Distinguishing Humans from Bots in Web Search Logs

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

ABSTRACT

Cleaning workload data and separating it into classes is a necessary pre-requisite for workload characterization. In particular, the workload on web search engines is derived from the activities of both human users and automated bots. It is important to distinguish between these two classes in order to reliably characterize human web search behavior, and to study the effects of bot activity. However, available workload data is not accompanied by labels that can be used as a basis for learning and generalization. To cope with the lack of labeled data, we suggest using two mechanisms. The first is to employ two thresholds for each criterion, enabling the identification of users who are most probably human or most probably bots according to need, and avoiding ambivalent cases. The second is the notion of "strong" criteria, which identify levels of activity which are highly unlikely or even impossible for humans to achieve. We then use an iterative process of refining the thresholds to combine the results of multiple metrics in a mutually consistent manner. Results using the AOL log identify over 92% of the users as human, and only a small fraction (0.6%) are probable bots. The humans tend to display relatively consistent behavior, whereas bots may exhibit markedly different behaviors. In particular, it is not uncommon for a bot to be very different from typical human behavior according to one criterion, while being indistinguishable from a human according to another.

1. INTRODUCTION

World-wide web search logs are maintained by all large scale web search engines. These logs are an important tool for studying and understanding web search behavior, for uncovering the effect of web search on the activity seen at web sites, and for the design and optimization of search engines [26, 1, 14]. However, upon inspection of these logs, one finds that they contain various anomalies that are unlikely to be representative of human search behavior. Such anomalous records may have a large effect on the perceived statistics of user behavior, so they need to be identified and handled prior to the analysis [6, 15].

The anomalous records may have diverse origins. One is bots, which are software agents that issue requests to a search engine in order to collect data, reverse engineer its index, or affect its rankings. Other non-human behavior may include using network address translation (NAT), where the source address actually represents a whole network with multiple independent users, or meta-search engines attempting to achieve improved results by pooling the results of multiple basic search engines. For brevity we will refer to all of these non-human behaviors as "bots".

Bots are not unique to web search activity, and the need to distinguish real humans from bots is widespread. For example, e-commerce sites want to rank product popularity based on real human actions, and avoid manipulations by software agents that mimic legitimate clients. Similar considerations apply to preventing bogus comments on blogs and in online chat rooms, and the ability to open accounts on email services, because bots may distribute spam and malware and pose a serious threat to other users [13]. Another example is massively multiplayer online games, where bots may be used to automate actions in a game without actually playing, thereby endangering the game providers' business plan [20]. The common approach to thwart such bots is to use CAPTCHAs, which are a challenge-response test that is easy for humans but hard for computers [30].

Our goal is not to protect web search engines, but rather to clean data regarding searches that have occurred. This may 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 [4]. Second, we want to facilitate the creation of a catalog describing different bot profiles that are encountered in real systems, so as to provide a basis for assessing their impact on search engines and the services they provide.

Our approach is to quantify the behavior of search users along multiple different axes, and set thresholds that distinguish humans from bots. Regrettably, no reliable labeled data may be available in order to learn about bot behavior and cross-validate the results. We therefore start with intuition (such as the often-used criterion that humans do not submit more than 100 queries in a day), and strengthen the classification by correlating the results obtained using the different criteria. In this, we give special status to "strong" criteria, which represent activity patterns that are believed to be well beyond those characteristic — or even possible — for humans (e.g. humans cannot submit 15 queries within a single minute). In addition, we use two thresholds for each criterion instead of one, thus leaving a small unclassified set about which we cannot be sure.

The next two sections describe the data we have at our disposal, and previous work on identifying bots. Our ideas are developed in Section 4, which presents criteria for distinguishing human and bot behavior, and Section 5, which presents the iterative analysis methodology. Section 6 then presents the results of applying this to real web search logs, in articular the one made public by AOL. Section 7 concludes the paper.

2. DATA SOURCES

While all search engine companies most probably maintain voluminous activity logs, few logs have been made publicly available for research. The most recent and largest ones are the infamous AOL log [22] and the Microsoft log [33], both from 2006. The provided data includes:

- Timestamp with seconds resolution, including time of day and day of the week.
- Anonymized source ID, based on an IP address or a cookie deposited with the user's browser.
- The query string.
- Clickthrough URL and rank, if any.

We focus on the AOL log, which is by far the most extensive that is freely available for research, and moreover, has undergone only minimal modifications in the interest of privacy. This log covers a period of 3 months and contains 36 million entries representing the activity of over 650 thousand users. But it can't be used safely for user studies without a serious attempt to identify and filter out those users who are actually bots.

Note that the precise semantics of the source ID are sometimes murky, and it is not clear whether a source ID corresponds to a single user [14]. The problem is exacerbated when the source ID field is altered occasionally to improve anonymization. The motivation for this practice is the desire to reduce the cumulative amount of information that is associated with any single user [7]. Based on the number of queries associated with users there is strong evidence that this approach was used in the Microsoft log. As this has a strong effect on metrics of user behavior, we therefore decided not to use this log.

The format of the other fields in the AOL log is also somewhat problematic. New queries and requests for additional pages of results that *did not* lead to a click are listed using only three fields: the user ID, the query, and the timestamp. But queries and requests for additional pages that *did* lead to clicks are not listed separately. Rather they are repeated with each click record. As a result the timestamp that appears in a click record is not the timestamp of the click, but the timestamp of the query that returned the result that was clicked upon. This practice loses the data about when clicks occurred. But by focusing on all *unique* tuples of ⟨user, query, timestamp⟩, and ignoring the click data, we can correctly identify all instances of new queries and requests for additional pages of results.

We note in passing that web server logs contain significant data that can be used to help identify bots, e.g. do they download images and possibly even including explicit identification in the "user agent" field or by virtue of access to the robots.txt file [12, 27]. Regrettably, this data is not available to us.

3. RELATED WORK

Most of the work on characterizing and detecting web bots focuses on web spiders [8, 21] and bots that operate in the context of e-commerce sites [28, 12, 27]. Such work obviously does not use specific features related to web search, and was based on web server logs, not search logs.

Early work on characterizing web search by Jansen, Spink, and their co-workers simply used a threshold of 100 queries to distinguish humans from bots: a user who submitted less than 100 queries was assumed to be a human, and one who submitted more was assumed to be a bot [18, 16, 26]. This was justified on the grounds that typical web sessions (assumed to be dominated by humans) are much shorter. It was also implicitly based on the assumption that logs are short (all the logs they used typically spanned about one day). This approach was also adopted by others (e.g. [11]).

Another approach that has been suggested is to try and identify agents by the number of unique queries submitted in a sliding window. Results showed that the window size is not very important, so one hour was selected. The suggested threshold was then 5–7 unique queries: users who submitted less were considered human, and those who submitted more were flagged as bots [5].

More sophisticated approaches, e.g. based on machine learning, have also been suggested [27] (and it is likely that search companies actually perform additional such studies that are not published). In supervised learning, a training set of labeled data is given, and the task is to generalize this to the full dataset. Regrettably, we do not have any labeled data to start with. Unsupervised learning is more limited; for example, it may be used to provide a clustering of the data. This too is hard to apply in our case, because the distributions of the behavioral metrics we identify below tend to be long tailed, without any distinctive modes or correlations. It is also possible to combine the two approaches into semi-supervised learning. Here a small training set is used to initialize the process, and then unlabeled data is used to extend it. This has some similarity with our work, where "strong" characteristics can be used to identify bots.

In particular, our work is somewhat similar to co-training [2]. However, co-training cannot be used with our data, because the sub-populations are not well-separated, but rather merged with each other. In fact, bots may even intentionally mimic humans. Thus a correct classification according to one attribute (e.g. more than 200 queries in a day is a bot) does not necessarily translate into a correct classification via another attribute (the bot could have done this in 30 minutes, but not all other users who were active for just 30 minutes are also bots).

Another relevant approach is fuzzy classification. This is based on fuzzy characterization of attributes, meaning that the characteristics of different classes may overlap, and that characteristics provide fractional degrees of evidence (e.g. a 60-year old person is "0.7 old") [32]. While this allows for expressing uncertainty and qualitative considerations, it still requires labeled data for the derivation of rules and a quantitative evaluation [23]. Our current work is concerned more with basic steps towards identification of the relevant variables and ranges of values; possible use of fuzzy classification to extend this work is considered as future work.

Several related works have been published recently by Microsoft authors. Yu et al. present SBotMiner, which focuses on the identification of bot nets, where each bot mimics human behavior but their aggregate behavior betrays them [31]. A key observation is that the fact that a set of users exhibit correlated behavior identifies them as bots. Kang et al. propose a semi-supervised setting, where access to the system is used to present a small subset of users with CAPTCHAs so as to obtain definitive identifications [19]. They then use an EM algorithm to learn a classifier based on several characteristic measurements. The prior work that is closest to ours is the study by Buehrer et al. [3]. They also look at distributions of behavioral characteristics, and define thresholds for bot activity. Their evaluation is based on some labeled data of unspecified origin. Our main contributions relative to this work are that we develop a methodology for combining the (possibly conflicting) results of different characteristics, and for the iterative refinement of thresholds.

4. CLASSIFICATION CRITERIA

Various criteria may be used to distinguish humans from bots, similar to the 100-query threshold mentioned above. But before we list all our criteria, two methodological issues should be discussed. These are the use of multiple thresholds, and the notion of strong criteria.

The work of Jansen, Spink, and their co-workers used a single threshold to make a binary decision: users who submitted up to 100 queries were classified as humans, whereas those who submitted more than 100 queries were classified as bots. Such crisp thresholds have also been applied by others to this and other criteria [3, 11]. However, in reality we may expect to have some overlap between the two populations. Therefore such a sharp boundary may be inappropriate, and lead to too many false negatives and false positives. As an alternative, we suggest the use of *two* thresholds, leading to a *three-way classification*[1]. Those users that are below the bottom threshold will be classified as humans. Those that are above the top threshold will be classified as bots. But those that are between the two thresholds will remain unclassified. Our goals in setting the thresholds are to balance the increase in our confidence in the classifications with an effort to reduce the number of users who remain unclassified.

In some cases, the threshold for bot activity may be placed at such high values so as to be highly improbable or even surpass the physical capacity of humans. For example, humans cannot achieve a rate of submitting 15 different queries within one minute, but bots can. Using such thresholds essentially precludes the danger of false positives, where humans are accidentally classified as bots. We thus call such criterion-threshold combinations *strong* criteria. Note that this is one-sided: there are no corresponding strong criteria for humans.

The obvious problem is that bots can easily imitate human behaviors, and actually do so in order to avoid detection. In this case, it will be impossible to identify the bot using the limited data that is available in the search log. We therefore may expect false positives on the human side, where bots are erroneously classified as humans. However, at least for the limited objective of characterizing typical human behavior, bots that are proficient at human mimicry do not distort the results and thus do not pose a problem.

Maximal Queries per Day.

The most widely used criterion for identifying bots is volume of activity. The fundamental problem with this criterion is that different log files cover different time durations, and one may expect more queries in a longer log. We therefore choose a single day as a useful unit of time that can be used consistently. We assume that this was also implicitly meant by Jansen, Spink, and their co-workers [18, 16, 26], as most of their logs were for durations of about one day.

In the context of a multi-day log, "queries per day" may be interpreted as the average number. However, real activity patterns tend to be non-homogeneous. To identify bots it is therefore natural to single out the day with the highest level of activity. Note that queries here are both new queries and repetitions of previous queries, i.e. requests for additional pages of results.

Maximal Queries per Minute.

The number of queries in a day can indicate the average rate in which queries are submitted. But queries are not uniformly distributed, and users tend to exhibit spurts of activity in which they make multiple queries and request many pages of additional results. We can thus gain extra data by considering the maximum number of queries in a minute.

A rapid rate will be used to indicate that the user is a bot [15]. Note that the query rate may in principle be expected to also depend on the response time from the search engine. However, typical response times are sub-second, and thus the interval between queries is dominated by human actions (think time and typing) and possible network delays. Bots require minimal if any think times and may not wait for responses.

A variant of this metric is to count the number of characters that were sent in a minute, in order to gauge the rate at which users type. This may sound like a potentially strong criterion, because professional typists typically achieve no more than around 200 characters per minute. However, we have no way to know whether the user actually typed all the query, or perhaps used copy/paste, and therefore we do not use this approach.

Minimal Interval between Different Queries.

A related metric to the previous one is the minimal interval between successive queries [9]. We only consider different queries here, because repeated submissions of the same query only require a single click, so two clicks may lead to recording two queries within one second by accident. Humans are expected to require several of seconds to submit a new query. In contrast bots can send a few queries within a second, leading to 0 intervals (recall that the logs employ timestamps with second resolution). A potential problem with this criterion is that it may be susceptible to log errors. It is possible that sometimes two different queries will get the same timestamp, even though they were sent at different times. The danger can be reduced by requiring multiple repetitions of 0 intervals.

An alternative metric that may overcome such errors is to use the median time interval. The drawback is that it is

[1] This may be regarded as a simple special case of fuzzy classification, which in principle also allows certain ranges of values to remain unclassified.

less usable as a strong criterion. The reason for this is that although it is possible for the median time interval of some user to be 0, and then we can be quite sure that this user is a bot, there is nothing preventing a bot from sometimes waiting between two queries, and then the median will not be 0 anymore. The result of using 0 as a bot parameter will be losing a significant number of bots.

Average Number of Words in a Query.

For this criterion we count the number of search terms in each query and take the average. This criterion is based on the assumption that bot queries tend to be longer and more sophisticated, while humans tend to send short and simple queries. Indeed, the average number of terms in queries has been between 2 and 3 in practically all studies to date (e.g. [25, 18, 33]). Better fidelity may be achieved by splitting it into two different sub-criteria: regular queries and questions queries. The reason for this is that questions queries tend to be longer than regular queries. This criterion has the obvious drawback that bots may send short queries as well, so it is hard to distinguish between humans and bots using this criterion alone. Related criteria that have been proposed by Buehrer et al. consider the diversity of words used, abundance of spam words, and whether successive queries tend to be alphabetized [3].

Maximal Number of Repetitions.

This criterion counts the number of repetitions of the same query. repetitions may happen normally as users try to refind things [29], or, more commonly, they may be just requests for additional pages of results. However, users who submit the same query an extremely large number of times are assumed to be bots. For example, in the AltaVista log from 2002, 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”. The reason for this behavior might be an attempt to manipulate query word ranking, or downloading all of the links pertaining to a certain query by requesting page after page of results.

In principle, the number of repetitions should be normalized to log length just like the total number of queries discussed above. However, repetitions that continue across more than a day may be of special interest. We therefore use the total number of repetitions as our metric. This implies that thresholds should be adjusted for each log according to its length.

Repetitions with Precise Periodicity.

This criterion is similar to the previous one, with one difference: now we are looking for users that submitted the same query at precisely measured intervals. This strengthens the previous criterion, as it only matches the behavior of scheduled bots. The chance that a human will send exactly the same query over and over with the same time intervals between the queries is very low, whereas bots may use measured intervals to pace their activity and avoid overloading the server. Note however that it is likely that there are many bots that will not fall into this criterion.

Despite the intuition against repeated intervals by humans, we have observed cases where users who appear human nevertheless exhibit several identical intervals. This was probably the result of skimming page after page of results, each taking just a few of seconds; as the result pages all have the same size, it may happen that the several requests for an additional page come at identical intervals.

Maximal Continuous Session Length.

In this criterion we check the duration of continuous activity. This refers to the period of time that the user continues submitting queries. It is based on the assumption that while humans need to rest for a few hours in a day, bots don’t. For example, if a user is active continuously for more than a day, it is most probably a bot [15].

The problem with defining continuous sessions is that queries have intervals between them. The length of a continuous session therefore depends on the maximal interval that is not considered a break [6]. In particular, the maximal break should not be too high. For example, if we say that a bot is a user that is active for 20 hours with intervals of no more than 5 hours, then it is possible that some human user will send one query every 5 hours, and will be considered as a bot by mistake. we typically use 10-minute intervals as a measure of continuity, but also checked 1-hour intervals.

Note that we consider only time intervals in our definition of sessions. This is because we want to use long stretches of activity to identify bots. Other researchers have used a combination of intervals and similarity between queries to define sessions [11], with the goal of characterizing the activities involved in trying to satisfy a single information need. Our sessions may include work on several information needs one after the other.

Correlation with Time of Day.

Humans are expected to be much more active during daytime. This is a behavioral criterion. In fact, both humans and bots may be active both during the day and at night. This criterion also interacts with the previous one. Long continuous sessions are more likely during the day, and long breaks are more likely during the night. For example, if a user had a break of 10 hours between two bursts of activity then it will be considered as a human. But if the bursts of activity occurred at 3 AM and 1 PM then it looks less like human behavior.

Clicking on Search Results.

Humans typically scan only a limited number of pages of results, and may click on a few of them. Thus behavior that deviates from this pattern may indicate the presence of a bot. Specifically, this includes the following behaviors [3]:

- Not clicking on any search results. This is not expected to provide a strong classification, because humans may refrain from clicking on results too.
- Clicking on all the results one after the other. This may characterize a bot that is using the search engine to collect information.
- Viewing very many pages of results. As mentioned above, this is not typical of humans.

As clickthrough information does not exist in all logs this criterion is not always available. However, we can use it for evaluation and characterization of the resulting classification. In particular, the AOL log does contain clickthrough data.

5. METHODOLOGY

Given the classification criteria, the question is how to set thresholds that will lead to the best discrimination between humans and bots. The basic problem is verifying the quality of the results, since we don't have any reliable source of information about which users are indeed humans and bots. We therefore need to develop heuristic methods to assess and improve our belief in the results.

The approach we use to achieve this is iterative. We initially use one criterion to create a classification of users into supposed humans and bots. Given this classification, we can compare the distributions of other criteria for the two groups. The original classification is graded according to how well these induced distributions are separated. If these distributions tend to be well-separated, the criteria support each other and our belief in the classification is increased. If they are not, meaning that a classification according to one criterion does not lead to a good separation for other criteria, then maybe the original classification was of inferior quality. This means that the thresholds need to be adjusted, or maybe the criterion is not useful for classification.

Once we find thresholds that can be used for classification using the different criteria, we still need to decide on the final classification of each user. If all the classifications using the different criteria agree, this is easy. If they do not, we give priority to strong criteria as described above.

5.1 Tools

To perform the analysis, we developed a log analyzer utility. This can parse the different logs, and supports setting thresholds on the criteria listed in Section 4 in order to partition the users into humans, bots, and unknown. It also supports further analysis of these groups, with both tabular and graphical outputs. In particular, given a classification based on one criterion, one can get the distributions of other criteria for the different groups of users. Full details are given in [10].

In addition, we created a script for automated searching for good thresholds. This systematically runs over a list of reasonable threshold values, and creates classifications based on them. Each classification is then graded based on how well it separates other criteria, as described below. This enables the thresholds that lead to the highest grades to be identified.

5.2 Progression of the Analysis

We started the analysis with the criterion of number of queries per day and a threshold of 100, for both bots and human, which means that we will classify users as human if they sent less than 100 queries in a day; otherwise we will classify them as bots. Thus we use the classification assumption of Jansen and Spink as our starting point.

Using this initial threshold led to a classification of 99.72% of the users as humans, and only 0.28% were classified as bots. We then used this classification to further investigate the distribution of queries per day and seek potentially better thresholds (Fig. 1). In this and subsequent figures, we plot the two distributions as histograms using a logarithmic scale. This is done because practically all the criteria have positive skewed distributions. The bin boundaries are set according to the Fibonacci series, so bins are 1, 2, 3, 4–5, 6–8, 9–13, 14–21, 22–34, and so on. Note that the histograms for humans and bots are independent: the sum of the bars for each class separately sum to 1.0, even though in reality there are many more humans than bots.

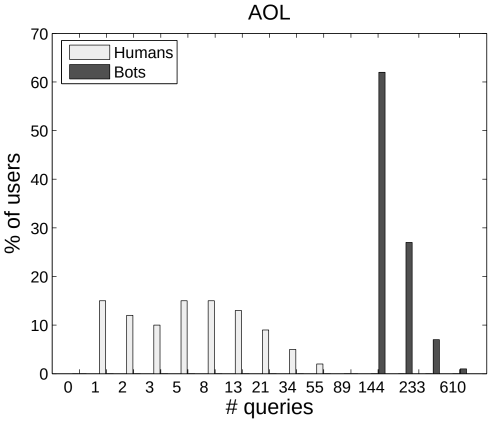

Figure 1: *Distributions of number of queries per day actually submitted by users identified based on a threshold of 100.*

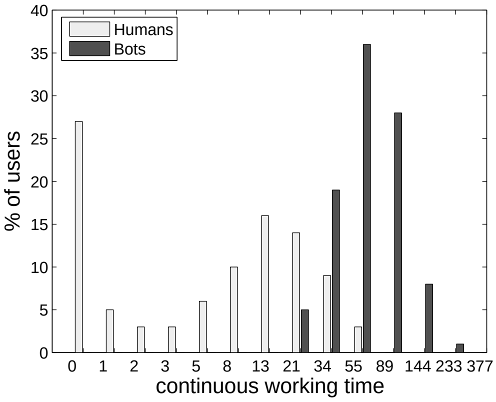

Figure 2: *Distributions of continuous working time, with less than 10 minutes breaks, of users identified as humans or bots according to their number of queries in a day.*

As shown in the figure the result was that the vast majority of users classified as human actually submitted considerably fewer queries than 100. This motivated the use of two thresholds, where humans are users who submit up to 50 queries a day, and bots are those that submit at least 100. This reduced the fraction of users classified as human to 98.57%, leaving only 1.15% unclassified (the fraction of bots naturally remained 0.28%).

The next step is to check the human and bots behavior using another criterion. We chose the criterion of continuous working time, with breaks of no more than 10 minutes. The results are shown in Fig. 2. The users previously classified as humans tend to work for up to about half an hour continuously, with few (around 3%) going up to one hour. Those classified as bots tend to work continuously for a longer period of time, from 21 minutes up to 4 hours (in this case around 4% work for less than 20 minutes, but more than

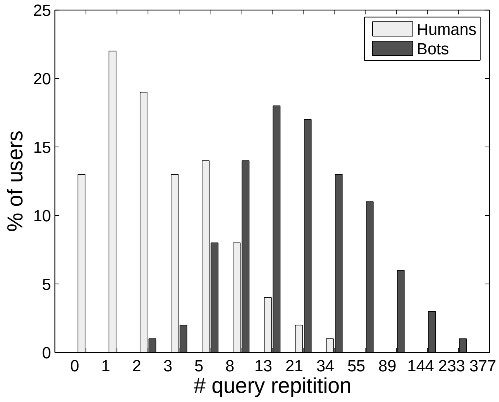

Figure 3: *Distributions of number of repeated queries by users identified as humans or bots according to their contin-ues working time.*

9). These results makes sense because a typical behavior for humans is to use the search engine in order to search for a specific item, a process that most of the times takes less than half an hour, while bots may use the search engine for other purposes that may take a longer time, like gathering information.

Based on these results, we can set tentative thresholds on continuous work at 20 minutes for humans and 40 for bots. This avoids the large overlap that exists in the bin of 21–34 minutes. Using these thresholds led to the classification of 84.06% of the users as humans, 3.26% as bots, and 12.67% remained unclassified. It also led to a good separation of the distributions of queries per day.

To assess the quality of the classifications based on the different criteria it is useful to check their intersections. The results for humans are very good: nearly all the users who were tagged as humans by continuous work were also tagged as humans by queries per day. But the results for bots are not so clear cut: while nearly $\frac{2}{3}$ of the bots by queries per day were also classified as bots by continuous work, this second analysis identified a much larger group. Thus nearly $\frac{5}{6}$ of them were new and not tagged as bots in the first analysis. In particular, some of these were actually classified as humans in the first analysis. We consider dealing with such contradictions below.

Given the classification based on continuous work, we can continue with the practice of looking at distributions this induces on other criteria. For example, we can look at the distribution of the number of repetitions of the same query by humans vs. bots (Fig. 3). Again we find that the humans and bots do behave differently. 93% of the users that we tagged as human re-send their queries up to 8 times, while 98% of the users that we tagged as bots re-send their queries 4 times or more. While the overlap is significant in this case (22% of humans and 24% of bots are in the range 4–8), the bulks of the distributions are nevertheless separated.

While this description gives the gist of how thresholds can be adjusted iteratively in order to derive a good separation, in order to mechanize the process we need to quantify the quality of the classification. We do this by measuring the relative size of the intersection between the classes, as described next.

5.3 Grading a Classification

We grade a classification using one criterion by its effect on other criteria. If the distributions of values attributed to humans and bots are well-separated, the original classification is considered good. If there is considerable overlap, the classification is not good. By quantifying the overlap we can grade the original classification, and in particular conduct a search for thresholds that lead to better classifications.

We quantify the separation based on histograms like those shown in Figs. 1, 2, and 3. The possible values (as represented by the histogram bins) are divided into three: values that are unique to humans, values that are unique to bots, and values that are shared by both. A good separation means that most users in each class are characterized by unique values, and only few can be confused with the other class because they are characterized by shared values. Thus the separation is measured by the fraction of users in the class characterized by unique values. Given that we have two classes, the final metric is the average of the fraction of humans with unique values and the fraction of bots with unique values. Symbolically, denote the bins by $b_1, \ldots b_n$, the fraction of humans in bin i by $H(b_i)$, and the fraction of bots by $B(b_i)$. Further denote by $I\neg B$ the set of indices i such that $B(b_i) = 0$, and by $I\neg H$ the set where $H(b_i) = 0$ The grade is then

$$G = \frac{1}{2}\left(\frac{\sum_{i\in I\neg B} H(b_i)}{\sum_i H(b_i)} + \frac{\sum_{i\in I\neg H} B(b_i)}{\sum_i B(b_i)}\right).$$

The grading as described so far is very strict about the dangers of confusion. If one bot has a value that falls in the same bin as 100,000 humans, it is enough to contaminate the bin and suggest that all those humans are subject to confusion. This is unreasonable. We therefore add two heuristics to alleviate such rigidity. The first is that if the fraction of users from one class that fall in a bin is less than 1%, they are not considered to contaminate the bin. The second is that if the fraction of users from one class is 10 times or more higher than the fraction of users from the other class, we consider the bin unique from the point of view of the class that has the larger fraction.

5.4 Combining Classifications

After iteratively searching for good thresholds on the different criteria and grading them based on the degree of separation they induce on other criteria, we are left with a set of classifications. Due to the diversity in bot behaviors, these classifications may be contradictory. The question is then how to combine then into one coherent classification.

Based on the fact that we employ a three-way classification, we classify a user as unknown if we have contradictory inputs about him. This is illustrated in Fig. 4. Thus if some user was marked as bot by some criterion, it will be tagged as bot only if he wasn't classified as human by any other criterion. The same works for humans: a human by one criterion will be tagged as human in the final classification if and only if it was not classified as a bot by any other criterion.

The only deviation from this approach is when strong criteria are involved. Strong criteria identify bots based on attributes that are believed to be highly improbably of humans. Therefore all users that were classified as bots by some strong criterion are tagged as bots for the final classi-

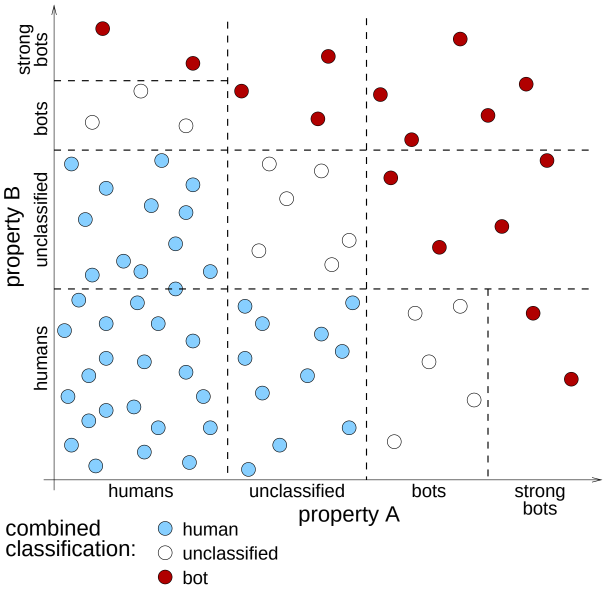

Figure 4: *Illustrative example of classification based on mixing two properties.*

fication, regardless of their classification by other criteria — even if they were classified as humans.

6. RESULTS

The description given above of the progression of the analysis included only a brief manual part of the whole process. In addition we conducted an automated search for thresholds that lead to good separations. In this section we summarize the final results of this process.

6.1 Useful Criteria

Initially we had a set of promising criteria for the classification of humans and bots, as listed in Section 4. In retrospect, some of these proved more useful than others. The main problem is that some criteria proved to be intrinsically not well separated. Regardless of how we obtained a putative partitioning of the users into humans and bots, there was always considerable overlap between the two groups. Specifically, this was the case for the following:

- The distributions of minimal time intervals always have a large overlap for humans and bots, even though bots are unique in the very lowest bins.
- The distributions of length of query were practically the same for both groups of users.
- The correlation with time of day was about the same for both groups as well, with the majority of users active during day time.

In grading a classification we therefore ignore these criteria. In addition, there were some criteria that were useful for initial classification, but still had enough overlap so that they degraded the grading unnecessarily. For example, this was the case with the maximal number of repetitions. The opposite situation also occurred, for example with the average number of queries in a day. This criterion was not good for classification, and we couldn't find good thresholds for it. But it proved useful in the grading. The final grades were therefore calculated as the average grade for the following criteria: maximum number of queries in a day and in a minute, average queries in a day, maximum periodic repetitions, and maximum continuous working time with breaks of up to 10 minutes.

The results also identified strong criteria that were not useful, simply because they were so strong that they didn't identify any users. These included high thresholds on queries per day and continuous work. In some cases we therefore used thresholds that are very high, but not necessarily beyond the physical capabilities of energetic humans.

6.2 Classification Thresholds and Results

The main results of the analysis are summarized in Table 1. The main criteria used are listed across the top of the table. For each criterion we have the following data:

- The best thresholds that were found to identify humans and bots, using the iterative process and mechanized search for good threshold values. For example, for the queries per day criterion, users with <25 were classified as human, and those with >50 as bots.
- The number of users in each class, and their percentage out of the total population. The percentages do not sum to 100% because of the interim group that was left unclassified.
- The grade of the classification, calculated as described above based on the separation of the two resulting user groups.
- A rough characterization of the resulting separation, in the form of estimated thresholds for *other* criteria. These thresholds indicate ranges of values that are dominated by the humans and by the bots, respectively, as classified by the criterion at the top of the column.

By looking at the table we can see that a reasonably consistent behavioral pattern emerges for the human users, and to a somewhat lesser extent also for the bots. This is reflected in the consistency between the thresholds that were used for classification and the thresholds that were found to characterize the achieved separation. For example, in the "repetitions" row we see the thresholds that were found to characterize the separation of users when looking at the distribution of number of repetitions. For humans, the obtained thresholds were <14, <13, <10, <14 and <5. Except for the last one, these agree with (and are slightly less strict than) the threshold of <10 that was used for classification according to this criterion. The pattern for the other criteria is similar. Somewhat surprisingly, even the classification based on the minimal interval criterion produced reasonably consistent results.

The results shown in the table are for each criterion independently. These classifications were then combined as described above in Section 5.4 (except for the minimal interval criterion, which was not used due to its low grade). As explained above, in this combined classification users will be marked as humans only if they were classified as human by at least one criterion, and were not marked as bot by any criterion; The same applies to bots the other way round. After performing the combined analysis, we got that 92.3% of the users were classified as humans, 0.3% as bots, and

criterion	queries/day		queries/min		min. intrvl [s]		repetitions		periodic rep.		cont. work [m]	
classification	human	bot	human	bot	human	bot	human	bot	human	bot	human	bot
thresholds	<25	>50	<5	>10	>9	<1	<10	>30	<1	>3	<20	>35
number	613639	9712	619894	2633	387851	18413	574470	21090	652868	4560	551473	33098
of users	93.3%	1.48%	94.3%	0.40%	59.0%	2.80%	87.4%	3.21%	99.3%	0.69%	83.9%	5.03%
grade	85		76		47		62		63		77	
induced thresholds:												
queries/day	–		<30	>45	<21	>30	<21	>21	<30	>40	<21	>21
queries/min	<4	>4	–		<3	>5	<3	>5	<5	>5	<3	>3
repetitions	<14	>20	<13	>20	<10	>20	–		<14	>25	<5	>8
periodic rep.	<1	>1	<1	>1	<1	>1	<1	>1	–		–	

Table 1: *initial thresholds, classifications, and resulting suggested thresholds for other criteria.*

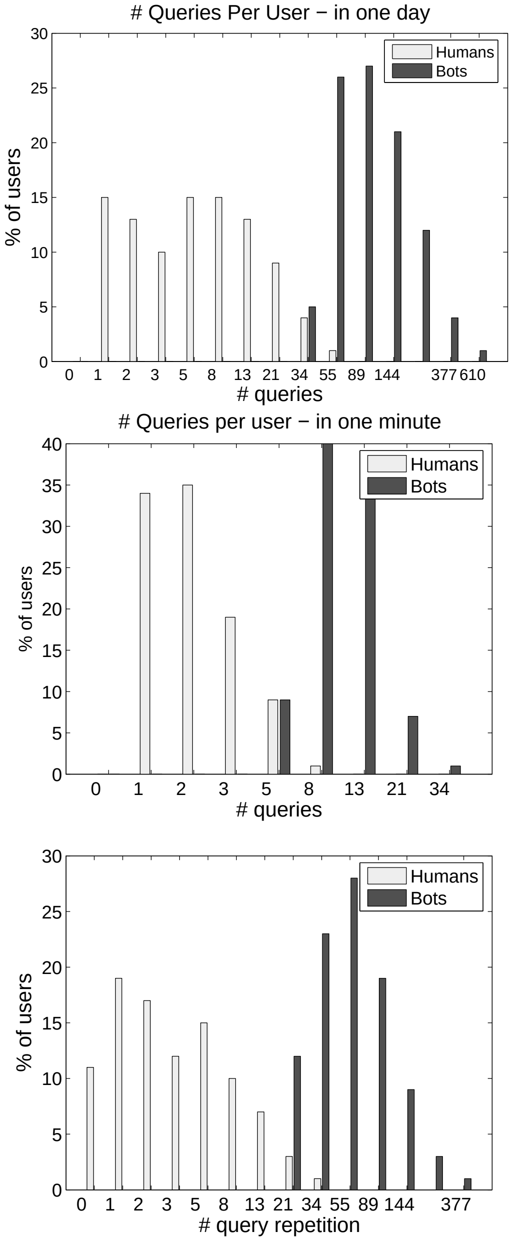

Figure 5: *Distributions of criterion values for humans and bots, as identified based on combining the results shown in Table 1 for the individual criteria.*

criterion	thresh.	bots	unclass.	grade
none		0.27%	7.30%	89
queries/day	200	0.30%	7.27%	89
queries/min.	15	0.34%	7.23%	85
min. inter.	0×3	0.40%	7.21%	84
repetitions	150	0.39%	7.18%	78
periodic rep.	7	0.28%	7.29%	89
all five		0.58%	7.03%	77

Table 2: *The effect of strong criteria.*

7.4% of the users remained unclassified. The grade of the combined classification was 88, which is considerably higher than that of most of the individual classifications, and also pretty good on an absolute scale. The actual distributions for several criteria are shown in Fig. 5.

6.3 The Effect of Strong Criteria

Note that in the above combined results all the criteria are equal, and we do not consider any of them to be strong. Enforcing the bots classification according to strong criteria has two effects: it more than doubles the size of the bots group from 0.27% to 0.58% of all users, but it also degrades the grade of the separation between humans and bots from 89 to 77 (Table 2). This means that users that we classified as bots using a strong criterion may behave like humans in another criterion. We couldn't find thresholds that classify these users as bots in all of the criteria simultaneously.

Table 2 also shows the effect of individual strong criteria. Because of the strong criterion definition, the decision of which criterion is strong influenced only the sizes of the bots group and the unclassified users group; the humans group remains unchanged. Initially, without considering any criterion as strong, only a consensus group of users are identified as bots. These are users who are far from human behavior in all criteria. When defining some strong criterion, more users are classified as bots, at the price of reduced separation. Each strong criterion by itself identifies some more bots, but this additional set is different for the different strong criteria. Note that we can use the minimal interval as a strong criterion even if we do not use it for the consensus group; the threshold of 0×3 means that we require 3 repetitions of a 0 interval.

It should be noted that the number of additional users identified as bots depends on the selected strong threshold. In some cases, the threshold can be set to values that are so high that they are beyond the physical capabilities of humans. In other cases we only see some effect if the thresholds

log	human	unclass.	bots	grade
AOL06	92.39%	7.03%	0.58%	77
AtW01	95.95%	3.79%	0.26%	92
AV02	96.36%	3.07%	0.56%	75

Table 3: *Comparison of results for different search logs.*

are lower than this lofty goal. However, the thresholds are always substantially higher than the regular bot threshold, such that the probability that a human reached this level of activity is very low.

To summarize, there are two subgroups of users that end up classified as bots: a group who are in the consensus (0.27% of all users), and an additional group whose classification as bots hinges on a strong criterion (0.31% of all users for the selected criteria and thresholds). This implies that a tradeoff is involved: we are classifying a larger fraction of the users at the price of using strong criteria to settle conflicting classifications. Assuming our definitions of strong criteria indeed capture behavior that is unpractical for humans, we have grounds to trust this classification and claim that we are actually not trading off any accuracy.

Given the final group of users classified as bots, it is interesting to note a few characteristics of their behavior. One finding is that the users identified as bots were also found to make clicks on results. Moreover, some just made few clicks, while others made many. Thus the conjecture that bots may be identified based on the fact that they perform many queries but no clicks is called into question. This matches the results of Kang et al. [19], who also found that the distributions of clicks by humans and bots are indistinguishable. Similar results pertain to the number of queries: for example, a full 12.4% of the users identified as bots performed less than 100 queries.

6.4 Results for Other Search Logs

After analyzing the AOL log, we turned to analyzing two other logs: those from AlltheWeb in 2001 and AltaVista in 2002. These were chosen because they are relatively large and recent, more so than the Excite logs from the 1990s.

In analyzing the new logs, we expect essentially the same threshold values to work, as they are supposed to reflect general human search behavior. And indeed, the resulting classifications led to good grades for the separation, so it seems that the same thresholds can indeed be used.

In fact, the results for the two additional logs are somewhat more extreme than in the AOL log (Table 3). Specifically, in AtW fewer users are identified as bots: only 0.26% of the total number of users, instead of 0.58%. On the other hand, more users are classified as human: 95.95% and 96.36%, respectively. Finally, the separation between the groups is better in AtW than in AOL, with a grade of 92.

7. CONCLUSIONS

Distinguishing between human and bot users from web search logs is difficult, and often no ground truth is available for training or evaluation of classification results. In particular, this is the case with the AOL log, raising questions regarding the possible use of this valuable resource. Previous work is split into two approaches: those who do have or can derive some labeled data can use machine learning to generalize this, while those who do not typically used a simple threshold of submitting 100 queries a day. We have extended this and cope with the lack of labeled data in the following ways:

- We considered multiple additional criteria for classification, including the instantaneous query submittal rate, the number of repetitions, and the duration of continuous work.
- We suggest the use of two thresholds instead of one, leading to a three-way classification: human, bot, or unknown.
- We performed an iterative process of refining the thresholds, using the results of a classification based on one criterion to learn about appropriate thresholds for other criteria. This is used to cluster humans and bots separately, with the thresholds optimized to improve the separation between the clusters.
- We identify some of the criteria as strong criteria, where an appropriate threshold identifies bots by specifying behavior improbable for humans.
- We identify some of the proposed criteria as not useful, based on observations that they do not lead to a good separation that is consistent with other criteria.
- We developed a methodology for combining the results of the different criteria, by accepting all classifications that are either strong or do not conflict with other classifications.

Our results indicate that the vast majority of users, estimated at between 92.4% and 96.3% (depending on the log), can be safely classified as humans. These human users exhibit relatively consistent and moderate behavior, submitting up to about 30 queries a day, at a moderate rate, with few repetitions, and in sessions that are up to 30 minutes long.

The bots, on the other hand, are much more diverse, and may exhibit extreme and unique behavior. Overall we classify from 0.26% to 0.58% of the users as bots. Of these, about half are in the consensus. The rest received conflicting classifications, with some criteria classifying them as bots while others classify them as humans. The final verdict about their classification as bots is based on the use of strong criteria. Thus it is not uncommon for bots to exhibit extreme behavior according to one criterion but moderate behavior according to another.

In each log, a certain fraction of the users remain unclassified, but we claim this is the most appropriate decision about them given the lack of precise information. This unclassified group is only 3–7% of the total users, leaving the vast majority of the data classified and suitable for use. However, it should be noted that this unclassified set most probably contains human users that exhibit extreme levels of activity, which are not well represented by those users that were positively identified as humans.

The main threat to the validity of our results is the lack of ground truth. This is inherent in this line of research, as our main motivation was to see what can be done when labeled data is not available. However, a future line of research is to verify our approach using a dataset that does indeed include labeled data, as has apparently been available in some cases (but not made public) [3, 19]. This will also extend the verification that our approach generalizes to other datasets besides the AOL log.

Given our classification of users into humans and bots, several avenues of additional research present themselves. The first is to repeat previous work on web search behavior (such as [25, 15, 17]), and check its sensitivity to the classification of users. In order to facilitate such research, we have made the list of users we have identified as bots available to others at http://www.cs.huji.ac.il/~feit/papers/RoboAOL/. Another is creating generative models of web search, which can be used to explain observed behavior and to drive evaluations (similar to and extending the work of Shriver et al. [24]). Finally, our methodology can be extended by considering additional behavioral criteria, and more sophisticated approaches such as fuzzy classifiers based on fuzzy quantifiers of the different attributes.

Acknowledgments

This research was supported by a grant from the Israel Internet Association. Many thanks to AOL for making the web search logs available.

8. REFERENCES

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno, "*Learning user interaction models for predicting web search result preferences*". In 29th *SIGIR Conf. Information Retrieval*, pp. 3–10, Aug 2006.

[2] A. Blum and T. Mitchell, "*Combining labeled and unlabeled data with co-training*". In 11th *Proc. Workshop Computational Learning Theory*, pp. 92–100, Jul 1998.

[3] G. Buehrer, J. W. Stokes, and K. Chellapilla, "*A large-scale study of automated web search traffic*". In 4th *Intl. Workshop Adversarial Information Retrieval on the Web*, pp. 1–8, Apr 2008.

[4] K. Burnett and E. G. McKinley, "*Modeling information searching*". *Interacting with Computers* **10(3)**, pp. 285–302, Jun 1998.

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

[6] 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.

[7] A. Cooper, "*A survey of query log privacy-enhancing techniques from a policy perspective*". *ACM Trans. Web* **2(4)**, art. 19, Oct 2008.

[8] M. D. Dikaiakos, A. Stassopoulou, and L. Papageorgiou, "*An investigation of web crawler behavior: Characterization and metrics*". *Computer Communications* **28(8)**, pp. 880–897, May 2005.

[9] O. Duskin and D. G. Feitelson, "*Distinguishing humans from robots in web search logs: Preliminary results using query rates and intervals*". In *Workshop on Web Search Click Data*, pp. 15–19, Feb 2009.

[10] O. M. Duskin, *Distinguishing Humans from Robots in Web Search Logs*. Master's thesis, The Hebrew University, Dec 2009.

[11] D. Gayo-Avello, "*A survey on session detection methods in query logs and a proposal for future evaluation*". *Information Sciences* **179(12)**, pp. 1822–1843, May 2009.

[12] 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. Lect. Notes Comput. Sci. vol. 4065.

[13] S. Gianvecchio, M. Xie, Z. Wu, and H. Wang, "*Measurement and classification of humans and bots in Internet chat*". In 17th *USENIX Security Symp.*, pp. 155–169, Sep 2008.

[14] C. Grimes, D. Tang, and D. M. Russell, "*Query logs alone are not enough*". In *WWW'07 Workshop on Query Log Analysis*, May 2007.

[15] 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, Nov 2006.

[16] B. J. Jansen and A. Spink, "*An analysis of web searching by European AlltheWeb.com users*". *Inf. Process. & Management* **41(2)**, pp. 361–381, Mar 2005.

[17] 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.

[18] B. J. Jansen, A. Spink, and J. Pedersen, "*A temporal comparison of AltaVista web searching*". *J. Am. Soc. Inf. Sci. & Tech.* **56(6)**, pp. 559–570, 2005.

[19] H. Kang, K. Wang, D. Soukal, F. Behr, and Z. Zheng, "*Large-scale bot detection for search engines*". In 19th *Intl. World Wide Web Conf.*, pp. 501–510, Apr 2010.

[20] D. Kushner, "*Playing dirty*". *IEEE Spectrum* **44(12INT)**, pp. 30–35, Dec 2007.

[21] A. Oke and R. Bunt, "*Hierarchical workload characterization for a busy web server*". In *TOOLS*, T. Field et al. (eds.), pp. 309–328, Springer-Verlag, Apr 2002. Lect. Notes Comput. Sci. vol. 2324.

[22] G. Pass, A. Chowdhury, and C. Torgeson, "*A picture of search*". In 1st *Intl. Conf. Scalable Information Syst.*, Jun 2006.

[23] J. A. Roubos, M. Setnes, and J. Abonyi, "*Learning fuzzy classification rules from labeled data*". *Information Sciences* **150(1-2)**, pp. 77–93, Mar 2003.

[24] E. Shriver and M. Hansen, *Search Session Extraction: A User Model of Searching*. Tech. rep., Bell Labs, Jan 2002.

[25] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz, "*Analysis of a very large web search engine query log*". *SIGIR Forum* **33(1)**, pp. 6–12, Fall 1999.

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

[27] A. Stassopoulou and M. D. Dikaiakos, "*Web robot detection: A probabilistic reasoning approach*". *Computer Networks* **53(3)**, pp. 265–278, Feb 2009.

[28] P.-N. Tan and V. Kumar, "*Discovery of web robot sessions based on their navigational patterns*". *Data Mining & Knowledge Discovery* **6(1)**, pp. 9–35, Jan 2002.

[29] J. Teevan, E. Adar, R. Jones, and M. Potts, "*History repeats itself: Repeated queries in Yahoo's logs*". In 29th *SIGIR Conf. Information Retrieval*, pp. 703–704, Aug 2006.

[30] L. von Ahn, M. Blum, and J. Langford, "*Telling humans and computers apart automatically*". *Comm. ACM* **47(2)**, pp. 57–60, Feb 2004.

[31] F. Yu, Y. Xie, and Q. Ke, "*SBotMiner: Large scale search bot detection*". In 3rd *Intl. Conf. Web Search and Data Mining*, Feb 2010.

[32] L. A. Zadeh, "*Fuzzy logic*". *Computer* **21(4)**, pp. 83–93, Apr 1988.

[33] Y. Zhang and A. Moffat, "*Some observations on user search behavior*". In 11th *Australasian Document Computing Symp.*, Dec 2006.