2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

The Ethicality of Web Crawlers

Yang Sun
AOL Research
Mountain View, USA
Email: yang.sun@corp.aol.com

Isaac G. Councill
Google Inc.
New York, USA
Email: icouncill@gmail.com

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

***Abstract*—Search engines largely rely on web crawlers to collect information from the web. This has led to an enormous amount of web traffic generated by crawlers alone. To minimize negative aspects of this traffic on websites, the behaviors of crawlers may be regulated at an individual web server by implementing the Robots Exclusion Protocol in a file called "robots.txt". Although not an official standard, the Robots Exclusion Protocol has been adopted to a greater or lesser extent by nearly all commercial search engines and popular crawlers. As many web site administrators and policy makers have come to rely on the informal contract set forth by the Robots Exclusion Protocol, the degree to which web crawlers respect robots.txt policies has become an important issue of computer ethics. In this research, 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 test the behaviors of web crawlers in terms of ethics by deploying a crawler honeypot: a set of websites where each site is configured with a distinct regulation specification using the Robots Exclusion Protocol in order to capture specific behaviors of web crawlers. We propose a vector space model to represent crawler behavior and a set of models to measure the ethics of web crawlers based on their behaviors. The results show that ethicality scores vary significantly among crawlers. Most commercial web crawlers receive fairly low ethicality violation scores which means most of the crawlers' behaviors are ethical; however, many commercial crawlers still consistently violate or misinterpret certain robots.txt rules.**

***Keywords*-robots.txt; web crawler ethics; ethicality; privacy**

I. INTRODUCTION

Previous research on computer ethics (a.k.a. machine ethics) primarily focuses on the use of technologies by human beings. However, with the growing role of autonomous agents on the internet, the ethics of machine behavior has come to the attention to the research community [1], [3], [8], [9], [12], [18]. The field of computer ethics is especially important in the context of web crawling since collecting and redistributing information from the web often leads to considerations of information privacy and security.

Web crawlers are an essential component for many web applications including search engines, digital libraries, online marketing, and web data mining. These crawlers are highly automated and seldom regulated manually. With the increasing importance of information access on the web, online marketing, and social networking, the functions and activities of web crawlers have become extremely diverse. These functions and activities include not only regular crawls of web pages for general-purpose indexing, but also different types of specialized activities such as extraction of email and personal identity information as well as service attacks. Even general-purpose web page crawls can lead to unexpected problems for Web servers such as the denial of service attack in which crawlers may overload a website such that normal user access is impeded. Crawler-generated visits can also affect log statistics significantly so that real user traffic is overestimated.

The Robots Exclusion Protocol[1] (REP) is a crawler regulation standard that is widely adopted on the web and provides a basis for which to measure crawler ethicality. A recent study shows more than 30% of active websites employ this standard to regulate crawler activities [17]. In the Robots Exclusion Protocol, crawler activities can be regulated from the server side by deploying rules in a file called robots.txt in the root directory of a web site, allowing website administrators to indicate to visiting robots which parts of their site should not be visited as well as a minimum interval between visits. If there is no robots.txt file on a website, robots are free to crawl all content. Since the 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. Therefore, the usage of the Robots Exclusion Protocol and the behavior of web crawlers with respect to the robots.txt rules provide a foundation for a quantitative measure of crawler ethics.

It is difficult to interpret ethicality in different websites. The unethical actions for one website may not be considered unethical in others. We follow the concept of crawler ethics discussed in [7], [18] and define the ethicality as the level of conformance of crawlers activities to the Robots Exclusion Protocol. In this paper, we propose a vector model in the REP rule space to define ethicality and measure the web crawler ethics computationally.

The rest of the paper is organized as follows. Section 2 discusses prior work related to our research. An analysis of crawler traffic and related ethical issues is presented in section 3. Section 4 describes our models of crawler behavior and definition of ethicality to measure crawler ethics. Section

[1] http://www.robotstxt.org/norobots-rfc.txt

978-0-7695-4191-4/10 $26.00 © 2010 IEEE
DOI 10.1109/WI-IAT.2010.316

5 describes an experiment on real-world crawlers based on their interactions with our honeypot and presents the results of ethicality defined in section 4. We conclude in section 6.

II. Related Work

Much research has been discussing the ethical issues related to computer and the web [3], [5], [9], [10], [11]. The theoretical foundation for machine ethics is discussed by [3]. Prototype systems are implemented to provide advice on ethically questionable actions. It is expected that "the behavior of more fully autonomous machines, guided by this ethical dimension, may be more acceptable in real-world environments."[3] Social contract theory is used to study the computer professionals and their social contract with society[9]. The privacy and piracy issues of software are discussed in [5]. The need for informed consent in Web related information research has been advocated in [11] and debated in [10].

Ethical problems that relate to web crawlers are discussed in [7], [18]. In [18], the ethical factors are examined from three perspectives: 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 the perspective of crawler ethics and expands it to a computational measure. The ethical issues of administrating web crawlers are discussed in [7]. It provides a guideline for ethical crawlers to follow. The guideline also gives great insights to our research of ethicality measurements.

Since our goal is to develop a computational model to measure the ethicality of web crawlers automatically based on their behavior on websites. Web server access log analysis is one of the key technologies used in our study. Prior research has been conducted to study crawler behaviors using web access logs [2], [6], [14]. [2] analyzes the characterization of workload generated by web crawlers and studies their impact on caching. [6] characterizes and compares the behavior of five search engine crawlers. A set of metrics is also proposed to describe qualitative characteristics of crawler behavior. The measure reveals the indexing performance of each web crawler for a specific website. [14] studies the change of web crawler behavior when a website is decaying. The results show that many crawlers adapt to site changes, adjusting their behavior accordingly.

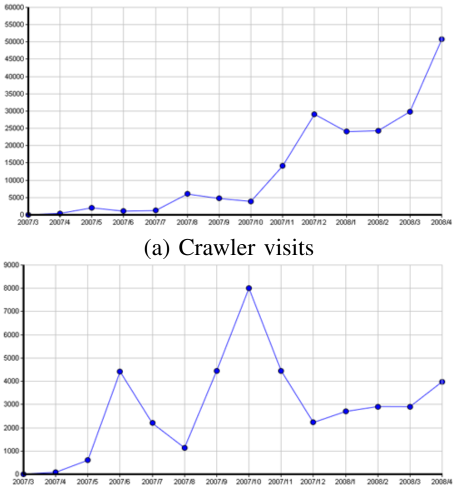
The legal issues surrounding network measurements are also discussed recently in [4], [13]. Since many web data mining research involves crawling, the privacy issues are also ethical issues regarding the crawlers used by researchers. These two papers provide some guidelines to consider ethical issues.

None of the above mentioned work provides a quantitative measure of the ethical factors (ethicality) of web crawlers.

III. Crawler Traffic Analysis

Web crawlers have become an important traffic consumers for most websites. The statistics of web crawlers visits show the importance of measuring the ethicality of crawlers. We analyze crawler-generated access logs from four different types of websites including a large scale academic digital library, an e-commerce website, an emerging small scale vertical search engine, and a testing website. The detailed statistics for each website are listed in Table II.

The access statistics show that more than 50% of the web traffic is generated by web crawlers on average and web crawlers occupies more bandwidth than human users in certain types of websites. Although the portion of crawler generated traffic varies for different types of websites, the contribution of crawler traffic to the overall website traffic are non-negligible. The crawler traffic is especially important for emerging websites. Figure 1 compares the crawler visits and user visits for an emerging specialized search engine as a function of date. Each point in Figure 1 corresponds to the total visits in that month. There are two dates (2007/06 and 2007/10) associated with news releases of the website. The comparison shows that user responses to news releases are much faster than crawlers. However, crawlers are more likely to revisit the website after they find it.



(a) Crawler visits

(b) User visits

Figure 1. The comparison of crawler visits and user visits as a function of date.

The crawler traffic statistics also checks whether each crawler visit is successfully regulated by the robots.txt files. The results show that up to 11% of crawler visits violated or misinterpreted the robots.txt files.

The *User-Agent* field is typically embedded in the HTTP request header according to the HTTP standards. This field is used for the webmasters to identify crawlers. A crawler user-agent string list[2] is used in our research to identify

[2] http://user-agents.org/

	Unique IPs	Crawler	User	Total Visits	Crawler	Violating REP	User
Website1	256,559	5,336 (2.08%)	251,223 (97.92%)	3,859,136	2,750,761(71.28%)	305,632(11.11%)	1,108,402(28.72%)
Website2	240,089	5,499(2.29%)	234,590(97.71%)	5,591,787	1,778,001(31.80%)	0	3,813,786(68.2%)
Website3	8,881	3,496(39.36%)	5,385(60.64%)	219,827	179,755(81.77%)	11,041(6.14%)	40,072(18.23%)
Website4	2,289	1,250(54.61%)	1,039(45.39%)	21,270	13,738(64.59%)	1,319(9.6%)	7,532(35.41%)

Table I
STATISTICS FOR IPS ADDRESSES VISITING FOUR WEBSITES IN ONE MONTH.

obvious crawlers. Since the field is specified from the client side, not all web crawlers can be identified in this manner. There are crawlers specifying themselves as web browsers intentionally [3]. Such crawlers have to be identified from their visiting patterns to the website including the average session length, the visit frequency, and occupied bandwidth. The sessions are identified by correlating the request and reference fields in the request header with page structures in websites. Furthermore, crawlers are not only "faking" browsers. Commercial crawlers' identities have also been used by other crawlers. For example, 46 crawlers named themselves as *Googlebot* whose IP addresses cannot be associated to *Google* by reverse DNS lookup in one-day's CiteSeer access log (see Figure 2). We use the blue color to circle out *Googlebots* that can be associated to *Google* and red color to circle out fake *Googlebots*.

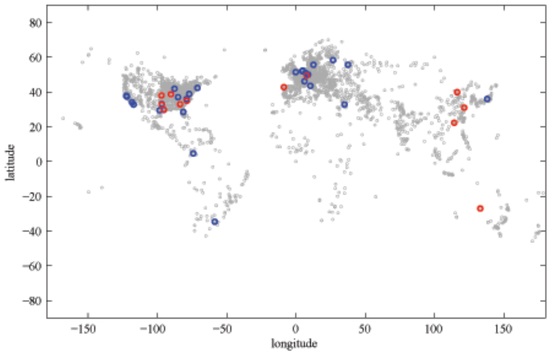


Figure 2. The geographical distribution of web crawlers named as *Googlebot*. The blue and red circles point out the well behaved and bad behaved *Googlebot* respectively.

The analysis of web crawler generated traffic shows that web crawlers become an important part of the web usage. Thus, measuring the ethicality of web crawlers is necessary.

IV. MODELS

This section describes the modeling of crawler behavior and ethical factors in our research. The notions and definitions are described in the following subsections.

A. Vector Model of Crawler Behavior

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. The notions of basic concepts related to ethicality are introduced as the following:

- A *rule* is a dimension in the rule space. Each rule is defined to be an item from a complete set of rules that describe all the regulations of web crawlers.
- For a given subset of N rules $R = \{r_1, r_2, ..., r_N\}$, a web crawler's behavior is defined as a N-vector $C = \{c_1, c_2, ..., c_N\}$ such that $c_i > 0$ if the crawler disobey the rule i and $c_i = 0$ otherwise.
- For a given subset of N rules $R = \{r_1, r_2, ..., r_N\}$, ethical weight is defined as a N-vector $W = \{w_1, w_2, ..., w_N\}$ in the rule space where w_i is the cost for disobeying rule i.
- For a given rule r, $N_v(C)$ is defined as the number of visits generated by crawler C violates or misinterprets the rule r and $N(C)$ is defined as the total number of visits generated by crawler C. $P(C|r) = \frac{N_v(C)}{N(C)}$ is the conditional probability of crawler C violates rule r.

B. Models of Measuring the Ethicality

The statistics of web access logs show that there are significant problems in crawler ethics. As discussed in [18], there can be consequences including denial of service, cost, and privacy if web crawlers do not crawl ethically. From a technical perspective, denial of service and cost refer to the same crawler behavior - generating overwhelming traffic to a website. Privacy refers to the behavior of crawlers accessing restricted content of a website. Web crawler generated ethical issues can be interpreted differently with different philosophy of determining ethicality. We introduce two models of ethicality measures to show how ethicality can be evaluated differently.

1) Binary Model: In certain cases, a crawler being unethical once is considered an unethical crawler. The binary model is based on this strong assumption and reflects whether a crawler has ever been violating any rules (see Eq1).

$$E_{bin}(C) = \mathbf{1}\left(\sum_{r \in R}(N_v(C)) > 0\right) \quad (1)$$

where $N_v(C)$ is the the number of visits generated by crawler C violating or misinterpreting robots.txt files, $\mathbf{1}(x)$ is indicator function where $\mathbf{1}(x) = 1$ when crawler visits have violations or misinterpretations, and $\mathbf{1}(x) = 0$ otherwise. Because of the strong assumption in the binary model, only

[3] http://www.munax.com/crawlingfaq.htm

clearly described rules without ambiguity in REP should be considered. For binary model, a crawler is ethical only if it obeys all rules at any time. Any exception classifies the crawler to unethical crawlers.

2) Cost Model: More generally, an ethical cost vector can be defined in the rule space. Each element of the cost vector represents the cost of violating corresponding rule. The ethicality $E_{cost}(C)$ of a crawler C is then defined as the inner product of the crawler vector and the ethical cost vector (see Eq2).

$$E_{cost}(C) =< C \cdot W >= \sum_i c_i \times w_i. \tag{2}$$

The ethical weight vector can be defined based on the actual costs of disobeying regulation rules in different circumstances. The elements of an ethical weight vector are not necessary to be a fixed value. Instead, cost functions can be applied to measure the ethicality for different crawlers in a better granularity. For small websites with limited internet bandwidth, crawlers that do not set proper visiting frequency or repeatedly crawling the same content disregarding the recency of the content may be considered unethical. Crawlers failed to update the restriction policy and resulted in keeping the outdated content to the public may considered unethical in e-commerce websites. Crawlers trying to access restricted information or information protected with authentication that is designed for the use in a limited group may also be considered unethical for many websites. Because of the differences in the content and functionality of websites, it is hard to measure the ethicality as a universal factor.

We propose two quantitative cost functions namely content ethicality E_c and access ethicality E_a to evaluate the ethicality of web crawlers from general regulation perspective. In content ethicality, cost is defined as the number of restricted webpages or web directories being unethically accessed. More specifically, for a set of websites $W = \{w_1, w_2, ...w_n\}$, $D(w_i)$ is the set of directories restricted in its robots.txt file of w_i. $V_C(w_i)$ is a subset of $D(w_i)$ which is visited by crawler C by violating or misinterpreting the rules in the robots.txt file. The content ethicality is defined in Equation 3.

$$E_c(C) = \sum_{w_i \in W} \frac{||V_C(w_i)||}{||D(w_i)||}. \tag{3}$$

The definition of content ethicality is intuitive. The more rules a crawler is violating, the higher the content ethicality score it will receive. According to the definition, the content ethicality score is a real number between 0 and 1.

Access ethicality can be defined as the visit interval of a crawler to a website with respect to the desired visit interval of the website. The desired visit interval for a website can be obtained from the *crawl-delay* rule in its robots.txt file. Since the web crawlers are automated programs that traverse the web, the visit interval for each website depends on the crawling policy of crawlers. In the access ethicality, we assume the visit interval for each crawler is proportional to the incoming links of a website (inlinks for each website can be estimated by the link results from Google). Thus, we can estimate the visit interval of each crawler for all websites by observing the visit interval for one website. For a set of sample websites $W = w_1, w_2, ...w_n$, the visit interval $interval_C(w_i)$ of crawler C for site w_i can be estimated by the visit interval $interval_C(w_a)$ for site w_a, $interval_C(w_i) = \frac{inlink(w_a)}{inlink(w_i)} \times interval_C(w_a)$. If a crawler obeys the *crawl-delay* rule, the lower bound of the interval is determined by the *crawl-delay* rules specified in robots.txt files. The access ethicality can be defined as Equation 4.

$$E_a(r) \quad = \sum_{w_i \in W} \frac{e^{-(interval_C(w_i) - delay(w_i))}}{1 + e^{-(interval_C(w_i) - delay(w_i))}} \tag{4}$$

For websites without *crawl-delay* entries, the default delay value is set to 0 since there is no desired crawl interval. According to the definition, the access ethicality score is normalized from 0 to 1 with lower scores representing crawlers that respect the desired crawling interval. When the crawlers obey the *crawl-delay* rule, the access ethicality scores are less than $1/2$.

V. Experiments

A. Crawler Behavior Test: Honeypot

The ideal way to get the ethicality scores for each web crawler would be to collect the access logs of all websites. However, this is impossible since websites typically do not share their access logs for a variety of reasons. To address the data collection issue, we set up a honeypot which records crawler behavior under different circumstances. The behavior traces collected from the honeypot can be used to induce the ethicality of all web crawlers visiting the honeypot, and used to estimate crawler ethicality at large.

The honeypot is designed based on the specifications in the Robots Exclusion Protocol and common rule cases derived from our prior study of 2.2 million sample websites with robots.txt files, including all cases where the robots.txt rules can be violated by crawlers. Table II shows the usage of rules in our robots.txt collection. Each website in the honeypot tests one or more specific case against each crawler.

Honeypot 1 This website tests crawlers' interpretation of $Disallow$ rule. There are two hyperlinks on the root page of the honeypot 1, $/d1/$ and $/d1/d01/$. The robots.txt file specifies two rules, $Disallow : /d1/$ and $Allow : /d1/d01/$. The rules should be interpreted as all files under directory $/d1/$ including $/d1/d01/$ are restricted based on the REP although the second rule conflicts the first one. The REP

Rule	Frequency
Contain disallow rules	913,330
Contain conflictions	465
Contain octet directives	500
Use variations of user-agent name	45,453
Use multiple user-agent name in one line	42,172
Use multiple directory in one line	3,464
Use crawl-delay rule	39,791
Use visit-time rule	203
Use request-rate rule	361

Table II
STATISTICS OF THE RULES IN ROBOTS.TXT FILES OF 2.2 MILLION WEBSITES.

specifies that “The first match found is used.” Thus, the confliction should be resolved by following the first rule. (To allow /d1/d01/ and disallow other directories in /d1/, the correct setting should be list the allow rules prior to the disallow rules.) In honeypot 1, if a crawler visits $/d1/$, it does not obey $Disallow$ rule. If a crawler visits $/d1/d01/$ but not $/d1/$, it does not resolve the subdirectory confliction correctly.

Honeypot 2 This website tests the crawler behavior when the rules are directly conflicting with each other. In this case, a hyperlink is pointing to a page under directory $/d1/$. The robots.txt file includes $Disallow : /d1/$ and $Allow : /d1/$ in the given order. There is obviously a confliction between the two rules. If the crawler visits pages under $/d1/$, it does not resolve the confliction.

Honeypot 3 This website has the same page structure as honeypot 1 but a different robots.txt file. There is only one rule, $Disallow : /d1/$. REP requires crawlers excluding any URLs starting with $/d1/$. Thus, $/d1/d01/$ being restricted is implied by the robots.txt file. If a crawler visits $/d1/d01/$, it fails to parse the rule correctly.

Honeypot 4 This website tests the robots.txt update policy of crawlers. The rules in robots.txt change after a period of time. Crawlers should update the robots.txt files periodically to respect the latest access regulation. If a crawler does not update robots.txt files as needed, restricted content may be accessed unethically. There are two pages $/d1/$ and $/d2/$ linked by the homepage of the website. $/d1/$ is restricted in time period 1 and allowed in time period 2. $/d2/$ is allowed in time period 1 and disallowed in time period 2. The timestamp of crawler requests to these two pages shows the crawlers’ update policy. This site provides evidence to study how search engines deal with web pages that change permissions over time.

Honeypot 5 This website tests how crawlers deal with errors in robots.txt. A few error lines are place before the rule $Disallow : /d4/$ in the robots.txt file. Thus, if a crawler obeys disallow rules in other honeypots but still visits web pages under $/d4/$, it is considered fail to handle the errors. It is not the crawlers’ responsibility to correct the errors. However, ethical robots should at least ignore the errors and still obey the correctly formatted rules.

Honeypot 6 This website tests how crawlers match the $User-Agent$ field. The robots.txt file in this site is dynamically generated by a server side script which parses the User-Agent string from the HTTP request header and generates substrings. The Robots Exclusion Protocol specifies that the $User-Agent$ field is case insensitive and a crawler should match a token that describes it. Substring and variations of crawler names are often used in the robots.txt files to specify the regulation rules. For example, $google$ as a crawler name is appeared in 16,134 robots.txt files in our collection. However, crawlers may ignore the name because it is not the exact name. There are six directories linked by the homepage of this website. Each directory is restricted to one substring of the incoming crawler except one open directory. In such settings, if a crawler requests the open directory, the one directory the crawler does not request shows the matching mechanism of the crawler. If all directories are requested by a crawler, it is considered unethical.

Honeypot 7 This website tests the rule for status codes regulation. A request to the robots.txt file in this website will receive a status code response of 403. According to REP, when the request to the robots.txt receives an HTTP status code of 403 or 401, “a robot should regard access to the site completely restricted.” If a crawler ignores the rule and visits any pages in the website, it is considered unethical.

Honeypot 8 This website tests whether a crawler is a spam crawler. In this website, there is no explicit link to the $/email/$ directory. However, $/email/$ is restricted in the robots.txt file. Thus, spambots can be identified if they try to explore the hidden directory $/email/$.

Honeypot 9 This website tests octet conversion related rules in the REP. If a %xx encoded octet is encountered, it should be un-encoded prior to comparison unless it is %2f. The page $/a-d.html$ is linked by the homepage of the website and the robots.txt file disallows $/a\%2dd.html$. Thus, crawlers should not crawl the file $/a-d.html$.

Honeypot 10 This website tests whether a crawler respects $META$-Tag rules specified in a webpage. The rule *¡META NAME=“ROBOTS” CONTENT=“NOINDEX,NOFOLLOW”¿* is specified in a webpage. According to the rule, any hyperlinks on this page should not be followed and corresponding webpages should not be requested.

Honeypot 11 This website tests whether a crawler respects the desired crawl delay. The robots.txt of this site sets the delay time to 200 seconds and then to 18000 seconds after two months. The interval between consecutive visits from one crawler is compared to the desired delay time. The access ethicality can be computed based on crawler behavior from this site.

Our settings of honeypot websites only test currently available robots regulation rules.

Rule	Probability of being violated
Conflict Subdir	4.6%
Conflict Dir	5.569%
Disallow Dir	0.969%
Timely Update Regulation	1.937%
Ignore Errors	0.726%
Crawl-Delay	6.78%
Substring of crawler name	1.695%
401/403 restriction	9.443%
Hidden Dir	0
Meta Tag	0.242%
Octec Conversion	0.968%
Case Sensitivity	3.39%

Table III
PROBABILITY OF A RULE BEING VIOLATED OR MISINTERPRETED.

User-agent	Content Ethicality
hyperestraier/1.4.9	0.95621
Teemer	0.01942
msnbot-media/1.0	0.00632
Yahoo! Slurp	0.00417
charlotte/1.0b	0.00394
gigabot/3.0	0.00370
nutch test/nutch-0.9	0.00316
googlebot-image/1.0	0.00315
Ask Jeeves/Teoma	0.00302
heritrix/1.12.1	0.0029
googlebot/2.1	0.00282
woriobot	0.00246
panscient.com	0.00202
Yahoo! Slurp/3.0	0.00136
msnbot/1.0	0.00049

Table V
CONTENT ETHICALITY SCORES FOR CRAWLERS VISITED THE HONEYPOT.

User-agent	Access Ethicality
msnbot-media/1.0	0.3317
hyperestraier/1.4.9	0.3278
Yahoo! Slurp/3.0	0.2949
Yahoo! Slurp	0.2949
Teemer	0.2744
Arietis/Nutch-0.9	0.0984
msnbot/1.0	0.098
disco/Nutch-1.0-dev	0.0776
ia_archiver	0.077
ia_archiver-web.archive.org	0.077
gigabot/3.0	0.0079
googlebot/2.1	0.0075
googlebot-image/1.0	0.0075

Table VI
ACCESS ETHICALITY SCORES FOR CRAWLERS VISITED THE HONEYPOT.

B. Results

We monitors the web access logs from within the honeypot and generates the statistics based on the crawlers that visits the honeypot. The probability of violating or misinterpreting each rule specified in our honeypot is listed in Table III.

The 401/403 restriction and the $crawl\text{-}delay$ rules have a high probability to be violated. Since $crawl\text{-}delay$ rule is an extension to the REP and 401/403 restriction rule is not explicitly specified in robots.txt files, the violation results are expected. However, the results also suggest the importance of making REP official and complete. There is also a high probability for crawlers misinterpreting the conflicting rules. It shows the needs for the author of robots.txt files to design and check their rules more carefully.

1) Binary Ethicality: The robot access logs for all the websites in our honeypot are analyzed to extract the binary ethicality vector of each crawler. The ethicality vector is constructed with each row representing a specific crawler and each column representing a specific rule (see Table IV).

Since not all crawlers visit every sub-site of our honeypot, crawlers should not be compared directly for the number of rules they violated or misinterpreted. For example, $Googlebot$ violates or misinterprets 7 rules in the rule space and $HyperEstraier$ violates 6. However, $HyperEstraier$ violates every rule it encountered. The binary ethicality vector shows that $MSNbot$ is more ethical than $Yahoo\ Slurp$ and $Yahoo\ Slurp$ is more ethical than *Googlebot*. The results also show that the interpretation of robots.txt rules for each crawler varies significantly.

2) Cost Ethicality: Table VI lists the content and access ethicality results for top crawlers that visited our honeypot during the time of the study.

The cost ethicality measure not only considers the penalty for not strictly obeying the rules in robots.txt files, but also grants additional ethical points to crawlers that exceed regulations in order to conduct themselves ethically. The additional cases include the following. (1) If a crawler encounters errors in robots.txt files but still parses and obeys the intention of the rules while crawling the site, it is considered more ethical than crawlers that ignore the errors. (2) Webmasters sometimes change the robots.txt rules during a crawl. Accessible directories may be restricted after the change. If a search engine respects the updated rules, it is considered more ethical than those that ignore the updates. (3) Many robots.txt files contain variations of robot names such as $Googlebot/1.0$. If a crawler notices the variations, it is considered more ethical than those ignore the variations. These behaviors are not measurable for all crawlers. Thus, we only show the results for known search engine crawlers (see Table VII).

User-agent	Parsing Errors	Update policy	Variation of Botname
Googlebot/2.1	yes	remove from index	no
ia_archiver	yes	n/a	yes
msnbot/1.0	yes	remove cache (still searchable)	no
Yahoo! Slurp	yes	remove from index	no

Table VII
ADDITIONAL ETHICAL BEHAVIOR OF CRAWLERS.

User-Agent	Conflict Subdir	Conflict Dir	Disallow Dir	Timely Update Regulation	Ignore Errors	Crawl-Delay	Substring of crawler name	401/403 restriction	Hidden Dir	Meta Tag	Octec Conversion	Case Sensitivity
Googlebot/2.1	1	1	0	1	0	1	0	1	0	0	1	1
Googlebot-image	1	1	0	1	0	1	0	1	0	0	1	1
hyperestraier/1.4.9	1	1	1	-	1	1	1	-	-	-	-	-
yahoo! slurp	1	1	0	1	0	1	0	1	0	0	0	1
msnbot/1.0	1	1	0	0	0	1	0	1	0	0	0	0
msnbot-media/1.0	1	1	0	0	0	0	1	0	0	0	0	0
nutch-0.9	0	0	0	0	0	1	1	1	0	0	0	1
gigabot/3.0	0	0	0	1	0	0	0	1	0	0	0	0
heritrix/1.12.1	0	0	0	0	0	1	1	1	0	0	0	0
panscient.com	0	0	0	0	0	1	0	1	0	0	0	1
charlotte/1.0b	0	0	0	0	0	1	-	1	0	0	0	0
visbot/2.0	0	0	0	0	0	0	0	1	0	0	0	0
arietis/nutch-0.9	0	0	0	0	0	0	0	1	0	0	0	0
Teemer Netseer	0	0	0	1	0	0	0	-	-	-	-	-
ask jeeves/teoma	0	0	0	0	0	1	0	-	-	-	-	-
netcraft web server survey	0	0	0	0	0	0	0	0	0	0	0	0
ia_archiver	0	0	0	0	0	1	0	0	0	0	0	0
botseer/1.0	0	0	0	0	0	0	0	0	0	0	0	0

Table IV
The violation table of web crawlers derived from honeypot access logs.

C. Discussion of Unethical Behavior

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, obvious reasons may be inferred from the type of rules these crawlers fail to obey. For example, *Googlebot* ignores the $Crawl - Delay$ rule. However, the access ethicality for *Googlebot* is very good which means the crawling interval between two visits are longer than the actual $Crawl - Delay$ rule. It is possible that the policy makers of *Googlebot* believe that *Googlebot* has a better policy than the webmasters. *Googlebot* also resolves conflictions by matching the rules allowing it to crawl more web pages. There is an obvious advantage of obtaining more information by interpreting the conflictions the $Google$ way. Since the Robots Exclusion Protocol is not an official standard, they can be interpreted with their own understandings.

VI. Conclusions

With the growing role of autonomous agents on the internet, the ethics of machine behavior has become an important factor in the context of web crawling. Web crawlers are highly automated with built-in intelligence and are seldom regulated manually. Crawler-generated visits may result in significant issues of ethics, which until now have not been studied quantitatively.

We propose a set of ethicality models to measure the ethicality of web crawlers and to some degree the intelligence of web crawlers. A honeypot, a set of websites where each site is configured with a distinct regulation specification using the Robots Exclusion Protocol, is constructed to capture specific behaviors of web crawlers. The results generated by our determination of robot ethicality show that commercial crawlers are typically very ethical. However, many commercial crawlers were found to violate some rules of the REP. The comparison between the ethicality and crawler favorability scores shows that there is no clear correlation between the two measures.

We propose the first computational models to evaluate the ethicality of web crawlers. The results may present some bias due to limited sampling. Crawlers that did not visit or only visited part of our honeypot result in missing data and an incomplete, though representative, account of the behavioral ethics of crawlers on the web. Future research will focus on bringing more crawler traffic to a larger honeypot in order to collect a larger sample of crawler behavior and induce the rules that govern robot behaviors.

References

[1] C. Allen, W. Wallach, and I. Smit. Why machine ethics? *IEEE Intelligent Systems*, 21(4):12–17, July 2006.

[2] V. Almeida, D. A. Menasce, R. H. Riedi, F. Peligrinelli, R. C. Fonseca, and W. M. Jr. Analyzing robot behavior in e-business sites. In *SIGMETRICS/Performance*, pages 338–339, 2001.

[3] M. Anderson, S. L. Anderson, and C. Armen. Towards machine ethics. In *Proceedings of AAAI 2005 Fall Symposium*, 2005.

[4] M. L. Boonk, D. R. A. d. Groot, F. M. T. Brazier, and A. Oskamp. Agent exclusion on websites. In *Proceedings of The 4th Workshop on the Law and Electronic Agents*, 2005.

[5] V. J. Calluzzo and C. J. Cante. Ethics in information technology and software use. *Journal of Business Ethics*, 51(2):301–312, May 2004.

[6] M. D. Dikaiakosa, A. Stassopouloub, and L. Papageorgiou. An investigation of web crawler behavior: characterization and metrics. *Computer Communications*, 28:880–897, 2005.

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

[8] L. Floridi. Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology*, 1(1):33–52, 1999.

[9] D. G. Johnson. *The Blackwell Guide to the Philosophy of Computing and Information*, chapter Computer Ethics, pages 65–75. Wiley-Blackwell, 2003.

[10] R. A. Jones. The ethics of research in cyberspace. *Internet Research: Electronic Networking Applications and Policy*, 4(3):30–35(6), 1994.

[11] D. Lin and M. C. Loui. Taking the byte out of cookies: privacy, consent, and the web. In *ACM POLICY '98: Proceedings of the ethics and social impact component on Shaping policy in the information age*, pages 39–51, New York, NY, USA, 1998. ACM.

[12] M. A. Pierce and J. W. Henry. Computer ethics: The role of personal, informal, and formal codes. *Journal of Business Ethics*, 15(4):425–437, 1996.

[13] D. C. Sicker, P. Ohm, and D. Grunwald. Legal issues surrounding monitoring during network research. In *IMC '07*, 2007.

[14] J. A. Smith, F. McCown, and M. L. Nelson. Observed web robot behavior on decaying web subsites. *D-Lib Magazine*, 12(2), 2006.

[15] Y. Sun, I. G. Councill, and C. L. Giles. Botseer: An automated information system for analyzing web robots. In *ICWE*, 2008.

[16] Y. Sun, Z. Zhuang, I. G. Councill, and C. L. Giles. Determining bias to search engines from robots.txt. In *International Conference on Web Intelligence*, 2007.

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

[18] 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.