Abstract

The identification of web spam has been identified as a major challenge for web search engines. Spam web sites deliberately manipulate their placement by paying customers in search engine rankings. One of the techniques used by spammers is so-called link spam, where farms of interlinked web sites are used to give high PageRank to certain web. These link farms tend not to have any legitimate content and so do not have incoming links from sites outside the farm. Therefore, if one page within a link farm can be identified, we can reasonably suspect that any pages that point to it are also web spam.

BadRank is a method for detecting spam web sites, based on the premise that a page is spam if it points to another spam page; i.e., the BadRank score of a page is the weighted sum of the BadRank scores of the pages that it links to. BadRank method need to modified to make BadRank score is converge by ensure the matrix is stochastic . Additionally, we can consider methods for incorporating knowledge about trusted (known non-spam) sites into the BadRank calculation. In this final project used WEBSPAM-UK 2006 dataset to test BadRank with stochastic modified and trust

From the result testing we can analyze that for the datasset web spam uk 2006, badrank that modified with leafbadlinks with trust can detects spam better than another modification. And also badrank that added by trust variable more effective to detect web spam than without it.

Keywords: *Web Spam,Badrank,Link Farm*