Distinguishing Humans from Robots in Web Search Logs: Preliminary Results Using Query Rates and Intervals

Omer Duskin Dror G. Feitelson
School of Computer Science and Engineering
The Hebrew University of Jerusalem
91904 Jerusalem, Israel

ABSTRACT

The workload on web search engines is actually multiclass, being derived from the activities of both human users and automated robots. It is important to distinguish between these two classes in order to reliably characterize human web search behavior, and to study the effect of robot activity. We suggest an approach based on a multi-dimensional characterization of search sessions, and take first steps towards implementing it by studying the interaction between the query submittal rate and the minimal interval of time between different queries.

1. INTRODUCTION

Web search logs, which are maintained by all large scale web search engines, are an important tool for studying and understanding web search behavior, and for the design and optimization of search engines. However, upon inspection of these logs, one finds that they contain various anomalies that "don't make sense". One simple example, that is also easy to clean, is replicated lines, as if the same user issued the same query twice within less than a second. Another example is "users" that are actually a meta-search engine, and forward queries on behalf of many real users. Such actions have a large effect on the perceived statistics of user behavior, so they need to be identified and handled prior to the analysis [2, 5].

Meta-engines are just one example of the general problem of multi-class workloads. In general, various types of robots my masquerade as users and issue queries to search engines. Correct separation of humans from such robots has important implications for both web search analysis and for the design of future systems. For example, it has been suggested that software agents use Boolean operators extensively while human users do not [1]. This would have implications for designing the user interface and for providing a separate programmatic interface for agents.

Our goal is to design and investigate ways to distinguish between human and robot search users. This will be used for two purposes. First, we want to filter out all non-human activity, in order to enable a more reliable characterization of human search behavior. Second, we want to catalog and describe the different robot profiles that are encountered in real systems, in order to provide a basis for assessing their impact on search engines and the services they provide.

Somewhat surprisingly, there has been very little previous research regarding this problem in the context of web search. Most previous work we know of is in the context of identifying robots in web sites such as e-commerce sites [4, 8]. Techniques used are often irrelevant for identifying robots in web search logs, because these logs do not contain the same information as that contained in general web server logs. For example, we do not know whether a user is identified is a robot in the user agent field, or whether a request was made for the robots.txt file. Other approaches, such as the observation that robots typically avoid downloading images, are also inapplicable.

Jansen and Spink, in their analyses of web search logs, simply use a threshold of 100 queries to distinguish humans from robots: a user who submitted less than 100 queries is assumed to be a human, and one who submitted more is assumed to be a robot [7, 5]. This is based on the argument that such a threshold is much higher than the average session length, but still there exist sessions that are much longer. Noting that their logs typically span a single day of activity, this can also be interpreted as setting the threshold at 100 queries per day. Buzikashvili suggested a somewhat lower threshold of 5–7 unique queries per hour [1]. To the best of our knowledge more involved approaches have not been studied.

Our approach is to study the activity patterns of search users along multiple dimensions, and try to build a classifier that distinguishes humans from robots by considering the complete profile of activity of each user. This is made more challenging by the fact that we do not have any precise information to begin with, as available logs do not include an indication of the nature of each user. Therefore we need to use heuristics such as those mentioned above. The dimensions that we intend to investigate include the following:

- The average rate or number of queries submitted. We prefer rate over an absolute number to accommodate logs of different durations, but in this work we use the number.
- The minimal interval between successive queries. Humans are expected to require tens of seconds or more to submit new queries.
- The rate at which users type, with humans limited to

Workshop on Web Search Click Data '09 Barcelona, Spain
Copyright 2009 ACM 0-89791-88-6/97/05 ...$5.00.

Table 1: *Data sources.*

Log	Timespan	Total entries
AtW01 (AlltheWeb)	6 Feb 2001	1,257,942 clicks
AV02 (AltaVista)	8–9 Sep 2002	2,659,315 queries
MSN06	1–31 May 2006	14,921,285 queries

around 200 characters per minute. This is similar to the above, but incorporates the lengths of the different queries rather than just using their number.

- The duration of sessions of continuous activity. If a user is active continuously for say more than 10 hours, we'll assume it is a robot.
- Correlation with time of day, which humans expected to be much more active during daytime. Note that this is based on the tacit assumption that users are in about the same time zone as the search engine, which may not always be true.
- The regularity of the submitted queries. Users who submit the same query thousands of times, or submit queries at precisely measured intervals, are assumed to be robots.

In this paper we take initial steps to characterize the first two items and the interaction between them. Specifically, we set thresholds on one attribute (say the number of submitted queries), and investigate the differences between the resulting groups of users in terms of the distribution of the other attribute (say the minimal interval between successive queries). We also consider using two thresholds rather than only one. For example, users who submit fewer than the bottom threshold are considered to be human, those who submit more than the top threshold are considered to be robots, and those that fall in between are left unclassified.

2. DATA SOURCES AND METHODOLOGY

Several web search logs are now available for research. Previous work has shown that these logs differ in many aspects [6]. It is therefore important to use multiple datasets when investigating search behavior, and to try and find invariants that are supported by all of them. In this work we use three datasets as described in Table 1.

All the logs record each transaction together with a timestamp at second resolution. Thus it is highly unlikely that two identical lines will be logged, as this implies that the same user issued the same query (or clicked on the same link) twice within one second. However, such repetitions do in fact occur. For example, in the AltaVista log, 39,461 of 2,659,315 lines were exact repetitions, which is 1.48% of the total. Possible explanations of such phenomena are that communication problems caused the same request to be delivered more than once, or that they are simply logging errors. In either case, exact repetitions are easy to filter out. However, even exact repetitions may actually be requests for additional pages of results by a robot (in some logs, these are not clearly distinguished). Therefore it is not clear that it is advisable to filter out repetitions at all.

Another problem is the definition of user or "source". All the logs tag entries with an anonymized source ID, which may be based on the user's IP address or on a cookie deposited with the user's browser. When looking at the queries, it is sometimes relatively easy to see that the source is actually a software agent. For example, in the AltaVista log, there is one source that has 22,580 queries, typically in sequences that arrive exactly 5 minutes apart (many such sequences are interleaved with each other). Of these, 16,502 are the query "britney spears", 868 are "sony dvd player", and 615 are the somewhat more surprising "halibut". Other cases are less obvious, and seem to be a random interleaving of multiple request streams. This is conjectured to arise due to either of two possible scenarios. The first is that network address translation (NAT) is at work, and the source address actually represents a whole network. The second is that this source is actually a meta-search engine which forwards the queries of multiple real human users. This is done to achieve improved results by combining the results of multiple basic search engines [3].

3. RESULTS

We analyzed our logs according to the above criteria, in an attempt to find a definition that leads to consistent results regarding the behavior of the robot and the human users. Note that we always removed repeated entries, and we aren't using it as a criterion. In the following, we use one user attribute for the classification and check the effect on the other attribute. For instance, we used the number of queries criterion to distinguish between human users and robots, and than we compared the humans' delay between queries and the robots' delay between queries.

3.1 Classification by Number of Queries

First, we analyzed the different logs with the number of queries criterion. This is essentially the same criterion that was used by Jansen and Spink, except that we experiment with different thresholds. This was done in order to overcome the fact that logs may have different semantics. For example, AlltheWeb 2001 provides clickthrough data, so one would expect more entries per user relative to the other logs that just record queries. There may also be a difference between logs that record the source IP address versus those that provide a cookie-based session ID. The parameters we used initially and the resulting classification are described in Table 2.

Given the tentative query-based classification into humans and robots, we calculated the distribution of the minimal intervals between queries for the users in each group. The logic behind this metric is that we expect it to take longer for a human to submit another query. We distinguish between different queries and resubmitting the same query a number of times (which happens when a user clicks on a few links from the result page, or requests additional results pages). The results show considerable overlap between the distributions, but the distributions for humans do tend to include higher values (Fig. 1; note that the X axis is logarithmic). Also, the intervals between repetitions of the same query are different than between different queries for humans, but the two distributions are very similar for robots, as may be expected. This raises our confidence that the original classification is meaningful.

The results for the AV02 log exhibit a peculiar peak at around 300 seconds (5 minutes), which exists only for the distribution of repetitive submissions of the same query, and is especially pronounced for users that have been classified as robots (Fig. 2). This is quite obviously the result of robots

Table 2: *Results of classification by number of queries.*

Log	Humans		Robots		Others
	Definition	Number	Definition	Number	Number
AtW01	<50 queries	150,828 (98.03%)	>300 queries	152 (0.10%)	2,876 (1.87%)
AV02	<10 queries	252,207 (87.00%)	>10 queries	33,239 (11.5%)	4,446 (1.53%)
MSN06	<10 queries	7,370,868 (98.66%)	>10 queries	76,666 (1.02%)	23,380 (0.31%)

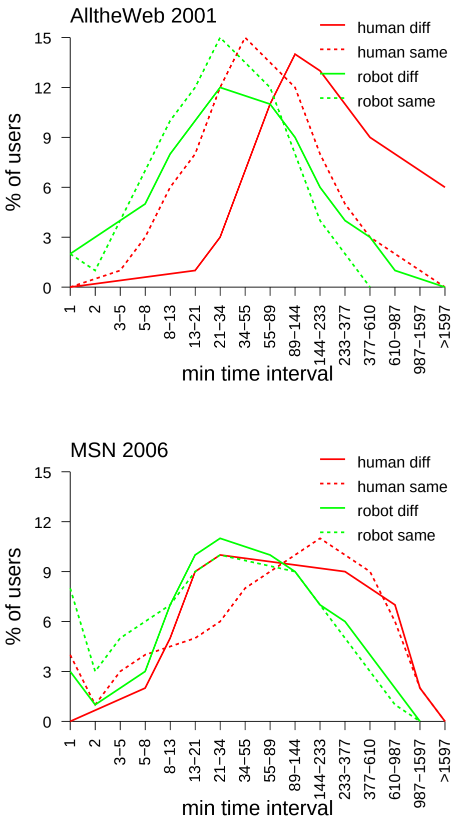

Figure 1: *Distributions of intervals between successive queries for users identified as human or robots according to their query rate.*

that indeed submit the same queries repeatedly at 5 minute intervals.

The precisely repetitive behavior of these robots provides testimony that they are indeed robots and not humans — testimony that is generally lacking from our data. This also indicates that the original classification is somewhat deficient, as several robots with this repetitive behavior had been classified as humans by mistake. Specifically, the "extra" peak in the humans distribution contains about 8% of the queries submitted by users classified as humans, so this is a lower bound on the error for this particular log.

3.2 Classification by Minimal Interval Between Queries

We next classified users using the minimal interval between different queries. The assumption is that humans cannot submit a new query within a second of a previous different query (we used intervals between different queries because they tend to be higher than intervals between repetitions of the same query). Therefore users with an interval of up to 1 second are considered robots. To be classified as humans we use a threshold of more than 10 or 25 seconds between different queries (we used a lower threshold for the AV02 log because of the short delays in that log). This leads to the results shown in Table 3.

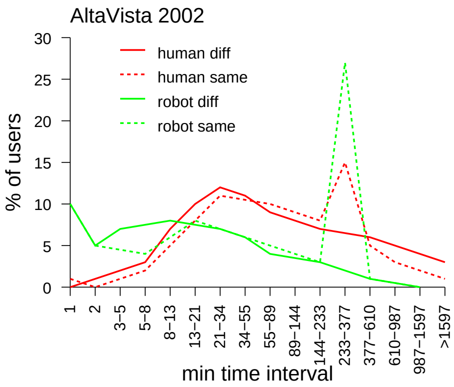

Figure 2: *Distributions for AltaVista exhibit a peak at 300 seconds, which probably represents robots with a periodic behavior.*

Based on this classification, we now checked the effect on the distribution of the number of queries submitted by users in the two classes (Fig. 3). Note that both axes are log scaled, and that users who submitted only a single query are not included, as they do not have an interval that could be used for classification. While the distributions have significant overlap, it is clear that the distribution for those users classified as humans has much more weight at low values, indicating that most of them only submit few queries. Users classified as robots tended to submit many more queries. This is especially noticeable in the AtW01 log. However, in the MSN06 log, there were much fewer users who submit many queries, including those classified as robots.

3.3 Effect of parameters

The above results are for given choices of the threshold parameters. We also experimented with other threshold values. The results are shown in Table 4. Based on this data, we can conclude that in any case a very large fraction of the users are most probably human, and only a very small fraction are robots. However, due to the high level of activity by robots, the number of queries they submit is much higher than their relative number. Also, it seems that the threshold values need to be specifically adjusted for each log.

To better understand the effect of the thresholds and the

Table 3: *Results of classification by minimal interval.*

Log	Humans		Robots		Others
	Definition	Number	Definition	Number	Number
AtW01	>25 seconds	147,575 (95.92%)	<1 seconds	852 (0.55%)	5,429 (3.53%)
AV02	>10 seconds	258,388 (89.13%)	<1 seconds	3,140 (1.08%)	28,364 (9.78%)
MSN06	>25 seconds	6,633,581 (88.79%)	<1 seconds	25,725 (0.34%)	811,608 (10.86%)

Table 4: *Comparison of different thresholds.*

Log	Humans		Robots		Others
	Definition	Number	Definition	Number	Number
AtW01	<50 queries	98.00%	>300 queries	0.01%	1.90%
	<60 queries	98.49%	>80 queries	1.00%	0.51%
	>25 seconds	95.92%	<1 seconds	0.55%	3.53%
	>10 seconds	98.30%	<1 seconds	0.55%	1.14%
MSN06	<50 queries	99.98%	>250 queries	0.00%	0.01%
	<10 queries	98.66%	>10 queries	1.02%	0.10%
	>25 seconds	87.31%	<1 seconds	0.34%	12.34%
	>10 seconds	95.97%	<1 seconds	0.34%	3.67%
	>5 seconds	95.98%	<1 seconds	1.20%	2.83%
AV02	<10 queries	87.00%	>10 queries	11.47%	1.53%
	<8 queries	82.87%	>9 queries	13.00%	4.13%
	>15 seconds	83.15%	<1 seconds	1.08%	15.77%
	>10 seconds	89.13%	<2 seconds	1.58%	9.29%
	>10 seconds	89.13%	<1 seconds	1.08%	9.78%

interaction between the number of queries and the minimal interval between them, we created scatter plots showing all the users on these two axes (Fig. 4). Both of the axes are log scale. For each log we used two definitions of minimal time interval: either between same queries or between different queries. An example of the results for the MSN06 log shows that there is no observable structure, and thus no natural threshold to use to differentiate humans from robots. However, we do see that for users with very high number of queries the intervals tend to be extremely low, and for users with high intervals there tend to be only few queries. Similar results hold for the other logs.

To better understand the interaction between the two criteria we use, Fig. 5 shows the relative sizes of the groups of users identified by each one and their intersection (again using the MSN06 log as an example). For humans, this indicates that practically all users whose minimal interval is longer than 25 seconds also submit fewer than 10 queries, whereas about 2/5 of the users who submit few queries have shorter intervals. Nevertheless, the intersection contains the majority of candidate users. For robot candidates, on the other hand, the intersection is small: the vast majority of users who submit more than 10 queries have non-zero intervals, and a slight majority of users with zero intervals only submit few queries.

4. CONCLUSIONS

Web search activity is actually multi-class, with a mixture of activity by humans and various types of robots. It is crucial to be able to differentiate between the different classes in order to characterize each one and study its effect on the system and the service it receives.

We have taken first steps in a systematic classification of users as humans or robots. Our results indicate that it may be impossible to find a simple threshold that can be used to perform a reliable classification. Therefore we advocate the use of several metrics in combination. As a start, we investigated thresholds for the number of queries submitted and the minimal interval between different queries. We also suggest using two thresholds to improve the confidence in the results, by leaving an unclassified group between the humans group and the robots group.

Future work includes a more systematic study of the different threshold values and their effect, considering using different thresholds for queries and for clicks, and integration of the additional user behavior attributes listed in the introduction.

Irrespective of the above, we note that a big problem in identifying robots from web search logs is the lack of "ground truth" — a reliable way to compare and test classification results. This may be improved by cross-referencing search logs and web server logs, and using various known techniques to identify web robots in the server logs. Thus it is desirable that in the future web server logs will be made available together with web search logs.

Acknowledgment

This research was supported by a grant from the Israel Internet Association.

5. REFERENCES

[1] N. Buzikashvili, "*Sliding window technique for the web log analysis*". In 16th *Intl. World Wide Web Conf.*, pp. 1213–1214, May 2007.

[2] N. N. Buzikashvili and B. J. Jansen, "*Limits of the web log analysis artifacts*". In *Workshop on Logging Traces of Web Activity: The Mechanics of Data Collection*, May 2006.

[3] O. Etzioni, "*Moving up the information food chain:*

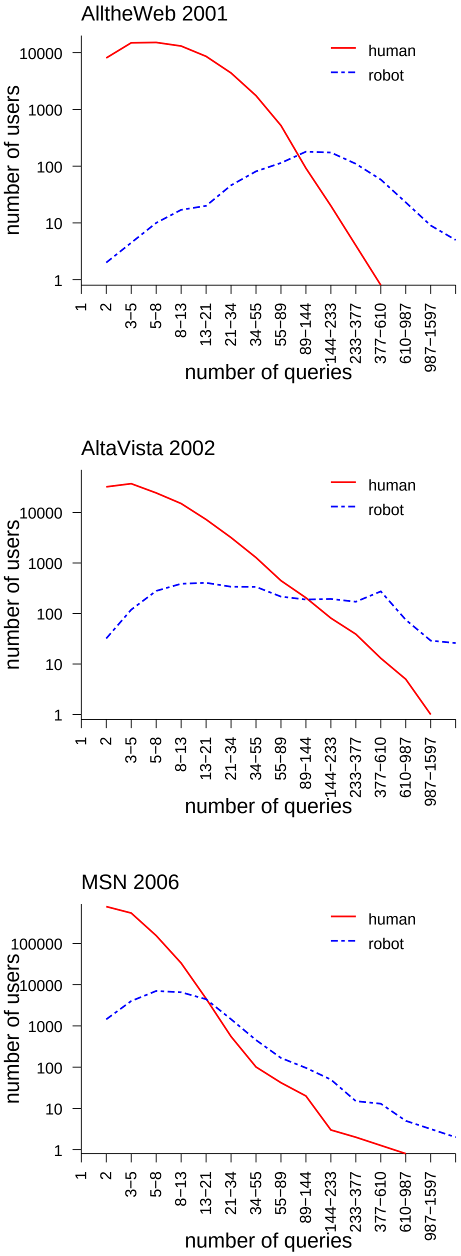

Figure 3: *Distributions of number of queries submitted by users identified as humans or robots according to their minimal inter-query interval.*

deploying softbots on the world wide web". *AI Magazine* **18(2)**, pp. 11–18, Summer 1997.

[4] N. Geens, J. Huysmans, and J. Vanthienen, "*Evaluation of web robot discovery techniques: a benchmarking study*". In 6th *Industrial Conf. Data Mining*, pp. 121–130, Jul 2006. (LNCS vol. 4065).

[5] B. J. Jansen, T. Mullen, A. Spink, and J. Pedersen, "*Automated gathering of web information: an in-depth examination of agents interacting with search engines*". *ACM Trans. Internet Technology* **6(4)**, pp. 442–464,

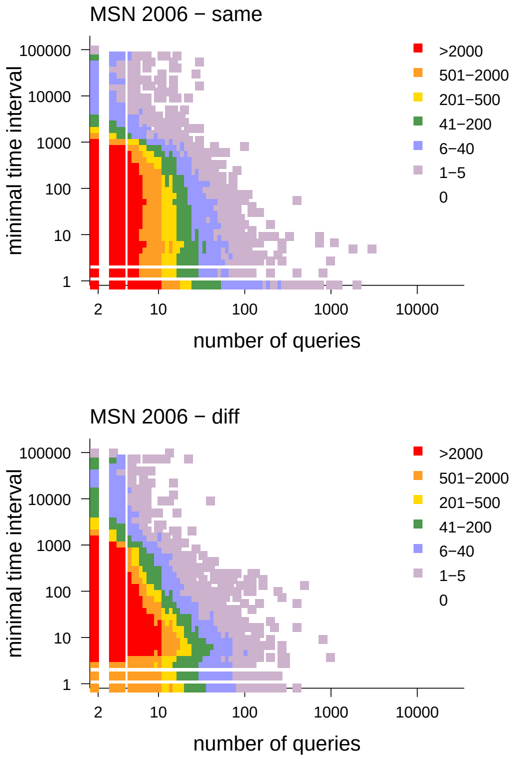

Figure 4: *Scatter plots showing the distribution of users in number-of-queries and minimal-interval axes.*

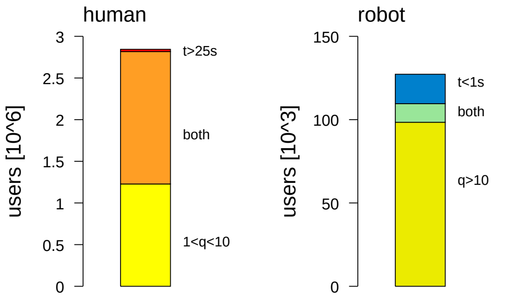

Figure 5: *Intersections of the two criteria used, for the MSN06 log.*

Nov 2006.

[6] B. J. Jansen and A. Spink, "*How are we searching the world wide web? a comparison of nine search engine transaction logs*". *Inf. Process. & Management* **42(1)**, pp. 248–263, Jan 2006.

[7] A. Spink and B. J. Jansen, *Web Search: Public Searching of the Web*. Kluwer Academic Publishers, 2004.

[8] A. Stassopoulou and M. D. Dikaiakos, "*Web robot detection: a probabilistic reasoning approach*". *Computer Networks*, 2009. (to appear).