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

Hybrid Genetic Algorithm based Decision Tree " , International Journal of Computer
Applications, Volume 133 – No.10,pp 20-27
Different Features", International Journal of Soft Computing and Engineering, Volume-1,
Issue-3, pp 70-73.
5. Apostolis Zarras, Antonis Papadogiannakis, Sotiris Ioannidis and Thorsten Holz ,2015,"
Revealing the Relationship Network Behind Link Spam ", online at
6. Maria Soledad Pera," A Structural, Content Similarity Measure for Detecting Spam
Documents on the Web", https://pdfs.semanticscholar.org/.../2836c81e3c0d1802e76f1acee604.

Arabic Web Spam Detection System", International Journal of Advanced Computer Science and
8. Heider A. Wahsheh, Mohammed N. Al-Kabi and Izzat M. Alsmadi ,2013, " A link and
Content Hybrid Approach for Arabic Web Spam Detection ", Intelligent Systems and
Applications, Published Online December 2012 in MECS (http://www.mecs-press.org/),
pp30-43.
9. Mohammed N. Al-Kabi, Izzat M. Alsmadi and Heider A. Wahsheh, 2015, " Evaluation of
Spam Impact on Arabic Websites Popularity", Journal of King Saud University – Computer and
Information Sciences, pp 222–229.
10. Tarek Amr Abdallah and Beatriz de La Iglesia ,2015, "URL-Based Web Page
Classification: With n-Gram Language Models",Springer International Publishing Switzerland,
CCIS 553, pp. 19–33.
and Recall as Appearance Space Quality Measure for Simplified Aggregate Details",
Eurographics Symposium on Rendering.
p.219.
15. Powers, David M W ,2011, "Evaluation: From Precision, Recall and F-Measure to ROC,
Informedness, Markedness & Correlation", Journal of Machine Learning Technologies. vol.2
,pp. 37–63.
16. Vili Podgorelec, Ivan Rozman and Peter Kokol,2002," Decision Trees: An Overview and
Web Mining Techniques to Block Spam Web Sites

Index Terms

Computer Science  Security

Keywords

Web Mining, Spam Web Sites, Decision Tree.