International Journal of Research
ISSN: 2348-6848 Vol-3, Special Issue-3
International Conference on Research and Recent Trends in Engineering and Technology. (ICRRTET)
Held on 27th January 2016 organized by **Sai Polytechnic College,** Kinhi Jawade, Yavatmal, Maharastra, India.

A Survey on Various Indexing Technique for Record Linkage

Pranay Tambekar[1]; Roshan Moharle[2]& Komal Kopare[3]
Department of Computer Engineering, Bapurao Deshmukh College of Engineering, Sevagram, Wardha.
[1]pranaytambekar02@gmail.com
[2]ramoharle@gmail.com
[3]komalkopare123@gmail.com

Abstract *— Record linkage is the problem of identifying similar records across different data sources. Record linkage is an important process in data integration, which is used in merging, matching and duplicate removal from several databases that refer to the same entities. De-duplication is the process of removing duplicate records in a single database. In recent years, data cleaning and standardization becomes an important process in data mining task. 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 de-duplication. This paper presents an analysis of record de-duplication techniques and algorithms that detect and*
remove the duplicate records.

Keywords — Data linkage; record linkage; data mining; clustering; classification.

1. INTRODUCTION

Generally data mining is the process of analysing data from different perspective and summarizing it into useful information. Data mining is one of the analytical tool for analysing data. It allows user to analyse data from many different dimensions or categories and summaries the relationship identified. Technically data mining is the process of finding correlation or patterns among dozens of fields in large relational databases. An increasingly important task in the data preparation phase of many data mining Projects is linking or matching records relating to the same entity from several databases As often information from multiple source need to integrate and combine in order to enrich data and allow more detailed data mining studies The aim of such linkage is to match and aggregate all records relating to same entity, such as a patient, a customer, a business, a consumer product or a genome sequence.

Record linkage is the process of matching records from several databases. Record linkage can be used to increase data quality and data integrity, to allow reuse of existing data sources for new studies, and to reduce the cost and efforts in data acquisition. Record linkage and de-duplication can be used to identify people who register for benefits multiple times or who work and collect unemployment money. Another application of current interest is the use of data linkage in crime and terror detection.

Statistical agencies have employed record linkage for several decades on a routinely basis to link census data for further analysis. Many businesses use de-duplication and record linkage techniques with the aim to de-duplicate their databases to improve data quality or compile mailing lists, or to match their data across organizations, for example, for collaborative marketing or e-Commerce projects. Many government organizations are now increasingly employing record linkage, for example within and between taxation offices and departments of social security to identify people who register for assistance multiple times, or who work and collect unemployment benefits. Other domains where linkage is of high interest are fraud and crime detection, as well as national security. Security agencies and crime investigators increasingly rely on the ability to quickly access files for a particular individual under investigation, or cross-check records from disparate databases, which may help to prevent crimes and terror by early intervention.

The problem of finding records that relate to the same entities not only applies to databases that contain information about people. Removing duplicate records in a single database is a crucial step. De-duplication can be achieved more efficiently by using indexing techniques. One or more (blocking) indexes need to be built with the aim of grouping together records that potentially match and thus reducing the huge number of possible comparisons. While this grouping should reduce the number of comparisons made as much as possible, it is important that no potential match is overlooked because of the indexing process. After indexes are built, records within the same index block are compared by using field comparison functions, resulting in a weight vector for each record pair compared. These weight vectors are then given to a classifier that decides if a record pair constitutes a match, non-match or possible match. Section 2 describes literature survey, Section 3 describes proposed architecture, Section 4 describes conclusion and Section 5 describes references.

2. LITRETURE SURVEY

Peter christen invented A Survey of Indexing Techniques for Scalable Record Linkage and De-duplication published on ieee transaction on knowledge and data engineering in the year 2011. This paper presents a survey of twelve variations of six indexing techniques. Their complexity is analysed, and their performance and scalability is evaluated within an experimental framework using both synthetic and real data sets. No such detailed survey has so far been published. The number of candidate record pairs generated by these techniques has been estimated theoretically, and their efficiency and scalability has been evaluated using various data sets. These experiments highlight that one of the most important factors for efficient and accurate indexing for record linkage and de-duplication 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, it is commonly up to domain and linkage experts to decide how such blocking keys are defined.[1]

Nishand. K. Ramasami. S. and T. Rajendran invented An Efficient way of record linkage system and de-duplication using indexing technique, classification and FEBRL framework published on International Journal of Emerging Science and Engineering ISSN:2319-6378, Issue-7 in the year May-2013.In this paper indexing techniques along with classification and comparators are used. All these are implemented in FEBRL framework. In that paper the cluster algorithm are used to classify the record pair into link ,non-link ,or if this decision should be done by a human review, possible links. Datasets used are real data and artificially generated data. Because in real datasets it is difficult to identify the deviations in results. So artificial datasets are also used. Artificial data are generated using FEBRL framework. This generator first creates original records based on frequency tables that contain real name and address values, as well as other personal attributes, followed by the generation of duplicates of these records based on random modifications such as inserting, deleting, or substituting characters, and swapping, removing, inserting ,splitting, or merging words. The types and frequencies of these modifications are also based on real characteristics. The true match status of all record pairs is known. The original and duplicate records were then stored into one file each to facilitate their linkage.[2]

Lalitha.L1, Maheswari. B2, Karthik S3 invented A Detailed Survey on Various Record De-duplication Methods published on International Journal of Advanced Research in Computer Engineering & Technology, in the year October 2012.In this paper they suggested an approach for duplicate record detection and removal. In this approach, they first convert the attributes of data into numeric form. Then, this numeric form is used to create clusters by using K-Means clustering algorithm. The use of clustering reduces the number of comparisons. After that the divide and conquer technique is used in parallel with these clusters for identification and removal of duplicated records. Here, this technique identifies all type of duplicated records like fully duplicated records, erroneous duplicated records and partially duplicated records. This technique is only applicable for single table instead of multiple sorted tables. The performance is measured by using the terms like true positives, false positives, false negatives, precision, recall and F-Score.[3]

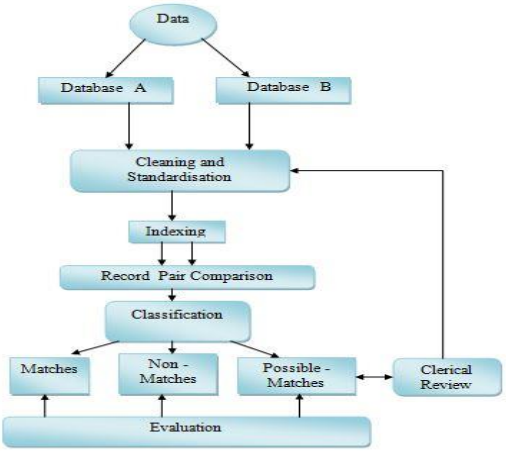
Peter Christen invented advanced record linkage method and privacy aspects for population reconstruction published on Research School of Computer science. The Australian National University Canberra ACT 0200Australia. In this paper, they describe these major challenges of record linkage in the context of population reconstruction, outline recent developments of advanced record linkage methods, and provide direction for future research.. The main challenges of record linkage are (1) scalability to the increasingly large databases common today (2) accurate and efficient classification of compared records into matches and non-matches in the presence of variations and errors in the data; and (3) privacy issues that occur when the linking of records is based on sensitive personal information about individuals. The first challenge has been ad- dressed by the development of scalable indexing techniques, the second through advanced classification techniques that either employ machine learning or graph based methods, and the third challenge is investigated by research into privacy-preserving record linkage.[4]

Sunita Yeddula & K. Lakshmaiah invented an Investigation of Techniques For Efficient & Accurate Indexing For Scalable Record Linkage & De-duplication published on Dept of CSE, Madanapalle Institute of Technology and Science, Madanapalle. This paper presents a survey of variations of six indexing techniques. Their complexity is analysed, 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 de-duplication is the proper definition of blocking keys. The number of candidate record pairs generated by these techniques has been estimated and their efficiency and scalability has been evaluated using various data sets. These experiments highlight that one of the most important factors for efficient and accurate indexing for record linkage and de-duplication 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 in to different blocks.[5]

Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis and Vassilios S. Verykios invented A survey on Duplicate Detection Algorithm published on ieee transaction on knowledge and data engineering, vol.19, no.1 in the year January 2007. In this paper, we present a thorough analysis of the literature on duplicate record detection. We cover similarity metrics that are commonly used to detect similar field entries, and we present an extensive set of duplicate detection algorithms that can detect approximately duplicate records in a database. We also cover multiple techniques for improving the efficiency and scalability of approximate duplicate detection algorithms. We conclude with coverage of existing tools and with a brief discussion of the big open problems in the area duplicate record detection techniques are crucial for improving the quality of the extracted data. The increasing popularity of information extraction techniques is going to make this issue more prevalent in the future, highlighting the need to develop robust and scalable solutions. This only adds to the sentiment that more research is needed in the area of duplicate record detection and in the area of data cleaning and information quality in general.[6]

M. Karthigha and S. Krishna Anand invented A Survey on Removal of Duplicate Records in Database published on Indian Journal of Science and Technology. This paper presents a thorough analysis of similarity metrics to identify similar fields in records and a set of algorithms and duplicate detection tools to detect and remove the replicas from the database. In this paper a detailed analysis of subsisting methodologies, which were used for record matching and de-duplication are confronted. It has been estimated from the above techniques that de-duplication is a cumbersome process and requires both time and memory. An effective and efficient de-duplication algorithm, which requires minimum number of comparisons for records with less memory and time need to be developed in future.[7]

3. PRAPOSED ARCHITECTURE



.

Fig. 1 Proposed Architecture of Record Linkage.

The general steps involved in the linking of two databases. Because most real- world data are dirty and contain noisy, incomplete and incorrectly formatted information. In any record linkage or de-duplication project is data cleaning and standardization. The main task of 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. Indexing is the topic of this survey, in which the indexing step generates pairs of candidate records.

These records are compared in detail in the comparison step using a variety of comparison functions appropriate to the content of the record fields (attributes). Approximate string comparisons, which take (typographical) variations into account, are commonly used on fields that for example contain name and address details, while comparison functions specific for date, age, and numerical values are used for fields that contain such data. Several fields are normally compared for each record pair, resulting in a vector that contains the numerical similarity values calculated for that pair. The next step in the record linkage process is to classify the compared candidate record pairs into matches, non-matches, and possible matches, depending upon the decision model used. Record pairs that were removed in the indexing step are classified as non-matches without being compared explicitly. If record pairs are classified into possible matches, a clerical review process is required where these pairs are manually assessed and classified into matches or non matches. Measuring and evaluating the quality and complexity of a record linkage project is a final step in the record linkage process.

A. *Extraction*

Data extraction is the act or process of retrieving data out of (usually unstructured or poorly structured) data sources for further data processing or data storage (data migration). The import into the intermediate extracting system is thus usually followed by data transformation and possibly the addition of metadata prior to export to another stage in the data workflow. Usually, the term data extraction is applied when (experimental) data is first imported into a computer from primary sources, like measuring or recording devices. Today's electronic devices will usually present an electrical connector (e.g. USB) through which 'raw data' can be streamed into a personal computer. Typical unstructured data sources include web pages, emails, documents, PDFs, scanned text, mainframe reports, spool files etc. Extracting data from these unstructured sources has grown into a considerable technical challenge where as historically data extraction has had to deal with changes in physical hardware formats, the majority of current data extraction deals with extracting data from these unstructured data sources, and

from different software formats. This growing process of data extraction from the web is referred to as Web scraping.

B. *Data Cleaning and Standardisation*

The cleaning and standardization of a data set using is currently done separately from a linkage or de-duplication project. A data set can be cleaned and standardized and is written into a new data set, which in turn can then be de-duplicated or used for a linkage. When a user selects the Standardization project type, and has initialized a data set on the data page, she or he can define one or more components standardizes on the Standardize page. Currently, component standardizes are available for names, addresses, dates, and telephone numbers. The name standardize uses a rule-based approach for simple names (such as those made of one given- and one surname only) in combination with a probabilistic hidden Markov model (HMM) approach for more complex names (Churches et al. 2002), while address standardization is fully based on a HMM approach (Christen and Belacic 2005). Each standardize requires one or several input fields from the input data set (shown on the left side of a standardize in the GUI), and cleans and segments a component into a number of output fields.

C. *Indexing*

When two databases are linked, each record from one database potentially has to be compared with all records from the other database. The vast majority of these comparisons will be between records that are not matches (i.e. refer to different entities). Indexing is the process of reducing this possibly very large number of record pairs that need to be compared in detail between databases by splitting each database into smaller sets of blocks or clusters, or by sorting the databases. The aim is to identify candidate record pairs from records in the same blocks or clusters that likely correspond to true matches, and that need to be compared in detail, generally using approximate string comparison functions (Christen, 2012a).The traditional blocking approach employs a blocking criteria (a single or set of attributes) to insert each record into one block (Fellegi and Sunter, 1969).

D. *Classification Techniques*

Record linkage classification is to decide if a pair or group of records are a match or a non-match. In the traditional probabilistic record linkage approach, each compared record pair classified independently into one of three classes. The third class of possible matches are those pairs that require manual classification through a clerical review process. Besides requiring an often time consuming manual clerical review step, this traditional approach has several drawbacks. First, it assumes independence between attributes. Statisticians have investigated approaches that allow dependencies between some attributes to be modelled, and have achieved improved classification outcomes in some situations. Second, the estimation of the parameters needed for the probabilistic record linkage approach is a non-trivial undertaking and requires knowledge about the error rates in the databases to be linked.

Following are the two main classification techniques:-

1) *k-means algorithm*

The algorithm clusters observation into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works:

Step a : The algorithm arbitrarily selects k points as the initial cluster centers ("means").

Step b : Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.

Step c : Each cluster centre is recomputed as the average of the points in that cluster.

Step d : Steps b and c repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps b and c are repeated or that the changes do not make a material difference in the definition of the clusters.

2) *Support Vector Machine (SVM)*

SVM is a supervised learning model. This model is associated with a learning algorithm that analyses the data and identifies the pattern for classification. The concept of SVM algorithm is based on decision plane that defines decision boundaries. A decision plane separates group of instances having different class memberships. For example, consider an instance which belongs to either class Circle or Diamond. There is a separating line which defines a boundary. At the right side of boundary all instances are Circle and at the left side all instances are Diamond.

E. *Evaluation*

The evaluation process is used to calculate efficiency and complexity of the three classes that is matches, non-matches and possible matches.

4. CONCLUSIONS

An analysis of the existing record de-duplication techniques and frameworks is done here. De-duplication and record linkage is a crucial step in data integration. From this survey, it is possible to conclude that the existing algorithms require more memory for de-duplication. It is also time consuming

process. In future a de-duplication algorithm can be designed for reducing the number of comparison between the records such that it reduces time consumption and utilises less memory space.

REFERENCES

[1] Peter Christen "A Survey of Indexing Techniques for Scalable Record Linkage De-duplication" IEEE Transaction on Knowledge and Data Engineering,VOL. Z, NO. Y, ZZZZ 2011.

[2] Nishand. K Ramasami. S and record linkage system and de-duplication using indexing technique, classification and FEBRL framework" International Journal of Emerging Science and Engineering (IJESE) ISSN:2319-6378,Volume-1,Issue-7 in the year May-2013.

[3] Lalitha. Maheswari.B2, Karthik.S3 "A Detailed Survey on Various Record De-duplication Methods'' International Journal of Advanced Research in Computer Engineering &Technology (IJARCET) Volume 1, Issue 8,in the year October 2012.

[4] Peter Christen "Advanced record linkage method and privacy aspects for population reconstruction" Research School of Computer science. The Australian National University Canberra ACT0200, Australia.

[5] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis and Vassilios S. Verykios "A survey on Duplicate Detection Algorithm" IEEE transactions on Knowledge and Data Engineering,VOL. 19, NO.1 in the year January 2007.

[6] M. Karthigha and S. Krishna Anand "A Survey on Removal of Duplicate Records in Database" IJOST.

[7] Sunita Yeddula & K. Lakshmaiah "Investigation of Techniques For Efficient & Accurate Indexing For Scalable Record Linkage & De-duplication'' Dept of CE, Madanapalle Institute of Technology and Science, Madanapalle.