ANALYSIS OF THE USAGE STATISTICS OF ROBOTS EXCLUSION STANDARD

Ajay, Smitha
Graduate Student
Department of Computer Science Indiana University, Bloomington
sajay@cs.indiana.edu

Jaliya Ekanayake
Graduate Student
Department of Computer Science Indiana University, Bloomington
jekanaya@cs.indiana.edu

ABSTRACT

Robots Exclusion standard [4] is a de-facto standard that is used to inform the crawlers, spiders or web robots about the disallowed sections of a web server. Since its inception in 1994, the robots exclusion standard has been extensively used. In this paper, we present our results of the statistical analysis of the usage of robots exclusion standard. Based on the results obtained, we propose that organizations like W3C should adopt the Robert Exclusion Standard and make it an official standard.

KEYWORDS

Robots Exclusion standard, robots, hidden web

1. INTRODUCTION

Robots Exclusion standard [4] is a de-facto standard that is used to inform the crawlers about the disallowed sections of a web server. It has been in use since 1994 to limit the access of pages or sections such as very deep virtual trees, duplicated information, temporary information, or cgi-scripts with side-effects (such as voting). From the aspect of a crawler, this is a voluntary standard since it does not provide a mechanism to stop crawlers from accessing disallowed sections of the web. However, most crawlers adopt and conform to the robots exclusion standard. Although the standard has been there for almost a decade, extensive research regarding its usage has not been done. We have performed a statistical analysis of the usage of the above standard and our intention is to propose that this be made an official standard. Our analysis covers the following areas namely:

- Usage of the standard; Fraction of the web that uses this standard
- Hidden web: Percentage of the web hidden by robots.txt
- Validation of robot.txts to find the percentage of robots.txt that has no errors or warnings.

Section 2 of the paper is about the related work in this area. In section 3, we discuss the methodology adopted to collect the results and tools employed to validate the robots.txt. In section 4 we present the results obtained followed by a discussion and analysis of the same in section 5.

2. RELATED WORK

Although the robots exclusion standard has been in vogue for nearly a decade and has been accepted as a defacto standard, there has not been any study conducted as regards the usage of the same. The commonly accepted standard was proposed in 1994 and a revised internet draft specification of the robots exclusion standard is still in progress [4]. Apart from a discussion group [5] dedicated to robots.txt and the proposed standard, we did not find any other related work.

3. DATA COLLECTION

3.1 Data Collection

In order to perform the analysis, we needed a large set of websites of various high levels domains (com, edu, org, info, gov, etc) that had robots.txts. For this purpose, the Open Directory Project (ODP)'s RDF was parsed to extract the websites of these domains. Of those websites, we randomly selected up to a maximum of 50,000 sites per high level domain and sent HTTP GET requests for each of these sites, to determine the existence of robots.txts. Of those sites that had robots.txts, random sites (up to 1000 sites per second level domain) were selected and crawled two levels completely. The open source crawler, Jobo [3] was customized and used for this purpose. In the process of crawling, we did violate the robots exclusion standard in order to collect the hidden pages. The alternative approach that we tried initially was to use only the HTTP GET request to determine the size of the hidden web pages, but the main drawback of this approach was that we could not estimate the fraction of hidden portion of a website beyond the first level which defeats the main purpose of our analysis. For each site, we collected the sizes of all the pages and whether they are hidden or allowed and persisted this information.

3.2 Validation

From the results collected above, the web sites that had robots.txts were validated using an online web based robots exclusion standard validator [6]. Since we had a huge number of websites pertaining to various domains, we used an automated testing tool (iMacros browser [2]) with a customized script that read the data from a file and wrote the validation results into an output file. These results are discussed in section 4.

4. RESULTS AND ANALYSIS

In this section, we present the results obtained and the analysis of the same.

Table 1. Usage statistics and percentage of the hidden web

High Level Domain	% of Usage	% of Hidden
com	24	19
org	20	9
net	22	12
edu	26	13
gov	43	11
info	27	15
Total	22	14

The usage results are obtained over a data set comprising of 128,000 second level domains and the percentage of the hidden web is calculated over a dataset of 30,000 websites that have robots.txts. The results reveal that 22% of all websites use robots.txt and their robots.txts hides 14% of their WebPages. That is, overall 3% of the entire web is hidden by robots.txts.An interesting observation was that the '.gov' domains

used robots.txts extensively and a manual inspection of many '.gov' sites revealed that the robots.txts is detailed and accurate. A p-value calculated as a result of a T-test returned a value of 0.009 that indicates that there is a significant statistical difference between the usages of the standard in the '.gov' domain with respect to the other high level domains.

Table 2. Percentage of errors and warnings

High Level Domain	% of Errors	% of Warnings
com	21	31
org	18	36
net	21	37
edu	19	35
gov	13	22
info	21	42
Total	20	34

The above results reveals errors and warnings in robots.txts.Errors indicate invalid syntax and inappropriate usage of robots.txt. The validator [6] generates warnings when the robots.txts contain features that are part of the new revised draft of the robots exclusion standard which is still in progress [4]. The robots or spiders may not guarantee compliance to the websites whose robots.txts has features of the revised draft. The errors and warnings have been classified in detail in figures 1 and 2 respectively. Again, an interesting observation was that the robots.txt contained in the '.gov' domain had the least number of errors and warnings.

Figure 1. Error types and their percentages

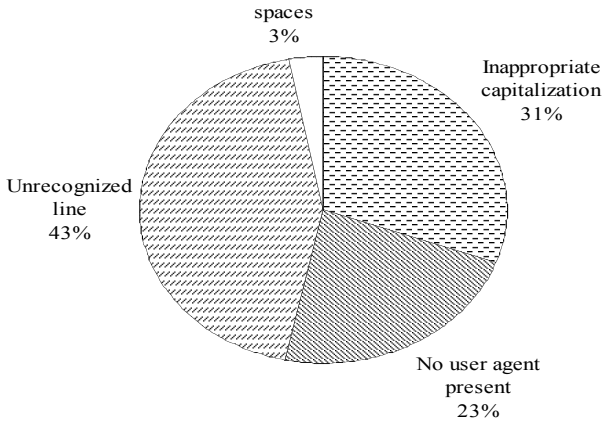
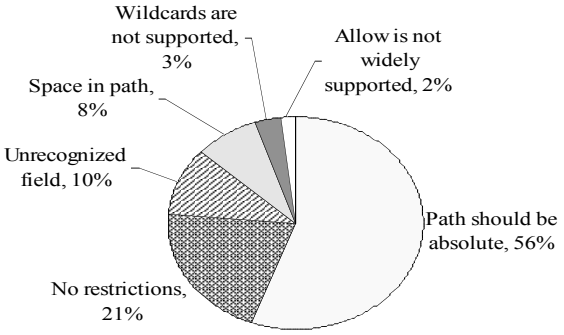


Figure 2. Warning types and their percentages



5. DISCUSSION

If we assume that the size of the web is 11.5 billion pages [1], according to our results, about 3 % of the total pages are hidden by the Robot Exclusion Standard which is about 330 million pages. This observation highlights the fact that the standard is widely accepted and used in all the top level domains. At the same time an error percentage of 20% shows that although the usage is wide spread across all the top level domains, there is a lack of knowledge of using the standard which defeats the purpose of the standard. We see the main cause for this problem is mainly due to the lack of a well defined official standard and hence suggest that organizations like W3C should adopt this as an official standard.

6. REFERENCE

[1] A. Gulli and A. Signorini, 2005, The Indexable Web is more than 11.5 billion pages. *In Poster proceedings of the 14th international conference on World Wide Web, Japan, ACM Press*. Japan.

[2] iMacros Browser, http://www.iopus.com/imacros/

[3] JoBo Crawler, http://www.matuschek.net/software/jobo/

[4] Robots Exclusion standard, http://www.robotstxt.org/wc/exclusion.html.

[5] Robots.txt news group, http://www.webmasterworld.com/forum93/

[6] Robots.txt Syntax Checker, http://www.sxw.org.uk/computing/robots/check.html.