National Technical Symposium on Advancements in Computing Technologies

Number 1 - Article 1

Year of Publication: 2011

Authors:

Snehal M. Shewale

Trupti S. Patil

{bibtex}ntst034.bib{/bibtex}

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-pageprogramming-

language-dependent. In this paper we reviewed a novel vision-based approach that is Web-pageprogramming- 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, and Webs,” Proc. Int’l Conf. Data Eng. (ICDE), pp. 24–33, 1998.
- D. Buttler, L. Liu, and C. Pu, “A Fully Automated Object Extraction System for the World Wide Web,” Proc. Int’l Conf. Distributed Computing Systems (ICDCS), pp. 361-370, 2001.
- D. Cai, X. He, J.-R. wen, and W.-Y. Ma, “Block-Level Link Analysis,” Proc. SIGIR, pp. 440–447, 2004.
- D. Cai, S. Yu, J. Wen, and W. Ma, “Extracting Content Structure for Web Pages Based on Visual Representation,” Proc. Asia Pacific Web Conf. (APWeb), pp. 406–417, 2003.
- C.-H. Chang, M. Kayed, M.R. Girgis, and K.F. Shaalan, “A Survey of Web Information Extraction Systems,” IEEE Trans. Knowledge and Data Eng., vol. 18, no. 10, pp. 1411–1428, Oct. 2006.
- 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, pp. 129– 147, 2003.
- V. Crescenzi and G. Mecca, “Grammars Have Exceptions,”Information Systems, vol. 23, no. 8, pp. 539-565, 1998.
- V. Crescenzi, G. Mecca, and P. Merialdo, “RoadRunner: Towards Automatic Data Extraction from Large Web Sites,” Proc. Int’l Conf. Very Large Data Bases (VLDB), pp. 109–118, 2001.
- D.W. Embley, Y.S. Jiang, and Y.-K. Ng, “Record-Boundary Discovery in Web Documents,” Proc. ACM SIGMOD, pp. 467– 478, 1999.
- W. Gatterbauer, P. Bohunsky, M. Herzog, B. Krpl, and B. Pollak, “Towards Domain Independent Information Extraction from Web Tables,” Proc. Int’l World Wide Web Conf. (WWW), pp. 71–80, 2007.
- J. Hammer, J. McHugh, and H. Garcia-Molina, “Semistructured Data: The TSIMMIS Experience,” Proc. East-European Workshop Advances in Databases and Information Systems (ADBIS), pp. 1–8, 1997.
- C.-N. Hsu and M.-T. Dung, “Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web,” Information Systems, vol. 23, no. 8, pp. 521–538, 1998.
- Testbed for Information Extraction from Deep Web at: http://daisen.cc.kyushu-u.ac.jp/TBDW/, 2009.
- A vocabulary and associated APIs for HTML and XHTML at: http://www.w3.org/html/wg/html5/, 09.
- N. Kushmerick, “Wrapper Induction: Efficiency and Expressiveness,” Artificial Intelligence, vol. 118, nos. 1/2, pp. 15–68, 2000.
- A. Laender, B. Ribeiro-Neto, A. da Silva, and J. Teixeira, “A Brief Survey of Web Data

Extraction Tools," SIGMOD Record, vol. 31, no. 2, pp. 84–93, 2002.

- B. Liu, R.L. Grossman, and Y. Zhai, "Mining Data Records in Web Pages," Proc. Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 601–606, 2003.

Index Terms

Computer Science Ubiquitous Computing

Key words

Deep Web Data mining Data Extraction Visual Features