Limits of the Web Log Analysis Artifacts

Nikolai (Nick) Buzikashvili
Human-Computer Interaction Laboratory, Institute of System Analysis, Russian Academy of Science
9, prospect 60-Letiya Oktyabrya, Moscow, 117312 Russia
buzik@cs.isa.ru

Bernard J. Jansen
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA, 16801, USA
jjansen@ist.psu.edu

ABSTRACT
In this technical paper, we estimate instrumental limits of an unexacting comparison of results reported in different Web log studies. We consider sensitivity of results of log analysis to 4 controllable factors: a log sampling technique, an observation period and two cut-off variables peculiar to the Web log analysis (a LAN cut-off to exclude local area networks clients and a temporal cut-off to detect temporal search sessions). It is shown that 3 of these factors may lead to multiple differences in the case of marginal values of the factors whereas an effect of usual factors combinations is limited by 30%. These limits overcover differences between the results of Excite and AltaVista studies.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Storage and Retrieval.

General Terms
Measurement, Experimentation, Human Factors.

Keywords
Controllable variables, sensitivity, Web transaction log analysis.

1. INTRODUCTION
Numerous studies of different search engines logs were conducted to date. Some of parameters reported in these studies have the same name (session length, terms per query, etc.). This is a sufficient reason to compare results of these studies (c.f. [3]) to discover differences of interaction with different engines. Table 1 shows significant differences of results reported in the *Fireball* [2], *AltaVista* [4], *Excite* [5], *Fast* [5] and *Yandex* [1] studies.

Table 1. Results of the Web transaction log studies

	***Fireball* 1998**	***Excite* 1999**	***Excite* 2001**	***Fast* 2001**	***Yandex* 2005**
terms/query	1.7	2.4	2.6	2.3	2.9
AND	2.4%	3%	10% all Bool	—	0.3%
PLUS	24.8%	2%		—	0.5%
quotations	8.6%	5%	9%	—	2.6%

Differences of results may follow from (1) features of interfaces, query languages and search techniques used in different search engines, (2) differences in contexts, including the a recommended usage manner, the overlap of query language and real-life written language, (3) hypothetic cultural differences, and (4) *differences between log analysis methods used in different studies*.

Contrary to first 3 factors the instrumental factor (4) is not only out of a rigorous investigation but also was not mentioned so far.

If this factor is *really* significant, then to compare different logs results we should exclude it and conduct a *cross-analysis*, in which the same tool under the same conditions analyzes different logs. However, the first step is to estimate an influence of the instrumental factor. If this factor may explain only small (less than *reported* differences) differences we may confine ourselves to comparison of published results. On the contrary, if method varying explains big differences (bigger than *reported* ones, ~30%), then (a) the role of this factor should be taken into account in a comparison of *reported* results, and (b) this is another sound reason to conduct a cross-analysis.

In the *Web* Transaction Log Analysis (Web TLA), the obvious sources of results differences are *controllable variables* arbitrary assigned by researchers. These are 1) *a temporal session cut-off* and 2) *local area network detection cut-off* measured in unique queries or in transactions. Sensitivity to these variables is investigated in Chapter 8. We also consider results dependency on 3) a *sampling technique* (Chapter 6) and 4) an *observation period duration* (Chapter 7).

2. WEB TLA PALLIATIVES
Web clients instead classic users. The individual *user* detection was not typically a problem in the study of classic information retrieval (IR) systems, but it is a significant issue with Web HTTP-based IR. In contrast to classic transaction logs, Web transaction logs record *clients*' transactions rather than individual *users*' ones. The Web client is either a single PC or Local Area Network (LAN). In contrast to users of a single PC, users of LAN may work *simultaneously* and a sequence of transactions of the same LAN client is a temporal mix of transactions of different users. Users of a single PC client work *sequentially* (one by one), so their queries may not alternate each other. Since Web logs provide no opportunity to recognize different users of LAN we need to exclude all LAN clients. The only criterion of the LAN client detection is a number of unique queries (or transactions) submitted by this client during some period. However, this is only probabilistically reliable way to detect LAN.

Sessions. In classic transaction logs, we see clear sequence of *<login>, <a search session (i.e. a time series of transactions)>, <logout>*. HTTP-based Web transactions do not support such sequences, and researchers usually use a temporal session notion.

Palliative cut-off variables. Web TLA uses 2 palliative cut-off values: 1) *a number of submitted queries cut-off* to exclude LANs and 2) a *temporal cut-off* to cut a time series of client transactions into a sequence of temporal sessions. Web researchers by the default consider and use these variables as objects of arbitrary setting rather than subjects for discussion. An effect of these variables was not measured and investigated.

3. TERMS USED

1. We distinguish *unique queries* submitted by a client during some time interval or during all interaction with engine, *submitted queries*, and *transactions*. We distinguish *queries unique* per a temporal session, *submitted queries* and *transactions*. A transaction may be either initially submitted query or paging of retrieved results (in the latter a page number field in the logged transaction is not 0). It should be noted that any transaction is frequently referred to as a query. We avoid this manner.

Let's present user's transactions with a single Search Engine (SE) in a form of transactions sequence $\{(q, p)\}$, where q is a query string and p is a retrieved page of results (0 corresponds to the first page). First and third transactions in the observation period (Fig. 1) are *head-disrupted* (i.e. non-0 page transactions which are not preceded during the observation period by 0-page transaction containing the same query string). First transaction (3,2) after the observation period presents a *tail-disrupted* query 3 (0-page transaction of this query is inside the observation period while the rest of transactions is performed after the period). We can detect and exclude head-disrupted transactions but we can't detect tail-disrupted transactions.

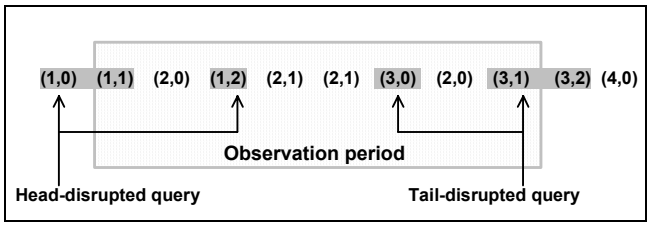

Figure 1. Sequence of logged transactions.

Next, inside the observation period, we see two 0-page transactions of query 2. It is very probable that second 0-page transaction corresponds to a new query submission rather than a return to 0 page viewing. Thus, we can account *two submissions* of query 2. However, we account only one *unique* query 2 in the observation period.

Finally, inside the period, we account only 5 transactions and only 2 unique queries per period. Sometimes researchers don't exclude head-disrupted transactions. In this case we should account 3 unique queries and 7 transactions. The shorter an observation period the more sensitive results to the method of head-disrupted queries accounting and to unavoidable tail-disrupted queries.

2. In this paper, we consider temporal sessions and task sessions:

A *temporal session* as any sequence of the single user's transactions with the same search engine cut from previous and successive sessions by some cut-off interval;

A *task session* is all transactions of a temporal session aimed at the same task. We used the following method of automatic task session detection. First, a relation of orthographic similarity is automatically constructed in the pairwise comparison of all unique queries belonging to a single temporal sessions. Next, a transitive closure of this relation is built up. All queries belonging to the same connected component are considered as a single task session.

4. DATA

We use logs of 2 search engines: the Russian-language *Yandex* (7 days, 2005, 175,000 clients), and the *Excite*: 1999 log full fragment (8 hours, 537,639 clients) and 2001 log sample (24 hours, 305,000 clients). The *Excite* data are not actual, less detailed and were often used previously. However, the previous usage of these data is an obvious advantage because the *Excite* data are *real benchmark* datasets. We use the Excite data to yield the results comparable with the results previously yielded *on the same logs* in the *Excite* project. In this paper, we report the results of the analysis elaborated on the *Excite* data. The conclusions followed from the *Yandex* log analysis are the same.

Contrary to the 2001 sample, the 1999 dataset is a whole 8-hour *fragment* of the *Excite* log. We use the *Excite*-1999 data to investigate sensitivity of Web TLA results to a sampling technique and to observation period.

5. SLIDING WINDOW TECHNIQUE

Observation periods of different samples are different. For example, the *Excite-1999, Excite-2001* and the *Yandex* samples used in our study cover correspondingly 8 hours, 24 hours and a week. Let we assign 10 unique queries as the LAN cut-off for a client series from the *8-hour Excite-99* sample. What compatible LAN cut-offs should be assigned to the *day* and *week* samples?

Instead attempts to solve this unsolvable problem, we use the *sliding temporal window* technique. Namely, if duration of the shortest observation period is equal to T hours and N unique queries LAN cut-off is assigned for this sample, we use the same *T hours temporal window* to slide over other samples and compare a number of client queries in this window with N.

In this study, we use various LAN cut-off values measured in *unique queries* submitted under *1-hour sliding window*.

6. EFFECT OF SAMPLING

The last summer the author analyzed the *Yandex* sample [1]. Since it was a first experience of the Russian search engine team in the log provision I tried to get an aswer: if the data are sampled what sampling was used and what fraction was sampled. However, our multistep interaction was a megilla finished by the answer "the data were somehow sampled because full data (~40 Gb) are too hudge". Whereas only ~3% of full data were sampled the sampling technique was nothing for gays who prepared the data.

A rigid rule of the log sampling is: all transactions of sampled clients should be sampled over the observation period. However, methods of clients sampling vary. The simplest method is to move through a raw transaction log and mark each new (or every n-th new) client. Transactions in a raw log are time-ordered, so all marked clients are time-ordered by first occurrences. When the number of marked clients is sufficient, one selects all transactions of these clients over a observation period grouping transactions of the same client. This is "sampling by first occurrence" method.

Unfortunately, this convenient sampling is not random. Let all single users are described by the same probabilistic distribution of submissions over time. Let a number of clients including k users (call him k-client) is $n(k)$. When we sample first client by the "sampling by first" method a probability to sample k-client equals to $kn(k)/\sum kn(k)$ rather than $n(k)/\sum n(k)$, i.e. is proportional to k. While following steps decrease a bias, a fraction of sampled

multi-user clients, in particular LANs is overestimated, so a fraction of clients submitted more queries is too big. For example, a sample of even 50,000 clients significantly differs from the whole *Excite-99* dataset (Fig. 2) in fractions of first transactions.

As seen from Fig. 3, even results elaborated on 50,000-client sample (9%, cf. only 3-percent sample of the *Yandex* log) enormously deviate from the whole log results. *The "by first" sampling changes all quantitative characteristics up to 10%.*

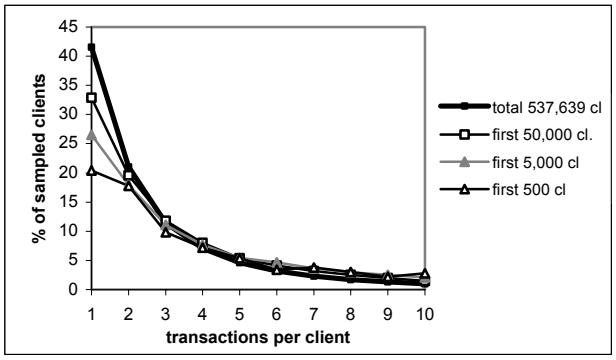

Figure 2. "By first" sampling from the 8-hour Excite-99 log.

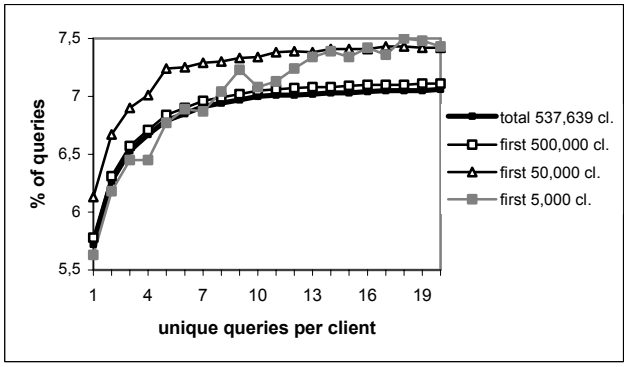

Figure 3. Cumulative fractions of quotation queries in 3 first-clients samples and on the whole Excite-1999 log.

7. EFFECT OF OBSERVATION PERIOD

Different previous Web TLA studies [1, 2, 4, 5] used different observations periods: 12 days (*AltaVista*), 8 and 24 hours (*Excite*) and 7 days (*Yandex*). The shorter an observation period the more sensitive results to the method of head-disrupted queries accounting and to unavoidable tail-disrupted queries. In particular, a period shortening increases (overestimates) fractions of clients submitted smaller number of transactions. The effect of time-partitioning is opposite to the effect of sampling "by first".

We cut the raw *Excite*-1999 log into 1-hour portions: from 9am to 10am and so on. Contrary to "by first" sampling, partitioning increases only 1-transaction fraction. So a contribution of 1-query differences to differences between the whole log results and results yielded on 1-hour fragment are the most significant, and a plot of any characteristic elaborated on 1-hour fragments is similar to a vertically shifted plot of the whole log. This shift is set by the difference in 1-query point. For example, Fig. 4 shows fractions of queries including quotations for the whole 1999 log and for its 1-hour fragments (cf. Fig. 3).

Variations of the observation period lead up to 10% change of results only for short periods (e.g., 1-2 hours). Thus, an observation period should be greater than 2 h. However, real log samples used in the previous studies meet this requirement and *we can exclude the observation period from factors affecting results*.

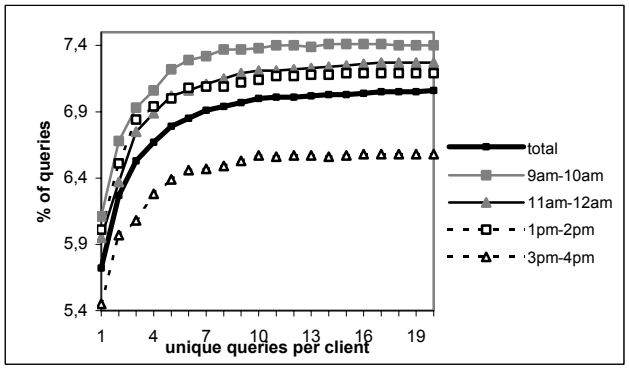

Figure 4. Cumulative fractions of quotation queries in four 1-hour fragments and on the whole Excite-99 log.

8. EFFECT OF 2 CUT-OFF VARIABLES

The Web TLA radically differs from classic TLA by usage of 2 controllable cut-off variables (a *LAN cut-off* measured in unique queries or in transactions submitted by a client during some period and a *temporal cut-off* measured in time units).

Considering differences induced by a temporal cut-off we should take into account that most frequently used (but rarely reported) values of temporal cut-off are 15 min and 30 min. So differences between characteristics estimated for these cut-off values are especially significant. At the same time, LAN cut-off variable has no such nearly "standard" values. First, it may be assigned by either a number of transactions or a number of unique queries submitted during different observation periods. Next, nobody use "standard" value for certain unitary period (e.g., a hour or 24 hours) to calculate comparable LAN cut-off value for another period. So we should consider all intervals of "reasonable values".

"*Per transaction*" characteristics (a query length or fractions of some kind of queries, e.g., Boolean, advanced queries, etc) *don't depend* on a temporal cut-off.

The same characteristics considered "*per unique query*" also *don't depend* on this cut-off when we consider unique queries *per client* but these characteristics depend when "unique queries *per temporal or task session*" are considered. For example, when a user submits 9-term query Q_1 and 1-term query Q_2 during 2 temporal sessions, a query length per client is 5. If first temporal session includes both queries while another includes only Q_2, a query length per temporal session is 3 while it equals to 5 if a user submits both queries in both sessions.

The influence of temporal cut-off is predictable: the greater cut-off, the longer temporal session and the greater such metrics of a temporal session as a number of unique queries, transactions or task sessions, etc. They should increase and do increase. However, several temporal-dependent characteristics (e.g. a number of unique queries per task session) are less predictable. For example, a number of unique queries per *task* session slightly increase for sequential sessions, and it is stable for parallel tasks.

At the same time all characteristics depend on a LAN cut-off.

As seen from Table 2, the influence of a temporal cut-off is small (*less than 10%* difference between 15 min and 30 min temporal cut-off values). Furthermore, increasing of a temporal cut-off, e.g. to 2 h gives no significant results. *Thus, to compare search behavior metrics corresponding to different logs we can relay on reported results of studies, which differ only by temporal cut-offs*.

Table 2. Characteristics for 3 temporal cut-offs and 4 LAN cut-offs measure in unique queries per 1-hour sliding window (*Excite* 2001 sample)

Temporal cut-off	**15 min**				**30 min**				**60 min**			
LAN cut-off	**3**	**5**	**10**	**20**	**3**	**5**	**10**	**20**	**3**	**5**	**10**	**20**
K temporal sessions	300	341	374	384	285	320	347	354	275	306	329	334
transactions/ temp.session	2.14	2.42	2.70	2.83	2.26	2.58	2.92	3.06	2.34	2.70	3.07	3.24
uniq.queries/ temp.session	1.25	1.40	1.56	1.63	1.32	1.50	1.68	1.76	1.37	1.56	1.77	1.86
task sess-s / temp.session	1.11	1.16	1.21	1.23	1.14	1.19	1.25	1.27	1.15	1.22	1.29	1.31
uniq.queries/ task session	1.12	1.21	1.29	1.33	1.16	1.26	1.35	1.38	1.18	1.28	1.38	1.42

We use a number of unique queries per a sliding window as a measure for LAN client detection. As seen from Table 2, Web TLA results depend on LAN client cut-off variable. However, changes are very significant only if one of the results corresponds to a *small* cut-off value. For big LAN cut-off values results change insignificantly. Fig. 5 shows significant dependency of fractions of special queries on LAN cut-off. A fraction of *AND* queries initially increases over a number of unique queries and than decreases whereas fractions of queries which use *PLUS* or quotations operators monotonically increase.

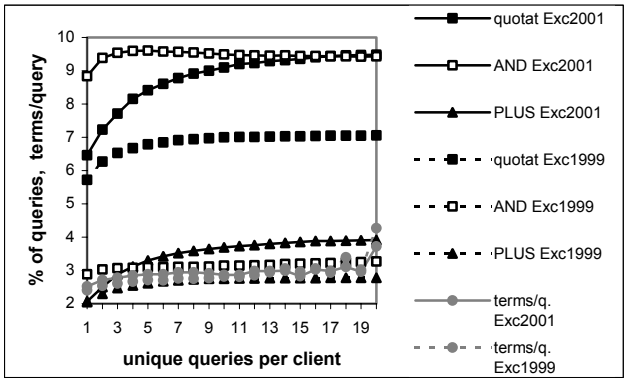

Figure 5. Cumulative fractions (%) of special queries and terms per query in the Excite logs as functions of LAN cut-off.

9. CONCLUSIONS

In this paper, we try to answer: are frequent 20-30-percent differences between the results reported by different web log studies real differences? Or should we consider these *statistically significant* differences as possible *artifacts* induced by the methods of these studies. We considered sensitivity of results of log analysis to 4 controllable factors: a log sampling technique, an observation period and two cut-off variables used in the Web log analysis. The observation period was excluded as insignificant. Other 3 factors may lead to multiple differences in the case of *marginal* values and combinations of factors (e.g. one of studies uses *poor sampling* or the *LAN cut-off* not greater than *1 unique query per 1-hour sliding window*) whereas an effect of usual factors combinations is *limited by 30%*. Thus, comparing the results of different Web studies, we have no reasons to consider 30-percent difference as significant. In particular, method variations overcover differences between the *Excite*-1999, *AlataVista* and *FAST* logs which so far have been commonly considered as significant. At the same time, the differences between *Excite*-1999 and *Excite*-2001 are too big and may not be explained by controllable variables varying.

From the sensitivity analysis point of view, it is very interesting to compare the results elaborated on the benchmark *Excite* data and previously elaborated in the *Excite* project *on the same logs*. While the results of re-evaluation are similar to results of the *Excite project*, this similarity is a similarity within the limits induced by cut-off controlled variables and is far from the match.

However, more interesting question is: *what combinations of the factors correspond to the results of the previous Web log studies?* Let's consider the *Excite* log characteristics shown in Table 1. Since the "per query" characteristics don't depend on a temporal cut-off, we should find out only the LAN cut-off values. The 3–5 unique queries during 1-hour sliding interval seem to be likely values. Since "per unique query" metrics give estimations surprisingly similar to estimations of "per transaction" metrics used in the *Excite* project, we can use the results of our "per unique query" measurements to estimate the LAN cut-off corresponding to the results of the *Excite* project. Fig. 5 gives a very unexpected answer: the *one* (!) *unique query LAN cut-off* leads to the results close to the results of the original *Excite* studies, i.e. we come across that very case which we considered as 'marginal'.

10. ACKNOWLEDGMENTS

The authors thanks Ian Ruthven, Inna Fëdorova and Natalia Ponomarëva for comments.

11. REFERENCES

[1] Buzikashvili N. The Yandex study: First findings. *Internet-mathematics.* Yandex (2005), 95-120.

[2] Holscher C., Strube G. Web search behavior of internet experts and newbies. International Journal of Computer and Telecommunications Networking, 33 (1-6) (2000), 337-346.

[3] Jansen B.J., Spink A. How are we searching the World Wide Web? An analysis of nine search engine transaction logs, *Inf. Processing & Management*, 42(1) (2006), 248-263.

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

[5] Spink A., Jansen B.J. Web search: Public Searching on the Web, ICKM 6, Springer (2004), 199 pp.