ISSN:0975-9646

Text Data Linkage of Different Entities Using Occt-One Class Clustering Tree.

E.Afreen Banu.
PG Scholar- CSE dept
Vel Tech Multitech Engg college.
Anna univ- Chennai.
India

R.Karthikeyan,
Assistant professor-CSE dept
Vel Tech Multitech Engg college.
Anna univ-Chennai.
India

Abstract - **A new one to many and many to many data linkage is based on a One-Class Clustering Tree (OCCT) which characterizes the entities that should be linked together. It is evaluated using datasets of Data leakage prevention, Recommender system and Fraud detection. The tree is built such that it is easy to understand and transform into Association rules. The Data Linkage is closely related to entity resolution problem and goal is to identify non-identical records and merge them into single representative record. Non-matching entities in certain domains can tend to fraudulent access. Knuth Morris pratt algorithm is used for fast pattern matching in strings. Pre-Pruning and Post-pruning are made in decision tree that reduce the time complexity of algorithm by reducing the size of tree.**

Keywords— **Text record linkage, One class Clustering, Knuth Morris Pratt algorithm, pruning.**

I. INTRODUCTION

Text mining is a burgeoning new field that attempts to glean meaningful information from natural language text. It may be loosely characterized as the process of analyzing text to extract information that is useful for particular purposes. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. Nevertheless, in modern culture, text is the most common vehicle for the formal exchange of information.

Text categorization is a kind of "supervised" learning where the categories are known beforehand and determined in advance for each training document. In contrast, document clustering is "unsupervised" learning in which there is no predefined category or "class," but groups of documents that belong together are sought. For example, document clustering assists in retrieval by creating links between similar documents, which in turn allows related documents to be retrieved once one of the documents has been deemed relevant to a query.

With massive amounts of data being collected by many businesses, government agencies and research projects, techniques that enable efficient and automatic sharing of large databases between organisations are of increasing importance in many data mining projects. Data from various sources often has to be linked and aggregated in order to improve data quality and integrity, or to enrich existing data with additional information . The aim of such linkages is to match all records that refer to the same entity, for example a customer, a patient, or a business. A related task is finding duplicate records that refer to the same entity within one database, as such duplicates can significantly affect data quality.

In this proposal a new data linkage method aimed at performing one-to-many and many to many data text linkage that can match entities of different types. The inner nodes of the tree consist of attributes referring to both of the tables being matched (TA and TB). 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 non-match.

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 many linkages between entities of same or different types. Another main advantage of the proposed system is using a one-class approach.

II. RELATED WORK

A. Graphical Models for existing Record-Linkage:

The record-linkage problem is the classification task of assigning the record-pair feature vectors to a label\matching or\non-matching. Denote the match-class by a binary variable M , where M = 0 indicates a non-match and M = 1 indicates a match. The goal of probabilistic record-linkage is to formulate a probabilistic model for the match-class M and the feature vector f , and use the same to estimate the probability of the match class given the record-pair feature vector. In an unsupervised setting, this amounts to estimating a generative model for (f ;M)

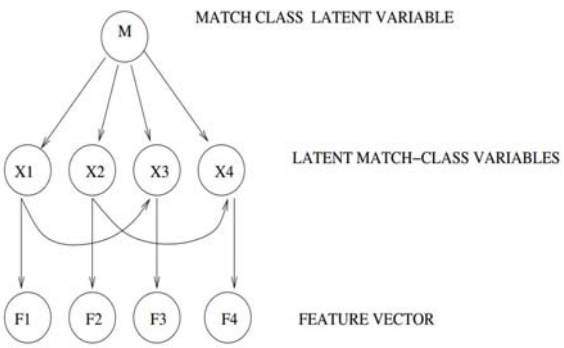

Fig 1: Hiearcical graphical model for record linkage

Specifically, one could interpret the binary-valued middle layer x nodes in Figure 1 as latent match variables for each field. Thus, each node xi in the middle layer corresponds to the match-class of a single field-pair distance feature fi. The top node in Figure 1 is the record-match class latent variable, which gives the match class of the entire record-pair, and which depends on the latent match class variables xi of the individual fields.

B. Fuzzy-CMeans (FCM)

FCM is a representative algorithm of fuzzy clustering which is based on K-means concepts to partition dataset into clusters. The FCM algorithm is a "soft" clustering method in which the objects are assigned to the clusters with a degree of belief. Hence, an object may belong to more than one cluster with different degrees of belief. It attempts to find the most characteristic point in each cluster, named as the centre of one cluster; then it computes the membership degree for each object in the clusters.

C. Indexing for record linkage and deduplication:

When two databases, A and B, are to be matched, potentially each record from A needs to be compared with every record from B, resulting in a maximum number of |A| × |B| comparisons between two records. Similarly, when deduplicating a single database A, the maximum number of possible comparisons is |A| × (|A| − 1)/2, be-cause each record in A potentially needs to be compared with all other records.

The performance bottleneck in a record linkage or deduplication system is usually the expensive de-tailed comparison of field (attribute) values between records making the naive approach of comparing all pairs of records not feasible when the databases are large. For example, the matching of two databases with one million records each would result in 10 ^12(one trillion) possible record pair comparisons.

At the same time, assuming there are no duplicate records in the databases to be matched (i.e. one record in A can only be a true match to one record in B and vice versa), then the maximum possible number of true matches will correspond to min(|A|, |B|). Similarly, for a deduplication the number of unique entities (and thus true matches) in a database is always smaller than or equal to the number of records in it. Therefore, while the computational efforts of comparing records increase quadratically as databases are getting larger, the number of potential true matches only increases linearly in the size of the databases.

III. PROPOSED WORK

A. Inducing linkage model:

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.

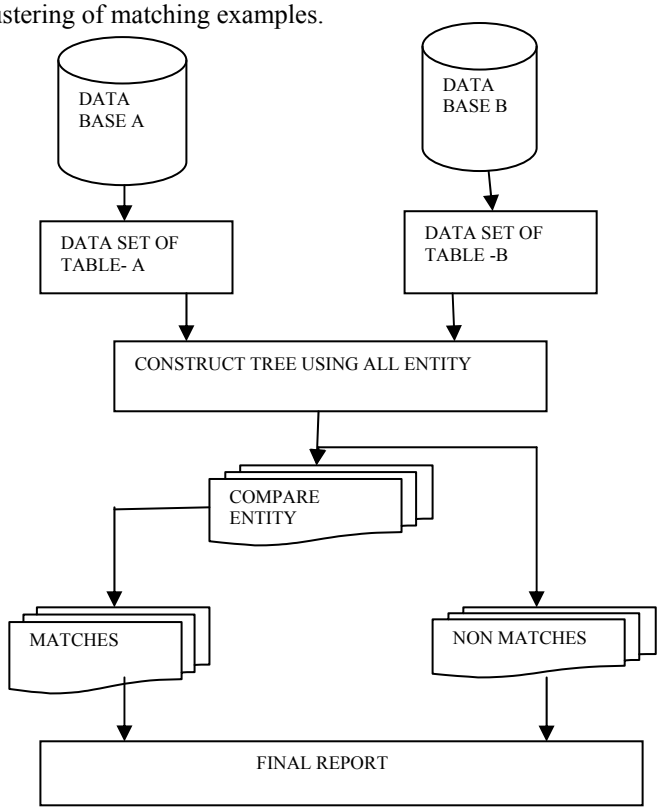

Fig 2: Architecture Diagram

During the linkage (i.e., testing) phase, each pair of records in the testing set is cross validated against the linkage model. The output is a score representing the probability of the record pair being a true match. The score is calculated using MLE. The tested pair is classified as a match if the A describing the attributes of tableTA, and B describing the attributes originating from table TB.

B. One Class Cluster

The one class models separates the desired class of data instances (the core) from other data instances(other median).The goal is to distill a subset of relevant examples, in the space with many outliers. It is the simplest, sequential co-clustering algorithm ,where words are clustered prior to clustering documents. In fig 2, the data set of table A and the data set of table B are constructed into a single one class tree where each of the node has its own number of clusters.

One-class problems are traditionally approached using vector-space methods, where a convex decision boundary is built around the data instances of the desired class, separating it from the rest of the universe. Most of the one-class approaches employ geometrical concepts to capture the notion of relevancy(or irrelevancy) using either hyperplanes or hyperspheres. The latter approach is proposed for One class cluster and in text as an optimization task of identifying the most coherent subset(the core) of k documents drawn from a given pool of

n > k documents. the level of topicality of a word w in terms of the ratio ρ(w) =p(w)/q(w), where p(w) is w's empirical probability (in D), and q(w) is its estimated probability in general English.

C. Knuth-Morris-Pratt algorithm

The Knuth-Morris-Pratt algorithm is based on finite automata but uses a simpler method of handling the situation of when the characters don't match. It helps in fast pattern matching of strings. In the Knuth-Morris-Pratt algorithm, we label the states with the symbol that should match at that point. We then only need two links from each state, one for a successful match and the other for a failure. The success link will take us to the next node in the chain, and the failure link will take us back to a previous node based on the word pattern. Each success link of a Knuth-Morris-Pratt automata causes the "fetch" of a new character from the text. Failure links do not get a new character but reuse the last character fetched. If we reach the final state, we know that we found the substring.

D. 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.

E. 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, a branch is pruned during the induction process if none of the possible splits are found to be more beneficial than the current node. In post-pruning, the tree is grown completely, followed by a bottom-up process to determine which branches are not beneficial. In our system, we follow the prepruning approach. This approach was chosen to reduce the time complexity of the algorithm. The decision whether to prune the branch or not is taken once the next attribute for split is chosen. It is proposed using one Maximum likelihood estimation (MLE) and Least Probable intersections (LPI).

IV. CONCLUSIONS

In this system we have represented a one class clustering tree approach which performs one-to-many and man 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. The Data Linkage is closely related to entity resolution problem and goal is to identify non-identical records and merge them into single representative record. Non-matching entities in certain domains can tend to fraudulent access. Knuth Morris pratt algorithm is used for fast pattern matching in strings. Pre-Pruning and Post-pruning are made in decision tree that reduce the time complexity of algorithm by reducing the size of tree.

For future work, the OCCT can be compared with other linkage methods and to evaluate the masquerade detection between different entities that are matched or unmatched and also to improvise the fast retrieval of text data documents.

ACKNOWLEDGMENT

The author wishes to acknowledge R.Kathikeyan (Asst prof- CSE DEPT) and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this Paper.

REFERENCES

[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] P. Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication" IEEE Trans. Knowledge and Data Engg , vol. 24, no. 9, pp. 1537-1555, Sept. 2012, doi:10.1109/TKDE. 2011.127.

[3] M.B. Salem and S.J. Stolfo, "Modeling User Search Behavior for Masquerade Detection," Proc. 14th Symp. Recent Advances in Intrusion Detection, 2011.

[4] M.Yakout Elmagarmid, H. Elmeleegy, M. Quzzani, and A.Qi, "Behavior Based Record Linkage," Proc. VLDB Endowment,vol. 3, nos. 1/2, pp. 439-448, 2010.

[5] A. Bouza, G. Reif, A. Bernstein, and H. Gall, "Semtree: Ontology-Based Decision Tree Algorithm for Recommender Systems,"Proc.Int'l Semantic Web Conf.,2008

[6] D.E. Knuth, J.H. Morris Jr., and V.R. Pratt, "Fast Pattern Matching in Strings," SIAM J. Computing,vol. 6, no. 2, pp. 323-350, 1977.

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

[8] Ron Bekkerman, Koby Crammer"One-Class Clustering in the Text Domain"-EMNLP '08 Proceedings of the Conference on Empirical methods in general language processing.