Web Robot Detection - the Influence of Robots on Web Mining

Christian Bomhardt and Wolfgang Gaul

Institut für Entscheidungstheorie und Unternehmensforschung
University of Karlsruhe (TH)
76128 Karlsruhe, Germany

Abstract. Web usage mining relies on web server logfile data. Parts of this data originate from web robots. This can - with repsect to the original aims of web mining - lead to contradicting decisions based on distorted results. We describe possibilities of web robot detection and give examples how, e.g., e-metrics and results of association rule algorithms can differ based on raw logfiles versus those that consist of requests of human users or web robots.

1 Introduction

1993 is the year where some of the first known web robots appeared in the internet [24], e.g., "Wanderer" by Matthew Gray (measuring web growth) and "JumpStation" by J. Fletcher (indexing). At that time, commercial web usage played a minor role, instead overloaded web servers or waste of bandwith were areas for robot deployment problems. Nevertheless, those problems together with a growing number of newly used robots led to a standard for robot exclusion ([18]). Today, most robots adhere more or less strictly to existing guidelines for robots. With about 16% of the web traffic originating from robots ([19]), nowadays, robot detection must be considered for serious web mining efforts. While cooperative robots can be detected with the help of a simple heuristics (as we will see later) malignant robots ignore the guidelines mentioned above and may even apply stealth technologies. Generally, there is not much known about malignant robots as their usage on the net is somehow "unethical" (e.g., extraction of mail addresses for spamming ([17]), unauthorized usage of US Government's Weather Service ([3])).

There exist four major categories of widely used robot detection technologies: *Simple methods*, *traps*, *web navigation behavior analysis*, and *navigational pattern modeling*. Simple methods check the [request], [agent], and [IP address] fields in webserver logfile entries against lists of known robot identifications ([5]). Traps consist of links within the HTML pages that are invisible for a human user. If such a link is visited, it must have been visited by a robot ([20]). Web navigation behavior analysis searches for typical navigation characteristics based on the objectives of the different classes of robots ([2]). Navigational pattern modeling calculates session attributes and applies

data mining algorithms that try to detect robots on the basis of selected session attributes ([22], [23], [10]). Malignant robots can be detected via traps, web navigation behavior analysis, and navigational pattern modeling. Simple methods are unlikely to detect malignant robots which are relatively sparse in the data. Thus, their detection requires additional efforts. This is why we concentrate on cooperative robots in this paper. With the increasing number of robots and changes of the identification information of known ones, the problem of keeping robot lists current and complete becomes more and more laborious. We faced this problem with the development of the robot detection tool RDT ([8]) - a specialized web logfile preprocessing software enabling researchers to effectively work with and understand large logfile data - one of the main requirements to effectivly remove robot requests from web usage data and to accomplish further web mining steps.

In the follwing the *Web Data Preprocessing* process will be divided into the substeps sessionizing, and robot detection. Both steps are supported by the RDT software. It speeds up preprocessing and therefore enables researchers to focus on their specific mining tasks.

2 Web Data Preprocessing

Every webserver records served HTTP-requests in its logfile. In the following, the wide spread combined logfile format ([4]) is used, which most HTTP servers can create. This format contains the following nine fields: [IP address] as client IP address, [name] as name of the user (usually unused), [login] as login-name of the basic HTTP-authentication, [date] as date and time of the request, [request] as HTTP-request containing the request method, the URL of the requested resource (page), and the desired HTTP-protocol, [status] as 3-digit status code returned by the server, [size] as number of bytes actually returned by the server, [referrer] as URL of the referencing page and [agent] as name of the client agent. HTTP-requests are the basis for web usage mining ([13], [14]).

2.1 Sessionizing

Each HTTP-request is written in the order of occurance into the server logfile and the construction of contiguous user requests requires further efforts. An overview of different types of sessionizers and their performance is given in [7]. In [13] a navigational path construction algorithm as basis for establishing sessions is described. For this paper, we selected a widely used heuristics from [12] where requests with the same agent and IP address are grouped together as long as the maximum idle time between two requests is smaller than 30 minutes. A common user session consists of two kinds of requests: main requests as result of an user action and auxiliary requests. Auxiliary requests are automatically issued by browsers to retrieve objects referenced

by the main request (e.g., stylesheets, images, background sound). If possible, we try to assign auxiliary requests - on basis of the referrer field - to their main requests.

2.2 Robot Detection

Sessions are seen as sessions of human users until hints for the opposite are found. Sessions with requests for files that are known to be never requested by human users (e.g., `robots.txt`, some hidden linked files from traps ([20] or typical files from worm attacks (e.g., `cmd.exe` for Nimbda ([16])) provide good hints for robot detection. All these files are stored in the *trapfile list.* Robots that obey to the robot exclusion standard identify themselves with an agent tag that can be recognized as the agent of a robot. Those tags are collected in the *robot agent list.* Some IP adresses are known to be solely used by robots (e.g., the google bots) and can be saved in the *robot IP list.* The composition of these lists can be simplified by assimilating the list of known robots available from `http://www.robotstxt.org/wc/robots.html` ([24]). Figure 1 summarises what could be called robot detection heuristics.

```
function IsRobotSession( Session )
{
if (session contains a request of a file from the trapfile list)
   then return TRUE;
if (session agent is contained in the robot agent list)
   then return TRUE;
if (session IP is contained in the robot IP list)
   then return TRUE;
return FALSE;
}
```

Fig. 1. Robot detection heuristics

The magnitude of logfile data requires the application of web mining tools. Specialized software like our RDT can support researchers with respect to data understanding and functions of data preprocessing. Because of space restrictions properties of RDT have to be described in a different paper.

3 Empirical Results

We tried to find out whether robots influence web mining results. The logfile used in this paper originates from a mid-sized online shop. The whole dataset (all sessions), the part without robots (user sessions, 73%), and the part without users (robot sessions, 27%) were examined and compared. First, we selected some simple but wide-spread features (e.g., top N, top entry, top exit

pages) which are used by many common logfile analysing tools like webalizer ([6]). Second, we calculated selected e-metrics from [21] and micro conversion rates from [15] for the three datasets. Third, as a more specific application, we looked for association rules ([1],[13],[14]) within the different datasets and compared the results.

The results of the simple features are summarized in table 1 (Top N list of most frequently requested pages), table 2 (Top entry list of most frequently requested entry pages), and table 3 (Top exit list of most frequently used exit pages). The shadowed outcomes show that 7 positions out of the first 16 highest ranked pages of the top N list are affected by robots. A similiar, even stronger effect can be observed in the top entry list (11 differences out of the first 16 positions) and also in the top exit list (6 differences out of 16). Results of this kind can be used for improvements of the website.

Table 1. Top N list of most frequently requested pages

URL of page	All sessions Position	User sessions Position	Robot sessions Position
/	1	1	1
/shop/show/de	2	2	5
/shop/show/show_basket.php3	3	3	3
/shop/templates/basket.php3	4	4	11
/shop/show/show_search.php3	5	5	65
/zaubertricks.htm	6	6	17
/de/termine.htm	7	13	2
/de/gewinnspiel.htm	8	7	6
/de/index.htm	9	10	4
/shop/templates/order_adrform.php3	10	8	-
/shop/templates/order_finish.php3	11	9	-
/de/information.htm	12	11	7
/de/kontakt.htm	13	12	8
/shop/images/tr.gif/	14	14	-
/admin/onedit/fr_leer	15	15	-
/de/newsletter.htm	16	16	10

The influence of robots on the generic e-metrics hit-to-visit(1) (percentage of sessions with more than 1 request), hit-to-visit(2) (percentage of sessions with more than 2 requests), and avg. visit depth (number of pageviews) is rather small in contrast to the strong distortion with respect to the avg. visit duration (cp. table 4). The influence of robots on micro conversion rates as visit-to-basket (percentage of visits with basket usage), basket-to-buy (percentage of sessions with basket usage that lead to purchases), and visit-to-buy (percentage of visits with purchases) is small but relevant (cp. table 5). Here, one can see that robots never buy.

Table 2. Top entry list of most frequently requested entry pages

URL of entry page	All sessions Position	User sessions Position	Robot sessions Position
/	1	1	1
/shop/show/de	2	2	4
/zaubertricks.htm	3	3	12
/de/gewinnspiel.htm	4	5	5
/de/index.htm	5	8	3
/shop/show/show_basket.php3	6	6	16
/shop/show/show_search.php3	7	4	148
/shop/templates/order_adrform.php3	8	7	-
/de/information.htm	9	10	7
/kartentricks.htm	10	9	32
/de/termine.htm	11	12	2
/de/kontakt.htm	12	11	6
/de/newsletter.htm	13	13	9
/de/links.htm	14	17	13
/de/agb.htm	15	19	11
/taschenspielertricks.htm	16	14	46

Table 3. Top exit list of most frequently used exit pages

URL of exit page	All sessions Position	User sessions Position	Robot sessions Position
/	1	1	1
/shop/show/de	2	2	3
/zaubertricks.htm	3	3	14
/shop/show/show_basket.php3	4	4	18
/shop/show/show_search.php3	5	5	60
/de/gewinnspiel.htm	6	6	7
/de/kontakt.htm	7	7	6
/de/index.htm	8	8	4
/de/termine.htm	9	12	2
/kartentricks.htm	10	9	34
/shop/images/tr.gif/	11	10	-
/de/information.htm	12	11	5
/de/agb.htm	13	13	10
/de/links.htm	14	14	12
/de/newsletter.htm	15	16	9
/de/wir_ueber_uns.htm	16	18	13

Table 4. Results of e-metrics analysis

Selected e-metrics	All sessions	User sessions (73%)	Robot sessions (27%)
Number of sessions	52295	38227	14068
Hit-to-visit(1)	37363 (71%)	29728 (77%)	7635 (54%)
Hit-to-visit(2)	29700 (56%)	24971 (65%)	4729 (33%)
Avg. visit depth	16	16.6	14.5
Avg. visit duration	580s	236s	1512s

Table 5. Results of micro conversion rate calculations

Micro conversion rate	Sessions with more than 1 pageview		
	All sessions	User sessions	Robot sessions
Visit-to-basket	3.5%	4.2%	0.7%
Basket-to-buy	31.4%	32.7%	0%
Visit-to-buy	1.0%	1.4%	0%

Table 6. Selected results of association rule mining

	All sessions	User sessions	Robot sessions
Number of rules found	9	19	3
Number of interesting rules found	4	6	3
Association rule	(Support,Confidence)		
{/shop/show/de} ⇒ {/}	(30.1%,58.7%)	(37.6%,59.4%)	-
{/zaubertricks.htm} ⇒ {/}	(10.2%,68.1%)	(12.8%,69.8%)	-
{/de/information.htm} ⇒ {/shop/show/de}	(3.1%,65.5%)	(3.6%,62.9%)	-
{/de/information.htm, /shop/show/de} ⇒ {/}	(3.1%,65.3%)	(3.5%,64.3%)	-
{/shop/show/show_search.php3} ⇒ {/shop/show/de}	-	(3.7%,50.5%)	-
{/de/gewinnspiel, /shop/show/de} ⇒ {/}	-	(3.0%,64.5%)	-
{/de/newsletter.htm} ⇒ {/de/gewinnspiel.htm}	-	-	(3.0%,53.1%)
{/de/newsletter.htm} ⇒ {/de/index.htm}	-	-	(3.0%,52.2%)
{/de/newsletter.htm} ⇒ {/de/kontakt.htm}	-	-	(3.0%,51.3%)

AprioriPre ([9]) and apriori ([11]) were used for association rule mining. Support was set to 3% and confidence to 50% to limit the number of rules. Table 6 displays the rules which where found. Uninteresting rules reproduce facts already known (order_finish ⇒ show_basket is such a rule if one cannot purchase without usage of the basket). Based on the selected support and confidence parameters the underlying dataset contained 4 interesting rules for all sessions. Two additional rules were found when only user sessions were evaluated. For the robot dataset completely different rules would have to be taken into consideration.

4 Conclusions and Outlook

Empirical results show that robots can seriously influence web mining. Some new aspects and distortions were discovered thanks to our robot detection efforts. The developed robot detection tool RDT can simplify and accelerate the logfile preprocessing step. It enabled us to effectively remove robot activities in large logfiles and it facilitated data understanding. The problem of detecting malignant robots will be addressed in a forthcoming paper.

References

1. Agrawal, R., Srikant, R. (1994) Fast Algorithms for Mining Association Rules, Proc. 20th Int. Conf. Very Large Data Bases, VLDB
2. Almeida, V., Riedi, R., Menascé, D., Meira, W., Ribeiro, F., Fonseca, R. (2001) Characterizing and Modeling Robot Workload on E-Business Sites, Proc. 2001 ACM Sigmetrics Conference, `http://www-ece.rice.edu/~riedi/Publ/RoboSimg01.ps.gz`
3. Anaconda Partners LLC: Anaconda! Foundation Weather, `http://anaconda.net/ap_wxdemo.shtml`
4. Apache HTTP Server Documentation Project: Apache HTTP Server Log Files Combined Log Format, `http://httpd.apache.org/docs/logs.html\#combined`
5. Arlitt, M., Krishnamurthy, D., Rolia, J. (2001) Characterizing the Scalability of a Large Web-Based Shopping System, ACM Transactions on Internet Technology, `http://www.hpl.hp.com/techreports/2001/HPL-2001-110R1.pdf`
6. Barrett, B. (2001) Webalizer, `http://www.mrunix.net/webalizer/`
7. Berendt, B., Mobasher, B., Spiliopoulou, M., Wiltshire, J. (2001) Measuring the Accuracy of Sessionizers for Web Usage Analysis, Proceedings of the Web Mining Workshop at the First SIAM International Conference on Data Mining, Chicago
8. Bomhardt, C. (2002) The Robot Detection Tool, `http://www.bomhardt.de/bomhardt/rdt/produkt.html`
9. Bomhardt, C. (2003) AprioriPre, `http://www.bomhardt.de`
10. Bomhardt, C., Gaul, W., Schmidt-Thieme, L. (2003) Web Robot Detection - Preprocessing Web Logfiles for Robot Detection, Working paper, Institut für Entscheidungstheorie und Unternehmensforschung

11. Borgelt, C. (2003) Apriori, a Program to Find Association Rules With the Apriori Algorithm, http://fuzzy.cs.uni-magdeburg.de/~borgelt/software.html
12. Catledge, L., Pitkow, J. (1995) Characterizing Browsing Strategies in the World-Wide Web, Computer Networks and ISDN Systems
13. Gaul, W., Schmidt-Thieme, L. (2000) Frequent Generalized Subsequences - A Problem from Webmining, in: Gaul, W., Opitz, O., Schader, M. (eds.): Data Analysis, Scientific Modelling and Practical Application, Springer, Heidelberg , 429-445
14. Gaul, W., Schmidt-Thieme, L. (2002) Recommender Systems Based on User Navigational Behavior in the Internet, Behaviormetrika, Vol. 29, No.1, 1-22
15. Gomory, S., Hoch, R., Lee, J., Podlaseck, M., Schonberg, E. (2000) E-Commerce Intelligence: Measuring, Analyzing, and Reporting on Merchandising Effectiveness of Online Stores, Working paper, IBM T.J. Watson Research Center
16. Heng, C. Defending Your Web Site / Server From the Nimbda Worm / Virus http://www.thesitewizard.com/news/nimbdaworm.shtml
17. Ipaopao.com software Inc. Fast Email Spider For Web, http://software.ipaopao.com/fesweb/
18. Koster, M. (1994) A Standard for Robot Exclusion, http://www.robotstxt.org/wc/norobots-rfc.html
19. Menascé, D., Almeida, V., Riedi, R., Ribeiro, F., Fonseca, R., Meria, W. (2000) In Search of Invariants for E-Business Workloads, Proceedings of ACM Conference on Electronic Commerce, Minneapolis, MN, http://www-ece.rice.edu/~riedi/Publ/ec00.ps.gz
20. Mullane, G.S. (1998) Spambot Beware Detection, http://www.turnstep.com/Spambot/detection.html
21. NetGenesis (2000) E-Metrics: Business Metrics for the New Economy, NetGenesis Corp.
22. Tan, P.-N., Kumar, V. (2000) Modeling of Web Robot Navigational Patterns, Proc. ACM WebKDD Workshop, 2000
23. Tan, P.-N., Kumar, V. (2001) Discovery of Web Robot Sessions Based on their Navigational Patterns, http://citeseer.nj.nec.com/cache/papers/cs/22262/http:zSzzSzwww.cs.umn.eduzSz~ptanzSzdmkd.pdf/tan02discovery.pdf
24. The Web Robots Pages www.robotstxt.org/wc/robots.html