International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 11, November - 2013

An Efficient Technique for Removing Duplicates in A Dataset

Dyuthi Raj S
P G Scholar
Department Of Computer Science and Engg
College Of Engineering, Perumon
Kollam, Kerala, India

Remya R
Assistant Professor
Department Of Computer Science and Engg
College Of Engineering, Perumon
Kollam, Kerala, India

Abstract— **In this modern era, every organization compete in order to provide high quality services to their clients. To achieve this they have to maintain replica free data's in their repositories. Government and private organizations invest a large amount of money for removing the replicas from their data repositories. The main reason behind this is the fact that replica free repositories not only provide higher quality data but also it saves the valuable time and resources to process the data. In this paper, record deduplication is performed by using a genetic programming approach that combines the best evidences available from the data set to find a deduplication function that identifies whether two records are replica or not The deduplication function that we selected are less computationally demanding since they use best evidences. This approach frees the user from the burden of selecting the best deduplication function since it automatically combines the best evidences available from the dataset. With this deduplication function, the fitness value for each record is calculated and it is used for performing the record deduplication.**

***Keywords*— Database administration, genetic algorithm, database integration**

I. INTRODUCTION

Databases play an important role in today's IT based market. Many organizations and systems depend on the accuracy and quality of databases for performing their operations. So the quality of the data stored in the databases, can have considerable cost entanglement to a system that relies on information to function and conduct the business. Database administration is a job whose primary function is to provide the overall aid for a computer database. These operations are carried out by a person called administrator or database administrator. Databases require consistent administration and maintenance, and a DBA is specially practiced to perform all of the functions essential to do so. As the amount of information increases the database administrator faces several problems such as how to maintain data availability, security, quality assurance, privacy, searching etc. The data that is available in the data repositories such as digital libraries and e-commerce brokers are obtained by gathering data from different data sources and these data may be of different structures.

The existence of dirty data (i.e. replicas, data without any standard representation etc) in data repositories can cause several problems like performance degradation, quality loss and increased operational cost and time. Dirty data is not always bad. It depends on the correct management of that data. In order to avoid the above specified problems, it is necessary to study the reason of "dirty" data in repositories. During the aggregation or integration of different data sources there occurs duplicates, quasi replicas or near duplicates in these repositories and that is the major root for the existence of dirty data. So it is necessary to detect and remove the duplicate entries in the data repositories. This problem is known as record deduplication

Actually record deduplication is the mission of identifying, in a data repository, whether two records that refer to the same real world item or object in spite of misspelling terms, typos, different writing styles or even different schema representations or data types.

In this paper, we present a genetic programming (GP) approach to record deduplication that combines the best pieces of evidences extracted from the dataset to produce a deduplication function which can be able to identify whether two or more entries in a repository are refer to the same real world object or not. Actually record deduplication is a time consuming task even for small repositories and it is difficult for the user to select best evidences from the dataset. So our aspire is to find a method that finds an apt combination of the best pieces of evidence that is present in the dataset, thus obtain a deduplication function that maximizes performance using a small typical portion of the corresponding data for training purposes. This resultant function can also be used on the left over data or even applied to other repositories with similar characteristics. With that deduplication function we

have to calculate the gene value for each record for performing the record deduplication.

II. LITERATURESURVEY

Record deduplication is a budding research topic in database and related fields such as digital libraries. When data is collected from different data sources using different information description styles and metadata standards, it gives a way to occur this problem. In order to get rid from these inconsistencies it is essential to design a deduplication function that properly combines the evidences available in the data repositories in order to identify whether a pair of record entries refers to the same real-world entity. In the area of bibliographic citations, this difficulty was extensively discussed by Lawrence et al. [1], [2]. Based on word matching, phrase matching, edit distance and subfield extraction they propose a number of matching algorithms. Word and phrase matching gives improved result.

Initially the user has to go through the entire dataset in order to perform the record deduplication. But it was a time consuming and complex work as the dataset contains thousands of records. Later many works have been proposed many approaches to combine and use the evidences extracted from the dataset since several strategies for extracting evidences become available. Elmagarmid et al. [3] grouped these approaches into the following two categories: 1) Ad-Hoc or Domain Knowledge Approaches—It mainly includes the approaches that depends on the domain knowledge. Techniques that make use of declarative languages [3] can be also classified in this category; 2) Training-based Approaches—Approaches that needs some supervised or semi supervised training comes under this category. Probabilistic and machine learning approaches fall into this category. Newcombe et al. [4] proposed the first approach to automatically handle the replicas by considering record deduplication problem as a probabilistic problem. But then also there occurs a problem of how to combine the evidences. Therefore Fellegi and Sunter proposed an approach that uses two boundary values in order to identify whether two records are replicas or not. Febrl [5] implement this method by using the following boundary values

1. Positive identification boundary—the records are considered as replicas if the similarity value lies above this boundary;
2. Negative identification boundary— the records are considered as not being replicas if the similarity value lies below this boundary,

If the similarity value lies in between the two boundaries then the records are classified as possible matches.

Our work closely related to approaches that makes use machine learning technique for achieving record level similarity from field level similarity 6], [7], [8], [9]. For training purposes these approaches use a small portion of the available data. The training and testing data should have similar characteristics so that obtained result can be used to unseen data also. In Marlin [6], [7] the evidences that are extracted are used to train SVM classifier so that they can be combined properly to identify the replicas. Here the similarity between the attributes is the probability of finding the score of their best alignment. *Active Atlas* [9] is a system that computes the similarity score between the fields in the record and base on that the mapping rules are learned. It reduces the training process and before Marlin it is used as the state of the art.In [10], a GP-based approach is used to improve results produced by the Fellegi and Sunter's method [5]. Here the characteristics of similar records are trained during the GP training phase. Then the best tree obtained from the training set are used to identify replicas in a set of records different from the one used in the training set. Thus a better combination of evidences is obtained. Here the deduplication framework is integrated with Febrl and the genetic operation crossover is also not used.

In [11], the number of evidences for comparison can be reduced since GP is used as a method for replica identification independently of any other technique. It requires a similarity metric between the evidences and requires one boundary value. In comparison with all the other techniques our method can automatically suggest the best deduplication function even when the similarity metric between the evidence is not known in advance. Also the suggested function can automatically adapts to changes in the boundary. After obtaining the function the fitness value for each record is calculated .So the number of comparison can also be reduced.

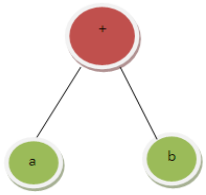
III. PROPOSED SOLUTION

The proposed solution is based on evolutionary programming that influences the natural selection. Genetic programming is the best known evolutionary technique whose ideas come from the properties of genetic operation and natural selection [12][13]. GP provides good performance on searching over the search space which is normally infinite in size. Here a population of individuals is taken rather than a single deduplication function. The reason behind this is the fact that each individual is considered as valuable genetic content and which can be used to create next pool of valuable genetic content after applying the genetic operations. During the generational process the functions are modified by using the genetic operations such as reproduction, cross over and mutation. The population contains individuals or solution to the given problem which are represented as a tree. The leaf of tree are input, constants etc. and internal nodes represent the operators, statements etc. that are used to manipulate the tree leaves.

Reproduction: Copies the selected individual to the new population without modifying it.

Crossover: Create new offspring program(s) for the new population by exchanging the randomly chosen parts from two selected programs.

Mutation: Create one new child program for the new population by replacing a selected subtree in the parent tree by a randomly selected subtree. It results into distinct individuals in the population.



tree(a,b)=a+b

Fig 1. Example of a function mapped as a tree

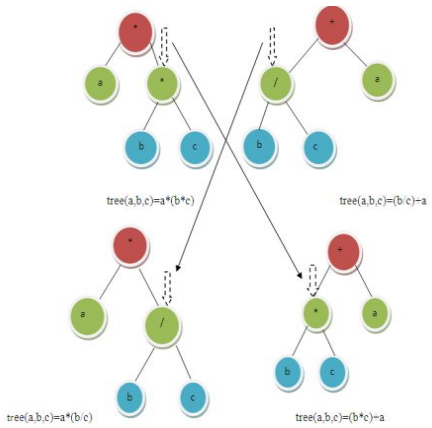


Fig 2. Random Subtree Crossover

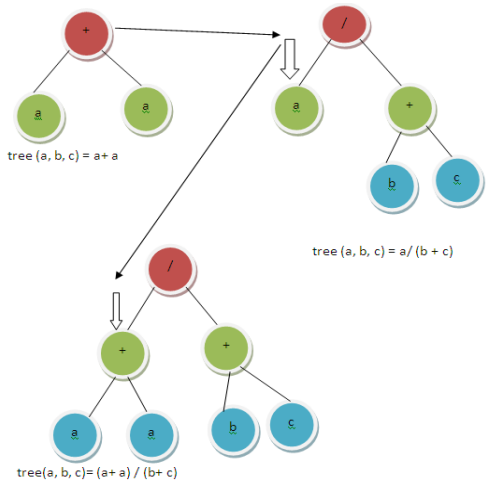


Fig 3. Random subtree Mutation

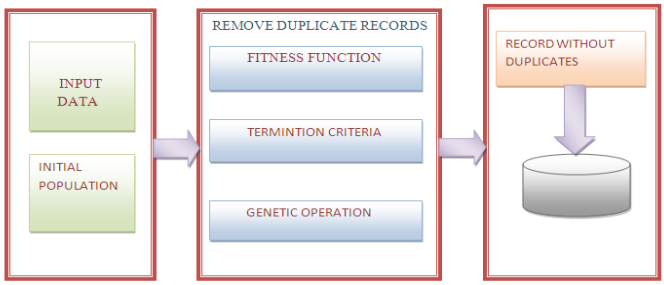


Fig 4. System Architecture

The fig 4 shows the architecture of our proposed system. Initially we have a dataset with duplicate records and a population that contains the selected evidences. Then we will find the best fitness function if it satisfies the condition. Otherwise we have to perform the genetic operations like reproduction, mutation and crossover. With the obtained deduplication function the records without replicas are stored into the database. The proposed algorithm is mentioned below.

Let D be a dataset that contains records of different fields denoted as D= $\{R_1, R_2, R_3, R_n\}$ where each record R_i contains s fields such as Name, Address, City, Phone no, Type, Class. The proposed method has two phases such as training phase and duplicate detection and removal phase. During training phase the best deduplication function is automatically calculated and it is used to detect and remove replicas in the data set during the second phase. The detailed algorithm is mentioned below.

Step 1: Calculate the similarity among the records in the dataset.

In step 1 the similarity between each field of the record pair is calculated using the similarity functions like Levenshtein distance and Cosine similarity.

Levenshtein distance: The Levenshtein distance is obtained by calculating the minimum number of operations such as insertion, deletion, substitution etc. needed to transform one string to another string. Here the record is considered as a single string.

Cosine Similarity: Cosine similarity between the record fields is obtained by first calculating the union of dimension of these fields. Then calculate the frequency of the occurrence vector for both the fields. Finally the dot product and magnitude of both fields is calculated. Hence we obtain the cosine similarity among the records.

Step 2: Extract the best evidence from the available data based on the similarity among the records.

Step 3: Find the fitness function.

The genetic algorithm is used in this step. There are well defined and discrete generation cycles. The individuals are

formed by the combination of extracting evidences. The detailed algorithm is shown below.

1. Initialize the population with random or user selected individuals.
2. Evaluate all the individuals in the current population by assigning value to each individual.
3. If the termination criterion is reached, then go to step 7. Otherwise continue.
4. Reproduce the best n individuals into the next generation.
5. Select the m individuals that will form the next generation with the best parents.
6. Apply the genetic operations to all the individuals selected. Their children form the next population. Replace existing generation by the generated population and return to step 2.
7. Bestow the best individuals in the population as the solution.

Step 4: Remove Duplicate Records.

After selecting the fitness function we have to calculate the fitness value for all the records. The selected fitness function can efficiently remove the duplicate records. If the fitness value of current record matches with the fitness value of other records, then we can eliminate the given record, because it is a duplicate record. Otherwise we can store the record .At the end we get a dataset without duplicate records.

In the genetic algorithm the assigned value is called fitness value and the evaluation function is called fitness function. When we use a genetic approach for our task, then certain requirements should be satisfied. They are mentioned as follows.

1. Every solution should be represented as a tree.
2. After applying the genetic operations, the resultant individual should also be a valid tree.
3. Each tree must be automatically evaluated.

By using this approach the deduplication function can be automatically identified without searching the entire search space which is normally infinite in size. Here our termination criteria is the two boundary values of the dataset. Since the deduplication function is automatically calculated it can also automatically adapt to changes in the boundary values. Also here we will calculate the fitness or gene value of each record by using the obtained deduplication function. Then record deduplication is performed by comparing the fitness value of the record .So the number of comparisons can be reduced and hence improves the performance, can efficiently mine better knowledge from the dataset and reduces the time and cost to process the data.

After doing these comparisons for all record pairs, the whole number of incorrect and correct identified duplicates can be computed. This information is later used for evaluating the fitness function. The fitness function f1 harmonically combines the precision (P) and recall(R) for evaluating the system.

P=TotalNumberOfCorrectlyIdentifiedDuplicatedPairs/TotalNumberOfIdentifiedDuplicatedPairs

R=TotalNumberOfCorrectlyIdentifiedDuplicatedPairs/TotalNumberOfTrueDuplicatedPairs

F= (2 *P *R)/ (P+R)

This metric can be used to evaluate how a specific individual performs in the task of identifying the replicas.

IV .RESULT AND DISCUSSION

In this paper, we presented record deduplication using a genetic programming (GP) approach. This 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. The input to the project is a restaurant dataset that contains duplicate records. The dataset contains fields like name, address, city, phone no:, type and class For extracting the best evidence from the dataset we have to calculate the similarity between the records. For that we use two similarity measures Levenshtein distance and Cosine similarity. These two measures are computed for all attributes of record pairs because different similarity operations have varying significance in different domains. Based on the cosine similarity between the records the values for each attribute in the dataset are calculated .Then based on the values, the best evidences are provided to the population. Also here the fitness values for the deduplication function are calculated based on the values of the attributes. If the fitness value for the deduplication function does not reach the expected value then genetic operations are performed to calculate new deduplication function. Otherwise it is selected as the best deduplication function. After selecting the fitness function we have to calculate the fitness value for all the records. If the fitness value of current record matches with other records then we can remove the given record, because it is a duplicate record. Otherwise we can store the record .At the end we get a dataset without duplicate records.

V.CONCLUSION AND FUTURE SCOPE

Deduplication is a key problem for guaranteeing the quality of the information made available by modern DLs. Digital libraries rely on the integrity of its data content and may be affected by data duplication. In this paper, we presented the record deduplication using the GP approach. Based on the evidence present in the data repositories our approach can automatically suggest the best deduplication function by appropriately combine the best evidence available in order to classify whether two or more distinct record entries are replicas (i.e., represent the same real-world entity) or not. With this function the fitness value for each record is calculated. Then deduplication is performed by comparing the fitness value of the records present in the data set. Thus the number of comparisons needed for performing record deduplication can be reduced. It also improves the performance and mine better knowledge from the dataset. In future we are using PSO algorithm as an alternative for the

genetic algorithm. By using particle swarm optimization algorithm instead of genetic algorithm it can provide better performance and accuracy. In future we can apply this approach to dataset in other domains also. Thus we can extend the range of use of our approach.

ACKNOWLEDGEMENT

First and foremost we would like to thank God almighty for having showered upon us His kindest blessings enabling us to fulfill this task. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them.

REFERENCES

[1] S. Lawrence, C.L. Giles, and K.D. Bollacker, "Autonomous Citation Matching," Proc .Third Int'lConf. Autonomous Agents , pp. 392-393, 1999.

[2] S. Lawrence, L.Giles and K. Bollacker,"Digital Libraries and Autonomous Citation Indexing, " Computer, vol. 32, no. 6, pp. 67- 71, June 1999.

[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] "Freely Extensible Biomedical Record Linkage," http:// sourceforge.net/projects/febrl, 2011.

[6] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg, "Adaptive Name Matching in Information Integration,"IEEE Intelligent System, 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] W. W. Cohen and J. Richman, "Learning to Match and Cluster Large High - Dimensional Data Sets for Data Integration , " Proc . Ninth ACM SIGKDD Int'lConf Knowledge Discovery and Data Mining, pp.475-480.2002

[9] S. Tejada, C.A. Knoblock, and S. Minton, "Learning Object Identification Rules for Information Integration," Information Systems, vol. 26, no. 8, pp. 607- 633, 2001.

[10] M. G. de Carvalho, M. A. Gonc¸alves, A.H.F. Laender, and A.S. da Silva, "Learning to Deduplicate," Proc. Sixth ACM / IEEE CS Joint Conf. Digital Libraries, pp. 41- 50, 2006.

[11] M. G. de Carvlho, A. H. F.Laender, M.A. Goncalves , and A .S. da Silva, " Replica Identification Using Genetic Programming, " Proc. 23rd Ann. ACM Symp. Applied Computing (SAC) , pp.1801 -1806, and 2008.

[12] J.R. Koza, Gentic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, 1992.

[13] 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 KaufmannPublishers, 1998.