INFO- 2153
Primljeno / Received:2015-09-13

UDK: 004.451.54:004.62:001.103
Authors Review / Pregledni rad

IMPLEMENTING MULTICAST DATA LINKAGE WITH ONE CLASS CLUSTERING TREE (OCCT)

IMPLEMENTACIJA POVEZIVANJA PODATAKA SA OCCT-om

S.Singaravelan[1], D.Murugan[2], R.Mayakrishnan[1]
Department of Computer Science and Engineering, P.S.R Engineering College, Sivakasi, India [1]; Manonmaniam Sundaranar University,Tirunelveli, India [2]

Abstract
Record linkage is traditionally performed among the entities of same type. It can be done based on entities that may or may not share a common identifier. In this paper we propose a new linkage method that performs linkage between matching entities of different data types as well. The proposed technique is based on one-class clustering tree that characterizes the entities which are to be linked. The tree is built in such a way that it is easy to understand and can be transformed into association rules. The data is split using four splitting criteria. The proposed system results better in performance of precision and recall.

Sažetak
Snimanje veza tradicionalno se izvodi između entiteta istog tipa. To se može učiniti na temelju subjekata koji mogu ili ne moraju dijeliti zajedničko identifikator. U ovom radu predlaže se nova metoda za povezivanje podudarnih subjekata ili različitih vrsta podataka. Predložena metoda temelji se na jednoj klasi klastera stabla koja karakterizira entitete koji su povezani. Stablo je izgrađen na takav način da ga je lako razumjeti i može biti pretvoren u pravila asocijacije. Podaci su podijeljeni pomoću četiri kriterija dijeljenja. Predloženi sustav rezultira boljim rezultatima u preciznosti i opozivu.

1. INTRODUCTION

Record linkage is a process of identifying different data items that refer to the same entity among different data sources. The main goal of record linkage is to join datasets that do not share a foreign key or a common identifier. Record linkage is usually performed to reduce the large data into smaller data. It also helps in removing duplicate records in the datasets. This technique is known as data deduplication. The record linkage can be divided into two types: deterministic record linkage and probabilistic record linkage. Deterministic record linkage is the simplest record linkage and it is also known as rules-based record linkage. Probabilistic record linkage is also known as fuzzy matching.

Record linkage can also be divided into: one-to-one and one-to-many record linkage. In one-to-one record linkage, an entity from one dataset has a single matching entity in another dataset. In one-to-many record linkage, an entity from first dataset has a group of matching entities from another dataset. Most of the previous works focuses on one-to-one record linkage. In this paper, a new record linkage method which performs one-to-many linkage is proposed. This method links the entities using a One-Class Clustering Tree (OCCT) /**1**/ [1]. A clustering tree is a tree in which each of the leaves contains a cluster whereas a normal tree consists of a single classification. Each cluster in the clustering tree is generalized by a set of rules. The OCCT can used in different domains like fraud detection, recommender systems and data leakage prevention. In fraud detection domain, the main aim is to find the fraudulent users. In recommender systems domain, the proposed system can be used for matching new users with their product expectations. In data leakage prevention domain, the main aim is to detect the abnormal access to the database records that indicates data leakage or data misuse.

The contribution of the proposed work is it allows performing one-to-many linkage between entities of same or different types. Another main advantage of the proposed system is using a one-class approach. Fig 1 describes the general outline of the record linkage process.

The rest of the paper is organized as follows. In Section 2 we review related works on the record linkage and decision trees. Section 3 deals with the proposed linkage model induction. Section 4 deals with the linkage using OCCT and finally Section 5 concludes the paper.

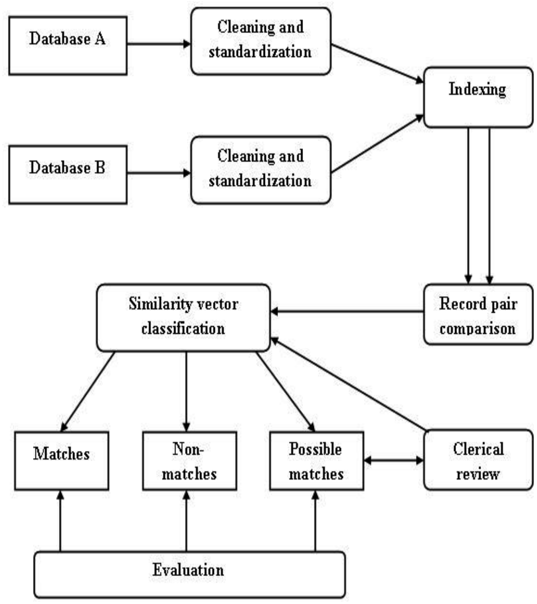

Fig 1: Outline of general record linkage process

2. RELATED WORK

Record linkage is a process of matching entities from two different data sources that may or may not share a common identifier (i.e., foreign key). One-to-one record linkage was implemented using algorithms like SVM classifier, Maximum Likelihood Expectation and performing behaviour analysis /**2**/. These methods assume that entities in the datasets are linked and try to match records that refer to the same entity. Only a few previous works have dealt about one-to- many record linkages. Storkey et al. /**3**/ used the Expectation Maximization algorithm for two purposes. They are, calculating the probability of a given record pair that is matched and to learn the characteristics of the matched records. A Gaussian mixture model was used to model the conditional magnitude distribution. The drawback in this system is no evaluation was conducted on this work. Ivie et al. /**4**/ used one-to-many linkage for genealogical research. In that work, data linkage was performed using five attributes: a person's name, gender, date of birth, location 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 specific attributes and therefore it is very hard to generalize.

Christen and Goiser /**5**/ used a C4.5 decision tree to determine which records must be matched to one another. In their work, different string comparisons methods are built and compared using different decision trees. However, their method performs the matching of attributes that are only predefined. Moreover only one or two attributes are usually used. In this paper, we propose a new record linkage method that performs one-to-many linkage that match entities of different data types along with the time calculation for the linkage process. The inner nodes of the tree consist of attributes that are in both of the tables being matched (TA and TB). The leaves of the tree will determine whether a pair of records described in the end of the tree with the current leaf as a match or non-match. Decision trees are used for regression tasks and for classification. However, the training set used for the induction of tree must not be unlabeled. Yet, acquiring a labelled dataset is a costly work. Therefore, we thought that using examples of one class in a decision model is highly preferable than using training set with labelled dataset.

When compared with traditional decision trees, clustering trees are different based on their structure /**6**/. In traditional decision

trees, each node represents a single classification. Whereas, in clustering trees, each node represents a cluster or a concept. The tree on the whole can be considered as a hierarchy. Then, each leaf of the tree is characterized by a logical expression, which represents the instances that belongs to it.
The OCCT is a decision model which resembles to a clustering tree. It is a one-class model that learns and represents only positive examples. This method differs from other clustering trees by linking two different data types.

3. LINKAGE MODEL INDUCTION

In the proposed method, linkage model induction is the first step. The linkage model gets the knowledge about records that are expected to match each other. The process includes deriving the structure of the tree. The tree building requires the decision of which attributes must be selected at each level of the tree. The inner nodes of the tree consist of attributes from table TA. The selection of attributes is actually done by using any one of the splitting criteria. The splitting criteria ranks the attributes based on their clustering of matching examples. A pre-pruning approach is implemented in this proposed method. When using this approach, the algorithm stops expanding a branch whenever the sub-branch does not improve the accuracy of the given model. The inducer is actually trained with matching examples only. The OCCT can be derived using any one of the splitting criteria. The splitting criterion is used to determine which attribute must be used in each step of constructing the tree. Our main goal is to achieve a tree that contains less number of nodes, as smaller trees easily generalize the data by avoiding over fitting. It will also be simpler for the human eyes to understand the tree structure /**7**/. The two types of splitting criteria used in this system are: Maximum Likelihood Estimation (MLE) and Least Probable Intersections (LPI).

3.1 Maximum Likelihood Estimation (MLE)

This particular splitting criterion uses the Maximum Likelihood Estimation (MLE) /**8**/ for choosing the attribute that is most appropriate to serve as the next splitting attribute for the forthcoming attributes that are yet to be split. We aim to choose the split that achieves the maximum likelihood and hence we choose the attribute that has the highest likelihood score as the next splitting criterion in the tree. The computational complexity of building a decision model using the MLE method is dependent on the complexity of building the model and time taken to calculate the likelihood. The complexity varies according to the method chosen for representing the model, size of the input dataset and to the number of attributes.

3.2 Least Probable Intersections (LPI)

Gershman et al. /**9**/ proposed an optimal splitting criterion which relies on cumulative distribution function (CDF). In this method, the main aim is to find a splitting attribute which has least amount of identifiers that are shared. That splitting attribute must be least probable to generate the subsets randomly. Hence, the splitting attribute with highest score is chosen as the next attribute for the split. The consecutive splitting attribute of the tree would be the attribute which has achieved the highest score. In terms of computational complexity, building a tree using the LPI method is found to be cheap when compared with other methods.

3.3 Coarse-Grained Jaccard (CGJ) Coefficient

The Jaccard similarity coefficient, a measure that is commonly used in clustering, measures the similarity between clusters /**10**/. The goal is to choose the splitting attribute which leads to the smallest possible similarity between the subsets (i.e., an attribute that generates subsets that are different from each other as much as possible). The computational complexity of building the model using the CGJ criterion, the values of the fields from TB can be expressed as a single (concatenated) string. Then, a string matching algorithm can be used to find the intersection between the two subsets of records.

3.4 Fine-Grained Jaccard (FGJ) Coefficient

The fine-grained Jaccard coefficient /**11**/ is capable of identifying partial record matches, as opposed to the coarse-grained method, which identifies exact matches only. It not only considers records which are exactly identical, but also checks to what extent each possible pair of records is similar. There are |A| possible attributes that are candidates for splitting, and therefore, the total complexity of identifying the first splitting attribute is $(|A|\cdot|B|\cdot|T|$. If the tree is not pruned, there would be |A| levels in the tree, therefore the process of selecting a splitting attribute is performed |A| times. Thus, the overall complexity of building the model using the FGJ criterion is bounded by$(|A| \cdot|B|\cdot|T|$.

3.5 Pruning

In a tree induction process, pruning is considered to be an important task. The necessity of using pruning is to build a tree with accuracy and also to avoid over fitting. Pruning can be done in two ways: 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 our system we have followed a pre-pruning approach. It was chosen for the reason that it reduces the time complexity of the algorithm. The decision to prune the branch or not is taken once the next attribute for split is chosen. In this proposed system, two pre-pruning methods are used. They are maximum likelihood estimation (MLE) and least probable intersections (LPI).

4. LINKAGE USING OCCT

Linkage is a process in which a pair is determined match or not. During this phase, each possible pair of test records is tested against the linkage model to determine if the pair is a match or not. This process results in calculating a score which represents the probability of the record pair if it is a true match. The initial score is calculated using maximum likelihood estimation /**12**/.

The input to the algorithm is an instance from table A i.e., TA and an instance from table B i.e., TB. The output of this algorithm is a Boolean value determining whether the instances should be matched or not. The likelihood score for a match between the records is calculated by using the probability of each value, given all other values and appropriate model.

Eventually, the determination of the given records is found match or not by comparing the likelihood score which was calculated earlier with the threshold value. The pair is found to be matched if the pair's score is greater than the threshold value. It is considered as a non-match if the pair's score is less than the threshold value.

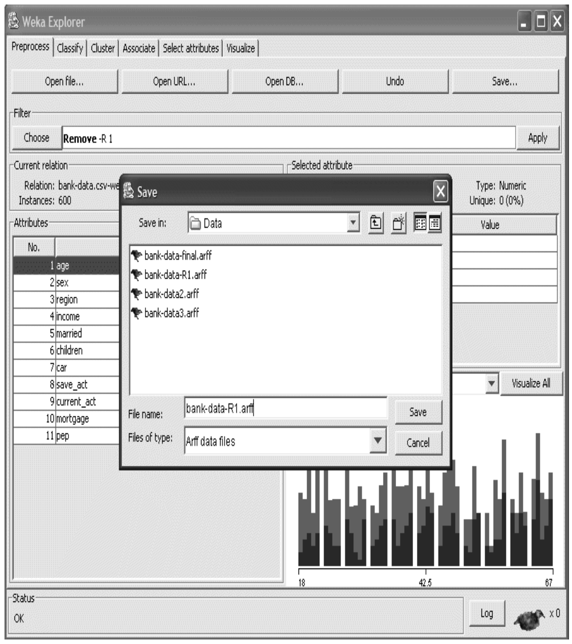

Fig 2: Linkage results

Finally, the pairs that are found to be matched are listed in the output. Also the time taken for the linkage process is calculated and displayed in the output.

5. CONCLUSIONS AND FUTURE WORK

In this system we have represented a one class clustering tree approach which performs one-to-many record linkage. This method is based on a one class decision tree model which sums up the knowledge of which records to be linked together. To summarize, this method allows performing one-to-many linkage while the traditional methods followed one-to-one linkage. Then, we have used a one-class approach which results in matching pairs are only required in the training set, as more number of non-matching (negative) pairs will confuse the model and it will lead to a less accurate model. Another advantage of using OCCT model is that the solution can be easily transformed to rules.The future work may include comparing the OCCT with the other data linkage methods. Also it can be extended to perform many-to-many linkage.

Notes

/1/ M.Dror, A.Shabtai, 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, 2013.

/2/ 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.

/3/ 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.

/4/ S.Ivie, G.Henry, 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.

/5/ P.Christen and K.Goiser, “Towards Automated Data Linkage and Deduplication,” Australian National University, Technical Report, 2005.

/6/ P.Christen and K.Goiser, “Towards Automated Data Linkage and Deduplication,” Australian National University, Technical Report, 2005.

/7/ S.Guha, R.Rastogi and K.Shim, “Rock: A Robust Clustering Algorithm for Categorical Attributes,” Informat- ion Systems, vol. 25, no. 5, pp. 345-366, July 2000.

/8/ D.D.Dorfmann and E.Alf, “Maximum-Likelihood EstiMation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals-Rating-Method Data,” Journal of Math Psychology, vol. 6, no. 3, pp. 487-496, 1969.

/9/ A.Gershman et al., “A Decision Tree Based Recomme- nder System,” in Proc. the 10th Int. Conf. on Innovative Internet Community Services, pp. 170-179, 2010.

/10/ S. Guha, R. Rastogi, and K. Shim, "Rock: A Robust Cluster-ing Algorithm for Categorical Attributes," Information Sys-tems, vol. 25, no. 5, pp. 345-366, July 2000.

/11/ Ibidem

/12/ D.D.Dorfmann and E.Alf, “Maximum-Likelihood EstiMation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals-Rating-Method Data,” Journal of Math Psychology, vol. 6, no. 3, pp. 487-496, 1969.