Abstract

The goal of the data mining process is to extract information from various data sources. Different sources can provide documents that contain data with different structure may be considered as representing the same conceptual information. Solution to this is duplication detection. Duplicate detection is detection of same real world entity in the data sources. Duplicate detection is a necessary task in data cleansing. Various algorithms are proposed for detection of duplicates in relational data, but very few solutions are focused on hierarchical data like XML. Duplicate Detection exactly identifies whether the data is duplicated or not. A peculiar method XMLDup is introduced for duplicate detection in XML data. XMLDup uses Bayesian network to evaluate probability of two XML elements being duplicates. It considers not only the content within the elements but also the way that content is structured. To improve
the run time efficiency of network evaluation, a lossless pruning strategy is used. The algorithm achieves high accuracy and recall score in several data sets. The XMLDup perform state-of-the-art in duplicate detection in terms of both effectiveness and efficiency.

References


Index Terms

Computer Science Algorithms

Keywords

Duplicate Detection Xml Bayesian Networks Data Cleaning And Optimization.