Abstract

The proliferation of web data sources increasingly demands the integration of these sources. To facilitate the integration process, a pre-analysis step is required to classify and group data sources into their correct domains. In this paper, we propose a feature-based clustering approach for clustering web data sources without any human intervention and based only on features extracted from the source schemas. In particular, we make use of both linguistic and structural schema features. We experimentally demonstrate the effectiveness of the proposed approach in terms of both the clustering quality and runtime.

References


Index Terms

Computer Science

Algorithms

Keywords

Web data source  Data integration  Clustering  Performance