ICIT 2015 The 7th International Conference on Information Technology
doi:10.15849/icit.2015.0085

A Query Log-Based Study of Cross-Nation Perception

Nikolai Buzikashvili
Institute of System Analysis
Russian Academy of Sciences
Moscow, Russia
buzik@cs.isa.ru

***Abstract*— Query logs are a huge and solid source for sociological analysis. However, they are insufficiently used in the sociological analysis, in particular in the comparative studies of different audiences. The paper presents a study of search images of Japan in queries of Russian and U.S. Web searchers. One-day logs of the *Yandex*, the Russian search engine, and the U.S. *Excite* were automatically analyzed to detect several categories of queries referring to Japan. Users submitting Japan-referring queries were attributed to these categorized. The study (*a*) compares rates of categorized Japan-referring users among Russian and U.S. searchers, (*b*) analyzes cross-linking between categories. The findings are: (*a*1) the Russian searchers are more interesting in Japan-referring topics, (*a*2) differences depend on categories: Russians show much more consumer interests, while U.S. are superior in masscult interests; (*b*1) the users submitting consumer queries less frequently search other topics referring to Japan; (*b*2) the users submitting queries relating to Japan culture more frequently search other Japan-referring topics; (*b*3) a Russian searcher searches several different Japan topics more frequently than U.S. searcher; (*b*4) the Russian and U.S. audiences significantly differ by the topic co-occurrence.**

Keywords— query log analysis; cross-cultural perception

I. INTRODUCTION: SOCIOLOGY OF SEARCH

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 data for sociological analysis. Query logs of Web search engines are a capacious but very special source of knowledge on public interests. Logs as such give no way to reveal either attitudes or origins of interests (except when a query is a result of another query).

While sociology of the Web mainly answers "*Who searches the Web?*" (age, gender, education, 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 searched topics ([1], [2], [4], [5], [7]). More sociologically sophisticated query log based studies such as [6], [9] are so far rare.

The paper presents a comparative study based on the logs of queries submitted by two national audiences (Russian and U.S) searching for the topics related to the third state (Japan) and its culture. In the study we use query logs of the *Yandex* (2007) and *Excite* (2001) search engines. The *Yandex* is the main Russian search engine and the *Excite* was a popular U.S. search engine in the early 2000s.

We study topic categories of queries related to Japan and corresponding categories of users submitting these queries. In this study, we compare two collective subjects: population of Russian and population of U.S. searchers. First of all, we will try to detect differences of search images of Japan among these populations. Another question is co-relation between searching different Japan-referring thematic classes.

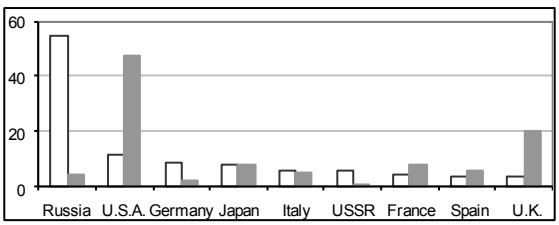

Fig. 1. Rates (%) of countries among 9 states referring to in Yandex-07 (white) and Excite-01 (gray) logs

We process the Japan-referring queries. While any country may be chosen as a perceived object, the reasonable questions are: is a number of queries referring to a country sufficient for statistically significant conclusions and how frequent are these queries among the queries related to other countries in the compared logs? We compare the rates of queries directly referring to 9 countries (Fig. 1). Fractions of Japan-referring queries are big enough and approximately equal in both logs.

doi:10.15849/icit.2015.0085

II. RESEARCH QUESTIONS

The subject of the study is a *user* i.e. a set of all queries submitted by him rather than a *query* or a (task or temporal) *session*. In the study, we investigate fractions of Russian and U.S. users submitting Japan-referring queries. The same user may submit queries belonging to different Japan-referring classes. The research questions are:

- How frequently do Russian (both *Yandex* logs) and U.S. searchers submit queries of Japan-referring classes?
- How do Japan-referring classes co-occur in a set of queries submitted by the same Russian or U.S. user?
- A comparison of fractions and classes' co-occurrence of Russian and U.S. Japan-referring searchers (*Yandex*-07 vs. *Excite*-01, a cross-time cross-nation comparison)

III. DATASETS

In this study we use 24-hour query logs of the Russian-language *Yandex* search engine (March 2007, 890,897 users) and 24-hour log sample of the U.S. *Excite* (May 4, 2001, 305,360 users). The datasets are different:

(1) *in time* (2007 vs. 2001; a "search image" varies over time). Since observation periods of the both datasets are equal to 24 hours we can ignore intraday variations. However, we cannot ignore week and season variations (e.g. in tourist queries) and we cannot ignore a long-term trend, which is particularly important and reflects changes of interests and change of the available Web services as alternative ways to get Japan-referring information. The datasets which we use are spaced far apart in time and search images may be shifted considerably over the years. As a result, can the comparison be valid? This is a crucial question for any, even time-synchronized, comparative study. Ideal comparative study of parallel social processes should be study of time series rather two time slices, even made in the same time.

(2) *in audience* (mainly Russian vs. mainly U.S.) and in population structure. Of course, not only Russians use *Yandex* and not only U.S. searchers used *Excite* in 2001. However, we can suppose that a 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 "American day" (0am – 6am, Pacific time zone) when non-American users are active.

(3) *in language* (mainly Russian in the *Yandex* log vs. mainly English in the *Excite* log). The queries submitted to the *Excite* are queries in English (~90%), German and Spanish. The queries submitted to the *Yandex* contain words in Russian, two other Slavonic languages (Ukrainian and Belorussian) and in English. The *Yandex* users *commonly* use English spelling of Japanese brand names.

IV. JAPAN-REFERRING VOCABULARY CREATION

To detect and categorize Japan-referring words, queries and users we use two crucially different kinds of categories:

(1) *basic categories* corresponding to *both* aspects (a general reference to Japan and a certain thematic denotation, e.g., culture) and

(2) two *subsidiary categories-filters* used to detect those Japan-referring queries, which cannot be classified by perfect theme. These subsidiary categories are *general* (corresponding queries contain *japan** stem, e.g. <*Japan*>, <*Japanese culture*>) and *geography* (Japanese geographic and administrative names). Queries attributed to subsidiary categories should be categorized into basic categories in the next steps.

To detect and categorize Japan-referring queries we use the automatic procedure based on the Japan-referring vocabulary (hereafter only words from this vocabulary are referred to as "vocabulary words"). It contains both Russian and English words related to Japan. About 300 word-combinations, words and stems were selected (both Russian and English spelling for each word; and some words in each language were presented in different writings, e.g. *mitsubishi* and *mičubisi*). Table 1 exemplifies initial categories of Japan-referring words used in the preliminary analysis. Some words were attributed to multiple categories during the preliminary analysis.

TABLE I. EXAMPLES OF JAPAN-REFERRING WORDS

Category, Number of Words and Word-Combinations in		**Examples**
Subsidiary Categories		
General	17	Japan, Japanese, Nippon, Nihon
Geography	107	Chugoku, Tokyo, Kyoto
Basic Categories		
Religion & ethic	50	satori, shinto, tsukuyomi, zen, todaiji
Traditional art, theater	55	hokusai, netsuke, origami, utamaro
History & interstate relations	85	edo, hojo, meiji, samurai, taisho, tokugawa, yamato
Traditional lifestyle	45	kimono, ryokan, tatami, yakuza
Literature	37	haiku, kanji, mukai, renga, kobo abe, miyamoto musashi
Masscult, movies	16	anime, manga, pokemon
Martial art	24	aikido, budo, judo, karate, kendo, kyudo, sumo
Traditional food	26	sake, sashimi, sushi, tsukemono
Cars	30	mazda, tyota
Consumer Goods	59	Marubeni, canon

A serious problem in the query processing is a lot of typos and a confusing spelling. For example, while a Russian spelling of *Mitsubishi* is *Mitsubisi*, 433 Russian searches type Russian *Mitsubisi*, 51 searchers use Russian *Mitsubishi* (and 909 Russian searchers type *Mitsubishi* in English). To avoid a confusable spelling problem we use all probable variants of spelling.

Multi-categorization of vocabulary words. The original categorization allows a multi-valued word attribution, e.g. *kotatsu* belongs to both *religion* and *lifestyle* categories. However, since one of our goals is a study of cross-category dependency among all queries submitted by a user, this manifold word attribution is undesirable since it leads to artifactual detection of cross-category dependencies. Some

doi:10.15849/icit.2015.0085

words have only one sense in any occurrence but senses of different occurrences are different. For example, *Hiroshima*, *Nagasaki* may occur either as historic or geographic terms, while *Pearl Harbor* is also a 2001 movie and a lot of *Excite*-2001 queries refer to the movie and a big fraction of historic queries is provoked by the movie.

V. PRELIMINARY ANALYSIS

The aims of the preliminary analysis are (1) a rough detection of categories among users' queries, (2) disambiguation of variants of words use, and (3) detection of necessity and possibility to combine different categories into non-overlapping classes. There are 2 reasons to combine different categories: (a) irremovable co-occurrence of different categories for some words and (b) too small rates of several categories among queries.

Each query are attributed to all categories of the vocabulary words contained in the query and a user is attributed to all categories of the vocabulary words contained in all queries submitted by him. Users submitting queries containing multi-attributed vocabulary words are attributed to all categories of these words. 29,208 (3.28%) of 890,897 *Yandex* users and 4,553 (1.49%) of 305,360 *Excite* users submitting queries containing the Japan-referring words. Fig. 2 shows the distributions of the *Yandex* and *Excite* users among Japan-referring categories.

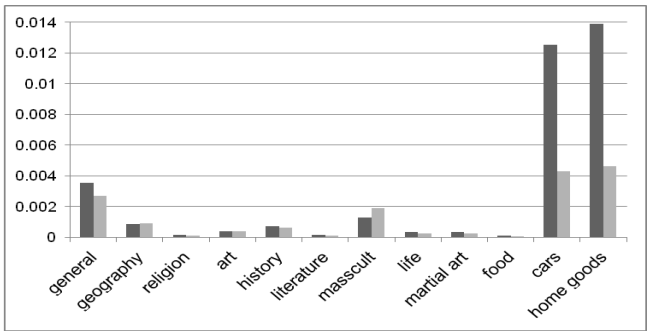

Fig. 2. Rates (%) of categories of Japan-referring users in the Yandex (black) and Excite (gray) logs

Table 2 shows number and fractions of users submitting queries containing words of corresponding categories. To compare these fractions in the *Yandex* and *Excite* logs we use *z*-test in form of:

$$z = \frac{| \widehat{p}_{Yandex} - \widehat{p}_{Excite} |}{\sqrt{\bar{p}(1-\bar{p})\left(1/n_{Yandex} + 1/n_{Excite}\right)}} \qquad (1)$$

where p_{Ynadex} and p_{Excite} are sample rates for the category in each log and p is a sample rate in a combined population. Fractions of consuming categories (*cars*, home *electronics*, *other* consumer *goods*) and *masscult* category are enormously different. Fractions of two categories (*history* and *literature*) are different for $z_{0.95}$ = 1.96 but we cannot discard the null hypothesis at $z_{0.99}$ = 2.58. Fractions of *geography*, *religion&ethics*, *arts* and *traditions* categories are equal even at 0.95.

TABLE II. COMPARISON OF FRACTIONS OF USERS ATTRIBUTED TO CERTAIN JAPAN-RELATED CATEGORIES AMONG 29,208 *YANDEX* AND 4,553 *EXCITE* USERS SUBMITTING JAPAN-RELATED QUERIES

Category	Yandex users		Excite users		*z* test
	Number	Fraction (%)	Number	Fraction (%)	
general	3,158	0.355	820	0.269	7.12
geography	812	0.091	272	0.089	0.33
religion&ethics	125	0.014	35	0.015	1.06
arts	247	0.028	71	0.023	1.31
traditions	120	0.013	43	0.014	0.25
history&interstat	643	0.072	185	0.061	2.10
e literature	159	0.018	37	0.012	2.14
masscult,movies	1,152	0.129	585	0.192	7.80
life	328	0.037	74	0.024	3.27
martial art	312	0.035	76	0.025	2.69
meal	112	0.013	20	0.066	2.73
cars	11,160	1.253	1,308	0.428	38.71
electronics	10,485	1.177	1,026	0.336	41.08
other goods	1,920	0.216	383	0.125	9.80

VI. RE-CATEGORIZATION: COMPOUND CLASSES

We take into account the preliminary analysis results regarding (1) a size of users categories (size of some categories is small for statistical inferences) and (2) a ambiguous categorization:

(*a*) closely related categories are combined into compound classes. The resulted 7 classes (5 basic classes and subsidiary *general* and *geography*) are shown in Table 3. We do not change attributes of the vocabulary words assigned in terms of 12 initial categories. Only processing is changed: if a word belongs to any category it accounted as belonging to corresponding class.

(*b*) vocabulary words belonging to different *new* classes are re-attributed to avoid a multiple categorization in terms of classes. (The only exception is *Pearl Harbor* which is frequently used both as *masscult* (the movie) and as *history*. Queries containing *Pearl Harbor* are attributed manually either to *masscult* or to *history*). Since some words and queries initially attributed to *geography* and *history* were re-attributed, data in Table 4 differs from data in Table 3. Now, if a user attributed to several classes, queries of this user really contain different words belonging to these classes.

TABLE III. NON-OVERLAPPED CLASSES OF WORDS

Class	Categories included into Class
general	general
geography	geography
culture	religion&ethics, arts, traditions, literature, life, food
history	history & interstate_relations
martial ort	martial art
masscult	masscult, movies
goods	cars, home electronics, other consumer goods

Fig. 3 and Table 4 show rates of users submitting queries containing reclassified Japan-referring words. Now all fractions are different at $z_{0.95}$ for Russian and U.S. searchers (cf. Table 2). Fractions of all classes among *all* Russian searchers are bigger than corresponding fractions among *all*

doi:10.15849/icit.2015.0085

U.S. searchers. At the same time, fractions of all non-consuming classes among *Japan-referring* searchers are significantly smaller than corresponding fractions among U.S. *Japan-referring* searchers. The Russian Japan-referring search is mainly consuming.

Fig. 3. Rates (%) of classes of the Japan-referring users among all Yandex (white) and Excite (gray) users

TABLE IV. COMPARISON OF FRACTIONS OF USERS ATTRIBUTED TO CERTAIN JAPAN-RELATED CLASSES AMONG 29,208 YANDEX AND 4,553 EXCITE USERS SUBMITTING JAPAN-RELATED QUERIES

Category	**Yandex users**		**Excite users**		***z* test**
	Number	**Fraction**	**Number**	**Fraction**	
general	3,158	0.355	820	0.269	7.12
geography	687	0.077	158	0.052	4.55
culture	890	0.100	204	0.067	5.22
masscult	1,152	0.129	585	0.192	7.80
history	606	0.068	174	0.057	2.06
martial art	312	0.035	76	0.025	2.68
goods	23,029	2.585	2,663	0.872	56.35

Co-occurrence of classes. If a user submits queries of several classes he is attributed to all these classes ("multi-class user"). A few users were attributed to two classes as a maximum. Table 5 shows the contingency table of users automatically attributed *only to basic classes* (as a result, diagonal values in Table 5 are less than values in Table 4 since users attributed to subsidiary classes are frequently attributed to other classes too); users attributed (also) to *general* and *geography* subsidiary classes are not included and will be re-attributed in the next chapter.

TABLE V. CONTINGENCY TABLE FOR 25,332 YANDEX AND 3,591 EXCITE "NON-SUBSIDIARY" USERS

Yandex	culture	history	martArt	masscult	goods
history		576	2	10	37
martial arts			296	1	11
masscult				1,116	15
goods					22,705
Excite	culture	history	martArt	masscult	goods
culture	183	1	0	5	2
history		158	1	5	10
martial arts			74	0	0
masscult				564	2
goods					2,638

VII. RE-ATTRIBUTION OF SUBSIDIARY CLASSES

Now we should classify users who were automatically recognized as belonging to two subsidiary classes (*general* and *geography*). The rates of these users are big enough (Table 4) and they should be automatically or manually categorized into basic classes.

The idea is that non-vocabulary words which co-occur with a vocabulary word may be related to the basic class assigned to this word. Two types of the co-occurrence were considered: (1) a *narrow* query-based co-occurrence in the same query and (2) a *wide* user-based co-occurrence in a whole set of queries submitted by the same user (e.g., if a user attributed just to one class *martial art* submits two queries <*tortie cat*> and <*jujutsu*>, then words *tortie* and *cat* co-occur with *jujutsu* and are considered as possible associated words of the *martial art* class). The first step of the automatic classification is mining of *non-vocabulary* words associated with any class. To mine them we use only those items (Japan-referring queries in the case of the narrow co-occurrence or Japan-referring users in the case of the wide co-occurrence), which are attributed just to one class. Next, if extracted co-occurred words more frequently occur in Japan-referring queries (a narrow co-occurrence) or in any query of Japan-referring users (a wide co-occurrence) attributed to the class these co-occurred words are considered as associated non-vocabulary words of this class. Let associated words be extracted for each basic class. Then queries attributed to subsidiary classes may be re-attributed to basic classes by occurrence of associated words.

To extract the non-Japan-referring terms, which represent the classes of queries we use following class-based metrics:

— *tf* (*term T* | *item_of_class_Cl*) — *"class frequencies"* of the non-vocabulary *term T* in *item_of_class_Cl*, i.e. the ratio of the number of *term T* occurrences in *item_of_class_Cl* to the total number of all words occurrences (a total length) in all unique queries belonging to *item_of_class_Cl*. The *item_of_class_Cl* is either any query containing vocabulary words belonging to *class_Cl* (the narrow co-occurrence) or any query of a user submitting at least one query containing vocabulary words belonging to *class_Cl* (the wide co-occurrence).

— *cf (term T)* — "*collection frequency"* of the non-vocabulary *term T* in all unique queries of the query collection, i.e. the ratio of the number of *term T* occurrences in all unique queries of the query collection to the number of all words occurrences in all unique queries of the collection.

— *contrast (term T | item_of_class_Cl) =*
tf (*term T* | *item_of_class_Cl*) / *cf (term T)*. If this ratio is significantly bigger than 1, then *term* represents *class_Cl*.

We count *contrast* (*term T* | *class Cl*) to detect non-vocabulary words closely connected to the vocabulary classes: terms which occur either (1) in the queries belonging to the class more often than to the other queries (narrow co-occurrence) or (2) in all queries submitted by users attributed

doi:10.15849/icit.2015.0085

to the class more often than in the queries of all other users (wide co-occurrence).

Fig. 4 shows results of re-classification of 3195 (of 3845) Russian and 703 (of 955) U.S. searchers primarily recognized as subsidiary classes (other "subsidiary users" submitted too general queries such as <*Japan*> were not re-classified). Fractions of *sex* and *cars* classes are enormously different and look like mirror images: At first sight it may be interpreted as a result of the *from-e-sex-to-e-commerce* tendency (Spink et al., 2002a). However we do not discover this tendency in *Japan-referring* queries comparing the *Excite* logs (2001 vs. 1999) or *Yandex* logs (2007 vs. 2005). Too low fraction of Japan-referring sex searchers among *Yandex* users may be partly explained by the fact that *Yandex* covers only the Russian Web domain.

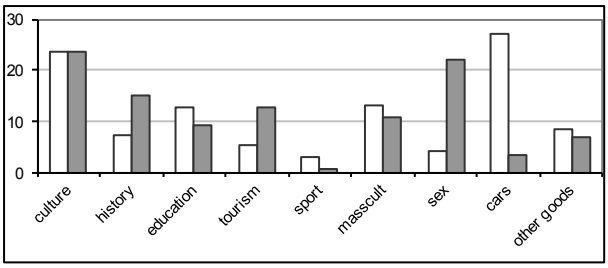

Fig. 4. Rates (%) of basic classes among Yandex and Excite users initially attributed to subsidiary classes

Even forth query in the *general* class is <*Japanese autos*>. Topics of *history* queries are really different between Russian and U.S. queries. While the latter are focused on World War II, the former practically ignore the WW II period but search for such topics as *Japanese ethnos forming* or *constitution of 1899*.

Let's present "subsidiary" Japan-referring users in terms of basic classes to add these re-classified users to other Japan-referring users automatically attributed to the basic classes in the previous chapter. Namely, we combine *culture*, *education* and *tourism* classes into basic *culture* class, *sex* is added to *masscult*, *cars* and *restaurants* are added to *goods*. Since any user is attributed to two classes as a maximum, to group classes we use inclusion-exclusion rule for two sets:

$$n(Group) = \sum_{class \in Group} n(class) - \sum_{class1,class2 \in Group} n(class1 \cap class2)$$

$$n(Group1 \cap Group2) = \sum_{class1 \in Group1,\, class2 \in Group2} n(class1 \cap class2) \qquad (2)$$

Classes co-occurrence. Table 6 presents a contingency table of manually re-classified users initially attributed to *general* and *geography* subsidiary classes. Since not all queries containing words of subsidiary classes (*general* and *geography*) may be recognized in terms of 5 basic classes some of "subsidiary users" were not re-attributed.

TABLE VI. CONTINGENCY TABLE FOR 3,195 *YANDEX* AND 703 *EXCITE* MANUALLY RE-ATTRIBUTED SUBSIDIARY USERS

Yandex	culture	history	martArt	masscult	goods
culture	1,305	104	15	30	19
history		234	3	4	6
martial art			103	0	3
masscult				543	14
goods					1,198
Excite	culture	history	martArt	masscult	goods
culture	311	11	1	9	3
history		108	0	1	0
martial art			5	0	1
masscult				225	0
goods					79

VIII. CONSUMERISM VS. "MASSCULTURISM"

Now we can add re-classified "subsidiary" searchers to searchers classified by the five basic classes (*culture*, *history*, *martial art*, *masscult* and *goods*). Fig. 5 shows rates of 5 basic classes among *users classified as Japan-referring* rather among all users (all rates among *all users* of the search engine are bigger for Russians).

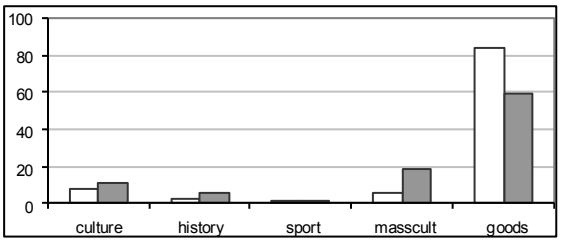

Fig. 5. Fraction (%) of users belonging to basic classes among Yandex and Excite Japan-referring users

TABLE VII. CONTINGENCY TABLE FOR 28527 YANDEX AND 4294 EXCITE JAPAN-REFERRING USERS

Yandex	culture	history	mArts	massc	goods
culture	2,100	121	26	43	58
history		810	5	14	43
martial arts			399	1	14
masscult				1659	29
goods					23903
Excite	culture	history	mArts	massc	goods
culture	494	12	1	14	5
history		266	1	6	10
martial arts			79	0	1
masscult				789	2

The combined categorization of manually and automatically classified users shows that Russian searchers demonstrate much more consumer interests. This is in accordance with the difference between U.S. "teenagers" and Russian "steadies" which is enormous among re-classified users. While we do not know age of the *Excite* users and only

doi:10.15849/icit.2015.0085

partly know distribution of the *Yandex* population by age ([11]), we suppose the revealed difference of Japan-referring searches is not explained by the age difference between Russian and U.S. searchers.

IX. CLASSES CO-RELATION

Table 7 presents a final contingency table for all users attributed to basic Japan-referring classes, i.e. Table 7 is a sum of contingency tables of users automatically attributed to basic classes (Table 5) and re-attributed users (Table 6). How do basic Japan-referring classes co-occur in a set of queries submitted by the same user? To detect closely interrelated classes we estimate the probability of a random co-occurrence of classes among independent classes of Japan-referring users. Our goal is to detect such cases of intersections of classes that infract the assumption about independency of classes.

Let n_i be the number of users attributed to the class i (diagonal elements in Table 7), $obs(i,j)$ be the number of users attributed to both classes i and j (non-diagonal elements in Table 7), and N be the number of all considered users. To measure the strength of the interrelation between two classes we use a probably $p(k \geq obs\ (i, j))$ that a number of random co-occurrences k of independent classes i and j (containing n_i and n_j users) is not less 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 smaller $p(k \geq obs\ (i, j))$, the stronger the interrelation is.

$$p(obs(i,j), n_i, n_j, N) = \sum_{k=obs(i,j)}^{k=\min(n_i, n_j)} p(k,\ n_i, n_j, N) \qquad (3)$$

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"

$$p(k, n_i, n_j, N) = \binom{n_i}{k}\binom{N-n_i}{n_j-k} \Big/ \binom{N}{n_j} = \frac{n_i!\, n_j!\,(N-n_i)!\,(N-n_j)!}{k!\,N!\,(n_i-k)!\,(n_j-k)!\,(N+k-n_i-n_j)!} \qquad (4)$$

We consider probabilities of class co-occurrence *among all* users (N_{Yandex}=890,897 users, N_{Excite}=305,360 users). This approach to class co-relation is absolutely correct but is not very expressive. Indeed, we can a priori suppose the co-relation of Japan-referring classes. As a result, we can expect that probabilities $p(k \geq obs\ (i, j))$ that a number of random co-occurrences k of independent classes i and j is not less than the observed intersection are small. Rather, when we consider probabilities of co-occurrence *among all* users, the non-small probabilities are surprising and should be of special interest as "symptoms of independence" of classes.

We also consider probabilities of class co-occurrence *only among the users attributed to Japan-referring basic classes* (N_{Yandex} = 28,527 users, N_{Excite} = 4,294 users). While this opposite "over-strong" approach is surplus (in particular, it elaborates the same ordering of probabilities), it visualizes differences between strong interclass relations. It is very expressive when we want emphasize the closest connections between classes, i.e. to differ strong co-relations (small probabilities of observed co-occurrence) from over-strong ("the smallest" probabilities). Tables 8, 9 shows estimations of probabilities $p(k \geq obs\ (i, j))$ that a number of random co-occurrences of independent classes is not less than the observed intersection for both considered sets of users.

1. "Independence criterion" (probabilities of Japan-referring class co-occurrence among *all* searchers, Table 8). While *goods* and *masscult* classes, at first sight, should be more co-related than, for example, *goods* and *martial art*, these biggest classes are practically independent in both audiences (more than 0.99 probability of a random co-occurrence). In general, *goods* class is the most independent in both audiences. The only exception presents the Russian audience for which *goods* and *history* are strong co-related classes (in contrast with independence of these classes in the U.S. audience). Co-relations of the *martial art* significantly differ among audiences: this class is more co-related with *goods*, *masscult* and partly *culture* classes in the Russian audience (0.189 vs. 0.506, 0.524 vs. 1 and 0 vs. 0.12

TABLE VIII. PROBABILITIES OF THE RANDOM CLASS CO-OCCURRENCE AMONG ALL USERS (INDEP. CRITERION)

Dataset	890,897 *Yandex* users, 305,360 *Excite* users			
Yandex	history	martial arts	masscult	goods
culture	0	**0**	0	0.42974
history		0.00004	0	**0.00003**
martial arts			**0.52473**	**0.18939**
masscult				0.99503
Excite	history	martial arts	masscult	goods
Culture	0	**0.12008**	0	0.64105
History		0.06654	0.00008	**0.90730**
martial arts			**1**	**0.50646**
masscult				0.99305

2. "Over-strong co-relation" criterion (probabilities of Japan-referring class co-occurrence among *Japan-referring* searchers, Table 9) reveals the big difference between a strong co-relation of the *culture* and *history* classes in U.S. audience and the *strongest* co-relation of these classes in the Russian audience (0.999 vs. 0).

TABLE IX. PROBABILITIES OF THE RANDOM CLASS CO-OCCURRENCE AMONG JAPAN-REFERRING USERS

Dataset	28,527 *Yandex* users, 4,294 *Excite* users			
Yandex	history	martial arts	masscult	goods
culture	**0**	0.76908	1	1
history		0.98917	1	1
martial art			1	1
masscult				1
Excite	history	martial arts	masscult	goods
culture	**0.99999**	0.99994	1	1
history		0.99390	1	1
martial art			1	1
masscult				1

doi:10.15849/icit.2015.0085

On all occasions, interdependency between classes is stronger in the *Yandex* audience and this is not an artifact.

X. CONCLUSION

We have investigated (1) differences between Russian and U.S. search images of Japan and (2) interdependency between searching for different Japan-referring classes: how frequently searchers of one Japan-referring topic also search for other topics.

1. Fractions of all classes *among all* Russian searchers are bigger than the fractions *among all* U.S. searchers. At the same time, fractions of the non-consuming classes *among Japan-referring* Russian searchers are significantly smaller than fractions *among Japan-referring* U.S. searchers. The Russian Japan-referring searchers are mainly consuming, whilst the U.S. Japan-referring searchers are much more masscult-oriented. A fraction of culture-oriented searchers is small in both audiences.

2. The Japan-referring *goods* class primary relate to goods rather than to Japan, and users submitting Japan goods queries do not frequently submit other Japan-referring queries. On the contrary, the *masscult* class is compatible with non-consuming classes and surprisingly is not compatible with *goods* in both audiences and *sport* in the Russian audience.

3. The Russian searchers submitting queries referring to Japan culture relatively frequently submit other Japan-referring queries, especially queries related to the history of Japan. Furthermore, all Japan-referring classes are more co-related in the Russian audience.

REFERENCES

[1] S. Beitzel, E. Jensen, A. Chowdhury, D. Grossman, and O. Frieder, "Hourly analysis of a very large topically categorized Web query log", in proceedings of 27th ACM SIGIR conf. on research and development in information retrieval, 2004, ACM Press, pp. 321-328.

[2] S. Beitzel, E. Jensen, A. Chowdhury,, O. Frieder, and D. Grossman, "Temporal analysis of a very large topically categorized Web query log". J. of the Association for Information Science and Technology, vol. 58, No. 2, 2007, pp. 166-178.

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

[4] B.J. Jansen, A. Spink, and T. Saracevic, "Real life, real users, and real needs: a study and analysis of user queries on the Web". Information Processing & Management, vol. 36, no. 2, 2000, pp. 207-227

[5] D. Lewandowski, "Query types and search topics of German Web search engine users", Information Services&Use, vol. 26, 2006, pp, 261-69

[6] M. Richardson., "Learning about the World through Long-Term Query Logs". ACM Trans. on the Web, vol. 2, no. 4, 2008, pp 21-27

[7] A Spink., S. Ozmutlu., H.C. Ozmutlu., and B. Jansen, "U.S. versus European Web searching trends", ACM SIGIR Forum, vol. 36, no. 2, 2002, pp. 32-38

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

[9] I. Weber, V. Garimella, and E. Borra, "Mining Web Query Logs to Analyze Political Issues", in proceedings of the WebSci 2012 conference, June 22–24, 2012, Evanston, Illinois, USA, ACM Press, 2012, 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

[11] [RU-Net] (2014 and earlier) Project "The internet in Russia/Russia on the internet".