ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Implementation of Many-to-Many Data Linkage using OCCT for Matching and Non-Matching Pairs

S.Mohanapriya[1], J.Mannar Mannan[2]

PG Scholar, Dept of Information Technology, Regional Centre of Anna University, Coimbatore, India.[1]

Teaching Assistant, Dept of Information Technology, Regional Centre of Anna University, Coimbatore, India.[2]

ABSTRACT: Data linkage is a process performed among entities of the same type or different type. It is necessary to develop the data linkage techniques for different types as well. In this paper, we propose a many-to-many data linkage and it is used to perform link between matching entities of different types. The proposed method is based on One-Class Clustering Tree (OCCT) for implementing many-to-many data linkage. The OCCT is built in such a way that it is easy to understand and can be transformed into association rules. The inner node consists of features from the first data set. The leaves of the tree represent features from the second data set that is matching with the first data set entities. The proposed method uses maximum-likelihood estimation for pre-pruning process which is used to create One-Class Clustering Tree effectively. Threshold value is used for decision making either the record pair is match or non-match.

KEYWORDS: clustering; data linkage; decision tree; prepruning.

I. INTRODUCTION

Data linkage (record linkage) is combining two or more records of independent source information. It is a technique which is used to connect the information from several data sources. It allows different types of information to be more readily available and so reduces the length of time looking for data. The goal of data linkage that does not share common identifier when joining two data sets. Data de-duplication is a data compression technique for eliminating redundant data. It is a pre-processing for data mining tasks identifying individuals across different data sets.

Data linkage techniques consist of two types: Deterministic linkage and Probabilistic linkage. Deterministic linkage is a simplest linkage and it is also known as rule based linkage. It is classified into two linkages: Exact linkage and rule-based linkage. Exact linkage is used when a unique identifier of high quality is available. Rule-based linkage is complex to build and maintain. Probabilistic linkage uses available attributes for linkage (eg: personal information) and also known as fuzzy matching. Data linkages can be mainly divided into three types: one-to-one, one-to-many and many-to-many data linkage. In one-to-one data linkage, two data sets of data are compared and goal is to identify all the best pairs between the data sets. In one-to-many data linkage, two data sets are compared and goal is to identify all individuals of first data set that matches to a particular element of a second data set.

In this paper, we proposed a data linkage method which performs many-to-many data linkage. In many-to-many data linkage, it contains two data tables or entities have multiple rows that are connected to one or more rows in the other table. The proposed method is implemented using One-Class Clustering Tree (OCCT) [1]. A clustering tree is a tree which contains a cluster instead of single classification. In clustering tree, each cluster is generalized by a set of rules which is stored in the appropriate leaf. OCCT is preferable because it can be easily translated to linkage rules. It is used to evaluate in different domains such as: data leakage, recommender systems and fraud detection systems.

The contribution of the proposed work is it allows performing many-to-many data linkage for linking between entities of different types. In existing methods, only linkage between entities of same type is performed. Another advantage of proposed work is performing many-to-many data linkage using one-class approach and the method we

10.15680/ijircce.2015.0304122

used for this approach is one-class clustering tree (OCCT). This is an important advantage because to obtain meaningful non-matching examples in some domains is difficult.

In this paper, the rest of topics organized as follows: Section 2 we review related works based on data linkage and decision trees. Section 3 deals about many-to-many data linkage using OCCT and its methodology and finally Section 4 concludes the paper.

II. RELATED WORK

Data linkage is a process of matching entities from two different sources that do not share a common identifier. It is performed among entities of the same type or different type. It is divided into one-to-one, one-to-many and many-to-many data linkage. Fig. 1. describes about the record linkage process. In one-to-one data linkage, the process is to associate one record in table T_A with a single matching record in table T_B. In one-to-many data linkage, the process is to associate one record in T_A with one or more matching records in T_B.

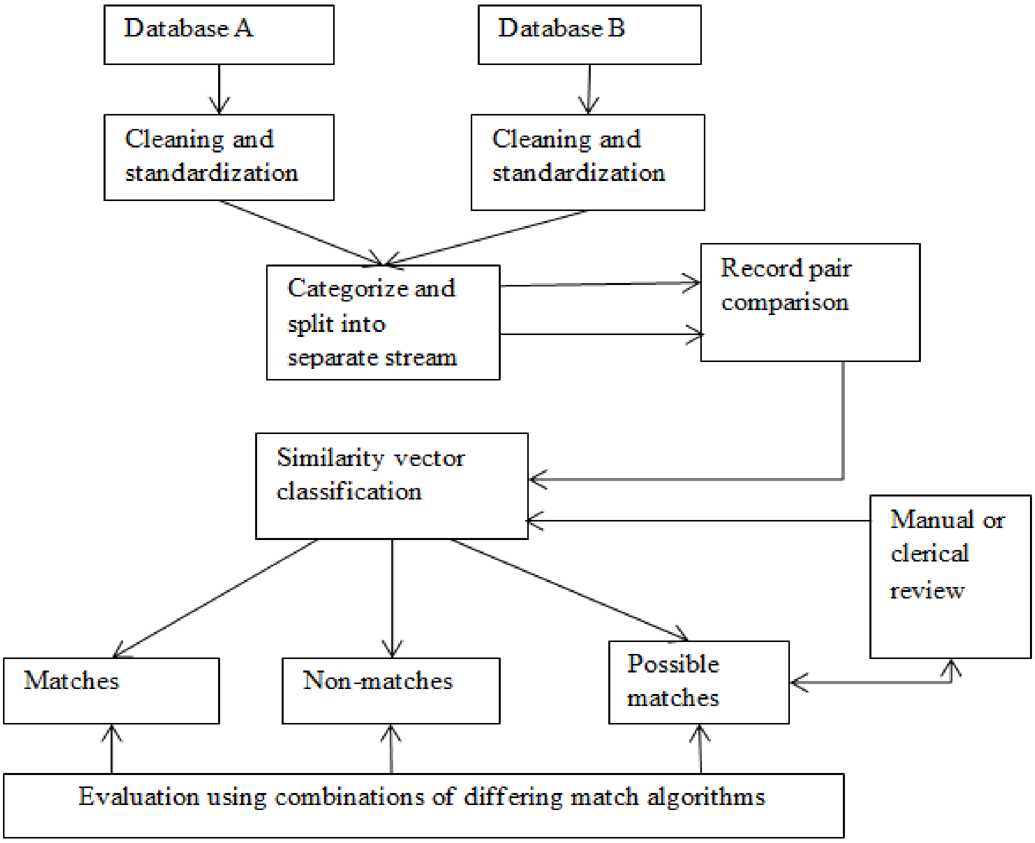
Database A
Database B
Cleaning and standardization
Cleaning and standardization
Categorize and split into separate stream
Record pair comparison
Similarity vector classification
Manual or clerical review
Matches
Non-matches
Possible matches
Evaluation using combinations of differing match algorithms

Fig.1: Record Linkage process.

In many-to-many data linkage , the process is to associate one-to-one data linkage was implemented using algorithms like SVM classifier, calculating Expectation Maximization or Maximum-Likelihood Expectation (MLE) [2] which is used to calculate the probability being record pair match. Record linkage or data linkage approaches that uses entity behaviour and is the process of identifying record that refer to the same real world entity. These methods refer to two different datasets that link between only the same entity, therefore , this is less relevant to data linkage that link between different entities.

Only a few previous works have dealt about one-to-many record linkage. Data linkage and deduplication [3] can used to improve quality and integrity, re-use of existing data sources and reduce costs. By using data linkage, the true matches or true non-matches can be classified. To improve the quality, precision and recall quality measures is used.

Ivie, Henry and Gatrell[4] used genealogical record linkage which handles one-to-many relationships by using four basic data types: name, gender, date, and location. GRL is used for determining whether two pedigrees or individuals refer to the same individual or not. It uses only specific attributes for performing the matches and it is very hard to generalize. Blockeel, Raedt and Ramon[5] constructed TIC (Top down Induction of Clustering trees) methodology which is based on clustering. Decision trees are based on classification; the leaves of the tree contain the classes and the branches represent the conditions for classification. A clustering tree is a decision tree in which the leaves of the tree contain clusters. The clustering tree can be induced by using instance based learning and decision tree induction. TIC approach is implemented using TIC system.

Data linkage is closely related to entity resolution [6]. In data linkage, the goal is to link between related entries in one or more data sources. In entity resolution, the goal is to identify non-identical records that represent the same real world entity and to merge them into a single record (deduplication).Record linkages is a process of identifying matching records that refers to same entity from several data sources and deduplication process is applied on single database. Removing duplicates from single database is complex step in the data cleaning process. Six indexing techniques are used for data linkage and deduplication process [7].

Torra and Domingo[8] analyzed record linkages techniques such as probabilistic and distance-based record linkages which are compared against numerical and categorical data. Distance-based record linkage is more appropriate for numerical data and probabilistic record linkages are more appropriate for categorical data. Guha, Rastogi and Shim[9] developed the clustering algorithm for both Boolean and categorical attributes. ROCK clustering algorithm is proposed which is based on linkages not on distances. A. Gershman et al. [10] constructed the decision tree which produces lists of recommended items at its leaf nodes, instead of single items and this leads to reduced amount of search. Splitting method is used for constructing the decision tree and the splitting is based on a new criterion - the least probable intersection size.

III. MANY-TO-MANY DATA LINKAGE USING OCCT

The two tables or data sets are the inputs and tables are pre-processed. The data sets are joined and data deduplication task is performed. Many-to-many data linkage is carried out by using One-Class Clustering Tree (OCCT). OCCT construction consists of finding the best split attribute from first data set and prepruning process is carried out for avoiding repetition and replication. Splitting criteria uses maximum likelihood estimation which chooses the values of the parameters that will maximize the probability of the particular sample. Representing the leaves is done by using second data set and the probability value is calculated for each sample. Threshold value is defined for determining either the record pair match or non-match.

Fig.2. describes about the architecture of the many-to-many data linkage using OCCT. If the probability value is greater than the threshold value then it is said to be record pair match and the probability value is lesser than the threshold value then it is said to be record pair non-match.

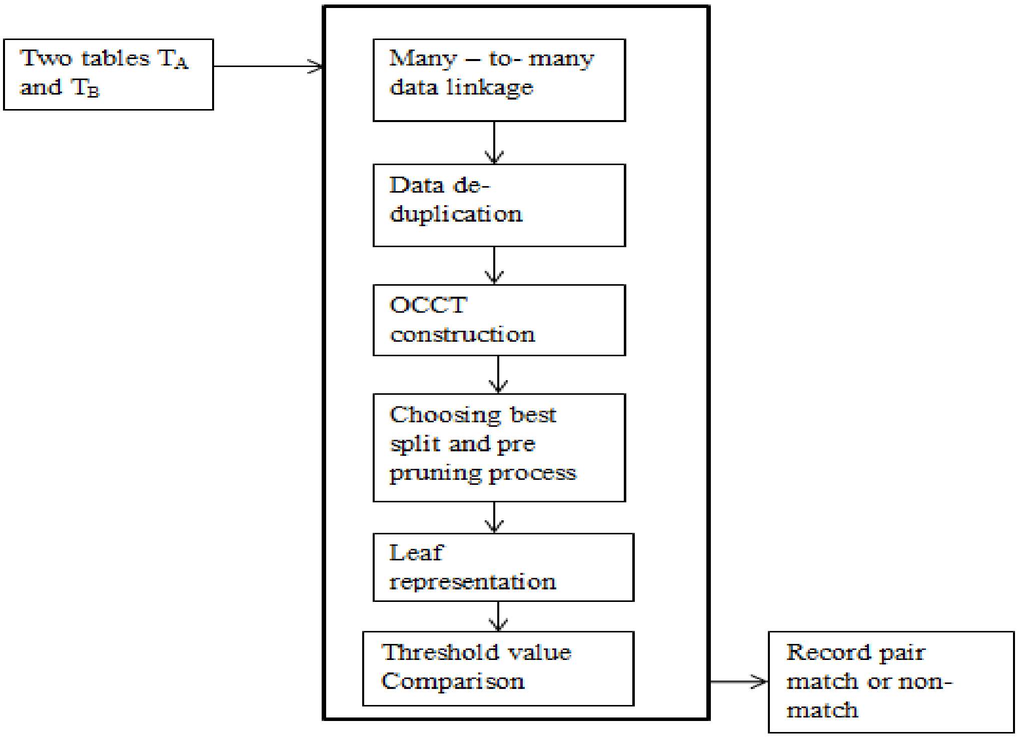
Two tables T_A and T_B
Many – to- many data linkage
Data de-duplication
OCCT construction
Choosing best split and pre pruning process
Leaf representation
Threshold value Comparison
Record pair match or non-match

Fig. 2: Architecture of many-to-many data linkage using OCCT.

The methodologies used in proposed method are

- Data set collection and data deduplication
- OCCT construction: choosing the best split attribute
- Prepruning process
- Leaf representation
- Applying OCCT for data linkage

A. *Data set collection and data deduplication:*

Data set is collected and pre-processing is carried out. Data deduplication is performed for deleting the duplicates from the two tables. It is used to increase the linkage process effectively and time complexity of linkage process is reduced. Deduplication is for improving the OCCT construction. The table A and table B data are updated for new data linkage process. Updating data will result in dynamic many-to-many data linkage and OCCT construction.

B. *OCCT construction: Choosing the best split attribute:*

The decision tree induction process includes deriving the structure of the tree. To build the tree, we must decide which attribute should be selected at each level of the tree. The inner nodes of the OCCT consist of attributes from table T_A only. For selecting the attribute maximum likelihood estimation method is used. The splitting criteria are used to rank the attributes based on how good they are in clustering the matching examples.

The splitting criteria used in proposed method are maximum- likelihood estimation (MLE). It is used to choose the attribute that is most appropriate to serve as the next splitting attribute. Once the probabilistic model has been induced, the probability of each record given these models is calculated. By using the spitting criteria, our goal is to choose the split that achieves the maximal likelihood, that is, we choose the attribute with the highest likelihood score as the next splitting attribute in the tree.

C. *Prepruning process:*

Prepruning process is implemented in tree induction process which is used to improve the accuracy of the model and avoids over fitting. There are two approaches of pruning a tree: prepruning and postpruning. Pre-pruning that stops growing the tree before it perfectly classifies the training set. Post-pruning that allows the tree to perfectly classify the training set, and then post prune the tree. In our proposed method, we follow the prepruning approach which is used to reduce the time complexity of the algorithm. Maximum likelihood estimation is computed by using equation (1)

$$L\left(r_{(b)}\right) = \sum_{i=1}^{|B|} \log \left(p(b_i = v_i | M_i; b_j = v_j, j = 1. \ldots, n, j \neq i\right) \qquad (1)$$

It is carried out once the next attribute for the split is chosen. For this process, maximum likelihood estimation is used which computes MLE score for each of the possible splits. If none of the candidate attributes achieve an MLE score which is greater than the current node’s MLE score, then the branch of the tree is pruned and the current node becomes a leaf.

D. *Leaf Representation:*

Once the model of the tree is built which is based on the attributes from table T_A and each leaf contains a data set containing the matching records from table T_B. Probabilistic models are induced for each of the leaves in the tree. Each model M_i in the tree is used for deriving the probability of a value of attribute $b_i \epsilon B$ from table $T_{B,}$ given the values of all other attributes in the table. There are two main motivations for performing the leaf representation in this model. The compact representation of the OCCT model is achieved by using set of probabilistic models. Second motivation is representing the matching records as a set of probabilistic models, this model achieve better generalization and avoids over fitting.

A feature selection process is executed on the leaf data set to choose the attributes that will be represented. The goal of the feature selection process is identifying the attributes that best represent the records that appear in a leaf. A different set of attributes might be chosen for representing each of the leaves.

E. *Applying OCCT for Data Linkage:*

During the data linkage or testing phase, each possible pair of the test records is tested against the linkage process to determine if the record pair is a match or non-match. This testing process produces a score which represents the probability of the record pair being a true match and the score is calculated using maximum likelihood estimation. The level of linkage is provided as a number between 0 and 1.

To reach a final binary decision (i.e., match or non-match) a threshold has to be defined. Threshold value is a predefined value which is used to determine the whether the record pair is either match or non-match. If the record pair’s score is greater than threshold value, then it is classified as a match otherwise it is classified as non-match. In proposed method, the decision making can be done by setting the threshold value as 0.5 and it can be used for effective linkage process.

IV. Simulation And Performance Evaluation

The performance evaluation is based on time of both one-to-many data linkage and many-to-many data linkage. Fig 3 describes about performance evaluation of two data linkages. The x-axis describes the type of linkage and y-axis describes about the time taken to retrieve the data. The accuracy of retrieval of data is more in many-to-many data linkage when compared with one-to-many data linkage.

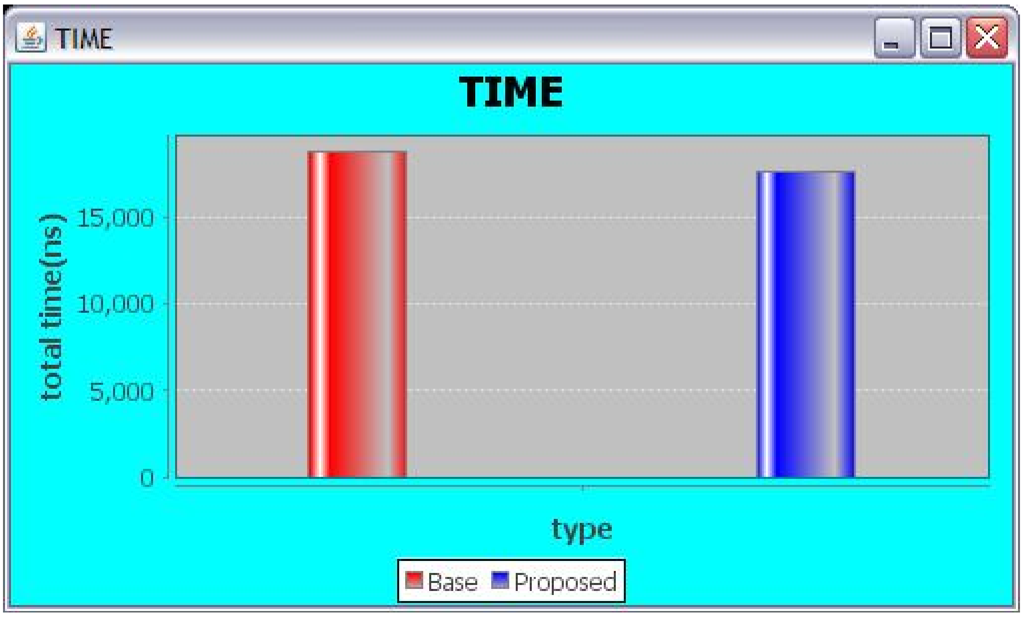

Fig. 3:Performance evaluation

V. Conclusion and Future Work

Data linkage is a process of linking between same entities or different entities. In this proposed system, we have constructed a one class clustering tree approach which performs many-to-many data linkage. Many-to-many data linkage is used to link records between different entities. Classification error is minimized and true matching pairs are identified by using this data linkage. This method is based on a one-class decision tree model which sums up the knowledge of which records should be linked to each other.

To summarize, this method allows performing many-to-many linkage while the traditional methods followed one-to-one data linkage and one-to-many data linkage. Another advantage of using OCCT model, it can be easily translated to linkage rules. Threshold value is defined for decision making whether the record pairs match or non-match.

In future work, OCCT model can be used for continuous attributes in data linkage process and the evaluation on training sets that contain non-matching examples.

References

1. Maayan Dror, Asaf Shabtai, Lior Rokach, and Yuval Elovici, “OCCT: A One-Class Clustering Tree for Implementing One-to-Many Data Linkage”, IEEE Trans. Knowledge andData Eng., VOL. 26, NO. 3, doi: 10.1109/TKDE.2013.23, 2014.
2. D.J. Rohde, M.R. Gallagher, M.J. Drinkwater, and K.A. Pimbblet,“Matching of Catalogues by Probabilistic Pattern Classification,” Monthly Notices of the Royal Astronomical Soc., vol. 369, no. 1, pp. 2-14, May 2006.
3. P. Christen and K. Goiser, “Quality and Complexity Measures for Data Linkage and Deduplication,” Quality Measures in Data Mining, vol. 43, pp. 127-151, 2007.
4. S. Ivie, G. Henry, H. Gatrell, and C. Giraud-Carrier,“A Metric-Based Machine Learning Approach to Genealogical Record Linkage,” Proc. Seventh Ann. Workshop Technology for Family History and Genealogical Research, 2007.
5. H. Blockeel, L.D. Raedt, and J. Ramon, “Top-Down Induction of Clustering Trees,” ArXiv Computer Science e-prints, pp. 55-63, 1998.
6. O. Benjelloun, H. Garcia, D. Menestrina, Q. Su, S. Whang, and J. Widom, “Swoosh: A Generic Approach to Entity Resolution,” The VLDB J., vol. 18, no. 1, pp. 255-276, 2009.
7. P. Christen, “A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication,” IEEE Trans. Knowledge and Data Eng.,vol. 24, no. 9, pp. 1537-1555, doi:10.1109/TKDE. 2011.127 ,Sept. 2012.
8. V. Torra and J. Domingo-Ferrer, “Record Linkage Methods for Multidatabase Data Mining,” Studies in Fuzziness and Soft Computing, vol. 123, pp. 101-132, 2003.

9. 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.

10. A. Gershman et al., “A Decision Tree Based Recommender System,” Proc. 10th Int’l Conf. Innovative Internet Community Services, pp. 170-179, 2010.

BIOGRAPHY

S.Mohanapriya has obtained B.Tech Information Technology in Sri Ramakrishna Engineering College, Coimbatore and doing M.Tech Information Technology in Regional Centre of Anna University ,Coimbatore.

J. Mannar Mannan has completed M.Tech Information Technology and working as Teaching Assistant in Anna University Regional Centre, Coimbatore. His research works towards Ontology and Information Retrieval.