Detecting Webspam Beneficiaries Using Information Collected by the Random Surfer

Thomas Largillier, LRI, Université Paris-Sud, F-91405, France
Sylvain Peyronnet, LRI, Université Paris-Sud, F-91405, France

ABSTRACT

Search engines use several criteria to rank webpages and choose which pages to display when answering a request. Those criteria can be separated into two notions, relevance and popularity. The notion of popularity is calculated by the search engine and is related to links made to the webpage. Malicious webmasters want to artificially increase their popularity: the techniques they use are often referred to as Webspam. It can take many forms and is in constant evolution, but Webspam usually consists of building a specific dedicated structure of spam pages around a given target page. It is important for a search engine to address the issue of Webspam; otherwise, it cannot provide users with fair and reliable results. In this paper, the authors propose a technique to identify Webspam through the frequency language associated with random walks among those dedicated structures. The authors identify the language by calculating the frequency of appearance of k-grams on random walks launched from every node.

Keywords: Data Exploration, Heuristics, Information Filtering, Networks, Web Technologies

INTRODUCTION

The Web has grown so big (Alpert & Hajhaj, 2008) that users cannot afford the patience to go all over it, or even through a rather small part of it. Except for their favorites sites they have to (and they do) use search engines that answer to billions of requests per day. Search engines are thus facing the issue of providing their users with good results as quick as they can, results being web pages taken from a huge index (billions of web pages) and under a tremendous load (billions of requests each day).

To ensure good results to a particular request, search engines mostly use a relevance metric for web pages and requests. Being relevant regarding a certain query does not ensure being in the first places of a search engine result list. To arbitrate between equally relevant pages, search engines have to use other metrics. One of them is popularity. Most search engines are using a popularity mechanism that is not content related. Indeed, with a popularity independent of the content of webpages, computational issues linked to ranking webpages are less prevalent. Thus, popularity often depends on the links a webpage receive from other web pages. The Google’s PageRank algorithm (Brin, Page, Mot-
wani & Winograd, 1999) compute the popularity of all pages independently of the query while Kleinberg HITS algorithm (Kleinberg, 1999) has a query dependent version of the popularity called the authority score.

The economic model sustaining most websites is such that webmasters want their sites to appear on the first places of a search engines regarding specific requests. Indeed, the income of a website is directly correlated to unique visitors a website receives. Since search engines are the major sources of visitors on the web, to attract as many visitors as possible one has to maximize its exposure on them. Being close to the top will redirect a huge amount of traffic if the targeted request is wisely chosen.

Artificially increasing the relevance towards a request without quickly using spamming techniques and being spotted by search engines is almost impossible. Plus spammers often want to boost a legitimate page so they don’t need to manipulate its relevance. So spammers aim to increase their popularity to move to the top of the list. The most effective way to increase the popularity of a given webpage is to create a set of dull pages organized in a specific architecture whose goal is to boost the target page. This is a borderline technique, which is far beyond the guidelines of most search engines. Structures intending to maximize the pagerank of one specific page are well known (Gyongyi & Garcia-Molina, 2005). Those structures can no longer be use efficiently since it is hard for them to avoid automatic detection. Webspammers then slightly modify those structures to increase their rank while avoiding automatic detection.

There is thus a weapon race between search engines and spammers. It is really important for the first ones to deal with Webspam since it can pollute their results, then without fair results they will lose the users’ confidence and visits. Having less visitors search engines will then lose their income due to advertising and sponsored links. Fighting spam is then an economic necessity for the search engines. For spammers this is the same problem: their income is correlated to their exposure in the SERPs (Search Engines Results Pages). So each time a search engine adapts itself to lower the incidence of Webspam, they have to renew their techniques to stay one step ahead.

In this paper we present a method whose goal is to identify malicious structures amongst web pages. The intuition behind our method is that spammers use specific pages architecture to route the PageRank around the target page in order to maximize its score while avoiding automatic detection. Since the PageRank can be seen as related to the behavior of a random surfer, it seems that using random walks to reproduce the behavior of this random surfer we will be able to expose paths created by spammers in order to manipulate and increase their pagerank.

The main results of this paper are:

- A method based on strong statistical results (borrowed from (Fisher, Magniez & de Rougemont, 2006)) that allows identifying malicious patterns in the random walks.
- A methodology that classify random walks in similar categories in order to identify spammers and pages benefiting from their manipulations. Strong experimental results showing the efficiency of our approach.

This paper is organized as follow, in the next section, we present some related work regarding Webspam detection and demotion. The following section introduces our method to identify pages benefiting from Webspam. In the section after that we detail our experiments to verify the validity of our approach and show our results before concluding in the last section.

**RELATED WORK**

Since search engines are based on well known popularity metrics such as the PageRank (Brin, Page, Motwani & Winograd, 1999), spammers are trying to artificially boost their web sites with respect to these metrics. At the same time, Webspam has been extensively studied since it is really important for search engines to be able to deal with it.
Adaptive Neuro-Fuzzy Control Approach Based on Particle Swarm Optimization
Gomaa Zaki El-Far (2012). Innovations and Developments of Swarm Intelligence Applications (pp. 81-98).
www.igi-global.com/chapter/adaptive-neuro-fuzzy-control-approach/65807?camid=4v1a

Weighted Average-Based Multi-Objective Optimization of Tube Spinning Process using Non-Traditional Optimization Techniques
www.igi-global.com/article/weighted-average-based-multi-objective-optimization-of-tube-spinning-process-using-non-traditional-optimization-techniques/99654?camid=4v1a