Abstract

Deduplication is the process of determining all categories of information within a data set that signify the same real life / world entity. The data gathered from various resources may have data high quality issues in it. The concept to identify duplicates by using windowing and blocking strategy. The objective is to achieve better precision, good efficiency and also to reduce the false positive rate all are in accordance with the estimated similarities of records. Various Similarity metrics are commonly used to recognize the similar field entries. So the main focus of this paper is to applying appropriate similarity measure on appropriate data to properly identifying the duplicates.
Validation of Deduplication in Data using Similarity Measure


Index Terms

Computer Science

Information Sciences
Keywords

Deduplication; Similarity Measure; Sorted Neighborhood Method (SNM); Windowing; Blocking.