Abstract

This paper presents a survey on data deduplication on large scale data. Deduplication is nothing but finding the duplicate records or duplicate data when compared with one or more data base or data sets. The data deduplication task has attracted a considerable amount of attention from the research community in order to provide effective and efficient solutions. Matching records from several data bases is known as record linkage. Those matched data contains important and useable information. These information is too costly to acquire because of which data deduplication process getting more attention day by day. Removing duplicate records during data cleaning process in a single database is a critical step, because the outcomes of subsequent data processing or data mining may get greatly influenced by duplicates. As database size increases day by day the matching process’s complexity becoming one of the major challenges for data deduplication.
To overcome this problem we propose a Two Stage Sampling Selection (T3S) model which has two stages, in which, the strategy is proposed to produce balanced subsets candidate pairs which are to be labeled is done in the first stage and in the second stage we produced a smaller and more informative training sets than in the first stage. An active selection is incrementally invoked for removing the redundant pairs which are created in the first stage. This training set can be effectively used for identifying where the most ambiguous pairs lie and to configure the classification approaches. When compared with state-of-the-art deduplication methods in large datasets, our evaluation shows that T3S is able to reduce the labeling effort substantially while achieving a competitive or superior matching quality.

References

3. R. J. Bayardo, Y. Ma, and R. Srikant, Scaling up all pairs similarity search, proceedings of 16th international conference in world wide web(2007)131-140

Index Terms
Keywords

Deduplication, T3S, FS-Dedup