

A STUDY AND SURVEY ON VARIOUS PROGRESSIVE DUPLICATE DETECTION MECHANISMS

Ashwini.V. Lakote¹, Lithin K²

¹M.Tech Computer Science and Engineering, REVA Institute of Technology and Management, Bangalore, India
ashwini.lakote@gmail.com

²Dept of Computer Science and Engineering, REVA Institute of Technology and Management, Bangalore, India
lithinkumble@gmail.com

Abstract

One of the serious problems faced in several applications with personal details management, customer affiliation management, data mining, etc is duplicate detection. This survey deals with the various duplicate record detection techniques in both small and large datasets. To detect the duplicity with less time of execution and also without disturbing the dataset quality, methods like Progressive Blocking and Progressive Neighborhood are used. Progressive sorted neighborhood method also called as PSNM is used in this model for finding or detecting the duplicate in a parallel approach. Progressive Blocking algorithm works on large datasets where finding duplication requires immense time. These algorithms are used to enhance duplicate detection system. The efficiency can be doubled over the conventional duplicate detection method using this algorithm. Several different methods of data analysis are studied here with various approaches for duplicate detection.

Index Terms: Data Duplicity Detection, Progressive deduplication, PSNM, Data Mining

1. INTRODUCTION

A. Data Mining

Data mining is also called as KDD or knowledge discovery in database.[1][2] The concept of data mining evolved from several researches that include statistics, database systems, machine learning concepts, neural networks, visualization, rough set, etc.[3][4] Both traditional and latest areas like businesses, sports, etc use the data mining concepts. For translating the raw data into valuable information, the companies use a process. By knowing the details about the customers and by developing efficient marketing policies, the sales and costs can be increased or decreased in the businesses. The efficient collection of data, warehousing and computer processing all have their influence on data mining concepts.[5] The data is the most essential important asset of any company but incase the data is changed or a bad data entry is made certain errors like duplicate detection arises. [6]

B. Duplicate Detection Problems

Duplicate detection denotes to the process of recognizing different representations of the real world objectives present in an information source [7][8]. It is not possible to ignore several qualities of duplicate detection like effectiveness and scalability due to the database size. [9]

There are two features in the problems of duplicate detection which are as follows:

- Several representations generally are not same and have certain differences like misspelling, missing values, changed addresses, etc which makes the detection of duplicates very difficult.

- The detection of duplicates is very costly because the comparison among all possible duplicate pairs is required.
- Progressive duplicate detection algorithms are as follows:-
- PSNM or Progressive Sorted Neighborhood Method working over clean and small datasets.
- PB or Progressive Blocking working over unclean and large datasets.

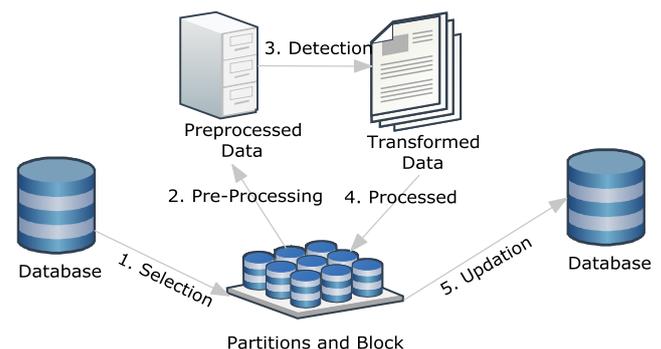


Figure 1: System Architecture

Above system architecture explains the process of duplicate data detection using progressive mechanism. This architecture is discussed in detail in section 3 of this paper.

Definitions:

- **Duplicate Detection:** It is the process of recognizing several representations in a matched real world item.
- **Data Cleaning:** It is known as Data Scrubbing which denotes a process of detection, correction and removal of

corrupted and inappropriate records present in the databases, tables, record sets, etc.

- **Progressiveness:** It improves the results, efficiencies and scalability of the algorithms used in this existing model. Techniques like window interval, look ahead, partition caching, Magpie Sort are used for delivering the results faster.
- **Entity Resolution:** It is also called as de-duplication or record linkage which identifies the accounts corresponding to similar entity of a real-world.
- **Pay-As-You-Go:** It is a technique where the candidate pairs are theoretically ordered by the matching chances. Then comparison on records using the match pairs are performed using the ER algorithm.

2. RELATED WORKS

P. G. Ipeirotis et al. proposed the following concepts in [9] which states that the ER algorithm is used in this paper for focusing on determine the expected records that are alike first. This technique gives various hints like the other general techniques. But still many problems are yet to be solved. There are three different types of hints which match the several ER algorithms called sorted list of record pairs, hierarchy of record partitions and order list of records. The hints are used to maximize the count of similar records recognized with less work and to increase ER quality.

S. E. Whang et al. [10] stated a survey on the active methods and non identical duplicate entries present in the records of the database records are all investigated in this paper. It works for both the duplicate record detection approaches. 1) Distance Based technique that measures the distance among the individual fields, by using distance metrics of all the fields and later computing the distance among the records. 2) Rule based technique that uses rules for defining that if two records are same or different. Rule based technique is measured using distance based methods in which the distances are 0 or 1. The techniques for duplicate record detection are very essential to improve the extracted data quality.

U. Draibach et al. in [11] denoted a Duplicate Count Strategy is used which become accustomed to the window size depending on the count of duplicates detected. There are three strategies:

- **Key similarity strategy:** The associations of the sorting keys influence the window size which is improved when the sorting keys are alike. Then we can expect several related records in this model.
- **Record similarity strategy:** The associations of the records influence the window size. The replacement of the real resemblance of the records is present inside the window.
- **Duplicate count strategy:** The count of the known duplicates influence the window size. DCS++ algorithm proves to be trustworthy than the SNM algorithm without losing the effectiveness. The algorithm of DCS++ is used to calculate the transitive closure and then save comparisons.

U. Draibach and F. Naumann in [12] proposed two major methods called blocking and windowing used to reduce the comparisons are studied in this paper. Sorted Blocks that denotes a generalization of these two methods are also analyzed here. Blocking divides the records to disjoint subsets and windowing slides a window on the sorted records and then comparison is made between records within the window. The sorted Blocks have advantages like the variable size of partition size instead of the size of the window.

A. Thor et al. [13] proposed a theory of deduplication which is also known as Entity Resolution which is used for determining entities associated to similar object of the real world. It is very important for data integration and data quality. Map Reduce is used for SN blocking execution. Both blocking methods and methods of parallel processing are used in the implementation of entity resolution of huge datasets.

Map Reduce steps:-

1. Demonstrating how to apply map reduce for a common entity having blocking and matching policies.
2. Identifying the main challenges and proposing two JobSN and RepSN approaches for Sorted Neighborhood Blocking.
3. Evaluating the two approaches and displaying its efficiencies. The size of the window and data skew both influences the evaluation.

3. PROPOSED SYSTEM

The proposed solution uses two types of novel algorithms for progressive duplicate detection, which are as follows: PSNM – It is known as Progressive sorted neighborhood method and it is performed over clean and small datasets. PB – It is known as Progressive blocking and it is performed over dirty and large datasets.

Both these algorithms improve the efficiencies over huge datasets. Progressive duplicate detection algorithm when compared with the conventional duplicate persuades two conditions which are as follows [1]:

- **Improved early quality:** The target time when the results are necessary is denoted as t . Then the duplicate pairs are detected at t when compared to the associated conventional algorithm. The value of t is less when compared to the conventional algorithm's runtime.
- **Same eventual quality:** When both the progressive detection algorithm and conventional algorithm finishes its execution on the same time, without terminating t earlier. Then the produced results are the same.

As demonstrated in Fig. 1 i.e. System Architecture, initially a database is picked for deduplication and for practical processing of data, the data is split into numerous partitions and blocks. Clustering and classification is used after sorting the data to make it more ordered for efficiency. Next step the pair wise matching is done to find duplicates in blocks and through new transformed dataset is generated. Finally

the transformed data is updated in database after all filtrations.

When the time slot of fixed is given then the progressive detection algorithms works on maximizing the efficiencies. Thus PSNM and PB algorithms are dynamically adjusted using their optimal parameters like window sizes, sorting keys, block sizes, etc. The following contributions are made which are as follows:

- PSNM and PB are two algorithms that are proposed for progressive duplicate detection. It exposes several strengths.
- This approach is suitable for a multiple pass method and an algorithm for incremental transitive closure is adapted.
- To rank the performance, the progressive duplicate detection is measured using a quality measures.
- Many real world databases are evaluated by testing the algorithms previously known.

There are three stages in this workflow which are as follows:

- Pair selection
- Pair wise comparison
- Clustering

Only the pair selection and clustering stages should be modified for a good workflow.

CONCLUSION

Several duplicate detection approaches are studied in this paper. The existing techniques which have algorithms to detect duplicity in records improve the competence in finding out the duplicates when the time of execution is less. The process gain within the available time is maximized by reporting most of the results.

REFERENCES

- [1]. "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.
- [2]. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases" (PDF). Retrieved 17 December 2008.
- [3]. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.
- [4]. Witten, Ian H.; Frank, Eibe; Hall, Mark A. (30 January 2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3 ed.). Elsevier. ISBN 978-0-12-374856-0.
- [5]. Think Before You Dig: Privacy Implications of Data Mining & Aggregation, NASCIO Research Brief, September 2004
- [6]. Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.
- [7]. M. A. Hernández and S. J. Stolfo, "Real-world data is dirty: Data cleansing and the merge/purge problem," *Data Mining and Knowledge Discovery*, vol. 2, no. 1, 1998.
- [8]. Thorsten Papenbrock, Arvid Heise, and Felix Naumann, 'Progressive Duplicate Detection' *IEEE Transactions on Knowledge and Data Engineering(TKDE)*,vol . 25, no. 5, 2014.
- [9]. A.K. Elmagarmid, P. G. Ipeirotis, and V. S.Verykios, "Duplicate record detection: A survey," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 19, no. 1, 2007.
- [10].S. E. Whang, D. Marmaros, and H. Garcia-Molina, "Pay-as-you-go entity resolution," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 25, no. 5, 2012.
- [11].U. Draisbach, F. Naumann, S. Szott, and O. Wonneberg, "Adaptive windows for duplicatedetection," in *Proceedings of the International Conference on Data Engineering (ICDE)*, 2012.
- [12].U. Draisbach and F. Naumann, "A generalization of blocking and windowing algorithms for duplicate detection." in *International Conference on Data and Knowledge Engineering (ICDKDE)*, 2011.
- [13].L. Kolb, A. Thor, and E. Rahm, "Parallel sorted neighborhood blocking withmapreduce," in *Proceedings of the Conference Datenbanksysteme in Büro, Technik und Wissenschaft (BTW)*, 2011.