A fast template-based approach to automatically identify primary text content of a web page

Nguyen D.Q., Nguyen D.Q., Pham S.B., Bui T.D.
Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam

Abstract: Search engines have become an indispensable tool for browsing information on the Internet. The user, however, is often annoyed by redundant results from irrelevant web pages. One reason is because search engines also look at non-informative blocks of web pages such as advertisement, navigation links, etc. In this paper, we propose a fast algorithm called FastContentExtractor to automatically detect main content blocks in a web page by improving the ContentExtractor algorithm. By automatically identifying and storing templates representing the structure of content blocks in a website, content blocks of a new web page from the website can be extracted quickly. The hierarchical order of the output blocks is also maintained which guarantees that the extracted content blocks are in the same order as the original ones. © 2009 IEEE.

Author Keywords: Data mining; Template detection; Web mining
Index Keywords: Fast algorithms; Hierarchical order; Indispensable tools; Template detection; Template-based; Text content; Web Mining; Web page; Information retrieval; Knowledge engineering; Search engines; Systems engineering; Websites

Year: 2009
Source title: KSE 2009 - The 1st International Conference on Knowledge and Systems Engineering
Art. No.: 5361702
Page : 232-236
Link: Scopus Link
Correspondence Address: Nguyen, D. Q.; Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam
Sponsors: College of Technology; Vietnam National University
Conference name: 1st International Conference on Knowledge and Systems Engineering, KSE 2009
Conference date: 13 October 2009 through 17 October 2009
Conference location: Hanoi
Conference code: 79895
DOI: 10.1109/KSE.2009.39
Language of Original Document: English
Abbreviated Source Title: KSE 2009 - The 1st International Conference on Knowledge and Systems Engineering
Document Type: Conference Paper
Source: Scopus
Authors with affiliations:

- Nguyen, D.Q., Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam
- Nguyen, D.Q., Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam
- Pham, S.B., Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam
- Bui, T.D., Human Machine Interaction Laboratory, College of Technology, Vietnam National University, Hanoi, Viet Nam

References:

- Gibson, D., Punera, K., Tomkins, A., The volume and evolution of web page templates (2005) Special Interest Tracks and Posters, 14th Int. Conf. on WWW, pp. 830-839