ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

A Survey of Record Deduplication

Shital Shankar Gujar, Avinash Shrivas

Abstract- 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 out-dated 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.

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

I. 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 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.

II. RELATED WORK

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 GPbased 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.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

III. 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 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.

1. **User Authentication:** This module will facilitate validation of the user details to authenticate the admin and the employee user. This it will facilitate only valid users to access the application.

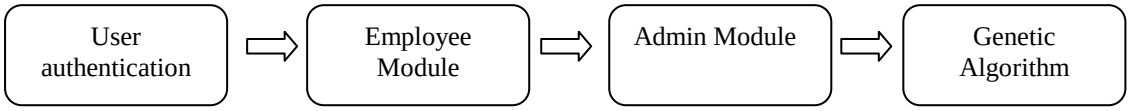

Fig 1 Overall Deduplication Process

1. **Employee Module:** This user will be responsible for the creating the dynamic data for his branch which will be used for de-duplication mechanism. It will consist of the patient details admitted into the system consisting of parameters like name, location, age, gender, blood group and blood type etc. These details will be transferred to the temporary database of the system which will be validated by the admin for duplicity.

2. **Admin Module:** This user will be responsible to validate the details entered by the employees about various patients based on the De-duplication mechanism utilized in the system. The rectified details will then be transferred to the final database of the application. This user will be given option to manage various branches and employees for the particular branch.

4. **De-duplication using Genetic Algorithm:** The system will analyze each of the records to identify duplicity of data using the concepts of Entropy and Gain Value in genetic programming to identify duplicity of data in terms of various attributes captured for each of the patient details stored within the system. Therefore, the system will display the admin with the duplicate entries within the system and give an option to send notification to the employee of the branch to verify and update the exact details into the system.

IV. 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 [8]. 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[11]. This special characteristic 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.

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[1]. 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.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

V. 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. For accomplishing this, we plan experiments with data sets from other domains. More specifically, we intend to investigate in which situations (or scenarios) our proposed GP-approach would not be the most adequate to use. Since record deduplication is a very expensive and computationally demanding task, it is important to know in which cases our approach would not be the most suitable option. In addition, we intend to improve the efficiency of the GP training phase by selecting the most representative examples for training. By doing so, we can minimize the training effort required by our GP-based approach without affecting the quality of the final solution.

REFERENCES

[1] B. Corona, M. Nakano, H. Pérez, “Adaptive Watermarking Algorithm for Binary Image Watermarks”, Lecture Notes in Computer Science, 1. Banzhaf W, Nordin P, Keller R E and Fran cone F D (1998), Genetic Programming-An Introduction Automatic Evaluation 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, Ganjam K, Ganti V and Motwani R (2003), “Robust and Efficient Fuzzy Match for Online Data Cleaning”, Proc.Ninth ACM SIGMOD Int’l Conf anagement 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.Ninth ACM SIGMOD Int’l Conf anagement of Data, pp. 802-803.

[8] Koza J R (1992), Genetic Programming: On The Programming of Computers by Means of Ntural 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.cioasia.com, August.

[11] A. Chatterjee and A. Segev, “Data Manipulation in Heterogeneous Databases,” ACM SIGMOD Record, vol. 20, no. 4, pp. 64-68, Dec. 1991.

[12] IEEE Data Eng. Bull., S. Sarawagi, ed., special issue on data cleaning, vol. 23, no. 4, Dec. 2000.

[13] J. Widom, “Research Problems in Data Warehousing,” Proc. 1995 ACM Conf. Information and Knowledge Management (CIKM ’95), pp. 25-30, 1995.

[14] A. McCallum, “Information Extraction: Distilling Structured Data from Unstructured Text,” ACM Queue, vol. 3, no. 9, pp. 48-57, 2005.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

[15] 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.

[16] H.B. Newcombe and J.M. Kennedy, "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information," Comm. ACM, vol. 5, no. 11, pp. 563-566, Nov. 1962.

[17] H.B. Newcombe, "Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories," Am. J. Human Genetics, vol. 19, no. 3, pp. 335-359, may1967

AUTHOR BIOGRAPHY

SHITAL SHANKAR GUJAR B.E COMPUTER AND M.E INFORMATION TECHNOLOGY PURSUING IN VIDYALANKAR INSTITUTE OF TECHNOLOGY, WADALA, MUMBAI.

AVINASH SHRIVAS B.E COMPUTER AND M.TECH COMPUTER ENGG, VIDYALANKAR INSTITUTE OF TECHNOLOGY, WADALA, MUMBAI.