# Sliding Window Technique for the Web Log Analysis

Nikolai Buzikashvili
Institute of System Analysis, Russian Academy of Science
9, Prospect 60-Let Oktyabrya, Moscow,
117312 Russia
+7 495 135-5357
buzik@cs.isa.ru

## ABSTRACT

The results of the Web query log analysis may be significantly shifted depending on the fraction of agents (non-human clients), which are not excluded from the log. To detect and exclude agents the Web log studies use threshold values for a number of requests submitted by a client during the observation period. However, different studies use different observation periods, and a threshold assigned to one period is usually incomparable with the threshold assigned to the other period. We propose the uniform method equally working on the different observation periods. The method bases on the sliding window technique: a threshold is assigned to the sliding window rather than to the whole observation period. Besides, we determine the sub-optimal values of the parameters of the method: a window size and a threshold and recommend 5-7 unique queries as an upper bound of the threshold for 1-hour sliding window.

## Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval –*search process.*

## General Terms

Measurement, Experimentation, Human Factors.

## Keywords

Web log analysis, agent, client discriminator.

## 1. INTRODUCTION

The aim of the Web log researchers is to extract a human searcher behavior from the Web search engines logs. An original log of the Web search engine contains not only individual users transactions and not all logged users are humans. As a result, a researcher should exclude: (1) a client who is a "mixture" of individual users (a local area network) since a sequence of transactions of this client is an interlace of transactions of different users; (2) a client which is an agent rather than a human.

The reliable method to exclude local networks is drawing only clients accepted cookies. However, some of "cookied" clients are programs (agents) rather than humans. Agents frequently "assume a mask of a human" and accept cookies. To detect agents the Web analysis [2] uses the rule: if a number of requests submitted by a client is greater than a certain threshold the client is detected as an agent and is excluded. The criterion of the agent detection is only probabilistically reliable.

We refer to a pair *(threshold, period)* as a *client discriminator*. If a number of requests submitted by a client during *period* is greater than *threshold* a client is excluded from the further analysis. A *threshold* is measured either in transactions or in unique queries.

The most studies use a whole observation period as *period*. But different log samples are drawn during different periods: for example, 12 days (*AltaVista* [2]), 7 days (*Yandex* [1]) 24 and 8 hours (e.g., *Excite*). As a result, we cannot say what condition is more stringent: 50 transactions per 8-hour period or 100 transactions per a day; we have no ground to compare the results yielded under different discriminators. Another topic is threshold validity: a threshold is assigned arbitrary and a usual goal of this setting is coverage of humans rather a trade-off between leaved out humans and biases caused by agents accepted as humans.

We propose a uniform method of agent detection, which creates equal conditions regardless the observation period. This preprocessing method was used previously (e.g. [1]) but it was not described and investigated. Here, we describe two variants of the technique and investigate the sub-optimal values of the both parameters of the client discriminator. Next, a reasonable upper bound of the threshold parameter is very important since an influence of non-rejected agents increases with the threshold and may significantly bias the results of the log analysis. We have no means to recognize humans or agents and cannot detect this value directly. However, we can do it indirectly.

## 2. SLIDING WINDOW TECHNIQUE

The uniform agents detection is based on the *sliding temporal window* technique. We select a sliding window size *T* (smaller than the observation period), assign a certain threshold *N* to this window and slide the window over a time series of the client transactions comparing a number of client requests covered by the window with *N*. If in some position of the window the number of requests covered by the window is bigger than *N* we exclude the client as an agent. The sliding technique for a threshold measured in unique queries slightly differs from the sliding technique for a threshold measured in transactions. Fig. 1 shows how a sliding window moves over the same time series of client transactions in the case of the threshold measured in unique queries (Fig. 1A) and in the case of the threshold measured in transactions (Fig. 1B).

We refer to the transaction retrieved *p*-th page of the results as *p*-page transaction. While 0-page transactions may be either a query submissions or returns to the 0 page of the results, a non-0-page transaction may not be a submission.

**A. Threshold Measured in Unique Queries**. The technique for a threshold measured in unique queries uses only 0-page transactions. At first we put the *left* margin of the sliding window

*WWW 2007*, May 8-12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

on the first 0-page transaction. At each next step of the procedure, we move the left margin to the next 0-page transaction and recalculate a number of unique queries covered by the window. If this number is greater than a threshold we reject a client.

**B. Threshold Measured in Transactions**. The procedure for is different: we move the *right* margin of the sliding window to the next uncovered transaction and account a number of transactions leaving the window.

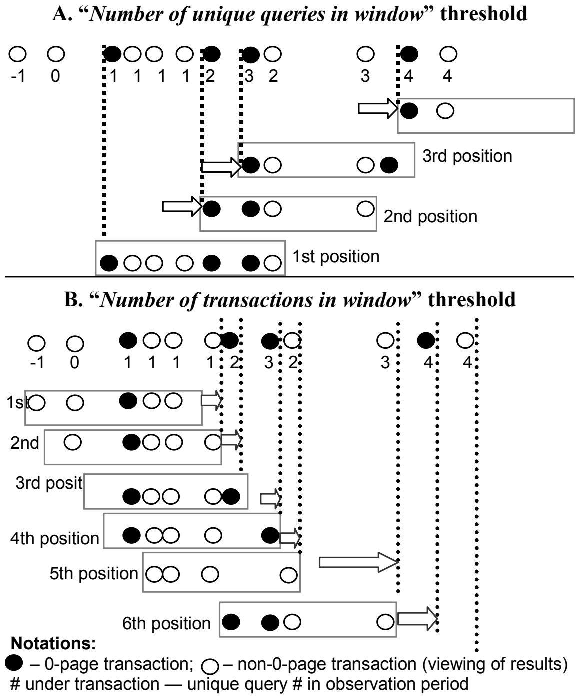
A. "*Number of unique queries in window*" threshold
-1 0 1 1 1 1 2 3 2 3 4 4
3rd position
2nd position
1st position
B. "*Number of transactions in window*" threshold
-1 0 1 1 1 1 2 3 2 3 4 4
1st
2nd
3rd posit
4th position
5th position
6th position
Notations:
● – 0-page transaction; ○ – non-0-page transaction (viewing of results)
# under transaction — unique query # in observation period

**Figure 1. A series of client transactions during observation period and two types of the sliding window movements.**

## 3. SUB-OPTIMAL PARAMETERS

We can manipulate both parameters of the client discriminator: a window size and a threshold. What values of the parameters are good enough? We consider how the sliding window technique works on the *Excite* 2001 log sample (24 hours, 305,000 clients). We consider only a threshold measured in unique queries and use different combinations of both parameters: (*a*) from 15 min to 24 h as a window size and (*b*) several grades of the threshold.

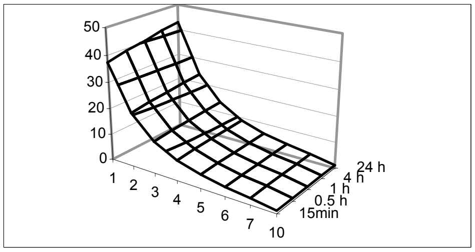

**Figure 2. Rates of clients excluded as agents depending on combination of sliding window size and a threshold.**

### 3.1 Window Size

Fig.2 shows rate of the excluded clients as a function of both discriminator parameters: a sliding window size and a threshold assigned to the window. While the client discriminator with the same threshold and the bigger window size rejects a larger fraction of clients, the distributions of excluded clients are very similar among different window sizes. Furthermore, if a client discriminator *A* excludes more clients than a client discriminator *B*, the clients excluded by *B* are also excluded by *A*. As a result, we can select any convenient size of the window lesser than usual observation periods. For example, [1] use 1-hour window.

### 3.2 Threshold Upper Bound

There is no threshold which separates humans and agents: any reasonable upper bound for the human requests is bigger than the lower bound for agent requests. Since humans submitted more requests than the threshold are ignored we are interested to increase the threshold. Since an influence of non-rejected agents should be not too big we are interested to decrease the threshold. Clients are not marked as humans and agents and we cannot determine a trade-off value of the threshold directly. Fortunately, agents behave similarly and they "like" to use query language syntax. To determine the upper bound of the threshold we can investigate how syntax-based metrics (e.g. fractions of queries containing Boolean or quotation operators) behave as a function of the threshold value for *both* classes of clients (*recognized* as humans or agents). Starting with a certain threshold a contribution of the true agents became visible and is evident in the changes of the metrics as a function of the threshold.

We study behavior of several metrics calculated for clients accepted as humans and rejected as agents by 1-h sliding window. While the most metrics monotonically increase for both classes of clients, Fig. 3 shows a prompting behavior of the fraction of unique queries containing *AND* operator. The bumps on both graphs may be explained by the influence of agent fraction and the corresponding threshold value (5-7 unique queries) should be considered as the threshold *upper bound* for the 1-h sliding window. This bound (a) very slow increases over the window size (e.g. 6-8 queries is a bound for 4-h window) and (b) is less surprising than 100-transaction threshold per 24 h observations frequently used in the Web log analysis.

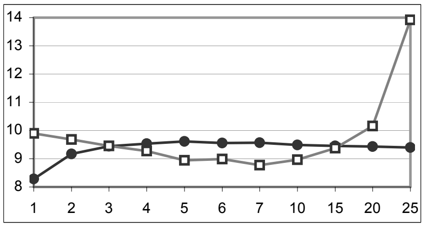

**Figure 3. Fraction of *AND* -queries as a function of threshold for client accepted as humans (●) and rejected as agents (□).**

## 4. CONCLUSIONS

We have proposed a unified method of agents' detection in the Web logs, which does not depend on the observation period. We have discovered that (1) a window size plays no significant part; (2) the upper bound of the threshold equals to 5-7 unique queries if a threshold is measured in queries per 1-hour sliding window.

## 5. REFERENCES

[1] Buzikashvili, N. Comparing Web Logs: Sensitivity Analysis and Two Types of Cross-Analysis. AIRS'06 (Singapore, 2006), LNCS 4182, Springer, 2006, 508–513.

[2] Silverstein, C., Henzinger, M., Marais, H., Moricz, M.: Analysis of a very large web search engine query log, SIGIR Forum, 33 (1), ACM, 1999, 6–12.