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. De-duplication is a property which provides additional information of similarities between the two entities. Thus, in today's data centric environment there are
huge numbers of defects in similarity measure. As a result to identify the duplicates is always been a challenging task. In this paper the primary focus is given on exact identification of duplicates in the database by applying concept of windowing & blocking. 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.

References

- L. Leitao, P. Calado, and M. Herschel, "Efficient and Effective Duplicate Detection in Hierarchical Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 5, May 2013
- U. Draisbach and F. Naumann, "A comparison and generalization of blocking and windowing algorithms for duplicate detection", in Proceedings of the International.
Workshop on Quality in Databases (QDB), 2009.

Index Terms

Computer Science

Information Sciences

Keywords

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