Abstrak

Beberapa masalah yang sering ditemukan pada data adalah ketidakkonsistenan data, duplikasi data, *human errors*, atau mungkin data telah rusak pada penyimpanan data. Hal ini menyebabkan overlapping atau data yang tumpang tindih. Untuk itu diperlukan cara untuk meminimalisir masalah pada data, salah satu caranya adalah data cleaning. Data cleaning adalah sebuah langkah untuk mendeteksi dan mengkoreksi (atau menghapus) sejumlah record, tabel, dan database yang kurang atau tidak akurat, setelah itu masalah – masalah yang ditemukan akan diganti, dimodifikasi atau dihapus dari database.

Pada tugas akhir ini dikembangkan suatu sistem untuk melakukan data cleaning dalam meng identifikasi duplikasi pada data. Dengan menggunakan metoda Multi-Pass Neighborhood, akan mengidentifikasi record yang duplikat pada database lalu record tersebut akan dibandingkan record lain untuk mendapatkan record yang konsisten. Pengujian dilakukan untuk melihat kualitas hasil identifikasi berdasarkan nilai recall dan nilai false-positive.

Berdasarkan pengujian yang sudah dilakukan, metoda Multi-Pass Neighborhood dapat menghasilkan nilai recall dan false-positive yang cukup baik dengan parameter ukuran lebar window, kombinasi rule dan jumlah passes yang digunakan.

Kata kunci : *Data Cleaning, Multi-Pass Neighborhood*, identifikasi data duplikat