Record Linkage Using Jaro-Winkler Technique Implemented In Database Misuse Domain

V.Vennila, S.Savitha, T.Sathya

Abstract - **The contemporary status in the area of record linkage is to bring the records from the various databases that do not have the common element and to link them. One-to-many record linkage is used for connecting a record from the first database with a set of matching records from the second database. And it is executed to classify the common user and the malicious user. The proposed system implements a splitting criteria namely jaro-winkler technique to choose the best attribute among all the available attributes at each node for the tree construction. A threshold is calculated which plays a main role in the categorization of the users. A pruning technique is used to reduce the tree by avoiding the unnecessary branches. Thus, the proposed system is enhanced in terms of time complexity and accuracy.**

Keywords: **Jaro-Winkler, Pruning, Splitting criteria, Record linkage.**

I. INTRODUCTION

In general, the major task of the record linkage is to recognize the various records that refers to the same record within a single database or when combining together two or more databases. Currently, there is a obligation to find out the methods that will link the database that does not share the common element, that is, a foreign key. It is also an important process while preprocessing the datasets. The record linkage is split into two types and they are one-to-one record linkage and one-to-many record linkage. In one-to-one record linkage, two databases are compared to spot all the elements of the first dataset that match to a particular element of a second dataset. In one-to-many record linkage, two datasets are evaluated to spot all the best pairs or matches in the second with the first dataset.

The record linkage practice is easiest when there is unique identification numbers such as aadhar card number, pan card number and voter identification number are available. The process is more challenging when the elements such as given first name, last name, birth date and resident address are available. By combining two or more elements of these kinds will identify an individual. Historically, record linkage was allocated to clerks who would seek and review the lists to bring together the approximately matching pair of records for comparison.

Record linkage is a tricky task because of errors, variations and missing data in the information used to link records, differences in data captured and maintained by different databases, e.g. age versus Date Of Birth, regularly and routinely change over time, name changes due to marriage, often no unique record elements are available and training data is not available in many linkage applications.

A tree is constructed where the leaves do not contain the labels instead a group of matching values is attached to each leaf. The tree as a whole describes a hierarchy. Each leaf of the tree is characterized by a valid expression representing the instances belonging to it. The main advantage of using this kind of trees is that they provide a description for each of the leaf.

In this paper, a new record linkage method is proposed that is aimed at performing one-to-many linkage that can match records of various types. For example, Let T_X and T_Y are the two tables of different types. The internal nodes of the tree consist of attributes that are referring to both of the tables being matched, that is Table X and Table Y. The leaves of the tree will determine whether a pair of records described by the path in the tree ending with the current leaf is a match or a mismatch.

The proposed method is implemented using the database misuse domain used to identify the common and malicious users. In this domain, the goal is to identify anomalous access to database records that may indicate a possible data leakage, loss of data integrity or data mishandling.

II. RELATED WORK

Record linkage is the process of matching entities from two different data sources that may or may not contain a common element. Shabtai et al. [1] used one-to-many linkage in different domains like fraud detection, recommender systems and data leakage prevention by developing a One-Class Clustering Tree. Four splitting criteria and pruning methods are implemented to construct the tree and to avoid the needless branches of the tree correspondingly. The shortcoming of this approach is that it is difficult to reduce the linkage computation time and it is a one-class approach.

Christen and Goiser [2] used a C4.5 decision tree to determine which records must be matched to one another. In their work, various string comparison methods are used and compared by constructing different decision trees. However, their method performs the matching of attributes that are only predefined. Moreover only one or two attributes are usually used.

Henry et al. [3] used one-to-many linkage for genealogical research. Record linkage was performed using five attributes: name of the person, birth date, place, gender and the relationships between the persons. Using these five attributes a decision tree was induced. The drawback of this approach is that it performs matching using the specific attributes and therefore it is very hard to generalize.

III. PROPOSED METHOD

ISSN: 2278 – 1323

The process flow of the proposed system is as follows: First the user dataset and the server log dataset for linkage process is taken into account. Then preprocess both the data by removing the null values or missing values and those attributes that have error data. After preprocessing, create a training table based on both the server log and user dataset. For training table creation, consider the elements such as requesting location, requesting day, requesting time from the first database and in the other table consider the elements such as user location and the business type.

Then measure the similarity score based on the elements of requesting location, requesting day, requesting time with user location and the business type by using the jaro-winkler technique. Then apply the post pruning method to remove the unnecessary branches that are not used.

A total probability value (sum all possible values of an attribute) is calculated which is compared with the probability calculated for each value of an attribute and that is used in the classification of the users as the common user and the malicious users. At last, the performance of the existing and the proposed system is evaluated in terms of execution time and the accuracy of the result.

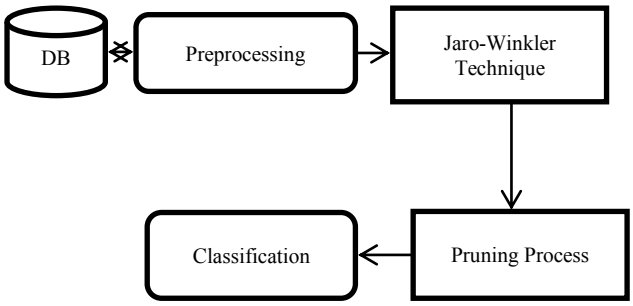

Fig. 1 Architecture diagram.

IV. SPLITTING CRITERIA

The splitting criteria is used to select the best attribute at each node using the following technique:

A. Jaro-Winkler Technique

It is used to measure the similarity between two strings. This algorithm is mainly developed for record linkage process and it was designed to calculate the similarity for short strings. It calculates a normalized score on the similarity between two strings. The computation is based on the number of matching characters held within the string and the number of transpositions.

To compare character sequences cs1 and cs2. The Jaro-Winkler distance is calculated using the below steps.

Step 1: The match range for the identifying the common characters is calculated as:

matchRange=max(cs1.length(),cs2.length())/2-1 (1)

Step 2: The Jaro distance of two given strings cs1 and cs2 is

$$\begin{cases} 0 & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|cs1|} + \frac{m}{|cs2|} + \frac{m-t}{m}\right) & otherwise \end{cases} \quad (2)$$

In the above equation, m is the number of matching characters; t is the number of transpositions. The matching characters are found within and after the match range, that is, the values are divided into two sections for finding the matching characters. The transpositions are the number of characters that are mismatched within the matching characters.

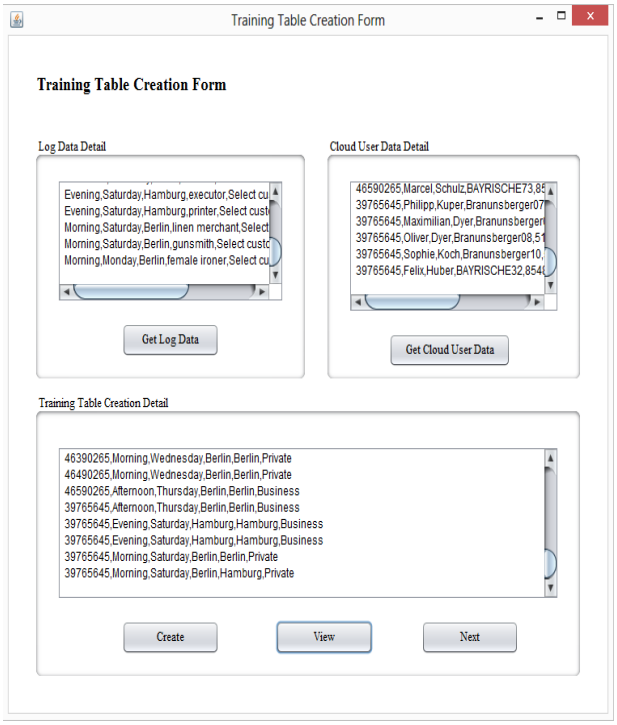

Fig. 2 Training Table Creation

V. PRUNING

Pruning is considered to be a significant action during the tree construction process. The necessity of pruning is to build a tree with accuracy and also to avoid over fitting. Pruning can be classified into two types: pre-pruning and post-pruning.

In pre-pruning, the branches are pruned during the induction process if there are no possible splits found. In post-pruning, the tree is built completely followed by a bottom-up approach to determine which branches are not beneficial. In the proposed system, post-pruning approach is implemented.

After pruning, classification is done based on the following rules: the user's requesting location and the user's city should be the same and also the working hours of the user should not be after hours; if all these conditions are satisfied then the user is labeled as common user who does not do any malicious activities. Otherwise, all the other records which do not satisfy the above condition are termed as the malicious user.

ISSN: 2278 – 1323

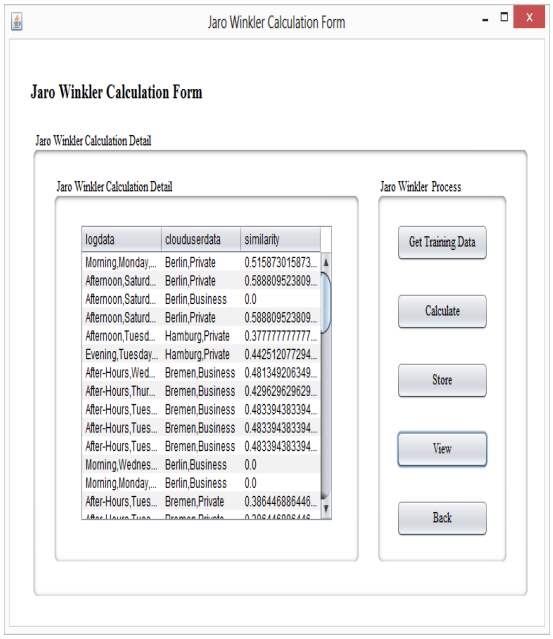

Fig. 3 Jaro-Winkler calculation

VI. CONCLUSION

A new method is proposed to link the data that do not have the common elements. A tree is constructed to accomplish the one-to-many record linkage and the database users are classified as normal and abnormal user. Improved efficiency is attained by means of the time complexity and accuracy of the record linkage process.

REFERENCES

[1] A.Shabtai, M.Dror, L.Rokach, Y. Elovici, "OCCT: A One-Class Clustering Tree for Implementing One-to- Many Data Linkage," IEEE Trans. on Knowledge and Data Engineering, TKDE-2011-09-0577, March 2014.

[2] P.Christen and K.Goiser, "Towards Automated Data Linkage and Deduplication," Australian National University, Technical Report, 2005.

[3] G.Henry, S.Ivie, H.Gatrell and C.Giraud-Carrier, "A Metric Based Machine Learning Approach to Genea- Logical Record Linkage," in Proc. of the 7th Annual Workshop on Technology for Family History and Genealogical Research, 2007.

[4] A.J.Storkey, C.K.I.Williams, E.Taylorand R.G.Mann, "An Expectation Maximisation Algorithm for One-to- Many Record Linkage," University of Edinburgh Informatics Research Report, 2005.

[5] M.Yakout, A.K.Elmagarmid, H.Elmeleegy, M.Quzzani and A.Qi, "Behavior Based Record Linkage," in Proc. of the VLDB Endowment, vol. 3, no 1-2, pp. 439-448, 2010.

[6] S.Guha, R.Rastogi and K.Shim, "Rock: A Robust Clustering Algorithm for Categorical Attributes Information systems, vol. 25, no.5, pp. 345-366, July 2000.

[7] D.D.Dorfmann and E.Alf, "Maximum-Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals-RatingMethod Data," Journal of Math Psychology, vol. 6, no. 3, pp. 487-496, 1969.

[8] P.Langley, Elements of Machine Learning, San Franc- Isco, Morgan Kaufmann, 1996.

[9] A.Gershman et al., "A Decision Tree Based Recommender System," in Proc. the 10th Int. Conf. on Innovative Internet Community Services, pp. 170-179, 2010.

[10] J.R.Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81-106, March 1986.

AUTHORS

VENNILA.V working as an Assistant Professor in the department of Computer Science and Engineering in K.S.R. College of Engineering, Namakkal, Tamilnadu, India.

SAVITHA.S working as an Assistant Professor in the department of Computer Science and Engineering in K.S.R. College of Engineering, Namakkal, Tamilnadu, India.

SATHYA.T received the B.E degree and currently pursuing M.E in Computer Science and Engineering in K.S.R. College of Engineering, Namakkal, Tamilnadu, India.

ISSN: 2278 – 1323