MJRET
Open Access
Multidisciplinary Journal of Research in Engineering and Technology
Journal homepage: www.mjret.in
ISSN:2348 - 6953

SCALABLE RECORD LINKAGE AND DEDUPLICATION FOR ELECTION DATABASE

[1]S.B.Kadus, [2]H.A.Sawant, [3]S.S.Tilekar, [4]H.D.Zendage

Department of Computer Engineering, University of Pune
Imperial College Of Engineering and Research
Pune, India
[1]kadussonali@gmail.com, [2]harshalisawant11@gmail.com,
[3]swaralitilekar@rediffmail.com,[4]hrushiz111@gmail.com

Abstract: *Record linkage is the technique of finding same data from collection of database that has same contents. This technique is when tested on single database that is known as de-duplication. Matching data is very much important in so many fields, because they have so many valuable information which is not possible to acquire from anywhere or it may be too costly to get. In data cleaning method removing same or duplicate data is very much essential step which applied on singe database because, in data mining same data can badly affect result of that process. With rising nature of today's database, in record linkage and de-duplication the difficulties of matching data become one of the challenges. For record linkage and de-duplication there are various indexing techniques have been developed in last few years. Their purpose is to reduce the number records by comparing matching records with non-matching one, at the same time must maintain high matching quality. This paper presents a system in which windowing Technique is applied on database, in which records are matched, their complication is analyzed and their work and flexibility is calculated using both fake and real data set.*

Keywords *- data linkage**,** blocking data, matching entity resolution**,** windowing techniques, scalability.*

1. INTRODUCTION

Record linkage [6] refers to the task of finding records in a data set that refers to the same entity across different data sources (e.g., data files, books, databases). As many businesses and research projects collect huge amount of data, techniques and the methods that allow efficient processing and analyzing of such massive databases have in recent years attracted interests from both industries and academia. Information from many sources needs to be integrated in order to improve quality of data, to enrich data that provides the facilities for

more detailed data analysis. Record linkage has the aim of providing the best possible match given with the available information.

1.1 The Record Linkage Process [6]

Most real-world data are dirty and contain noisy, incorrect and incomplete information; a first step in record linkage project is data cleaning [4] and standardization [5] because lack of good quality of data can be one of the biggest obstacles in the linking of two databases. The main task of data cleaning and standardization is to convert the raw input data into well defined, proper mannered and into consistent forms.

The second step is the 'indexing' [7]. The indexing step generates the pairs of candidate records which are compared in detail using comparison functions which are appropriate to the content of the record fields. Approximate string comparisons, which take variations, for example that contains name and address details. Several fields that are normally compared for each record pair, which results in a vector that contains the numerical similar values which are calculated for that pair. The third step in the record linkage process is to classify compared candidate record pairs into matches, possible matches, and non-matches. Record pairs are removed in the indexing steps which are classified as non-matches. Estimating and measuring the complexity and quality of a record linkage [6] project is a final step in the record linkage process.

2. LITERATURE SURVEY

2.1 Pattern search methods for Suffix arrays

Here the system consists of compact number of disk blocks for indexing using suffix arrays [7]. Here the disk space is reduced because the methodology is based on prefix and suffix. Though this search patterns were compact, a major disadvantage was these methods were not cost effective. Also the FM-INDEX method [5] has drawback that it requires more memory to place the whole structure.

2.2 Sorted Neighborhood Techniques for Record Linkage

These techniques are widely used for matching records [2] in a large database. This is a classical method which is easy to implement and has running time O(wn).The major steps were:

1. Create key *k* for each record.
2. Sort all the records using k.
3. Fix a window size for matching a record sequentially with the previous one and merge the records.

2.3 Use of Blocking [1] and Windowing [3] methods for Detecting and Removing Duplicates

Here the Windowing [3] and Blocking is compared. The Sorted blocks [1] method is better than both of them. Here the partition size is changeable but it requires right configuration settings. The number of comparisons depends on the number of related values.

2.4 AA-SNM method for record pair comparison

This is another method which is used for hastily and correctly finding boundary record pairs in variable sized blocks. It automatically performs window enlargement [4] and retrenchment and has varying error rate of duplicates.

Suffix array method [7] has limitation of more spare requirements as we need more space to store suffixes related to each record during process of de duplication [2], so in real world application large amount of processing space will be required to process this algorithm.

3. PROPOSED SYSTEM

In our system we will use IA-SNM method for data de duplication [2]. In this method we will compare each record with next record until we get similarity between them and we will use window which will be expanded when records are similar else shrunk if no similar records are there. We will compare records like name of person to find similarity between records. We will consider 80% similarity between names as there is a possibility of spelling mistake during data entry in database.

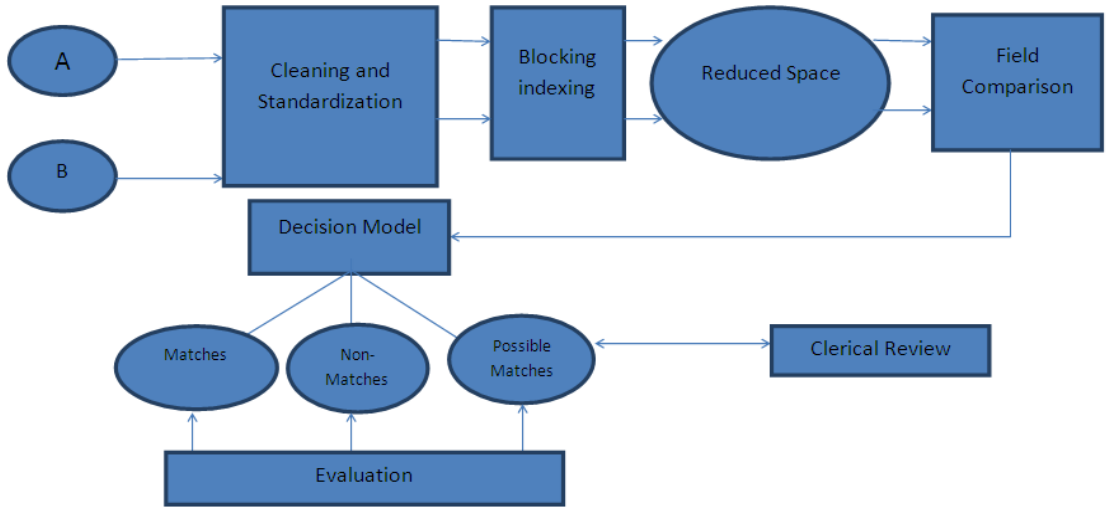

Fig.1.System architecture

3.1 User classes and characteristics

1. **Voter:** In democracy, all adult citizens i.e. age>18 must be eligible to participate in elections. Eligibility rules focus on age, citizenship and residence. Exact requirements also include name, address, mobile number, date-of-birth, PAN/AADHAR card number, address proof, email id. Voters should know how to register, login, update details and vote.

2. **Government Database:** The government database contains voter IDs and voter's details, voting regions. It is possible to create, update and delete entries and get information from this database.

3. **Standardization:** It will contain voter details, Index for each entry and validation rules. The actual cleaning and validation will be done in this class.

4. **IASNM:** This class will be implemented using window size, cache (logical), keyword to be matched and index. It gets record set, set, reset or increment window size, performs record pair comparison [2], slide window and adds keyword to cache.

5. **Evaluation:** It will store Index, matches, possible matches and non-matched record pairs. It includes windowing, comparison, similarity vector classification [5] and clerical review.

6. **Admin:** Admin has name, address, email and unique username and password. He will be responsible to maintain data, control system, check information and check de duplication [2], login and logout.

7. **Processed database:** It will have index, voter details and matches information. Here we can save, highlight matches, remove duplicates, update changes, add new index, retrieve records and add duplication information [2].

4. ADVANTAGES

a. De duplication solves the issues by locating identical content and handling it appropriately.
b. The use of de duplication using windowing algorithm results in more efficient use of both memory cache and storage capacity.
    i. It is most effective in applications where many copies of very similar or even identical data are stored on single disk.
    ii. It reduces the amount of storage needed for given set of files.
    iii. De duplication ultimately reduces redundancy.

5. CONCLUSION

As record linkage [6] is being used immensely in industries, a large number of algorithms are developed and analyzed successfully. Presented solutions have a set of constraints whose values are manually set by professional and are fixed at the time of execution. We are going to use the blocking [1] and windowing algorithms along with adjustable window size. So it provides more accuracy and efficiency for record linkage. This solution will adaptively and dynamically changes constraints of record linkage algorithms at the time of execution. The speed of performing actions on records will be improved using these new methods.

ACKNOWLEDGMENT

We would like to sincerely thank Mr. Gadekar, our guide, for his support and encouragement**.**

REFERENCES

[1]. Tim Churches, Rohan Baxter and Peter Christen, In SIGKDD Workshop on Record Linkage ,Data Cleaning, and Object Consolidation, 2003 for A comparison of fast blocking methods for record linkage.

[2]. Karl Goiser and Peter Christen .Complexity and Quality measures for de duplication and data linkage. In Springer, 2007 volume 43 of Studies in Computational Intelligence and In Quality Measure as well as in Data Mining,

[3]. Salvatore J. Stolfo and Mauricio A. Hern´andez .Data Mining and Knowledge Discovery, .Real-world data is dirty: Data cleansing and the merge/purge problem2 (1), 1998

[4]. Charles Elkan and Alvaro E. Monge .In Proceedings of the Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997. For detecting approximately duplicate database records an efficient domain-independent algorithm.

[5]. W. E.Winkler, "Methods for creating and evaluating data quality," Elsevier Information Systems, no. 7, vol. 29, pp. 531–550, 2004.

[6]. D. E. Clark, "Practical introduction to record linkage for injury Research," Injury Prevention, vol-10, pp-186–191, 2004.

[7]. C. W. Kelman,J. Bass and D. Holman , "Research use of linked Health data – A best practice protocol" Oust NZ Journal of Public Health, vol-26, pp-251–255, 2002.