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

- G.O. Arocena and A.O. Mendelzon, “WebOQL: Restructuring Documents, Databases,
- D. Cai, S. Yu, J. Wen, and W. Ma, “Extracting Content Structure for Web Pages Based
Extraction Systems,” IEEE Trans. Knowledge and Data Eng., vol. 18, no. 10, pp. 1411–1428,
- C.-H. Chang, C.-N. Hsu, and S.-C. Lui, “Automatic Information Extraction from
Semi-Structured Web Pages by Pattern Discovery,” Decision Support Systems, vol. 35, no. 1,
- V. Crescenzi and G. Mecca, “Grammars Have Exceptions,” Information Systems, vol. 23,
- V. Crescenzi, G. Mecca, and P. Merialdo, “RoadRunner: Towards Automatic Data
- D.W. Embley, Y.S. Jiang, and Y.-K. Ng, “Record-Boundary Discovery in Web
Experience,” Proc. East-European Workshop Advances in Databases and Information Systems
(ADBIS), pp. 1–8, 1997.
- Testbed for Information Extraction from Deep Web at:
- A vocabulary and associated APIs for HTML and XHTML at:
- N. Kushmerick, “Wrapper Induction: Efficiency and Expressiveness,” Artificial

Index Terms

Computer Science
Ubiquitous Computing

Key words
Deep Web Data mining
Data Extraction
Visual Features