International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Record Deduplication Approaches and Algorithm for Removing Duplicate Data

Nikita A. Pande[1], Namrata D. Ghuse[2]

[1]M.E. Ist year (CSE), P. R. Pote COET, Amravati, India

[2]Assistant Professor, P. R. Pote COET, Amravati, India

Abstract: ***In today's world, by increasing the volume of information available in digital libraries, most of the system may be affected by the existence of replicas in their 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. This paper presents an record deduplication techniques and algorithms that detect and remove the duplicate records.***

Keywords: Record Deduplication, Record Linkage, Record deduplication approaches, genetic programming, 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.

In a data repository, a record that refers to the same real world entity or object is referred as duplicate records. And that duplicate record is also called as "dirty data". Due to this dirty data many problem are occurred as follows:

1) Performance degradation—as additional useless data demand more processing and more time is required to answer simple 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) Increased cost —because of the additional volume of useless data, investments are required on more storage media and extra computational processing power to keep the response time levels acceptable.

The problem of detecting and removing these duplicate records from a repository is known as record deduplication. It is also referred as record linkage [1], data cleaning [2]. Data deduplication can be used to improve data quality and integrity, which helps to re-use of existing data sources for new studies, and to reduce costs and efforts in obtaining data. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. In the deduplication process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. Deduplication is a key operation in integrating data from multiple sources.

2. Literature Survey

Record deduplication is a growing research topic in database. Today, this problem arises mainly when data are collected from disparate sources using different information description styles and metadata standards. Other common place for replicas is found in data repositories created from OCR documents. These situations can lead to inconsistencies that may affect many systems such as those that depend on searching and mining tasks. To solve these inconsistencies it is necessary to design a deduplication function that combines the information available in the data repositories in order to identify whether a pair of record entries refers to the same real-world entity.

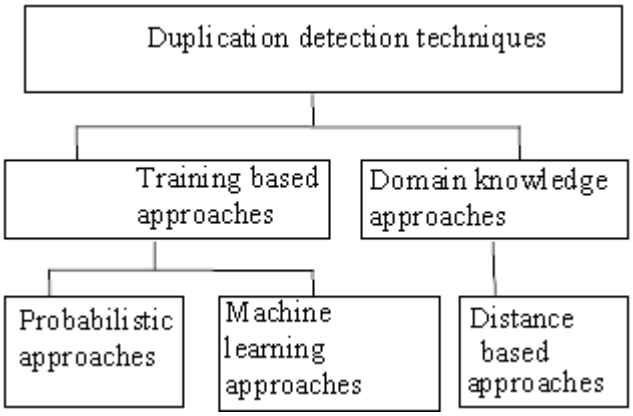

Figure 1: Category of Record Deduplication Approaches

There are various approaches to record deduplication. Elmagarmid et al. [3] classify this approaches in two category:

1) Training based approaches - based on supervised or semi supervised learning.
2) Domain knowledge approaches - based on domain knowledge and uses declarative languages.
3) Following approaches for record deduplication

2.1 Probabilistic Approach

Newcombe et al. [4] were the first ones to address the record deduplication problem as a Bayesian inference problem i.e, a probabilistic problem and proposed the first approach to instinctively handle duplicates. However, their approach was considered empirical [3] since it lacks statistical ground. After Newcombe et al."s work, Fellegi and Sunter [5] proposed a more elaborated statistical approach to deal with this problem. Their method depends on the definition of two boundary values that are used to classify a pair of records as being duplicates or not. It is implemented with Bayes"s rule and Naive based classification.

This method is implemented in the tool such as, Febrl [2], usually work with two boundaries as follows:

a. Positive identification boundary—if the similarity value lies above this boundary, the records are considered to be duplicated.
b. Negative identification boundary—if the similarity value lies below this boundary, the records are considered not to be duplicated.

For the situation in which similarity values lies between the two boundaries, the records are classified as "possible matches or considered as their exist replicas" and, in this case, a human judgment is necessary.

2.1.1 Limitations of Probabilistic Approach

- This method depends on the two boundary values definition that are used to classify a pair of records as being duplicates or not.
- Bad boundaries may increase the number of identification errors.

2.2 Machine Learning Approach

This method apply machine learning techniques for deriving record level similarity functions that combine field-level similarity functions, including the weights of records [6], [7], [8], [9]. It uses a small portion of the available data for training. This training data set is assumed to have similar characteristics to those of the test data set, which makes feasible to the machine learning techniques to generalize their solutions to unseen data. It uses this approach to improve both the similarity functions that are applied to compare record fields and the way the pieces of evidence are combined. The main idea behind this approach is that, given a set of record pairs, the similarity between two attributes (e.g., two text strings) is the probability of finding the score of best alignment between them, so the higher the probability, the bigger the similarity between these attributes.

The adaptive approach presented in [8] consists of using examples for training a learning algorithm to evaluate the similarity between two given names, i.e., strings representing identifiers. We use the term attribute to generally refer to table attributes, record fields, data items, etc. This approach is applied to both clustering and pair-wise matching. During the learning phase, the mapping rule and the transformation weights are defined. The process of combining the transformation weights is executed using decision trees. This system differs from the others in the sense that it tries to reduce the amount of necessary training, depending on user-provided information about the most relevant cases for training. Active Atlas is an object identification system that aims at learning mapping rules for identifying similar records from distinct data sources. The process involves two steps as follows: 1) First, a candidate mapping generator proposes a set of possible mappings between the two set of records by comparing their attribute values and computing similarity scores for the proposed mappings. 2) Then, a mapping rule learner determines which of the proposed mappings are correct by learning the appropriate mapping rules for that specific application domain. This learning step is executed by using a decision tree.

2.2.1 Limitations of Machine Learning Approach:

- It requires large computation and memory storage requirement is high.
- Machine-learning techniques are data oriented i.e, they model the relationships contained in the training data set.

3. Methods / Approach

3.1 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 [10].

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-ofthe- art SVM-based system for record deduplication which is implemented using RBF kernel.

The proposed system uses the two steps:

1) Genetic Programming Framework chooses one file for training purposes.
2) Genetic Progamming 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 [11].

3.1.1 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 [12].

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 [13].

3.1.2 Limitations of genetic programming approach

- 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.

3.2 Firefly Algorithm for Record Deduplication

V. P. Archana Linnet Hailey, proposed the firefly algorithm (FA) is a Meta heuristic algorithm, stimulated 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. Xin-She Yang formulate this 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 .

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 [14].

3.2.1 Advantages of firefly algorithm:

- 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.

4. Discussion

The objective of this paper is to give the different approaches to record deduplication. Finally, it gives the advantages of the firefly algorithm where firefly algorithm gives the better performance result when compared to others.

5. Conclusion

Because of enormous collection of data, duplicate records in the organization are increasing. Thus to remove replica in the repository the record deduplication process is introduced. 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 the limitations or disadvantage of deduplication approach, genetic programming like it requires more memory for deduplication and how the efficiency gets improved by using the firefly algorithm which uses the optimization technique.

6. Future scope

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. In future a deduplication algorithm can be designed for reducing the number of comparison between the records such that it reduces time consumption and utilizes less memory space. Looking into ways to combine different deduplication approaches into a smarter system

References

[1] N. Koudas, S. Sarawagi, and D. Srivastava, “Record Linkage: Similarity Measures and Algorithms,” Proc.

ACM SIGMOD Int'l Conf. Management of Data, pp. 802-803, 2006.
[2] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani, "Robust and Efficient Fuzzy Match for Online Data Cleaning," Proc. ACM SIGMOD Int"l Conf. Management of Data, pp. 313-324, 2003.
[3] 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.
[4] H.B.Newcombe, J.M. Kennedy, S. Axford, and A. James, "Automatic Linkage of Vital Records," Science, vol.130, no. 3381, pp. 954-959, Oct. 1959
[5] I.P. Fellegi and A.B. Sunter, "A Theory for Record Linkage," J. Am. Statistical Assoc., vol. 66, no. 1, pp. 1183-1210, 1969.
[6] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg, "Adaptive Name Matching in Information Integration," IEEE Intelligent Systems, vol. 18, no. 5, pp. 16-23, Sept./Oct. 2003.
[7] 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.
[8] J.R. Koza, Gentic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, 1992.
[9] W. Banzhaf, P. Nordin, R.E. Keller, and F.D. Francone, Genetic Programming - An Introduction: On the Automatic Evolution of Computer Programs and Its Applications. Morgan Kaufmann Publishers, 1998.
[10] 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.
[11] 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.
[12] 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"
[13] 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.
[14] 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

Nikita A. Pande received her B.E (CSE) from P. R. Pote college of engineering and Technology, Amravati Affiliated to Sant Gadge Baba Amravati University, Amravati, Maharashtra, India in 2013. Currenty she is pursuing M.E. in Computer Science and Engineering from P. R. Pote College of Engineering And Technology Amravati, Maharashtra, India.

Namrata D. Ghuse received her M.E. (CSE) from P.R.M.I.T Badnera, Maharashtra, India in 2014. Currently she is working as assistant professor in P. R. Pote College of Engineering and Technology Amravati, Maharashtra, India.