One to Many Data Linkage and Clustering using Maximum Likely-Hood Estimations

V. A. Kiruthika
PG Scholar,
Department of Computer Science and Engineering,
J.K.K. Nattraja College of Engineering and Technology,
Kumarapalayam, Tamilnadu, India

D. Mahesh
Assistant Professor,
Department of Computer Science and Engineering,
J.K.K. Nattraja College of Engineering and Technology,
Kumarapalayam , Tamilnadu, India

Abstract- **Record linkage is the process of matching records from several databases that refer to the same entities. When applied on a single database, this process is known as deduplication. Increasingly, matched data are becoming important in many applications areas, because they can contain information that is not available otherwise, or that is too costly to acquire. Removing duplicate records in a single database is a crucial step in the data cleaning process, because duplicates can severely influence the outcomes of any subsequent data processing or data mining. With the increasing size of today's databases, the complexity of the matching process becomes one of the major challenges for record linkage and deduplication. In recent years, various indexing techniques have been developed for record linkage and deduplication. They are aimed at reducing the number of record pairs to be compared in the matching process by removing obvious nonmatching pairs, while at the same time maintaining high matching quality. This paper presents a survey of variations of six indexing techniques. Their complexity is analyzed, and their performance and scalability is evaluated within an experimental framework using both synthetic and real data sets. These experiments highlight that one of the most important factors for efficient and accurate indexing for record linkage and deduplication is the proper definition of blocking keys.**

Keywords—Clustering, classification, mapping, data matching, decision tree induction

I. INTRODUCTION

Data Mining refers to extracting or mining knowledge from large databases. Data Mining and knowledge discovery in the databases is a new interdisciplinary field, merging ideas from statistics, machine learning, databases and parallel computing. Data Mining is the non-trival process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise.

Data Mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of Data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining works to analyze data stored in data warehouses that are used to store that data that is being analyzed. That particular data may come from all parts of business, from the production to the management.

The actual Data Mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the Data Mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional step.

Data linkage or Record linkage refers to the task of finding records in a data set that refers to the same entity across different data sources. It is necessary when joining data sets based on entities that may or may not share a common identifier. Record Linkage is also called as the Data Linkage which is the same process. The process of linking and aggregating records from one or more data sources representing the same entity which is also called as Data Matching, Data Integration, Data Scrubbing, ETL(Extraction, Transformation and Loading), Object Identification, Merge-purge, etc. The goal of the Data Linkage task is to joining datasets that do not share a common identifier (i.e., a foreign key). Common Data Linkage scenarios include: linking data when combining two different databases and Data Deduplication.

Data Deduplication is a data compression technique for eliminating redundant data which is commonly done in as a preprocessing step for data mining tasks for identifying individuals across different data sets. It is common to divide data linkage into two types**:** **one-to-one** and **one-to-many**. In **one-to-one** data linkage, the goal is to associate an entity from one dataset with a single matching entity in another dataset. In **one-to-many** data linkage, the goal is to associate an entity from the first dataset with a group of matching entities from the other dataset. Major benefits of data Deduplication is to reduce the storage space on tape. Deduplication process is of different types namely: Post-process deduplication, In-line deduplication, Source versus target deduplication. Deduplication methods are Chunking, Client backup deduplication, Primary storage and secondary storage.

Decision tree learning uses a decision tree as a predictive model which maps o about an item to conclusions about the item's target value. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. More descriptive names for such tree models are classification trees or regression trees.

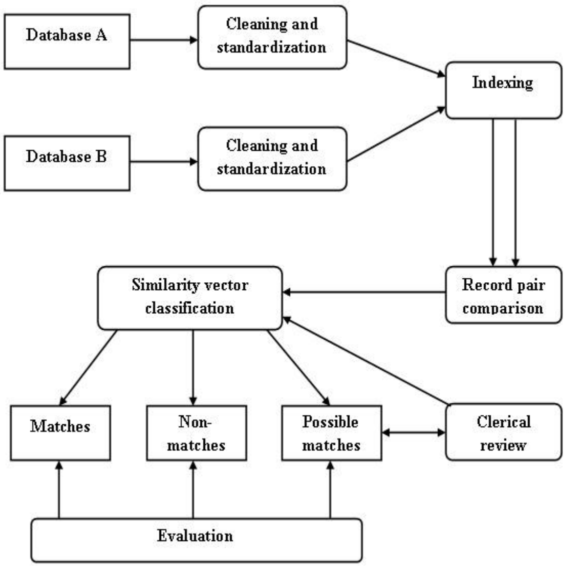

Fig. 1. Outline of general record linkage process.

In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. A Decision Tree is a flowchart-like structure in which each node denotes a test on an attribute. Each branch represents an outcome of the test and the leaf nodes represent classes or class distributions. Unknown samples can be classified by testing attributes against the tree. The path traced from root to leaf holds the class prediction for that sample. Decision tree uses different algorithm's one of the earliest algorithm is Hunt's Algorithm, and other algorithm's are CART, ID3, C4.5, C5.0, SLIQ, SPRINT are the other decision tree algorithm's.

II. LITERATURE SURVEY

A. Probabilistic data generation for deduplication and data linkage.

In many data mining projects the data to be analyzed contains personal information, like names and addresses. Cleaning and preprocessing of such data likely involves deduplication or linkage with other data, which is often challenged by a lack of unique entity identifiers. In recent years there has been an increased research effort in data linkage and deduplication, mainly in the machine learning and database communities. Publicly available test data with known deduplication or linkage status is needed so that new linkage algorithms and techniques can be tested, evaluated and compared. However, publication of data containing personal information is normally impossible due to privacy and confidentiality issues. An alternative is to use artificially created data, which has the advantages that content and error rates can be controlled, and the deduplication or linkage status is known. Controlled experiments can be performed and replicated easily. This paper present a freely available data set generator capable of creating data sets containing names, addresses and other personal information.

Finding duplicate records in one, or linking records from several data sets are increasingly important tasks in the data preparation phase of many data mining projects, as often information from multiple sources needs to be integrated, combined or linked in order to allow more detailed data analysis or mining. The aim of such linkages is to match all records related to the same entity, such as a patient or customer. As common unique entity identifiers (or keys) are rarely available in all data sets to be linked, the linkage process needs to be based on the existing common attributes. Data linkage and deduplication can be used to improve data quality and integrity, to allow re-use of existing data sources for new studies, and to reduce costs and efforts in data acquisition. In the health sector, for example, linked data might contain information that is needed to improve health policies, and that traditionally has been collected with time consuming and expensive survey methods. Artificially generated data can be an attractive alternative. Such data must model the content and statistical properties of comparable real world data sets, including the frequency distributions of attribute values, error types and distributions, and error positions within these values.

B. *Improved One-to-Many Record Linkage using One Class Clustering Tree.*

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 inner nodes of the tree consist of features of the first set of entities. The leaves of the tree represent features of the second set that are matching. The data is split using two splitting criteria. Also two pruning methods are used for creating one-class clustering tree. The proposed system results better in performance of precision and recall.

C. *Top-down induction of clustering trees.*

An approach to clustering is presented that adapts the basic top-down induction of decision trees method towards clustering. The resulting methodology is implemented in the TIC (Top down Induction of Clustering trees) system for first order clustering. The TIC system employs the first order logical decision tree representation of the inductive logic programming system Tilde. Various experiments with TIC are presented, in both propositional and relational domains. A clustering tree is a decision tree where the leaves do not contain classes and where each node as well as each leaf corresponds to a cluster.

To induce clustering trees, we employ principles from instance based learning and decision tree induction. More specifically, we assume that a distance measure is given that computes the distance between two examples. Furthermore, in order to compute the distance between two clusters. First order logical decision trees are similar to standard decision trees, except that the test in each node is a conjunction of literals instead of an test on an attribute. They are always binary, as the test can only succeed or fail. Clustering can also be done in an unsupervised manner however. When making use of a distance metric to form clusters, this distance metric may or may not use information about the classes of the examples. Even if it does not use class information, clusters may be coherent with respect to the class of the examples in them.

This principle leads to a classification technique that is very robust with respect to missing class information. A system for top-down induction of clustering trees called TIC has been implemented as a subsystem of the ILP system Tilde. TIC employs the basic TDIDT framework as it is also incorporated in the Tilde system. The main point where TIC and Tilde differ from the propositional TDIDT algorithm is in the computation of the tests to be placed in a node, for details. Furthermore, TIC differs from Tilde in that it uses other heuristics for splitting nodes, an alternative stopping criterion and alternative tree post-pruning methods. In this first experiment we want to evaluate the effect of pruning in TIC on both predictive accuracy and tree complexity. We have applied TIC to two databases: Soybeans (large version) and Mutagenesis. We chose these two because they are relatively large (as noted before, the pruning strategy is prone to random influences when used with small datasets). Future work on TIC includes extending the system so that it can employ first order distance measures, and investigating the limitations of this approach (which will require further experiments).

III. EXISTING SYSTEM

As many businesses, government agencies and research projects collect increasingly large amounts of data, techniques that allow efficient processing, analyzing and mining of such massive databases have in recent years attracted interest from both academic and industry. One task that has been recognized to be of increasing importance in many application domains is the matching of records that relate to the same entities from several databases. Often, information from multiple sources needs to be integrated and combined in order to improve data quality, or to enrich data to facilitate more detailed data analysis. OCCT, a one-class decision tree approach was used for performing one-to-many and many-to-many data linkage. The proposed method is based on a one class decision tree model that encapsulates the knowledge of which records should be linked to each other. This paper contains a proposed of four possible splitting criteria and two possible pruning methods that can be used for inducing the data models. The task of record linkage is now commonly used for improving data quality and integrity, to allow re-use of existing data sources for new studies, and to reduce costs and efforts in data acquisition.

A. *Disadvantages*

- Training set requires only matching pair not with a non-matching pair.
- It is more critical to reduce the linkage computation since inducing the OCCT model can be done offline while the linkage phase.
- Splitting the tree by using attributes from both tables would increase the linkage time.

IV. PROPOSED SYSTEM

It can be believed that when enough non-matching examples are available, the **J48** model is preferable and would probably work better. An important advantage of the OCCT model over a decision tree-based data linkage solution is the simplicity of the model which can easily be transformed to rules of the two tables. This is not the case in other decision tree based linkage models where the inner nodes of the tree consist of attributes from both tables TA and TB, thus making them difficult to read and almost impossible to translate into rules.

There are number of candidate record pairs that can be generated by techniques that have been estimated and their efficiency and scalability also been evaluated using various data sets. Thus, the experiment highlight that one of the most important factors for efficient and accurate indexing for record linkage and deduplication is the proper definition of blocking keys. Because training data in the form of known true matches and non-matches is often not available in real world applications. The indexing techniques in this investigation are heuristic approaches that aim to split the records in a database into blocks such that matches are inserted in to the same block and non-matches into different blocks.

A. *Advantages*

- Data cleaning and standardization is the conversion of the raw input data into well defined, consistent forms, as well as the resolution of inconsistencies in the way information is represented and encoded.
- Records are compared in detail in the comparison step using a variety of comparison functions which is appropriate to the content of the record fields.
- Training Data sets includes the non-matching record pairs.

V. EXPERIMENTS

A. *Self-service Authentication Acces*

In this module we design the windows for the project. These windows are used to send a message from one peer to another. We use the Swing package available in Java to design the User Interface. Swing is a widget toolkit for Java. It is part of Sun Microsystems' Java Foundation Classes an API for providing a graphical user interface for Java programs. In this module mainly we are focusing the login design page with the Partial knowledge information. Application Users need to view the application they need to login through the User Interface GUI is the media to connect User and Media Database and login screen where user can input his/her user name, password and password will check in database, if that will be a valid username and password then he/she can access the database.

B. *Cleaning and Standardization with Dataset.*

After Successful completion of the first module now evolve into second module is cleaning and standardization of data sets. Here what we are going to consider as a Datasets means that which we want to merge as a combination and we can form a single table with efficient data sets for the effective mining. If we consider data mining in to two major parts means one is useful data extraction from the data ware house or database and the another one is data arrangement in this concept we are majorly concentrate and achieve the data clustering of different datasets for the better and efficient retrieval of data mining.

C. *Indexing of Record Pair Comparision*

This is the third and major module of our application in this module what we are going to achieve means, record wise comparisons of each and every table what we are going to mention table lists. But before this module the pre requisite thing is we need to give very clean and standardized data.

Why we are going to perform this module means we need to form a separate table for efficient mining. After comparisons of each and every thing in both of the tables what we are going to mention. It will form a separate index for the newly evolved or created table.

D. *Similarity Vector Classification*

This is the fourth and important module in our application. In this module what we are going to perform means similar data of the tables what we are given and we can consider the indexing of their keywords. And here we can form a fresh and separate table which will yield the better results compared to the existing mechanisms. Because of this number of user queries which they are requesting data will reduce the time complexity why because in existing there are number of tables available so, the user given query should be verify with those tables and evolve results but now there is only one efficient table which can reduce the query response time.

E. *Overall Performance Evaluation*

This is the last and final module of our application, in this module what we are going to perform means, the evaluation of the entire datasets, keywords, tables, requests, requestors and how much we are going to perform efficiently without getting any late to the end user or requestor. This module also performs the functionalities like user requested queries and what the information getting to them and what the cluster we need to join to the another table to the efficient handling of queries and performing operations like analysis, evaluation of the application.

VII. CONCLUSION

A one-class decision tree approach for performing one-to-many and many-to-many data linkage. The proposed method is based on a one-class decision tree model that encapsulates the knowledge of which records should be linked to each other. In addition, we proposed four possible splitting criteria and two possible pruning methods that can be used for inducing the data models. Our evaluation results show that the proposed algorithm is effective when applied in different domains. Our goal is to link a record from a table TA with records from another table TB. The generated model is in the form of a tree in which the inner nodes represent attributes from TA and the leafs hold a compact representation of a subset of records from TB which are more likely to be linked with a record from TA, whose values are according to the path from the root of the tree to the leaf

For future work, we plan to compare the OCCT with other data linkage methods. In addition, we plan to extend the OCCT model to the many-to-many case and to handle continuous attributes. We also propose evaluating the results on additional domains, and characterizing which splitting criterion and pruning methods should be applied for each type of domain.

REFERENCES

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

[2] J.Domingo-Ferrer and V.Torra, “Disclosure RiskAssessment in Statistical Microdata Protection via Advanced Record Linkage,” Statistics and Computing, vol. 13, no. 4, pp. 343-354, 2003.

[3] M.D.Larsen and D.B. Rubin, “Iterative Automated Record Linkage Using Mixture Models,” J. Am. Statistical Assoc., vol. 96, no. 453, pp. 32-41, Mar. 2001

[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] 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.

[6] 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.

[7] L. Gu and R. Baxter, “Decision Models for Record Linkage,” Data Mining, vol. 3755, pp. 146-160, 2006.

[8] Frank, M.A. Hall, G. Holmes, R. Kirkby, and B. Pfahringer, “WEKA - A Machine Learning Workbench for Data Mining,” The Data Mining and Knowledge Discovery Handbook, pp. 1305-1314, Springer, 2005.

[9] P. Christen and K. Goiser, “Towards Automated Data Linkage and Deduplication,” technical report, Australian Nat’l Univ., 2005.

[10] De Comite´, F. Denis, R. Gilleron, and F. Letouzey, “Positive and Unlabeled Examples Help