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 Design Paper on Indexing Technique for Medical Record Linkage

Prachiti Deshpande[1]; Shraddha Atalkar[2]; Deepa Jagtap[3]& Prof. V.G.Bhujade[4]
Department of Computer Engineering, Bapurao Deshmukh College of Engineering, Sevagram, Wardha.

[1]prachiti_01@rediffmail.com
[2]shraddhaatalkar@gmail.com
[3]jagataparadhana@gmail.com
[4]vaishali.bhujade@rediffmail.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 which we are using in medical domain*.

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. PRAPOSED SYSTEM

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.

PROGRESS OF PROPOSED WORK

The application is under development we are developing this application for the medical domain.Until now we have develop 30% of total work. We have different modules which are:

There are seven main modules , those are :-

- ❑ To design GUI.
- ❑ Database Creation.
- ❑ Cleaning and Standardisation.
- ❑ To implement Sorted Neighborhood Algorithm.
- ❑ To implement Indexing Technique.
- ❑ Classification
- ❑ Evaluation

Out of which we made Login as Admin and Database of Patient.

As we open the application we seen a home page contains various tags in it. As we click on the Admin tag the window appears containing user-id and password field. We have introduce a Admin who has all the rights to add the records of patient. As seen in the bellow screenshots.

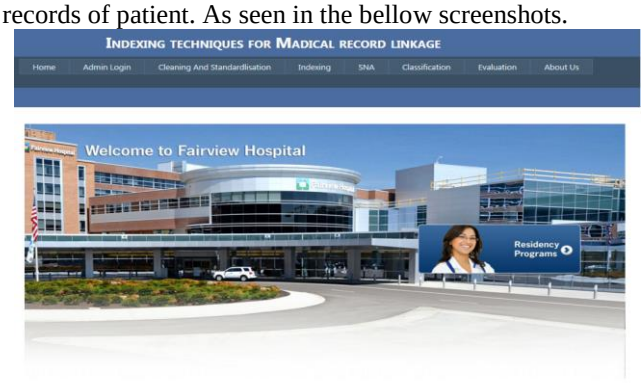


Fig. 1 Home Page

The Fig. 1 shows the snap shot of the Home Page of Application. Home Page contains the various tags like Admin Login, Cleaning and Standardization, Indexing, SNA, Classification and Evaluation.

Log-in as Admin -

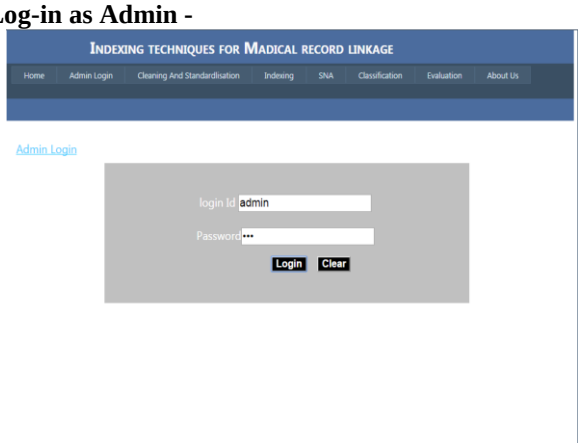


Fig. 2 Admin Log-in.

The Fig. 2 shows the snapshot of admin log-in page. The Admin has the authority to enters the patient record in the database also he has the access to the saved patient records too.

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.

Entering Patient Record In Database A -

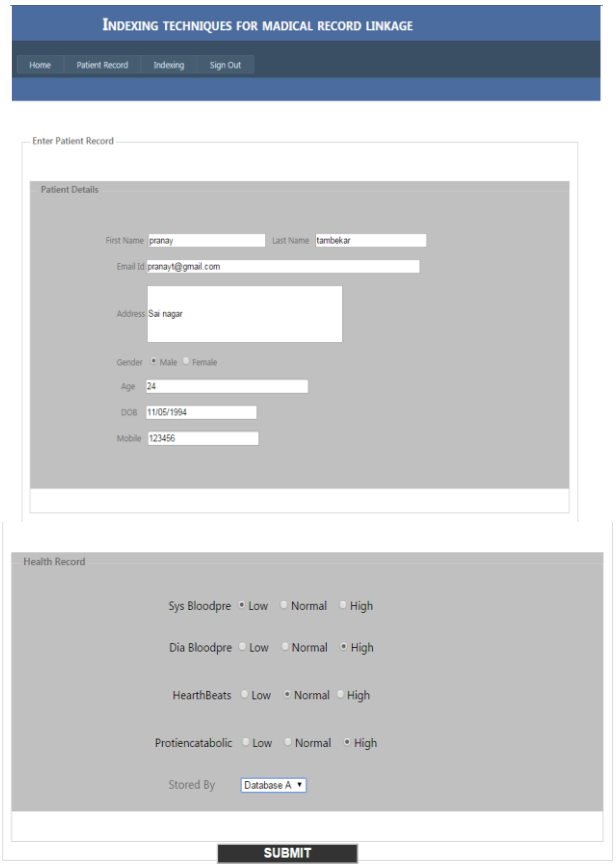
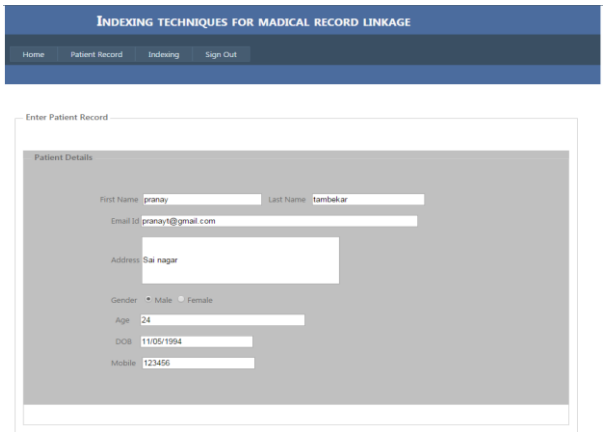


Fig. 3.1 Patient Registration (Database A).

Entering Patient Record In Database B -



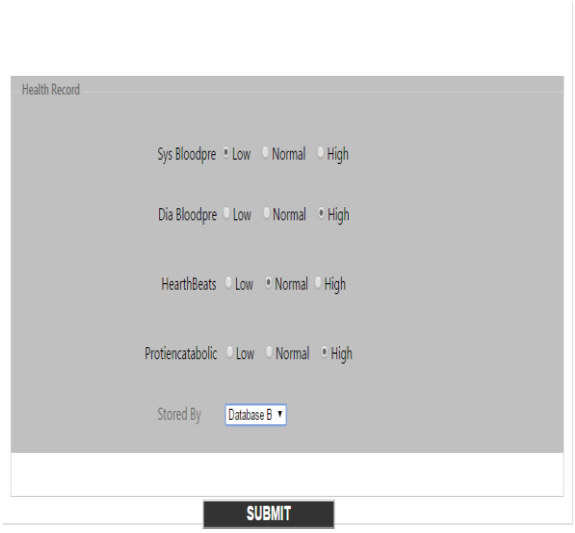


Fig. 3.2 Patient Registration (Database B).

The Fig. 3 shows the snapshot of registration of patient. The admin registers the patient and saves the information in one of two database i.e. in Database A or in Database B.

The admin has also the authority to search any particular record from both the database. Following figures shows records stored in Database A and Database B.

Searching Records -

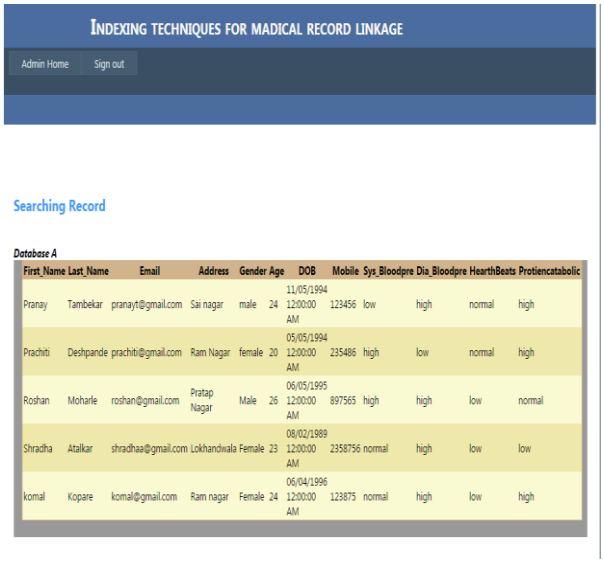


Fig. 4.1 Patient Records (Database A).

Database B

First_Name	Last_Name	Email	Address	Gender	Age	DOB	Mobile	Sys_Bloodpre	Dia_Bloodpre	HearthBeats	Protiencatabolic
Pranay	Tambekar	pranayt@gmail.com	Sai nagar	male	24	11/05/1994 12:00:00 AM	123456	low	high	normal	high
Prachiti	Deshpande	prachiti@gmail.com	Ram Nagar	female	20	05/05/1994 12:00:00 AM	235486	high	low	normal	high
Roshan	Moharle	roshan@gmail.com	Pratap Nagar	Male	26	06/05/1995 12:00:00 AM	897565	high	high	low	normal
Shradha	Atalkar	shradhaa@gmail.com	Lokhandwala	Female	23	08/02/1989 12:00:00 AM	2358756	normal	high	low	low
komal	Kopare	komal@gmail.com	Ram nagar	Female	24	06/04/1996 12:00:00 AM	123875	normal	high	low	high

Fig. 3.1 Patient Records (Database B).

3. CONCLUSIONS

As the above shows we have completed almost 30% of our work. We are going to work on the indexing techniques for de-duplication of the records also we are going to compare these techniques too.

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.