Towards Query Log-Based Sociology of Search

Nikolai Buzikashvili
Institute of System Analysis, Russian Academy of Sciences, Moscow, Russia

Abstract

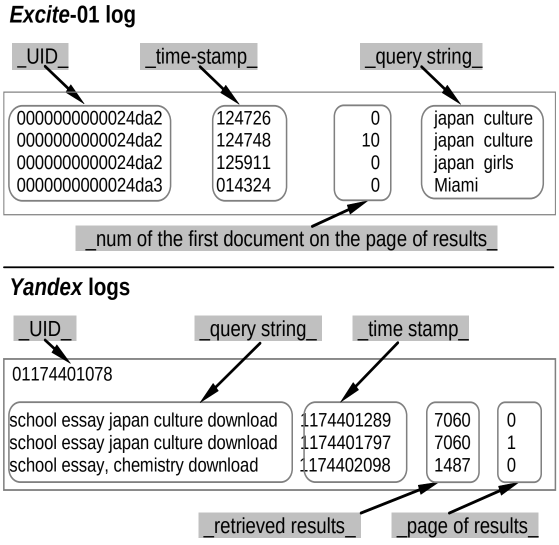
Query logs of Web search engines are a narrow huge source for a full-valued sociological analysis. The paper describes a common sequence of steps and tools of the query log-based sociological study. As an example, search images of Japan in the queries of Russian and U.S Web searchers are investigated. Two logs of the Russian search engine and one log of the U.S. search engine are automatically analyzed to detect and classify users submitted queries referring to Japan. User classes, co-relations between the classes and dynamics in both national audiences are studied.

1. Introduction

Among three questions considered by the researchers of the Web search, “Who searches the Web?” (subjects), “What do they search for?” (objects) and “How do they search?” (search tactics) the first two questions primarily relate to the applied sociology and should be formulated and answered consistently. The Web era has opened not only a new field of social activity but also a huge source of the very special data for a sociological analysis of public interests. The ways to interpret a huge collection of the very small units (queries) are obviously limited. Logs as such give no possibility to reveal either the attitudes or the origins of interests (except when a query is the result of another query). The query logs are a huge, representative, straight (vouched by submitted queries, cf. declarations in polls), long-term but narrow base for causal inferences.

While sociology of the Web mainly answers “Who searches the Web?” (age, gender, etc.) and uses polls, the query log-based sociology answers “What do they search for?” and uses query logs. The common subject of the Web log based sociology is a classification of queries by the topics searched, e.g., “sex”, “commerce” [1], [2], [4], [5], [7], and the manual attribution of queries was used in the early studies. More sociologically sophisticated studies such as [6], [9] are rare.

Objects of query log-based sociology. Anonymized “query log” datasets are the main (and the only) base for query log sociology (another useful source is users’ personal data available only for search engines and used in their targeting advertising). The datasets are combinations of records of several logs of the search engine. The datasets contains transaction descriptions grouped by user and ordered by time for each user. All datasets include: user identifier (UID), timestamp of the transaction, page number of the retrieved results (or number of the first document on the results page) and query string (Fig. 1).



```
0000000000024da2   124726    0   japan culture
0000000000024da2   124748   10   japan culture
0000000000024da2   125911    0   japan girls
0000000000024da3   014324    0   Miami
```

```
01174401078

school essay japan culture download   1174401289   7060   0
school essay japan culture download   1174401797   7060   1
school essay, chemistry download      1174402098   1487   0
```

Figure 1. Examples of “query log” datasets used in the study

Goals of query log-based sociological study and its tools. The primary goal of any query log-based sociological study is an attribution of users/queries according to the classification used in the study. A machine learning approach based on the creation of learning samples of queries is not only too expensive but also inappropriate when the objects of interest are too rare. For example, Japan-referring queries are rare neighbors in search sessions and the majority of them is so rare that manual detection of a single query *‘mukai’* for learning sample passes into a manual detection of all *‘mukai’* queries. An alternative way is the use of initial topical vocabulary of keyword[-combination]s, its modification during a study and a usage of simple classification rules. The task and tools are similar to the task and tools of the (keyword-based) contextual advertising. The aim is a perfect unambiguous attribution of a query (a sequence of user’s queries) to classes considered.

When objects (queries, users) are attributed, the further processing is a usual sociological study of the classes' rates, interrelations and dynamics, and does not depend on the specific nature of the objects considered. The main technical problem of a query log-based analysis is an auxiliary task of supervised creating of tools for an automatic classification of queries.

The paper describes from scratch all steps and tools of a full-value query log-based sociological study exemplified by a comparative study of queries submitted by two national audiences (Russian and U.S.) searching for the topics related to the third state (Japan). The logs of the Yandex (2005, 2007) and of the Excite (2001) search engines are used. The Yandex is the main Russian search engine and the Excite was a popular U.S. search engine in the early 2000s.

We consider topic categories of queries related to Japan (e.g. Japan culture, Japan goods, etc.) to categorize users submitting these queries in terms of Japan-referring classes. It is of particular interest to study co-relation between topic classes of users, dynamics of the same audience's classes, and differences between different audiences.

There were two reasons for choosing Japan as a "perceived object". The first reason is that fractions of Japan-referring queries are big enough and approximately equal in the logs of both audiences (Fig. 2). The other reason is that the choice of a "language-exotic" object minimizes the problem of recognition and disambiguation of topical queries and extremely simplifies a supervised creation of tools and an automatic classification of queries.

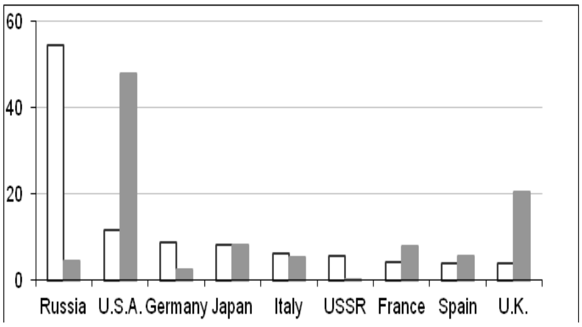


Figure 2. Rates (%) of 9 states referring to in the Yandex-07 (white) and Excite-01 (gray) logs

2. Research questions of the particular sociological study

The subject of the particular study is a *user* i.e. a set of all queries submitted by him rather than a *query*. A user who submitted queries belonging to several Japan-referring classes is attributed to all these classes (that is user classes intersect). Japan-referring classes of Russian and U.S. searchers are studied. The research subjects are:

(1) Frequency of Japan-referring classes in Russian and U.S. search audiences.

(2) Co-occurrence of Japan-referring classes in Russian and U.S. audiences.

(3) Changes of classes' frequency and co-occurrence in Russian audience over two years (Yandex-07 vs Yandex-05, a long-term intra-audience dynamics study)

(4) A comparison of fractions and co-occurrence of Japan-referring classes in Russian and U.S. audiences (Yandex-05 vs. Excite-01)

In the study, (1) and (2) are executed completely, an intra-audience comparison (3) may be considered as a fragment of a longitude study of the same audience, but an inter-audience comparison (4) of the datasets spaced far apart in time is less factually grounded.. The serious and unavoidable problem of query log-based sociology is that researchers have a very limited access to search engines' logs and available logs are rare and irregular.

However, the primary goal of the paper is to demonstrate the potentials and techniques of a query-based sociology, so the weakness of the particular dataset corpus is far from being the weakness of the method.

3. Datasets and differences between them

Three datasets are used in the study: (1) 24-hour sample of the U.S. *Excite* search engine (May 4, 2001, Friday), (2) 7-day sample of the Russian *Yandex* search engine (March 9-15, 2005, a week from Wednesday to Tuesday), (3) 24-hour complete query logs of the *Yandex* search engine (March 20, 2007, Tuesday). Three preprocessed datasets after elimination of users with broken log records and users detected as robots are described in Table 1.

Table 1. Preprocessed datasets

	***Excite*-01**	***Yandex*05**	***Yandex*07**
Sampling	sample	sample	whole set
Observat. period	day	week	day
Users in dataset	305360	176187	860618
Japan-referring users	4553	9486	28439
Rate of Japan-referring users	1.49%	5.38%	3.30%

Comparison of over-a-day *Excite*-01 and *Yandex*-07 frequencies with over-a-week *Yandex*-05 frequencies is not valid (if probability *p(day)* to submit 1+ Japan-referring queries throughout a day was equal to 0.03 for any user then probability to submit Japan-referring queries during 7 days would be *p(7days)* =

1 - (1-$p(day)$)^7 = 0.1911. But since probability to submit Japan-referring queries once again is bigger for users submitted such queries earlier, observed 7-day frequency of Japan-referring users $p_{obs}(7days)$=0.0538 is less than $p(7days)$, see rates in Table 2).

Fortunately, due to random sampling of users in the *Yandex*-05 dataset we can consider 7 one-day datasets instead of a week dataset (Table 2).

Table 2. *Yandex*-05 week dataset by days

	Users	Japan-ref.users	Rate (%) of Japan-referring users
Wed	59044	1768	2.99
Thu	59597	1840	3.09
Fri	58277	1791	3.07
Sat	36763	1149	3.09
Sun	36295	1133	3.13
Mon	61411	1898	3.12
Tue	60851	1774	2.92
Sum over:			
5 work days	299180	9071	3.03
2 off days	73058	2282	3.12
all week	372238	11353	3.05

Reported in Table 3 Yates corrected z-values (see formula (3) in Chapter 5.2) for rates of Japan-referring users in each day-to-day pair and in work-days sum vs off-days sum are smaller than critical $z_{0.95}$ = 1.96. Since no significant difference between frequences of Japan-referring users in any pair of days is observed we can joint one-day sets and use a sum of 7 one-day sets as a *one-day representation* of the *Yandex*-05 week data. As a result of this statistically correct trick, we use the virtual one-day set of 372238 virtual users instead of a set of initial real 176187 users.

Table 3. Z-values for each pair of Yandex-05 7 days and work-days vs off-days sums

	Thu	Fri	Mon	Tue	Sat	Sun
Wed	0.92	0.77	0.96	0.79	1.13	1.09
Thu		0.12	0.02	0.79	0.31	0.28
Fri			0.16	1.58	0.43	0.40
Mon				1.78	0.29	0.25
Tue					1.85	1.81
Sat						0.01

The *Yandex*'s audience is mainly Russian. The queries submitted to the *Yandex* contain words in Russian (~95%). Next, we can suppose that the majority of the *Excite*-2001 queries were submitted by U.S. searchers because (1) about 90% of queries are in English and (2) only 11% of queries are submitted during the "American night" (0am – 6am, Pacific time) when non-American users are active. Thus, the *Excite*'s audience is mainly U.S.

The *Yandex* and *Excite* logs are spaced far apart in time. When we compare two different time slices of different audiences we cannot ignore long-term trends, which are important and reflect changes of interests and the change of the available Web services as an alternative to search engines ways to get Japan-referring information. Can the comparison be valid? This is a crucial question in any comparative study. Honestly the answer should be that an ideal comparative study of parallel social processes should be study of time series rather than two time slices, even made at the same time. Fortunately, (a) the compared dimensions of the Russian and U.S. audiences seem to be low time-variable and (b) this comparison mainly demonstrates the techniques usage.

4. Steps of the topical classification

In the chapter, the universal sequence of steps of any query log-based sociological study is described. At the same time, the specific nature of the Japan-related topics significantly simplifies a tools creation (mainly non-dictionary words) and an application (mainly unambiguous topical attribution).

The basic tool of the study is a thematic Japan-referring bilingual near-parallel Russian and English vocabulary containing topically attributed keywords and keyword-combinations is named "the *vocabulary*". The words not included in the vocabulary are referred to as *non-vocabulary*. Standard Russian and English dictionaries are referred to as "*dictionaries*", and words non-included in them are referred to as *non-dictionary*.

All vocabulary creation/modification actions below are 2-stage: (a) an automatic extraction of words/combinations as candidates for inclusion in the vocabulary, (b) a manual approval of candidates and attribution of them to vocabulary topics.

4.1. Step 1. Japan-referring topics selection and initial thematic Japan-referring bilingual vocabulary creation

We identify searchers curious about Japan-referring topics by their submission of at least one Japan-referring query. To detect these queries we need to a create Japan-referring thematic vocabulary that includes (near-)parallel Russian and English words and word-combinations marked by the word topic categories.

4.1.1. Word categories. To detect and categorize Japan-referring words, queries and users we set up two kinds of categories: (1) ten *basic categories* corresponding to *both* aspects (a general reference to Japan and a certain thematic denotation, e.g., *religion, lifestyle*, etc) and (2) two *subsidiary categories*, *General* and *Geographic names,* used to detect those Japan-referring queries, which cannot be attributed to basic categories. Queries attributed to subsidiary categories should be re-categorized where possible into the basic categories in the next steps.

4.1.2. Initial Vocabulary creation.

Data. A small near-parallel bilingual Russian-English corpus of Japan-referring texts and a big Russian-language corpus of non-Japan-referring texts.

Tools for non-dictionary words extraction: Russian and English computer dictionaries

Procedure

(1) *Two-way automatic extraction of candidates*:

(1.1) *non-dictionary words and word-combinations* in near-parallel corpus of Japan-referring texts.

(1.2) *words significantly more frequent* in Russian Japan-referring texts than in Russian non-Japan-referring texts.

(2) *Manual approval of candidates, bilingual parallelization and categorization in terms of basic categories.* Almost all non-dictionary words and word-combinations detected by (1.1) and only a few too-frequent words detected by (1.2) were approved as Japan-referring words and (two-)word combinations. Extracted Russian and English spellings were parallelized and attributed to basic categories.

Besides, words from the *General* category and brand names (*Cars* and *consumerGoods* categories) were included manually.

Table 4. 12 initial intersecting categories of Japan-referring words and examples of them

Category and number of items in it		Examples
Subsidiary:		
General	17	*Japan, Japanese, Nippon, Nihon*
Geographic names	107	*Chugoku, Tokyo, Kyoto*
Basic:		
Religion & ethic	50	*satori, shinto, tsukuyomi, zen, todaiji*
traditional Arts & Theater	55	*gadaku, hokusai, koto, netsuke, origami, utamaro*
traditional Lifestyle	45	*kimono, ryokan, tatami, yakuza*
Literature	37	*haiku, kanji, mukai, renga, miyamoto musashi*
traditional Food	26	*sake, sashimi, sushi, tsukemono*
History & interstate relations	85	*edo, hojo, meiji, samurai, taisho, tokugawa, yamato*
Martial Arts	24	*aikido, budo, judo, karate, kendo, kyudo, sumo*
Masscult & movies	16	*anime, manga, pokemon*
Cars	30	*mazda, toyota*
consumer Goods	59	*marubeni, canon, nec*

Distributions of words among categories are shown in Table 4. The initial categorization allows a multi-valued word attribution, e.g. *kotatsu* belongs to both *Religion* and *Lifestyle* categories. Thus, initial word categories are overlap.

4.2. Step 2. Trial run

Initial detection of Japan-referring users and vocabulary expansion. In this step, each query is attributed to all categories of the vocabulary words contained in the query and a user is attributed to all categories of queries submitted by him.

Query processing. The vocabulary contains words and (two-)word-combinations. Word-combinations are detected (and eliminated from the further processing of this query) first, and then the vocabulary words are detected in the rest of the query. E.g., a combination “*japan culture*” (which will be added to the vocabulary only in Step 3 among other combinations) rather than “*japan*” word should be detected first in the query <*school essay japan culture download*>.

The aims of Step 2 are:

(1) *a rough evaluation of number of users attributed to each category.*

(1.1) If the number of users marked by subsidiary categories is big enough then the words from these categories should be specified by their collocations in queries and as possible be re-categorized in terms of the basic classes. E. g., such combinations as <*kyoto temples*> or <*yokohama tiers*> should be attributed to the basic classes.

(1.2) If the number of users attributed to a basic category is small then the category should be combined with another categories according to their thematic similarity. Another reason for combining categories is that we need to construct the vocabulary with *non-overlapping* word classes.

(2) *an automatic detection of frequent misprints and non-unique spellings* of the vocabulary words to include these spellings in the vocabulary. If a word in a query is non-Russian and differs from some Japan-referring vocabulary word by one or two (for more than 6-symbol long words) symbols, the word is detected as a probable misprint or non-unique spelling (e.g., *mitsubishi*, *mitsubisi* and *mičubisi* in Russian). All detected words are inspected for a mass presence in the Web and if presented they are checked manually. Approved spellings are added to the vocabulary, and *only exact match* with vocabulary words is taken due account in the *further* study.

Tables 5.1 and 5.2 show categorization of the users according to the queries submitted by them.

Table 5.1. Initial Japan-related user categories

	***Excite*01**	***Yandex*05**	***Yandex*07**
Japan-referring users	4553	11353	28439
General	820	1377	3089
Geographic names	158	120	611
Religion & ethic	35	55	118
Art & Theater	71	93	243
traditional Lifestyle	43	72	250

Literature	37	99	154
traditional Food	20	27	112
History	174	192	595
Martial Arts	76	152	300
Masscult&Movies	585	445	1117
Cars	1308	2897	10905
Goods	1365	6187	11847

Table 5.2. Rates of initial Japan-related user categories among all users

	***Excite*01**	***Yandex*05**	***Yandex*07**
Japan-referring users	1.491	3.050	3.304
General	0.269	0.370	0.359
Geographic names	0.052	0.032	0.071
Religion & ethic	0.011	0.015	0.014
Art & Theater	0.023	0.025	0.028
traditional Lifestyle	0.014	0.019	0.029
Literature	0.012	0.027	0.018
traditional Food	0.007	0.007	0.013
History	0.057	0.052	0.069
Martial Arts	0.025	0.041	0.035
Masscult&Movies	0.192	0.120	0.130
Cars	0.428	0.778	1.267
Goods	0.447	1.662	1.377

4.3. Step 3. Categories re-combination and vocabulary expansion

4.3.1. Combining categories into compound classes. Small (in terms of the users attributed to) and closely topically related basic categories are aggregated into compound classes. Two subsidiary categories are aggregated into the *General* class. Table 6 shows the resulting compound classes and the number of users attributed to them. Since user categories intersect, the number of users in the compound class is smaller than the sum of users in the combined categories. Now, the *word classes* do not intersect while the corresponding *classes of users* may intersect and do intersect.

Table 6. Classes of words and number of users attributed to corresponding user classes

Class	**Categories included into class**	***Excite*01**	***Yandex*05**	***Yandex*07**
General	General, Geographic names	955	1477	3639
Culture	Religion, Art, Lifestyle, Literature, Food	204	342	869
History	History	174	192	595
Martial Arts	Martial arts	76	152	300
Masscult	Masscult	585	445	1117
Cars	Cars	1308	2897	10905
Goods	Goods	1365	6187	11847

4.3.2. Vocabulary mutation

(1) Expansion of subsidiary words *by (non-Japan-referring) dictionary words and re-attribution of detected combinations to basic classes*. Since the number of users attributed to subsidiary categories is big, the words from these categories should be specified by their *in-query* collocations and these combinations should be re-categorized as possible in terms of the basic classes. E.g, a combination *Kyoto castles* should be attributed to the *Culture* class against to attribution of *Kyoto* to *General*.

Procedure.

(a) A non-Japan-referring dictionary word w_i more frequently collocated in queries with a *General* class word W than w_i appears in all other queries is automatically extracted as a candidate for a combination < w_i W> attributed to some basic class.

(b) The list of extracted word-combinations is checked manually and approved candidates are inserted into the appropriate word class. When < w_i W> is approved the status of W is not affected.

The results of the procedure are very successful and give 43 word-combinations such as *Japan history*, *Japan culture* (it is worthy of note that *school essay* is a very frequent co-locator of these combinations). An impressive example is an expansion of *yokohama*. This geographic name frequently appears in the *Yandex* logs. However, all 23 *yokohama* occurrences are occurrences of *yokohama tiers* combination (*Cars*).

The dramatic change in the sizes of user classes in Table 7 (cf. Table 6) is the result of the subsidiary words expansion with dictionary non-vocabulary words.

Definitions. A *temporal session* is defined as a sequence of the user's queries cut from previous and successive sessions by a certain time gap. A *task session* is technically defined as a connected component of the lexical similarity graph of queries submitted during a temporal session. When we extract task sessions we use 30-min time gap. But as a result of lexical task session detection, it is unlikely that queries containing unknown Japan-referring words will be included into the task session detected as Japan-referring. For this reason, we also use *short-gap temporal sessions* with a 5-min time gap. One can expect that Japan-referring queries are close in time.

(2) New Japan-referring words extraction. One can expect that *non-vocabulary* Japan-referring words more frequently appear in (a) queries containing words from the vocabulary, (b) task sessions containing such queries, and (c) short-gap temporal sessions containing such queries. Just as in the candidates extraction from the Japan-related texts in Step 1, we suppose that (non-vocabulary) Jana-referring words in queries are non-dictionary. As a result, all non-dictionary words are automatically extracted from Japan-referring queries and task and short-gap sessions.

The results of the procedure are far from being successful. While a lot of non-dictionary spellings are automatically detected in Japan-referring queries and task sessions, *neither of them* is Japan-referring words. Several non-dictionary words detected in short-gap temporal sessions are really Japan-referring words attributed to *Culture*, *History* and *Martial arts* classes. But these word classes expansion yields no results in the user classes since only users earlier detected as Japan-referring use these words.

5. The user classes co-relation measure and tests used

5.1. Inter-class co-relation measure

To detect closely interrelated and “incompatible” classes in the dataset we estimate the probability of a random co-occurrence for each pair of classes of Japan-referring users. The bigger the probability that an intersection of two classes is not bigger than observed intersection is the stronger inter-class co-relation is.

Let n_i be the number of users attributed to the class i (diagonal elements in “contingency table of classes”), *obs*(i,j) be the number of users attributed to both classes i and j (non-diagonal elements), and N be the number of all considered users. To measure the strength of the interrelation between two classes we use the probability $p(k \le obs\,(i, j))$ that the number of random co-occurrences k of the independent classes i and j (containing n_i and n_j users) is not bigger than the observed intersection $obs\,(i, j)$. This measure shows to what extent the observed interrelation is incompatible with the assumption of independence of the classes. The bigger $p(k \le obs\,(i, j))$, the stronger the interrelation is.

$$p(obs(i,j), n_i, n_j, N) = \sum_{k=0}^{k=obs(i,j)} p(k, n_i, n_j, N) \qquad (1)$$

where $p(k,\ n_i,\ n_j,\ N)$ is a hypergeometric probability of k co-occurrences of n_i marks of the type i and n_j marks of the type j which are independently used to mark N “cells”

$$p(k, n_i, n_j, N) = \binom{n_i}{k}\binom{N-n_i}{n_j-k} \Big/ \binom{N}{n_j} = \frac{n_i!\,n_j!\,(N-n_i)!\,(N-n_j)!}{k!\,N!\,(n_i-k)!\,(n_j-k)!\,(N+k-n_i-n_j)!} \qquad (2)$$

When classes are small the measure of co-relation is of low reliability since even a small (1-2 items) variation in the number of co-occurred items greatly changes the measure

Two variants of the co-relation measure. Besides of an obvious population of *all users* we also calculate the measure for the *only Japan-referring users* population. The reason for using the latter, “artificial” variant is that Japan-referring classes are expectedly co-related and we want to visualize the difference between the strong and the strongest co-relations. While both measure variants give the same ranking of co-relations, the former variant (“*among all users*”) clearly shows independence (and even “incompatibleness”) of classes and the latter variant (“*among only Japan-referring users*”) is more appropriate for visualization of strong co-relations.

Small probabilities in the *“among all users”* measure are markers of *incompatibility* of classes. Significant probabilities in the *“among only Japan-referring users”* measure clearly show the closest co-relation between classes.

5.2. Tests

To compare the rates of the same Japan-referring class among all users in two logs we use z-test in the Yates’ corrected form:

$$z = \frac{|\,\widehat{p}_1 - \widehat{p}_2\,| - 0.5(1/n_1 + 1/n_2)}{\sqrt{\widehat{p}(1-\widehat{p})\left(1/n_Y + 1/n_2\right)}} \qquad (3)$$

where p_1 and p_2 are sample rates for the corresponding class in each of two datasets and p is a sample rate in a combined population.

To test the similarity of the proportions of k user classes in two logs we use χ^2 *test* with (k -1) degrees of freedom. Since expected intersections of independent classes are smaller than 5 we do not use χ^2 *test* (with (k+1)k/2-1 degrees of freedom) to test for similarity of classes’ co-occurrence in two datasets.

6. Results

User classes detected under the mutated vocabulary are shown in Tables 7.1-7.3.

Table 7.1. User classes

	***Excite*-01**	***Yandex*05**	***Yandex*07**
Japan-ref. users	4553	11353	28439
General	252	205	454
Culture	497	824	2083
History	268	261	800
Martial arts	79	191	388
Masscult	789	627	1624
Cars	1333	3026	11588
Goods	1394	6426	12069

Table 7.2. User classes rates among all users (%)

	***Excite*-01**	***Yandex*05**	***Yandex*07**
Japan-ref. users	1.491	3.050	3.304
General	0.083	0.055	0.053
Culture	0.163	0.221	0.242
History	0.078	0.070	0.093
Martial arts	0.026	0.051	0.045
Masscult	0.258	0.168	0.189
Cars	0.437	0.813	1.347
Goods	0.457	1.726	1.402

Table 7.3. Z-values of user classes rates among all users (similar rates are given in bold)

	***Yandex*-07 vs *Yandex*-05**	***Excite*-01 vs *Yandex*-05**
Japan-ref. users	7.44	43.65
General	**0.50**	4.24
Culture	**2.20**	5.52
History	4.20	**1.14**
Martial arts	**1.42**	5.39
Masscult	**2.48**	7.91
Cars	27.70	19.87
Goods	13.05	51.64

Since the subsidiary *General* class before the vocabulary mutations was too big, we planned to use a *session-level* approach that decreases the rate of users attributed to the *General* class. The approach uses the "*Basics cover*" heuristic: if at least one query in a short-gap session belongs to the basic class then the whole session is attributed only to basic classes. As a result, a user may be attributed to the *General* class only if one of his sessions contains a query from the *General* class but does not contain queries from basic classes. Since the number of users attributed not only to the *General* but also to basic classes is small under the mutated vocabulary we do not need to use the heuristic.

6.1. Rates and proportions of user classes

Rates of user classes among all users. We assign the critical level of difference for Japan-referring classes among all users to 0.99 ($z_{0.99}$ = 2.58). The rates with sample z-value smaller than $z_{0.99}$ are considered as similar and are given in bold in Table 7.

Intra-audience comparison (Yandex-07 vs *Yandex*-05*). Dynamics of Japan-referring interests of the Russian audience.* There are no changes in the rates of *Culture*, *Masscult* and *Martial arts* classes. The *History* rate grows. However, the mostly impressing are the changes in the rates of "consumer classes": enormous growth of the *Cars* rate against significant decrease of the *Goods* rate during two years.

Inter-audience comparison (*Excite*-01 vs *Yandex*-05). The *History* rates are practically identical. The *Masscult* rate in the U.S. audience is significantly bigger whilst all other rates are bigger in the Russian audience.

Proportions between user classes among Japan-referring users. Sample χ^2 values for distributions of 7 classes are 1301.6 (*Excite*-01 vs *Yandex*-05), 805.9 (*Yandex*-05 vs *Yandex*-07). No similarity of proportions is observed even in the logs of the same audience.

6.2. Co-occurrence of user classes

The co-relations of classes is the most interesting part of the study. Table 8 shows co-occurrence of Japan-referring user classes. The differences between the sums of diagonal elements (sizes of classes) and sums of elements below (or above) the diagonal (sizes of pairwise intersections of classes) are only slightly less than the number of Japan-referring users. So we can ignore co-occurrence of 3+ classes.

Table 8. Co-occurrence of Japan-referring classes

***Excite*-01**							
	Gen	Cult	Hist	MArts	Mass	Cars	Goods
Gen	252	2	0	0	1	0	1
Cult		497	12	1	14	2	3
Hist			268	1	6	9	1
MArts				79	0	0	1
Mass					789	0	2
Cars						1333	12
Goods							1394
***Yandex*-05**							
	Gen	Cult	Hist	MArts	Mass	Cars	Goods
Gen	205	1	2	0	1	1	0
Cult		824	48	11	16	2	11
Hist			261	2	4	4	6
MArts				191	0	3	3
Mass					627	6	11
Cars						3026	90
Goods							6426
***Yandex*-07**							
	Gen	Cult	Hist	MArts	Mass	Cars	Goods
Gen	454	2	1	0	0	5	1
Cult		2083	121	26	43	39	26
Hist			800	5	13	32	14
MArts				388	1	8	7
Mass					1624	17	19
Cars						11588	257
Goods							12069

6.2.1. Classes Independence ("Incompatibility").

Table 9 presents the probabilities $p(k \le obs\ (i, j))$ of the classes co-occurrence in the *all users* population. They are, for the most part, too high for a random co-occurrence of independent classes. But some pairs show strong *incompatibility*. This fact cannot be interpreted as an artifact caused by the small size of classes (*Masscult* and *Cars* are big classes).

Classes' incompatibility and its dynamics in the Russian audience. Both *Yandex* logs show "incompatibility" of the *Culture* and *Goods* classes. At the same time, while big *Culture* and *Cars* classes are incompatible in 2005, they are compatible in 2007. Other two-year changes are: *Masscult* changed its co-relations with consuming classes from independent in 2005 to incompatible in 2007 and oppositely changed co-relation with the *Martial arts* from incompatibility to independence. The former may be an artifact caused by the small size of the *MartialArts* class. But changes in *Culture* and *Cars*, neither incompatibility of the *Masscult* and *Cars* and *Goods* in 2007 cannot be explained by small size of the classes but can be explained by age differences. We can suppose that *Masscult* class corresponds to young people while consumer classes, especially *Cars*, correspond to adults.

Table 9. Independence of basic user classes in the all users population (incompatible are given in bold)

***Excite*-01**

	Hist	MArts	Mass	Cars	Goods
Cult	~1	0.889	~1	0.361	0.604
Hist		0.928	0.999	~1	0.372
MArts			**0**	**0**	0.697
Mass				**0**	**0.125**
Cars					0.978

***Yandex*-05**

	Hist	MArts	Mass	Cars	Goods
Cult	1	1	1	**0.001**	**0.159**
Hist		0.992	0.999	0.835	0.703
MArts			**0**	0.796	0.358
Mass				0.599	0.481
Cars					~1

***Yandex*-07**

	Hist	MArts	Mass	Cars	Goods
Cult	1	1	1	0.971	**0.249**
Hist		~1	1	1	0.762
MArts			0.480	0.844	0.696
Mass				**0.121**	**0.185**
Cars					1

Incompatible classes in the U.S. audience. Incompatibility between *Masscut* and *Goods* is similar to the incompatibility in the Russian audience in 2007, incompatibility of *Masscut* and (small) *Martial arts* is the same as in the Russian audience in 2005. However, the most interesting is the strong *tripolar mutual repulsion* of the *Masscult*, *Martial arts* and *Cars* classes.

6.2.2. "Over-strong co-relation" (probabilities of classes' co-occurrence among *only Japan-referring* searchers, Table 10) reveals the great difference between Russian and U.S. audiences. The Russian audience shows some strong co-related classes, particularly the strong *tripolar mutual gravity* of *Culture*, *History* and *Martial arts* in 2005. On the contrary, the U.S. audience shows no strong classes' co-relations.

Table 10. Strong co-relation in Japan-referring users population (strongest are given in bold)

***Excite*-01**

	Hist	MArts	Mass	Cars	Goods
Cult	~0	~0	0	0	0
Hist		0.01	0	0	0
MArts			0	0	0
Mass				0	0
Cars					0

***Yandex*-05**

	Hist	MArts	Mass	Cars	Goods
Cult	**1**	**0.173**	0	0	0
Hist		**0.063**	~0	0	0
MArts			0	0	0
Mass				0	~0
Cars					0

***Yandex*-07**

	Hist	MArts	Mass	Cars	Goods
Cult	**1**	**0.180**	0	0	0
Hist		0.014	0	0	0
MArts			0	0	0
Mass				0	0
Cars					0

7. Conclusion

We have investigated (1) the rates of the different topical classes and inter-class relations in each of Russian and U.S. search images of Japan presented in two *Yandex* and one *Excite* logs, (2) two-year changes in the Russian search image, and (3) structural difference between Russian and U.S. search images of Japan. The findings of the particular query log-based sociological study are:

1. The fractions of all classes *among all* Russian searchers are bigger than the fractions *among all* U.S. searchers. The only exception is the *Masscult* class. At the same time the rates of the classes significantly (and sometimes paradoxically, see mixed dynamics of the "consumer classes") differ even during two years.
2. Both audiences show similar incompatibility of classes (not only *Culture* vs *goods* but also *Masscult* vs *goods*). The U.S. audience shows not only binary but also tripolar incompatibility of classes (mutually incompatible *Masscult*, *Martial arts* and *Cars* classes).
3. The audiences are the opposites in "strong classes' co-relations". While the U.S. audience shows no strong co-related classes, the Russian audience shows even tripolar mutual gravity of *Culture*, *History* and *Martial arts* classes.

However the main goal of the study is to demonstrate capabilities of the query log-based sociology rather than the specific results on the particular datasets. The study shows than regardless of the very short data units query logs provide a sufficient base for full-value sociological conclusions. The obvious advantage of the query log-base sociology is *reliable* data. When a polls-based study has to do with 5-year repetitive 50-percent answer of 500 respondents "*I need to know what is Kyogen*" (cf. [10]) a query log-based study deals with a perfect permille of the queries "*kyogen*".

8. References

[1] Beitzel, S., Jensen, E., Chowdhury, A., Grossman, D., Frieder, O. (2004) ‘Hourly analysis of a very large topically categorized Web query log’, in Proceedings of ACM SIGIR conf. on research and development in information retrieval, 2004, ACM Press, pp. 321-328.

[2] Beitzel, S., Jensen, E., Chowdhury, A., Frieder, O., Grossman, D. (2007) ‘Temporal analysis of a very large topically categorized Web query log’**.** Journal of the Association for Information Science and Technology, 58, 2 **(2**007), pp. 166-178.

[3] Blei, D., Lafferty, J.D. (2006) ‘Dynamic topic models’, in Proceedings of 23rd Int. Conf. on Machine Learning ICML (Pittsburg, USA, June 2006), ACM Press, pp. 113-120.

[4] Jansen, B.J., Spink, A., Saracevic T. (2000) ‘Real life, real users, and real needs: a study and analysis of user queries on the Web’, Information Processing & Management, 36(2), pp. 207-227.

[5] Lewandowski, D (2006) ‘Query types and search topics of German Web search engine users’. Information Services & Use, 26, pp. 261-269.

[6] Richardson, M. (2008) ‘Learning about the World through Long-Term Query Logs’. ACM Transactions on the Web, Vol. 2, No. 4, 2008, pp. 21-27.

[7] Spink, A., Ozmutlu., S., Ozmutlu, H., Jansen, B.J. (2002) ‘U.S. versus European Web searching trends’. ACM SIGIR Forum, 2002, 36(2), pp. 32-38.

[8] Wang, X., McCallum, A. (2006) ‘Topics over time: a non-Markov continuous time model of topical trends’, in Proceedings of KDD '06 (Philadelphia, USA, 2006), ACM Press, pp. 138-145.

[9] Weber, I., Garimella, V.R.K., Borra, E. (2012) ‘Mining Web Query Logs to Analyze Political Issues’, in Proceedings of the 4th ACM Web Science Conf. 2012, June 22–24, 2012, Evanston, Illinois, USA, ACM Press, pp. 330-334.

[10] [US-to-Japan-Polls] (2014 and earlier) The U.S. Polls on opinions toward Japan.http://www.mofa.go.jp/region/n-america/us/survey/index.html