Abstract

World Wide Web has more and more online Web databases which can be searched through their Web query interfaces. The number of Web databases has reached 25 millions according to a recent survey. All the Web databases make up the deep Web (hidden Web or invisible Web). Often the retrieved information (query results) is enwrapped in Web pages in the form of data records. These special Web pages are generated dynamically and are hard to index by traditional crawler-based search engines, such as Google and Yahoo. This kind of special Web pages deep Web pages. Deep Web contents are accessed by queries submitted to Web databases and the returned data records are enwrapped in dynamically generated Web pages.
(they will be called deep Web pages in this paper). Extracting structured data from deep Web pages is a challenging problem due to the underlying intricate structures of such pages. Until now, a large number of techniques have been proposed to address this problem, but all of them have inherent limitations because they are Web-page-programming-language dependent. As the popular two-dimensional media, the contents on Web pages are always displayed regularly for users to browse. This motivates to seek a different way for deep Web data extraction to overcome the limitations of previous works by utilizing some interesting common visual features on the deep Web pages. The paper, a novel vision-based approach that is Web-page programming-language-independent is proposed. This approach primarily utilizes the visual features on the deep Web pages to implement deep Web data extraction, including data record extraction and data item extraction. We also propose a new evaluation measure revision to capture the amount of human effort needed to produce perfect extraction. Experiments on a large set of Web databases show that the proposed vision-based approach is highly effective for deep Web data extraction. The approach consists of four primary steps: Visual Block tree building, data record extraction, data item extraction, and visual wrapper generation. Visual Block tree building is to build the Visual Block tree for a given sample deep page using the VIPS algorithm. With the Visual Block tree, data record extraction and data item extraction are carried out based on our proposed visual features. Visual wrapper generation is to generate the wrappers that can improve the efficiency of both data record extraction and data item extraction. Highly accurate experimental results provide strong evidence that rich visual features on deep Web pages can be used as the basis to design highly effective data extraction algorithms.

References

- Wei Liu and Weiyi Meng, "Vision based approach for deep web data extraction" IEEE trans. on Knowledge and Data Engineering 2010.
- Gustavo O. Arocena, Alberto O. Mendelzon, "WebOQL: Restructuring Documents, Databases and Webs";
- Jayant Madhavan, Shawn R. Jeffery, Shirley Cohen, Xin (Luna) Dong, David Ko, Cong Yu, Alon Halevy, "Web-scale Data Integration: You can only afford to Pay As You Go";
- Ron Bekkerman and Andrew McCallum, "Disambiguating Web Appearances of People in a Social Network";
- James Caverlee, Ling Liu, and David Buttlar, "Probe, Cluster, and Discover: Focused Extraction of QA-Pagelets from the Deep Web"; IEEE2004
- Jer Lang Hong, "Deep Web Data Extraction"; IEEE2010
- Chia-Hui Chang, Chun-Nan Hsu, Shao-Chen Lui, "Automatic information extraction from semi-structured Web pages by pattern discovery"; 0167-9236/02/$ - see front matter D 2002 Elsevier Science B. V. All rights reserved. PII: S0167-9236(02)00100-8 E-mail addresses: chia@csie. ncu. edu. tw(C.-H. Chang), chunnan@iis. sinica. edu. tw(C.-N.
A Framework for Deep Web Data Extraction using Vision and Novel based Approach

Hsu), anyway@db.csie.ncu.edu.tw (S.-C. Lui).

- Bing Liu, Robert Grossman, Yanhong Zhai Kai Simon, Georg Lausen, "Mining Data Records in Web Pages" liub@cs.uic.edu, grossman@uic.edu, yzhai@cs.uic.edu SIGKDD '03, August 24-27, 2003, Washington, DC, USA Copyright 2003 ACM 1-58113-737-0/03/0008. $5.00.

- Bing Liu, Robert Grossman, Yanhong Zhai in "Mining Data Records in Web Pages" liub@cs.uic.edu, grossman@uic.edu, yzhai@cs.uic.edu, SIGKDD '03, August 24-27, 2003, Washington, DC, USA Copyright 2003 ACM 1-58113-737-0/03/0008. $5.00.

- Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, Clement Yu in "Annotating Structured Data of the Deep Web" {ylu0, haihe, hkzhao, meng}@cs.binghamton.edu, yu@cs.uic.edu, 1-4244-0803-2/07/$20.00 ©2007 IEEE.

Index Terms

Computer Science Data Mining

Keywords

Deep Web /invisible Web Deep Web Search Engine Web-page Programming Language