ABSTRACT

We present two algorithms for calculating string Dis-Similarity percentage of De-duplication system. Our algorithms are multiple levels of clustering to incorporate constraints for reducing the volume of data and Information Gain (IG) for calculating Dis-Similarity. In our proposed system, we will first separate the records into block sized subset by using clustering algorithm and applying the subset value to IG. Most of the existing algorithm systems depend on generic or manually tuned distance metrics for estimating the similarity. We ran extensive experiments with huge data and compared them with various versions of existing algorithms and reported that the new system reduces the time consumption for string comparison and higher average accuracy than that of the existing systems.

None of the existing system produces the dis-similarity percentage between pair of string in given data set. Here we have presented an efficient solution for calculating string dis-Similarity percentage of De-duplication system by using Multi Level Clustering (MLC), Information Gain. Our algorithms work in two phases: Multi Level Clustering construction and Text Dis-Similarity calculation. Our methods reduce the time consumption for finding a duplicate record and using smaller amount of memory than the existing method.
The goal of this work is to identify groups of similar entities in the presence of linked environment and searching methods should reduce the number unwanted comparison during de-duplication. It will maximize the performance of data de-duplication. But de-duplication and group detection of similar entities have mostly been dealt with separately and as unrelated problem. We argue that the two problems occur for most real world application and this calls complete frame work for addressing them. In order to achieve this goal, in this research work I have proposed a new technique.

A turnaround measure is significant between the duplicate detection with grouping and duplicate detection without grouping. It clearly shows duplicate detection with grouping is much better than duplicate detection without grouping. And shown duplicate detection with grouping has taken less time for Dis-Similarity calculation with Single attribute and single Token with respect to any volume of data. Our executed result clearly shows for calculating dis-similarity value is more accurate irrespective of the token count. If we increase the dataset volume the performance will increase for with grouping and decrease for without grouping and there is no relevance for any number of attribute and token.