Abstract

The purpose of this research was to review, analyze and compare algorithms lying under the empirical technique in order to suggest the most effective algorithm in terms of efficiency and accuracy. The research process was initiated by collecting the relevant research papers with the query of “duplication record detection” from IEEE database. After that, papers were categorized on the basis of different techniques proposed in the literature. In this research, the focus was made on empirical technique. The papers lying under this technique were further analyzed in order to come up with the algorithms. Finally, the comparison was performed in order to come up with the best algorithm i.e. DCS++. The selected algorithm was critically analyzed in order to improve its working. On the basis of limitations of selected algorithm, variation in algorithm was proposed and validated by developed prototype.

After implementation of existing DCS++ and its proposed variation, it was found that the proposed variation in DCS++ producing better results in term of efficiency and accuracy. The algorithms lying under the empirical technique of duplicate records deduction were focused. The
An Efficient Duplication Record Detection Algorithm for Data Cleansing

research material was gathered from the single digital library i.e. IEEE. A restaurant dataset was selected and the results were evaluated on the specified dataset which can be considered as a limitation of the research. The existing algorithm i.e. DCS++ and proposed variation in DCS++ were implemented in C#. As a result, it was concluded that proposed algorithm is performing outstanding than the existing algorithm.

References


Index Terms

Computer Science Information Sciences

Keywords

Duplication Records Detection Algorithm, DCS++, Windowing, Blocking