Abstract

The aim of this paper is to introduce a system based on web mining techniques to prevent spamming web pages. The system relies on content analysis, used features are Uniform Resource Locator(URL), Number of words in page Title, Globally Popular Keywords(GPK) and N-GRAM. The proposed system used Decision Tree(DT) rules ; which is the best classifier to detect Web spam content. It produces accuracy of .97 % in detecting spam web sites.

References


6. Maria Soledad Pera," A Structural, Content Similarity Measure for Detecting Spam Documents on the Web", https://pdfs.semanticscholar.org/.../2836c81e3c0d1802e76f1acee604.


Index Terms

Computer Science  Security

Keywords

Web Mining, Spam Web Sites, Decision Tree.