Cleaning and Analyzing Internet Search Logs

Final Research Report to the Israel Internet Association

PI: Dror Feitelson
School of Computer Science and Engineering
The Hebrew University, 91904 Jerusalem, Israel

Web search logs are an important tool for studying and understanding user search behavior, which is important for the evaluation of existing systems and the design of new approaches. Our research included two projects:

- Distinguishing human users from bots, in order to enable more focused and accurate characterization of the behavior of each, and
- A characterization of long-term human serach behavior.

Data Logs

When writing the proposal for this research we had 8 search logs available to use. Most were short (a single day), but one, from AOL, was 3 months long with over 20 million queries. During the beginning of the work we also acquired access to the MSN resarch log, which contains data about a whole month of activity and nearly 15 million queries.

In the end we decided to focus mainly on the AOL log from 2006. Other logs are either too short or do not seem to add any significant information. The MSN 2006 log was especially disapointing. The reason we had to refrain from using it is that user privacy considerations had caused Microsoft to partition long user sessions into multiple short sessions that cannot be identified as belonging to the same source [2]. This prevents the use of the log for user studies.

Distinguishing Humans from Bots

The question of correctly classifying humans and bots looks like a classic machine learning problem. However, there are two difficulties. First, we have no labeled training data. Second, the two groups are not necessarily well-separated. We therefore try to perform the classification in the following way:

1. Define a set of criteria that may be used to distinguish humans from bots. For example, in the past, Jansen and Spink have advocated using a threshold of 100 queries to make such distinctions: users who submit more are classified as bots and not used in the analysis of (human) user behavior [9, 7].

2. For each criterion, define not one but *two* thresholds. For example, we may suggest that humans are those users that submit less than 50 queries, bots are those that submit more than 100, and leave the range between 50 and 100 undefined.
3. Combine the results of different criteria by using one criterion to create a classification, and then checking its effect on other criteria. This enables an iterative search for good threshold values that lead to good separation.
4. Use extreme cases (e.g. a user who submits many thousands of queries) as obvious bots, and use them to override conflicting classifications based on different criteria. These are called "strong" criteria.

We have identified the following criteria for possibly distinguishing humans from bots.

- Maximum number of queries in a day. Note that we turn an absolute number ("users who submit 100 queries") into a rate ("100 queries in one day") to allow this to be applied to logs of different lengths. A threshold of 200 was set for strong identification of bots, as passing this threshold would mean submitting a query every 2.5 minutes on average for 8 hours running.
- Maximum number of queries per minute. This extends the above but looks at the maximal momentary rate rather than at the daily average, which is expected to be higher for bots [6]. A threshold of 15 queries in one minute was set for strong identification of bots.
- Minimum time interval between different queries. Note that we require the queries to be different, to avoid possible effects from double logging of the same query. Bots were identified by an interval of 0, i.e. two queries in the same second.
- Number of repetitions of the same query. While human users may repeat a query several times [11], very large numbers of repetitons most probably indicate a bot.
- If repetitions of a query come at precise intervals, this bolsters our belief that they are generated by a bot. A threshold of more than 7 repetitions with the same time intervals was set for strong identification of bots.
- Humans need to rest and sleep, while bots do not. Therefore extended continuous work is indicative of bots [6].

Some additional criteria were also checked, but turned out to be less useful.

Using these criteria we created a log-analysis utility. The utility's user interface allows one to define thresholds on a criterion, and displays the resulting classification and the distributions of other criteria. An example of the user interface is shown in Fig. 1. This corresponds to the above example, where humans are identified as those users who submit up to 50 queries a day, and bots as those that submit more than 100. The analysis shows that using these thresholds 98.6% of the users are identified as human, and only 0.3% as bots; the remaining 1.1% are left unclassified.

Fig. 2 shows the effect of such a classification on the distribution of another criterion, in this case the number of times the same query is repeated. The discrimination appears to be pretty good,

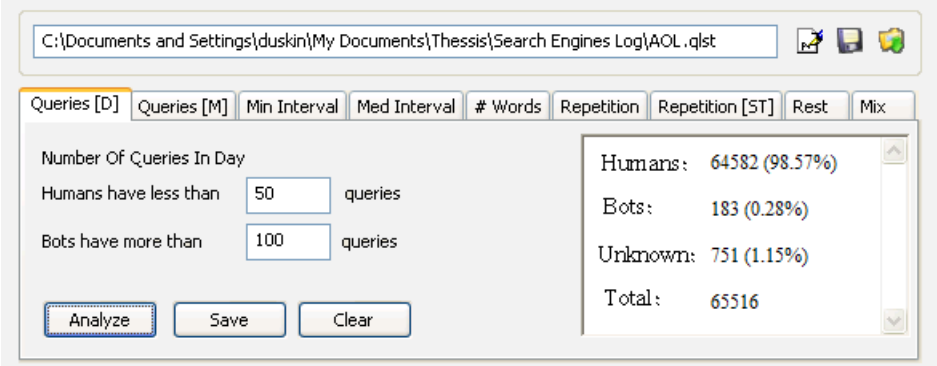

Figure 1: *Part of the user interface of the log analysis utility.*

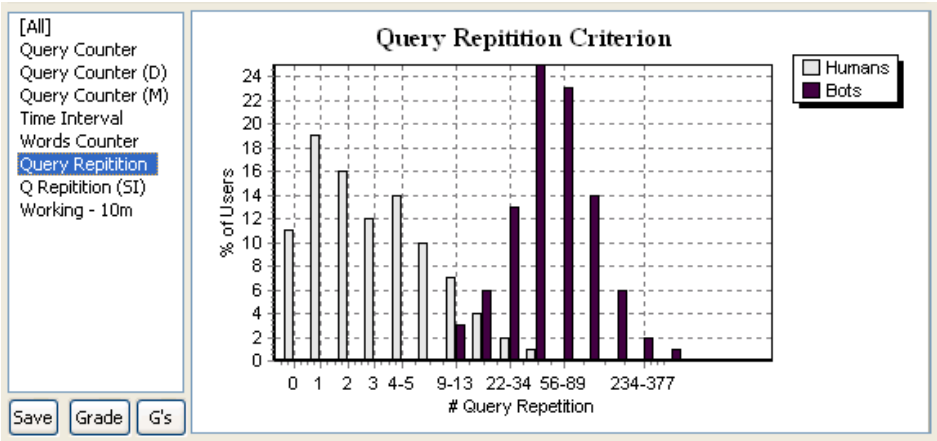

Figure 2: *Effect of classification on the distribution of another criterion.*

with users identified as human repeating queries up to about 10 times, while those identified as bots repeat queries dozens or hundreds of times. By quantifying the quality of this separation, we created an automated iterative process that systematically searched for good threshold values.

Once we have good thresholds for each criterion, the evidence is combined as illustrated in Fig. 3. Users who are identified as human by some criterion and are not identified as a bot by any other criterion are classified as human. The same (in reverse) is done for bots. But if the user is identified as a bot according to a strong criterion, this trumps any possible conflicting identification as a human.

The result of applying this methodology is the identification of 92.4% of the users as human, and 0.58% as bots. Of the bots, about half are in the consensus, and did not have conflicting classifications as humans. The rest were identified based on some strong criterion. This means that the humans tend to display relatively consistent behavior, whereas bots may exhibit markedly different behaviors. In particular, it is not uncommon for a bot to be very different from typical human behavior according to one criterion, while being indistinguishable from a human according to another.

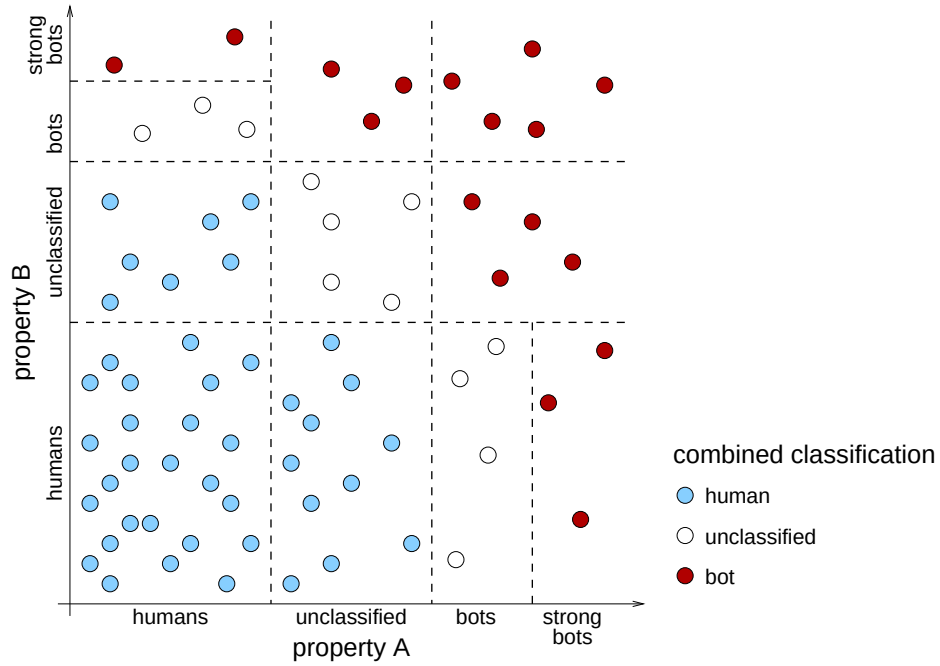

Figure 3: *Illustrative example of classification based on mixing two properties.*

Long-Term Human Search Behavior

Based on the identification of human users, we set out to derive a new long-term human search model (similar to the work of [10]). In particular, our goal is to model the "average" user, including the arrival process (user sessions) and the patterns of submitting queries (length of queries, how many are submitted, and how they are modified). This work is in advanced stages but not complete yet. We expect it to be finished towards the end of the calendar year (Dec 2010).

Our first results pertain to the popularity of different queries. We found that popular queies can be classified into two easily distinguishable groups: those that are popular all the time, and those that are only popular for a short time. The first are overwhelmingly navigational queries. The latter are related to news stories or sports events.

The question of distribution of query types (informations, navigational, and transactional) has received mixed results in the past [1, 5]. We developed a technique to automatically identify navigational queries based on either of two criteria:

- The query has a URL structure, e.g. it begins with "www." and/or ends with ".com".
- The query is a single term, and this term appears as a basic component of the clicked URL. For example, the query may be "bank of america", and the user then clicked on www.bankofamerica.com.

Using this we show that the fraction of navigational queries drops with popularity: they are nearly 80% of the most popular queries, but drop to less than 10% of the less popular ones.

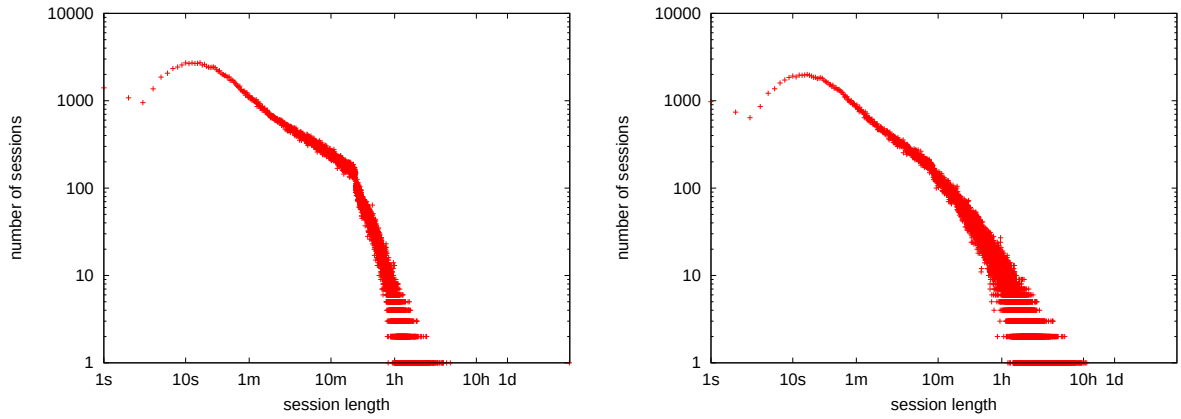

Figure 4: *Distributions of sessions lengths obtained by using a global threshold of 20 minutes (left), or by using per-user thresholds (right).*

Another issue that has been investigated in the past is that or user sessions. Sessions have been defined in different ways, sometimes focusing on sequences of queries with only short intervals between them, and sometimes emphasizing the shared query terms [4]. We reconcile these approaches by definning "sessions" to be sequences that are done in one sitting, and "quests" to be sequences that are part of satisfying the same informationn need. There can be multiple quests in a single session, or alternatively, a single quest may extend across multiple sessions.

The common way to identify sessions is to use a threshold of around 30 minutes [3, 8]. Then a break in the sequence of queries that is longer than 30 minutes signifies the start of a new session. We noticed that using such an approach leads to a distribution of session lengths with a noticeable break at the value of the selected threshold, which indicates an artifact (left of Fig. 4). To correct this, we devised an algorithm that adjusts the threshold for each user. Applying this leads to a continuous distribution of session lengths, without any obvious artifacts (right of figure).

Quests are easier to identify — we do so by comparing the query terms of successive queries by the same user. The part of the research that is not completed yet concerns the relationship between sessions and quests. Specifically, we want to characterize the lengths of quests, and the ways in which users re-formulate their queries during a quest.

Preliminary results are shown in Fig. 5. Each such graph corresponds to the activity of a single user. Both axes represent the serial numbers of queries executed by this user during the duration of the log. The grey scale represents the degree of similarity between queries. Thus the diagonal just shows that each query is identical to itself, and off-diagional elements indicate the repetition of a query or at least a similar query (if it is grey instead of black).

One thing that is obvious from these plots is that successive queries are often independent of each other, but sometimes we see a block of successive queris that are highly similar. These blocks represent non-trivial quests, where the query was reformulated until the desired information was found. By tabulating the sizes of such blocks, we can quantify the frequency and number of such reformulations. The thin vertical lines between queries represent session breaks as identified by the

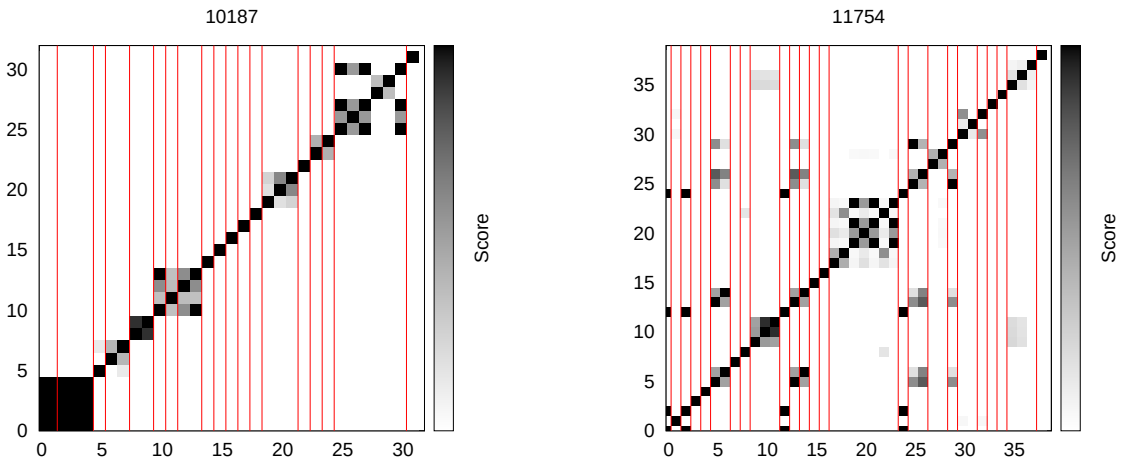

Figure 5: *Representation of the relationship between queries and sessions.*

per-user thresholds. As may be expected, such breaks tend to occur between independent queries and between quests. However, we find that occasionaly they occur within a quest, indicating that the quest persisted in a non-continuous manner over a long time.

References

[1] A. Broder, "*A taxonomy of web search*". *SIGIR Forum* **36(2)**, Fall 2002.

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

[3] D. Downey, S. Dumais, and E. Horvitz, "*Models of searching and browsing: Languages, studies, and applications*". In 20th *Intl. Joint Conf. Artificial Intelligence*, pp. 1465–1472, Jan 2007.

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

[5] B. J. Jansen, D. L. Booth, and A. Spink, "*Determining the informational, navigational, and transactional intent of web queries*". *Inf. Process. & Management* **44(3)**, pp. 1251–1266, May 2008.

[6] 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.

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

[8] B. J. Jansen, A. Spink, C. Blakely, and S. Koshman, "*Defining a session on web search engines*". *J. Am. Soc. Inf. Sci. & Tech.* **58(6)**, pp. 862–871, Apr 2007.

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

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

[11] 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.

Publications

So far our work has led to the following:

1. O. Duskin and D. G. Feitelson, "Distinguishing humans from bots in web search logs: preliminary results using query rates and intervals". In *Workshop on Web Search Click Data*, pp. 15-19, Feb 2009. DOI http://doi.acm.org/10.1145/1507509.1507512.
2. O. Duskin, M.Sc. thesis, Dec 2009.
3. O. Duskin and D. G. Feitelson, "Distinguishing humans from bots in web search logs" (submitted).
4. Listing of user IDs in the AOL search log that are believed to be non-human or bots. URL http://www.cs.huji.ac.il/~feit/papers/RoboAOL/
5. David Mehrzadi, M.Sc. thesis (in preparation).