Abstract

This paper presents a robust unsupervised approach for extraction of data records from dynamic web pages using tag tree comparison. Extracting data records from the web pages involves following sequences. We first download the related web pages of interest on our system. Next we construct DOM trees for those pages using a parser. We then compare two or more web pages to eliminate the noisy unwanted data such as header, menu bar, navigation bar, advertisements, etc and find the region of interest called Data region or Object region. We then traverse subtrees of data region to detect individual data record and pull them in the XML file. The main contribution of this paper is in developing a fully unsupervised approach for extracting structured as well as semi-structured data records from the web pages. Our proposed system can extract data records from many commercial websites more precisely. Hence it can serve as a source for integrating information from various web sources which can be used for providing value added services such as comparative shopping, market intelligence, meta-querying and search.
Data Record Extraction using Tag Tree Comparison

- Zhai, Y., & Liu, B. 2006. Structured data extraction from the web based on partial tree alignment. Knowledge and Data Engineering, IEEE Transactions on, 18(12), 1614-1628.

Index Terms

Computer Science

Database Management

Systems
Keywords
Data Record Detection  Information Extraction  Automatic Extraction  Web Mining
Semi-Structured data
Wrapper Generation.