Exploration of Data Mining Techniques in Record Deduplication

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

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

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

Abstract: ***In today's business world, the database plays a vital role in decision making. As the organization grows, the size of the database also gets increased. This enormous growth in the database size leads to a problem of dirty data. Dirty data is the replicated data in the database which causes some issues like performance degradation, increasing operational cost and the lack of quality. This can be removed by the process of record deduplication. The record deduplication refers to identifying the same entity with different representations. Further cleaning and removing of replica in the repository become a mandatory work. Thus this paper surveys some of the record deduplication approaches. Also it compares with three approaches to record deduplication such as genetic programming, Modified BAT algorithm, and firefly algorithm approach with its limitation and advantages on all the three got discussed.***

Keywords: Record Deduplication, preprocessing, Cleaning, Dirty data, genetic programming, mbat algorithm, firefly algorithm.

1. Introduction

Database is the important source for every organization and that can be derived from different sources. Each heterogeneous source has different representation for same entity, which leads to replica in the repository. Thus large investments are made by organizations to clean the replica from the repository. Data mining is the popular technology which extracts the useful information needed by the organization for taking a better decision. This is the step of KDD (Knowledge Discovery in Databases) process. Fayyad et al. states that, "KDD is the process of identifying a valid, potentially useful and ultimately understandable structure in data" [7]. In KDD process, Preprocessing is the data cleaning stage where the unnecessary information's will be removed. The data cleaning is the process of identifying and correcting the records from the database [4]. It includes parsing, transformation, and removal of duplicates. One common approach to avoid replica, is the record deduplication (also termed as record linkage or data linkage) [3].

The record deduplication is the process of identifying same entity across different data sources. There are various approaches to record deduplication. They are:

1. Adhoc or domain knowledge approaches - based on domain knowledge and uses declarative languages.
2. Training based approaches - based on supervised or semi supervised learning.

This training based approach further classified as Probabilistic Approach and Machine Learning Approach [1]. Figure 1, gives a visual idea of record Deduplication approaches.

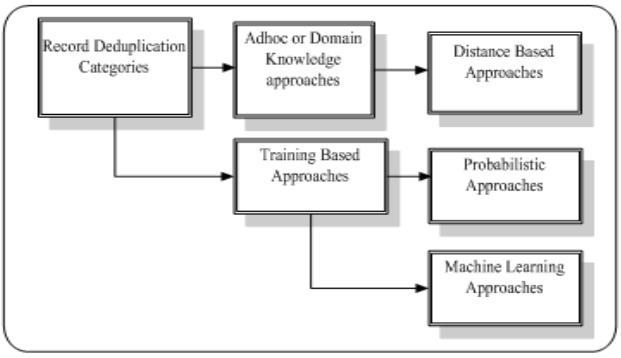

Figure 1: Category of Record Deduplication Approaches

2. Data Preprocessing

Data in the real world may be a dirty data. Dirty data are caused by incomplete, noisy and inconsistent data present in the database. The incomplete data are caused due to lack of attribute values or contains only aggregate data. Noisy data arises when the database contains errors or outliers and few data may be inconsistent with discrepancies in codes or names. Thus the data pre-processing is important step in the data mining process which helps to clean the dirty data from the database or repository.

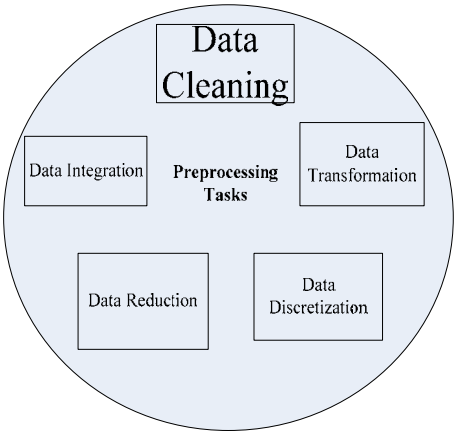

Figure 2: Preprocessing Tasks in KDD Process

3. Genetic Programming Approach to Record Deduplication

Moise's G. de Carvalho proposed a genetic programming approach to record deduplication. In this approach, several different pieces of evidence are extracted from the data content to find a deduplication function. This function helps to identify whether there exist a replica in the repository with only fewer evidence. Genetic programming is used to adapt functions to a given fixed replica identification boundary without the user intervention. The proposed approach has two real data sets. In addition, three additional data sets are created using the synthetic data set generator [1].

The first real data set the Cora Bibliographic data set with the collection of different citations. These citations were divided into multiple attributes by an information extraction system and second real data set, named as restaurant data set, contains 864 entries with 112 duplicates which are grouped from Fodor and Zagat's guidebooks. The synthetic data sets were created using the Synthetic Data Set Generator (SDG) which is available in Febrl Package. In the first set of experiments, the proposed method compares the results between GP-based approach and Marlin. Marlin is a state-of-the-art SVM-based system for record deduplication which is implemented using RBF kernel. The proposed system uses the two steps:

1. GP Framework chooses one file for training purposes.
2. GP Framework tests the results of the training step in all remaining files.

Gabriel .S. Goncalves proposed an approach based on deterministic technique to automatically suggest training phase of de carvalho's al's method based on genetic programming. They used synthetic datasets which show that only 15% of the example suggested by their approach. The proposed work saves training time of up to 85%. The experimental results show that it is possible to use reduced set of training examples without affecting the quality of the obtained solutions and also reduces time necessary for the execution of the training phase. It uses positive and negative pairs of records where positive pairs of records are replicas [2].

Experiments were based on three ways:

Reduction in the percentage of records pairs with positive, negative, positive and negative pairs of records. Thus their proposed work tries to automatically suggest training phase based on genetic programming with less time effort. In future, they suggest placing GUI to get incorporated so that it helps the experimental users to work in easy way.

Baoping Zhang shows on how the combination of citation based information and structural content helps to improve text document classification into predefined categories. They used Genetic programming techniques which indicates as it can discover similarity functions superior to those based on single type of evidence. The empirical shows that the genetic programming has able to discover better similarity functions than genetic algorithm. In Genetic algorithm, the representation will be a fixed length bit string and real numbers where Genetic programming, it is represented as more complex structures. Ex: trees, linked list or stack [11]. Thus it is concluded that their experimental results demonstrates the use of GP framework to discover better similarity functions on two different sets of documents from each level of the ACM Computing Classification System. They also showed in their experiment about the better results on both traditional content-based and combination-based SVM classifiers. Thus their future work includes some parallel computation, testing with different document collections, better citation matching for fixing some OCR errors and also using some different matching strategies.

An´ısio Lacerda proposed a new framework using genetic programming for associating ads with web pages. The use of genetic programming here is to learn functions from the given web page content which select the most appropriate ads. These ranking functions are designed to optimize overall precision and minimize the number of misplacements. They used a real ad collection of web pages from a newspaper with the gain of about 61.7% in average precision [12].

Limitations

- 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.

4. Modified Bat Algorithm for Record Deduplication

Faritha. Banu. A, proposed a novel technique called Modified BAT algorithm for record deduplication where the scheme is initialized and explores with an inhabitant of optimal random solutions by updating bat inventions. The crossover or mutation operators of genetic programming are not used in the proposed work [10].

Contributions of the proposed work

- Keywords are collected from the collection of documents which they gathered at the first stage.
- Uses optimization problem with max and min input set and then computes the value.
- IBAT is a Meta heuristic algorithm that optimized a problem by iteratively improves a candidate solution with regard to a quality measure.

It is based on echolocation activities of microbats with changeable pulse rates of emission and loudness with Doppler Effect.

The experimental results compare the precision, F-Measure, Recall and time comparison between SVM, GP and Modified Bat Algorithm and proved that BAT algorithm was better. Thus this paper concluded that their research work enlarges the optimization of procedure and increases the most represented data samples to get selected. S. Subi

proposed an new meta heuristic method, the Bat Algorithm, based on the echo sound behavior 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 behavior of biological systems and/or physical systems in nature [14]. Experimental results show that the performance evaluation with mbat algorithm is better when compared to genetic programming with the various parameters.

Advantages

- MBAT divides a number of similarities with evolutionary computation procedures such as Genetic Algorithms.

5. Firefly Algorithm For Record Deduplication

V. P. Archana Linnet Hailey, proposed the firefly algorithm (FA) for record deduplication which is a Meta heuristic algorithm, moved by the flashing behavior 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 behavior of the each fireflies and their movements from i to j.

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 [15].

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.

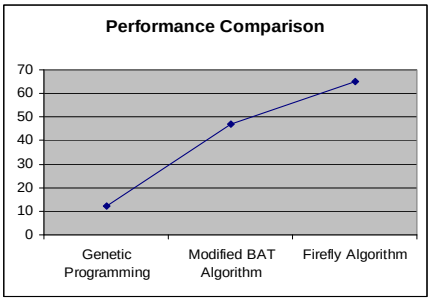

Figure 3: Performance comparison of various algorithms adopted in record deduplication

6. Critique of Existing Approaches

In most cases of real time record deduplication, data mining is the best choice which is more need on the realistic issues of requirements, constraints, and commitment towards removing the replica in the data warehouse.

- There has not been any efficient empirical evaluation of commercial data mining tools for record deduplication.
- Lack of knowledge in handling incomplete dataset.
- Though there is a tremendous growth in the information collection and extraction there is a lack of proper approaches which leads to more time and cost efforts.

7. Discussion

The primary objective of this paper is to give the different approaches to record deduplication, which summarizes, compares and categorizes the relevant approaches in record deduplication. Finally, highlights the advantages of the firefly algorithm where it gives the better performance result when compared to other two algorithms. Also, the second objective of this paper is to compare and states the performance evaluation among all those algorithms.

8. Conclusion

Duplicate records in the organization are increasing nowadays due to enormous collection of data. Thus the record deduplication is the process to remove replica in the repository. In this paper, we discussed on some of the approaches for removing replica in the repository with various scenarios on duplication problems. It also covers some of the disadvantage of genetic programming, modified bat algorithm and how the efficiency gets improved by using the firefly algorithm which uses the optimization technique. In future, the modified firefly algorithm can be implemented for record deduplication with improved efficiency. As this firefly algorithm can be implemented with the flashing behavior, the deduplication of record can be done efficiently when compared to other approaches.

References

[1] Moises G. de Carvalho, Alberto H.F. Laender, Marcos Andre Goncalves, Altigran. S. da Silva, “A genetic programming approach to record deduplication”, IEEE Transactions on knowledge and data engineering, Vol.24, No. 3, March 2012.

[2] Gabriel S. Goncalves, Moises G. de Carvalho, Alberto H. F. Laender, Marcos. A. Goncalves, “ Automatic selection of training examples for a record deduplication method based on genetic programming”, Journal of Information and Data Management, Vol. 1, No. 2, June 2010, pp. 213-228.

[3] http//en.wikipedia.org/wiki/Record_linkage

[4] Lalitha. L, Maheswari. B, Dr.Karthik. S, “A Detailed Survey on Various Record Deduplication Methods”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 8, October 2012, pp. 160-163.

[5] http://en.wikipedia.org/wiki/Data_pre-processing

[6] http://www.techopedia.com/definition/14650/data-preprocessing.

[7] Arun . K. Pujari, Data Mining Techniques.

[8] Glenn Wright, M.P.A, Kaiser Permanente, Oakland, CA, “Probabilistic Record Linkage in SAS”.

[9] http://en.wikipedia.org/wiki/Machine_learning

[10] Faritha Banu, A, Chandrasekar C, “An Optimized Approach of Modified BAT Algorithm to Record Deduplication”, International Journal of Computer

Applications (0975 – 8887) Volume 62– No.1, January 2013.

[11] Baoping Zhang, Yuxin Chen, Weiguo Fan, Edward A. Fox , Marcos Gonc¸alves, Marco Cristo, P´avel Calado, "Intelligent GP Fusion from Multiple Sources for Text Classification"

[12] An´ısio Lacerda1 Marco Cristo1 Marcos Andr´e Gonc¸ alves, "Learning to Advertise", "SIGIR'06, August 6–11, 2006, Seattle, Washington, USA. Copyright 2006 ACM 1595933697/06/0008.

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

[14] S. Subi, P. Thangam, "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.

[15] 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.

Author Profile

R. Gayathri received her MCA and M. Phil 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 is working 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.