International Journal of Engineering & Technology, 7 (1.1) (2018) 294-296

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Record linkage and deduplication using traditional blocking

Somasekhar G [1] *, Sesha Sravani K [2], Keerthi P [2], Sai Sandeep G [2]

[1] Asst. Professor, CSE Dept., KL Deemed to be University, Guntur, AP
[2] Student, CSE Dept., KL Deemed to be University, Guntur, AP
**Corresponding author E-mail: giddalurisomasekhar@kluniversity.in*

Abstract

Record Linkage and Deduplication are the two process that are used in matching records. Matching of records is done to remove the duplicate records. These duplicate records highly influence the outputs of data mining and data processing. If the matching of records is done on the single database, it is called Deduplication. In Deduplication we check for the duplicate records in the single database. Unlike deduplication if the matching of the records is done on the several databases it is called as record linkage. In this paper we also discuss about the indexing technique called as traditional blocking which is used to remove non matching pairs that leads to the less number of record pair to be compared.

Keywords: *Blocking; Blocking Key; Blocking Key Value; Deduplication; Record Linkage; Traditional Blocking.*

1. Introduction

As research projects and business collects huge data, mining techniques with processing, and analyzing on of massive databases have attracted both the industry and academia. A single task that attains increased importance in multiple application is the records matching of similar entities acquired from numerous databases. Information acquired from numerous sources is integrated to improve the data quality. To improve data quality data from multiple sources are combined. As the data quality increases it is easy to match frequent corresponding to entities and for easier data analysis.

This process is done to improve the data quality and integrity. For example, in medical policies matched data may contain important information that is collected with the time and survey methods [4], [5]. Today, many businesses use record linkage and deduplication techniques to duplicate their databases to match their data and to enhance data quality, example for this is e commerce marketing and collaborative projects.

Domains where record linkage are used heavily are crime detection and fraud. Crime investigators and Security organizations highly depend on the ability to rapidly access files of that person under investigation, or cross-check records from different databases. So that we can prevent terror and crimes at the earliest [1].

2. Literature survey

The problem in identifying the record linkage relates to similar entities but does not apply on databases that possess information related to people. Other entities for matching include consumer products, bibliographic citations, publications, web pages and search results. In bioinformatics with large database, record link age finds genome sequences, which is similar unknown sequence. In information retrieval, duplicate documents like web page and citations are removed.

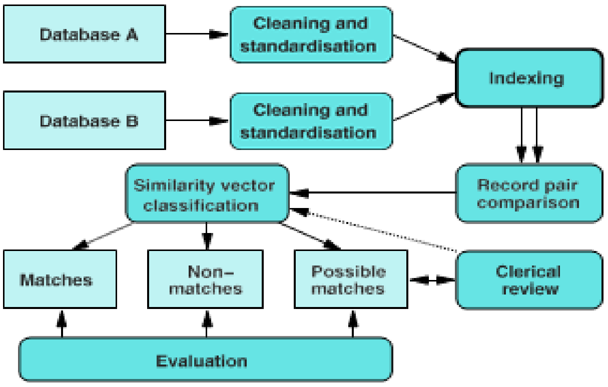

Fig. 1: General Record Linkage Process.

The results from search engines, digital libraries, text indexing automated systems is very much beneficial using the record linkage technique. Other application like consumer products in online shops involves finding the product and comparing it with relevant other shops in other sites. This requires record linkage and deduplication procedure. However, the product descriptions often vary slightly, hence the record matching seems a challenging one

2.1. Indexing for record linkage and deduplication

The matching of two different databases leads to comparison between each record in the database. This results in maximum number of comparisons of two different record from different databases. Similarly, while deduplication the individual database, the available record in an individual database is compared potentially with other data records. The bottleneck in deduplication or record linkage performance is its expensive field comparison between the data records. This makes the records pair comparisons not feasible for larger databases. Similarly, if there is no matched duplicate record, then the high number of true matched record correspond to minimum value in the databases. Likewise, in deduplication, the unique entities are always equal or less than the records in the database. Hence, there is a quadratic increment in the computation of records comparison with increased databases. However, the available true matches increase linearly with database size. Given that, the most comparisons between the data records are not considered as true matches. Hence, the indexing process reduces such potential comparisons through the elimination of non-matching record pairs. The record linkage approach otherwise called as blocking uses indexing technique to splits the given databases into blocks of non-matching sets and then the records of each block is compared with other individual blocks.

The complexity will be estimated for this process by the number of candidate records pairs generated. This is most time taking step in deduplication and record linkage project this step helps users to predict how much time will be taken for the project

2.2. Blocking key value

When we want to divide the records into non overlapping blocks it should be done based on a criteria it can be either a single attribute or the concertation of attributes. The records that have similar values are placed in one block, here the similarity depends on the similar looking values or similar sounding based on the characteristics of data to be matched [1].

They are lot of issues that needs to be considered when we are selecting an attribute as blocking key. The first issue that need to be considered is the quality. The field that we are selecting as the blocking key should not contain missing values and must have less errors and there must be variation in the value for the records rather than having the same value. The quality of the field is considered important because when the blocking key is formed with the field that consist errors then the record will be inserted into wrong block which cause missing of the true matches.

The other thing that must be considered when selecting a blocking key is the frequency distribution of the values in the field [1]. This is considered because it affects the size of the block suppose the field that we selected as blocking key consists of value that is repeated in almost all records. Then the result will be formation of a big block due to all records being placed in one block and it directly affects the execution time. If we take the gender as the blocking key for the set of records two blocks of large size will be formed and it takes comparatively more time than the selection of other field as blocking key. If in database A there are m records and in database B there are n records having the same BKV, then m x n record pairs will be created from that block, So it may be profitable if we use the fields that contain uniformly distributed value [3].

When we are selecting a blocking key we must also consider one trade off between many small blocks and large blocks. When we take the small blocks into consideration there will be less number of record pair and some true matches might miss. If you go with large blocks, then the true matches will be covered but comparing of many records is to be done as many record pairs are generated.

2.3. Phases in traditional blocking

The traditional blocking and other indexing techniques are generally divided into two phases Build and Retrieve

2.3.1. Build

In Build phase all the records that are present in database are read and the blocking key values are generated, and they are inserted in to the index data structure that is appropriate. If the inverted indexing is used, then the blocking key values become the key for inverted indexing and the records having same blocking key values are inserted into one inverted index list. For every database the Index data structures is built. In other case the database from which records origination is learned by flag which is generated by the record identifier.

The built phase also consists of entering the attributes which are used in comparison step into another data structure which will be more efficient in case of field comparisons to access the records. This is usually done by selecting an hash table or appropriate indexed database.

2.3.2. Retrieve

In each block the record identifier list is retrieved from inverted index, from the list the generation of candidate record is done. In deduplication in this phase every record in the block will be matched with other records in the same block, where as in record linkage the records in one block of the database will be paired with the records that contain same blocking key value from other blocks in other databases and After the comparison step is completed the resulting vector which contain the similarity value is given to classifier in classification step [1].

3. Traditional blocking

The records are divided based on the blocking key values. Records having the same Blocking key values are kept in one block and records in the that block are compared with each other. When we are comparing the records in the same block the number of records that are to be compared to find the matching record are decreased. Though this technique is oldest technique it gives the result faster than the rest of the indexing techniques when you use good Blocking key value.

Table 1: Example of Records in Database

Identifiers	Surname	BKV's (Sound encoding)
R1	Smith	S530
R2	Millers	M460
R3	Peter	P362
R4	Myler	M460
R5	Smyth	S530
R6	Millar	M460
R7	Smyth	S530
R8	Miller	M460

Here the Table consists of the records and the surnames are taken as blocking key the similarity between them is measured using soundex encoding and the value of the surnames becomes the blocking key value and the inverted index data structure for these records by using traditional blocking will be like.

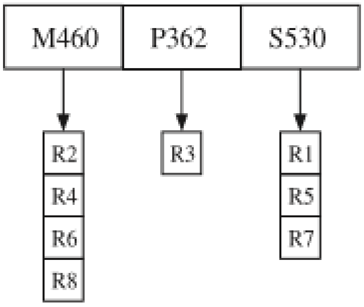

Fig. 2: When Traditional Blocking Is Appiled.

The records are divided based on the blocking key value and all the records that contain same blocking key value are in one block. The records that contain smyth and smith are one side because they are similar in sounding. In deduplication the records are compared in each block which results in the less number of comparisons.

The traditional blocking has its share of disadvantages when we made the record field that has errors and missing values as blocking key value there are good amount of chances that the record will not be inserted into the right block. Another disadvantage of traditional blocking is that when we select the record field that is repeated more times as the blocking key that results in forming the blocks with large size. When we retrieve data from such blocks high number of record pairs will be generated making the technique slower. Suppose in the student record list when we divide

the records based on the gender two blocks will be formed that will be large in size when lot of records are present. So, the number of records that are to be compared will be more than when we use the section of the student or the year he is studying as the blocking key. To increase the efficiency of traditional blocking technique it is better to use the less repetitive field as blocking key. If we use the uniform distributed blocking key which gets us the blocks of uniform size, then the number of records that are generated is
For Deduplication

$$U_{TBD}= b*(n_A / b * (n_A / b -1) / 2) = n_A /2(n_A / b -1) \quad (1)$$

For record linkage

$$U_{TBRL} = b *(n_A /b * n_B /b) = n_A n_B / b \quad (2)$$

Where
na/ b, nb/b is the records in the databases.
b is the number of blocking key values
Having blocks of uniform size is more advantageous than non-uniform size as the uniform size will generate less number of record pairs than other. For example, let's take two blocks of uniform size containing records of number 'a' then the pairs generated is 2 * a^2. On the other hand, let's take a non uniform block consists of records in the number (a - 2) and (a+2) respectively.
The number of comparisons will be

= (a-2) ^2 + (a+2) ^2

= (a^2 – 4a + 4) + (a^2 +4a +4)

= 2* a^2 +8

From this we comparison we can say that non uniform block generate large number of record pairs than uniform blocks and it is better to generate uniform blocks in traditional blocking.

4. Conclusion

Through traditional blocking we can decrease the time taken for the execution of record linkage and deduplication as it is decreasing the number of pairs to be compared. Even though it is the oldest indexing technique it gives result quickly when we use the good blocking key value and split the records into uniform blocks. In future the traditional blocking should be improved as the similarity between different blocks of records must be less than the minimum similarity and the similarity in the same block of records must contain good amount of similarity between them so that the true matching records will be in one block when we are dividing them as blocks in traditional blocking.

References

[1] Peter Christen, "A Survey of Indexing techniques for Scalable Record Linkage and Deduplication," Journal of Knowledge and Data Engineering, Vol 24, September 2012.

[2] J. Jonas and J. Harper, "Effective Counterterrorism and the Limited Role of Predictive Data Mining," Policy Analysis, no. 584, pp. 1-11, 2006.

[3] Carlo Batini, Monica Scannapieco, "Data and Information Quality: Dimensions, Principles and Techniques " pp 228.

[4] D.E. Clark, "Practical Introduction to Record Linkage for Injury Research," Injury Prevention, vol. 10, pp. 186-191, 2004. https://doi.org/10.1136/ip.2003.004580.

[5] C.W. Kelman, J. Bass, and D. Holman, "Research Use of Linked Health Data—A Best Practice Protocol," Australian NZ J. Public Health, vol. 26, pp. 251-255, 2002. https://doi.org/10.1111/j.1467-842X.2002.tb00682.x.