Abstract

The data available on the web is so voluminous and heterogeneous. Deep Web, contains magnitudes more and valuable information than the surface Web. Deep Web contents are accessed by queries submitted to Web databases and the returned data records are enwrapped in dynamically generated Web pages. A large number of techniques have been proposed to address this problem, but all of them are Web-page programming-
language-dependent. In this paper we reviewed a novel vision-based approach that is Web-page programming-language-independent. ViDE utilizes the visual features on the deep Web pages to implement deep Web data extraction, including data record extraction and data item extraction. Our experiments on a large set of Web databases show that the proposed vision-based approach is highly effective for deep Web data extraction.

Reference

- Testbed for Information Extraction from Deep Web at:
- A vocabulary and associated APIs for HTML and XHTML at:

**Index Terms**

Computer Science
Ubiquitous Computing

**Key words**

Deep Web Data mining
Data Extraction
Visual Features