Scientific Journal of Impact Factor (SJIF): 4.14

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 12, December -2016

"Survey on Detection of duplicate records using Windowing and Blocking techniques"

[1]Miss. Pooja Appasaheb Tippanna, [2] Prof. K. S. Kadam

[1,2]*Department of Computer science Engineering,D. K.T.E. Society's Textile and Engineering Institute, Ichalkaranji.*

Abstract — *Duplicate detection is process of finding multiple representation of same real word entity. In very large data sets, like malls and website data, the data is difficult to manage and if there are duplicate entries it is all the more time consuming. Process is required to find duplicate entities in very short time by maintaining the quality of dataset. In this proposed approach a novel method is used namely progressive blocking that considerably increases the proficiency of finding the duplicates and maintain quality of dataset.*

I. INTRODUCTION

Identifying multiple representations of same real word entities is duplication detection. The large databases usually face the challenges of duplicate data due to erroneous data entry. For example the data in super markets, retailing industry face these problems. The problem of detecting duplicate entities is an important data cleansing task, necessary to improve data quality. When merging data from different sources, the result is unique i.e. ensure that an entity has only one representation in result.

Firstly the data set is going to partitioned by using partitioning method like range, hash. The further process is going to implemented on these partitions.

After doing the partitioning, develop an algorithm for duplication detection, one based on neighborhood comparisons in predefined window size and the other based on formation of blocks in the dataset are proposed

In sorted neighborhood techniques partitioning the input data into windows of records and restricting entity resolution to entities of the same window thus become necessary to reduce then number of entity comparisons whilst maintaining match quality. After that it sorts all entities using an appropriate sorting key and only compares entities within a predefined window.

In the blocking method which constitute of creation of blocks holding similar records and comparing the records consisting in them. In blocking the records are presorted so as to obtain the similar blocks.

Sorted neighborhood method:

Sorted neighborhood is a popular windowing approach that works as follows. A sorting key K is determined for each of n entities. Typically the concatenated prefixes of a few attributes form the sorting key. Afterwards the entities are sorted by this sorting key. A window of a fixed size w is then moved over the sorted records and in each step all entities within the window, are compared. Sorted neighborhood is a popular blocking approach that works as follows. The Sorted neighborhood approach is very popular for entity resolution due to several advantages. First, it reduces the complexity. Thereby matching large datasets becomes feasible and the window size w allows for a dedicated control of the runtime. Second, the SN approach is relatively robust against a suboptimal choice of the sorting key since it is able to compare entities with a different sorting key

.

Blocking method:

In the blocking method, like Sorted neighborhood method, it also pre-sorts the records to use their rank-distance in this sorting for similarity estimation. Blocking algorithms assign each record to a fixed group of similar records (the blocks) and then each block is compared with each record of the other block, and the duplicated records are determined. This process is iterated over the entire dataset. The duplicates obtained are pruned to get the clean dataset.

Blocking technique is approach that builds upon an equidistant blocking technique and the successive enlargement of blocks. In this method first creates blocks and then progressively extends a fine-grained blocking.

II. LITERATURE REVIEW

This system Near-uniform Range Partition Approach for Increased Partitioning in Large Database by Jie Song et al. allows Database partitioning technique which adopts divide and conquer method can efficiently simplify the complexity of managing massive data and improve the performance of the system. According to the "divide and conquer" method, the table is partitioned into several parts.

The system Top-k Set Similarity Joins by Chuan Xiao et al. proposed the problem of answering similarity join queries to retrieve top-k pairs of records ranked by their similarities. Traditional approaches for the similarity joins with a given threshold will have to make guesses on the similarity threshold and incur much redundant calculation. In this an efficient algorithm that computes the answers in a progressive manner, upper bound score and the K^{th} temporary result score are exploited to develop several optimizations and to improve the space and time efficiencies of the algorithm. This algorithm provides the top-K pair records ranked by their similarities by eliminating guess work of users.

Framework for Evaluating Clustering Algorithms in Duplicate Detection by Oktie Hassanzaeh et al. proposed approach for duplicate detection using clustering framework. In this approach entity resolution is used as a part of the data cleaning process to identify records that possibly refer to the same real-world entity. It provides an evaluation framework for understanding what hurdles remain towards the goal of truly scalable and general purpose duplication detection algorithms. It generate results using partitioning of the similarity graph which is the common approach in many early duplicate detection techniques, confirms the common wisdom that this scalable approach results in poor quality of duplicate groups and it also shows that this quality is poor even when compared to other clustering algorithms that are as efficient. However this approach will work only in sequential manner, due to this time required for duplicate detection is large as compared to other state-of-art approaches and results are also not satisfactory.

The system Parallel Sorted Neighborhood Blocking with Map Reduce by Lars Kolb et al. proposed approach for identifying replication of data using Map Reduce technique. In this record linkage is applied to determine all entities that are referring to the same real world object and it also shows how entity resolution workflows with a blocking strategy and a match strategy can be realized with Map Reduce. Normally it is focusing on Sorted neighborhood blocking and proposed two Map Reduce-based implementations. This algorithm works fine and gives satisfactory results. However there are some limitations of Map Reduce technique.

III. MOTIVATION

In very large data sets, like malls and website data, the data is difficult to manage and if there are duplicate entries it is all the more time consuming.

The users generally clean data in very short time and that too without bothering about the technical specification of the data. Therefore there is need to perform a process that will effectively remove the document duplication and retrieve the clean data.

IV. SYSTEM ARCHITECTURE

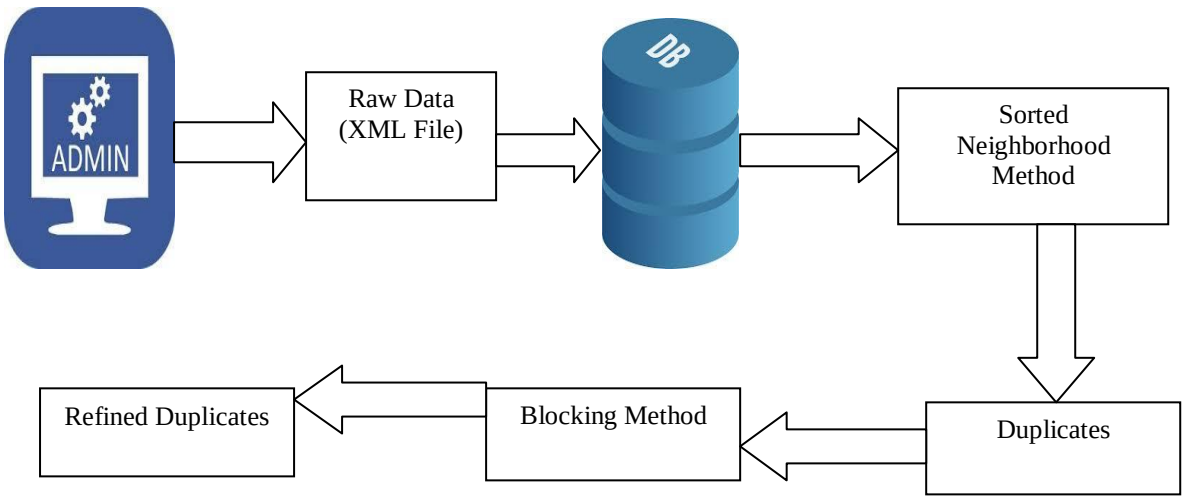

Fig1. Architecture of duplication detection

The above diagram illustrates architecture of duplication detection. In the initialization phase convert the xml data into the database. Then apply the Sorted neighborhood method on the database. As a result of this method, obtain duplicate records. For getting refined duplicated records apply Blocking method, in this method the comparisons is takes place between the blocks. And get the refined duplicate records.

.

REFERENCES

[1] O. Hassanzadeh, F. Chiang, H. C. Lee, and R. J. Miller, "Framework for evaluating clustering algorithms in duplicate detection," in Proceedings of the International Conference on Very Large Databases (VLDB), 2009.

[2] S. Yan, D. Lee, M. yen Kan, and C. L. Giles, "Adaptive sorted neighborhood methods for efficient record linkage," in International Conference on Digital Libraries (ICDL), 2007.

[3] Jie Song, Yu-bin Bao "Near-uniform Range Partition Approach for Increased Partitioning in Large Database",2010.

[4] U. Draisbach and F. Naumann, "A generalization of blocking and windowing algorithms for duplicate detection." in International Conference on Data and Knowledge Engineering (ICDKE), 2011.

[5] H. B. Newcombe and J. M. Kennedy, "Record linkage: making maximum use of the discriminating power of identifying information," Communications of the ACM, vol. 5, no. 11, 1962

[6] 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.

[7] J. Madhavan, S. R. Jeffery, S. Cohen, X. Dong, D. Ko, C. Yu, and A. Halevy, "Web-scale data integration: You can only afford to pay as you go," in Proceedings of the Conference on Innovative Data Systems Research (CIDR), 2007.

[8] S. R. Jeffery, M. J. Franklin, and A. Y. Halevy, "Pay-as-you-go user feedback for dataspace systems," in Proceedings of the International Conference on Management of Data (SIGMOD), 2008.

[9] O. Hassanzadeh and R. J. Miller,"Creating probabilistic databases from duplicated data," VLDB Journal, vol. 18, no.5, 2009.

[10] F. Naumann and M. Herschel, An Introduction to Duplicate Detection. Morgan & Claypool, 2010.

[11] U. Draisbach, F. Naumann, S. Szott, and O. Wonneberg, "Adaptive windows for duplicate detection," in Proceedings of the International Conference on Data Engineering (ICDE), 2012.

[12] P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 24, no. 9, 2012.

