CS239 Project Proposal: Web-spam Detection

Yong Kyun (Paul) Kwon and Ka Cheung (Richard) Sia
nasa@ucla.edu, kcsia@cs.ucla.edu

January 26, 2005

1 Introduction

Search engines play an important role in helping user to locate information on the Web nowadays. Having a higher rank in the result returned by popular search engines essentially draw a lot of traffic to the corresponding website. Thus, it is of high commercial value to make a website visible in the search engine's results. Consequently, spammers (or search engine optimizers, SEO) arise and create spam pages in order to fool the ranking method of search engines. These spam pages cause threat to both the search engine and users:

- If search engines continue to rank these malicious website high, which are usually related to pornography, pirated software, mortgage consolidation, medication (Viagra), popular consumer search (free iPod), users will get frustrated and depreciate the usability of the search engine.
- Users will get directed to these website, which, may be malicious in some sense and be convinced to reveal their personal information (phishing).

Spam pages are evolving like virus, spammers keep changing the techniques to fool search engines and avoid being detected. This is like a battle between search engines and spammers, thus, very few academic papers related to search engine ranking methods and spam techniques are published. Among some of them are: Gyongyi et.al. [3] uses a TrustRank propagation technique to assign trust-worthy score to webpages, Davison [1] and Fetterly et.al. [2] apply statistical techniques to detect both content and link spams. We consider spam page detection a valuable technology for both search engines and Internet user and worth investigating.

2 Methodology

In this project, we would like to:

- Survey on some of the common techniques used by spammers
- Implement one or two efficient approaches to detect spam pages
- Experiment the algorithms in a web repository to test the performance

The following are two techniques we would like to investigate:

Detecting web communities

Example: *http://newerabasketball.blogspot.com/*
Weblogs or guest-book entries are easily editable webpages, they are cheap (practically no cost) resource for spammers to create links pointing to their desired websites. These artificially created hyperlinks usually follow a specific pattern according to our preliminary studies, we would like to explore a scalable method in detecting these kind of link spam.

Detecting abnormal content

Example: *http://www.iosonline.org/free-cell-phone-no-credit-check.html*
This is an example spam page not suppose to be read by human, it exists simply to direct users to and boost up the ranking of this website *https://www.freecreditprofile.com/* . These pages try to exploit the popular PageRank method used in ranking search result. When we take a closer look at the page, the word "credit","free" occurs abundantly, which is very rare to be happened on a legitimate webpage. We propose to use efficient statistical techniques to filter out such spam pages.

3 Timeframe

- Week 4-5: Study both webspam and spam-detection techniques.
- Week 5-8: Implement proposed spam-detection algorithms.
- Week 8-10 Experiment on collection of webpages.

References

[1] Brain Davision. Recognizing Nepotistic Links on Web. In *Proceedings of the AAAI Workshop on Artificial Intelligence for Web Searc*, 2000.

[2] Dennis Fetterly, Mark Manasse, and Marc Najork. Spam, Damn Spam, and Statistics. In *Proceedings of the International WebDB Workshop*, 2004.

[3] Zoltan Gyongyi, Hector Garcia-Molina, and Jan Pedersen. Combating Web Spam with TrustRank. In *Proceedings of the International Conference on Very Large Data Bases*, 2004.