Abstract

Today's important task is to clean data in data warehouses which has complex hierarchical structure. This is possibly done by detecting duplicates in large databases to increase the efficiency of data mining and to make it effective. Recently new algorithms are proposed that consider relations in a single table; hence by comparing records pairwise they can easily find out duplications. But now a day the data is being stored in more complex and semi-structured or hierarchical structure and the problem arose is how to detect duplicates on XML data. Also due to differences between various data models, the algorithms which are for single relations cannot be applied on XML data. The objective of this project is to detect duplicates in hierarchical data which contain textual data and multimedia data like images, audio and video. It also focuses on eliminating the duplicates by using elimination technique such as delete. Here Bayesian network is used with modified pruning algorithm for duplicate detection, and experiments are performed on both artificial and real world datasets. The new XMLMultiDup method is able to perform duplicate detection with high efficiency and effectiveness on multimedia datasets. This method compares each level of XML tree from root to the leaves computing probabilities of similarity by assigning weights. It goes through the comparison of structure, each descendant of both datasets and find duplicates despite difference in data.
Efficient Duplicate Detection and Elimination in Hierarchical Multimedia Data

References

- http://www.hpi.uni-potsdam.de/naumann/projekte/repeatability/.
- https://www.hpi.uni-potsdam.de/fileadmin/hipi/FG_Naumann/
Efficient Duplicate Detection and Elimination in Hierarchical Multimedia Data

- http://www.morganclaypool.com/doi/abs/10.2200/S00262ED1V01Y201003DTM003
- http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.70.8263&rep=rep1&type=pdf#page=14
- http://www2.cs.uni-paderborn.de/cs/ag-boettcher/lehre/SS05/sem-ss05/SIGMOD05

Index Terms

Computer Science
Multimedia

Keywords
XML Data
Bayesian network
Pruning.