IJERST

International Journal of Engineering Research and Science & Technology

ISSN : 2319-5991
Vol. 2, No. 2
May 2013

www.ijerst.com

Email: editorijerst@gmail.com or editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
Vol. 2, No. 2, May 2013

Research Paper

RECORD DEDUPLICATION USING GENETIC PROGRAMMING APPROACH

S Kavitha[1], K Arunadevi[1], S V Sivaranjani[2] and V Karthika[1]

Corresponding Author: **V Karthika, ✉ kvkarthikavelu@gmail.com*

In the upcoming growing of technology the use of databases are very high. As the use of databases grows higher the dirty data on the other side is the biggest disadvantage with the databases. Dirty data can contain such mistakes as spelling or punctuation, incorrect data associated with a field, incomplete or outdated data or even data that is duplicated in the database. Various data cleaning software's are used to remove the dirty data. In our paper we are proposed a concept of Genetic programming approach to record Deduplication that 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. In addition, our genetic programming approach is capable of automatically adapting these functions to a given fixed replica identification boundary. We are applying this genetic programming approach for the blood bank database management to deduplicate the records.

Keywords: Database integration, Evolutionary computing and Genetic algorithms, Database integration

INTRODUCTION

Several systems such as digital libraries and other database systems like organization databases are affected by the duplicates. We propose a genetic programming approach to find a deduplication function that is able to identify whether two entries in a repository are replicas or not. Deduplication is a task of identifying the duplicate data in a repository that refer to the same real world entity or object and systematically substitutes the reference pointers for the redundant blocks; also known as storage capacity optimization. Dirty data is defined in various categories (1) performance degradation—as additional useless data demand more processing, more time is required to answer simple user 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) increasing operational costs—because of the additional volume of useless data, investments are required on more storage media and extra

[1] Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India.

[2] Department of Information Technnology, Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India.

computational processing power to keep the response time levels acceptable. To avoid these problems, it is necessary to study the causes of “dirty” data in repositories. A major cause is the presence of duplicates, quasi replicas, or near-duplicates in these repositories, mainly those constructed by the aggregation or integration of distinct data sources. The problem of detecting and removing duplicate entries in a repository is generally known as record deduplication.

In our project we remove the dirty date in the blood bank management system. As a part of genetic programming approach the gaining concepts and the entropy calculations are used to deduplicate the records.

RELATED WORKS

Record deduplication is a growing research topic in database and many other fields as we mentioned above. The data collected from disparate sources having the redundant data. Other replicas present because of the OCR documents. This leads to the inconsistent that may affect the originality of the database and the database management systems.

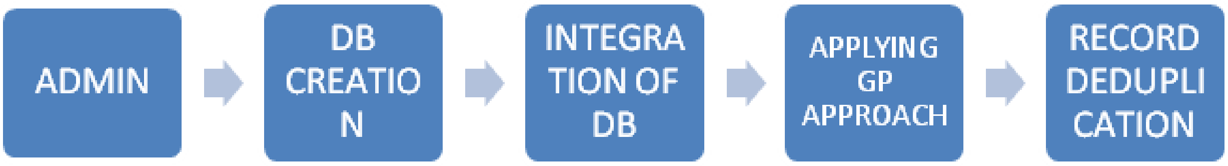
This could be overcome by the Genetic programming approach an evolutionary algorithm-based methodology inspired by biological evolution to find computer programs that perform a user-defined task. It is a specialization of Genetic Algorithms (GA) where each individual is a computer program. It is a machine learning technique used to optimize a population of computer programs according to a fitness determined by a program’s ability to perform a given computational task.

The main contribution of this paper is a GP-based approach to record deduplication that: Outperforms an existing state-of-the-art machine learning based method found in the literature; provides solutions less computationally intensive, since it suggests deduplication functions that use the available evidence more efficiently and frees the user from the burden of choosing how to combine similarity functions and repository attributes .This distinguishes our approach from all existing methods, since they require user-provided settings; frees the user from the burden of choosing the replica identification boundary value, since it is able to automatically select the deduplication functions that better fit this deduplication parameter.

PROPOSED WORK

In Figure 1 overview of our project record deduplication detection is shown. The Blood group data set in which we are going to find duplicates is taken. The entropy value is calculated for the data set as a whole, i.e., for positive as well as negative. Based on the entropy value the donor

Figure 1: Overview of Our Project of Record Deduplication



records will be displayed in a Tree structure in which the blood groups are grouped together. Entropy is the part of gain process. The entropy value is applied into the gain formula which is used to display the donor record with the highest priority.

DESCRIPTION

Administrator

The person who is responsible for setting up and maintaining the system is called as the system administrator. Administrator maintains the database in a secure manner. Responsible for installation, configuration, monitoring and administration and improving all the duplicates in the database performance and capacity. Admin are usually charged with installing, supporting, and maintaining servers or other computer systems, and planning for and responding to service outages and other problems. Administrator is also responsible for creating backup and recover policy monitor network communication.

Our project is implemented for blood bank system. Here the administrator maintains the whole database which contains the details like registration of users and their blood donation details. Also the various blood bank branches for that particular blood bank Admin can look for all those details with that he can insert, update any information into the database as well as he can delete any unwanted information from the database. He calculate the entropy and gain values in order to group the user details with priority. So that he can display the blood groups by order. From that the admin can come to know which blood group is required most. Finally, he merge/integrate all the blood bank branches to find out the duplicate entries in the database by means of gaining values. So that the duplicates will be displayed separately. After that he will send the mails to the blood banks which are having the duplicates in their database.

The admin create new branches with branch id’s, they can change the user name and password of each branch admin’s, lock/unlock their accounts, monitor the security over the truncations.

Creation of DB and Entropy Calculation

Entropy is one kind of measurement procedure in information theory, details about Entropy is in here. In here, we will see how to calculate Entropy of given set of data.

Entropy(S) = Σn = 1 – p(I)log2p(I)

p(I) refers to the category of the blood group.

S refers to the collection size we will see the example for entropy calculation from the following tables.

From Table 1 its known that there are two persons with positive type of AB blood group, one with negative type of AB and other with Positive type of B has been registered.

AB + = 2

AB - = 1

B+=1

Table 1: Master Table

Phone number	Blood Group	Blood Type
123456789	AB	+
1234567891	AB	-
1234567892	AB	+
1234567893	B	+

The entropy value is calculated for each of the blood group by means of the above formulae.

Entropy (AB+) = 1 – (2/5) log 2(2/5) = 0.960

Entropy (AB-) = 1 –(1/5) log 2(1/5) =0.590

Entropy (B +) = 1 – (1/5) log 2(1/5) = 0.789

Integration of Dataset and Detection Using Gaining Value

The gaining value is calculated for the records. Based on the gaining value the records which have the same key attribute values are grouped and they are displayed with their highest priority. Grouping records makes easier in identify the duplicate records and also this makes easy access of records. It improves the system performance in searching and retrieving the records. After finding entropy we next going to find gain value. Entropy is the part of gaining process. Information gain is G(S,A), where S is the collection of the data in the data set and A is the attribute for which information gain will be calculated over the collection S.

Table 2:Transaction Table (Excluding Dirty Data)

	Phone Number	Donation Date
B1	123456789	10/1/12
B2	1234567893	10/1/12
B1	1234567891	1/1/12
B1	1234567892	2/1/12
B2	123456789	1/7/12
B1	123456789	10/1/12
B2	1234567893	10/1/12
B1	1234567891	1/1/12
B1	1234567892	2/1/12
B2	123456789	1/7/12

$$\text{Gain}(S, A) = \text{Entropy}(S) - \Sigma(|Sv|/|S|) \times \text{Entropy}(Sv))$$

The entropy value is applied into the above formula in order to find the gain value for each blood group the gain value is calculated to the corresponding entropy value.

Table 2 is the transaction table which shows that blood donors can donate blood at different branches with their personal details.

Display the Duplicates

The administrator can merge the database to find out the duplicate entries for a whole-based on the gain value the admin can come to know the duplicate entries in the database. For example if the gain is negative value means the admin can know that the corresponding blood group is duplicated, most he will get the overall duplicates from all the database. To detect the duplicates from each branch he can split the databases. From each branch the duplicate entries will be displayed and the mails have to be sent to those branches.

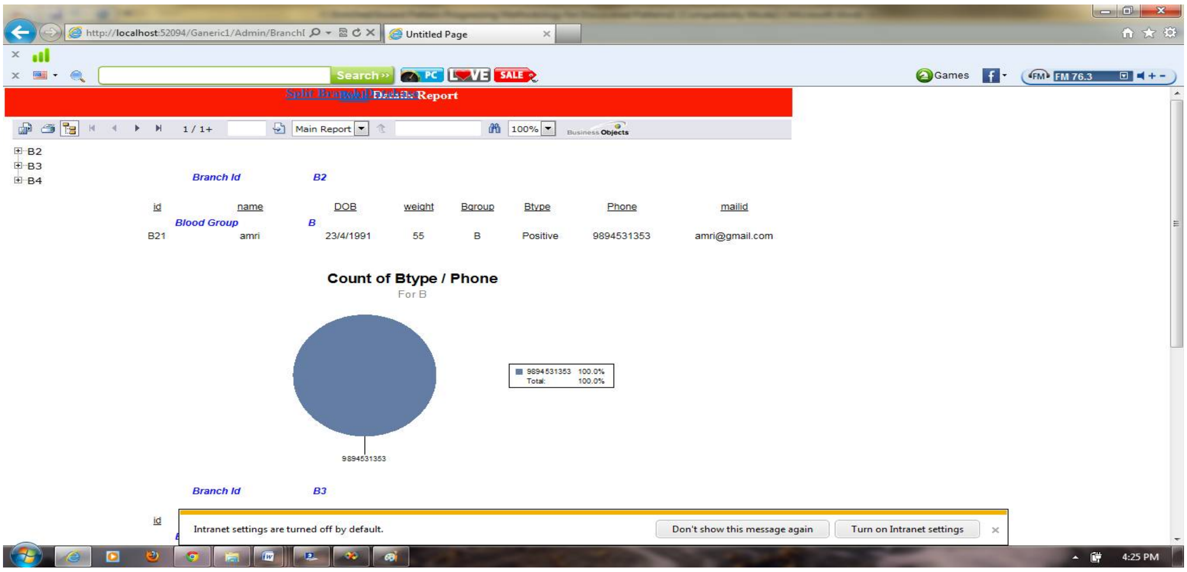
GENETIC PROGRAMMING APPROACH

The problem of record duplication is solved by some of the evolutionary techniques. Genetic programming is one of the best known evolutionary programming techniques. The main aspect that distinguishes GP from other evolutionary techniques 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 characteristics enables GP to model any other machine learning representation, another advantage of GP over other evolutionary techniques, its applicability to symbolic regression problems, since the representation structures are variable. Gp is able to discover the independent variables and their relationships with each other and with any dependent variable. Thus, GP can find the correct functional form that fits the data and discover the appropriate coefficients.

EXPERIMENTAL RESULTS

Figure 2 depict us about the duplicate records in each branch and Figure 3 is the output of genetic programming approach which clearly shows the duplicate entries from entire branch.

Figure 2: Duplicates at Each Branch



Branch Id B2

id	name	DOB	weight	Bgroup	Btype	Phone	mailid
Blood Group	B						
B21	amri	23/4/1991	55	B	Positive	9894531353	amri@gmail.com

Count of Btype / Phone

For B

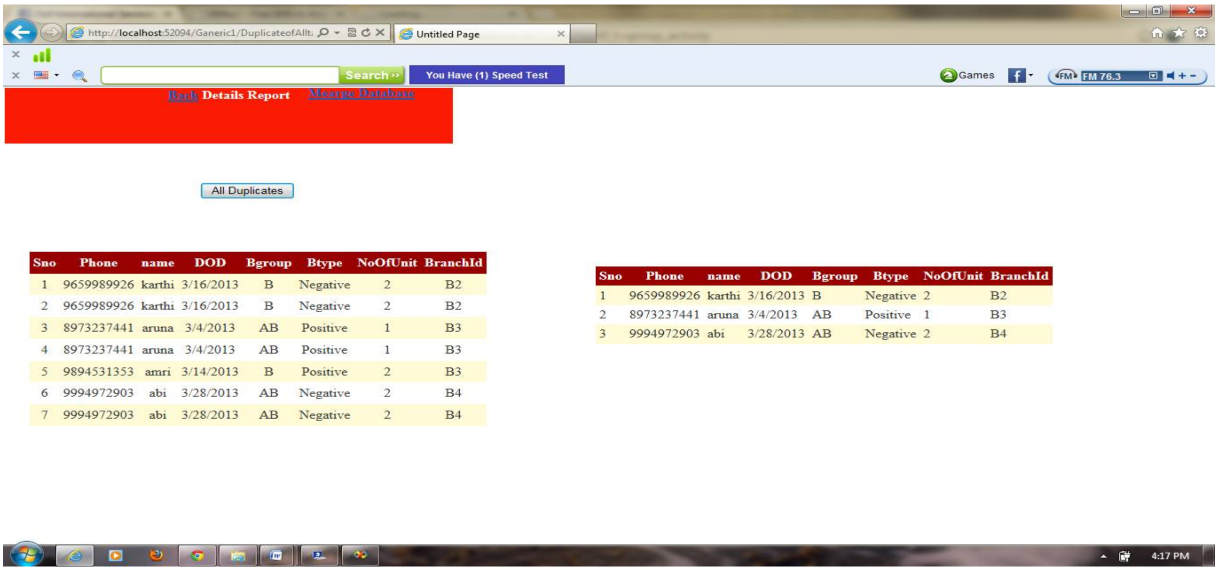
■	9894531353	100.0%
	Total:	100.0%

Branch Id B3

id

Intranet settings are turned off by default.

Figure 3: Duplicate Entries for all the Branches



Sno	Phone	name	DOD	Bgroup	Btype	NoOfUnit	BranchId
1	9659989926	karthi	3/16/2013	B	Negative	2	B2
2	9659989926	karthi	3/16/2013	B	Negative	2	B2
3	8973237441	aruna	3/4/2013	AB	Positive	1	B3
4	8973237441	aruna	3/4/2013	AB	Positive	1	B3
5	9894531353	amri	3/14/2013	B	Positive	2	B3
6	9994972903	abi	3/28/2013	AB	Negative	2	B4
7	9994972903	abi	3/28/2013	AB	Negative	2	B4

Sno	Phone	name	DOD	Bgroup	Btype	NoOfUnit	BranchId
1	9659989926	karthi	3/16/2013	B	Negative	2	B2
2	8973237441	aruna	3/4/2013	AB	Positive	1	B3
3	9994972903	abi	3/28/2013	AB	Negative	2	B4

CONCLUSION

Identifying and handling replicas is important to guarantee the quality of the information made available by the 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. Thus, for this reason, there have been significant investments from private and government organizations for developing methods for removing replicas from large data repositories. In this paper, we presented a GP-based approach to record deduplication. Our 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.

Our experiments show that our GP-based approach is able to adapt the suggested deduplication functions to different boundary values used to classify a pair of records as replica or not. Moreover, the results suggest that the use of a fixed boundary value, as close to 1 as possible, eases the evolutionary effort and also leads to better solutions.

As future work, we intend to conduct additional research in order to extend the range of use of our GPbased approach to record deduplication.

REFERENCES

1. Banzhaf W, Nordin P, Keller R E and Francone F D (1998), *Genetic Programming - An Introduction: On the Automatic Evolution of Computer Programs and Its Applications*. Morgan Kaufmann Publishers.
2. Bell R and Dravis F (2006), "Is You Data Dirty? and Does that Matter?," Accenture Whiter Paper, http://www.accenture.com.
3. Bhattacharya I and Getoor L (2004), "Iterative Record Linkage for Cleaning and Integration," *Proc. Ninth ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery*, pp. 11-18.
4. Chaudhuri S, GanjamK, Ganti V and Motwani R (2003), "Robust and Efficient Fuzzy Match for Online Data Cleaning", *Proc. ACM SIGMOD Int'l Conf. Management of Data*, pp. 313-324.
5. de Carvalho M G, Gonc¸alves M A, Laender A H F and da Silva A S (2006), "Learning to Deduplicate", Proc. Sixth ACM/IEEE CS Joint Conf. Digital Libraries, pp. 41-50.
6. Fellegi I P and Sunter A B (1969), "A Theory for Record Linkage," *J. Am. Statistical Assoc.*, Vol. 66, No. 1, pp. 1183-1210.
7. Koudas N, Sarawagi S and Srivastava D (2006), "Record Linkage: Similarity Measures and Algorithms", *Proc. ACM SIGMOD Int'l Conf. Management of Data*, pp. 802-803.
8. Koza J R (1992), *Gentic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press.
9. Verykios V S, Moustakides G V and Elfeky M G (2003), "A Bayesian Decision Model for Cost Optimal Record Matching," *The Very Large Databases J.,* Vol. 12, No. 1, pp. 28-40.
10. Wheatley M (2004), "Operation Clean Data", *CIO Asia Magazine*, http://www.cio-asia.com, August.

IJERST

International Journal of Engineering Research and Science & Technology
Hyderabad, INDIA. Ph: +91-09441351700, 09059645577
E-mail: editorijlerst@gmail.com or editor@ijerst.com
Website: www.ijerst.com

ISSN 2319-5991
02
9 772319 599001