Abstract

In the huge network of World Wide Web, web pages contained large amount of information. Web researches are always requiring main content (e.g., an article text) from the web pages to be gathered, processed and stored quickly and efficiently. Mining the data on the Web has become a major task for locating useful information from the Web. The Web information’s that are considered as useful information usually has huge amounts of noise data’s such as navigation bars, links, advertisements, copyright notices etc. Performance of Web mining can be improved by identifying and removing noises from Web pages. In this paper new method is proposed for removing noise content tag and extracts the information of main content tag from web pages.

References

2. Bing Liu, Web Data Mining (Exploring Hyperlinks, Contents, and Usage Data), Springer.
4. Thanda Htwe. Cleaning Various Noise Patterns in Web Pages for Web Data Extraction, International Journal of Network and Mobile Technologies ISSN 1832-6758 Electronic Version VOL 1 / ISSUE 2 / NOVEMBER 2010 © 2010 INTI University College.(University of Computer Studies, Yangon, Department of Software Technology, tdhtwe80@gmail.com)

Index Terms
Keywords

WPNC, Noise Block, HTML Tag, White Listed tags, HDT, LDT, Black Listed Tags.