Detecting Link Hijacking by Web Spammers

Young-joo Chung, Masashi Toyoda, and Masaru Kitsuregawa

Institute of Industrial Science, The University of Tokyo,
4-6-1 Komaba Meguro-ku, Tokyo, Japan
{chung,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract. Since current search engines employ link-based ranking algorithms as an important tool to decide a ranking of sites, Web spammers are making a significant effort to manipulate the link structure of the Web, so called, link spamming. Link hijacking is an indispensable technique for link spamming to bring ranking scores from normal sites to target spam sites. In this paper, we propose a link analysis technique for finding link hijacked sites using modified PageRank algorithms. We performed experiments on the large scale Japanese Web archive and evaluated the accuracy of our method. Detection precision of our approach was improved about 25% from a naive approach.

Keywords: Link analysis, Web spam, Information retrieval, Link hijacking.

1 Introduction

In the last decade, search engines have been essential tools for information retrieval. People rely heavily on these tools to find information on the Web, and as a result, most Web sites get a considerable number of visitors via search engines. Since the increase in visitors usually means the increase in financial profit, and approximately 50% of search engine users look at no more than the first 5 results in the list [1], obtaining a high ranking in the search results became crucial for the success of sites.

Web spamming is the behavior that manipulates Web page features to get a higher ranking than the page deserves. Web spamming techniques can be categorized into term spamming and link spamming. *Term spamming* manipulates textual contents of pages by repeating specific keywords that are not related with page contents and by adding irrelevant meta-keywords or anchor text. Search engines which employ textual relevance to rank pages will return these manipulated pages at the top of the result list. *Link spamming* manipulates the link structure of the Web to mislead link-based ranking algorithms such as PageRank [4]. Link-based ranking algorithms consider a link as an endorsement for pages. Therefore, spammers create numerous false links and construct an artificially interlinked link structure, so called a spam farm, where all pages link to the target spam page in order to centralize its link-based importance.

Links from external normal pages to spam pages are needed in order to attract the attention of search engines and feed ranking scores to spam farms. These

T. Theeramunkong et al. (Eds.): PAKDD 2009, LNAI 5476, pp. 339–350, 2009.

links that are created without any agreements of page owners are called *hijacked link*. To hijack link, spammers post comments including URLs to spam pages on public bulletin boards, buy expired domains and sponsor pages. Hijacked links affect link-based ranking algorithms significantly, when they are pointing to spam farms containing a large amount of spam pages.

In this paper, we propose a novel method for detecting Web sites hijacked by spammers. Most of previous research has focused on demoting or detecting spam, and as far as we know, there has been no study on detecting link hijacking that is important in the following situations:

- Hijacked sites are prone to be attacked continuously by various spammers (e.g. by repetitive spam comments on blogs). Observing such sites will be helpful for the prompt detection of newly created spam sites that might not be filtered by existing anti-spam techniques. Since spam detection has been an arms race, it is important to find out sites with new spamming methods.
- Once we detect hijacked sites, we can modify link-based ranking algorithms to reduce the importance of newly created links from hijacked pages in those sites. It makes the algorithms robust to newly created spam. Though it might temporally penalize links to normal sites, we can correct their importance after we invent spam detection methods for novel spamming techniques.
- Crawling spam sites is a sheer waste of time and resources. Most crawlers have spam filters, but such filters cannot quickly adapt themselves to new spamming methods. By reducing the crawling priority of new links from hijacked pages in detected sites, we can avoid collecting and storing new spam sites, until spam filters are updated.

In order to identify hijacked sites, we consider characteristics of the link structure around hijacked sites. As Figure 1 indicates, hijacked sites are supposed to have a certain number of links to both normal and spam sites, and exist at the boundary of them. To detect this boundary, we take account of trustworthiness and spamicity of whole sites. Normal sites would have high trustworthiness and low spamicity, and in contrast, spam sites would have low trustworthiness and high spamicity. These relations will be reversed at the link between normal sites and spam sites, or where link hijacking occurs. Based on this idea, we detect the point where trustworthiness and spamicity are reversed in order to extract hijacked sites.

In addition, we focus on the fact that hijacked sites have links pointing to both normal and spam sites. Out-neighbors of normal sites will show much more trustworthiness than spamicity, and vice versa. Thus, it would be assumed that overall trustworthiness and spamicity in out-neighbors of hijacked sites are equivalent compared to those of normal or spam.

Trustworthiness and spamicity of a site can be evaluated by some link-based ranking algorithms such as modified versions of PageRank. For each site, we calculate white and spam scores using two different modified PageRanks. Intuitively, these scores represent the degree of trustworthiness and spamicity of sites.