ISSN 2007-9737

Query Topic Classification and Sociology of Web Query Logs

Nikolai Buzikashvili

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

buzik@cs.isa.ru

Abstract. In the paper, the objects, tasks, and a general procedure of the sociological analysis of Web search engine query logs are described and illustrated by a methodologically complete study of the cross-nation search image changes based on two-year spaced query logs of the national search audience.

Keywords. Query log analysis, query expansion, sociology of query logs.

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. The Web era has opened not only a new field of social activity but also a huge source of the 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 not self-contained, and small-element base for causal inferences.

While sociology of the Web mainly answers the question "*Who searches the Web?*" (age, gender, etc.) and uses polls, the sociology of query logs answers the question "*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" [4, 5, 7, 10, 13]. The manual attribution of queries was used in the early studies. More sociologically sophisticated studies such as [12, 16] are rare.

Query log-based works are frequently insufficient both in terms of sociological tasks (in particular, interdependencies of the investigated factors are out of study) and statistical methods used. Also they are not always sophisticated in terms of query topic detection techniques. The aim of this paper is to describe a nearly standard procedure of topic detection as a part of a methodologically full-value query log-based sociological study.

Web search engines' query logs data may be used individually and in combination with data extracted from other sources (e.g. polls) in the form of *sociological data fusion*. Since information extracted from queries is obviously insufficient for explanation of dependencies and changes, the second way seems to be more appropriate when we should correctly explain interdependencies or changes. But the data fusion problems are out of this study and we follow the first approach without technically correct invoking external socioeconomic information.

The rest of the paper is organized as follows. In Section 2, goals and technical problems of query log sociology are stated. The research questions of the illustrative longitudinal query log-based sociological study are formulated in Section 3. Section 4 describes the datasets used. The typical sequence of steps of the topic detection-based sociological study of query logs is given in Section 5. Section 6 presents measures used in the sociological study of the changes of the cross-nation search image based on two-year spaced query logs of the same national audience, Section 7 describes the results of the study, and Section 8 presents conclusions.

doi: 10.13053/CyS-19-4-2325

2 Object and Tasks of Sociology of Query Logs and Query Topic Detection

Objects. Anonymized "query log" datasets are the main base for query log sociology (a useful source of query specification in terms of user age, gender, etc. is personal data available only for search engines' teams and used in their *targeting advertising*). The datasets are combinations of records of several logs of the search engine and contain transaction descriptions grouped by user and ordered by time for each user. The datasets usually include "user" (i.e. cookie) identifier (UID), query string, timestamp of the transaction, and page number of the retrieved results.

Goals and tasks of sociology of query logs. The primary task of any query log-based sociological study is a *perfect unambiguous* attribution of users/queries according to the classification used in the study.

When objects (queries, users) are attributed, the further processing is a usual sociological study of the classes' rates, interrelations, and changes, and does not depend on the specific nature of the queries. To explain the facts revealed in the query log analysis we need to use external, "out-log" knowledge and data, e.g. what social, cultural, political events may determine interests expressed in queries.

The main technical problem is an auxiliary task of automated creating of tools of an automatic classification of queries. Numerous works (e.g., [1, 3, 6, 9, 11, 14, 15]) started by [2, 8] are devoted to query topic detection, including query attribution to predefined thematic classes, and give particular methods of query categorization widely used in *contextual* online *advertising*.

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.

While the tools of the query log-based sociology are similar to the tools of the contextual online advertising, we do not observe query-log sociology flourish. The curious and unavoidable curse of query log-based sociology is that available logs, "the huge source for a longitude sociological analysis", are rare and irregular. Query log providing for non-adversarial researches is a matter of a good and disinterested will of search engines' teams.

3 Longitudinal Study and Its Research Questions

A complete study. The paper describes from scratch all steps of a methodologically complete query log-based sociological study exemplified by a study of queries submitted by the same national audience (Russian) searching for the topics related to the other state (Japan). Two logs (2005 and 2007) of the main Russian search engine Yandex are used.

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 and changes of the rates and classes co-relations during two years. The datasets used in the study are a good base for intra-audience estimates, particularly for detection of the two-year changes of the audience interests.

The reason for choosing Japan as a "perceived object" is that the choice of a "language-exotic" object *decreases* the problem of recognition and disambiguation of topical queries and simplifies an automatic classification of queries.

Research questions. The subject of the particular illustrative study is a *search engine user* as a set of all queries submitted from the same *cookie*. As a result, a single user may include all people who submit queries from the same cookie during the observation period. A user who submitted queries belonging to several Japan-referring classes is attributed to all these classes (that is, user classes intersect). The research questions are:

- Classes' frequencies and their changes in the Russian search audience from 2005 to 2007;
- Classes' co-occurrence and their changes over these two years.

The analysis of two *Yandex* logs may be considered as a methodologically complete fragment of a full-value longitude study of the Russian audience. However, the primary goal of the paper is to demonstrate the potentials and techniques, so the relative (but usual for log studies) weakness of the dataset corpus is far from being the weakness of the method.

4 Datasets and Two-Year Changes of Audience

Two datasets are used in the study: 24-hour complete query logs of the Russian *Yandex* search engine (March 20, 2007, Tuesday) and 7-day sample of the *Yandex* (March 9-15, 2005). The preprocessed datasets after elimination of users with broken log records and users detected as robots are given in Table 1.

Two questions arise: (1) about differences in conditions of observations (2005 *week sample* vs 2007 *one-day complete* dataset) and (2) about structural similarity of audiences (probable determiners are a sex-age and an access (office/home/mobile) structures, and a possible determiner is a geographical distribution (metropolitan vs province)). While rates of users from non-metropolitan areas and young users have increased ([18, 19]), we can consider 2005 and 2007 audiences as the same Russian audience.

A comparison of over-a-day 2007 frequencies with over-a-week 2005 frequencies is not valid (if the probability *p(day)* to submit 1+ Japan-referring queries throughout a day was equal to 0.03 for any user, then the probability to submit Japan-referring queries during 7 days would be *p(7days)* = 1 - (1-*p(day)*)^7 = 0.1911. But since the probability to submit Japan-referring queries once again is bigger for the users submitted such queries earlier, the observed 7-day frequency of Japan-referring users p_{obs}*(7days)*=0.0538 is less than *p(7days)*, see Table 1).

Table 1. 2005 and 2007 datasets description

	2005	**2007**
Sampling	sampled	whole population
Observ.period	week	day
All users	176187	860618
Japan-ref.users	9486	28439
Rate of Jap.-ref	5.38%	3.30%

Table 2. 2005 dataset by days

	All users	**Jap-ref. users**	**Rate (%)**
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
Sums over:			
- 5 work days	299180	9071	3.03
- 2 off days	73058	2282	3.12
- all 7 days	372238	11353	3.05

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).

Reported in Table 3 Yates corrected *z*-values (see Section 6.2) for rates of Japan-referring users in each day-to-day pair and in work-day sum vs off-day sum are smaller than the critical $z_{0.95}$ = 1.96. Since no significant difference between the frequencies of Japan-referring users in any pair of days is observed, we can join 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 372,238 virtual users instead of a set of 176,187 initial real users.

5 Steps of Query Topic Classification

In this section, the complete sequence of data mining steps of any query log-based sociological study is described from scratch.

It is worthy of note that the specific nature of the Japan-related topics significantly simplifies tool creation (mainly non-dictionary words) and an application (mainly unambiguous topical attribution).

The basic tool of the study is a thematic Japan-referring Russian vocabulary containing topically attributed keywords and keyword-combinations named "the *vocabulary*". The words not included in the vocabulary are referred to as *non-vocabulary*. The standard Russian electronic dictionary is referred to as "*dictionary*", and words non-included in this dictionary are referred to as *non-dictionary*.

All vocabulary creation/modification actions below are two-stage: (1) an automatic extraction of words/combinations as candidates for inclusion in the vocabulary, (2) a manual approval of candidates and attribution of them to vocabulary topics.

5.1 Step 1. Japan-referring Topics Selection and Initial Thematic Japan-referring 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 create a Japan-referring thematic vocabulary that includes words and word-combinations marked by the word topic categories.

a. 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.

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

	Off-day sum
Work-day sum	0.24

b. Initial Vocabulary Creation

(1) *Data.* A small Russian-language corpus of Japan-referring texts and a big Russian-language corpus of non-Japan-referring texts.

(2) *A tool* for non-dictionary word extraction: Russian electronic dictionary.

(3) *Procedure*

(3.1) *Two ways of automatic extraction of candidates*: (3.1.a) *non-dictionary words and word-combinations* in the corpus of Japan-referring texts; (3.1.b) *words significantly more frequent* in Russian Japan-referring texts than in Russian non-Japan-referring texts.

(3.2) *Manual approval of candidates, bilingual parallelization, and categorization in terms of basic categories.* Almost all non-dictionary words and word-combinations detected by (3.1.a) and only a few too-frequent words detected by (3.1.b) were approved as Japan-referring words and word combinations. The extracted Russian spellings were attributed to basic categories.

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, Nihon*
Geographic names	107	*Chugoku, Tokyo, Kyoto*
Basic:		
Religion & Ethics		*satori, shinto, tsukuyomi, zen, todaiji*
Tradit. Art & Theater	55	*gadaku, hokusai, koto, origami, utamaro*
Tradit. Lifestyle	45	*kimono, ryokan, tatami, yakuza*
Literature	37	*haiku, kanji, mukai, renga, miyamoto musashi*
Tradit. Food	26	*sake, sashimi, sushi, tsukemono*
Interstate Relations	85	*edo, hojo, meiji, samurai, taisho, 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*

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

The distributions of words among the 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.

5.2 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 her.

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" should be detected first in the query <school essay japan culture download>.

The aims of Step 2 are

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

(1.1) If the number of users marked by the 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.

Table 5. Initial Japan-related user categories and their rates among all users

	Absolute values		Rate (%) among all users	
	2005	**2007**	**2005**	**2007**
Japan-refer.users	11353	28439	3.050	3.304
General	1377	3089	0.370	0.359
Geograph. names	120	611	0.032	0.071
Religion & Ethic	55	118	0.015	0.014
Art & Theater	93	243	0.025	0.028
Tradit. Lifestyle	72	250	0.019	0.029
Literature	99	154	0.027	0.018
Tradit. Food	27	112	0.007	0.013
History	192	595	0.052	0.069
Martial Arts	152	300	0.041	0.035
Masscult & Movies	445	1117	0.120	0.130
Cars	2897	10905	0.778	1.267
Goods	6187	11847	1.662	1.377

Table 6. Classes of words and number of users in corresponding user classes of datasets

Class	Categories included in Class	2005	2007
General	General, Geographic names	1477	3639
Culture	Religion,Art,Lifestyle, Literature, Trad.Food	342	869
History	History	192	595
Martial Arts	Martial arts	152	300
Masscult	Masscult movies	445	1117
Cars	Cars	2897	10905
Goods	Goods	6187	11847

(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 symbols (for more than 6-symbol long words),

the word is detected as a probable misprint or non-unique spelling (e.g., *mitsubishi* and *mičubisi* in Russian). All detected words are inspected for a mass presence in the Web and if presented they are checked manually. The approved spellings are added to the vocabulary, and only exact match with vocabulary words is taken as due account in the further study.

Table 5 shows the categorization of the users according to the queries submitted by them.

5.3 Step 3. Category Re-Combination and Vocabulary Expansion

a. Combining Categories into Compound Classes

Small (in terms of the users attributed) 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 classes 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.

b. 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 the General class.

Procedure.

(a) A non-Japan-referring dictionary word w_i more frequently collocated in queries with a *General* class word W than w_i which 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 noting 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 word expansion with dictionary non-vocabulary words.

Definitions.

A *temporal session* is defined as a sequence of the user 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 word 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) Japan-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 the expansion of these word classes yields no results in the user classes since only users earlier detected as Japan-referring use these words.

6 The User Class Co-Relation Measure and Tests Used

6.1 Inter-Class Co-Relation Measure (One-Tailed Fisher's Exact Test 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 the observed intersection, 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 $p(k \leq obs\,(i, j))$ 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 $p(k \leq obs\,(i, j))$, the stronger the interrelation is.

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

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":

$(\; n \; \backslash (\; N - n \; \backslash \;\; // (\; N$

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

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. Beside 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.

6.2 Tests for Two-Year Change Detection

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:

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

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.

Table 7. User classes, their rates among all users, and Yates corrected *z* values (similar rates are in bold)

	Abs. values		**Rate(%) among all**		**z-values**
	2005	**2007**	**2005**	**2007**	
Jap-ref. users	11353	28439	3.050	3.304	7.34
General	205	454	0.055	0.053	**0.47**
Culture	824	2083	0.221	0.242	**2.15**
History	261	800	0.070	0.093	3.94
Martial Arts	191	388	0.051	0.045	**1.42**
Masscult	627	1624	0.168	0.189	**2.40**
Cars	3026	11588	0.813	1.347	25.12
Goods	6426	12069	1.726	1.402	13.58

Table 8. Co-occurrence of Japan-referring user classes in 2005 and 2007 datasets

2005	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

2007	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

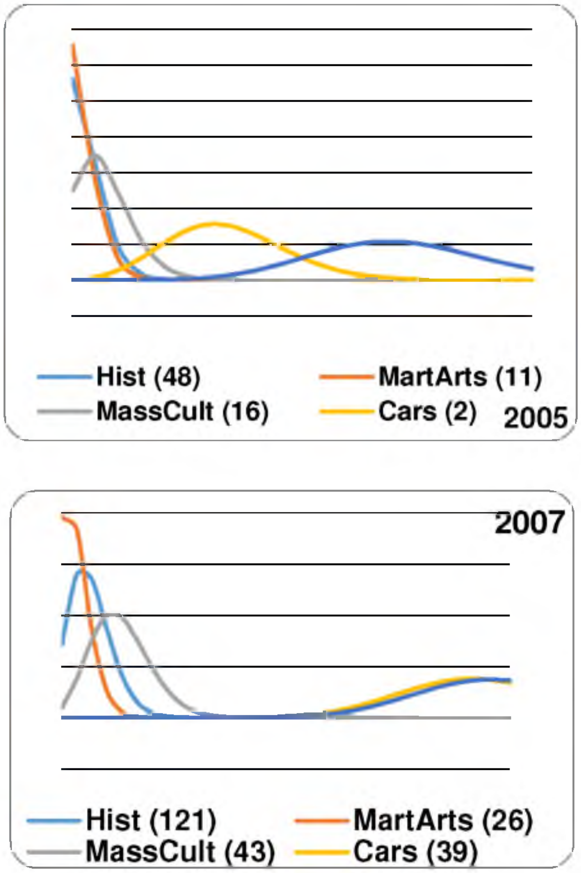

Fig. 1. Probabilities of intersections of the Culture class with other classes among all 2005 (top) and 2007 (bottom) users (observed intersections are given in brackets)

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.

7 Particular Study Results, Changes of Rates, and Co-Relations

User classes detected under the mutated vocabulary are shown in Table 7.

7.1 Changes of 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.

Two-year changes of Japan-referring interests. There are no changes in the rates of *Culture*, *Masscult*, and *Martial arts* classes. The *History* rate grows.

However, the mostly impressing are the opposing changes in the rates of "consumer classes": enormous growth of the *Cars* rate against a significant decrease of the *Goods* rate during two years.

Proportions between user classes among Japan-referring users. The observed χ^2 value for distributions of 7 classes equals to 805.9 while critical χ^2(*6 d.f., p=0.95*) = 12.6. Thus, no similarity of proportions in the 2007 and 2005 data is observed.

7.2 Changes of Co-Occurrence of User Classes

The co-relation 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 the diagonal elements (sizes of classes) and the sums of the 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 the co-occurrence of 3+ classes.

Probability distributions of intersections of all pairs of *independent* classes are near-symmetric but have a *cropped left tail* except the pairs containing big *Cars* and *Goods* classes (see Fig. 1 presenting probabilities of intersections of the *Culture* class with other classes). A center of the distribution for a pair of small classes is only a little bigger than 0 and even the minimal probability of intersection $p(k \leq 0)$ is big enough. This is an additional reason why the reliability of big $p(k \leq obs\ (i, j))$ for intersections of small classes is smaller than the reliability of the same or smaller values for big classes.

Table 9. Independence of basic user classes in the *all users* population (*incompatible* is given in bold italic, near independent are given in bold)

2005	Hist	MArts	Mass	Cars	Goods
Cult	1	1	1	***0.036***	**0.239**
Hist		0.999	0.999	0.937	0.832
MArts			0.723	0.928	0.580
Mass				0.749	0.600
Cars					0.999

2007	Hist	MArts	Mass	Cars	Goods
Cult	1	1	1	0.981	**0.315**
Hist		~1	1	1	0.839
MArts			0.833	0.918	0.818
Mass				**0.174**	**0.250**
Cars					1

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

2005	Hist	MArts	Mass	Cars	Goods
Cult	**1**	**0.260**	0	0	0
Hist		**0.181**	~0	0	0
MArts			~0	0	0
Mass				0	0
Cars					0

2007	Hist	MArts	Mass	Cars	Goods
Cult	**1**	**0.362**	0	0	0
Hist		0.037	0	0	0
MArts			0	0	0
Mass				0	0
Cars					0

a. Class Independence Changes over Two Years

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

Both 2005 and 2007 logs show incompatibility of the *Culture* and *Goods* classes. On the contrary, while the big *Culture* and *Cars* classes are strongly incompatible in 2005, they are co-related in 2007. Other two-year changes are: *Masscult* changed its co-relations with consuming classes from nearly independent in 2005 to incompatible in 2007. The former may be an artifact caused by the small size of the *Martial Arts* class. But changes in *Culture* and *Cars*, or incompatibility of the *Masscult* and *Cars* and *Goods* in 2007, cannot be explained by a 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.

b. "Over-Strong Co-Relation"

Probabilities of class co-occurrence among *only Japan-referring* searchers ("over-strong co-relation" measure) are given in Table 10. The Russian audience shows some strong co-related classes, particularly, the strong *tripolar mutual gravity* of *Culture*, *History*, and *Martial Arts* both in 2005 and 2007.

7.3 Surprising Dynamics?

Two-year big changes in Japan-referring interests of the Russian searcher audience are revealed. Is it a big surprise? To answer, we should consider changes in any similar sociological study. The only similar longitude is a Japan-referring US poll [17]. Fig. 2 shows 3-year dynamics of positive answers to "For each of the areas of Japanese culture tell me if you are interested in the topic" [17] (a question asked beginning from 2011). Although the verity of poll answers sometimes is not far from the verity of personal data in users' accounts, the results reported in [12] show a possibility of wide range of changes.

8 Conclusions

We have investigated (1) the rates of the topical classes and inter-class relations in two (2005 and 2007) Russian search images of Japan presented in the *Yandex* logs, and (2) two-year changes in the Russian search image. The findings of this particular query log-based sociological study of class rates, co-relations, and two-year changes in them are as follows.

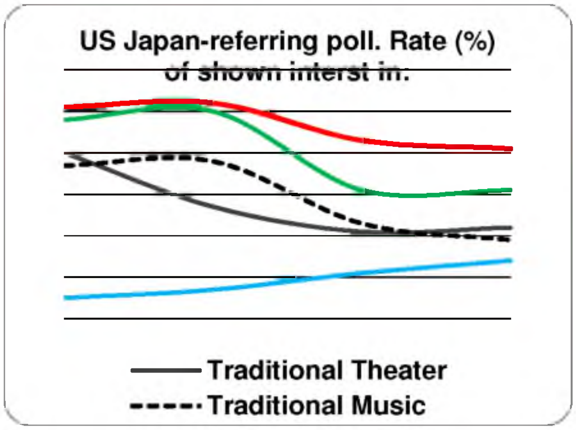

Fig. 2. Three-year changes of interest to Japan cultural topics in US audience

(1) **Rates of the Japan-referring classes and changes in rates (see Table 7)**. There are no two-year changes in the rates of Culture, Masscult and Martial Arts classes of Russian searchers. The rate of the History class grows. The "consumer classes" (Cars and Goods) demonstrate rapid, strong, and opposing changes: an enormous growth of the Cars rate against a significant decrease of the Goods rate over two years.

(2) **Class co-relations: independence and strong co-relation and changes in these relations**. There are three pairs of independent classes in 2005: Culture is "incompatible" with *Cars* and *Goods*, and *Masscult* is absolutely incompatible with *Martial Arts* (the latter incompatibleness may be an artifact resulted from the small size of the *Martial Arts* class). However, two years later the role of *Culture* is played by *Masscult*: the *Masscult* class is "incompatible" with both consumer classes, while *Culture* is compatible with *Cars* (see Table 9). On the contrary, there is no change in the strongest co-relation between the *Culture, History*, and *Martial Arts* classes (see Table 10).

Of course, the results of the exemplifying study may be not without interest. However, the primary goal of the undertaken study is to demonstrate capabilities of the query log-based sociology rather than the specific results on the particular datasets.

The study shows that regardless of the very short data units, query logs provide a sufficient base for sound sociological conclusions. The obvious advantage of the query log-base sociology over polls is *reliable* data. And anyway, Web search engines' query logs data deserves the right to be one of the sources for sociological data fusion.

References

1. **Aiello, L.M., Donato, D., Ozertem, U., & Menczer, F. (2011).** Behavior-driven clustering of queries into topics. *Proc. of the 20th Conference on Information and Knowledge Management (CIKM'11),* ACM, New York, NY, pp. 1373–1382. DOI: 10.1145/2063576.2063775.
2. **Allan, J., Papka, R., & Lavrenko, V. (1998).** On-line new event detection and tracking. *Proc. of the 21st ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'98)*, ACM, New York, NY, pp. 37–45. DOI: 10.1145/290941.290954.
3. **Alvanaki, F., Sebastian, M., Ramamritham, K., & Weikum, G. (2011).** EnBlogue: emergent topic detection in web 2.0 streams. *Proc. of the 2011 ACM SIGMOD/PODS Conf. (SIGMOD'11)*, ACM, New York, NY, 2011, pp. 1271–1274. DOI: 10.1145/1989323.1989473.
4. **Beitzel, S., Jensen, E., Chowdhury, A., Grossman, D., & Frieder, O. (2004).** Hourly analysis of a very large topically categorized Web query log. *Proc. of the 27th ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'04),* ACM, New York, NY, pp. 321–328. DOI: 10.1145/1008992.1009048.
5. **Beitzel, S., Jensen, E., Chowdhury, A., Frieder, O., & Grossman, D. (2007).** Temporal analysis of a very large topically categorized Web query log. *JASIST*, Vol. 58, No. 2, pp. 166–178. DOI: 10.1002/asi.20464.
6. **Chelaru, S., Altingodve, I.S., Siersdorfer, S., & Nejdl, W. (2013).** Analyzing, detecting, and exploiting sentiment in web queries. *ACM Transactions on the Web*, Vol. 8, No. 1. DOI: 10.1145/2535525.
7. **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. *Inf. Proc. & Management,* Vol. 36, No. 2, pp. 207–227. DOI: 10.1016/S0306-4573(99)00056-4.
8. **Lam, W., Mukhopadhyay, S, Mostafa, J., & Palakal, M. (1996).** Detection of shifts in user interests for personalized information filtering. *Proc. of the 19th ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'96)*. ACM, New York, NY, pp. 317–325. DOI: 10.1145/243199.243279.
9. **Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Polland, V., & Thomas, S. (2002).** Relevance models for topic detection and tracking. *Proc. of 2nd Conf. on Human Language Technology (HLT'02)*, Morgan Kauffman, pp. 115–121.
10. **Lewandowskl, D. (2006).** Query types and search topics of German Web search engine users. *Information Services & Use*, Vol. 26, No. 4, pp. 261–269.
11. **Li, L., Xu, G., Yang, Z., Dolog, P., Zhang, Y., & Kitsuregawa, M. (2013).** An efficient approach to suggesting topically related web queries using hidden topic model. *World Wide Web*, Vol. 16, No. 3, pp. 273–297. DOI: 10.1007/s11280-011-0151-3.
12. **Richardson, M. (2008).** Learning about the World through Long-Term Query Logs. *ACM Trans. on the Web*, Vol. 2, No. 4, pp. 21–27. DOI: 10.1145/1409220.1409224.
13. **Spink, A., Ozmutlu, S., Ozmutlu, H., & Jansen, B.J. (2002).** U.S. versus European Web searching trends. *ACM SIGIR Forum*, Vol. 36, No. 2, pp. 32–38. DOI: 10.1145/792550.792555.
14. **Shen, D., Sun, J-T., Yang, Q., & Chen, Z. (2006).** Building bridges for web query classification. *Proc. of the 29th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*, ACM, New York, NY, pp. 131–138.
15. **Allan, J. (ed.) (2002)**. *Topic Detection and Tracking. Event-based information organization*. Kluwer.
16. **Weber, I., Garimella, V.R.K., & Borra, E. (2012).** Mining Web Query Logs to Analyze Political Issues. *Proc. of the 2012 ACM Conference on Web Science (WebSci)*, Evanston, USA, ACM, New York, NY, pp. 330–334. DOI: 10.1145/2380718.2380761.
17. **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.
18. **Internet in Russia, Russia in Internet polls. (2008).** *Report 22*. http://bd.fom.ru/report/map/az/%D0%A0%E2%84%96/internet/internet0801/int08011.
19. **Development of the Internet in Russia's Regions (2013).** Yandex, 2013. http://download.yandex.ru/company/ya_russian_regions_report_2013.pdf.

Nikolai Buzikashvili received the M.Sc. in Applied Mathematics from the Moscow Institute of Economics and Statistics. He is a Senior Researcher at the Institute for System Analysis of the Russian Academy of Sciences. Areas of his interest are Web information search and retrieval. He is author of more than 50 publications in the area of computer science.

Article received on 09/02/2015; accepted 05/03/2015.
Corresponding author is Nikolai Buzikashvili.