Abstract

Today internet has made the life of human dependent on it. Almost everything and anything can be searched on net. Web pages usually contain huge amount of information that may not interest the user, as it may not be the part of the main content of the web page. To extract the main content of the web page, data mining techniques need to be implemented. A lot of
research has already been done in this field. Current automatic techniques are unsatisfactory as their outputs are not appropriate for the query of the user. In this paper, we are presenting an automatic approach to extract the main content of the web page using tag tree & heuristics to filter the clutter and display the main content. Experimental results have shown that the technique presented in this paper is able to outperform existing techniques dramatically.

Reference

- J. Hammer, H. Garcia-Molina, J. Cho, A. Crespo, R. Aranha, “Extracting semi structure information from the
A Heuristic Approach for Web Content Extraction

- Liu, B., Zhai, Y., “NET- “A System for extracting Web Data From Flat and Nested Data Records”, WISE-05 (Proceeding of 6th International Conference on Web Information System Engineering), 2005

Index Terms

Computer Science, Information Retrieval

Key words

HTML Parser, Tag Tree, Web Content Extraction, Heuristics