International Journal of Information Technology and Management

Vol. IX, Issue No. XIV, November-2015, ISSN 2249-4510

AN EFFICIENT DUPLICATION RECORD DETECTION ALGORITHM FOR DATA CLEANSING

AN INTERNATIONALLY INDEXED PEER REVIEWED & REFEREED JOURNAL

www.ignited.in

International Journal of Information Technology and Management
Vol. IX, Issue No. XIV, November-2015, ISSN 2249-4510

An Efficient Duplication Record Detection Algorithm for Data Cleansing

Radha Saini[1]* Dr. Omparkash[2]

[1]Research Scholar

[2]Associate Professor, Faculty of Computer Science, OPJS University, Rajasthan

Abstract – The purpose of this research was to review, analyze and compare algorithms lying under the empirical technique in order to suggest the most effective algorithm in terms of efficiency and accuracy. The research process was initiated by collecting the relevant research papers with the query of "duplication record detection" from IEEE database. After that, papers were categorized on the basis of different techniques proposed in the literature. In this research, the focus was made on empirical technique. The papers lying under this technique were further analyzed in order to come up with the algorithms. Finally, the comparison was performed in order to come up with the best algorithm i.e. DCS++. The selected algorithm was critically analyzed in order to improve its working. On the basis of limitations of selected algorithm, variation in algorithm was proposed and validated by developed prototype.

After implementation of existing DCS++ and its proposed variation, it was found that the proposed variation in DCS++ producing better results in term of efficiency and accuracy. The algorithms lying under the empirical technique of duplicate records deduction were focused. The research material was gathered from the single digital library i.e. IEEE. A restaurant dataset was selected and the results were evaluated on the specified dataset which can be considered as a limitation of the research. The existing algorithm i.e. DCS++ and proposed variation in DCS++ were implemented in C#. As a result, it was concluded that proposed algorithm is performing outstanding than the existing algorithm.

General Terms

Data Quality, Data Cleansing, Dirty Data

Keywords - Duplication Records Detection Algorithm, DCS++, Windowing, Blocking

----------X----------

1. INTRODUCTION

Now-a-days, the digital economy is totally dependent on the databases. Many industries and businesses have huge amount of data stored in different databases. In this fast world, it is necessary that data operations on the database are carried out smoothly and efficiently (Ahmed et. al., 2007).. However, to access the useful information that can help in decision making for industries and businesses, it is necessary to integrate large dataset. When data is integrated from different sources then it contains a huge part of dirty data. This dirty data contain mistakes in record values, duplication in records, spelling mistakes, null or illegal values, disobedience referential integrity and inconsistency in records (Ying et. al., 2009).

Quality assurance of data is necessary for fast retrieval of data, quick and smooth data processing, and right decision making. Business organizations are paying high attention towards data quality because dirty data can effect important decisions in businesses. In addition, cleansed data can improve the production because of quality decisions (Rehman and Esichaikul, 2009). Data cleansing is performed to get cleansed and quality data. Therefore, Data cleaning is important for business industry. The available data cleaning methods are not time and cost effective (Mansheng et. al., 2009). Duplication in data is one of the most important issues of Data cleaning. When data is gathered from different source then due to mistakes in spells or difference or inconsistency of format may cause presence of duplicate records in data (Hua et. al., 2010). Extraction of knowledge from huge databases is known as data mining (Hua et. al., 2010). Duplicate record deduction and data redundancy control are also hot topics of data mining and data integration (Huang et. al., 2008. Mansheng

et. al., 2009). With the increase of Quality data demand, many logical and statistical methods have been provided to resolve the problem (Gollapalli et. al., 2011). In this regard, there are three basic techniques of Duplicate records detection which are knowledge-based techniques, probabilistic techniques and empirical techniques (Rehman and Esichaikul, 2009). Many algorithms have been proposed under those techniques but all of them somehow lack in one of these parameters which are time efficiency, cost effectiveness, space consumption and accuracy (Gollapalli et. al., 2011). Duplication record detection is a very diverse field so this decision was made that one of its basic technique will be chosen and then focus will be on algorithms which lie within that technique. It was decided to select empirical technique and compared all the algorithms under this category. After comparison, most effective algorithm will be selected and improved accordingly. The objectives of this research study are as follows:

1. To study the algorithms of duplication records detection

2. To perform comparative analysis of duplicate records detection algorithms lying under the empirical technique

3. To implement the best selected algorithm after performing comparative analysis

4. To suggest improvement in the selected algorithm

2. LITERATURE REVIEW

This section provides the necessary background material that is required to understand this research theme.

2.1 Types of Data Sources

Data can be retrieved from single and multiple sources. Therefore, data quality is ensured in both cases.

Single source data can have lack of integrity constraints and poor schema design at schema level and mistakes in data entries or duplication in data at instance level. Multi source data faces the issue of numbering and structural conflicts at schema level and overlapping, contradiction, inconsistency of data at instance level (Rehman and Esichaikul, 2009).

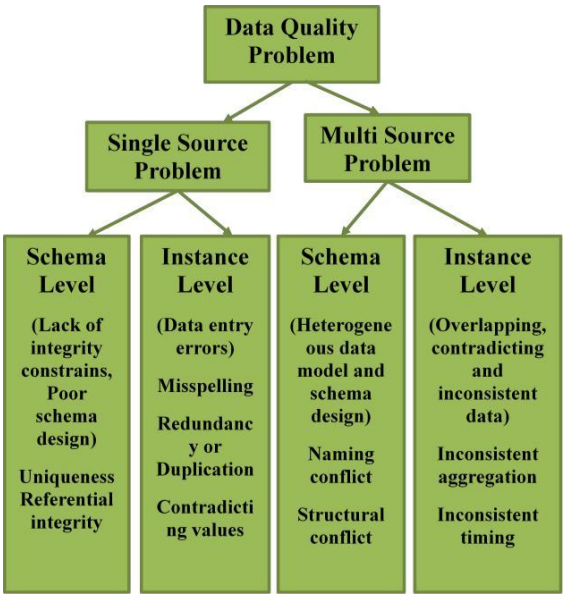

Fig1: Division of Data quality problems according to resources (Rehman and Esichaikul, 2009).

2.2 Dirty Data and Data Cleaning

When data is integrated from multiple sources then it contains a huge part of dirty data in it. This dirty data contain mistakes in records values, duplication in records, spelling mistakes, null or illegal values, disobedience referential integrity and inconsistency in records. This dirty data can infuse authentication of data. Therefore, it is necessary to clean data Data quality management is burning issue of enterprises because it has power to manipulate the decisions (Wei et. al., 2012). Therefore, data quality is spotted as bottleneck issue in businesses and industries (Zhe and Zhi-gang, 2010).

Operational databases and online analytical processing systems cannot avoid the issue of data quality while integrating data. These issues are caused by non-unified set of standards in distributed databases. Data cleaning plays an important role in providing quality data by detection and removal of inconsistencies from data [11].

2.3 Duplicates and Types of Duplicates

Duplicates are the records that represent the same real-world object or entries. Record matching is a state of art technique to find these duplicates [12].

Duplicates can be of two types that are exact or mirror duplicates and approximate or near duplicates. Exact duplicate records contain the same content but on the other hand content of near duplicate records vary slightly [13]. The records which contain syntax differences or typographical errors but represent the same real world entity are known as near duplicates [14].

2.4 Duplication Records Detection and Types

Duplicate record detection is one of the most important data quality problems [15]. Detection of Duplicate plays an important role in record linkage, near duplicate detection and filtering queue [16]. Duplication detection is used to identify the same real world entities which exist in different format or representation in database [17,18]. It is very common to find some non-identical fields or records that refer the same entity

Efficient and accurate detection of duplicates is hotspot of the data mining and online analyzer (Mansheng et. al., 2009). Now-a-day, duplication detection is the most popular topic in research (Gollapalli et. al., 2011). Duplication detection is based on two basic Stages. The first one is the outer stage in which record matching technique or duplication record matching technique is applied. The second one is the inner stage that is based on field matching techniques.

Duplication record detection algorithms are divided in three types i.e. knowledge-based techniques, probabilistic techniques, and empirical techniques (Rehman and Esichaikul, 2009). Empirical algorithms consist on sorting, blocking and windowing methods. Knowledge based algorithms demand training and the use of that training and reasoning skills in order to perform detection. Probabilistic algorithms are based on statistical and probability methods that are Bayesian networks, expectation maximization and data clustering. In this research study, focus is on empirical algorithms.

2.5 Empirical Algorithms

The general algorithms are as follows:

2.5.1 Blocking

Assign the *sorting key* to each record. Sort all the Records according to the key. Later, records are partitioned into disjoint partitions (means no record can be present in more than one partition) according to some *blocking key* (partition predicate).

Finally, comparison will be performed between records within the blocks. Using this technique, least comparisons will be performed [20].

2.5.2 Windowing

First of all, Merge two provided list of records. Sort all the records by *lexicon* order according to the attributes selected as a key. A fixed size slide window will be used. Records within the window will be compared with each other and first record will be released to select the next record in *fixed size* window.

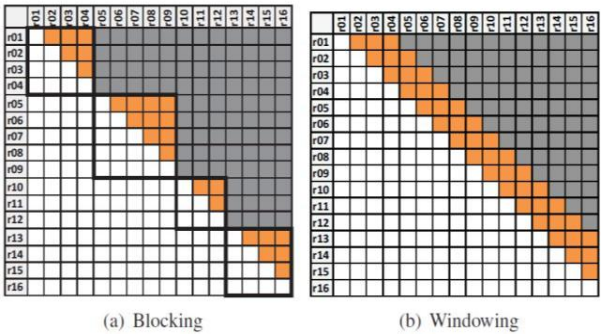

Fig2: Selection of elements for comparison in Windowing and Blocking [20]

2.6 Comparison among Windowing and Blocking techniques

Similarity and differences among these algorithms are discussed below:

Both algorithms try to perform reduction in number of record comparisons. For reducing comparisons, intelligent guesses are made about window / block sizes. In both algorithms, first of all data records are sorted and it is assumed that after sorting duplicates will be close to each other.

However, the mechanism of selection of records for comparisons is different from one another. In blocking algorithms, records are blocked in disjoint partitions. On the other hand, windowing algorithm works by sliding a window over the records.

The use of domain specific key for sorting can reduce the complexity of the algorithm but also cause domain dependency [21]. It is not even necessary to keep the key domain specific. Therefore, blocking and windowing methods such as sorted neighborhood are domain independent [22]. In this research, empirical algorithms are chosen due to the nature of domain independence.

2.7 Related Work

The algorithms have been discussed in detail below:

2.7.1 Sorted Blocks

Input Parameters: Records, key (may or may not be unique), overlapping value (o)

Records are blocked according to the partition predicate. After that records within the partition plus the overlapping records (Selected with the help of a fix size parameter) will be compared with each other.

*Output***:** Duplicate or Non-Duplicates

2.7.2 Duplicate Count Strategy++ Input Parameters: *Records, Sorting key (key), Window Size (w), Threshold ( )*

A growing window is slide over the records and records within the window are compared with each other. If a duplicate is found then it will be added to skipped list and will never be selected again for comparison which will ultimately reduce the number of comparisons.

Output: Duplicate or Non-Duplicates

2.7.3 Decision making algorithm

Input Parameters: Databases, Databases priorities values, Initial field priorities values, Initial threshold, Final threshold Match the field count of each record and assign each field of first database to the field of other database. Set the priorities of fields and sort them accordingly. Select a specific number of fields of all records and compare them if any two records cross a specific threshold then these records will be compared further.

International Journal of Computer Applications (0975 – 8887) Volume 127 – No.6, October 2015

Output: Exact Similar, Approximate Similar, Less Similar and Non Similar

2.7.4 Nested Blocking

Input Parameters: Data source, standardization rules, blocking fields and Threshold

Records are divided into partitions then partitions are further divided into sub-partitions. Afterwards, comparison will be performed within sub-partitions.

Output: Duplicate, Possible Duplicate or Non-Duplicates

2.7.5 PC-Filter+

Input Parameters: Database, blocking key value, Size of partition (s), threshold ( )

Records are blocked in equal size partitions. Records within the blocks will be compared. PCG (partition comparison graph) will be constructed for inter comparison. If number of blocks will be less than defined ratio then all blocks will be compared with each other. Otherwise, a defined number of neighboring blocks will be compared.

Output: Duplicate or Non-Duplicates

3. RESEARCH METHODOLOGY

The steps of research methodology adopted in this research study are shown in Fig 3 and their description is given below:

3.1 Set Aims and Objectives of Research

The main purpose of this research was to review different algorithms which have been proposed in the literature to suggest the most effective one in terms of efficiency and accuracy.

3.2 Preparation of Proposal

Some research articles were selected randomly from ACM and IEEE digital libraries. Based on these articles, proposal was written to defend and propose research topic.

3.3 Collection of Research Papers in the relevant domain i.e. duplicates records detection

Afterwards, it was decided that systematic review of literature will be followed. In order to perform the systematic review, different digital libraries were searched out for the research articles under duplication records detection keyword.

3.4 Search of research papers

While searching articles, it was found that IEEE digital library contains most relevant research articles. With the keyword of “duplicate records detection” total 61 articles were found.

3.5 Selection of relevant Research Papers and Division of Research Paper

Selected research articles were divided into four major categories. From these categories, there were three different techniques of duplication record detection and the articles which did not lie under these techniques were categorized as ‘Others’.

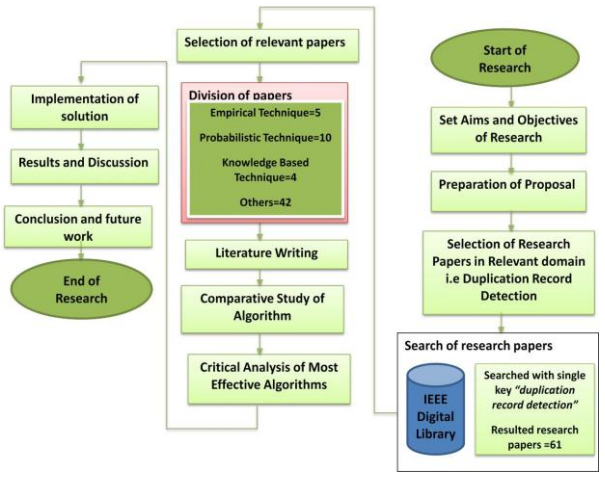

Figure 3: Research methodology

3.6 Literature Writing

Literature review was performed based on the selected articles and the main focus remained on empirical techniques and other techniques were ignored for the sake of this research.

3.7 Comparative Study of Algorithms under Empirical Technique

Comparative study of algorithms under the heading of empirical technique was performed in order to come up with comparative analysis. .

Critical analysis of most effective algorithm

Critical analysis of DCS++ was performed and suggestions were given for improvement of the algorithm.

3.8 Implementation of Solution

Solution is implemented for the existing DCS++ Algorithm and the proposed Algorithm.

3.9 Results and Discussions

The evaluation of algorithm was performed and the results have been discussed in detail below.

4. CRITICAL ANALYSIS DCS++

Windowing algorithm provides more accuracy instead of blocking. Therefore, DCS++ is selected because it is the most efficient windowing algorithm among all variants which are included in this study.

4.1 DCS++ Algorithm

Sort all the records according to the *sorting key*. Afterwards, put the *w* records in current window (*win*) sequentially. Now, select a record from all records sequentially and check whether the record is in skip records list (*SkipRecords*) or not. Compare the selected record with all the records within win and increase count of number of comparisons (*c*) by 1. If a record is found as a duplicate of Selected record then mark it as duplicate by adding it in to the *SkipRecords*, increase count of current duplicated record (*d*) by1and add the record in *win* sequentially till *win.lenght* < *duplicate record count* + *w-1* and *win.lenght with increase* < *records*. When all the records within the win are compared, remove the first record of *win*. If remaining records in *win* < *w* then add one record at the end Otherwise, remove records from the end till *win. Length* = *w* and move back to the step of selecting a record sequentially from all records. Continue the process till end of records [23].

4.2 Critical Point

Records are sorted according to single or composite key but not by all fields of the records. Therefore, it is possible that duplicate records lie in the same window but not consecutively. In Fig 4, full advantage of transitive property with DCS++ algorithm cannot be taken. It is clearly reflected by the Fig. 5, with any size of window, that record numbers 3, 4, 5, 6 even if add to the skip list in first window but will be compared again with record 2 in second window.

In such case, DCS++ will perform unnecessary comparisons. As shown in Fig 4. The problem can be resolved by increasing a single check in the algorithm. After that, algorithm will avoid those unnecessary comparisons.

4.3 Proposed Algorithm

Sort all the records according to the *sorting key*. Afterwards, put the *w* records in current window (*win*) sequentially. Now, select a record from all records sequentially and check whether the record is in skip records list (*SkipRecords*) or not. Compare the selected record with all the records within *win* that are not in *SkipRecords* and increase count of number of comparisons (*c*) by 1 with each comparison. If a record is found as a duplicate of Selected record then mark it as duplicate by adding it in to the *SkipRecords*, increase count of current duplicated record (*d*) by1and add the record in win sequentially *till win.lenght* < *duplicate record count* + *w-1* and *win.lenght with increase* < *records*. When all the records within the win are compared, remove the first record of win. If remaining records in *win* < *w* then add one record at the end. Otherwise, remove records from the end till *win. Length w* and move back to step of selecting a record sequentially from all records. Continue the process till end of records.

Now, the prototype of proposed algorithm is developed to find that whether with the improvement in DCS++ have retained its accuracy. Secondly, an attempt is being made to see whether with a good String matching algorithm, is there any potential to have higher precision value.

5. RESULTS AND DISCUSSIONS

It is concluded by Table 2 and Table 3 that with the use of exact string matching algorithm, the accuracy of proposed algorithm is same as accuracy of DCS++ algorithm and there is no improvement in number of comparisons. The results are not bad as 100% Precision and 70.63% Recall is achieved. The most noticeable thing is that, there is not a single false detection of duplicate with naïve string matching

algorithm. Overall, results with naïve algorithm are satisfactory.

The Proposed Algorithm with Basic String Matching Algorithm requires reduced number of comparisons instead of DCS++ with Basic String Matching Algorithm. On the other hand, both algorithms have same accuracy. By Table 2, 3, 4 and 5, it can be concluded that the Recall value of both algorithms i.e. DCS++ and Proposed Algorithm with Basic String match algorithm by using the right threshold value is more than of naïve algorithm, but this gain requires the little compromise on the Precision value.

Table 6 and 7 shows that the Proposed Algorithm that is implemented with the modified Recursive Algorithm is performing more accurately and efficiently than of DCS++ with Recursive Algorithm with lower threshold values but with higher threshold values they have same performance. Another important aspect is the gain of 96.81% F-Score value. It can be concluded by taking look at Table 4, 5, 6 and 7 that DCS++ and Proposed Algorithm with Recursive Algorithm is performing much better than of Basic String matching algorithm, but the error percentage of DCS++ and Proposed algorithm with best F-Score is ((Numbers of duplicates actually exist - Numbers of duplicates detected)/Number of total records in dataset)*100= ((112-91)/865)*100=2.43%. This error percentage is extremely low so it is negligible.

With the help of above discussion, it can be concluded that the proposed algorithm which is implemented with the help of Modified Recursive Algorithm outperforms than of all other algorithms in term of accuracy and efficiency.

8. CONCLUSION

The most challenging task of this research study was to prove that after making changes in the basic algorithm of the DCS++, there is no loss of efficiency or accuracy instead of proving the improvement. Prototype of both original DCS++ algorithm and the new proposed algorithm is implemented. With the results of evaluation, it is concluded that with Exact String or Field match both algorithms work almost in similar manner. On the other hand, with Approximate String or Field match number of comparisons are reduced by the proposed algorithm.

Moreover, accuracy in terms of recall, precision and F-Score is almost similar for both algorithms, but in case where Proposed Algorithm is used with modified recursive algorithm with minimum threshold value, it produces more accurate results than of original DCS++.

It is also proved that it is mostly not possible in case of real data that all duplicates are detected with the use of exact string matching algorithm, even if the precision reached to 100% but the F-Score is lower. The reason of using two different approximate algorithms was to show that there is a room to gain higher rate of duplicate detection with the same record detection algorithm by using more efficient string matching algorithm. The recursive algorithm is used by calling twice for a single string match to gain high accuracy with a non-symmetrical algorithm. It increases the complexity but outperforms with the efficient choice of the threshold value.

The proposed algorithm is the best choice for the task of duplication record detection. It is domain independent but input dependency is there. The algorithm provides almost similar results than of DCS++ in terms of accuracy excluding some cases where accuracy of proposed algorithm is higher. On the other hand, efficiency of proposed algorithm is equal or higher in some cases.

6. RECOMMENDATIONS AND FUTURE WORK

The research is scoped to the empirical techniques only. However, other techniques can be explored with the same directions. For the approximate string matching, basic and recursive algorithms are improved and applied to see their effects. There are many other algorithms which exist for approximate string matching and yet 100% precision and recall with the approximate match is not achieved yet so other techniques can also be applied to produce better results. Moreover, Window can be slide on field along with records instead of sliding window only on the records. This means that instead of selecting only number of records in a window, the number of fields can also be reduced with respect to its important.

In this research, it was found that there exist some records which require knowledge base to detect duplicates correctly. For example, Arfa Sikander, Street number 5 iqbal roads, Daska and Arfa Sikander, Street number 5 iqbal road, Sialkot are the same records but in the first case nearby famous city name is mentioned. Moreover, there are also few other cases which are not being handled by the recursive algorithm. For example, 7th or seventh, these both cases cannot be handled without knowledge base.

10. REFERENCES

Ahmed K. Elmagarmid, P., G. Ipeirotis, and Vassilios S. Verykios (2007). "Duplicate Record Detection: A Survey," IEEE Trans. on Knowl. and Data Eng., vol. 19, pp. 1-16.

D. Bhalodiya, M., K. Patel, and C. Patel (2013). "An Efficient way to Find Frequent Pattern with," in Nirma University International Conference on Engineering.

L. Huang, P. Yuan, and F. Chu (2008). "Duplicate Records Cleansing with Length Filtering and

Dynamic Weighting," in Semantics, Knowledge and Grid, 2008. SKG '08. Fourth International Conference on, Beijing, pp. 95 - 102.

L. Zhe and Z. Zhi-gang (2010). "An Algorithm of Detection Duplicate Information Based on Segment," in International Conference on Computational Aspects of Social Networks.

M. Gollapalli, X. Li, I. Wood, and G. Governatori (2011). "Approximate Record Matching Using Hash Grams," in 11th IEEE International Conference on Data Mining Workshops.

M. Rehman and V. Esichaikul (2009). "DUPLICATE RECORD DETECTION FOR DATABASE CLEANSING," in Machine Vision, 2009. ICMV '09. Second International Conference on , Dubai, pp. 333 - 338.

P. Ying, X. Jungang, C. Zhiwang, and S. Jian (2009). "IKMC: An Improved K-Medoids Clustering Method for Near-Duplicated Records Detection," in Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on, Wuhan, pp. 1 - 4.

Q. Hua, M. Xiang, and F. Sun (2010). "An Optimal Feature Selection Method for Approximately Duplicate Records," in Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on, Chengdu.

X. Mansheng, L. Youshi, and Z. Xiaoqi (2009). "A PROPERTY OPTIMIZATION METHOD in SUPPORT of APPROXIMATELY DUPLICATED RECORDS DETECTING," in Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on.

Z. Wei, W. Feng, and L. Peipei (2012). "Research on Cleaning Inaccurate Data in Production," in Service Systems and Service Management (ICSSSM), 2012 9th International Conference on, Shanghai.

Corresponding Author

Radha Saini*

Research Scholar

E-Mail – sainiradha2010s@gmail.com