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: Scorpus 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
ISBN: 9.78E+12
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:

- Arasu, A., Garcia-Molina, H., University, S., Extracting structured data from web pages (2003) Proceedings of SIGMOD, pp. 337-348
- Kolcz, A., Yih, W., Site-independent template-block detection (2007) Proceedings of PKDD, pp. 152-163
- Cai, D., Yu, S., Wen, J.-R., Ma, W.-Y., Vips: A vision-based page segmentation algorithm (2003) MSR-TR-2003-79, Microsoft
- 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
- Song, R., Liu, H., Wen, J.-R., Ma, W.-Y., Learning block importance models for Web pages (2004) Thirteenth International World Wide Web Conference Proceedings, WWW2004, pp. 203-211. , Thirteenth International World Wide Web Conference Proceedings, WWW2004
- Hsu, J.Y., Yih, W., Template-based information mining from HTML documents (1997) Proceedings of AAAI-97, pp. 256-262. , AAAI Press
- Lerman, K., Getoor, L., Minton, S., Knoblock, C., Using the structure of web sites for automatic segmentation of tables (2004) Proceedings of SIGMOD, pp. 119-130
- Vieira, K., Da Silva, A.S., Pinto, N., De Moura, E.S., Cavalcanti, J.M.B., Freire, J., A fast and robust method for web page template detection and removal (2006) International Conference on Information and Knowledge Management, Proceedings, pp. 258-267. , DOI 10.1145/1183614.1183654, Proceedings of the 15th ACM Conference on Information and Knowledge Management, CIKM 2006
- Yi, L., Liu, B., Li, X., Eliminating noisy information in web pages for data mining (2003) Proceedings of 9th KDD, pp. 296-305
- Mehta, R., Madaan, A., Web page sectioning using regex-based template (2008) Proceedings of 17th WWW, pp. 1151-1152
- Debnath, S., Mitra, P., Pal, N., Giles, C.L., Automatic identification of informative sections of web pages (2005) IEEE Transactions on Knowledge and Data Engineering, 17 (9), pp. 1233-1246. , DOI 10.1109/TKDE.2005.138
- Debnath, S., Mitra, P., Lee Giles, C., Automatic extraction of informative blocks from webpages (2005) Proceedings of the ACM Symposium on Applied Computing, 2, pp. 1722-1726. , DOI 10.1145/1066677.1067065, Applied Computing 2005 - Proceedings of the 20th Annual ACM Symposium on Applied Computing
- Lin, S.H., Ho, J.-M., Discovering informative content blocks from web documents (2002) Proceedings of KDD, pp. 588-659. , ACM
- Xiao, X., Luo, Q., Xie, X., Ma, W.-Y., A comparative study on classifying the functions of web page blocks (2006) International Conference on Information and Knowledge Management, Proceedings, pp. 776-777. , DOI 10.1145/1183614.1183725, Proceedings of the 15th ACM Conference on Information and Knowledge Management, CIKM 2006
- Wang, Y., Fang, B., Cheng, X., Guo, L., Xu, H., Incremental web page template detection (2008) Proceedings of 17th WWW, pp. 1247-1248
- Bar-Yossef, Z., Rajagopalan, S., Template detection via data mining and its applications (2002) Proceedings of 11th WWW, pp. 580-591