Abstract

With the phenomenal growth of the web, there is an ever increasing volume of data and information published in numerous web-pages. It is said that web is noisy. A web page typically contains a mixture of many kind of information e.g. main contains, advertisements, navigational panels, copy right blocks etc… for a particular application only part of information is useful and
the rest are noise. These all seriously harm web mining. Advertisements and Sponsor images are not much important in surfing. As there is a need of technique that keep common navigation structure as it is but removes image advertisement and improve surfing efficiency. In this paper a small application HTML Tag Differentiator is created which removes image advertisement using rule based classifier.

Reference

- Web Mining http://www.cs.ualberta.ca/~tszhu/webmining.htm
- Learning To remove Internet Advertisement, Nicholas Kashmerick 3rd Int. Conf. of Autonomous Agent, 1999
- Filtering the Web using WebFilter http://www.math.uni-paderborn.de/~axel/NoShit/
- Editorial: Special Issue on Web Content Mining SIGKDD Explorations. Volume 6, Issue 2, Page 1, 2003
- Why Block Ads? http://feh.net/adbloc/adbloc.shtml

Index Terms

   Computer Science   Information Retrieval

Key words

Image Advertisement         HTML Tag Differentiator
Rule Based Classifier
Web Content Mining