Abstract

A Web page has huge information and the information in the Web pages is useful in real world applications. The additional contents in the Web page like links, footers, headers and advertisements may cause the content extraction to be complicated. Irrelevant content in the Web page is treated as noisy content. A method is necessary to extract the informative content and discard the noisy content from Web pages. An integration of textual and visual importance is used to extract the informative content from Web pages. Initially a Web page is converted in to DOM (Document Object Model) tree. For each node in the DOM tree, textual and visual importance is calculated. Textual importance and visual importance is combined to form hybrid density. Density sum is calculated and used in content extraction algorithm to extract the informative content from Web pages. Performance of Web content extraction is obtained by calculating precision, recall, f-measure and accuracy.
Web Content Extraction by Integrating Textual and Visual Importance of Web Pages


- Dandan Song, Fei Sun, Lejian Liao. &quot;A hybrid approach for content extraction with text density and visual importance of DOM nodes&quot;. In the proceedings of Springer Knowl Inf Syst, DOI 10.1007/s10115-013-0687-x, Verlag London 2013.


- Lan Yi, Bing Liu, Xiaoli Li. &quot;Eliminating Noisy Information in web pages for Data Mining&quot; . In the Proceedings of ACM 1-58113-737-0/03/0008, SIGKDD . 03, August 24-27, 2003, Washington, DC, USA

- Liang Chen, Shaozhi Ye, Xing Li. &quot;Template Detection for Large Scale Search Engines&quot; . In the proceedings of ACM 1-59593-108-2/06/0004SAC apr;06 April 23-27, 2006, Dijon, France.
- Pinto D, Branstein M, Coleman R, Croft WB, King M, Li W, Wei X (2002) QuASM: a system
Web Content Extraction by Integrating Textual and Visual Importance of Web Pages

for question answering using semi-structured data. In: Proceedings of JCDL ’02, pp 46–55

Index Terms

Computer Science

Information Sciences

Keywords

Web Content Extraction    Web content Mining    DOM tree    Vision based Page Segmentation.