The Hebrew University of Jerusalem
School of Computer Science and Engineering

Master Thesis

Distinguishing Humans from Robots in Web Search Logs

Author:
Omer M. Duskin

Supervisor:
Prof. Dror G. Feitelson

December 2, 2009

Acknowledgment

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

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. We present a few behavioral criterions, such as the maximal number of queries in a day, query submittal rate, minimal interval of time between different queries, and more. By studying the interaction between the behavioral criterions of the humans vs. the robots, we were able to classify the users in the log files. In order to find the best classification, we used a grading method that helped us to compare the results from several analysis. Our conclusion was that analysis that combines a few criterions may give the best classification between humans and robots.

Contents

1 Introduction 4

2 Data Sources and Methodology 6

3 Scope and Limitations 8

4 Criterions and Measurements 10

4.1 Introduction 10
4.2 Maximum Number of queries in a Day 12
4.3 Maximum Number of queries in a Minute / Average rate . . . 13
4.4 Average Number of queries in a Day 13
4.5 Minimum Time Interval 13
4.6 Median Time Interval 15
4.7 Average Number of words in search queries 15
4.8 Repetition 17
4.9 Periodic Repetition 17
4.10 Continuous Working time 18
4.11 Correlation with time of day 18

5 Analyzing Principles 19

5.1 Analyzing Methodology 19
5.2 The Log Analyzer utility 20
5.3 Basic Analysis Scenarios 25
5.4 Analyzing Confidence 34

6 Results 36

6.1 Method 36
6.2 Analysis 40
6.3 Conclusion 46

1 Introduction

Web search engines record and store the interactions between searchers and search engines in log file on the server. The log file contains data regarding to the user search activity, such as the computer's identifier of the user, user query, search engine access time, and the clicktrough. Once we have the log file, we can analyze it in order to answer research questions related to the users behaviors.

In many situations we face the need to distinguish between human and software applications that run automated tasks over the internet, known as bots or robots. There are some past works regarding the distinguishing between human and robots in different fields, although it seems that the field of web search log remains somewhat untouched. As we will briefly show, in each field the methods that are being used are practically different because the data we can use is different.

One example is websites who want to distinguish between human and robots. The need emerges when that website owner wants to check for example which product is the most wanted using the number of clicks, and he obviously doesn't care about robots clicks. In this case we may use a CAPTCHA, which is a type of challenge-response test to ensure that the response is not generated by a computer. We may ask the visitor to look at an image containing a series of distorted letters and numbers set against a busy background, and to type in the sequence of characters they see. The idea is that humans can easily recognize the shape of letters while a computer will find it difficult to do so.

Another example is massively multiplayer online games. There may be robots that are used in order to cheat by writing a script to automate actions in a game without actually playing. This has a severe effect on honest players and impacts motivation to continue the game, threatening online game providers' business model. In this case we can analyze the character's movement data, e.g. to extract waypoints and detect repeated paths. This allows us to find movement patters that appear frequently, indicating that a character is controlled by a script and not a human player [10].

Another related example is the need to detect robot users in on-line chat. Those robots target popular chat networks to distribute spam and malware. They pose a serious threat to Internet users [11].

Our research focused on distinguishing between humans and robots in web search engines log. 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. Web search engines are trying to better support humans through the development of new web search tools, based on the humans behaviors. We would also like to learn about the page access statistics such as the number of site visits originating from the search engine, or the keywords that visitors used to find some site in the search engine.

Another thing that can be learned from the web search log is the site entry pages, which are the pages that a visitor begins to view the site from. This may be the website homepage or some other secondary page. After arriving, the visitor will typically view a few pages and then exit the site. Therefore, knowing the entry points of your website is useful.

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 [6, 8].

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 [5]. 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 catalogue 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. 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

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
AOL06	1 Mar - 31 May 2006	28,898,361 queries

who submitted more is assumed to be a robot [13, 8]. This is based on the argument that such a threshold is much higher than the average session length. 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 [5]. To the best of our knowledge more involved approaches have not been studied.

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 [9]. 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 used four datasets as described in Table 1. We used logs of popular web search engines in order to ensure the generalization of our results. Since AOL was our biggest and newest web log, in most of our research we focused it. Another Advantage regarding the AOL log was the source ID, that seems to be constant and not dynamic like the MSN session ID field. This fact allows us to retrieve data regarding the users activity along the all logged period. Since the MSN log consists dynamic ID, called session ID, it's impossible to analyze user activity along the all log.

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.

In our research we distinguished between two cases. If the same query is repeated with exactly the same timestamp at second resolution and the same clicktrough, we decided to filter it because it doesn't make sense and may be caused due some error. But, if the repetitions included the same timestamp but not the same clicktrough, we didn't filter themo ut and we actually used them in order to identify robots.

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. The fact that we are not sure about the nature of the source ID field is one of our obstacles. The problem exacerbated when the source ID field is altered occasionally to improve anonymization, which may cause some robots will be classified as human while using some criterions, such us number of queries in a day.

Another field that exists in all of the logs is the query string. Sometimes we may find additional data regarding the query, such as the number of clicks in the page that resulted from that query. Basically, in order to develop a generic heuristic, we didn't use data that doesn't exist in all of the web logs as a standard. For this reason, the clicks data was not used in our research.

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 [7].

Our goal is to be able to identify those users that are not representing a human, meaning robots and meta search engines, and the only way to find them is by using complex analysis. Our approach is to find a group of robots characteristics that can be learned from the given web search logs. The purpose of this is that by analyzing a given log and identifying the user's characteristic we will be able to reliably distinguish between humans and robots in the web log.

The methodology for this research is based upon analyzing web search engines logs. In addition to the massive size of the logs, that contains hundreds of thousands of searchers and millions of queries, and therefore can minimize abnormal behaviors and give us reliable results. we enjoy the benefit of using real queries. Moreover, as Spink described in her analyses of web search logs [14], "the advantage of this kind of research over others is that the data comes from real users with real information needs submitting real queries. Accordingly, it provides a realistic glimpse into how web users search, without the altered behavior that can occur with lab studies or survey data."

3 Scope and Limitations

The basic problem in the research is verifying the truth of the results. Given that there is no way to accurately identify human from robots [12], researches that relays on search engines logs for data must either ignore the problem or assume some behavioral differences between humans and robots. Since we don't have some source of information that can verify our results, we have to use some heuristic methods to increase our confidence in the results. We have a few analysis methods that can be separate into two groups. One group will contain those methods that can identify only a small percentage of the robots, but with a high reliability, and another group will contain methods that can identify many more robots, but with a lower reliability. We wish to find a combination of methods from those two groups that will give us the best discrimination in both quantitative and reliability aspects.

Another basic challenge that we are dealing with is the available data for the analysis. There are many kinds of web robots and there are many ways to identify them, depending on the field. Most of the ways require some data regarding the robot. This data is usually available when we are talking about servers. Web server's logs contain significant data that can be used to identify robots. This includes

- Explicit identification as a robot in the user agent field.

- Making a request for the robots.txt file.
- Downloading web pages but not downloading embedded images.

Thus it is desirable that in the future web server logs will be made available together with web search logs. But given the fact that this is not the case, analyzing the web search logs is not trivial.

Our research is based upon analyzing and classifying the users according to their behavior. We are basically trying to identify robots according to their behaviors, as can be learned from the search logs. We have established a series of behavioral criterions. For each criterion we have two thresholds, one for humans and one for robots. Based on these two thresholds each user may be classified as a robot or as a human. Users may also remain unclassified in some cases, when the user's value falls between the two thresholds. Although this way has its reasoning and benefits, it also has some disadvantages.

The obvious problem is that robots can easily imitate human behaviors. In that case, if a robot will act exactly like a human, it will be impossible to trace it using the limited data that is available in the web log. This problem is reflected in some robot design manuals that are available on the web. In those manuals one can find specific instructions how to imitate human behaviors, such as:

- Make sure that your robot runs slowly.
- Don't loop or repeat.
- Run at opportune times.
- Don't run often.

Another obstacle that we are facing is the fact that we don't know the exact meaning of the source ID column in each web log, which means that for a specific IP address there may be a few rotating IDs in the log.

In order to overcome the above obstacles we needed to find enough robot characteristics. We realized that the more robot characteristics we will have, we will be able to retrieve more reliable results. By using several robot characteristics, we will be able to increase the chances of successful identification by using techniques of union and intersection of the results.

There are many web logs, and each one contains different data, and so different results may be learned from it. Basically, we aspired to use web logs that will contain many queries and many users. Another criterion for a good web log for our analysis is that the user source ID altering rate will be minimal, in order to have minimal effect upon our analysis. While we had a preferred web log that fitted to the above description there is an advantage of using several web logs in order to crosscheck the results so that the conclusions will be more representative. Therefore, we mainly used the web logs that fitted our descriptions, but verified the results using the others logs.

4 Criterions and Measurements

4.1 Introduction

Before getting into analyzing the search engine logs, we need to understand how we can distinguish between robots and humans. As mentioned before, Jansen and Spink, in their analyses of web search logs, simply use a threshold of 100 queries in a day to distinguish humans from robots. The question is, while given only the log file, can we do better?

Our approach, for distinguishing between humans and robots, is based upon two thresholds, one for humans and one for robots. Each user will get a parameter according to some criterion. This parameter will naturally be lower than one threshold, greater then the other, or between them. According to the criterion definition and to the user' parameter location regarding the thresholds we will classify the user as human, robot, or as unclassified user.

Our research is based on studying the activity patterns of search users along multiple dimensions, and trying 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 regarding the nature of each user. Therefore we need to start from heuristics such as those mentioned above.

The dimensions that we intend to investigate include the following:

- The maximum number of queries submitted in a day. We choose a day as a measure that can be consistent between different logs, in contrast with absolute number of queries in a log.
- The maximum number of queries submitted in a minute. When searching for the maximum number of queries in a minute we can get the maximum sending rate at which queries can be submitted. This criterion

is different from the previous one because it checks a different behavior character. It checks possible sending rate and not work quantity.

- Average Number of queries in a Day. As opposed to the first criterion, in this criterion we calculates the average number of queries in a Day.
- The minimal time interval between successive different queries. Humans are expected to require tens of seconds or more to submit new queries. This criterion may be referred as maximal momentary rate, as opposed to the maximal rate in a whole day.
- Average number of words in search queries. This is a behavioral criterion. There were many studies regarding the average number of words that are being used in a search query, and we may use it in our research.
- The regularity of the submitted queries. In this criterion we are counting the repetition of the same query. Users who submit the same query thousands of times are assumed to be robots.
- Periodic Repetition. 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.
- 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. Humans are 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.

During the search for appropriate criterions we realized that we can divide the criterions into two logically separated groups. One group will contain the criterions that with a certain parameter will indicate that the users that were classified as robots are indeed robots. If we can find a criterion with some parameter that a human can not possibly reach, we will say that this is a strong criterion. Strong criterions are those who make the classification by physical capabilities. They will help us to be surer that the users that are tagged as robots in this group are actually robots. On the other hand, if in some criterion we will distinguish the users by the typically expected behavior and not by an unreachable parameter, we will call it a weak criterion.

By separating the criterions into these two groups, we can find the users that are classified as robots by the strong criterions, and then we can do some analyzing using the other criterions and by comparing the results we will be able to adjust the right thresholds.

4.2 Maximum Number of queries in a Day

At first, counting the total number of queries may sound as a legitimate criterion. One can assume that by taking the users with maximum queries in a given log file we can identify the robots users. Although it is possible, the fundamental problem in this approach is that different log files covers different time durations, so we can not find some parameters that will indicate a robot or a human behavior pattern that will be generally correct, for any given log, which is our main goal. If the different logs cover different durations, there is no sense to learn from the total amount of queries of a user in one log to another.

Therefore we focused on counting the number of queries in some fixed period of time, say a day. This way, this criterion could be used for comparing user activities among different logs in order to find a pattern. For each user we will calculate the maximum number of queries in one day.

```
for (each iUser in log) {
    for (each iDay in log period) {
        num     = getNumOfQueriesInDay(iUser, iDay);
        maxNum  = max(maxNum, num);
    }
}
```

The reason for that is that in a big logs the average may be low in a way that won't reflect user abilities. The total queries in a day can tell us the intensity of the user requests, while assuming that it's unlikely for a human to send more than some amount of queries in a day because it means that the user works for all day without a break, which isn't a typically human behavior. For instance, we can set 100 as a robot parameter, because it means that the user sent, in average, a query every 3 minutes in 5 hours duration, and it's not a normal human behavior. After analyzing the logs we saw that human users tend to send even less than 50 queries in a day, so we were able to tighten our thresholds. We believe that With the above thresholds this criterion can be consider as a strong criterion.

4.3 Maximum Number of queries in a Minute / Average rate

We can gain extra data by retrieving the maximum number of queries in a minute. This criterion can tell us about the user's maximum sending rate. Counting the number of characters that were sent in a minute may seem like a good idea in order to verify the rate at which users type. Although it may sound as a strong criterion, because humans are limited to around 200 characters per minute, we do not use it. The reason for that is that we have no way to know whether the user actually typed all the query or perhaps he used copy & past, which may allow a him to send a significant amount of characters in one minute but not by typing them. If the user sends 20 queries in one minute, it says that the average rate of the user is one query every 3 seconds, which is unlikely to be a sending rate of a human, even by using copy & paste. Up to now we introduced two strong criterions. Although there are more strong criterions, not all of the criterions are strong.

4.4 Average Number of queries in a Day

In this criterion we will calculate the average number of queries in one day. As apposed to the maximum number of queries criterion, this criterion come to check the behavioral differences more than the ability of the user. We can use this criterion with high thresholds and by that use it as a strong criterion, but the problem is that the average tend to be low and therefore using a high threshold will lead to significant amount of unclassified users. As said before, the main problem of this criterion is the weakness to extreme values that are likely to exist in big logs, like the ones that we are using. Nevertheless, we decided to check whether this criterion can give us added value to the analysis.

4.5 Minimum Time Interval

In this criterion we are taking the minimum of all time intervals between two successive records of different queries. The logic behind this criterion is that a human can't send two queries without waiting some time between them, in contrast to robots that can send a few queries within a second. In this manner, we can add this criterion to the hard criterions group. In fact, the search logs are sorted by user ID & time, so the check for the minimum time interval is pretty simple. For each user we will save the minimum of all time intervals.

```
for (i=0; i<numOfUsers; ++i) {
   for (j=1; j<user[i].numOfRecords; ++j) {
      if (user[i].records[j].query != user[i].records[i-1].query) {
            timeInterval =  user[i].records[j].time -
                            user[i].records[j-1].time;
            user[i].minTimeInterval = min(  user[i].minTimeInterval,
                                            timeInterval );
      }
   }
}
```

The reason that we are taking the time between two different queries and not also the time between same queries repetition is to avoid a situation when a user clicks twice on the same query by accident, and it doesn't mean that this is a common behavior of that user and it also doesn't mean that this user is capable of sending two different queries in one second. Another kind of error that we wish to avoid is log errors. It is possible that sometimes two different queries will get the same time stamp, although they were sent in different times.

In order to avoid this kind of mistakes, that some user may have two queries in the log as if they were sent in the same time because of some error, we will choose some window with maximum queries and will examine it. By doing so, we are reducing the chance of using corrupted data. On one hand, this criterion can give us a pretty strong classification, since we are talking about the physical abilities of the robot vs. the human, rather than the common behavior of a human vs. a robot. On the other hand, there are a few problems while using this criterion. The first problem is the reliability of the data. Although we are trying to avoid it, it is still quite possible that as a result of some communication error, or some other error, two queries that weren't sent together will be recorded with the same time stamp. In this case, a human user that actually sent his queries with a delay from one to another will be classified as a robot just because of a one time error. Although this problem might exist, it might not be so common and therefore not significant in log files that are big enough. In addition, instead of taking the minimum time interval, we can use the median time interval, which will solve most of the occurrences of the problem.

The second problem of this criterion is that its classification is pretty good only for extreme parameters, such that a bot is a user with minimum time of 0 seconds and a human is a user with minimum time of 60 seconds. In that case, most of the chances are that the user that are classified as bots are actually bots, but the problem is that most of the users remain unclassified. We had two optional directions in order to overcome this problem. First, we tried to find a combination of parameters that will give us some significant amount of users that are considered as bots, while keeping the amount of unclassified users low. The second direction was to combine this criterion with another one, so that the other one will give us a wide classification and this one will be used to strength the classification of the first.

4.6 Median Time Interval

In this criterion, instead of just taking the minimum of the time difference between two successive different queries, we take the median time between two successive different queries. As mentioned before, the benefit of using this criterion is to overcome possible errors in the log. Since the minimum time interval criterion may lead to incorrect results we may use the median time interval instead. The drawback of using this criterion versus using the minimum time interval is that it is less useable as a strong criterion. The reason for that is that although it is possible that the median time interval of some user will be zero, and then we can be quite sure that this user is a robot, there is no prevention that robot will also wait sometimes between two queries, and then the median will not be zero anymore. The result of using zero as a robot parameter may be losing a significant number of robots.

4.7 Average Number of words in search queries

In this criterion we will calculate for each user the average number of words of search query. This criterion is based on the assumption that robots queries tend to be longer and more sophisticated, while humans tend to send short and simple queries. This criterion can be split into two different sub-criterions: regular queries and questions queries. The reason for that is the assumption that questions queries tend to be longer than regular queries. This criterion has obvious drawback, robots may send short queries as well, so it is hard to distinguish between humans and robots using that criterion alone. We may choose a relative large number of words as a parameter in

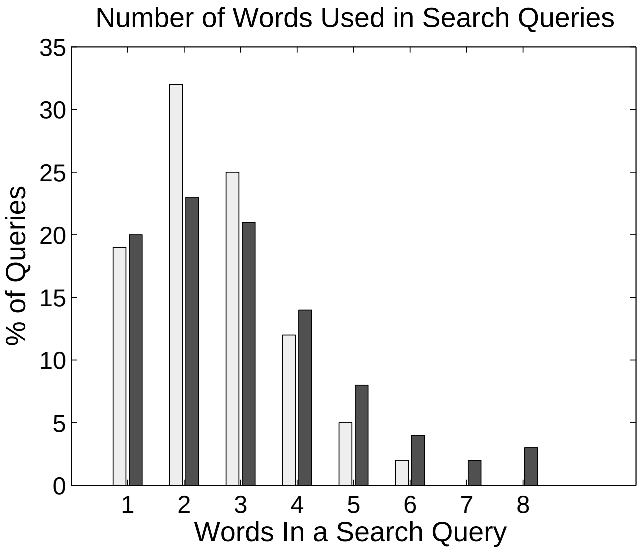


Figure 1: *number of words per search query 2004 vs. 2009.*

order to filter the humans, but by doing so we may also lose some robots. The conclusion of these drawbacks is that we need to be careful with this criterion and to do some research before using it.

The advantage of this criterion may be the fact that the average number of words in search queries in past few years was quite the same. It help us to know the expected humans behavior in order to find users with abnormal behavior in order to classify them as robots. In fact, there was an increase in the average number of words per search query just in the last year. According to The Marketing Pilgrim, Andy Beal, there is significant increase in 8-word search queries in last 5 years [4]. He compares the recent HitWise data to a OneStat report from 2004.

> According to hitwise.com, theres been a significant drop in the number of people using 2-and 3-word searches, while those using 5-words and above are becoming more common. Likely from those of us that cut-and-paste large quantities of text into Google (Fig. 1).
>
> One interesting observation: the number of people using 1-word searches has remained surprisingly similarin fact, showing a small increase in the past 5 years. This either means that the search

engines are getting better at delivering relevant search results, or more "brand" searches are being made, or some of us are just committed to being lazy with our searching.

4.8 Repetition

In this criterion we are using the maximum number of times that the user re-sent the same query string. There are several reasons for the existence of query repetitions in web log. One reason for that is the attempt to re-find previous result that was clicked during another search session. The base assumption of this criterion is that robots may resend their query without any change, while humans tend refine their queries and not to repeat over the same query, especially after a while. There might be a problem since human users may click on results that had not been clicked before or ask for the next results page of a query, which in some search engine will be translated to another record in the log with the same query. Nevertheless, we believe that it's not a big obstacle since the number of humans repetition should be low anyway, because of following reasons:

- Most of the users won't use more than a dozen results of a query [2].
- Most of the users that repeats the same query will click the same result, and won't start a big search session from the beginning. "People were clearly much more likely to click on results they themselves had seen before" [15].

Anyway, for some large number of repetitions it can be a useful criterion.

4.9 Periodic Repetition

In order to strengthen the previous criterion, of repetition of the same query, we might check for repetitions that come in the same time intervals. This fits to scheduled bots while the chance that a human send exactly the same query over and over with the same time intervals between the queries is very low. We can refer to this criterion as a strong one, although it is likely that there are many robots that will not fall into this criterion. Because we have timestamp at second resolution, user that have even three periodic repetition will be considered as a robot.

4.10 Continuous Working time

In this criterion we will check the maximum duration of continuous activity sessions. This refers to query session length, which is the period of time that the user continues submitting query records. We will search for the minimum of the maximum resting time between queries. This criterion is based on the assumption that while humans need to rest for a few hours in a day, robots don't.

We can deduce that if a user is active for a long period of time without any significant rest it's probably a robot. For example, if a user is active continuously for say more than 10 hours, we'll assume it is a robot.

There is still another thing that we need to take care in this criterion. When we are saying that some user was sending queries for 10 hours without a break, what does it exactly mean? There is always some time between two successive queries, so when is that time long enough to be regarded as a break? In other words, we need to decide what is the longest time interval between two successive queries that doesn't break the continuity of sending queries. Choosing the right factors for this criterion is not an easy task. We need to decide how long should a user be active, and what does it mean to be active, what is the time interval that doesn't break the continuity of sending. For example, if we say that a robot is a user that is active for 20 hours with breaks of 5 hours, then it is possible that some human user will send one query every 5 hours, and will be considered as a robot by mistake.

4.11 Correlation with time of day

This is a behavioral criterion. In case we will see a user that is active during night hours we will assume it's a robot. The basis for this criterion is the assumption that humans are more active during daytime. As said before, the problem using this criterion is the fact that there may be users that are not in the same time zone as the search engine, which in result will cause a misleading time stamp. The time zone problem means that we can't really know from the server log what is the time for the user. It's may be morning time in the server log while it is actually evening time for the user. This problem may be minor regarding the AOL search engine log, as a result of the massive log size and the probability that most of its users are within the same time zone, more or less.

5 Analyzing Principles

5.1 Analyzing Methodology

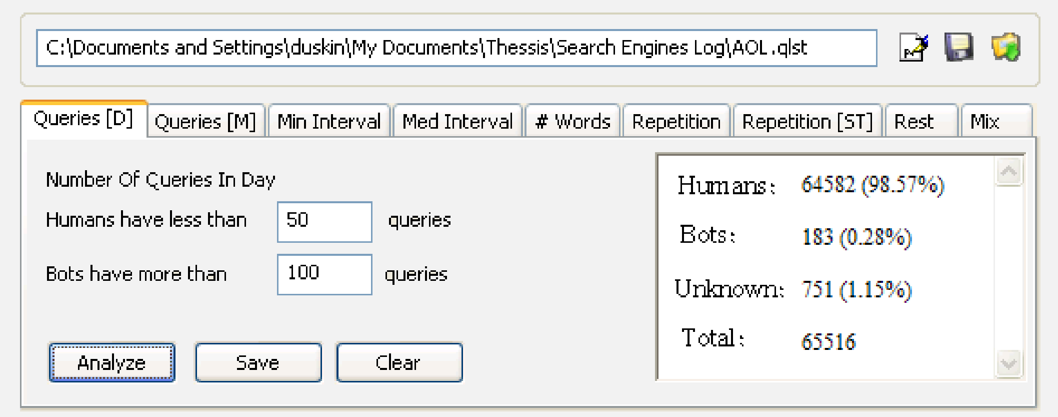
As previously said, we separated our analyzing criterions into two groups. One group contains the criterions that make the classification by physical capabilities of robots vs. the physical capabilities of the humans, and the other group makes the separation by the users' behavior. Because humans can not imitate robots' physical capabilities but robots can imitate human behavior, we refer to the first group of criterions as the strong criterions.

The analysis process is done by comparing the results of one criterion with the results of another. By combining several criterions we increase our confidence in the results. For example, if we have a criterion that can give us a distinction between humans and robots with some parameters, we can strengthen it by showing that the users that were marked as robots are also marked as robots by another criterion, that we are more positive about. The same goes for humans.

When we came to the analysis process we used the fact that we have two kinds of criterions. We started with the strong criterions, which gave us strong distinction between humans and robots. We used these criterions to create a group of users that we are most certain that they are robots. Then we proceeded to analyze the log with the rest of the criterions, while verifying the strength of each criterion by the amount of robots being discovered.

The reason that we were not satisfied by the first strong criterions is the number of users that were classified. The biggest problem with those criterions is that many robots do not necessarily use the maximum of their computing power, and therefore in the terms of physical capabilities criterions it is impossible to discover them. By combing the power of the behavior criterions group with the confidence of the physical capabilities criterions group we can achieve a much better classification, leaving less users unclassified.

For example, when counting the number of queries that some user sent in a minute, we can conclude that a user that sent more than 100 queries must be a robot since it's impossible for a human to send queries in such rate. Let's call the group of users that sent more than 100 queries in one minute fast-robots-group. Now, while we are searching for the right thresholds for the criterions from the behavior criterions group, if we were able to classify all the users from the fast-robots-group as robots we can be more confident about our thresholds. In general, the analysis steps are:



C:\Documents and Settings\duskin\My Documents\Thessis\Search Engines Log\AOL.qlst

Number Of Queries In Day

Humans have less than 50 queries

Bots have more than 100 queries

Humans:	64582 (98.57%)
Bots:	183 (0.28%)
Unknown:	751 (1.15%)
Total:	65516

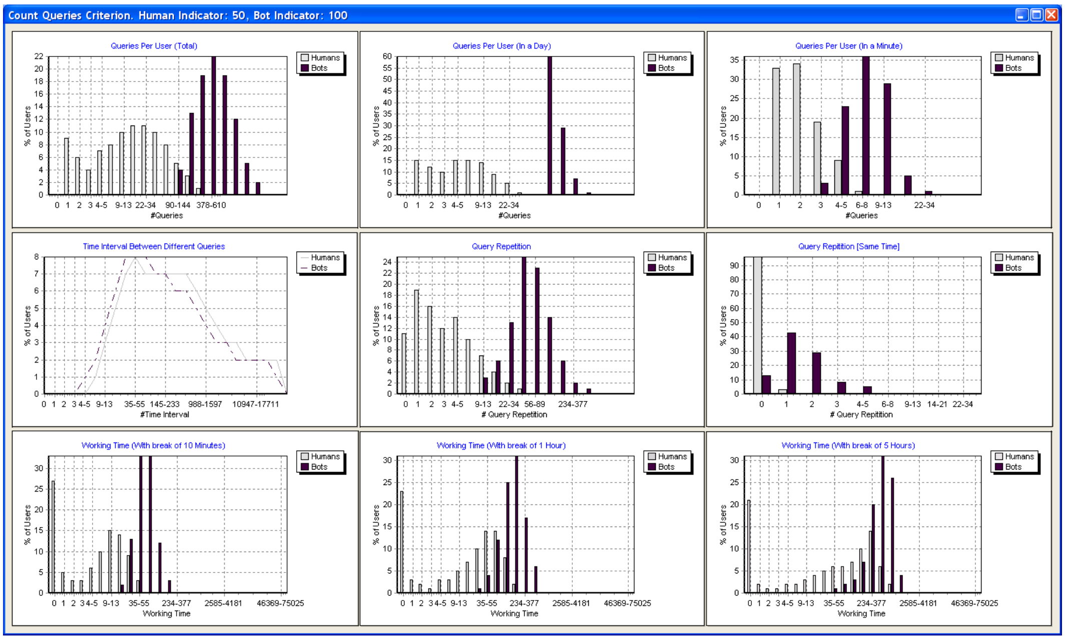
Figure 2: *screen shoot of Log Analyzer - Distinguising process.*

- Classifying the users using some strong criterion.
- Analyzing human behavior vs. robots behavior using other criterions.
- Repeating the previous stages in other logs.

5.2 The Log Analyzer utility

In this section we will describe the Log Analyzer utility. Please notice that unless indicated differently, the word "users" refers to the users in the web log and not to the users of the tool. In order to analyze the given logs, we developed a tool dedicated for this purpose. The tool, which we called "Log Analyzer", is pretty simple. For a chosen log file, the Log Analyzer gets as input a chosen criterion and the thresholds for it, and gives as output the users that were classified as robots, the users that were classified as humans, and those who were in the middle and remained unclassified. In the following picture we can see the first stage of the analysis where the user of the application chooses some log file to analyze and a criterion for the classification of the users into the human and robots groups. After giving the thresholds for the criterion we will get the number of users that were classified as human, and the percentage of this number from the total number of users in the log, and also the number of robots and the unclassified users (Fig. 2).

After classifying the users into the groups of human and robots, it is possible to analyze the human behavior vs. the robots behavior, using other criterions. The results are presented in graphs that make it easy to compare



Count Queries Criterion. Human Indicator: 50, Bot Indicator: 100

Figure 3: *screen shoot of Log Analyzer - The "All" window.*

the users' behavior. There is a list of criterions available for analyzing the users' behavior, and next to that list there is a graph according to the user criterion selection from the list (Fig. 4). An exception is the first line in the list, with the "[All]" mark. Clicking on this line will open a new window that contains all the available graphs in one place (Fig. 3).

In the first step we distinguished between humans and robots using some criterions with chosen thresholds (Fig. 2). By doing so we have created two users groups: those who are probably humans and those who are probably robots. In the next step we can look at the distribution of one criterion, such as the number of query repetition, in those two groups (Fig. 4). If the distributions that we got are separated, than the two criterions agree with each other. This means that the analyzing criterion in the first step gave us good classification, and that humans and robots do behave differently also in the second criterion. The above is right in case that the criterions from the first and second steps are not related, so the correlation between the results can be explained only by the fact that the classification was good.

Another ability that we added is cross analysis. After running a few classifications using different criterions, we can find out the intersection of

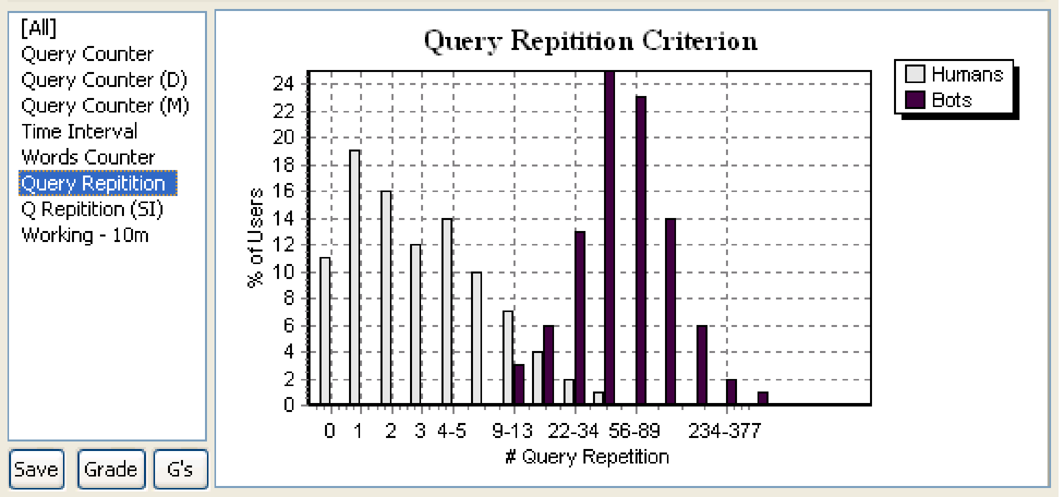


Figure 4: *screen shoot of Log Analyzer - Analyzing humans beahviour VS. robots beahviour.*

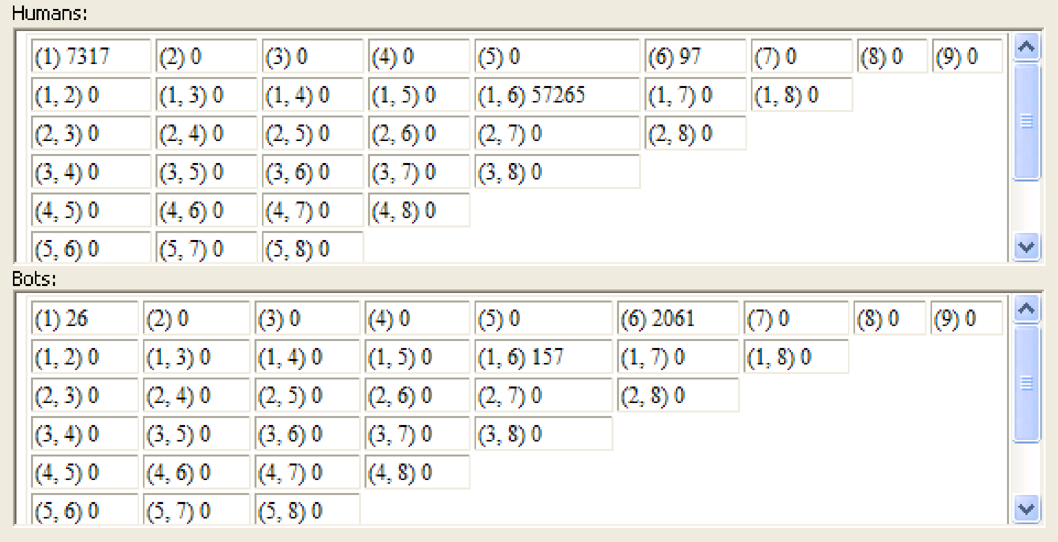
Table 2: *Classification sample.*

Criterion	Human	Robots
Number of Queries	64582 (98.57%)	183 (0.28%)
Query Repetition	57362 (87.55%)	2218 (3.39%)

all the previous analysis. This is useful when we are trying to learn about the strength of some criterion. We can take a strong criterion, that with some thresholds is giving us a reliable classification of robots, and another criterion which we want to check its credibility. The tested criterion will be considered more reliable if with some thresholds it succeeds to classify as robots all the users that were classified as robots using the reliable criterion.

For example, let's say that we made a classification analysis using the first criterion, which is the number of queries criterion, with the thresholds of 50 & 100. After this we made another classification analysis using the sixth criterion, which is the query repetition criterion, with the thresholds of 10 & 30. The summery of the classification analysis is shown in Table 2. As we can see, by using the number of queries criterion we classified more users as human and by using the other criterion we classified more users as robots.

Now, lets say that we would like to see the number of users that were classified as humans in both of the analysis, and the number of users that



Humans:

(1) 7317	(2) 0	(3) 0	(4) 0	(5) 0	(6) 97	(7) 0	(8) 0	(9) 0
(1, 2) 0	(1, 3) 0	(1, 4) 0	(1, 5) 0	(1, 6) 57265	(1, 7) 0	(1, 8) 0		
(2, 3) 0	(2, 4) 0	(2, 5) 0	(2, 6) 0	(2, 7) 0	(2, 8) 0			
(3, 4) 0	(3, 5) 0	(3, 6) 0	(3, 7) 0	(3, 8) 0				
(4, 5) 0	(4, 6) 0	(4, 7) 0	(4, 8) 0					
(5, 6) 0	(5, 7) 0	(5, 8) 0						

Bots:

(1) 26	(2) 0	(3) 0	(4) 0	(5) 0	(6) 2061	(7) 0	(8) 0	(9) 0
(1, 2) 0	(1, 3) 0	(1, 4) 0	(1, 5) 0	(1, 6) 157	(1, 7) 0	(1, 8) 0		
(2, 3) 0	(2, 4) 0	(2, 5) 0	(2, 6) 0	(2, 7) 0	(2, 8) 0			
(3, 4) 0	(3, 5) 0	(3, 6) 0	(3, 7) 0	(3, 8) 0				
(4, 5) 0	(4, 6) 0	(4, 7) 0	(4, 8) 0					
(5, 6) 0	(5, 7) 0	(5, 8) 0						

Figure 5: *screen shoot of Log Analyzer - Analyzing classification intersection.*

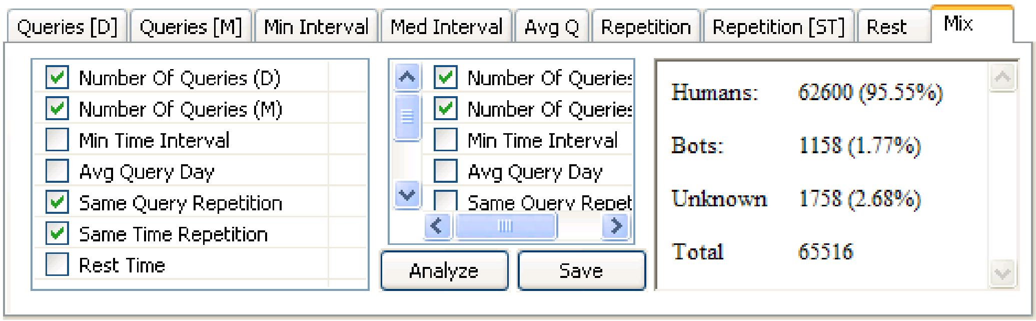
Table 3: *Classification of Intersection.*

Criterions intersection	Human	Robots
Only in number of queries	7317	26
Only in query repetition	97	2061
Both of them	57265	157

were classified as robots using both of the criterions. We can find it out using two tables, one for the humans and one for the robots (Fig. 5). Each table shows the cross analysis of the two criterions. In order to save place, each criterion is represented by its tab order in the main window and not by its name.

The results of the classification intersection, as shown in the print screen, are summerized in Table 3. As we can see there is a pretty big intersection between those two criterion analyses. We can consider using it in order to decide whether or not the best way is to use several criterions in order to get the best distinction.

The last ability of our tool is to create a mix of some criterions. The reasoning behind this ability is the desire to create a group that contains all the users that we are pretty sure that they are robots. When we will have some other analysis, we will be able to search for the intersection between



☑ Number Of Queries (D)
☑ Number Of Queries (M)
□ Min Time Interval
□ Avg Query Day
☑ Same Query Repetition
☑ Same Time Repetition
□ Rest Time

☑ Number Of Querie
☑ Number Of Querie
□ Min Time Interval
□ Avg Query Day
□ Same Query Repet

Humans:	62600 (95.55%)
Bots:	1158 (1.77%)
Unknown	1758 (2.68%)
Total	65516

Figure 6: *screen shoot of Log Analyzer - Distinguishing process, using union of criterions.*

the users that we just classified as robots, and that we are pretty sure that they are robots, to the users that are classified as robots in the new analysis, and by that we can be more confident about the results.

The "mix" tab, next to the criterions tabs, gives us the union of some other criterions (Fig. 6). This tab is usable only after we made a few analysis and we wish to combine a few of them. We have two lists of criterions. From the first list we can choose the criterions, that we have already analyzed using them, in order to mix their results While using a number of criterions in order to distinguish humans from robots, we may want to choose criterions that we regard as strong criterions. For this reason we have the second list of criterions, that allows us to choose the strong criterions from them.

The difference between the strong and regular criterions is in the analysis of the results. When analyzing the results of a strong criterion, all the users that were classified as robots are tagged as robots for the final over-all results of all of the criterions. In table 5 we can see an example of two criterions scenarios, when one of them is a strong criterion.

We have different operation when dealing with the rest of the criterions, that aren't strong. For the rest of the criterions, if some user was marked as robot, it will be tagged as robot only if he wasn't classified as human by any other criterion, as shown at table 4. The same works for humans. A human by one criterion won't be tagged as human if it was classified as robot by another criterion. The treatment for the human users is identical for the strong criterion and for the regular criterions.

Table 4: *mix criterion analysis*

		first criterion		
		human	?	robot
2'nd criterion	human	human	human	?
	?	human	?	robot
	robot	?	robot	robot

Table 5: *mix criterion analysis, including strong criterion*

		strong criterion		
		human	?	robot
2'nd criterion	human	human	human	robot
	?	human	?	robot
	robot	?	robot	robot

5.3 Basic Analysis Scenarios

We previously discussed the analysis methods that we used in our research. In order to clarify it and to give some sense to the final results that we will present later on, we will present some analyzing scenarios that we had. Some of the scenarios were more successful than others, but we do believe that we can learn from all of them. These scenarios were the base of our research because they gave us the opportunity to learn the importance of fine-tuning of the thresholds for the criterions.

We started our research with the criterion of number of queries per day with a threshold of 100, for both robots 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 robots. Without any previous knowledge about robots or human behavior, we decided that it will be good to start with the classification assumption of Jansen and Spink.

What we have done is to analyze the "All the Web" log using our log analyzer utility. As said before, we decide to start with the criterion of number of queries per day with a threshold of 100 for both human and robots. The output of the above was a classification of the total users in the log into three groups, the robots, the humans and the unclassified. Obviously the unclassified group was empty, because our threshold did not leave any place for uncertainty. As a result we got that 99.24% of the users were human, and the rest were robots. Please notice that although we did classify all of the users, the clear disadvantage is that the amount of confidence in our result is not very high, because we can not really be sure that every user that sends less than 100 queries in a day can not be a robot.

Another output that we got from the log analyzer utility is a graph showing us the segmentation of the robots and the human by any other chosen criterion. For example, after retrieving the previous results, we plotted a graph that showed us the number of queries per day that humans tend to send and the number of queries per day that robots tend to send. Obviously we get a clear distinction that humans tend to send less than 100 queries in a day, because this was the basis for our analysis, so we need to take this graph in the right perspective. But we can learn something out of it in order to sharpen our thresholds. We used both AllTheWeb and AOL logs in order to classify the users according to the max number of queries they sent in a day. After the classification, we checked the number of queries of the classified users. As we can see in Fig. 7, most of the human users tend to send only a few queries in a day, a lot less than 100. It seems that there aren't many users that sends more than 50 queries in a day, regardless of our thresholds. We can use this in order to run a new analysis where the threshold for the humans will now be around 50. The purpose of the process is to find the most extreme thresholds that will still give us a reasonable distinction for most of the users.

Before we proceed, please notice that this graph, and all the rest of the graphs, represents the behavior of the human vs. the behavior of the robots, which mean that the bright bars, for instance, represent 100% of the users that were tagged as human and not of the total number of users. Similarly, the dark bars represent 100% of the users that were tagged as robots.

By changing the human threshold to 50 we got that there were 98.03% human, which is not far away from the previous conclusion. So we actually narrow down the range of number of queries for which a user may be classified as a human, without effecting the final classification results by much.

As we can see, we can learn something from the distribution of the number of queries per day for human and robots, but we must remember that these results are somehow biased because we started by saying that human are those who have less than 100 queries, so it is not a big surprise that we got that.

In order to be more certain about the results we can check the behavior of the users on another criterion and to compare the behavior of the users that

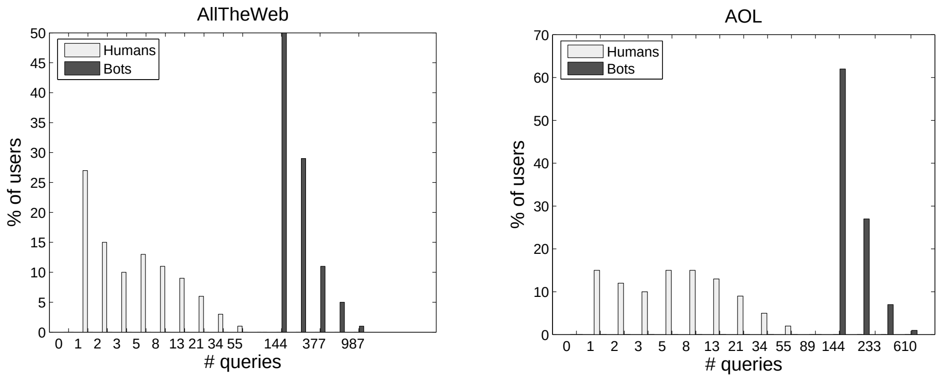


Figure 7: *Distributions of number of queries submitted by users identified as humans or robots according to their number of queries in a day.*

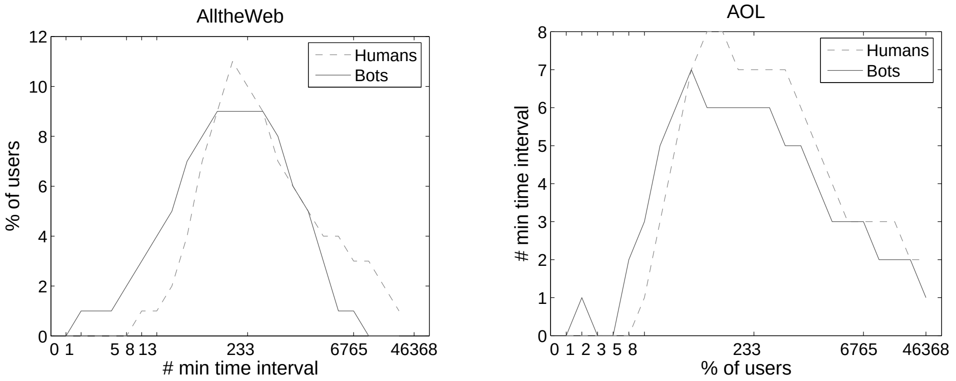


Figure 8: *Distributions of time interval between different queries of users identified as humans or robots according to their number of queries in a day.*

we tagged as human vs. the behavior of the users that we tagged as robots. We can check whether there is a difference in the behavior in selected fields and to think whether or not the difference makes sense. So the next logical step will be checking the behavior of the users that we previously classified as human and the robots, using another criterion.

We decided to continue by checking the time interval between different queries of the users who were marked as human versus the time intervals for robots. This way we can better learn about the robots and the human behavior.

As we can see from the min time intervals analysis (Fig. 8), the minimum time interval of the users who were marked as robots is 0 and the minimum time interval of the users who were marked as human is about 5. By using the criterion of number of queries per day in order to classify the users as human or robots and then analyzing there behavior by the time interval between different queries, we can see that robots can send queries faster than humans, which makes sense and can strengthen our thesis about the criterion and its parameters.

We started with checking the above thresholds on the AllTheWeb log, but of course it is not enough, we need to analyze also other logs and to compare our results. We continued by using the above criterions and thresholds on the AOL log, in order to verify the consistency of the results. The distribution of the classification wasn't so far from the one we got using the AllTheWeb log. We got that 98.57% of the users are human, but we got only 0.28 of the users are robots, while the rest of the users remain unclassified. Therefore we changed the robots threshold on number of queries to be 60, and then we got that 1% of the users are robots. By reducing the robots threshold we got more users that were now classified as robots, but the confidence of whether or not those users really are robots downgrades.

The reasonable next step is to check the human and robots behavior using yet another criterion. For the next analysis we used the criterion of continuous working time, without resting of more than 10 minutes. This criterion measures the longest searching bursts of a user, but a time interval of less than 10 minutes between two queries will be ignored regarding the continuousness of the current searching burst. Again, we wanted to figure out if the users that we tagged before as human are actually acting differently from those we tagged as robots.

We can see that regarding the working time of human vs. the working time of robots, there is a distinction between their behaviors (Fig. 9). The graph we got shows that humans tend to work up to about half an hour continuously (without a break of more than 10 minutes), while robots tend to work continuously for a longer period of time, up to 4 hours. The results makes sense because it seems like a typical behavior for humans 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 robots may use the search engine for other purposes that may take longer time, like gathering information.

Note that humans were those users who sent less queries, and this may indicate that they will use the search engine for less time. But the graphs show searching bursts and not total usages. The fact that some users sent

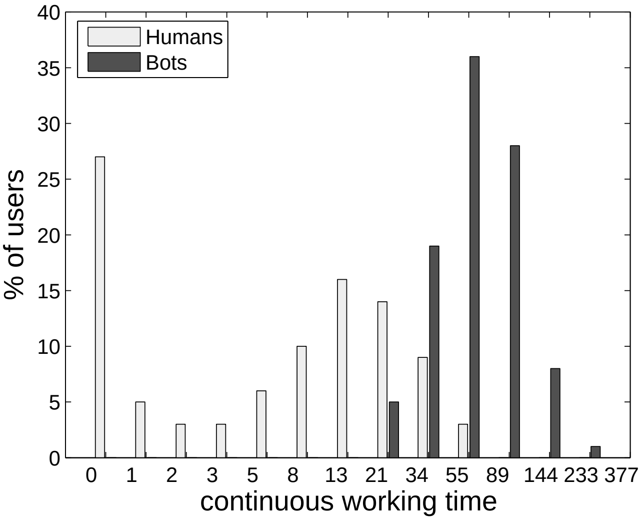


Figure 9: *Distributions of continuous working time, with less than 10 minutes breaks, of users identified as humans or robots according to their number of queries in a day.*

50 queries does not necessary mean that their searching burst time will be significantly shorter than those who sent more than 60 queries. It may have some influence, but it is possible that a user will send a query every few minutes, which will lead to a searching burst of more than an hour even with 20 queries. So the fact that the humans searching bursts tend to be shorter than the robots seems indicating of their behavior, and not just a results from the classification process. And as we said before, although it doesn't guarantee anything it does give us a good feeling about our chosen criterion and threshold.

After the last analysis we have some idea regarding the continuous working time of humans and robots. So we decided to continue analyzing the log using the continuous working time criterion with thresholds taken from the previous analysis. By observing the last graph, we can see that humans regularly works less than 30 minutes and robots works more than 30 minutes without a break. We took these observation and stretched them a little and decided to use 20 minutes and 40 minutes as thresholds for the new analysis. These thresholds mean that users that their maximum searching burst time is less than 20 minutes will be classified as humans, and those with more

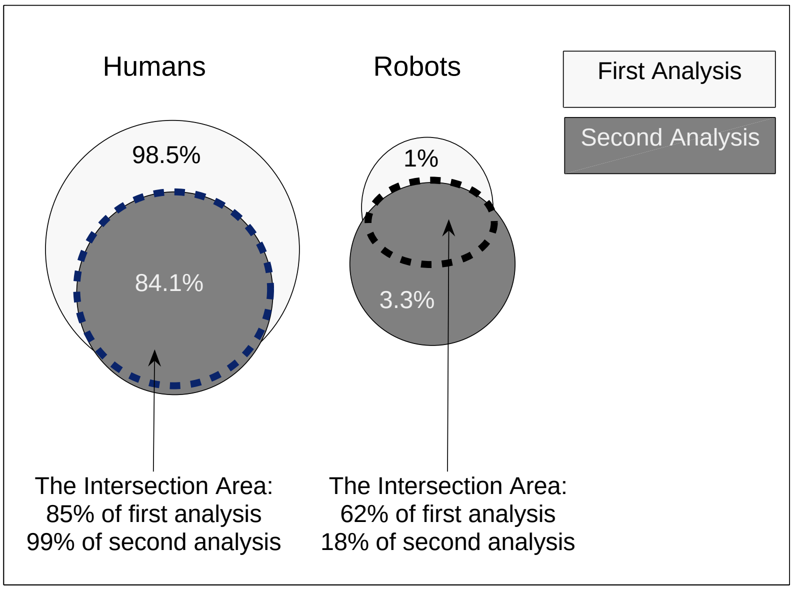


Figure 10: *Intersection of distinguishing by different analysis criterions.*

than 40 minutes will be considered robots.

After running the analysis of the continuous working time with the above thresholds we noticed that there is a significant intersection between this analysis and the previous one, using the number of queries in a day criterion. The intersection isn't surprising, given that the second analysis thresholds were taken from the first one, but it can help us to see how good is the distinction and how far we can stretch the thresholds from each other and still get good results. Almost all of the users that were tagged as human in the second analysis were also tagged as human in the first analysis, and most of the users that were tagged before as robots, were also tagged as robots now (Fig. 10).

The main difference between the analysis results is that more users were classified as robots in the second analysis. Therefore, the intersection part of the robot users have a smaller ratio regarding the number of robots in the second analysis. We may learn from this results that both of the criterion managed to recognized the humans by their behaviors, While the question about the robots behavior remains somehow unanswered.

Regarding the above conclusion, The percentages of the humans/robots groups are from the total number of users, while the percentages of the intersections areas were from the amount of classified users as humans/robots. So we should take into consideration that during the second analysis 12.5% of the users remained unclassified, compared to 0.5% in the first one. Since the percentages are from the classified users, the total amount of users that we took into consideration in the second analysis is smaller than the first one. This is why the 3.3% of the second analysis is smaller than the 1% of the first one, and therefore the robots from the second analysis that aren't in the intersection aren't necessarily humans by the first analysis.

The uncertainties of scenarios like this are an example of the disadvantages of the two thresholds methods, that cause the unclassified users. Nevertheless, we believe that this method gives us results with higher confidence than using one threshold for distinguishing.

After analyzing the log and distinguishing between humans and robots using the continuous working time criterion, with the thresholds that we took from the previous analysis, we can use the classification in order to learn about the difference between the humans and the robots behaviors in other fields. The more fields that we can recognize that there is separation between the humans and the robots behaviors, the more confident we can be about the analyzing process.

For example, we can look at the distribution of the number of re-sending the same query by humans vs. robots (Fig. 11). We found out that the humans' group and the robots' group do behave differently. Most of the users that we tagged as human tend to re-send their queries less than 6 times, while most of the users that we tagged as robots tend to re-send their queries more than that. Again we have the doubt that the behavioral differences that we have got are a result from the analysis thresholds. But it seems that the correlation is weak, because the fact that a user searches with bursts of less than 20 minutes or more than 40 minutes doesn't necessarily influence the number of query repetitions, especially when we are talking about a small number of repetitions. Moreover, The results make perfect sense because it is quite reasonable that robots will re-send the same query many times while humans tend to refine their queries if they are not get what they are looking for. These kinds of results are useful not only for the learning of the robots

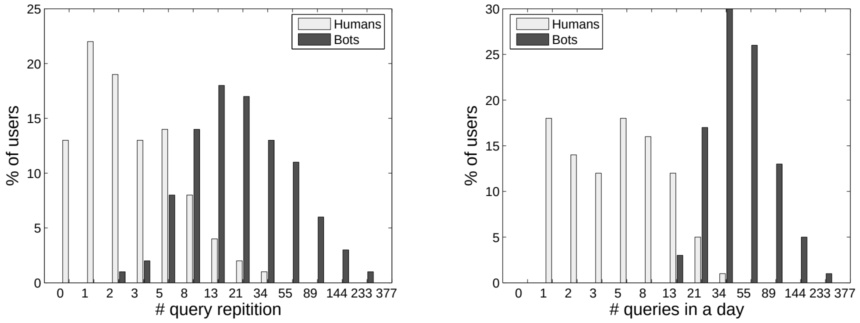


Figure 11: *Distributions of number of queries repetition and max number of queries in a day submitted by users identified as humans or robots according to their continues working time.*

and human behavior, but also to give us a hint on the correctness of the basic criterion that we used to classify the users.

We can also use the last classification in order to learn about the maximum number of queries per day of humans vs. robots. The second graph in Fig. 11 shows the maximum number of queries that were sent by the users, and we can clearly see that the two groups behave differently. As we can see, humans tend to send much less queries in a day than robots. Although the results seems reasonable, we do have some problems in this analysis. We still have almost 13% of unclassified users, and the criterion that we used in order to classify the users isn't the stronger one. Nevertheless, we can see some common-sense in the results, so we just need to search for analysis that will give us results with more confidence.

We can conclude from the above scenarios that the analysis by extreme threshold can give us a good distinction, and therefore a strong confidence that the users that were tagged as robots are actually robots. The drawback of analysis with extreme thresholds is that there may be much more unclassified users and therefore robots that we did not discover. Another drawback, as we can see from the continuous working time criterion, is that sometimes we can discover more robots but further investigation is needed in order to be sure that they really are robots.

Another analysis sample on the AOL log could be based on the time interval between different queries. We decided to classify the users using a strong criterion, such as time interval with the threshold of 60 seconds for human and 0 second robots. This means that users for whom the closest

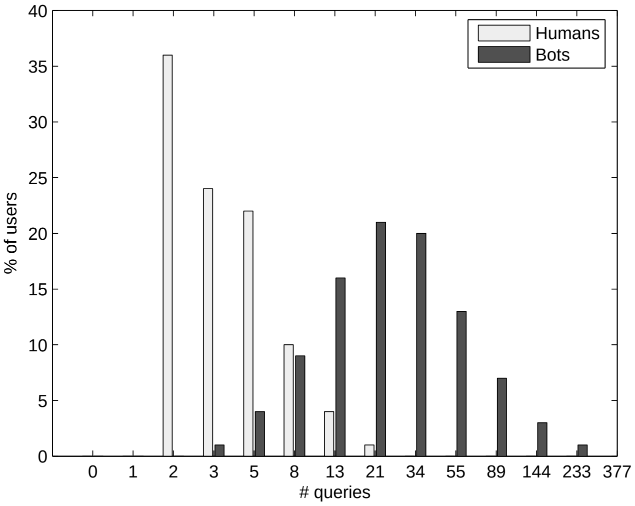


Figure 12: *Distributions of number of queries submitted by users in one day, identified as humans or robots according to their minimum time interval between different queries.*

2 different queries have at least 60 second between them will be tagged as human, and users that have 2 different queries that were sent within one second will be tagged as robots. After distinguishing between the users by the above criterion and thresholds, we used the criterion of number of queries in order to learn about the behavior of the humans and robots (Fig. 12).

As we can see in the graph, the groups of the human and robots are quite separated. Although we wish to retrieve more clear separation, it can suggest that there is some reasoning behind the chosen thresholds. Nevertheless, deeper investigation will uncover some problem with the above analysis. The problem is that the percentage of users that are unidentified is pretty high. By the last thresholds we got that 84% of the users are unidentified as neither human nor robots. This can tell us about the difficulty with choosing the right combination of thresholds and criterions.

Regarding the Correlation with time of day criterion, it seems that there is no difference between the humans and the robots behaviors. We ran an analysis over the AOL log, using the repetition criterion with the thresholds of less than 30 for humans and more than 50 for robots (Fig. 13). As we can see in the graph that we've got, the majority of the users are active

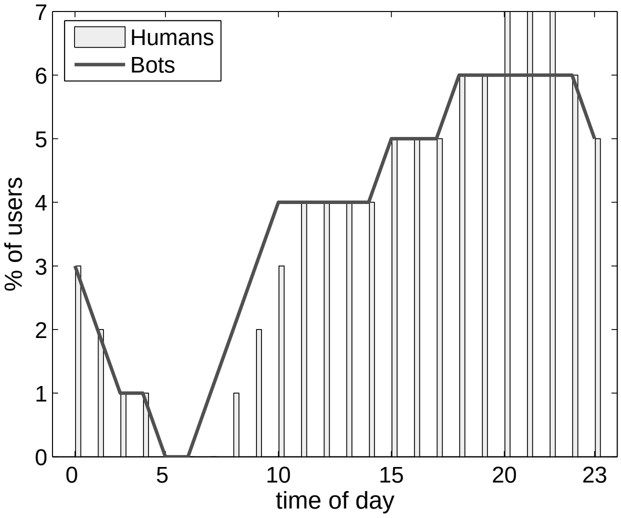


Figure 13: *Distributions of time of day of queries submitted by users, identified as humans or robots according to their maximal number of query repetition.*

during day time. The percentage of users that are active during the night is negligible and may be exceptional humans, robots, or even humans that live in another time zone than the search engine server.

5.4 Analyzing Confidence

After analyzing the log by different criterions we realized that there is a need for some way to quantify the confidence with the results. One of the main difficulties in the research is to know whether our classification is correct. In order to strength the assumption of some criterion and thresholds we used a group of users that we are pretty sure that they are robots and checked the intersection between the checked group and the robots group. By doing that, we can be more positive that the users that we tagged as robots are really robots.

As previously said, this group of robots users can be created using the strong criterions. The strong criterions are those that with certain thresholds are physically distinguishing between human and robots, meaning that humans just can not do it. For example, if we are using the criterion of number of queries per minute with a threshold of 60, which means sending a new query every second, a human just can not do it so this criterion can be considered as a strong criterion.

From all of the criterions that we mentioned before, we searched for those that with some thresholds we could use them as strong criterions. We choose the following criterions as strong one:

- Number of queries that were resent with identical time intervals.
- Min time interval between different queries.
- Number of queries in a minute.
- Number of queries in a day.
- Continues work

After choosing the strong criterions, we want to find the thresholds that for them we will be able to classify users as robots with the best confidence. Obviously we will want to maximize the number of users that were classified as robots, while maintaining the basic principle of confidence with the classification. Here are some examples of behavior that may distinguish robots in a robust way; the conversion to the right criterion and thresholds is obvious:

- More than 5 queries with the same time interval between them.
- Time interval between different queries of 0 seconds, that repeats a few times in order to reduce the chances of errors in the log time field.
- More than 20 queries in one minute.
- Day with more than 200 queries. It may seems a bit low for a strong criterion because a human can physically send 200 queries in a day, but it seems that it is pretty much for a human. For example, it may mean that the user send a query every 2.5 minutes for 8 hours continuously.
- More than 20 hours of constant work, with resting of less than one hour.

The users that were classified as robot in one of the previous analysis criterions and thresholds can create a group of users that we are pretty sure that they are all robots. We can use this group in order to check the correctness of other analysis. We can check the percentage of the “positively robots" group that are contained in the robots group from a given criterion analysis.

For example, if 100% from the "positively robots" group are included in the robots group of a given criterion, we can be pretty sure about that criterion and threshold combination. In contrast, if none of the users that in the "positively robots" group are included in the robots group of a given criterion, we can be pretty sure that the given criterion and thresholds fails in robots classifying. This method has one drawback. If 100% of the users will be classified as robots, then obviously they will contain 100% of the positively robots. So we need to use it with some thought and to make the classification of the users with some reasoning.

6 Results

6.1 Method

After having a list of criterions for the classification and the tools to analyze the quality of the classification, we searched for a criterion or a combination of criterions that with appropriate thresholds will give us reasonable results.

We started to analyze the logs using a single criterion. First we used the ones that we refer to as strong criterions, those criterions that with some thresholds will give us a reliable classification. Than, by looking at the results, we tried to learn about the human and the robots behavior and to come up with a more precise analysis using even more than one criterion for analysis.

The reasoning behind this method is that with a strong criterion we can identify the robots. Using the classification between humans and robots we can try to learn about the humans behavior. The assumption is that human behavior can be learned, in contrast to robots that might not have a common behavior and can easily imitate human behavior. Hence, while we are analyzing the results from the strong criterion classification there might be robots that were classified as humans, but we believe that they don't have influence on the statistics. There are several reasons why we believe that robots that were classified as human don't necessarily have bad influence on the results:

- There is a massive amount of human users in the logs. If a few robots were falsely classified as humans, it won't effect the global human behavior analysis.

- In case that some robot does behave like a human in some criterion, in some cases we won't even mind that it will be classified as one. Sometimes we are interested in the human behavior and not necessarily in the human users.

When using a few criterions in one analysis there is an option to select the strong criterions. The difference between the strong and regular criterions is illustrated in Analyzing Principles chapter.

Another method that we used was giving grades to the analysis. The grading process was based upon the behaviors of the users that were classified as human vs. the users that were classified as robots. When we looked at all of the graphs which displayed the behaviors of the users that were classified by some analysis, we tend to give more credit to an analysis that creates a better separation, between the humans' bars and the robots' bars, in its graphs. This grading was a way to automate our intuition. A better grade means a better separation, hence a better classification process. Basically, the grade indicates the percentage of intersection, and therefore a smaller grade is better. But, in the next chapters, in order to make the comparison of a few grades more intuitive, we will present the one's complement of the grade, so the bigger the better.

The grading method consists of a few values. We have a few graphs that we believe represent the success of an analysis. The process consists of giving grades to a few chosen graphs and then calculating the average of these graphs' grades. In order to retrieve a reasonable and representative grade for the analysis, we discarded un-useful graphs, those that always have very little separation between the humans and robots. For example, the min-time criterion graph. Because we saw that in the terms of min-time interval between different queries robots don't behave differently from humans, we didn't take this criterion into consideration while calculating the grades. It doesn't mean that the min-time criterion doesn't useful for the analysis. It's possible that with the right thresholds and the right mixing with other criterions, it will give some added value. But as a measurement for the success of other criterion, it's not good. The same works also for other criterions that we didn't use for our grading process.

The only place that we have fond a use for the average number of queries in a day is in the grading process. Since the average tend to be low, there is a significant overlapping between the humans bars and the robots bars. Because of the overlapping and the low scattering of the values, it's impossible to find thresholds for it. Nevertheless, we have found that when we're using other criterion for the analysis, we may find some differences between the humans and the robots average query counter. So we decided to use it in the grading process. The measurements we choose to use for the grading process are:

- Maximum number of queries in a day.
- Maximum number of queries in a Minute.
- Average Number of queries in a Day
- Maximum periodic Repetition.
- Maximum continuously working time, with breaks of 10 minutes.

The grade of each graph is the average of two values, the percentage of the intersection from the human bars and the percentage of the intersection from the robots bars. For example, in the left graph of Fig. 14, the intersection between the humans and the robots bars is 10% of the humans and 50% of the robots, which gives us an average of 30%.

In addition to the above, we also ignored small overlaps that may give us unjustly bad grade. As demonstrated in the second graph of Fig. 14, although the average intersection is about 53%, we can identify robots pretty well, and this is our goal. It seems that the question is when to ignore this kind of intersection.

The size of the bar that we decide to ignore depends on the other bar side. First, we will ignore any intersection that one of the bars in it is less than 1%. Since we are presenting the data in logarithmic scale, most of the bars are expected to have a significant value and bars of less than 1% probably indicate on anomaly and should be ignored. Moreover, if we have one bar that is larger than 40%, it is reasonable to ignore the intersection even if the other bar is 3%. In order to fix the above problem we decided to ignore all intersections in case that the smallest bar is smaller than $\lfloor$*(larger bar size) / 10*$\rfloor$. This is a rule of thumb that came from reasonable thinking and not from some analyzing results.

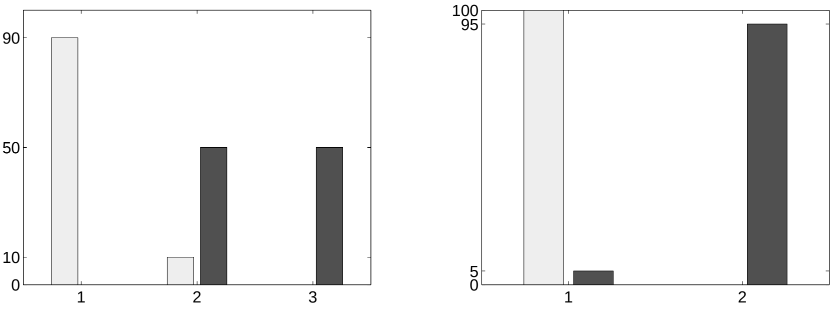


Figure 14: *Possible distributions between humans and robots.*

Summing up, we say that we have an intersection when the humans' bars overlapped the robots' bars. In order to quantify the results we have calculated the percentage of the intersection from the robots bars and from the human bars. The reasoning behind this is the assumption that a smaller intersection between the human bars and the robots bars means better separation between the user's behaviors. A better separation means that our analysis is better, because if there is a distinction between the humans' bars and the robots bars it means that these two groups actually behave differently, and if these two groups behave differently in many criterions than it means that the analysis was good. So our grading system just checks the graphs that correspond to some analysis and give it a grade based on the amount of intersection between the humans' bars and the robots' bars. Because none of the criterions is perfect, we aspire for a good separation in as many graphs as possible.

In order to find the best graded analysis, we created a small script that run over a list of reasonable thresholds and gave us the grades for each one. We can automatically evaluate many different combinations of thresholds. This method is supposed to accelerate and simplify the search for good thresholds. So in addition to reasonable selection of thresholds, we can use this script in order to help us find good thresholds, based on our grading method.

Another thing that needs to be taken into consideration is the percentage of unclassified users. If we will classify only the users that we are most sure that they are robots then we will probably get a very good classification and the grade will be respectively good, but we might just miss our goal in case that most of the users will be unclassified. So when compering several analyses, the selecting of the best one should be composed of the grade, which represent the intersection level, and the percentage of unclassified users.

The last part of the research will be trying to verify the results using other metrics. We found that there exists a search engine user behavior study using surveys. An example for a survey like that is the survey conducted by iProspect company [2], a search engine marketing firm. We will check if our results, that were based on our log analysis techniques, were similar to the iProspect results.

6.2 Analysis

As we said before, in order to get a grip on the difference between the human users behavior vs. the robot users behavior, we used a few strong criterions. In the Analyzing Principles chapter we saw a few analysis scenarios. In this section we will discuss the way we choose the criterions and thresholds in our research.

Although the min-time criterion doesn't give us a good separation, we believe that it's a good criterion to start with, because it's a really strong criterion. If we are using a threshold of 0 or even 1 for robots, we can be pretty sure about the type of the users that were classified as robots. The problem with this criterion is that it's sensitive to errors in the time stamps. We would like to minimize the probability that a human will be classified as a robot due to an error in the time stamp in the log. So, we decided that only users that were able to send two different queries within two seconds, at least two times, will be classified as robots.

Previously we used this criterion with the thresholds of 60 second for humans and 0 seconds for robots. When we used those thresholds, most of the users remained unclassified. The obvious solution for that is to lower the threshold of the humans. The question is to what value we can lower the humans threshold? We used our rating mechanism in order to check the analysis grades of a number of potential thresholds. For the robots threshold we used 1x2, which means sending two different queries with time difference of one second, for at least two times. The human threshold was the variable we wanted to find.

Since the grades are one's complement of the percentage of the intersection, we are looking for the highest grade. By looking at table 6, we can see that as we reduces the humans threshold, the percentage of the unclassified users decrease, together with the grade. We will want to choose the threshold that will give us good grade with reasonable percentage of the unclassified

Table 6: *human's thresholds for min time interval criterion.*

human's threshold	unclassified users	differ	grade	differ
12	49%		49	
11	47%	2	48	1
10	44%	3	45	3
9	41%	3	45	0
8	37%	4	42	3
7	34%	3	39	3

users. We can see that using this criterion alone is not a good idea, but it's a start. We wanted to find thresholds that will give us the best analysis grade but with at least 50% of classified users. We can see that choosing 9 seconds as a threshold is better than 10, since while the we get less unclassified users the grade is the same. So we decided to continue and check the threshold of 9 second for humans, in the minimum time criterion.

In order to check this threshold we looked at the graph of query repetitions (Fig. 15). It seems that although the robot users do not have a unique behavior, humans do. We noticed that about 70% of the humans resend their query less than 6 times, and about 95% of the humans resend their query less than 20 times.

In order to verify our results we looked at the latest search engine user behavior study by iProspect [2]. The IProspect study is based on self-reported behavior in response to a survey. IProspect research suggests that:

- 62% of search engine users click on a search result within the first page of results.
- 90% of users click on a result within the first three pages of search results.

If we take into consideration that each search engine page consists of 10 optional results, and that users doesn't always click on all the available results before they move to the next page, our results are very much alike. Although there are probably also robots that act like humans in that manner and those effect our results of the humans behavior, the number of human users should neutralize their effect.

In the last analysis we used the minimum time interval criterion in order to distinguish between the users, and than we analyzed the users behavior

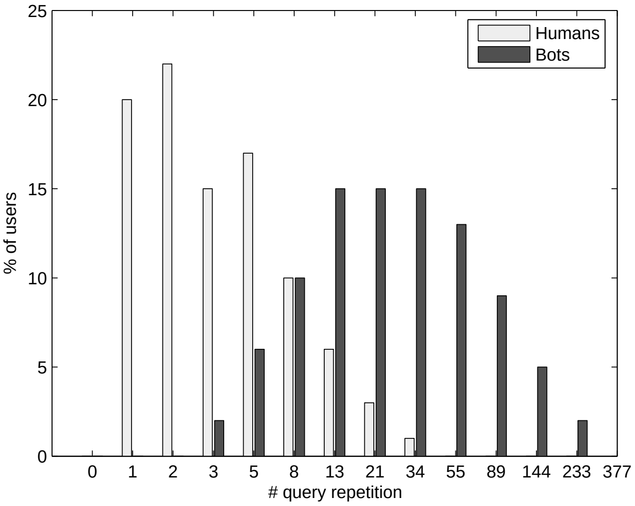


Figure 15: *Distributions of number of repetition of same query submitted by users identified as humans or robots according to their minimum time interval between different queries.*

using the query repetition criterion. Now we want to classify the users using the query repetition criterion. In order to choose thresholds we used the conclusion of the last analysis with some stretching for confidence. We decided to say that users who resend their queries less than 10 times are human, and those who resend their queries more than 30 time are robots. The total grade of that analysis was 43%. As we said before, the grade is one's complement, so it's means that we have intersection in more than half of the bars, which isn't so good.

Next we started to analyze the log by the rest of the criterions, using the previous results as a basis for the initial thresholds. We wanted to start with a strong criterion, so we started with the same time repetition, which means re-sending the same query at precise time intervals. The same time repetition criterion is a strong criterion because no humans user holds a stopper and tries to resend his queries with the same time intervals. For the thresholds we said that humans are those users that don't have any repetitions of the same query with the same time intervals, and users with at least three equivalent queries with the same time interval between them will be consider as robots. From that last analysis we moved along to the rest of the criterions. The

Table 7: *suggested thresholds, based on other criterions.*

	ST Repet'		Repetition		Min Time		#Queries,M		#Queries,D	
Thresholds	<1	>2	<10	>30	>9	<1	<5	>10	<25	>50
Grade	72		53		37		69		75	
#queries(D)	<30	>40	<21	>21	<21	>30	<30	>45	–	
#queries(M)	<5	>5	<3	>5	<3	>5	–		<5	>6
repetition	<14	>25	–		<10	>20	<13	>20	<14	>20
ST repetition	–		<1	>1	<1	>1	<1	>1	<1	>1

thresholds were based upon the results of previous analysis, with a confidence coefficient. This is why the thresholds seems to be stricter than needed. Since there is almost no separation in the min-time criterion, we didn't add it to the table. In this kind of criterions we suggest to give thresholds according to the humans capability and not by their behaviors. Our analysis and results are presented in table 7. For each criterion in the table we have the following:

- The thresholds that we used for each criterion. We wrote the humans threshold on the left and than the robots threshold.
- The grade for that analysis, according to the grading method we discussed before.
- An estimation for a good thresholds for the humans and for the robots. The thresholds that we picked doesn't necessarily represent the extreme values of the users behaviors, but the values that we believe that may be a good thresholds for others analysis.
- There is no estimation for the humans behavior in the criterion that the users were classified by, because it's biased from the analysis itself.

By looking at the table we can see that there is a behavioral pattern for the human users. Although this pattern is not compelling to all of the human users, it is something to work with. We also used all the above analyses and ran a mix analysis. As explained before, in the mix analysis users will be marked as humans only if they were classified as human by one criterion and weren't classified as robot by any other criterion. The same works also for robots. After running the mix analysis, we have got that 95.5% of the users are humans, 0.5% are robots, and 4% of the users remained unclassified. The grade of that analysis was 88, which is pretty good, specially relative to what

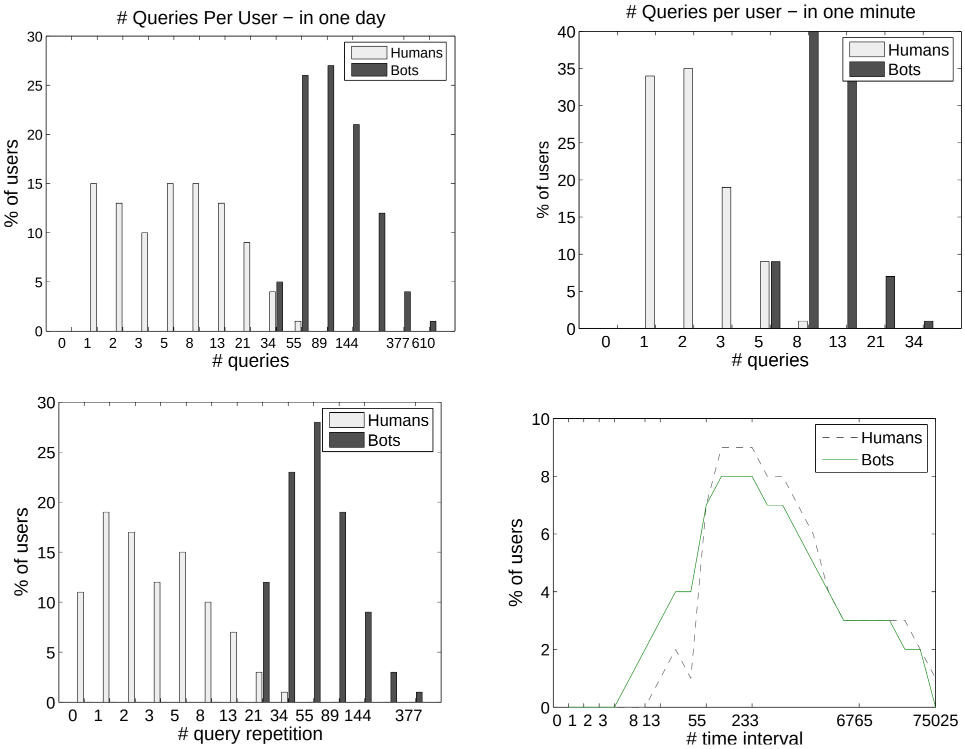


Figure 16: *Several distributions of users, identified as humans or robots according to the mix criterion analysis.*

we've got so far. In the next set of graph (Fig. 16) we can see the results from the mix analysis of the classification made by the criterions in table 7.. We can see a clear distinction between the humans and the robots behaviors.

When we enforced the robots' classification of the strong criterions, while using the mix criterion, we got lower grades. This means that users that we did classify as robots using one strong criterion, may behave like humans in other criterion. We didn't find thresholds that classify the same users as robots in all of the criterions. Basically we can either trust every strong criterion and classify the users as robots based on it, or we can decide to "trust" those classification only if the chosen users doesn't behave like humans in any other criterion. There isn't an absolute answer to that question and it's based on the user preferences. If we would like to be more confidence, we won't enforced the robots' classification if they contradict other criterions classification, in the price of more unclassified users. It depends how "strong"

Table 8: *influence of strong criterions.*

	AOL06			AlltheWeb01			AltaVista02		
Strong Criterions	Robo	UC	G'	Robo	UC	G'	Robo	UC	G'
None	0.30%	6.20%	89	0.02%	2.19%	95	0.09%	2.28%	92
#Queries, Day	1.54%	4.96%	75	1.97%	0.24%	71	1.32%	1.05%	65
#Queries, D & M	1.74%	4.76%	72	2.02%	0.19%	71	1.42%	0.95%	64
Repetition	3.40%	3.10%	51	0.38%	1.83%	75	0.46%	1.91%	68
All	6.50%	0%	47	2.21%	0%	69	2.37%	0%	56

is the criterion and how much one trust it. If we would like to classify as many users as possible, we will have to trust all of the classifications Using the above analysis.

After analyzing the AOL log, we decided to check our hypothesis on other logs as well. We decided to check the same thresholds we used on the AOL on the AlltheWeb 2001 and AltaVista 2002 logs. The results of the mix criterion with different strong criterion are given in table 8. The table contains the percentage of the robots, unclassified users, and grade for each combination of strong criterions. Because of the strong criterion definition, the decision of which criterion is strong influence only the sizes of the robots group and the unclassified users group.

We can see the influence of choosing some criterions as strong criterion, and tagging users as robots even in case of contradiction. The difference between the logs' analysis results may be explained by the changes during the years between the logs, in the search behavior or in the amount or capabilities of bots in the web.

Another criterion, that we seldom use, is the words counter criterion. The problem with that criterion is the massive intersection between the humans and the robots bars. As as matter of fact, it seems that in this field there is no difference between the humans and the robots behaviors (Fig. 17). The fact that we can't retrieve a good classification makes it hard to use it as a criterion. Although we can't use the words counter as a criterion, we can find some benefit from it. Regarding the number of words used by search engine users in the last years, there is unanimity that the average number of word phrases in search engine is about 2 words [3, 1]. The fact that our conclusion in that manner is suitable with the past studies is important. Even so we can't use the words counter as a criterion, it can give us indirectly confidence regarding our research.

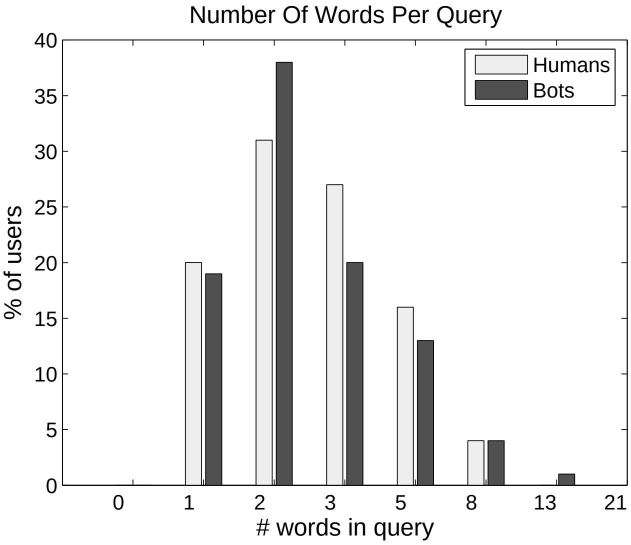


Figure 17: *Distributions of number of words per query, identified as humans or robots according to the mix criterion analysis.*

6.3 Conclusion

Our goal in that research was to investigate the distinction between human users and robots. In the past, Jansen and Spink have advocated using a threshold of 100 queries in a day to make such distinctions: users who submit more than these threshold will be classified as robots and won't be used in the analysis of humans user behavior [6, 8]. This threshold seems to be a reasonable choice if we want to be sure regarding the classified robots. But in order to get more accurate results and to be more confident about them, further work needs to be done. After all, it is unlikely that there is a single value that distinguish between humans and robots. Moreover, the humans behavior can change over time, so a fixed value won't work. We searched for a more flexible method, and decided to use two thresholds that can be evaluated during research upon a given log file.

We have investigated using two thresholds, one for humans and one for robots. In the case of the criterion of number of queries in a day, the first threshold indicates the maximum number of queries a user can send in order to be classified as a human. Any user that sent, in any given day, less queries than that threshold will be classify as human. The second threshold indicates the minimum number of queries a user need to send, in some day, in order to be classified as a robot. Users that fall between those thresholds will remain unclassified. We have two contradicting goals. On one hand, we want to separate these two thresholds as much as possible, but on the other hand we want to minimize the unclassified users. So we are willing to "absorb" a small number of unclassified users, in case we get a better classification.

In addition to the criterion of number of queries per day, we found a few more criterions that we can use in order to distinguish between humans and robots. We can refer to each criterion as a dimension, so what we have is a multiple dimension space with many points inside it. Each point in this space represent a user, and we seek a cut that will separate the points into two groups, the human users and the robots. The more we investigated, the more we realized that such a cut does not exist.

The basic problem is that there no absolute defined behavior for the robots. There is nothing to prevent the robots from behaving like humans. So even if we would be able to define some characteristic of humans behavior, it is possible that we will falsely classify some robots as humans. In case a robot completely behaves like a humans, in some cases we just might not care if it's a robot. For example, if we want to analyze the humans behavior just for research purposes, maybe we won't mind that there are some robots that act like humans, because they don't have a bad influence on the results. But there is still the problem that robots may imitate human behaviors in some criterions but not in others.

This is why we developed the "mix" criterion, as described before. By choosing the criterion we wish to mix we have the power to control the number of dimensions in the next analysis. By doing so, we reduced the problem of finding a cut in the given multi-dimensional space into finding a cut in a smaller multi-dimensional space.

We figured that it's not realistic to find a combination of all of the criterions in order to receive classification that will give us a relatively small percentage of unclassified user and a good separation between humans and robots. Instead, we believe that it is best to find thresholds for each criterion and to look for groups of criterions that may distinguish the users in the best way. The search for the thresholds for each criterion is made by the following steps:

- searching for a strong criterion, with no correlation to the one we investigate.

- use that criterion to classify the users into two groups, humans or robots.
- looking at the behavior of the users that were classified as humans.

The reason that we start with “strong” criterions, as we defined before, is that we can easily choose reasonable thresholds for them. In the strong criterion we can pick a threshold that will classify robots with high confidence, and it’s a start. The more repetition of the above steps the more data we will have on the humans behavior in the wanted field. By doing so, we may get a perspective on the humans behaviors in order to come with a range of reasonable thresholds for the criterions. When we will have a grip on the possible range for the thresholds, we will be able to use the grading method in order to find the one that seems most reasonable.

As we can see in the results chapter (Table 7), we looked over the humans behaviors according to other analysis and tried to come up with the best thresholds for the classification. What we found out is that the combinations of those analyses may give us a better results, in the means of the grades, while lowering the percentage of the classified users. Since our grades only represent the separation level between the humans bars and robots bar, one should decide the level of trust in the strength of the criterions that been used, in order to decide whether a given grade is good or bad.

One should take into consideration that our research was based on general-purpose search engines. In case of analyzing specialty search engines, such as Answers.com in the answers searching field or the MedicineNet.com int the medical search field, it may be preferable to use other thresholds and even other criterions.

Although we expect that there will be some similarity between the preferred thresholds of different logs from similar search engines, it may help to do some adjustment when analyzing different logs from different periods. We believe that there is no such thing as a perfect value for the thresholds. The criterions and thresholds should be chosen according to the research purpose and given logs.

References

[1] "*hitwise.com — search intelligence for online advertising and search marketing*". URL http://www.hitwise.com.

[2] "*iprospect.com — search engine user behavior study*". URL http://www.iprospect.com.

[3] "*onestat.com — website statistics and website metrics*". URL http://www.onestat.com.

[4] "*Marketingpilgrim — internet marketing news*". URL http://www.marketingpilgrim.com, Feb 2009.

[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] O. Etzioni, "*Moving up the information food chain: Deploying softbots on the world wide web*". *AI Magazine* **18(2)**, pp. 11–18, Summer 1997.

[8] 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.

[9] 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.

[10] S. Mitterhofer, C. Kruegel, E. Kirda, and C. Platzer, "*Server-side bot detection in massively multiplayer online games*". *IEEE Security and Privacy* **7(3)**, pp. 29–36, May 2009.

[11] G. S, X. M, and W. Z. andWang H, "*Measurement and classification of humans and bots in internet chat*". *Proceedings of the 17th conference on Security symposium* pp. 155–169, Sep 2008.

[12] 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.

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

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

[15] 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.