Measuring The Web Crawler Ethics

C. Lee Giles
College of Information Sciences and Technology
Pennsylvania State University
University Park, PA, USA
giles@ist.psu.edu

Yang Sun
AOL Research
888 Villa St
Mountain View, CA, USA
yang.sun@corp.aol.com

Isaac G. Councill
Google Inc.
76 Ninth Avenue 4th Floor
New York, NY, USA
icouncill@gmail.com

ABSTRACT

Web crawlers are highly automated and seldom regulated manually. The diversity of crawler activities often leads to ethical problems such as spam and service attacks. In this research, quantitative models are proposed to measure the web crawler ethics based on their behaviors on web servers. We investigate and define rules to measure crawler ethics, referring to the extent to which web crawlers respect the regulations set forth in robots.txt configuration files. We propose a vector space model to represent crawler behavior and measure the ethics of web crawlers based on the behavior vectors. The results show that ethicality scores vary significantly among crawlers. Most commercial web crawlers' behaviors are ethical. However, many commercial crawlers still consistently violate or misinterpret certain robots.txt rules. We also measure the ethics of big search engine crawlers in terms of return on investment. The results show that Google has a higher score than other search engines for a US website but has a lower score than Baidu for Chinese websites.

Categories and Subject Descriptors

K.4.1 [**Public Policy Issues**]: Ethics; K.4.1 [**Public Policy Issues**]: Privacy

General Terms

Measurement, Design, Experimentation, Algorithms

Keywords

robots.txt, web crawler ethics, ethicality, privacy

1. INTRODUCTION

Web crawlers have been widely used for search engines as well as many other web applications to collect content from the Web. These crawlers are highly automated and seldom regulated manually. With the fast growing online services relying on Web crawlers to collect Web pages, the functionalities and activities of web crawlers have become extremely diverse. Crawler activities typically include requests of web pages for general-purpose text indexing and searching, extraction of email and personal identity information for business purposes as well as for malicious purposes. Accessing the web information with automated Web crawlers can lead to ethical problems of privacy and security. For example, crawlers can extract personal contact information for spam purposes and identity theft. Crawlers may also overload a website such that normal user access is impeded. Web crawler activities can be regulated from the server side by deploying Robots Exclusion Protocol (a set of rules in a file called robots.txt) in the root directory of a website, allowing webmasters to indicate to visiting crawlers which parts of their sites should not be visited as well as a minimum time interval between visits. A recent study shows more than 30% of active websites employ this standard to regulate crawler activities [2, 3]. However, since the Robots Exclusion Protocol (REP) serves only as an unenforced advisory to crawlers, web crawlers may ignore the rules and access part of the forbidden information on a website. Violating the robots.txt rules can lead to serious privacy and security concerns. Thus, measuring crawler ethics becomes an important task to help detecting improper crawler behavior in early stages as well as identifying unethical crawlers. The issues of crawler ethics, however, did not bring enough attention to the research community and are under studied. Crawler ethics are not limited to whether crawlers obeying website rules, but also can be studied in terms of the value provide to websites. If a crawler provides zero value to the crawled website, it should also be considered less ethical than those who provide positive values.

In this research, we propose a vector space model of measuring web crawler ethics based on the Robots Exclusion Protocol. We define the ethicality metric to measure web crawler ethics. We also study the ethics of big search engine crawlers in terms of return on investment where crawler visits are considered investments from websites and corresponding search engine traffic is considered as returns. The results show that Google has a much higher score in US websites but has a lower score than Baidu in Chinese websites.

2. RELATED WORK

The ethical factors are examined from three perspectives [4] : denial of service, cost, and privacy. An ethical crawl guideline is described for crawler owners to follow. This guideline suggests taking legal action or initiating a professional organization to regulate web crawlers. Our research adopts these perspectives of crawler ethics and expands it to a computational measure. The ethical issues of administrating web crawlers are discussed in [1]. It provides a guideline for ethical crawlers to follow. The guideline also gives great insights to our research of ethics measurements. However,

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

none of the above mentioned work provides a quantitative measure of web crawler ethics.

3. CRAWLER BEHAVIOR MODEL

In our research, each web crawler's behavior is modeled as a vector in the rule space where rules are specified by Robots Exclusion Protocol to regulate the crawler behavior. If a crawler violates a rule, the corresponding vector element is larger than 0. Websites can also be modeled in the rules space that if a website includes a rule in its robots.txt file, the corresponding vector element is larger than 0. The actual value for a rule element can be defined based on the consequences or cost of violating such rule.

We define content ethicality E_c and access ethicality E_a scores to evaluate web crawler ethics. In content ethicality, cost is defined as the number of restricted web pages or web directories being unethically accessed (see Eq. 1).

$$E_c(C) = \sum_{w_i \in W} \frac{||V_C(w_i)||}{||D(w_i)||}. \quad (1)$$

Access ethicality is defined as how a crawler respects the desired visit interval (*crawl-delay* rule in robots.txt file) of the website(see Eq. 2).

$$E_a(r) = \sum_{w_i \in W} \frac{e^{-(interval_C(w_i) - delay(w_i))}}{1 + e^{-(interval_C(w_i) - delay(w_i))}} \quad (2)$$

A major advantage for websites allowing search engine crawlers to crawl their web pages is that the search engines bring traffic back to them. From this perspective, being ethical for a web crawler means bringing more visits back to the crawled websites. The effective ethicality of search engine S to a website can be defined as the ratio between the user visits referred by the search engine to the website and visits generated by the crawler r of the search engine to the website (see Eq. 3).

$$E_{effective}(r) = \frac{Referenced(S)}{Crawled(r)} \quad (3)$$

4. EXPERIMENTS

Rank	User-agent	Content Ethicality
1	hyperestraier/1.4.9	0.95621
2	Teemer	0.01942
3	msnbot-media/1.0	0.00632
4	Yahoo! Slurp	0.00417
5	charlotte/1.0b	0.00394
6	gigabot/3.0	0.00370
7	nutch test/nutch-0.9	0.00316
8	googlebot-image/1.0	0.00315
9	Ask Jeeves/Teoma	0.00302
10	googlebot/2.1	0.00282

Table 1: Content ethicality scores for crawlers visited our test site.

Table 1 and 2 list the content and access ethicality results for top crawlers that visited our test website during the time of the study. Higher ethicality scores represent unethical crawlers.

The effective ethicality of *Google*, *Yahoo*, *MSN* and *Baidu* are shown in Table 3. The data is collected between 2008/05/13 to 2008/06/21. Site 1 is $CiteSeer^x$, a large scale academic digital library for computer science. Site 2 is a Chinese movie information website. Site 3 is *guopi.com*, an online makeup retail store.

Rank	User-agent	Access Ethicality
1	msnbot-media/1.0	0.3317
2	hyperestraier/1.4.9	0.3278
3	Yahoo! Slurp/3.0	0.2949
4	Teemer	0.2744
5	Arietis/Nutch-0.9	0.0984
6	msnbot/1.0	0.098
7	disco/Nutch-1.0-dev	0.0776
8	ia_archiver	0.077
9	gigabot/3.0	0.0079
10	googlebot/2.1	0.0075

Table 2: Access ethicality scores for crawlers visited our test site.

5. CONCLUSIONS

We formally defined three ethicality scores to measure web crawler ethics. Results show that most commercial crawlers receive a good ethicality scores. However, it is surprising to see commercial crawlers constantly disobeying or misinterpreting some robots.txt rules. The crawling algorithms and policies that lead to such behaviors are unknown. However, obtaining more content is an obvious reason for most crawlers failing to obey certain rules.

	Website	Crawled	Referenced	E_{return}
google	Site 1	16799253	260898	0.01553
	Site 2	872001	46469	0.05329
	Site 3	368417	145115	0.39389
yahoo	Site 1	17375962	3919	0.00023
	Site 2	502584	1249	0.00249
	Site 3	315119	11819	0.03751
msn	Site 1	677181	362	0.00054
	Site 2	16330	5448	0.33362
	Site 3	51128	3801	0.07434
baidu	Site 1	27	37	1.37037
	Site 2	622667	61964	0.09951
	Site 3	1830847	844786	0.46142

Table 3: Comparison of the effectiveness of *Google*, *Yahoo*, *MSN* and *Baidu*.

The effective ethicality scores of search engines varies significantly for different websites. Ranking by the referenced visits, *Google* plays a dominating role in the US based site 1 and ranks the 2nd and 3rd in the two China based websites. *Baidu* leads in the search market in China.

6. REFERENCES

[1] D. Eichmann. Ethical web agents. *Computer Networks and ISDN Systems*, 28(1-2):127–136, 1995.

[2] S. Kolay, P. D'Alberto, A. Dasdan, and A. Bhattacharjee. A larger scale study of robots.txt. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 1171–1172, New York, NY, USA, 2008. ACM.

[3] Y. Sun, Z. Zhuang, and C. L. Giles. A large-scale study of robots.txt. In *WWW '07*, 2007.

[4] M. Thelwall and D. Stuart. Web crawling ethics revisited: Cost, privacy, and denial of service. *J. Am. Soc. Inf. Sci. Technol.*, 57(13):1771–1779, November 2006.