Abstract

Web spam denotes the manipulation of web pages with the sole intent to raise their position in search engine rankings. Since a better position in the rankings directly and positively affects the number of visits to a site, attackers use different techniques to boost their pages to higher ranks. In the best case, web spam pages are a nuisance that provide undeserved advertisement revenues to the page owners. In the worst case, these pages pose a threat to Internet users by hosting malicious content and launching drive-by attacks against unsuspecting victims. When successful, these drive-by attacks then install malware on the victims machines. In this paper we introduce a clustering and classification approach to detect spam web pages in the list of results that are returned by a search engine. Initially, we apply K-nearest neighbor approach for clustering. And then we will apply K-means classification over those links for
categorizing them as either spam or non-spam links.

References

- Van Lam Le, Ian Welch, Xiaoying Gao, Peter Komisarczuk School of Engineering and Computer Science, Victoria University of Wellington P.O. Box 600, Wellington 6140, New Zealand “Two-Stage Classification Model to Detect Malicious Web Pages” 2011 International Conference on Advanced Information Networking and Applications
- Lourdes Araujo and Juan Martinez-Romo “Web Spam Detection: New Classification features Based on Qualified Link Analysis nd Language Models” ieee transactions on information forensics and security, vol. 5, no. 3, september 2010 581
- Thomas Largillier, Sylvain Peyronnet “Lightweight Clustering Methods for Webspam Demotion” 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology
- Chakrit Likitkhajorn, Athasit Surarerks, Amon Rungsawang “A Novel Approach for Spam Detection Using Boosting Pages” 20ll Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE)

Index Terms

Computer Science      Engineering and Technology

Keywords

Data mining   K-nearest neighbor   K-means algorithm   Spam and Non-spam links
Search Engine
An evolved classification and clustering approach for the detection of web spam