International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Clustering Tree based Implementation of Record Linkage on Many-to-Many Relation

V. Balvannanathan[1], R. Siva[2]

[1]PG Student, Department of Computer Science and Engineering, K.C.G College of Technology, Chennai, India

[2]Associate professor, Department of Computer Science and Engineering, K.C.G College of Technology, Chennai, India

Abstract: ***Record linkage or entity resolution are emerging strategy to avoid duplication and other purposes. Recommender domain uses the linkage method to provide efficient results in terms of accuracy. This paper introduces a new Many-to-Many Record Linkage (MMRL) algorithm which links records from one table with a set of records from another table. MMRL algorithm is based on clustering tree which forms the group on each table separately that to be linked. Hierarchical structure such as tree is suitable to understand and execute the linkage process. Intermediate nodes are having less similarity value than end nodes. Each node of the clustering tree contains a cluster instead of a single classification. Prediction accuracy depends on the end node. Jaccard similarity and metaphone similarity are used as distance measures. Prediction result shows whether the records are matched or not. This result proves the efficiency of MMRL algorithm. A data set from movie recommender domain was evaluated for this paper. This MMRL algorithm gives better performance and results.***

Keywords: Record Linkage, Clustering Tree, Similarity, MMRL algorithm.

1. Introduction

Record linkage is the chore of discovering different entries that cite to the similar entity across variety of data sources [5]. The primary objective of the record linkage task is joining data sets that do not share the commonality (i.e., identifier). In general record linkage scenarios include: linking data when combining two different databases [9]; data de-duplication, which is commonly done as a pre-processing step for data mining tasks [8], [12]; determining individuals over various census data sets [6]; associate closely related DNA sequences [11]; and linking galactic objects from different catalogues [13]. It is more general to classify the record linkage into one-to-many and many-to-many [4]. In many-to-many record linkage, the objective is to associate the group of entities from one data set with the multiple matching entities in another data set. In one-to-many data linkage [7], primary objective is link an entity from source data set with a group of matching entities from target data set. Already so many works are done for one-to-many record linkage.

In this paper, a new record linkage method is proposed to performing many-to-many linkage. In addition, while data linkage is usually performed among entities of the same type, the proposed data linkage technique can match entities of different types. For example, in a book database we might want to link a book record with the authors who wrote the book (according to different features that describe the book and features describing the authors). The proposed method links between the entities using Clustering Tree. A clustering tree is a tree in which each of the leaves contains a cluster instead of a single classification. Each cluster is generalized by a set of rules (e.g., a set of conditional probabilities) that is stored in the appropriate leaf [14], [15].

The proposed method was evaluated using movie data set from recommender domain. In the recommender domain, the proposed algorithm is used to link naive users of the system with group of items and items are anticipated to like which is according to their demographic attributes. The result shows proposed algorithm executes well in linkage scenario. In addition, it performs phonetic algorithm such as metaphone to get better results.

2. Literature Survey

This section discusses about the related works to the proposed system. Statistical disclosure control and record linkage [8] methods for the microdata have been undergone with the database that holds variables and terminologies. Terminologies are nothing but the domain of the variables. With the help of this system individuals of the common variables could be identified. Recommender systems [1] have been widely used with the content-based, collaborative and hybrid methods. Rating estimations methods are implemented and it is mostly used in the area of industrial strength. But still this recommender system failed to suit up for wide range of applications. Data Linkage and de-duplication methods [4] widely used and the quality and complexity measures are implemented. The aim is to match the records with the identity. It has been used in the pre-processing stop of data mining works and the algorithms evaluated with the negative matches and more numerical values. Top down induction [15] for clustering tress admits the instance based learning, and implemented with the propositional and relational domains. The system is experimented with pruning, comparison with learners, prediction of multiple attributes, handling the missing information. Machine learning approach with metric based [11] applied to the genealogical record linkage for the names, dates and locations. Comparative methods are implemented with one-to-many relationships. High level of accuracy and precision could be achieved with the help of this implementation. The experimented result shows that algorithm performs well than already performed methods.

3. Terms and Definitions

Record Linkage (RL) is the task of different manifestations of real world entities in different records or mentions by identifying and linking/grouping. We must know the difference between record linkage and linked data. Both are different. Record Linkage is the focused area of determining entities which has more probability to be matched over different datasets. Linked Data is about structuring and publishing data to facilitate the discovery of related information. Record Linkage can be divided into two classes. One is deterministic record linkage which involves exact matches on unique identifiers or combinations of fields that uniquely identify given individuals. All identifiers must agree for a link to be made. It can be a better choice when the entities in the datasets are identified by a common identifier or when there are several identifiers whose quality of data is relatively high. Figure1 shows that difference between probabilistic linkage and deterministic linkage

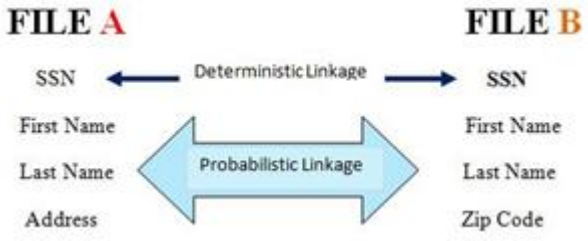

Figure 1: Types of linkage process

Another one is probabilistic record linkage that takes a different approach to the record linkage problem by taking potential identifiers, computing weights for each identifier based on its estimated ability to correctly identify a match or a non-match and using these weights to calculate the probability that two given records are the related entity. In this paper, probabilistic record linkage is more concentrated. To achieve probabilistic record linkage, threshold value should be used. Probabilities of record pairs which are above the upper threshold are considered to be matches and probabilities with below the lower threshold are considered to be non-matches. One more result from linkage process may be obtained that is possible matches. Under possible matches, record pairs are having probability value between upper and lower threshold. It should be done by manual process to get a result that whether the record pairs are matched or not.

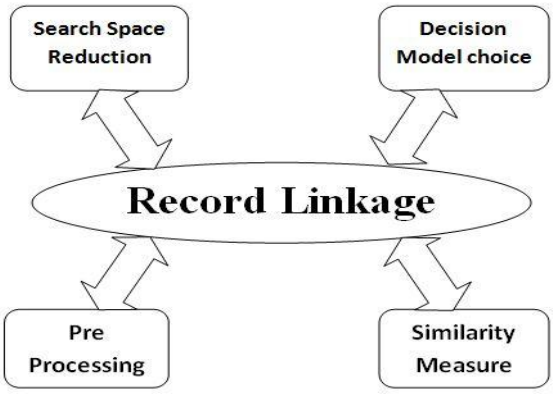

Figure 2: Components of linkage process

The collective components could make better results from linkage process. These components are such as (1) pre-processing, 1(2) search space reduction, (3) similarity measure, (4) decision model choice. Pre-processing is for making the inputs as suitable to run the model/algorithm without any error such as typo, null values, etc. Search space reduction can be achieved by implementing the methods like blocking/indexing. This increases the speed and efficiency of the model. Similarity measure determines how much entities are similar or related. Decision model is used for representing a structural design of the logic embodied by rules and statements.

4. Architecture of MMRL

This part shows that how the model is built to implement the MMRL record linkage process. Design of the architecture process comprises the various tasks. Fig3 explains the architecture of this process. Records from different database are collected and it might be involved in pre-processing work. Records may have unwanted fields or data. In the pre-processing work, missing values in the records are filled with the global value "UNKNOWN". Compound attributes like address are normalised. These processes could make the data as suitable to evaluate the model. After completing pre-processing work, data are split into two sets. Training set is used for develop the model. 2/3 of data are assigned as training set and remaining data are used for testing. Test dataset are used to check how the model works and the performance in terms of accurate or positive results.

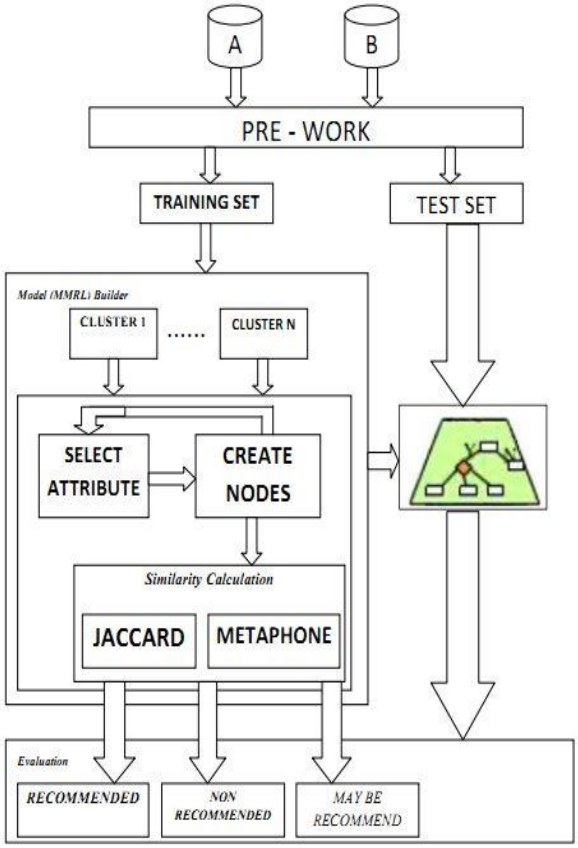

Figure 3: Architecture of MMRL

Now training data sets are going to use to create the model. In the first step of building the model, initial clustering process are performed. This process may be called as

indexing. It reduces the search space and time to improve the efficiency of the model. For example, if attribute age is chosen as a candidate, we can split the data into three category that young, middle, older. Now we can link the records corresponding to their category and no need to check the similarity of the records in other category. It creates n clusters and we may use which the model needs. Next step is attribute selection.

Attribute selection is the important step to make a good split during the tree construction. In c4.5 algorithm, information gain is used to attribute selection process. If the gain value for an attribute is greater than threshold value, then we can select this attribute as for split. After attribute selection process, starts the node creation process. Nodes are having clusters of value instead of single value. It forms the hierarchical structure as like tree. Many rules and conditions are applied for each node. This process (splitting nodes) is repeated until reach the leaf/end node.

Results from the above step are sent to the next step called similarity calculation. It defines how much amount records are similar. Here we use two similarity algorithms such as Jaccard similarity algorithm and metaphone phonetic similarity algorithm.

Jaccard similarity is a statistic used for comparing the similarity of different records. For example, C1={0,1,2}, C2={3,4}, C3={5,6},C4={7,8,9}. Let us consider C1 might represent action movies, C2 comedies, C3 documentaries, and C4 horror movies. Now we can represent A = {C1, C3} and B = {C1, C2, C3, C4} since A only contains elements from C1 and C3, while B contains elements from all clusters. The Jaccard distance of the clustered sets is now

$$\textbf{Jac\_Sim}(A, B) = \frac{|\{C1, C2\}|}{|\{C1, C2, C3, C4\}|} = 0.5$$

The phonetic algorithm is an algorithm used for indexing of words by their pronunciation. It converts the input words into codes. By comparing these codes, we can get similar records.

Final step of this work is evaluation. In this step, according to score value we can categories the records into three. Here we set the threshold value. The records having greater score value than threshold are more recommended. The records having less score value than threshold are not recommended. The records having equal score value to threshold may be recommended. It shows the efficiency and accuracy of the model.

5. Proposed MMRL Algorithm

The proposed Many-to-Many Record Linkage algorithm is used to link set of entities from one table with different entities from another table.

Linkage is a process in which a pair is determined match or not. During evaluation, each possible pair of test records is 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 input to the algorithm is a set of instances from table X (i.e., Tab_X) and a set of instances from table Y (i.e., Tab_Y). Output of this algorithm is determining whether the instances could be matched or not. The score for a match between the records is calculated by using jaccard coefficient and metaphone phonetic similarity calculation.

Eventually, the determination of the given records is found match or not by comparing the 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.

MMRL ALGORITHM

INPUT

X - Tab_A's attributes set
Y - Tab_B's attributes set
Tab_{XY} – Matching instance set
Th – Threshold value

OUTPUT

RS - Set of recommended values for each entity

STEPS

1) Create root node N_r.
2) Check the nodes whether it having minimum size.
3) If the condition satisfies, create next level nodes N_c under parent node.
4) Repeat steps 2 and 3 until no node having minimum size.
5) In step2, if condition fails, return root node as tree.
6) Then perform similarity calculation using jaccard coefficient.
7) Return records to be recommended which has score value above the threshold value.
8) Evaluate possible matching records using metaphone which has equal score value with threshold
9) Set the results under recommended category from metaphone which has positive score value.

Metaphone phonetic similarity calculation is carried out for the records which have equal score value with threshold. Positive results from metaphone similarity calculation are added to the recommended category. Finally, the pairs that are found to be matched are listed as recommended to the users.

6. Result

In this work, drawbacks of one-to-many data linkage were mitigated and improve the quality of many-to-many record linkage. By this record linkage, accuracy of the recommendation will be improved using mmrl algorithm. To achieve this linkage process, various algorithms such as jaccard metaphone similarity algorithms are used. Figure4 explains that how the many-to-many record linkage will link two tables with many entities using similarity score. The binary score helps to take a decision whether the records are matched or not.

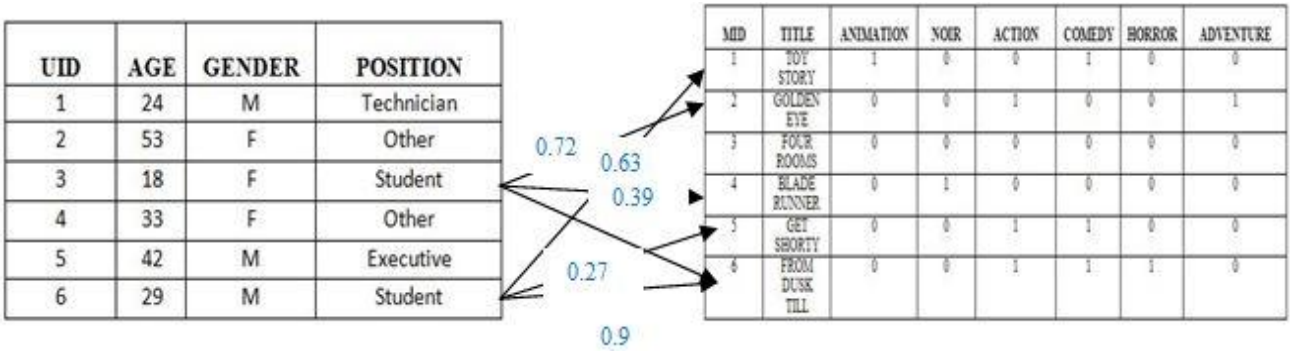

Figure 4: Pictotial representation of many-to-many task

Entities should be chosen to perform the linkage process between entities. In figure5, the page is shown which is for entities selection. TableX has many entities. Then similarity calculation will be performed for each selected entities from one table with different entities from another table.

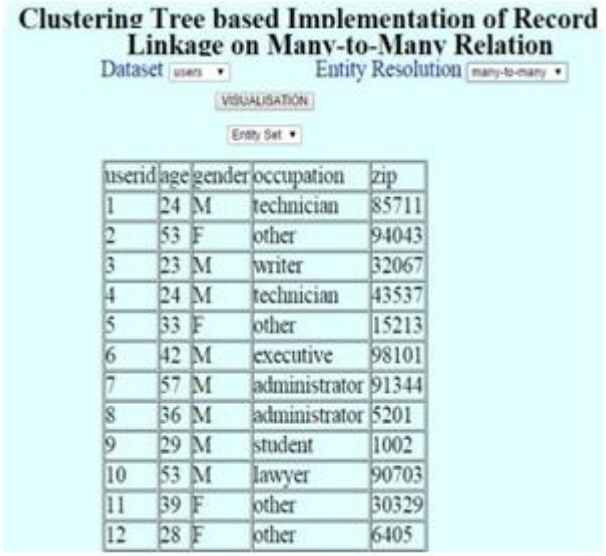

Figure 5: Entities Selection

After selecting entities, similarity calculation process is done. The result will be displayed as shown in the figure6.

Here jaccard similarity measure function is used. It also shows the distance from the selected entity. The records by satisfying threshold value are considered to recommend for the user.

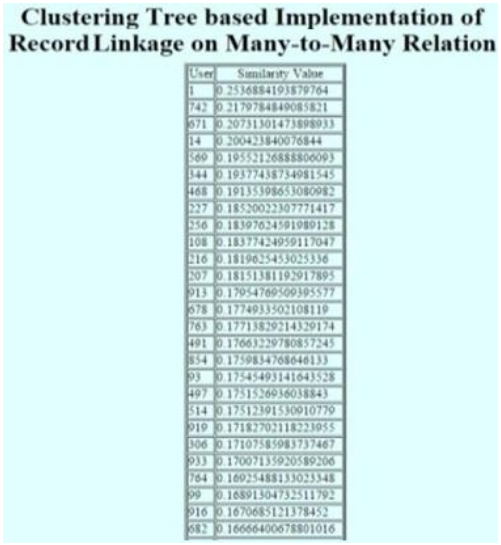

Figure 6: Similarity measure

$$\textbf{Jac\_Sim}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Jaccard similarity is defined as ratio between common in A and B by union of A and B.

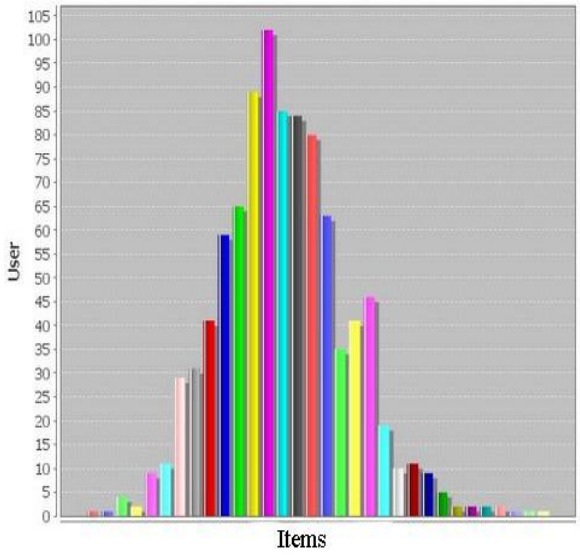

Figure 8: graph for user and items

Threshold value plays vital role for recommender domain. In figure8, the graph was plotted that shows particular item links with how many users.

7. Conclusion

In this work, implementation of many-to-many record linkage using clustering tree is carried out with effective method. To get good results, here better similarity measures such as jaccard and metaphone are used. De-duplication is obtained from this record linkage methodology. Attributes used in this work are discrete which is obtained from movie recommender domain. The proposed many-to-many record linkage using clustering tree can be used to obtain all records that match the given records. Movie dataset are focused for this many-to-many record linkage. This work can be proceed as a future work with various distance metrics and continuous attributes for better and accurate results.

References

[1] Adomavicius G and Tuzhilin A, 'Toward the Next Generation of Recommender Systems: A Survey of the

State-of-the-Art and Possible Extensions', *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no.6, pp. 739-749, June 2005.

[2] Bin Jiang, Jian Pei, Yufei Tao, and Xuemin Lin, 'Clustering Uncertain Data Based on Probability Distribution Similarity', *IEEE Transactions On Knowledge And Data Engineering*, Vol. 25, No. 4, April 2013.

[3] Christen P, 'A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication', *IEEE Transaction on Knowledge and Data Engineering*, vol. 24, no. 9, pp. 1537-1555, Sept. 2012.

[4] Christen P and Goiser K, 'Quality and Complexity Measures for Data Linkage and Deduplication', Quality Measures in Data Mining, vol. 43, pp. 127-151, 2007.

[5] Fellegi IP and Sunter, 'A Theory for Record Linkage', *J. Am. Statistical Soc.*, vol. 64, no. 328, pp. 1183-1210, Dec. 1969.

[6] Larsen MD and Rubin DB, 'Iterative Automated Record Linkage Using Mixture Models', *J. Am. Statistical Assoc.*, vol. 96, no. 453, pp. 32-41, Mar. 2001.

[7] Ma'ayan Dror, Asaf Shabtai, Lior Rokach, and Yuval Elovici, 'OCCT: A One-Class Clustering Tree for Implementing One-to-Many Data Linkage', *IEEE Transactions on Knowledge And Data Engineering*, Vol. 26, No. 3, March 2014.

[8] Torra V and Domingo-Ferrer J, 'Disclosure Risk Assessment in Statistical Microdata Protection via Advanced Record Linkage', *Statistics and Computing*, vol. 13, no. 4, pp. 343-354, 2003.

[9] Yakout M, Elmagarmid A.K, Elmeleegy H,Quzzani M, and Qi A, 'Behaviour Based Record Linkage', *Proc. VLDB Endowment*, vol. 3, nos. ½, pp. 439-448, 2010.

[10] Yung-Shen Lin, Jung-Yi Jiang, and Jue Lee, 'A Similarity Measure for Text classification and Clustering', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 7, July 2014.

[11] S. Ivie, G. Henry, H. Gatrell, and C. Giraud-Carrier, "A Metric-Based Machine Learning Approach to Genealogical Record Linkage," Proc. Seventh Ann. Work shop Technology for Family History and Genealogical Research, 2007.

[12] F. De Comite´, F. Denis, R. Gilleron, and F. Letouzey, "Positive and Unlabeled Examples Help Learning," Proc. 10th Int'l Conf. Algorithmic Learning Theory, pp. 219-230, 1999.

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

[14] P. Langley, Elements of Machine Learning. Morgan Kaufmann, 1996.

[15] H. Blockeel, L.D. Raedt, and J. Ramon, "Top-Down Induction of Clustering Trees," ArXiv Computer Science e-prints, pp. 55-63, 1998.