Detecting Soft Errors by Redirection Classification *

Taehyung Lee†, Jinil Kim†, Jin Wook Kim‡, Sung-Ryul Kim§, and Kunsoo Park†
†‡{thlee,jikim,jwkim,kpark}@theory.snu.ac.kr; §kimsr@konkuk.ac.kr

†School of Computer Science and Engineering, Seoul National University, Seoul 151-742, South Korea
‡HM Research, Seoul 151-742, South Korea
§Division of Internet & Media, Konkuk University, Seoul 143-701, South Korea

ABSTRACT

A *soft error* redirection is a URL redirection to a page that returns the HTTP status code 200 (OK) but has actually no relevant content to the client request. Since such redirections degrade the performance of web search engines in many ways, it is highly desirable to remove as many of them as possible. We propose a novel approach to detect soft error redirections by analyzing redirection logs collected during crawling operation. Experimental results on huge crawl data show that our measure can classify soft error redirections effectively.

Categories and Subject Descriptors: H.3.3 [Information Systems]: Information Search and Retrieval; H.3.5 [Information Systems]: Online Information Services–*Web based services*
General Terms: Experimentation, Measurement
Keywords: Search engine, Soft 404, Spam, URL redirection

1. INTRODUCTION

Behind the rapid proliferation, the web exhibits rapid decay as well. Recent studies have reported that a large fraction of the Web consists of dead pages, and especially, more than 25% of the dead links are involved in the (in)famous *soft error* redirections [2].

Soft error redirections occur in many ways. Many web servers today redirect the requested – but permanently or temporarily inaccessible – URL to a dedicated page informing a certain error such as 403 (forbidden), 404 (not found), or 5xx (server error), with the status code 200 (OK). Sometimes, miscofiguration causes inadvertent redirections to the site's root page.

Another important type of soft error is a redirection spam. Redirection methods have historically been abused by spammers in many ways. For example, when a domain name is allowed to lapse, a spammer often re-registers it and puts a redirect from this site into his own web site, in order to profit from the prior promotional works of the previous owners of the "parked" site [2]. Sometimes, a spammer generates enormous keyword stuffed URLs and redirects them to the target page. We regard these types of redirections as soft errors in a broad sense.

*Supported by Korea Research Council of Fundamental Science and Technology. The ICT at Seoul National University provides research facilities for this study. §This author is supported by Seoul R&BD Program(10581).

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

The soft error redirection seriously disturbs not only the web surfer's experience, but also the search engine's performance in many aspects. In crawling, for example, it leads to crawl a bunch of pages which are not related with the requested URLs, but whose contents are nearly identical. In indexing, these irrelevant and near-duplicate pages waste computation time and disk space. In ranking, they adversely affect link-based ranking to result in disproportionate promotion to undeserving pages. Then, in presentation, they distort displayed search results and lead to a poor user experience. Thus, for search engines, it is crucial to find out such unnecessary redirections in as early a stage as possible.

However, to the best of our knowledge, there have been few attempts to tackle this problem in the literature. Bar-Yossef et al. [2] developed a heuristic for detecting *soft 404*s, in which they induce a 404 error by fetching a URL with a *random* file name, and find out that the target server generates a soft error redirection to a certain page (soft 404). Their heuristic effectively finds soft 404 pages, but it does not cover soft errors related with other types of errors such as 403, or 5xx. Moreover, the assumption that roots of web sites are never redirected to soft 404 pages prohibits finding the soft errors related with redirection spams.

We propose a novel measure to detect soft error redirections by analyzing redirection logs collected during crawling operation. Since we only use the redirection logs, we neither require predefined white/black lists nor fetch actual web pages for the whole process. The results demonstrate that we can effectively identify not only soft 404, but also 403, and 5xx. Moreover, the proposed measure also catches redirection spams in the type of soft errors.

2. HEURISTICS AND SCORING MEASURE

Let u be the URL[1] of a page and $host(u)$ be the hostname of u, e.g., for $u =$ `http://www.foo.com/bar`, $host(u)$ is `www.foo.com`. Suppose we have a large collection U of *original* URLs along with V of *target* URLs, and a function $f : U \to V$, where a redirection[2] from $u \in U$ to $v \in V$ is denoted by $v = f(u)$. We are to find soft error redirections $(u, f(u))$ from a redirection log $L \subset U \times V$, where L is a set of redirection pairs $\{(u, f(u)) : u \in U\}$ collected by a crawler.

Our first observation is that if a host has a dedicated page v for soft error redirection (and this appears to be common), we can expect to find multiple occurrences of v as

[1]We assume that all input URLs are properly parsed and normalized according to the URI syntax defined in RFC 3986.
[2]End-to-end redirection up to 5 hops.

the target URL in redirection log L. For example, we observed prevalence of redirections to `http://errdoc.gabia.net/404.html`, which is a soft 404 page in one of the largest Korean web hosting sites. This simple observation gives the following heuristic. For a target URL $v \in V$, let I_v be the number of redirections in L whose target URL is v.

H1. ***A soft error redirection*** (u, v) ***is likely to have a large*** I_v**.**

Unfortunately, there are many redirections that have large I_v albeit they are not related with soft errors. For example, legitimate redirections to a site's root page v may have a large I_v due to soft error redirections to v, as well as legitimate redirections. We observed many instances that a reputable URL such as `http://www.google.com` is redirected not only from legitimate URLs such as `http://google.com`, but also from a lot of irrelevant URLs such as `http://buyonline.drugsmore.com`. Thus, a decision solely based on **H1** may filter out relevant redirections too. We address the challenge of distinguishing between relevant and irrelevant redirections sharing the same target URL by using the following two characteristics of original URL's host.

Since hosts generating a number of soft errors usually tend to redirect many URLs into a few target URLs, we characterize these hosts by means of *convergence* of redirections. For a host h, let N_h be the number of original URLs in h and M_h be the number of target URLs redirected from URLs in h. Then, N_h/M_h, the average number of original URLs in h sharing the same redirection target, represents the amount of *converging* redirections in h. The following heuristic employs this characteristic.

H2. ***A redirection from a host*** h ***with large*** N_h/M_h ***is likely to be a soft error redirection.***

On the other hand, legitimate advertising servers or portals' directory servers typically generate a large number of redirections whose target hosts are distinct. Using this characteristic, we can prevent their redirections from being misclassified as soft errors by **H1**. We define H_h as the number of distinct target hosts reached by redirections from a host h, i.e. $H_h = |\{host(y) : (x, y) \in L, \text{ s.t. } host(x) = h\}|$.

H3. ***A legitimate redirection server*** h ***is likely to have a large*** H_h**.**

Based on the above heuristics, we design a scoring measure μ as follows. For each redirection $(u, v) \in L$, we define

$$\mu(u, v) \triangleq k_1 \log_{10} I_v + k_2 \log_{10} \frac{N_{host(u)}}{M_{host(u)}} + k_3 \log_{10} \frac{1}{H_{host(u)}}$$

where k_1, k_2, and k_3 are nonnegative weight constants. This measure gives higher scores to soft error redirections.

3. EXPERIMENTAL VALIDATION

All the experimental data used for validation were collected by WisponBot [1]. Here, we only show results on the log with 13,549,460 redirections from 3.2 billion Korean web pages crawled in June, 2008. We remark that experimental results for other sets of crawl data with different sizes and timestamps show similar trends.

We say that a redirection is a *highly suspected* soft error, if any of distinct patterns suggesting soft errors, such as "404", "503", "error", *etc.*, are found in the target URL, but not in the original URL. We first estimate classification

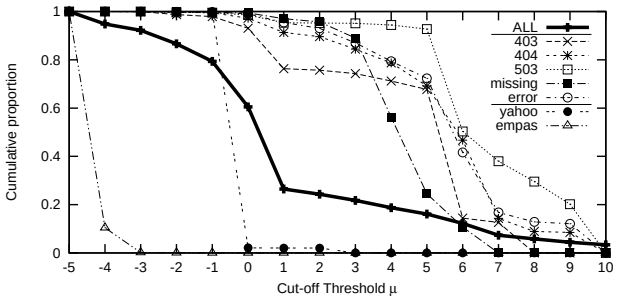

Figure 1: Reverse cumulative distribution curves for highly suspected soft error and normal redirections over μ $(k_1, k_2, k_3 = 1.0)$

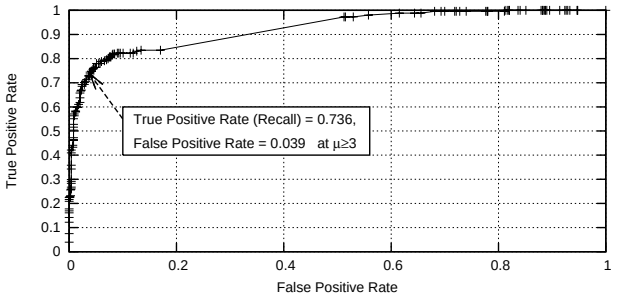

Figure 2: ROC Curve with the AUC=0.95

performance of our scoring measure by observing the reverse cumulative distribution of highly suspected soft error redirections across the score. Fig. 1 shows clear distinction between highly suspected soft error redirections and the entire redirections in the log (ALL). For example, if we set 3 as the cut-off threshold value, we can find almost 80% of soft errors having "404", "503", and "error" in the target URL. On the other hand, legitimate redirections from a directory server such as Yahoo directory or `empas.com` (a Korean portal) have a different shape of curve, which shows that almost all redirections have scores below 0.

Fig. 2 shows the Receiver Operating Characteristic (ROC) curve derived from manual inspection on randomly sampled 1,000 redirections. The area under the ROC curve has 0.95, which indicate a high accuracy. The cut-off threshold of 3 gives us precision of 0.866, recall of 0.736 and false positive rate of 0.039. Table 1 below shows composition of correctly classified soft errors with this threshold. We can effectively identify soft error redirections to sites' root page ("Root" in the table), soft 5xx and 403 as well as soft 404s. Note that 60.7% of redirection spams are also caught by our measure.

Table 1: Composition of correctly classified soft errors

(%)	404	Root	5xx	Spam	403	*etc.*	**TOT.**
Portion	30.5	24.1	21.9	9.1	0.5	13.9	**100.0**
Recall	69.5	71.4	100.0	60.7	50.0	68.4	**73.6**

4. REFERENCES

[1] Wispon. `http://www.wispon.com`.
[2] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins. Sic transit gloria telae: towards an understanding of the web's decay. In *WWW '04*.