A Larger Scale Study of Robots.txt

Santanu Kolay, Paolo D'Alberto, Ali Dasdan, and Arnab Bhattacharjee
Yahoo! Inc.
Sunnyvale, CA, USA
{santanuk,pdalbert,dasdan,arnab}@yahoo-inc.com

ABSTRACT

A website can regulate search engine crawler access to its content using the robots exclusion protocol, specified in its robots.txt file. The rules in the protocol enable the site to allow or disallow part or all of its content to certain crawlers, resulting in a favorable or unfavorable bias towards some of them. A 2007 survey on the robots.txt usage of about 7,593 sites found some evidence of such biases, the news of which led to widespread discussions on the web. In this paper, we report on our survey of about 6 million sites. Our survey tries to correct the shortcomings of the previous survey and shows the lack of any significant preferences towards any particular search engine.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Search Process

General Terms: Experimentation, Measurement

Keywords: Crawler, robots exclusion, robots.txt, search engine

1. INTRODUCTION

Search engines retrieve content from (web)sites using crawling agents called *robots* or *crawlers*. Since a site needs to deal with many crawlers using its resources, it needs to regulate their behavior. The *robots exclusion protocol* [2] is a partial solution to this regulation problem, providing an advisory regulation for crawlers to follow. To use this protocol, a site will typically specify its protocol rules in a file called robots.txt. The rules allow the site to allow or disallow part or all of its content to specific crawlers.

Despite the importance of this protocol for both content providers and search engines, the first reasonably large scale study for its usage was done only recently in 2007 [4, 5]. The study was performed over 2,925 distinct robots.txt files from 7,593 sites. This study reported many useful observations but the most surprising of them was the one that there is a bias towards specific crawlers, which was also shown to strongly correlate with the respective search engine market shares. This observation was also discussed widely in many well-known blogs following the search engine field. A web search using the query "robots.txt study bias" returns many pages discussing the findings of this study.

Intrigued by the observations of the previous study, we started our own investigation into the robots.txt usage. Using the crawler of Yahoo! search engine, we started our regular retrieval of about 2.2M non-empty robots.txt files from 6M sites (5M top and 1M randomly selected, where "top" is in terms of HostRank, a quality score similar to PageRank but computed on the host graph). Having access to many properties of these sites, we performed our analysis of the collected data in many ways, and also replicated the analysis of the previous study. Among our many findings, the main result is that some sites may have bias towards specific crawlers but overall the top two crawlers seem to have access to the same amount of content. This is contrary to the point made by the previous study.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. EXPERIMENTS AND RESULTS

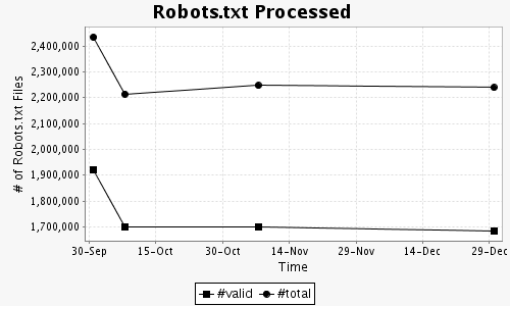
An ideal measure of the bias towards a crawler is through the value of the allowed and disallowed content. Unfortunately, this is almost impossible to measure because, e.g., a search engine cannot value the disallowed content without serving it to its customers. Hence, all the approaches to measuring the bias resort to approximating content via directory or URL counts.

In [4, 5], the bias was measured by counting the number of *directories* disallowed, i.e., the crawler with the highest (lowest) such count was regarded as having the most unfavorable (favorable) bias. The drawback of this approach is that the number of directories disallowed may not correlate well to the amount of content or the number of URLs disallowed.

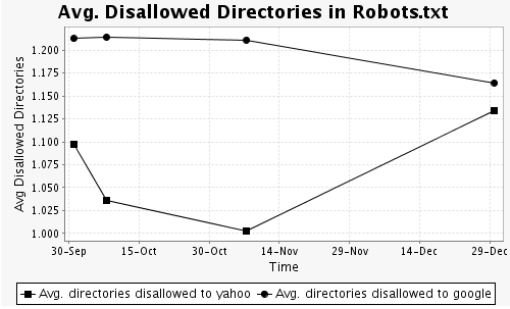
In our study, we replicated the previous approach (Figs. 1(b) and 1(c)) but we also tried to estimate the number of URLs disallowed by sending path queries to search engines. Though better than the directory counts, the URL counts still cannot scale due to daily search limits imposed by search engines. We, however, were able to get the counts for at least 1,500 sites per month.

Once we had the URL counts for two search engines Y and G, we computed the relative bias (Fig. 1(d)) of a site H towards these engines in two steps: (1) find the percentage PY (PG) of H's URLs that only Y (G) has, and then (2) report bias towards Y if PY is larger than PG (by some ϵ) or bias towards G otherwise.

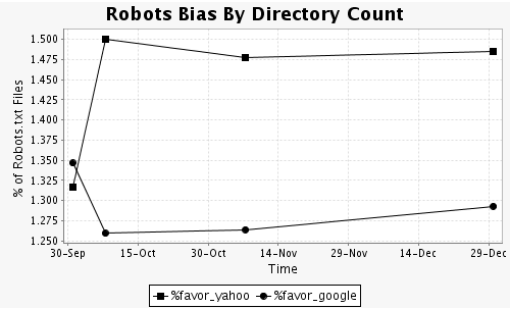
We performed other experiments to understand the significance of the robots.txt bias: the use of CrawlDelay (Fig. 1(e)), which may introduce bias if a crawler does not honor it, and the use of the sitemap protocol. The sitemap protocol [3] enables a site to inform a search engine about any of its content it prefers to be crawled. It was introduced in Apr. of 2007, and the big four search engines all support it. Our



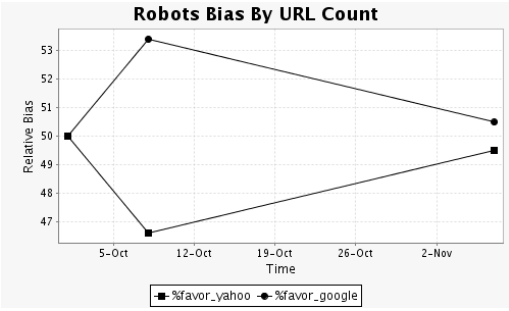
(a) # of non-empty robots.txt files processed.



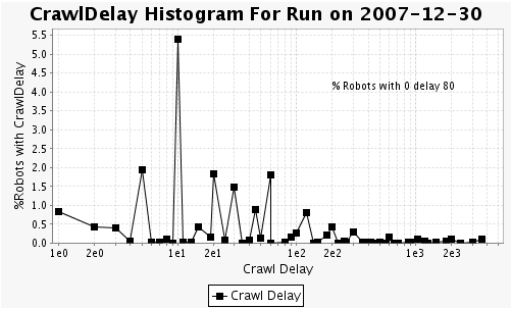
(b) Average (Avg.) # of disallowed directories disallowed per robots.txt file.



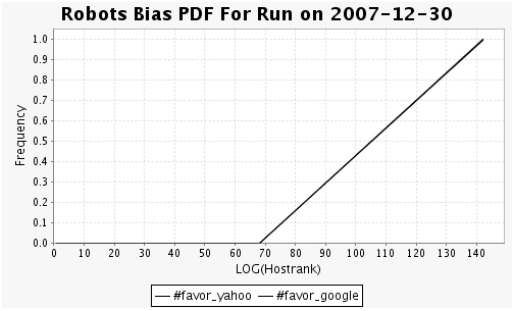
(c) Bias via directory count occurs in less than 1.5 % of valid robots.txt files.



(d) Bias via URL count is almost the same.



(e) Bias via CrawlDelay is mostly created by small values.



(f) Bias via HostRank is almost the same. The bias mostly occurs with low-rank hosts.

Figure 1: Our comparison of the robot.txt bias towards two major search engine crawlers. Note that these engines get almost the same share of bias in all cases and that the amount of bias in each case is actually very small.

evaluation [1] has shown that since Oct. 2007, the sitemap usage has been increasing by 10,000 robots.txt files a month. However, only about 6% out of 1.7M valid robots.txt files had the sitemaps link specified in Dec.

3. CONCLUSIONS

We have done a comprehensive and larger scale study of the robots.txt usage. Our results show that the bias towards search engines is not as serious as reported by the recent prior work [4, 5].

4. REFERENCES

[1] A. Dasdan, D. Pavlovski, and A. Bhattacharjee. A large scale study of sitemap usage. Available by email, Jan 2008.

[2] M. Koster. A method for web robots control. The internet draft, The Internet Engineering Task Force (IETF), 1996.

[3] The sitemap protocol. URL: http://www.sitemaps.org, Apr 2007.

[4] Y. Sun, Z. Zhuang, I. G. Councill, and C. L. Giles. Determining bias to search engines from robots.txt. In *Proc. of Int. Conf. on Web Intel. WI)*, pages 149–55. IEEE/WIC/ACM, 2007.

[5] Y. Sun, Z. Zhuang, and C. L. Giles. A large scale study of robots.txt. In *Proc. of Int. Conf. on World Wide Web (WWW)*, pages 1123–4. ACM, 2007.