A Dynamic Deduplication Approach for A Data Cleaning Process

[1] G.Rajeswari

[2] Dr V.Srinivas Rao , [3] Dr V. Hima Deepthi

Dept. of CSE, V.R. Siddhartha Engineering College (Autonoums), Kanuru

Affiliated to JNTU Kakinada, A.P India

Abstract:

Digital libraries, E-commerce brokers and similar vast information oriented systems rely on consistent data to offer high-quality services. But presence of duplicates, quasi replicas, or near-duplicate entries (Dirty Data) in their repositories asperses their storage resources directly and delivery issues indirectly. Significant investments in this field from interested parties prompted the need for best methods for removing replicas from data repositories. Prior approaches involved using SVM classifiers, or Genetic Programming (GP) approaches to handle these dirty data. Although performance wise GP systems are better than SVM's, both approaches suffered with processing overheads that requires a pre training to actually implement Deduplication process. So propose to use Active learning Genetic Programming Mechanism a query dependent record matching method that requires semi supervised data set.AGP uses an Active Learning approach in which a committee of multi attribute functions votes for classifying record pairs as duplicate or not. Results shows that AGP guarantees quality of record deduplication while reducing the number of examples was needed.

Index Terms:

Evolutionary computing and genetic algorithms, Information Retrieval, Ranking Functions, Machine Learning.

I. INTRODUCTION

Now a day's information gathering from different resources is main aspect for developing individual assurances but record redundancy is the concept for decreasing individual assurance. Usually built on data gathered from different sources, data repositories such as those used by digital libraries and e-commerce brokers may present records with disparate structure. We call each pair a *database descriptor*, because they tell how the images are distributed in the distance space. By replacing the similarity function, for example, we can make groups of relevant images more or less Compact, and increase or decrease their separation. Feature vector and descriptor do not have the same meaning here. The importance of considering the pair, feature extraction algorithm and similarity function, as a descriptor should be better understood. In CBIR systems, it is common to find solutions that combine image features irrespective of the similarity functions. Our motivation to choose GP stems from its success in many other machine learning applications. Some works, for example, show that GP

can provide better results for pattern recognition than classical techniques, such as Support Vector Machines. Different from previous approaches based on *genetic algorithms* (GAs), which learn the weights of the linear combination function, our framework allows nonlinear combination of descriptors. It is validated through several experiments with two image collections under a wide range of conditions, where the images are retrieved based on the shape of their objects. These experiments demonstrate the effectiveness of the framework according to various evaluation criteria, including precision--recall curves, and using a GA-based approach (its natural competitor) as one of the baselines. Given that it is not based on feature combination, the framework is also suitable for information retrieval from multimodal queries, as for example by text, image, and audio. The great majority of genetic programming algorithms that deal with the classification problem follow a supervised approach, i.e., they consider that all fitness cases (examples) available to evaluate their models are labeled. However, in certain applications, such as data Deduplication, spam detection, and text and protein classification, a lot of human effort is required to label the training data. In scenarios like the aforementioned, methods following a semi-supervised approach might be more appropriate, as they reduce significantly the time required for data labeling while maintaining acceptable accuracy rates. Semi-supervised methods work with a combination of labeled and unlabeled data, and can be used both in the contexts of classification and clustering. Here we focus on semi-supervised methods for classification. Many methods following this approach have been previously proposed, including self-training and co-training. Nonetheless, we are not aware of any classification method based on genetic programming following a semi-supervised approach, although genetic semi-supervised clustering methods have already been proposed. AGP was tailored to solve a challenging database problem: data Deduplication. The main goal of data Deduplication is to identify different records in a database referring to the same real-world entity. This problem was chosen because, given the size of the repositories involved (in the order of millions of records), the process of labeling data can be extremely expensive or even unpractical. Furthermore, in some cases it is hard even for humans to decide if two records are replicas or not in the absence of enough information.

II . RELATED WORK

In [1] record deduplication became the major problem for many of the information oriented systems.Many techiniques has been implemented for record deduplication.(KFINDMR using the most represented data samples) is a technique used to find most represented data samples to improve the accuracy of the classifier.The KFINDMR algorithm calculates the mean value of the most represented data samples in centroid of record members. It selects the first most represented data sample that closest to the mean value and calculates the minimum distance. The system removes the duplicate dataset samples in the system and find the optimization solution to the deduplication of records or data samples.The advantage is that it can achieves higher precision values.But it suffers with a lot of processing overheads.

In [2] the ABC Algorithm is implemented for record duplication problem.This algorithm is used to

generate optimal similarity measure.Once the optimal similarity measure is obtained the deduplication of remainig data sets is done with the help of optimal similarity measure using ABC algorithm. The ABC Algorithm explained with 4 steps. Firstly Similarity pair computation for all the records. Next Feature vectors are calculated. New similarity formulae generation using optimization algorithm.Finally Duplicate detection using the similarity new formulae .The advantage is that it provides better performance compared to previous technique but time consumption for implementing is more.

In [3] The Technique used is Unsupervised duplicate detection for record duplication.It is a query dependent record matching mechanism. It uses 2 co-operating classifiers such as Weighted component similarity summing and support vector machines. Feature vectors data record are calculated.later those are again computed using weighted component similarity summing classifier by assigning weights to the pairs of records.Based on the weights the results are sent to the another cooperating classifier Support Vector Machine Mechanism.This classifier finally updates the deduplicated records finally.The advantage of using this this algorithm is it provides better performance but time consuming is more as more iterations are to be done.

In [4] proposed an an approach based on a deterministic technique that automatically suggests examples for the training phase of de Carvalho et al.'s GP-based record deduplication method. Initially, we verify the real need of using all the training examples generated for the training phase. For this, we performed several experiments in which the examples of duplicated pairs of records were gradually reduced in order to verify how each reduction affected the effectiveness and performance of the process of generating deduplication functions. Next, a deterministic method was used to generate training examples for the deduplication process using GP, allowing an analysis of the viability of automatically selecting these examples. Our experimental results show that it is possible to use a reduced set of training examples without affecting the quality of the obtained solutions in the end of the process of generating deduplication functions, significantly reducing the time necessary for the execution of the training phase. The advantages is it reduces the need of human intervention in the process of creating training examples. The disadvantages are the functions such as edit distance function and jaro similarity function are most adequate for the data types and there is a need to use other deterministic classification methods such as k-means.

III . EXISTING APPROACH

The problem of detecting and removing duplicate entries in a repository is generally known as record Deduplication. Low-response time, availability, security, and quality assurance are some of the major problems associated with large data management. Existence of "dirty" data in the repositories leads to.

Performance Degradation: As additional useless data demand more processing, more time is required to answer simple user queries;

Quality Loss—The presence of replicas and other inconsistencies leads to distortions in reports and misleading conclusions based on the existing data;

Increasing Operational Costs—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. We Proposes a genetic programming (GP) approach to record Deduplication. When there is more than one objective to be accomplished, GP has capability to find suitable answers to a given problem, without searching the entire search space for solutions, which is normally very large. It 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. To reduce computational complexity, this Deduplication function should use a small representative portion of the corresponding data for training purposes. This function, which can be thought as a combination of several effective Deduplication rules, is easy and fast to compute, allowing its efficient application to the Deduplication of large repositories.

GENETIC PROGRAMMING:

Genetic Programming (GP), an inductive learning technique introduced by Koza as an extension to Genetic Algorithms (GA), is a problem-solving system inspired by the idea of Natural Selection. The search space of a problem, i.e., the space of all possible solutions to the problem, is investigated using a set of optimization techniques that imitate the theory of evolution, combining natural selection and genetic operations to provide a way to search for the fittest solution. The *main difference* between GA and GP relies on their internal representation---or data structure---of an individual. In general, GA applications represent each individual as a fixed-length bit string, like a fixed-length sequence of real numbers. In GP, on the other hand, more complex data structures are used.

Flowchart for Genetic Programming

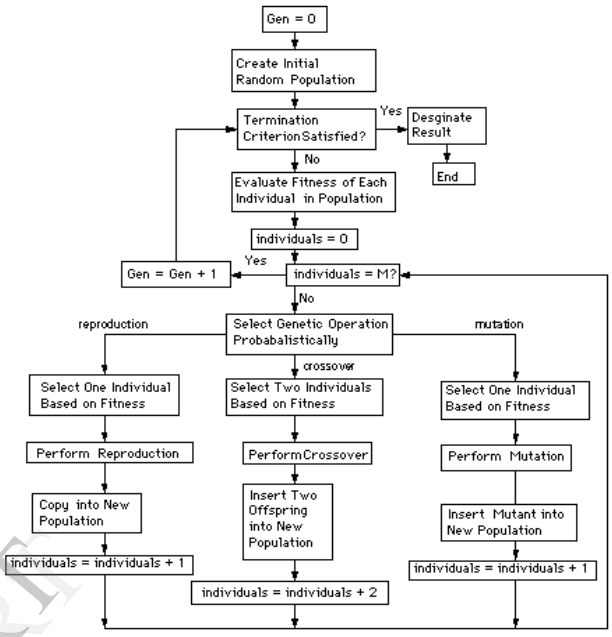

Figure 1: Flow chart for Genetic Programming.

GP searches for good combination functions by evolving a population along several generations. Population individuals are modified by applying genetic transformations, such as *reproduction*, *mutation*, and *crossover*. The reproduction operator selects the best individuals and copies them to the next generation. The two main variation operators in GP are mutation and crossover. Mutation can be defined as random manipulation that operates on only one individual. This operator selects a point in the GP tree randomly and replaces the existing sub tree at that point with a new randomly generated sub tree.

IV.PROPOSED APPROACH

As most of the traditional Deduplication methods that use learning for identifying replicas, AGP also works in three phases: (1) Generates all possible pairs

of candidate records for comparison, exhaustively or through blocking techniques. (2) Calculates a similarity metric between each pair based on their attributes. In this phase, each attribute is manually associated with a well-known distance metric according to its type (i.e., numerical, short or long string). (3) Uses the similarity of each pair to learn how to deduplicate. A semi-supervised approach based on genetic programming and active and reinforcement learning finds a committee (set) of multi attribute functions that classifies a pair as a duplicate or not. Note that, although we focus on the data Deduplication problem, the method proposed here can be easily adapted to any other application domain where labeling examples is an important and expensive process.

V.EMPIRICAL RESULTS

In this section we describe the performance of the active learning genetic programming in data redundancy. In this process we assign different data records into our data repository. Register for uploading file (e.g., text, pdf) in the sequential order with different names with same content present in the data sets. In this process every user can register with particular files in the above same process. After that we are checking the relevancy of the every file present in the user register.

Algorithm:

Evaluate F(Generation$_i$,committee$_i$,pairs)

For each f in generation$_i$ but not in committee$_i$ do

For each p in pairs do

Mp=label(f,p);

Switch(Lp,Mp) do

Case(+,+):$W_f = W_f + W_p$;

Case(-,+): $W_f = W_f - W_p$;

Figure 2: Similarity function release function.

We are applying AGP in the above sequence process for detecting data Deduplication from different files with same content distribution.

5.1 Individual

In the problem of data Deduplication, each individual represents a similarity function between records. The trees that represent the similarity functions are generated using the four basic mathematical operators.

5.2 Process Overview

Initially, a Preprocessing generates a set P of pairs of records from a database DB being deduplicate. Typically, not all possible pairs from DB are in P since some blocking strategy might be used for pruning unlikely pairs. Next, a similarity function sim is deployed for estimating the similarity between records in each pair.

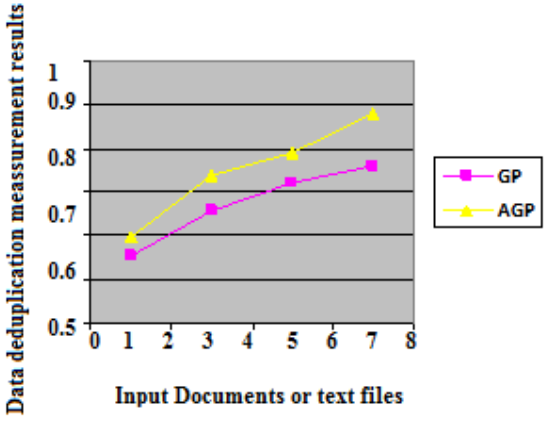

Figure 3: Comparison of data redundancy results in both GP&AGP with training and semi training data sets.

In this way we are calculating every user individuals and process state present in the every users. We are finding similarity function results of

every individual user perspective in the commercial way.

VI.CONCLUSION

Genetic Programming (GP) approaches to handle these dirty data. Although performance wise GP systems are better than SVM's, both approaches suffered with processing overheads that requires a pre training to actually implement Deduplication process. In this paper we propose a semi-supervised approach based on genetic programming and active and reinforcement learning finds a committee (set) of multi attribute functions that classifies a pair as a duplicate or not. In our approach we also increase the performance complexity.

VII.REFERENCES

[1] 1Deepa Karunakaran and 2Rangarajan Rangaswamy, Optimization Techniques To Record Deduplication, Journal of Computer Science 8 (9): 1487-1495, 2012.

[2] P.Shanmugavadivu#1, N.Baskar, "An Improving Genetic Programming Approach Based Deduplication Using KFINDMR",*International Journal of Computer Trends and Technology- volume3Issue5- 2012,*

[3] Gabriel S. Gonçalves, Moisés G. de Carvalho, Alberto H. F. Laender, Marcos A. Gonçalves"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, Pages 213–228.

[4] 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.

[5] 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.

[6] T.P.C. Silva, E.S. de Moura, J.M.B. Cavalcanti, A.S. da Silva, M.G. de Carvalho, and M.A. Gonc¸alves, "An Evolutionary Approach for Combining Different Sources of Evidence in Search Engines," Information Systems, vol. 34, no. 2, pp. 276-289, 2009.

[7] Bilenko, M. and Mooney, R. J. Adaptive duplicate detection using learnable string similarity measures. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, pp. 39–48, 2003.

[8] Christen, P. Febrl: a freely available record linkage system with a graphical user interface. In Proceedings of the Australasian Workshop on Health Data and Knowledge Management. Wollongong , NSW, Australia, pp. 17–25, 2008.

[9]Qingwei, Y., W. Dongxing, Z. Yu and W. Xiaodong, 2010. The duplicated of partial content detection based on PSO. Proceedings of the IEEE 5th International Conference on Bio-Inspired Computing: Theories and Applications, Sept. 23- 26, IEEE Xplore Press, Changsha, pp: 350-353. DOI: 10.1109/BICTA.2010.5645302

[10] Kumar, J.P. and P. Govindarajulu, 2009. Duplicate and near duplicate documents detection: A review. Eur. J. Sci. Res., 32: 514-527.

[11] Samanta , S. and S. Chakraborty, 2011. Parametric optimization of some non-traditional machining processes using artificial bee colony algorithm. Eng. Appli. Art. Intell., 24: 946-957. DOI: 10.1016/j.engappai.2011.03.009.

[12] Sarawagi and A. Bhamidipaty, "Interactive Deduplication Using Active Learning," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 269-278, 2002.

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

[14] T.P.C. Silva, E.S. de Moura, J.M.B. Cavalcanti, A.S. da Silva, M.G. de Carvalho, and M.A. onc¸alves, "An Evolutionary Approach for Combining Different Sources of Evidence in Search Engines," Information Systems, vol. 34, no. 2, pp. 276-289, 2009.

[15)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.

[16] Nick Koudas, Sunita Sarawagi, Divesh Srivastava," Record Linkage: Similarity Measures and Algorithms", *SIGMOD 2006*, June 27–29, 2006, Chicago, Illinois, USA. Copyright 2006 ACM 1-59593-256-9/06/0006.

[17] Ricardo da S. Torres, Alexandre X. Falcão, Marcos A. Gonçalves, João P. Papa, Baoping Zhang,"A genetic programming framework for content-based image retrieval", *R.S. Torres et al. / Pattern Recognition 42 (2009) 283 – 292.*

[18] Humberto Mossri de Almeida, Marcos Andr´e Gonc¸ alves, Marco Cristo, P´avel Calado," A Combined Component Approach for Finding Collection-Adapted Ranking Functions based on Genetic Programming", *SIGIR'07,* July 23–27, 2007, Amsterdam, The Netherlands. Copyright 2007 ACM 978-1-59593-597-7/07/0007.

[19] Humberto Mossri de Almeida, Marcos Andr´e Gonc¸ alves, Marco Cristo, P´avel Calado," A Combined Component Approach for Finding Collection-Adapted Ranking Functions based on Genetic Programming", *SIGIR'07,* July 23–27, 2007, Amsterdam, The Netherlands. Copyright 2007 ACM 978-1-59593-597-7/07/0007.

[20] Junio de Freitas, Gisele L. Pappa, Altigran S. da Silva, Marcos A. Gonc¸alves," Active Learning Genetic Programming for Record Deduplication", in Proc. of the 8th ACM SIGKDD, 2002, pp. 269–278.

[21] 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.

[22] D. A. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Mach. Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[23] A. Beygelzimer, S. Dasgupta, and J. Langford, "Importance weighted active learning," in *ICML '09: Proc. of the 26th Annual Int. Conf. on Machine Learning*. New York, NY, USA: ACM, 2009, pp. 49–56.