raise the following questions that are performance degradation and quality loss the presence of replicas, also increasing operational costs to keep the data at several places in the database. The problem of detecting and removing duplicate entries in a repository is generally known as record deduplication.

Genetic programming is used to identify the replica records or duplicate records in the dataset. GP approach combines numerous dissimilar pieces of proof that is extracted from the data contented to create a deduplication purpose to recognize whether two or additional entries in a warehouse are replicas or not. Genetic Programming is one of the greatest recognized evolutionary algorithms. Throughout the evolutionary procedure, the individuals are handled and adapted by genetic operation such as reproduction, crossover and mutation by iterative manner that is estimated to offspring better individuals in the subsequent number of generations performed by each process.

The steps of Genetic algorithm are the following:

Initialize the population with random or user provided individuals that is original records.

Estimate fitness value for all individual records in the population by assigning the numeric values randomly.

If the termination criterion is satisfied, then perform.

The last step. Otherwise continue step 5.

Repeat the best n individual's records into the next generation population.

Select m individuals to compile the next production with the best parents.

Perform genetic operations to all records selected at step 6. Their children will create the next population.

Replace the presented production by the generated population and go back to Step 2.

Present the best individual population as the output for deduplication.

The assessment of Step 2 is completed by assigning to an individual a value with the purpose to measure how appropriate that individual solves the problem. The resulting value is also called raw condition and the assessment functions are called fitness functions. The consequences are representing in tree arrangement in this case, the rule is that every probable solution found is located in the tree and evolutionary process is applied for each tree representation in the records. The fitness function is the GP part that is answerable for evaluating the generated individuals all along the evolutionary process. If the selection of the fitness function is wrong results also become poor solution to find the replicas in the dataset. GP was used to find the best grouping function for previously user-selected evidence and automatically selected evidence finally the GP results are tested with dissimilar replica data in the repository. It is able to automatically propose deduplication functions based on evidence at hand in the data repositories. The recommended functions correctly combine the greatest evidence obtainable by the way to recognize whether two or more different record entries are replicas or not, it performs than the existing supervised or semi supervised based learning methods since it is able to routinely select the deduplication functions that improved to fit this Deduplication parameter.

4. RECORD DEDUPLICATION WITH FIREFLY ALGORITHM

The firefly algorithm (FA) is a meta heuristic algorithm, stimulated by the flashing behaviour of fireflies. The most important reason for a firefly's flash is to act as a indicate system to be a focus for other fireflies and find the duplicate records based on the flashing behaviour of the each fireflies and their movements from i to j. Xin-She Yang formulate this firefly algorithm by presumptuous:

All fireflies are unisexual, so as to one firefly will be concerned to all further fireflies;

Attractiveness is comparative to their brightness and for any two fireflies, the fewer bright one will be concerned by the brighter one; conversely the brightness can reduce at the same time as their distance increase;

If there are no fireflies brighter than a specified firefly, it will move at random and selects the best duplicate records combination or evidences that are extracted from data content to find replica or not .

In this proposed Firefly algorithm based record deduplication, the objective function (f(x)) of a given is based on difference pieces of evidences that are extracted from the data comfortable. It helps the fireflies to travel towards best location of duplicate records identification or replica records identified and new attractive locations in order to obtain optimal record Deduplication results than the GP to embed the original without deprivation of quality and robustness. After the evaluation of the initial population the firefly algorithm enters its main loop, which represents the maximum number of generations of the iterations for each firefly to find the best results to record Deduplication task. For each production the firefly with the greatest light intensity ($I_j > I_i$) is chosen as the potential optimal solution. Each and every one firefly is

characterized by their light intensity related with the objective function. Each firefly vary attractiveness with distance r via 〖 exp〗 ⁡(-λ r); the population of n fireflies generates n solutions. Each firefly is changing its location iteratively. Finally Rank the fireflies and find the current best result;

1. Initialize the objective function f(xi), x= (x_1,x_2 ……,x_(d))
2. Generate an initial population for record deduplication with fireflies X_i (i=1,2,……,n)
3. Set Max number of iterations=Max generations, Set t=1
4. Formulate light intensity I so that it is associated with f (xi),

Define absorption coefficient λ
While (t<Maxgeneration)
For i=1: n (for all n fireflies)
For j=1: n (n fireflies)
If(I_j>I_i)
Move firefly i towards j;
x_i^(t+1)=x_i^t+βexp⁡[-γr_ij^2]
End if
Vary attractiveness with distance r via exp⁡(-λ r);
r_(ij)=|(|x_i-x_j |)|2 =√(∑_(k=1)^d▒ 〖(〖x〗 _(i,k)- 〖 x〗 _(j,k) 〗))^2
Evaluate new solutions and update light intensity for each firefly;
End for j
End for i
Rank the fireflies and find the current best duplicate records or replica data;

5. End while;
6. Post-processing the results and visualization of the location;
7. End process

5. EXPERIMENTAL RESULTS

Finally in this section measure the performance of the GP and FA algorithm for deduplication task .Measure the accuracy of the system the Cora dataset are used as experiments conduction process to deduplication. In our experiments, we used Cora dataset to found the duplicate records. It contains the multiple attributes and evidences it can be extracted from dataset with the following attribute consideration. These citations were divided into multiple attributes (author names, year, title, venue, and pages and other info) by an information extraction system.

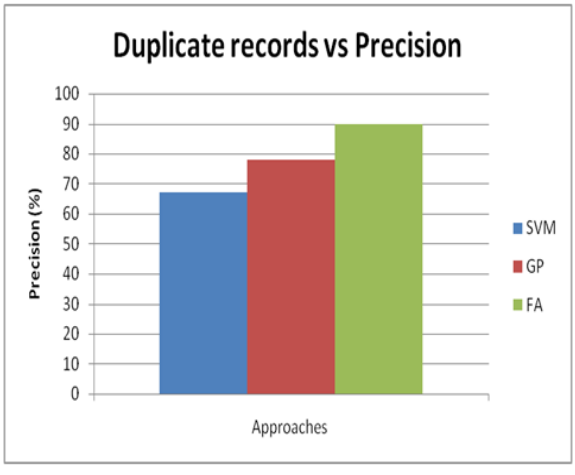

Figure 1. Precision comparison

Figure1 measures the precision value comparison of the records deduplication task with SVM, GP and FA.If the precision value is high more of the duplicate records found by the process, proposed FA are high precision result than the GP, SVM. Corresponding precision values are measured in Y-axis.

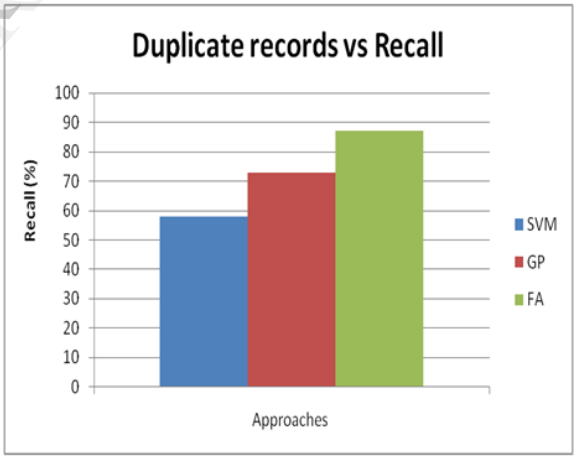

Figure 2. Recall comparison

Figure 2 measures the recall value comparison of the records deduplication task with SVM, GP and FA.if the recall value is high more of the duplicate records found by the process, proposed FA are high recall result than the GP, SVM. Corresponding recall values are measured in Y-axis.

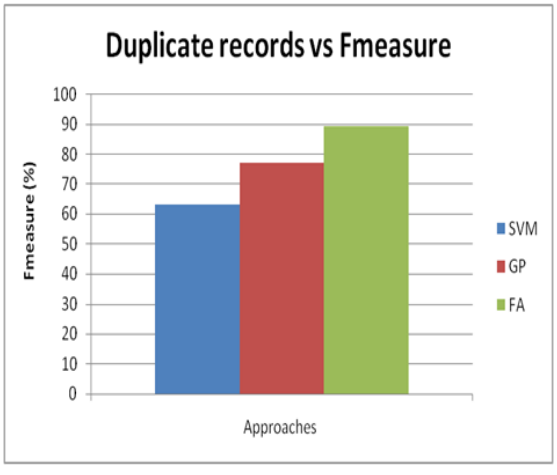

Figure 3. Fmeasure comparison

Figure 3 measures the recall value comparison of the records deduplication task with SVM, GP and FA.if the recall value is high more of the duplicate records found by the process, proposed FA are high recall result than the GP, SVM. Corresponding recall values are measured in Y-axis.

6. CONCLUSION

Deduplication is a very exclusive and computationally challenging task, it is significant to identify which cases our approach would not be the most suitable option. Thus there is a need to examine in which situation our GP-approach would not be the most adequate to use. It combines more than a few different pieces of evidence extracted from the data content and produces the deduplication function, improves the accuracy of the deduplication task result proposed system introduces a FA based deduplication result. FA based algorithm founds the deduplication results based on their firefly movements from i to j and their light intensity update values .It improves the efficiency of the result than the training phase by selecting the most representative examples for training.

REFERENCES

[1] W.Banzhaf, P.Nordin, R.Keller, and F.Francone, "GP – An Introduction; On the Automatic Evolution of Computer Programs and its Applications". Morgan Kaufmann, Jan. 1998.
[2]S.Sarawagi and A.Bhamidipaty, "Interactive deduplication using active learning," in Proc. of the 8th ACM SIGKDD, 2002, pp. 269–278.
[3] O. Chapelle, A. Zien, and B. Sch¨olkopf,Semi-supervised learning. MIT Press, 2006.
[4] C.Rosenberg, M.Hebert, and H.Schneiderman, "Semi-supervised self-training of object detection models," in Proc. of the 7th IEEE Workshop on Application of Computer Vision, 2005, pp. 29–36.
[5] A.Blum and T.Mitchell, "Combining labeled and unlabeled data with co-training," in Proc. of the 11th Annual Conf. on Computational Learning Theory, 1998, pp. 92–100.
[6] Y. Hong, S. Kwong, H. Xiong, and Q. Ren, "Genetic-guided semisupervised clustering algorithm with instance-level constraints," in GECCO '08: Proceedings of the 10th Annual Conf. on Genetic and Evolutionary Computation, 2008, pp. 1381–1388.
[7].Imran R. Mansuriimran@it.iitb.ac.in IIT Bombay, Sunita Sarawagi sunita@it.iitb.ac.in IIT Bombay, "Integrating unstructured data into relational databases".
[8].Yihong Ding, A Thesis Proposal Presented to the Department of Computer Science Brigham Young University, "Semiautomatic Generation of Data-Extraction Ontologies", July 3, 2001.
[9].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".
[10].M. Bilenko and R.J. Mooney, "Adaptive Duplicate Detection Using Learnable String Similarity Measures," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 39- 48, 2003.
[11].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.
[12].Weifeng Su, Jiying Wang, and Frederick H. Lochovsky," Record Matching over Query Results from Multiple Web Databases", Knowledge Discovery and Data Mining, VOL. 22, NO. 4, APRIL 2010.