A Survey on Record Linkage

[1]A K Kousthubha, [2]Dr. K Raghuveer

[1]PG Student, [2]Professor & Head
Department of ISE,
The National Institute of Engineering, Mysuru, India

***Abstract*: Record linkage is the process of identifying similar records present in same or different databases. It removes duplicate records from the database. This paper describes about different approaches followed for record linkage, the algorithms that can be applied for record linkage, it also summarizes the applications and advantages of record linkage.**

***Keywords*- Record linkage, Probabilistic, Deterministic, Token based algorithm, Edit based algorithm, Machine learning**

I. INTRODUCTION

A lot of data is generated and stored in databases. These data is heterogeneous in nature. The process of integrating data i.e. matching of records which are present in different databases is known as record. These record represent the same real world entity but they might have different attributes. The nomenclature of the record names may differ from one database to another.

The record linkage problem can be addressed by deterministic approach or probabilistic approach. Deterministic approach is applies for one-to-one comparison between records and probabilistic methods involve the calculation of linkage weights estimated given all the observed agreements and disagreements of the data values of the matching variable [1].

Record linkage can be applied across fields like data warehousing, data mining , knowledge discovery etc.

II. RECORD LINKAGE APPROACHES

- **Deterministic approach**

It is the simplest method of matching records from different databases. Deterministic approach can be used for attributes which have attributes which are unique. For example it can be used for matching SSN (social security number) or Unique identification number. The drawback is the data should be present in all the databases to be compared and the data should be exactly same as the data present in the comparing database [1]. The deterministic approach can't deal with data with erroneous attributes and missing values.

- **Probabilistic approach**

In probabilistic-based approach data is trained to compute a maximum likelihood estimate which determines whether a record pair is matching or not. The unsupervised Expectation Maximization (EM) algorithm is applied instead of using the train data [2]. The source records and the comparable records are merged and the record pairs are generated. Each matching records pairs are compared and a score is calculated for each pair. Based on the threshold score it is decided if the pairs exact match or not. If the score is above threshold it is exact match , if the score is below threshold it is not the match[1].

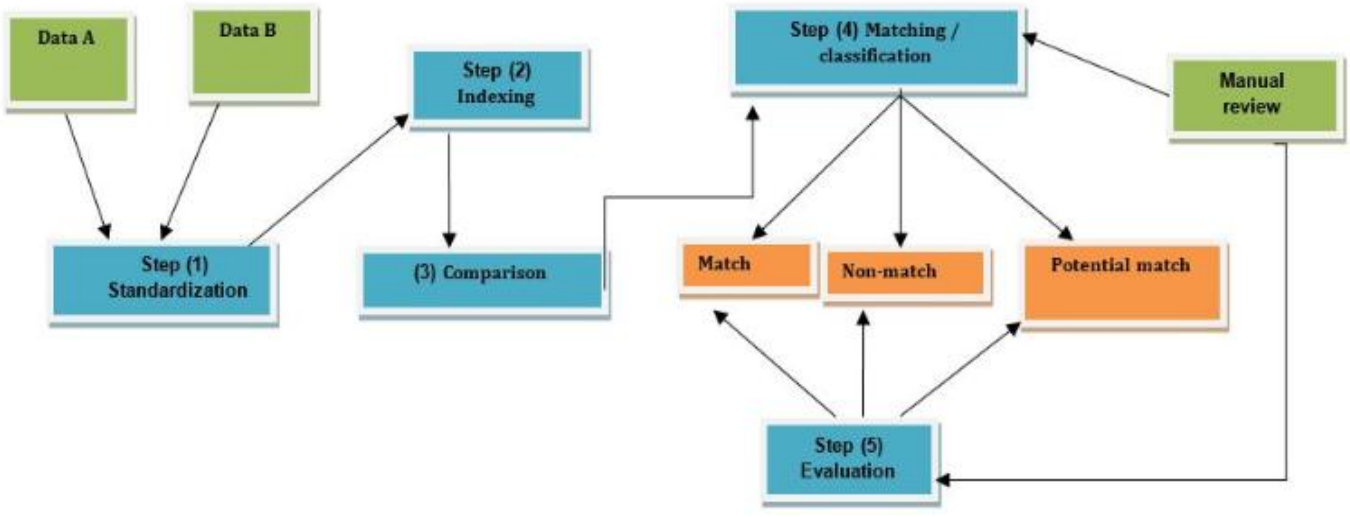

Figure 1: Probabilistic approach of Record linkage

- **Machine Learning**

Record Linkage can be considered as a classification problem .For linking data and deduplicating data machine learning methods like clustering methods, decision trees and support vector machines can be applied [3].

III. ALGORITHMS FOR RECORD LINKAGE

There are algorithms which are token based used for record linkage like Tf-idf-Cosine similarity and Jaccard Coefficient [4].

- **TF-IDF Cosine similarity**

The records/the strings are tokenized and they are converted to vectors. Term frequency(TF) and inverse document frequency(IDF) values for each record is generated. The TF*IDF values for each record is calculated. Then the cosine similarity score is calculated for the comparable record pairs. The pair with highest similarity score is the exact match [5].

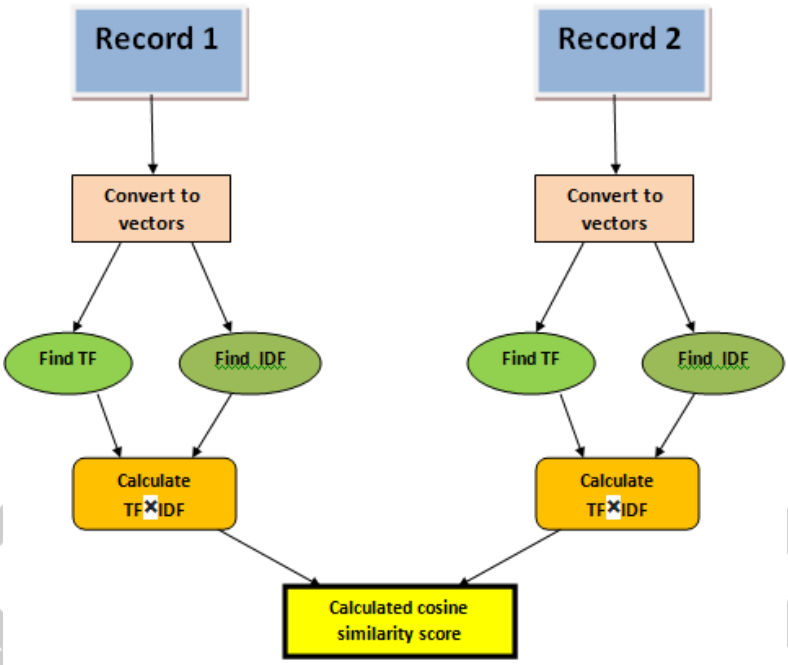

Figure 2: Record linkage using cosine similarity

- **Jaccard coefficient**

The similarity between words is determined accurately than the similarity between letters of the word in Jaccard coefficient. This algorithm is highly stable as it can handle misspelled words.
The formula to find the Index is [6]:

Jaccard Index = (the number in both sets) / (the number in either set) * 100

The same formula in notation is:

$$J(X,Y) = |X \cap Y| / |XUY|$$

There are several edit based algorithms for finding the match for the record like soundex, Jaro-Winkler [4].

- **Soundex algorithm**

In order to match records which sounds alike we use phonetic matching system. Soundex encoding values are assigned to the words in such a way that the words which sound the same have same values
For example consider names like Stefan, Steph, Steve, Stephen, Steven, Stove and Stuffin. Now we need to search the names which sounds same as Stephen[7].
The names which matches with Stephen are positives and the ones which doesn't match or rejected by the soundex search engine are the negatives[7].

- **Jaro-Winkler Algorithm**

Jaro-Winkler is string metrics used to determine the edit distance between two strings. The Jaro distance or edit distance is nothing but the number to transformations made to convert one string to another [8].

The Jaro-Winkler metric is widely used for matching census records. It can be applied to match short strings from equivalent database by comparing their edit distances.

IV. PHASES OF RECORD LINKAGE

- **Pre-processing**

The data retrieved from the database will not be in a standardized form. The two record to be compared may be in different syntax or formats. The records may contains error and they may be misspelled. Thus, records need to be standardized . In order to standardize the data, the records must undergo pre-processing. Some of the pre-processing techniques are conversion to lower case, stopword removal special character removal.

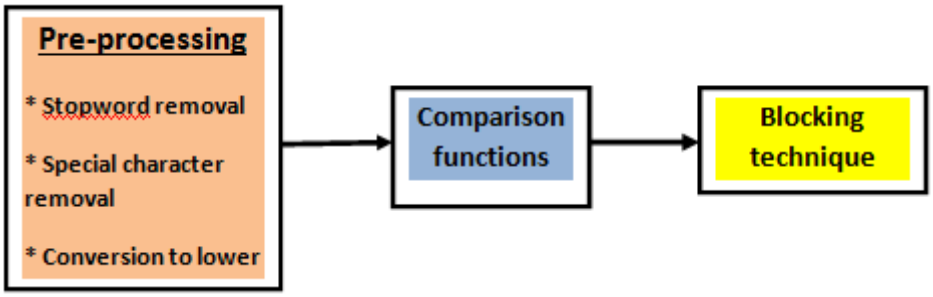

Figure 3: Phases of Record Linkage

- **Comparison Function**

The records can be compared based to several similarity measures like sequence ,set, phonetic[9]. The algorithms like edit distance or Jaro-Winkler can be chosen to express the distance between pairs of words. Cosine similarity algorithm can be applied to find the similarity between records with long descriptions.

- **Blocking Technique**

In order to reduce the number of record comparison pairs to a optimal number blocking technique is used. Blocking methods partition the data sets into blocks or clusters of records which share a blocking attribute or are otherwise similar with respect to a defined criterion [10].It reduces the memory as well as reduces the compilation time.

V. APPLICATION OF RECORD LINKAGE

- **Pharmaceutical databases**

The record linkage method can be applied to pharmaceutical databases in order to identify unique patients records present in different hospital databases.
The content of the drug exposure files is discussed, as well as methods of linking to other sources, including cancer registries, hospital databases, clinical laboratories, and birth registries [11].

- **E-Commerce databases**

Record linkage can be applied in ecommerce databases to match similar products from different companies.

- **Bank databases**

Bank data base contains detail of customers like name and addresses; if a bank wants to retrieve details of retail customers from mortgage database record linkage can be used [12].

VI . ADVANTAGES OF RECORD LINKAGE

- **Completeness**

Record linkage avoids missing attributes in the database or missing data in the database. Misspell information is handled while pre-processing. Insufficiency of information in the financial databases may lead to serious consequences, this can be avoided by record linkage [13].

- **Comparability**

In order to facilitate data's use in exploratory analysis, modeling and statistical estimation several databases need to be combined to a data warehouse. Record linkage is the process of combining data from different sources/databases.

- **Reduces Computation time**

Large amount of unclean data is present in database, the data needs to be cleansed before computation. As data is cleansed during pre-processing. The cleansing time is saved.

- **Scalability**

As large amount of data is being generated every day it is difficult of scale data.

VII . CONCLUSION

Deterministic approach works fine only if it finds and exact match i.e. deterministic approach can be applied for only high quality data. Probabilistic approach is better than deterministic approach as it can process the unclean data with missing values.As Jaro-Winkler algorithm gives the edit distances it can be applied for comparing words. Cosine similarity algorithm can be applied for comparing long strings.

Record linkage improves the quality of data, reduces the memory usage and the data retrieval time is reduced.

REFERENCES

[1] "An Introduction to Probabilistic Record Linkage", John 'Mac' McDonald,Centre for Longitudinal Studies

[2] "A Comparative Study of Record Matching Algorithms", Shirley Ong Ai Pei (276905).

[3] "The RecordLinkage Package: Detecting Errors in Data", Murat Sariyar and Andreas Borg

[4] "Record Linkage:Similarity Measures and Algorithms", Nick Koudas (University of Toronto), Sunita Sarawagi (IIT Bombay), Divesh Srivastava (AT&T Labs-Research)

[5] "A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method", HUANG Cheng-Hui1),2) YIN Jian1) HOU Fang2)

[6] "Using of Jaccard Coefficient for Keywords Similarity", Suphakit Niwattanakul*, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu

[7] " Phonetic Matching: A Better Soundex", Alexander Beider, Stephen P. Morse

[8] "Overview of Record Linkage and Current Research Directions", William E. Winkler

[9] "A Survey of Data Quality Issues inCooperative Information Systems", Carlo Batini, Tiziana Catarci, Monica Scannapieco, 23rd International Conferenceon Conceptual Modelling (ER 2004) and "Searching andIntegrating Information on the Web, Seminar 3", Chen Li,Univ. Irvine

[10] "A Comparison of Fast Blocking Methods for Record Linkage", Rohan Baxter, Peter Christen, Tim Churches

[11] "Pharmacy-Based Medical Record Linkage Systems", Ron M C Herings

[12] "Record linkage",Thomas H. Herzog,1 Fritz Scheuren2 and William E. Winkler3∗

[13] "Data Quality and Record Linkage Techniques", Thomas N. Herzog, Fritz J. Scheuren, William E. Winkler,Lars Pedersen