ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

A Survey of Sorted Neighbourhood Indexing Technique for DeDuplication

Rupali Vairagade, Shantanu Surve, Karan Naik, Ganesh Sonawane, Jayant Athawale

Assistant Professor, Dept. of Computer, SITS, Savitribai Phule Pune University, Pune, India

Student, Dept. of Computer, SITS, Savitribai Phule Pune University, Pune, India

Student, Dept. of Computer, SITS, Savitribai Phule Pune University, Pune, India

Student, Dept. of Computer, SITS, Savitribai Phule Pune University, Pune, India

Student, Dept. of Computer, SITS, Savitribai Phule Pune University, Pune, India

ABSTRACT:The process of matching records which are from several databases and which refer to the same entities is known as Record Linkage. When you apply this process on a single database, this process is known as de-duplication. Matched data are becoming an important in many application areas, because they can contain information that is not available, or it is too costly to acquire. Thus, removing duplicate records from a single database is an important step in the data cleaning process as duplicates can severely affect the outcomes of any subsequent data processing or data mining. Because of the constant increase in the size of today's databases, the difficulty of the matching process has become one of the major issue for record linkage and de-duplication. Recently, various indexing methods have been developed for record linkage and de-duplication. Thus, by removing the obvious non-matching pairs these methods have been aimed at reducing the number of record pairs to be compared during the matching process. This paper presents the survey of sorted neighbourhood indexing technique.

KEYWORDS: Record Linkage; Blocking Key Value(BKV); De-duplication; Similarity Vector Classification(SVC); Indexing technique

I. INTRODUCTION

In the recent years, techniques that allow efficient processing, analysing and mining have attracted the interest from academia and industry as many businesses, government agencies and research projects collect very large amounts of data. The matching of records that relate to the same entities from many databases is becoming a task of increasing importance.

In orderto improve data quality, or to enrich data to facilitate more detailed data analysis, information taken from different sources needs to be combined. The records that have to be matched frequently correspond to entities which refer to people, who can be clients or customers, patients, taxpayers, students, employers or travelers. Record linkage is now commonly used for improving data quality and integrity, to allow re-use of existing data sources for new studies, and to reduce costs and efforts in data acquisition. For example, in the health sector,matched data can contain information that is needed to improve health policies. Linked data is also used in health surveillance systems to enrich data that is used for the detection of suspicious patterns, such as outbreaks of contagious diseases. Another application where record linkage is of prime importance is fraud and crime detection as well as national security.
The main purpose of indexing technique is to improve the speed of data retrieval operations on a databases. An index is a copy of data from a table that can be searched very efficiently which also includes direct link to the complete row of data from where it was copied. So ideally an indexing technique for record linkage and de-duplication should be robust with regard to multiple databases. Thus, the aim of this paper is to fill the gap and provide both researchers and practitioners with the information about characteristics of sorted neighbourhood indexing technique.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

II. RELATED WORK

In this paper [1] provides background on record linkage methods that can be used in combining data from a variety of sources which consists of person lists, business lists. It also gives some areas of current research. In this paper [2], Record matching, which identifies the records that represent the same real-world entity, is an important step for data integration. Most state of the art record matching methods are supervised, which requires the user to provide training data. In this paper[3],the necessity of comparing records in a file with those in other file in an effort to find out which pair of records relate to the same data unit which arises many contexts most of which can be categorized as either the same record in main file storage. In this paper [4],often entities have two or more representation in the database. Duplicate records do not share a common key and/or they contain errors that make duplicate matching a difficult task.Indexing is required for real time entity resolution system so as to speed up the matching process .This is done by reducing the number of candidate records that need to be matched with a query record. In this paper[5] ,Indexing is required for real time entity resolution system to speed up the matching process by reducing the number of candidate records that need to be matched with a query record. In this paper Survey of twelve variations of six indexing techniques are given, from which we have selected Sorted Neighbourhood Indexing Technique.

III. PROPOSED ALGORITHM

A. *Design Considerations:*

- This technique was proposed in mid 1990s.
- In this algorithm databases are sorted according BKVs (Blocking Key Values)
- The window of a fixed size w, (w >1) moved over a data sets sequentially.
- Using this methodology, record pairs are generated in current window.

B. *Description of the Proposed Algorithm:*

Aim of the proposed algorithm is sort the database according to BKVs and generates a record pairs. After generating record pairs, linking is done.

Step 1: Take multiple databases(two or more)

Step 2: Sort the database using BKV's(blocking key value). Fix the window size w and it should be always greater than 1 i.e.(w>1)
This window is moved over data sequentially to get similar record pairs.

Step 3: Linking of records is done. For record linkage consider following terms,
(nA + nB): total no of records in both the databases
(nA + nB − w + 1): total no of window positions.
Total no of unique candidate pairs generated equals to:
uSNDSA = w(w − 1)/2 + (nA − w)(w − 1)
= (w-1)(nA-w/2)

C. *Proposed system:*

The proposed system have the following four modules along with functional requirements as shown in fig.1.

1. Cleaning and standardization
2. Record pair comparison
3. Similarity vector classification
4. Clerical review

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

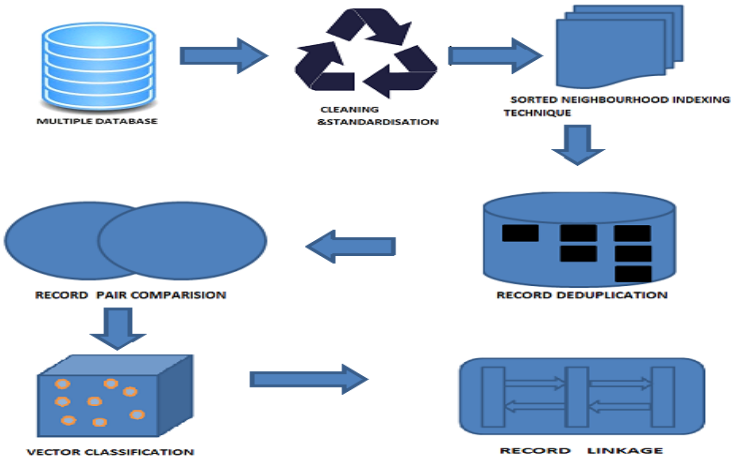

Fig.1 Architecture of Proposed System

a) Cleaning and standardization

Most of the real-world data are dirty and contain noisy, incorrectly formatted and incomplete information. So, if the users believe that the data is dirty, they are unlikely to trust the results of any data mining that has been applied. An essential first step in any record linkage or deduplication project is data cleaning and standardization. It has been known that a lack of good quality data can be one of the biggest hurdle to successful record linkage. The main purpose of data cleaning and standardization is the conversion of the raw input data into well defined, consistent forms.

.

b) Record pair comparison

The indexing step produces pairs of candidate records that are compared in detail with the comparison step using a variety of comparison functions suitable to the content of the record fields (attributes). Approximate string comparisons, which take (typographical) variations into account, are mostly used on fields that for example contain name and details of address, while comparison functions specific for age, date, and numerical values are used for fields that contain such data. Many fields are normally compared for each record pair, which results in a vector that contains the numerical similarity values calculated for that pair.

c) Similarity vector classification

Using these similarity values, depending on the decision model that has been used the next step to be followed in the record linkage process is to segregate the compared candidate record pairs into matches, non-matches, and possible matches. The record pairs which were removed in the indexing step are now classified as non-matches without being compared explicitly.

d) Clerical review

. After the data sets are classified into record pairs according to their dissimilar and similar properties a clerical review process is required where these pairs are manually evaluated and segregated according to their similar and dissimilar properties.

IV. PSEUDO CODE

Step 1: Start

Step 2: Take the multiple databases(two or more)

Step 3: Sort the databases using BKV's(Blocking key values)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

A.Fix the window size for comparison,(w>1),which will move sequentially over data.
B.Compare record pairs within current window.
Step 4: BKV's are inserted into array that is sorted alphabetically from left hand side of array
Step 5: Then window is moved over this array and candidate record pair are generated in current window.
Step 6: In case of record linkage, BKV's from both databases will be inserted into combine array and sorted alphabetically, such that for each pair one record is selected from each of two database.
Step 7: End

V. SIMULATION RESULTS

Two experiments were conducted in succession. The first using four 'real' data sets that had been previously used by the record linkage research community and the second using artificial data sets. Table 1 summarizes these data sets. The aim of the first experiments was to under seek how sorted neighbourhood indexing technique is able to handle various types of data, while the aim of second experiment was to investigate the scalability of the different indexing techniques to larger data sets.

The first three databases were taken from real world. Now, 'Census' contains records that were created by the US Census Bureau based on real census data, 'Cora' comprises of bibliographic records of machine learning publications and 'Restaurant' contains records that were extracted from the Fodor and Zagat restaurant guides. The 'CDDB' data set comprises of records of audio CDs, such as their title, artist, genre and year. This last data set was recently used in the assessment of a novel indexing technique. The true match status of all record pairs is available in all four data sets.

As shown in Table 1, two series of artificial data sets have been created. The 'Clean' data contain 80% original and 20% duplicate records, with up to three duplicates for one original record, a maximum of one modification per attribute, and a maximum of three modifications per record. The 'Noisy' data comprises of 60% original and 40% duplicate records, with up to nine duplicates per original record, a maximum of three modifications per attribute, and a maximum of ten modifications per record.

Dataset name	**Task**	**Number of records**	**Total number of true matches**
Census	Linkage	449+392	327
Restaurant	Deduplication	864	112
Cora	Deduplication	1295	17184
CDDB	Deduplication	9763	607
Clean	Linkage	1000-100000	200-20000
Noisy	Linkage	1000-100000	400-40000

Table 1. Datasets used in Experiments.

VI. CONCLUSION AND FUTURE WORK

Now days E-commerce market is growing rapidly. The business analytics that are used for implementing various market consumer trends require analysis and adoption to new trends in order to maximize the profit This paper has presented a technique of sorted neighbourhood indexing. The number of candidate record pairs generated by this technique has been estimatedtheoretically, and their efficiency and scalability has been evaluated using various data sets. This system highlights that one of the mostimportant 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 the domain and linkage experts to decide how such blocking keys are defined. The experimental results showed that there are large differences in the number of true matched candidate record pairs generated by the different techniques, but the sorted neighborhood indexing technique provides an optimal solution and thus it seems that the goal of deduplication & record linkage is achieved. Hence to overcome the complexities of multiple databases and data processing we designed this system.The

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

application will benefit many industrial and IT organizations dealing in large databases.It can also be implemented in e commerce and in various business analytics for reducing complexity of databases and reduce database space.

.

REFERENCES

1 W.E.Winkler," Overview of record linkage and current research direction," US Bureau of the Census, Tech. Rep. RR2006/02,2006.
2 W.Su,Wang, and F. H. Lochoysky,"Record matching over query results from multiple web database," IEEETransactions on knowledge and Data Engineering, vol. 22, no. 4, pp. 578-589,2009.
3 D.E.Clark,"Practical introduction to record linkage from injury research," Injury Prevention, vol. 10,pp. 186-191,2004.
4 P.Christen, R Gayler, and D.Hawking,"Similarity-aware indexing for real-time entry resolution," in ACM CIKM 09, Hong Kong 2009,pp. 1565-1568
5 Peter Christen " A survey of indexing techniques for scalable records linkage and deduplication" ,"IEEE Transaction on Knowledge and Data Engineering, vol. 24, no. 5, 2012