Duplicate Detection using Algorithms

Elizabeth Priya Mammen  
PG Student  
Department of Computer Science & Engineering  
Mount Zion College of Engineering, Kadammanitta, Pathanamthitta, Kerala

Sreelakshmi R  
Assistant Professor  
Department of Computer Science & Engineering  
Mount Zion College of Engineering, Kadammanitta, Pathanamthitta, Kerala

Abstract

Data are most important asset. If there are more number of data that data may contain some duplicates. For finding that duplicates we use some algorithms like Sorted Neighborhood Method (SNM), Progressive Sorted Neighborhood Method (PSNM) etc. Duplicate count strategy++ (DCS++) algorithm is the best and efficient algorithm for finding the duplicates in the data sets. By using the DCS++ algorithm we can find the distinct data.

Keywords: Data, DCS++, PSNM, SNM

I. INTRODUCTION

Progressive duplicate detection is a pay-as-you-go method that identifies most duplicate pairs timely in the process. Instead of dropping the overall time needed to finish the entire duplicate detection task, progressive methods try to decrease the average time after which a subjective duplicate is found. The aim of progressive duplicate detection is to recognize most matches early in the process. Therefore, this algorithm increases the efficiency of the duplicate detection workflow and not the class of the reported results. Usually, windowing techniques are used to reduction the difficulty of the duplicate detection process. Therefore, ending a progressive algorithm prematurely might be sensible to find slightly less duplicates in a much shorter time.

Hence, the progressive styles increase the efficiency particularly for use case situations that do not need the whole result. In evaluation to traditional duplicate detection, I can define progressive duplicate detection with the subsequent two criteria.

Improved Early Quality - May be an arbitrary objective time at which results are needed. Then the output of a progressive algorithm will be greater at t than the output of an equivalent traditional algorithm. Typically, t is much smaller than the complete runtime of the traditional algorithm.

Same Eventual Quality - If together a traditional algorithm and its progressive version finish implementation without early conclusion at t, then they will output the identical results.

II. EXISTING SYSTEM

In the existing system Sorted Neighborhood Method (SNM) is used for finding the duplicates in the dataset. SNM has 3 phases

1) Key selection: sorting key is assigned to each record. The key is generated by concatenating two or more values of attributes.

2) All records are sorted according to key.

3) Slides a window over sorted data within particular window all records pairs are compared and duplicates are marked.

A disadvantage of the sorted neighborhood method is fixed window size. Some duplicates might be missed when selected window size in too small. On other hand, unnecessary comparison is carried out when window size is too large. The efficient detection of duplicate records in challenging work because database cleaning is very complicated process.

III. PROPOSED SYSTEM

The proposed system is used to find duplicates in the large dataset. Duplicate Count Strategy++ (DCS++) algorithm is used in the proposed system. DCS++ is an efficient method for finding duplicates in the dataset. By using this we get distinct dataset. In DCS++ the dataset is sorted by a key and then it will slide a window over all the records and find all the duplicates in the dataset. In DCS++ the window size is variable.
**A. Architecture**

First system should upload records from database. After uploading these records are sorted by using sorting technique. From these sorted records duplicate records are detected by using progressive sorted neighborhood method (PSNM) and duplicate count strategy++ (DCS++).

In PSNM compares the records that are within a window of records in the sorted order that are already partitioned. The PSNM algorithm uses this procedure iteratively without varying the window size. PSNM dynamically changing the execution order of the comparison based on intermediate results. After completing this procedure using PSNM we can display the duplicates records in the uploaded input.

In DCS++ compares the records that are within a window of records in the sorted order. In DCS++ algorithm uses this procedure iteratively by variable window size. If there any duplicate records are identified the window size will be incremented. There should not be any partition is required. For each detected duplicates the next records of that duplicate are added to the window. The detected duplicate records are skipped for save comparisons. It does not miss any duplicate records from the dataset. DCS++ algorithm is the best and efficient method for finding duplicates in the dataset.

**B. System Architecture.**

This project is used to find the duplicates in the project. In existing system we use progressive duplicate detection is used for conclusion the duplicates in the data set. Progressive sorted neighborhood method (PSNM) algorithm is used. In this method the datasets are partitioned and then checks the copies in the dataset. The disadvantage in this technique is the checking is done within the partition only, between the partitions here is no checking is take place. Because of that occasionally the duplicates may waste. The disadvantage in progressive method is overcome by spending an adaptive technique for duplicate detection. For this duplicate count strategy ++ (DCS++) algorithm is used for discovery the duplicates in the dataset.

**IV. Conclusion**

This project is used to find the duplicates in the project. In existing system we use progressive duplicate detection is used for conclusion the duplicates in the data set. Progressive sorted neighborhood method (PSNM) algorithm is used. In this method the datasets are partitioned and then checks the copies in the dataset. The disadvantage in this technique is the checking is done within the partition only, between the partitions here is no checking is take place. Because of that occasionally the duplicates may waste. The disadvantage in progressive method is overcome by spending an adaptive technique for duplicate detection. For this duplicate count strategy ++ (DCS++) algorithm is used for discovery the duplicates in the dataset.

**REFERENCES**


