Deduplication of Hospital Data using Genetic Programming

Pallavi P. Gujar
Department of computer engineering
Thakur college of engineering and Technology, Kandiwali, Maharashtra, India

Priyanka Desai
Department of Computer Engineering
Thakur College of Engineering and Technology, Kandivali, Maharashra, India

Abstract: The record deduplication is the task of identifying, in a data repository, records that refer to the same real world entity or object in spite of misspelling words, types, different writing styles or even different schema representations or data types. Several systems that rely on consistent data to offer high-quality services, such as digital libraries and e-commerce brokers, may be affected by the existence of duplicates, quasi replicas, or near-duplicate entries in their repositories. Because of that, there have been significant investments from private and government organizations for developing methods for removing replicas from its data repositories. This is due to the fact that clean and replica-free repositories not only allow the retrieval of higher quality information but also lead to more concise data and to potential savings in computational time and resources to process this data. Genetic Programming, new methodology has been proposed to reduce the computation time and resources to process data. Genetic programming combines several different pieces of evidence extracted from the data content to find a deduplication function that is able to identify whether two entries in a repository are replicas or not.

Keywords – Data Deduplication, Duplicate Detection, Genetic Programming, Data Identification

I.Introduction

Database plays an important role in today's IT-based economy. Many industries and systems depend on the accuracy of databases to carry out operations. Therefore, the quality of the information stored in the databases can have significant cost implications to a system that relies on information to function and conduct business. The increasing volume of information available in digital media has become a challenging problem for data administrators. Record deduplication is the task of identifying, in a data repository, records that refer to the same real world entity or object in spite of misspelling words, typos, different writing styles or even different schema representations or data types. Thus, there have been large investments from private and government organizations for developing methods for removing replicas from data repositories. [1]

The quality of the data residing in databases gets degraded due to a multitude of reasons. Such reasons include typing mistakes(e.g., lexicographical errors, character transpositions) during insertion, lack of standards for recording database fields (e.g., person names, addresses), and various errors introduced by poor database design (e.g., update anomalies, missing key constraints). Data of poor quality can result in significant impediments to popular business practices: sending products or bills to incorrect addresses, inability to locate customer records during service calls, or inability to correlate customers across multiple services, etc. [2]

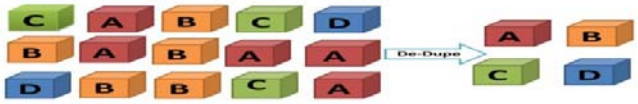

Figure 1: Data Deduplication

II. LITERATURE SURVEY

Record duplication mainly arises when data are collected from disparate sources using different information. The designed technique is a 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. [3]
The common problems are:
1) The existing structured databases of entities are organized very differently from labeled unstructured text.
2) There is significant format variation in the names of entities in the database and the unstructured text.
3) In most cases the database will be large whereas labeled text data will be small. Features designed from the databases should be efficient to apply and should not dominate features that capture contextual words and positional information from the limited labeled data. [5]
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 pair of record entries refers to the same real-world entity.
In the following section different deduplication techniques are discussed. Deduplication of data is mainly done to extract valuable information in spite of misspelling, typos, different writing styles or even different schema representation or data types. [6]

- Domain Knowledge Approaches

The idea of combining evidence to identify replicas has pushed the data management research community to look for methods that could benefit from domain-specific information found in the actual data as well as for methods based on general similarity metrics that could be adapted to specific domains

- Probabilistic Approaches

Newcombe et al. were the first ones to address the record deduplication problem as a Bayesian inference problem (a probabilistic problem) and proposed the first approach to automatically handle replicas. However, their approach was considered empirical since it lacks a more elaborated statistical ground. After Newcombe et al.'s work, Fellegi and Sunter proposed a more elaborated statistical approach to deal with the problem of combining evidence. Their method relies on the definition of two boundary values that are used to classify a pair of records as being replicas or not. Tools that implement this method, such as Febrl, usually work with two boundaries as follows:
1. Positive identification boundary—if the similarity value lies above this boundary, the records are considered as replicas;
2. Negative identification boundary—if the similarity value lies below this boundary, the records are considered as not being replicas.
For the situation in which similarity values stand between the two boundaries, the records are classified as "possible matches" and, in this case, a human judgment is necessary. [4]

- Machine Learning Approaches

The proposals that are more related to our work are those that apply machine learning techniques for deriving record level similarity functions that combine field-level similarity functions. These proposals use 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. The good results usually obtained with these techniques have empirically demonstrated that those assumptions are valid. a GP-based approach to improve results produced by the Fellegi and Sunter's method. [5] Particularly, we use GP to balance the weight vectors produced by that statistical method, in order to generate a better evidence combination than the simple summation used by it. Our experimental results with real data sets show improvements of 7 percent in precision with respect to the traditional Fellegi and Sunter's method. [7]

III.BASICS OF GENETIC PROGRAMMING

Genetic Programming (GP) is a type of Evolutionary Algorithm (EA), a subset of machine learning. EAs are used to discover solutions to problems humans do not know how to solve, directly. Free of human preconceptions or biases, the adaptive nature of EAs can generate solutions that are comparable to, and often better than the best human efforts.

Inspired by biological evolution and its fundamental mechanisms, GP software systems implement an algorithm that uses random mutation, crossover, a fitness function, and multiple generations of evolution to resolve a user-

defined task. GP can be used to discover a functional relationship between features in data (symbolic regression), to group data into categories (classification), and to assist in the design of electrical circuits, antennae, and quantum algorithms. GP is applied to software engineering through code synthesis, genetic improvement, automatic bug-fixing, and in developing game-playing strategies. [8]

The main aspect that distinguishes GP from other evolutionary techniques (e.g., genetic algorithms, evolutionary systems, genetic classifier systems) is that it represents the concepts and the interpretation of a problem as a computer program—and even the data are viewed and manipulated in this way. This special characteristic enables GP to model any other machine learning representation. Another advantage of GP over other evolutionary techniques is its applicability to symbolic regression problems, since the representation structures are variable

- Genetic Programming approach for data deduplication

The data is gathered from various resources. Thus it contains "dirty data". The data without any standard representation and presence of replicas are said to b dirty data. To deal with this problem approach based on Genetic programming is used. Evolutionary programming is based on ideas inspired on the naturally observed process that influence virtually all living beings, the natural selection. Genetic Programming is one of the best known evolutionary programming techniques. It is a direct evolution of programs or algorithms used for the purpose of inductive learning (supervised learning), initially applied to optimization problems. [5]

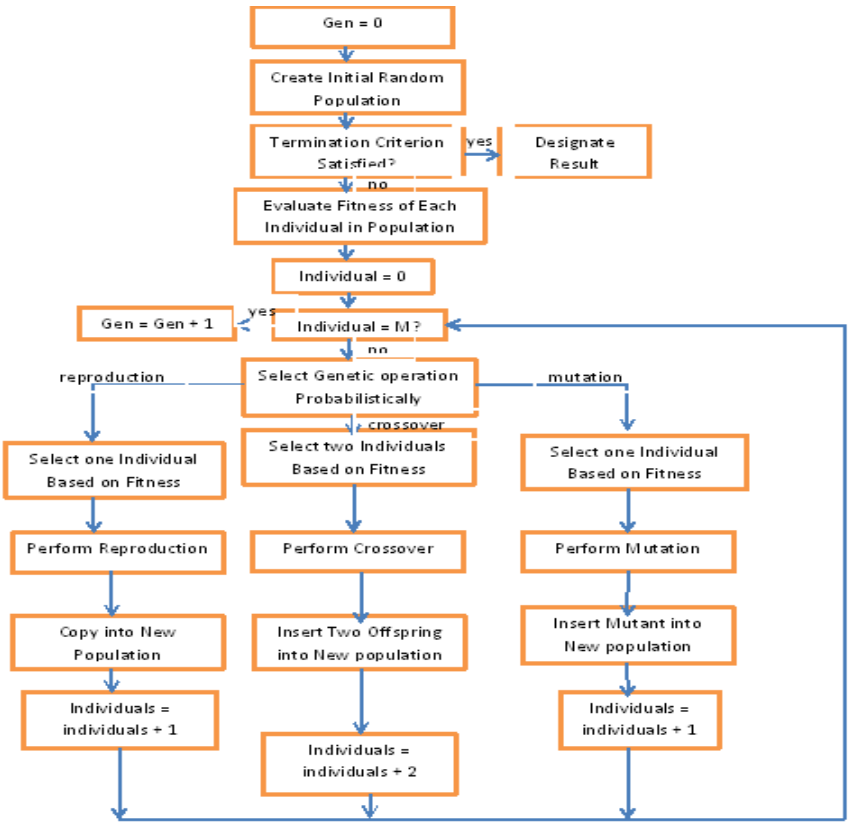

Figure 2: GP Executional steps

IV.EXPERIMENT

The experiment is conducted on the patient's data. Current database contains about 100 entries. The maximum time consumed by the algorithm to find the duplicate entry is about 66 seconds. And the maximum time to enter the details of a patient is about 100 seconds.

In this experiment Jaro-Winkler algorithm is been used. We will compare Jaro-winkler, Levenshtein and SmithWaterman algorithm. For comparison first we will use single string and will compare the algorithms.

Table 1: Comparison according to matching result

Strings	Jaro-Winkler	Levenshtein	SmithWaterman
Pallavi Pallavi	0.90476197	0	14
Ganesh Pallavi	0.43650794	6	2
9619443551 9619443551	0.93333334	0	20
9619443551 9867121343	0	8	4

Table 2: Comparison according to time required

Strings	Jaro-Winkler	Levenshtein	SmithWaterman
Pallavi Pallavi	9 sec	7 sec	7 sec
Ganesh Pallavi	8 sec	11 sec	7 sec
9619443551 9619443551	11 sec	7 sec	5 sec
9619443551 9867121343	8 sec	7 sec	8 sec

Now we will compare the above algorithms according to results and time required for strings “Pallavi” and “Pallavi”.

Table 3: Comparison according to result and time required

Agorithms	Results	Time Taken
Jaro-Winkler	0.90476197	9
Levenshtein	0	7
SmithWaterman	14	7

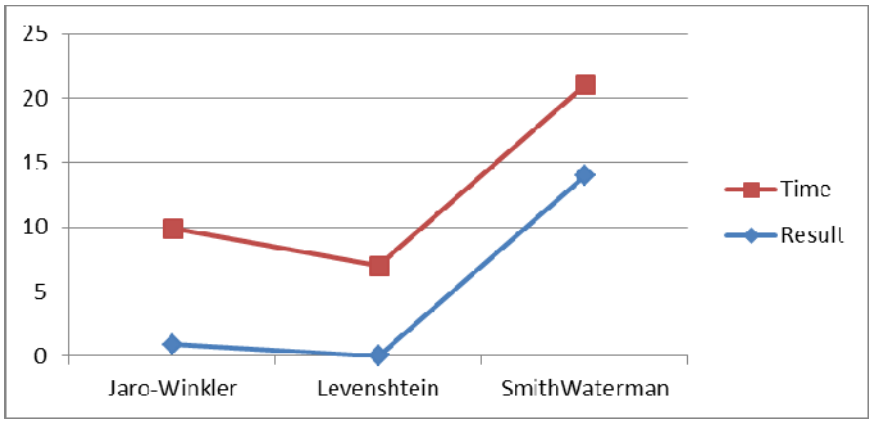

Figure 3: Comparison according to results and time required

V.CONCLUSION

Duplicate detection is an important problem in data cleaning. In real world, there are so many domains where records are getting saved again and again. For the purpose of memory storage in organizations record deduplication has become essential task.

Identifying and handling replicas is important to guarantee the quality of the information made available by data-intensive systems such as digital libraries and e-commerce brokers. These systems rely on consistent data to offer high-quality services, and may be affected by the existence of duplicates, quasi replicas, or near-duplicate entries in their repositories.

GP- based approach is able to automatically suggest deduplication functions based on evidence present in the data repositories. The suggested functions properly combine the best evidence available in order to identify whether two or more distinct record entries are replicas (i.e., represent the same real-world entity) or not.

One of the advantages of using GP is the fact that it allows us to analyse the final generated functions in order to infer the relevance of each evidence for the deduplication task. Moreover, at end of the GP processing, there is not only one solution, but a population of individuals (functions) that can execute the task in an equivalent or similar manner. This allows users to choose the best suitable option to fulfil their needs.

As we have seen in the experiments Jaro-Winkler algorithm is useful for finding the similar records from the database. Genetic programming allows the system to find similar records using crossover mechanism. Use of Jaro-Winkler algorithm allows finding the duplicate records from the database by giving numerical values between 0-9.

REFERENCES

[1] M.G. de Carvalho, M.A. Gonc¸alves, A.H.F. Laender, and A.S. da Silva, "A Genetic Programming Approach to Record Deduplicate" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3, MARCH 2012

[2] Mark R. Coppock, Steve Whitner, "Data De-Duplication for Dunmies", Wiley Publishing, Inc.

[3] Pallavi Gujar, Priyanka desai,"A Survey of Record Deduplication Techniques", International Journal of Latest Trends in Engineering and Technology(IJLTET)

[4] S. Guha, N. Koudas, A. Marathe, and D. Srivastava, "Merging the Results of Approximate Match Operations," Proc. 30th Int'l Conf.Very Large Databases (VLDB '04), pp. 636-647, 2004.

[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] http://www.genetic-programming.com/gpanimatedtutorial.html

[7] N. Koudas, S. Sarawagi, and D. Srivastava, "Record Linkage:Similarity Measures and Algorithms," Proc. ACM SIGMOD Int'lConf. Management of Data, pp. 802-803, 2006.

[8] 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.