Abstract

As we know internet detailed web pages contains information which are not considered as primary content such as advertisements, headers, footers, navigation links and copyright information. Also information on web pages such as comments and reviews are not preferred by search engines to index as informative content, thereby having an algorithm to extracts only main content could help better quality on web page indexing. Almost all algorithms have been proposed are tag dependent means they could only look for primary content among specific tags such as \texttt{div} or \texttt{table}. The algorithm in this paper simulates a web page user visit and how the user finds the main content block position in the page. The proposed method is tag independent and has two phases to accomplish the extraction job. First it transforms input DOM tree obtained from input HTML detailed web page into a block tree based on their visual representation and DOM structure in a way that on every node it will have specification vector, then it traverses the obtained small block tree to find main block having dominant computed value in comparison with other block nodes based on its specification vector values. The
introduced method doesn’t have any learning phases and could find informative content on any random input detailed web page. This method has been tested in large variety of websites and as we will show, it gains better precision and recall based on other compared method K-FE.

Reference

- Deng Cai, Xiaofei He, Ji-Rong Wen and Wei-Ying Ma: Block Level Link Analysis. In: Proc. 2004 Int. Conf. on Research and Development in Information Retrieval (SIGIR’04), Sheffield, UK (July 2004).
- Hwanjo Yu, AnHai Doan, and Jiawei Han: Mining for Information Discovery on the Web: Overview and Illustrative Research. In: Intelligent Technologies for Information Analysis, edited by Ning Zhong, Springer-Verlag, invited paper, pp. 135-168 (2004).
Index Terms

Computer Science

Information Retrieval

Key words

Web mining

Noise elimination

Informative content

Information retrieval

Information extraction