XPath-wrapper induction for data extraction

Tran N.-K., Pham K.-C., Ha Q.-T.
College of Engineering and Technology, Vietnam National University, Hanoi, Viet Nam; Computer Science Department, University of Illinois, Urbana-Champaign, United States

Abstract: The Web contains an enormous amount of information which is formatted for human beings. This makes it difficult for computer to extract relevant content from various sources. This paper presents an XPath-wrapper induction algorithm which leverages user queries and template-based sites for extracting structured information. Our experiments show average accuracy of 94%. © 2010 IEEE.

Index Keywords: Amount of information; Data extraction; Human being; Structured information; Template-based; User query; Wrapper induction; Natural language processing systems

Year: 2010
Source title: Proceedings - 2010 International Conference on Asian Language Processing, IALP 2010
Art. No.: 5681601
Page : 150-153
Link: Scopus Link
Correspondence Address: Tran, N.-K.; College of Engineering and Technology, Vietnam National University, Hanoi, Viet Nam; email: khanhtn09@vnu.edu.vn
Conference name: 2010 International Conference on Asian Language Processing, IALP 2010
Conference date: 28 December 2010 through 30 December 2010
Conference location: Harbin
Conference code: 83700
DOI: 10.1109/IALP.2010.33
Language of Original Document: English
Abbreviated Source Title: Proceedings - 2010 International Conference on Asian Language Processing, IALP 2010
Document Type: Conference Paper
Source: Scopus
Authors with affiliations:
• Tran, N.-K., College of Engineering and Technology, Vietnam National University, Hanoi, Viet Nam
• Pham, K.-C., Computer Science Department, University of Illinois, Urbana-Champaign, United States
• Ha, Q.-T., College of Engineering and Technology, Vietnam National University, Hanoi, Viet Nam
References:
• Crescenzi, V., Mecca, G., Merialdo, P., Roadrunner: Towards automatic data extraction from large web sites (2001) Proceedings of the 27th International Conference on Very Large Data Bases, pp. 109-118

• Kushmerick, N., Weld, D.S., Doorenbos, R., Wrapper induction for information extraction (1997) 15th International Joint Conferences on Artificial Intelligence, p. 246


• Craven, T.C., HTML tags as extraction cues for web page description construction (2003) Informing Science Journal, 6