

A Large-Scale Study of Robots.txt

Yang Sun, Ziming Zhuang, and C. Lee Giles
The Pennsylvania State University
University Park, PA, USA
{ysun, zzhuang, giles}@ist.psu.edu

ABSTRACT

Search engines largely rely on Web robots to collect information from the Web. Due to the unregulated open-access nature of the Web, robot activities are extremely diverse. Such crawling activities can be regulated from the server side by deploying the Robots Exclusion Protocol in a file called robots.txt. Although it is not an enforcement standard, ethical robots (and many commercial) will follow the rules specified in robots.txt. With our focused crawler, we investigate 7,593 websites from education, government, news, and business domains. Five crawls have been conducted in succession to study the temporal changes. Through statistical analysis of the data, we present a survey of the usage of Web robots rules at the Web scale. The results also show that the usage of robots.txt has increased over time.

General Terms

Experimentation, Measurement.

Keywords

crawler, robots exclusion protocol, robots.txt, search engine.

1. INTRODUCTION

Without robots, there would probably be no search engines. Web search engines, digital libraries, and many other web applications such as offline browsers, internet marketing software and intelligent searching agents heavily depend on robots to acquire documents. Robots, also called “spiders”, “crawlers”, or “bots”, are self-acting agents that navigate around-the-clock through the hyperlinks of the Web, harvesting topical resources at zero costs of human management [4]. Because of the highly automated nature of the robots, rules must be made to regulate such crawling activities in order to prevent undesired impact to the server workload or access to non-public information.

The Robots Exclusion Protocol has been proposed [3] to provide advisory regulations for robots to follow. A file called robots.txt, which contains robot access policies, is deployed at the root directory of a website and accessible to all robots. Ethical robots read this file and obey the rules during their visit to the website. Despite the criticality of the robots.txt convention for both content providers and harvesters, little work has been done to investigate its usage in

detail, especially at the scale of the Web. A study of the usage of robots.txt in UK universities and colleges investigated 163 websites and 53 robots.txt [2]. Robots.txt files were examined in terms of file size and the use of Robots Exclusion Protocol within the UK university domains. Drott [1] studied the usage of robots.txt as an aid for indexing to protect information on 60 samples from Fortune Global 500 company websites.

In this poster, we present the first large-scale study of robots.txt files covering the domains of education, government, news, and business. We present our observations on a considerably larger scale data than previous studies.

2. DATA COLLECTION

Our primary source to collect the initial URLs to feed our crawler is the Open Directory Project (DMOZ). Our collection from DMOZ covers three domains: education, news, and government. The university domain is further broken down into the American, European, and Asian university domains. We use the Fortune Top 1000 Company List as our data source in the business domain. Our crawler has performed five crawls for the same set of websites between Dec. 2005 and Oct. 2006.

3. RESULTS

Statistics: We crawled and investigated 7,593 unique websites including 600 government websites, 2,047 newspaper websites, 1,487 USA university websites, 1,420 European university websites, 1,039 Asian university websites, and 1,000 company websites.

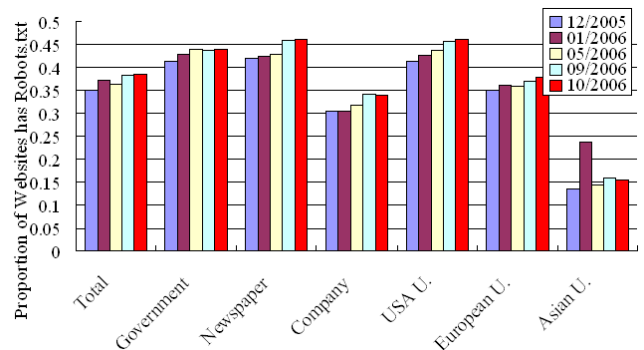


Figure 1: Probability of a website that has robots.txt in each domain.

Overall, the percentage of websites that have robots.txt has increased from 35% to 38.5% in the past 11 months (see Figure 1). Since search engines and intelligent searching agents become more important for accessing web information, this result is expected. The Robots Exclusion Protocol is more frequently adopted by government (44%), newspaper (46%) and university websites in the USA (45.9%). It is used extensively to protect information not to be offered to the public and balance workload for these websites.

There are 1056 named robots found in our dataset. The universal robot “*” is the most frequently used robot in the User-Agent field and used 2744 times, which means 93.8% of robots.txt files have rules for the universal robots. 72.4% of the named robots appeared only once or twice.

Size and Length: An interesting observation is that the sizes and lengths of the robots.txt files on governmental websites are significantly larger than those from the other investigated domains. There are 26 files at a length of 68 lines and 4 at 253 lines. A reasonable explanation is that government websites tend to adopt more sophisticated robot restrictions, which results in larger and longer robots.txt files (see Table 1).

	avg size with standard deviation
USA edu	625.1 ±158.1
European edu	422.5 ±86.7
Asian edu	270.1 ±108.0
Business	895.8 ±472.6
Gov	1551.2 ±760.6
News	509.7 ±41.3

Table 1: The average size (in bytes) and average length (in number of lines) of the collected robots.txt files.

Crawl Delay: The field name “Crawl-Delay” in robots.txt files has recently been used by web administrators. Web server administrators most likely use this field in their robots.txt files to arrange an affordable workload. The usage of Crawl-Delay increased from 40 cases (1.5%) in Dec. 2005 to 140 cases (4.8%) in Oct. 2006. The frequency of Crawl-Delay rules for different robots are shown in Table 2.

Incorrect Use: When we examine the content of the collected robots.txt files, a significant number of incorrect uses of the Robots Exclusion Protocol has been found. These incorrect uses include misnamed files, incorrect locations, and conflicting rules. Because of these incorrect uses, the access policy will be ignored by robots. We observe 13 cases of misnamed robots.txt files and find 23 files in which a specific name such as “crawler”, “robot”, or “webcrawlers” appear in the User-Agent field. General name in the User-Agent field is an incorrect use of the Robots Exclusion Protocol.

Robot Name	Number of Delay Rules
msnbot	42
slurp	36
yahooseeker/cafekelsa	12
googlebot	7
teoma	6

Table 2: The frequency of Crawl-Delay rules for different robots.

We also found 282 robots.txt files with ambiguous rules and 18 files with conflicting rules (e.g. a directory is disallowed first and then allowed or allowed first and then disallowed).

The actual method for how robots will access the robots.txt is not specified in the Robots Exclusion Protocol. Open source crawlers such as “Websphinx”, “Jspider” and “Nutch” checks the robots.txt file right before crawling each URL by default. We observe “Googlebot”, “Yahoo! Slurp” and “MSNbot” cache the robots.txt files for a website. During the modification of robots.txt file, these robots might disobey the rules. As a result, there are a few disallowed links appearing in these search engines.

Comments In Robots.txt: We have found cases in which comments in robots.txt files are not written for version or explanation but written for users or robot administrators¹. Even a blog has been found in a robots.txt file².

Issues: The rule “Disallow: ” can be understood as matching everything or nothing. Does the rule mean allow robots to crawl anything or nothing? Another issue is about the “Crawl-Delay” field. Since this field is not in the original Robots Exclusion Protocol, not every robot recognizes this rule. How should webmasters design the robots.txt if they do not have the knowledge whether a robot recognizes the rule or not? There is no further discussion in the Robots Exclusion Protocol about these issues.

4. CONCLUSIONS

We have presented a survey of the use of the Robots Exclusion Protocol on the Web through statistical analysis of a large sample of robots.txt files. Our study indicates that the usage of robots.txt has increased over the past 11 months in which 2,662 robots.txt files were found in the first crawl and 2,925 files were found for the last crawl. We observe that 46.02% of newspaper websites currently have implemented robots.txt files and the newspaper domain is the domain in which the Robots Exclusion Protocol is most frequently adopted. 45.93% of the USA university websites in our sample adopt the Robots Exclusion Protocol, significantly more than European (37.8%) and Asian (15.4%) sites. Many incorrect uses of the Robots Exclusion Protocol were found. This, in addition to the potential for ambiguity in robots.txt files, implies that a better-specified, official standard is needed.

5. ACKNOWLEDGMENTS

The authors would like to acknowledge partial support from the National Science Foundation and useful suggestions from Isaac Councill and Andrei Broder.

6. REFERENCES

- [1] M. Drott. Indexing aids at corporate websites: The use of robots.txt and meta tags. *Information Processing and Management*, 38(2):209–219, 2002.
- [2] B. Kelly and I. Peacock. Webwatching uk web communities: Final report for the webwatch project. *British Library Research and Innovation Report*, 1999.
- [3] M. Koster. A method for web robots control. In *the Internet Draft, The Internet Engineering Task Force (IETF)*, 1996.
- [4] G. Pant, P. Srinivasan, and F. Menczer. *Crawling the Web*, chapter Web Dynamics. Springer-Verlag, 2004.

¹<http://www.ebay.com/robots.txt>

²<http://www.webmasterworld.com/robots.txt>