Abstract

Problems that are often found in the data is data inconsistency, duplication of data, human errors, or data that is broken when storing the data. This results in overlapping data. Therefore a way is needed to minimize problems with data, one way is to perform data cleaning. Data Cleaning is the act of detecting and correcting (or removing) a number of records, tables, and databases that are less or not accurate. then those problems that was found is going to be replaced, modified or deleted from the database.

In this final task, a system is developed to do data cleaning to identify duplication in data. By using the Multi-Pass Neighborhood, the records which are duplicate will be identified, then those pairs of duplicate record would be compared with another records, to get the consistent data, which are called clean data. The testing phase was done to see the quality of the identification based on the value of recall and false-positive value.

Based on the testing that was done, the methods of the Multi-Pass Neighborhood can generate a good recall and false-positive value based on the window width parameter, the combination rule and the number of passes used in this final task

Keyword : *Data Cleaning, Multi-Pass Neighborhood, Identification duplicate data*