International Journal of Research, Science, Technology & Management
ISSN Online: 2455-2240, Volume 4 Issue II, August 2016

DETECTION OF DUPLICATE RECORD USING GENETIC ALGORITHM

Sheetal Rajoriya, Manu S

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 Dirty data can contain such mistakes as spelling or punctuation, incorrect data associated witha field, incomplete or outdated data or even data that is duplicated in the database.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*

1. INTRODUCTION

Genetic algorithms are ideal for these types of problems where the search space is large and the number of feasible solutions is small. To apply a genetic algorithm to a scheduling problem we must first represent it as a genome. One way to represent a scheduling genome is to define a sequence of tasks and the start times of those tasks relative to one another. Each task and its corresponding start time represent a gene[13]. A specific sequence of tasks and start times (genes) represents one genome in our population. To make sure that our genome is a feasible solution we must take care that it obeys our precedence constraints. We generate an initial population using random start times within the precedence constraints. With genetic algorithms we then take this initial population and cross it, combining genomes along with a small amount of randomness (mutation)[8]. We let this process continue either for a pre-allotted time or until we find a solution that fits our minimum criteria.Several systems such as digital libraries another database system likes organization databases are affected by the duplicates.

2. PROCEDURE FOR PAPER SUBMISSION

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 characteristic enables GP to model any other 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

3. MATH

The term entropy usually refers to the Shannon entropy, which plays a central role in information theory as a measure of information, choice, and uncertainty contained in a system consisting of a random variable

$$Entropy\ (\ Set\ ) = I\ (Set\ ) = -\sum P(value\ i\ ) \cdot \log 2 P(value\ i\ )$$

Information Gain is the expected reduction in entropy caused by partitioning the examples according to a given attribute Visualizing Information Gain

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

4. HELPFUL HINTS

- **Figures**

IJRSTM

International Journal of Research, Science, Technology & Management
ISSN Online: 2455-2240, Volume 4 Issue II, August 2016

DeDuplication | Welcome emp1 | Home | Manage Patients | Edit Profile | Logout

Patient Details

+ Add Patient

Sr No	Branch	Name	Address	Age	Gender	Blood Group	contact	Actions
1	Vashi	Rohan Shah	Sector 18	83	Male	B-	9865745635	Edit Delete
2	Vashi	Rahul Shama	Sector 15	35	Male	O+	9876543215	Edit Delete
3	Vashi	Robin DCosta	Sector 18	85	Male	B+	9865745630	Edit Delete
4	Vashi	Rakesh Verma	new street	26	Male	A+	9892012348	Edit Delete
5	Vashi	kirti pagare	kalyan	26	Female	A+	7208833563	Edit Delete

© 2014 Company, Inc. · Privacy · Terms Back to top

DeDuplication

localhost:8080/DeDuplication/rawData.jsp

DeDuplication | Welcome admin | Home | Manage Employees | Manage Branches | Raw Data | DeDuplication | Logout

Raw Data

Sr No	Branch ID	Employee ID	First Name	Last name	Address	City	State	Country	Age	Gender	Blood Group	Height	Weight	Disease	Contact	Date
1	2	2	John	Smith	Sector 16	Navi Mumbai	Maharashtra	India	52	Female	O-	150	50 kg	Diarrhea	9876543212	2014-04-30 14:20:59.0
2	1	1	Rohan	Shah	Sector 18	Mumbai	Maharashtra	India	83	Male	B-	148	75 kg	Fever	9865745635	2014-04-30 17:34:39.0
3	1	1	Rahul	Shama	Sector 15	Navi Mumbai	Maharashtra	India	35	Male	O+	150	50 kg	Malaria	9876543215	2014-04-30 14:20:59.0
4	2	2	Rahul	Sharma	Sector 16	Navi Mumbai	Maharashtra	India	35	Male	O+	135	35 kg	Diarrhea	9876543234	2014-04-30 14:20:59.0
5	1	1	Robin	DCosta	Sector 18	Navi Mumbai	Maharashtra	India	85	Male	B+	148	72 kg	Cold	9865745630	2014-04-30 17:34:39.0
6	4	4	Hugh	Jackman	New York	New York	New York	America	48	Male	O+	175	75 kgs	abcd	9865327415	2014-05-08 12:21:00.0
7	3	3	Steve	Waugh	Australia	Australia	Australia	Australia	68	Male	B-	168	68 kg	xyz	9876543215	2014-05-09 15:51:26.0
8	1	1	Rakesh	Verma	new street	MUMBAI	Maharashtra	India	26	Male	A+	125	65 kg	Malaria	9892012348	2014-07-12 14:09:07.0
9	2	2	Rakesh	Verma	new street	MUMBAI	Maharashtra	India	26	Male	A+	125	65 kg	cough	9892012348	2014-07-12 14:11:08.0

IJRSTM

DeDuplication | Welcome admin | Home | Manage Employees | Manage Branches | Raw Data | DeDuplication | Logout

Duplicate Data

Sr No	Branch ID	Employee ID	First Name	Last name	Address	City	State	Country	Age	Gender	Blood Group	Height	Weight	Disease	Contact	Date
1	1	1	Rahul	Shama	Sector 15	Navi Mumbai	Maharashtra	India	35	Male	O+	150	50 kg	Malaria	9876543215	2014-08-31 15:16:22.0
2	2	2	Rahul	Sharma	Sector 16	Navi Mumbai	Maharashtra	India	35	Male	O+	135	35 kg	Diarrhea	9876543234	2014-08-31 15:16:22.0
3	1	1	Rakesh	Verma	new street	MUMBAI	Maharashtra	India	26	Male	A+	125	65 kg	Malaria	9892012348	2014-08-31 15:16:22.0
4	2	2	Rakesh	Verma	new street	MUMBAI	Maharashtra	India	26	Male	A+	125	65 kg	cough	9892012348	2014-08-31 15:16:22.0

© 2014 Company, Inc. · Privacy · Terms Back to top

IJRSTM
ISSN ONLINE: 2455-2240

5. 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 GP based approach to record deduplication.

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.

IJRSTM