An Efficient Record Linkage Technique for Handling BIG DATA

Sneha Ambhore
Inurture Education Solutions
Banglore, Karnataka

Shailvi Maurya
Inurture Education Solutions
Banglore, Karnataka

ABSTRACT
The word BIG DATA is nothing but huge amount of data generated through all the sources including social networking sites like facebook, twitter, Intstagram etc. this data sometimes may be repetitive that is same person can have record in more than one databases, whereas it is belonging to a single person so those records should be merged. Also sometimes a situation may occur where you need entire history of a person in this case record linkage will make it possible. Many Researches has been done for efficiently linking the records as record linking is becoming important day-by-day since it increases the quality of the data. In this research we are going to focus on algorithm for efficiently linking the records and keeping records secure. The software called FEBRL is used for comparing our algorithm efficiency with previously defined algorithms.

General Terms
Algorithm,Complexity,Record Linkage,Indexing

Keywords
Efficient,Big data,Records,Clustering

1. INTRODUCTION
Matching records belonging to same entity is becoming important in the aspect of increasing data quality. Matching records is a task of identifying information that belongs to same entity from different databases. The record linkage technique is not a single step process, it requires data to

be complete but data in databases may sometimes be incomplete or error prone and thus linking cannot be done on such data thus data should be standardized before linking. In Record linking data from different databases is first standardize that is cleaned or corrected and then blocking is being applied on it and then the records are being compared and with the help of classifier it is being checked if records belonging from different databases are similar or dissimilar.

Finding blocking key value to check for the similar attribute between two tables and on the basis of that blocking key value get the count of similar attribute with the help of count and then based on the maximum count place the data with similar attributes from that particular datasets into a single block and check for similarity if it exists between two different datasets and if it do exists then will go for linking them.

The proposed work is about to have a technique to efficiently partition the datasets and based on similarity link the record of an individual if exists in two different datasets also assigning security labels to it so as to restrict the access of records to authenticated persons

2. LIERATURE SURVEY
As discussed earlier a survey on various indexing techniques was been carried out in year 2012 . The survey was been done on total of 6 indexing techniques and all the pros and cons of each of those 12 techniques was been discussed in[1].also the complexity as well as their performance and its scalability was being carried out with the help of the experimental framework on real and synthetic datasets as well. The survey described the drawback in those techniques with the solution on those drawback which when applied can improve the performance of the algorithms.

A new technique proposed in[2] tells us that relationship between the attributes of the different entities are very important when there are missing values as they provide accuracy. The missing values of dataset can be considered as a big problem since the data referred if are wrong the further process done on it will also go wrong which in turn will waste time as well as memory.

A clustering algorithm for record linkage was proposed in[3].The algorithm proposed here was to show how the clusters formed out of unloadable datasets. It is possible to handle The huge amount of data by clusters which was unmanagable normally with other existing techniques. first the representation of sample data is then the visualization is done for the same. The experiment carried out here shows that it is fast enough for both data partitioning and data visualization. The proposed algorithm shows the visual evidence of how big the cluster can be thus if the cluster formed out of single link clustering varies from the visual cluster image then the cluster formed can be discarded.

The efficient record linkage technique was being implemented in[4] which was based on multidimensional Euclidian space. The approach of this technique is to first take two records and merge them together and then map its value to multidimensional eculidian space. mapping the algorithm which was being used here was fast mapping technique. The reason lexographic reordering been used here is that the it is been assumed the strings with similar substring will be placed nearby when arranged in lexographic order. A similarity join is carried out on similar attribute to determine if there exist any similar record. The algorithm first takes two list of records and merge them and then based on lexographic technique sorting is carried out on it..

The technique for handling specially census data that is administrators data was been proposed in [5] where the encryption of identifiers for privarcy preserving was been considered with a novel blocking data approach . the idea proposed here was the Q-gram technique where the identifiers were converted from privarcy preserving to non-privarcy preserving so that it can find neighbours in binary space applied for finding neighbourhood data on normal unencrypted data. when it comes to linkage of census data the time required is more but the limitations from institution are link it in some particular time in such case the time should not exceed than some hours.

An overview of the techniques available so far now were being discussed in a paper[6]. Where the three problems that occurs while record linking that are being considered are being discussed here, also the very important aspect among them are confidentiality and the privacy of data. The aspects of linking technique the evaluation all such major problems which come across the record linking are being discussed here. also the recently developed taxonomy are being discussed here.

A comparative analysis of two blocking techniques were being proposed in this paper[7] where the recently introduced two techniques were being compared namely Canopy clustering and Bigram indexing. The canopy clustering was been evaluated

In [8]paper summarization of Algorithm three methods for managing large volumes of data efficiently with respect to online record linkage is being introduced.which are as follows BlockSketch,skip bloom,$blockSketch

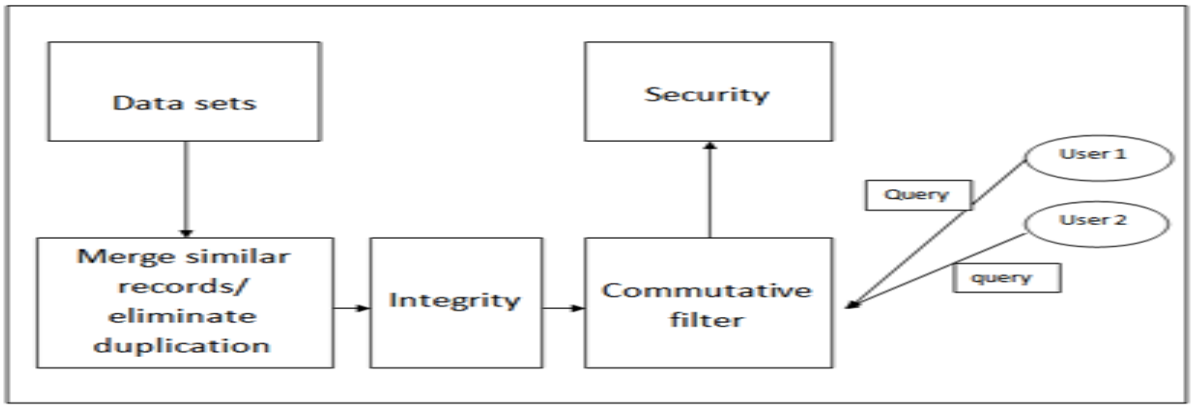

Fig 1: Architecture of the system

3. PROPOSED SYSTEM

The aim of this technique is to efficiently link the records with security. record linking can come across a problem where sensitive data should be proctected from unauthorized access/modification reason behind this is after linking the records it may happen that an unpriviledged user is accessing the most sensitive data, to avoid this sensitivity is needed on records after linking.

The data from different databases are firstly standardized and then blocking is done on them with the help of merge sort and after that the records are being compared to other records to see if it can be merged or linked with some other record. After merging the records sensitivity is being assigned to the merged record for security reasons and before giving the access to the for any record the accessibility check will be done and then depending on the record sensitivity and the users access right the decision will be taken to give the access to a particular record or not.

Let "T" denote the database

The record linking will consider the following before linking

If The databases have some attributes in common then we can link the records if they are matching

$T1 \cap T2 \cap T3 \cap ……..\cap Tn \neq \emptyset$

Else

$T1 \cap T2 \cap T3 \cap T4 \cap …..Tn\text{-}1 \cap Tn$

T1T2 T3T4……T……Tn-1Tn

T1T2T3T4……..Tn-1Tn

Where provided

$T1 \cap T2 \neq \emptyset$ Similarly

$Tn\text{-}1 \cap Tn \neq \emptyset$

Also

$T1 \cap T2 \subseteq$ key

ALGORITHM FOR LINKING RECORDS

Let D denote Datasets

D1,D2,D3,D4…….Dn

Count1,count2,count3 as per the number of comparison between datasets

Threshold for COUNT=4

Step1 :-compare D1 with all datasets until Dn for existence of similar column

D1D2,D1D3,D1D4,D1D5….D1Dn

(PASS1)

Loop until total number of datasets i.e Dn for similar attribute.

If (D1.colname=D2.colname)

for every comparison increment the count(i) value

else go for next comparison (D1.colname =D3.colname)

step 2:- if the count value of D1 and D3 >= COUNT then exclude those two datasets for next comparison.

Else will be used for comparison for next upcoming passes

Step 3:- after every pass compare for greatest count value and check if it ie Equal or greator than COUNT then those two record will be passed for comparison of similar records.

i.e D1D3,D1D7 etc

step 4:- if two datasets D1D3 has count value greator than COUNT and

D1D5 have count value equal to COUNT

D1D3.count > COUNT AND D1D5.coun t= COUNT

then the one with the highest count will be considered for similarity checking

step 5:- if the number of datasets are even then the comparison will be valid

D1D3,D2D4,……Dn-1,Dn

Else the one leftout will be forwarded two next round after merging to see if still it has any column in common or can be linked

Step 6:- apply sensitivity on records after linking

ALGORITHM FOR CHECKING ACCESS RIGHTS

Step1:- The user will login and ask query requesting data

User req q1.data

Step2:- Depending upon the sensitivity of record and access label of record and that of user is compared

if (users.acces rights >= records)

permission granted record provided

Else

Permission denied

4. RESULT & ANALYSIS

Table 1: Medical Data Set

Medical datasets		
Hospitals	**Attributes**	**No. of Records**
AIMS	Name, Age, BloodGroup,City	10000
CMC	Name , Address, city,age,sex	5000
VIT healthcare	FullName, Age, BloodGroup,City,coconut	6000
Vasan EYE care	Name, City, Age, Address	90000

Table 2: College Data set

College Datasets		
VIT	Id, Name, age, Branch, Degree ,Block	200000
SRM	Name, Branch, Year, City	50000
Anna University	Id, Age, City, Degree	80000

Table 3: Similar Records

	College	Hospitals
Max Attributes	6	5
Similar Attributes	3	4
Records	1000000	5000000
Similar Records	1000	2000

5. CONCLUSION

The paper aim is to have an efficient record linking technique which can link the records if the similar record exists so as to increase the quality of data. The similar records are compared on the basis of the attributes present in the different datasets, depending upon those attributes. The above paper is also considering the privacy of the record by assigning the security tags to the records after linking, so that the data of a record which is sensitive can be accessible to the unauthorized person.

6. REFERENCES

[1] peter christen,"A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication"IEEE Transaction on Knowledge and Data Engineering,vol 24,No.9,September 2012

[2] Gayan Prasad Hettiarachchi,Dhammika Suresh Hettiarachchi,Nadeeka Nilmini Hettiarachchi,Azusa Ebisuya,"Next Generation Data Classification and Linkage", Osaka University of Tokyo,Japan

[3] Timothy C Heavens[1],James C. Bezdek[2],Marimuthu Palsniwami[2], "Scalable Single Linkage Heirarchical Clustering for BigData", University of Melbourne,USA,Australia

[4] Liang Jin, Chen Li,Sharad Mehrotra," Efficient Record Linkage in Large Datasets", University of California,Irvine

[5] Rainer Schnell,"An Efficient Privacy Preserving Record Linkage Technique for Administrative data and census",Statistical Journal of (IAOS),2014

[6] Peter Christen, "Overview and taxonomy of Techniques for Privacy-preserving Record Linkage " (JSM) Joint Statistical Meeting, August 2013

[7] Rohan Baxtor ,peter christen, Tim churches, " A Comparison for Fast Blocking methods for Record Linkage", Australian National University, Australia

[8] Dimitrios Karapiperis,Aris Gkoulalas-Divanis,Vassilios S.Verykios,"Summarization Algorithms for Record Linkage" Hellenis Open University Patras Greece 2005