ICIT 2015 The 7th International Conference on Information Technology
doi:10.15849/icit.2015.0084

Similar Searching, Unsimilar Clicking

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

***Abstract*— In the study, a search engine log is partitioned into IP classes that differently present audiences of free and busy searchers. It is shown that searching behavior of users from different classes is practically identical in all characteristics except their click behavior. Differences in click behavior between classes are great. Free users click more frequently than busy users while search in the same manner on other counts.**

Keywords— click behavior; IP address; query log analysis; query reformulation

I. INTRODUCTION

A lot of studies look at click behavior and its uniform models. Some studies research individual and task differences in click behavior of different users ([3], [4], [5], [9]). Web log analysis reveals different manners of real-life search and clicking, but the effect of individual/task differences and environmental factors on these differences is not clear from this analysis.

Psychological studies (e.g., [8]) investigate the effects of work environment on the worker's performance. Several papers investigate web search behavior depending on different environmental conditions (e.g., [6], [7]).

In this web log-based study, we indirectly reveal the influence of real-life environmental factors and compare search behavior of users who are mainly "stressed-out office workers" or mainly "free homebodies". We can assume that when an office worker search is aimed at his office responsibilities rather than personal goals he must perform other functions and may be overloaded by them. We presume that his searching behavior under office stressors differs from his behavior in a relaxed atmosphere. We also presume that the shift of search behavior under varying environmental conditions reflects not so much individual manners but a common environmental dependency.

We investigate not the influence of individual features such as "advanced user" or "newbie" but the influence of the real-life environmental factors on search behavior of [the same] searchers.

The opportunity of this long-scale real-life comparison is provided by the "IP classes" consideration. The IP classes of users are defined in the following way: the IP-*N* class includes all users who share the same IP address with (*N*-1) other users. There are two reasons why a search engine detects different users operating from the same IP: (1) different browsers on the same PC, (2) different PCs sharing the same proxy server. As a result, IP-1 corresponds to individual users, IP-2 may correspond either to two users who use different browsers on the same PC or to a proxy presenting two PCs, and the IP-3+ classes mainly correspond to users sharing the same proxy.

A fraction of (from-)home users among the IP-1 users appears to be bigger than among the IP-3+ (i.e. proxy) users. Of course, there is no strong relation between home/office and IP-1/IP-3+. Some of the IP-99 users may be cliff dwellers whilst an IP-1 user may be a small enterprise worker. However, the rate of cookies created on weekend (see Table 1) is a good indirect indicator of fractions of home/office users, and this rate in the IP-1 class is bigger than in IP-3+ classes.

Also, while there is no a strong relation between an office and noisy environment, and home and relaxed atmosphere (the counterexamples are obvious just as some fishes can fly and some birds cannot fly), we can assume that on average there is more stressed-out search behavior at work and more free behavior at home [8].

In the study, we consider the characteristics of search behavior of different IP classes and discover that all of them are very similar except that of the click activity. Frequencies of clicks per query or per first page of the retrieved results are about 30% bigger for the IP-1 users than for the IP-4+ users. This difference can be explained neither by the topical difference between "home" and "office" queries nor by different ways of moving across the pages of the retrieved results. Thus, the study discovers a strong non-individual difference in click behavior between "free home" users and "busy office" ones. However, we do not know whether free home squirrels/ birds visit useless sites or busy office workers ignore the necessary ones.

doi:10.15849/icit.2015.0084

II. IP CLASSES

Web users may share the same IP address either if they use different browsers on the same PC or they operate via the same proxy server. If 3+ users have the same IP address, they probably use a common proxy rather than 3+ browsers on the same PC.

IP group. An IP group is a set of users who submit queries from the same IP address.

IP class. If an IP group includes *N* users, the users are attributed to the IP*N* class. We also refer to queries, sessions, clicks, etc. of users belonging to the IP*N* class as IP*N* class queries, sessions, etc

III. RESEARCH QUESTIONS AND ASSUMPTIONS

Research Question

Are real-life working conditions stressful enough to change user search behavior? In particular, are the following features different in (differently stressed) IP classes:

- *A query level:*
  - terms per a query,
  - clicks per a query,
  - viewed pages (screens) of the retrieved results per a query,
  - clicks per viewed pages,
  - clicks on the first page of the results
- *A search session level:*
  - task sessions per a temporal session,
  - queries per a task session,
  - click behavior in different types of task sessions:
    - single-query sessions,
    - sessions with linear query modification,
    - sessions with branching query modification
  - clicks in query narrowing and broadening modifications.

Besides, if the clicks and viewed pages of the retrieved results are actually different in the "office" IP-3, 4+ classes from the "home" IP-1 class then we should test two hypotheses "*once started they will continue*": if a user from the "office class" made the first click then he will continue clicking similarly to a user from the "home" IP-1 class; if a user from the "office class" moved to the next page of the retrieved results he will continue moving through pages just as the user from the "home" class.

In the study of the environment influence on search, we do not focus on "a user as a sequence of his operations during a long observation period". "User" is an irrelevant unit, not in the sense that different people can operate from the same UID but meaning that the same person may be busy or free. Even office work has breaks and pauses and even stay-at-homers may be unexpectedly busy. Whereas *user* is an irrelevant unit, *query* is a minimum unit and *task session* (and even *temporal session*) is a convenient unit representing stable ("busy" or "free") behavior.

Assumption

We assume that:

- a fraction of home users in the IP-1 class is significantly bigger than in the IP-4+ class, and a fraction of home users in the IP*K* class is not smaller than in the IP-*(K+1)* class,
- home users are less stressed-out by their environment.

At weekends and on holiday days the degree of activity of office workers decreases more than of home users, and it can be assumed that the rate of office cookies created at weekends is less than that of home users' cookies. The data in Table 1 support the first assumption: the rate of cookies created on Sunday decreases over the IP classes, and is twice as little for the IP-4+ class than for the IP-1 class.

TABLE I. FRACTIONS OF COOKIES CREATED ON SUNDAY

IP-1	**IP-2**	**IP-3**	**IP-4+**
8.5%	8.5%	7.5%	4.4%

IV. DATASETS

We use a complete one-day dataset (March 20, 2007) drawn from the logs of the *Yandex* search engine. The dataset combines three logs (a query log, a log of the results and a click-through log) and reports queries, retrieved results and clicks on them.

The users in the dataset are represented by unique UIDs where UID is a concatenation of <*hash (IP address)*> and 10-digit <*time_of_cookie_creation*>, which allows us to detect all users who share the same IP address (the later *Yandex* public datasets formats do not provide this useful feature). According to UIDs the dataset was separated into IP-1, IP-2, IP-3, IP-4+ sets presenting corresponding IP classes.

The users who submitted more than 40 unique queries per 1-hour are eliminated as "robots". To segment a logged time series of transactions into temporal sessions a 30 min cut-off was used. Table 2 shows cleaned-up datasets of IP classes.

TABLE II. GENERAL CHARACTERISTICS OF IP CLASS DATASETS

	IP-1	**IP-2**	**IP-3**	**IP-4+**
Users	656557	98790	35022	27434
Temporal sessions	1105496	163021	56064	43464
Queries unique in a temporal session	2463767	367193	125074	95815
Task sessions	1632796	241374	82264	63374

doi:10.15849/icit.2015.0084

V. TERMS AND METHOD

In the paper, we use the notions of:

— *a query skeleton* that include only those query terms that are nouns, names, acronyms and unknowns which may be attributed to these parts of speech, a skeleton includes neither features (adjectives,) nor actions (verbs, adverbs),

— *a query narrowing* here denotes an expansion a query by additional term(s), and a *query broadening* here denotes an exclusion of one or more terms from a query (e.g., when a user submits query <*cat*> after submission of <*red cat*> he narrows the initial query, a modification of <*cat toys*> into <*toys*> is a broadening, and a modification of <*cat toys*> into <*cat food*> is neither narrowing nor broadening),

— *a temporal session* as a sequence of the user's transactions with the search engine cut from previous and successive sessions by a 30 min time gap;

— *a task session* as all queries of a task session technically defined as a connected component of the similarity graph of queries submitted during a time session. To extract task sessions (a) a matrix of term-based pairwise similarity of all unique queries of the current temporal session is filled (two queries are defined as similar if they contain common skeleton terms), (b) a transitive closure of this similarity relation is made [1]. The method [1] covers misprints in queries.

VI. FEATURES ON ALL QUERIES

Query Level Features

The general query-level features of IP classes are reported in Table 3.

TABLE III. FEATURES OF IP CLASSES. QUERY-LEVEL

	IP-1	**IP-2**	**IP-3**	**IP-4+**
terms in a query	3.02	3.01	3.01	2.98
terms in a query skeleton	2.18	2.18	2.19	2.18
viewed pages per query	1.58	1.57	1.54	1.50
clicks per query	**1.64**	**1.50**	**1.35**	**1.26**
clicks per viewed page	**1.04**	**0.96**	**0.88**	**0.84**
clicks on the first page	**1.25**	**1.15**	**1.05**	**0.99**

Users from all IP classes formulate queries identically (see "terms in a query" and "terms in a query skeleton"). The number of moves between pages of the retrieved results only slightly decreases over IP classes.

On the contrary, the click-based characteristics demonstrate big differences among classes. A number of clicks per a submitted query, a number of clicks per a viewed page of the retrieved results and a number of clicks on the first page of the retrieved results monotonously decreases over IP classes.

Session Level Features

The general session-level features of IP classes are reported in Table 4. While a number of temporal sessions during a day steadily decrease over IP classes from 1.68 at IP-1 to 1.58 at IP-4+, the average numbers of task sessions in a temporal session are similar, and the average numbers of queries in a task session are practically identical. Thus, users in the IP-3+ classes start session less often but once started they behave similarly to IP-1 users.

TABLE IV. FEATURES OF IP CLASSES. SESSION-LEVEL

	IP-1	**IP-2**	**IP-3**	**IP-4+**
temporal sessions per user	1.68	1.65	1.60	1.58
task sessions per temporal one	1.48	1.48	1.47	1.46
queries per task session	1.51	1.52	1.52	1.51
broadening query reformulation:				
in linear reformulations	11.4%	11.4%	11.4%	11.4%
in branching reformulations	23.7%	23.6%	24.1%	24%
narrowing query reformulation:				
in linear reformulations	38.3%	38.2%	38.5%	38.3%
in branching reformulations	14.2%	14.2%	15.4%	13.6%

Fractions of broadening and narrowing query reformulations in task sessions are the same across IP classes.

Different Types of Query Reformulation

According to [2] a task session may include several branches and a complex search task may be decomposed into chains which merge in the final step. However, the most significant types are linear and branching query modification. Here we use Jaccard part-of-speech-based similarity metric to detect inter-query dependencies in a task session. For example, branching search will be detected in the sequence of 3 queries <*big red cat*>, <*red cat*>, <*big cat*>, whilst a linear modification is detected in the sequence <*red cat*>, <*cats food*>, <*cats veterinary*>.

Table 5 shows practically identical rates of 3 types of temporal sessions among temporal sessions of IP classes.

TABLE V. TYPES OF TEMPORAL SESSIONS

	IP-1	**IP-2**	**IP-3**	**IP-4+**
Total	1109481	163577	56225	43569
Single-query temporal sessions	52.14%	51.35%	51.84%	51.73%
Linear temporal sessions	41.27%	41.95%	41.85%	42.09%
Non-linear temporal sessions	6.59%	6.70%	6.31%	6.18%

Table 6 shows an average number of clicks per components of single-query, linear and branching task sessions. We see two sorts of differences: (1) the difference between IP-classes and (2) replicated in all IP classes differences between different structures of search and between query position in the same structure.

doi:10.15849/icit.2015.0084

Clicks dependency on a query position in a search structure is near identical among IP classes, but belonging to IP class shifts click values. For example, Fig. 1 shows an average number of clicks in positions of 2- and 3-query linear chains of linear query modification. And IP classes' plots look like results of a parallel shift of the same plot.

TABLE VI. CLICKS PER CHAIN OF QUERY REFORMULATIONS IN 3 STRUCTURES OF SEARCH

Chain Length	IP-1	IP-2	IP-3	IP-4+
Single-query task session	1.63	1.53	1.40	1.40
Linear chain of query modification				
2-query	3.38	3.08	2.70	2.56
3-query	4.62	4.11	3.72	3.54
Branches (including a root query)				
First branch				
2-query	3.14	2.84	2.74	2.22
3-query	4.38	3.81	3.78	3.66
Last branch				
2-query	3.74	3.34	3.08	2.52
3-query	4.92	4.41	4.26	3.45

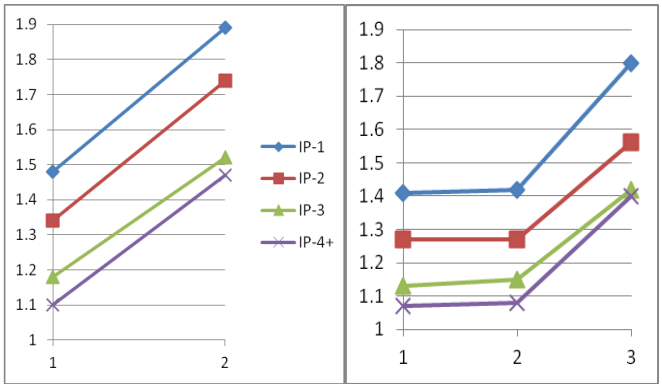

Fig. 1. Average number of clicks in positions of 2- and 3-query linear chains

Resume

No noticeable difference in search behavior is observed between classes except the only, but great difference in clicks. Any click-based query- or session-level characteristic (clicks per a query, per viewed pages, per first viewed page, clicks in any type of query modification chains) sharply decrease over IP classes.

VII. CLICKS AND VIEWED PAGES OF THE RETRIEVED RESULTS

Differences between IP Classes in Clicks and Viewed Pages Distributions.

Tables 3-6 present mean values. Let's consider clicking and moving in more detail. Distributions of clicks per query are reported in Tables 7. Also we consider "clicks per the first page of the retrieved results" and "viewed pages per query" distributions.

We compare clicks per query distributions of the IP classes on [L, 15+] intervals of clicks per query (L=0,.,14) by χ^2 test. A log-scaled Fig. 2 shows empirical and critical χ^2 values. A sample curve in a point L presents empirical χ^2 value for the interval [L,15+], and a critical curve shows critical χ^2 value at p=0.05 for degrees of freedom for <4 sets, 15-L set size>. Distributions on [L, 15+] are similar beginning with L=3.

TABLE VII. CLICKS PER QUERY DISTRIBUTIONS IN IP CLASSES

	0	1	2	3	...	15+
IP-1	1010108	650483	295077	172617	⋯	16671
IP-2	165656	91237	41113	23657	⋯	2195
IP-3	61506	29424	12994	7501	⋯	623
IP-4+	49190	22161	9470	5297	⋯	420

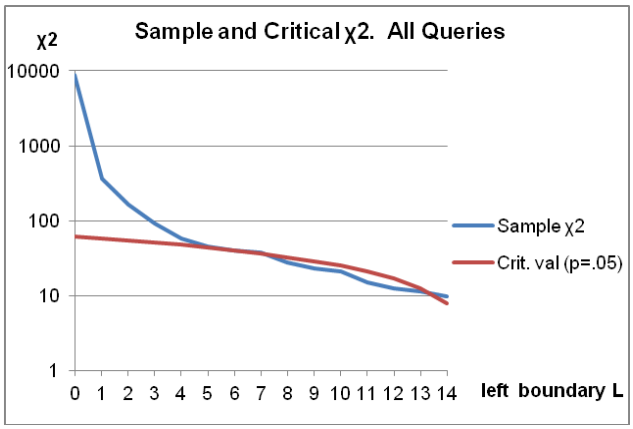

Fig. 2. Sample χ^2 ([L,15+]) and critical χ^2(p=0.05) as functions of the left boundary L=0,..,14

As Fig. 2 shows, click behaviors of IP classes strongly differ on the first clicks and are similar for subsequent clicks. Distributions of clicks on the first page of the results demonstrate just the same incoherence in the first clicks and coherence in the following clicks. At the same time, distributions of viewed pages are similar on all intervals [L, 9+], where L=1,,,,8.

Description in Terms of Transition Probabilities.

Let's return to the hypotheses "once started they will continue". As regards to click behavior, it means that empirical probabilities of the first click may differ between classes but probabilities of transition to the any following clicks are very similar.

Empirical transition probabilities of the next click on the results of the query and transition probabilities of the next click on the first page of the results are shown in Fig. 3, where

doi:10.15849/icit.2015.0084

p(c) is an empirical probability of the next click when *c*-1 clicks are made.

As Fig. 3 shows, probabilities for different IP classes greatly differ for the first click and are very similar for the further clicks. The trigger "once started" hypothesis is true for click behavior. A mainly stressful environment is not totally and continuously stressful and even a normally busy worker has some free time.

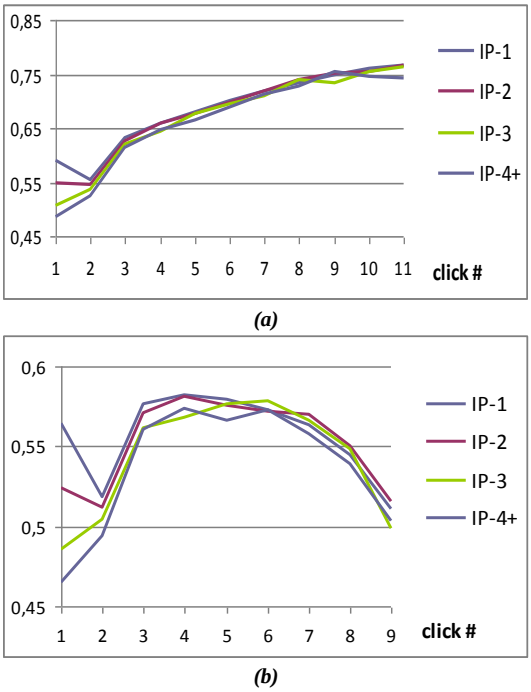

Fig. 3. Probabilities of transition to click# *(a)* on all viewed pages and *(b)* on the first page

Empirical transition probabilities of moving to the another page of the retrieved results are shown in Fig. 4, where *p(move#)* is an empirical probability of *move#*-th move to the another page (i.e. after viewing *move#* retrieved pages). The probabilities are similar among all IP classes and moving across the retrieved results does not depend on the IP class. This is not a surprise since viewed pages per query distributions are similar among classes.

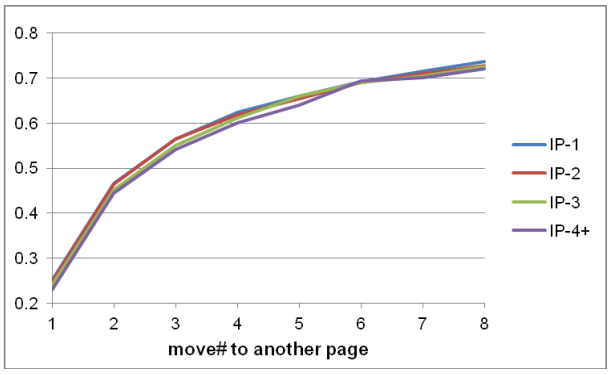

Fig. 4. Probability of move# to the another page of the retrieved results

VIII. FEATURES IN TOPIC DIMENSIONS

Different search topics lead to different search behavior. Topics may be differently presented in queries of different IP classes and between-class differences may be the result of the differences in the topic occurrence. To check a uniformity of the discovered dependencies we investigate search behavior on three topics.

We consider two topics – "*Travel*" and "*Education*". We choose 20 topic-specific terms for the *Travel* topic and 10 topic-specific terms for *Education*. If a query contains a topic term, this query is considered as the "topical" one. If a task session contains a topical query than the session is considered as a topical session. Table 8 shows the number of topical queries, topical sessions and queries belonging to the sessions in IP classes.

TABLE VIII. TOPICAL QUERIES AND QUERIES IN TOPICAL TASK SESSIONS

	IP-1	**IP-2**	**IP-3**	**IP-4+**
Travel queries	45689	6246	2103	1556
Education queries	52871	7548	2434	1766
Queries in Travel sessions	68412	9399	3187	2416
Queries in Education sessions	79499	11395	3671	2624

The results in Table 9 show the same dependencies that were observed on all queries: (1) similar values of queries per task session, (2) a slightly decreasing number of viewed pages and (3) a sharply decreased number of clicks over IP classes.

TABLE IX. QUERIES IN ALL AND TOPICAL TASK SESSIONS

	IP-1	**IP-2**	**IP-3**	**IP-4+**
Queries per task session in:				
All sessions	1.51	1.52	1.52	1.51
Travel sessions	2.23	2.25	2.25	2.36
Education sessions	2.45	2.45	2.45	2.37
Clicks per query in:				
All sessions	1.64	1.50	1.35	1.26
Travel sessions	1.77	1.65	1.62	1.35
Education sessions	1.95	1.72	1.56	1.61
Viewed pages per query in:				
All sessions	1.58	1.57	1.54	1.50
Travel sessions	1.59	1.57	1.56	1.57
Education sessions	1.74	1.76	1.66	1.68
Clicks per viewed pages in:				
All sessions	1.04	0.96	0.88	0.84
Travel sessions	1.12	1.05	1.03	0.86
Education sessions	1.08	0.98	0.91	0.96

doi:10.15849/icit.2015.0084

	IP-1	IP-2	IP-3	IP-4+
Clicks on the first page in:				
All sessions	1.25	1.15	1.05	0.99
Travel sessions	1.39	1.29	1.27	1.09
Education sessions	1.40	1.26	1.18	1.15

IX. CONCLUSION

The empirical study of IP classes gives answers to the questions about the influence of the environmental stressors on real-life search and click behavior of the Web user:

- a query formulation/reformulation does not vary across IP classes and does not depend on real-life stressors,
- the number of queries in task sessions, the number of task sessions in a temporal session and fractions of different types of task sessions do not vary across IP classes and do not depend on real-life stressors,
- the number of the viewed pages of the retrieved results decreases slightly over IP classes,
- the number of clicks (per query, per a viewed page of the retrieved results, on the first viewed page) and in all types of task sessions decreases over classes along with a fraction of free users.

The first click plays the trigger role: if a user in the "office IP class" started clicking he will continue clicking similarly to a "home" IP-1 class user (Fig. 3). 0-click behavior frequent in IP-3+ classes seems to be a bit strange: "a busy user" has no time to visit the retrieved pages but has enough time to move through 2+ pages of the retrieved results. Real-life environment stressors do not make searchers change query (re)formulation or significantly change the number of viewed pages but urge them to decrease the number of clicks.

The rates of IP classes have changed since 2007 and the fraction of IP-4+ class must have increased. However, it is not the IP classes which matter as such but the fact that since 2007 the share of home users (retired people, home workers), i.e. the share of "free users" with their non-stressed behavioral patterns has grown.

REFERENCES

[1] N. Buzikashvili. "Automatic Task Detection in the Web Logs and Analysis of Multitasking", 9th Conferece ib Asian Digital Libraries, 2006, Kyoto, Japan, 2006, LNCS 4312, Springer-Verlag, 2006, pp. 131-140.

[2] N. Buzikashvili. "Structure of the Web Searcher's Query Modifications: Sequential, Branching, Merging, Re-Merging and Non-Sequential Execution". Paper 569, 5th Int Conferencr on Inf/ Trchnology, 2011, Al Zaytoonah University, Amman, Jordan, 2011.

[3] C. Eickhoff, J. Teevan, R. White, and S. Dumais.. "Lessons from the journey: a query log analysis of within-session learning", WSDM'14, NY, USA ACM Press, ,2014, pp. 223-232.

[4] J. Jiang, D. He, and J. Allan. "Searching, Browsing, and Clicking in a Search Session: Changes in User Behavior by Task and Over Time", in proceedings of 37th ACM SIGIR conf. on research and development in information retrieval, ,Gold Coast,Austalia,2014, ACM Press, pp. 607-616.

[5] B. Hu, Y. Zhang, W. Chen, G. Wang, and Q. Yang. "Characterizing search intent diversity into click models", in proceedings of 20th World Wide Web Conference, Hyderabad, India. ACM Press, 2011, pp. 17-26.

[6] S.Y. Rieh. On the Web at Home: "Information Seeking and Web Searching in the Home Environment", J. of the Association for Information Science and Technology,vol. 55, no. 8, 2004, pp. 743–753.

[7] S. Sushmita, H. Joho, M. Lalmas, and R. Villa. "Factors affecting click-through behavior in aggregated search interfaces", in proceedings of 19th Conf. Information and Knowledge Managemen, Toronto, 2010, ACM Press, 2010, pp. 519-528.

[8] J.C. Vischer. "The effects of the physical environment on job performance: towards a theoretical model of workspace stress", Stress and Health, vol. 23, no 3, 2007, Wiley, pp. 175–184.

[9] R. White, S. Dumais, and J. Teevan. "Characterizing the influence of domain expertise on web search behavior", 2nd ACM Conf. on Web Search and Data Mining, Barcelona, Spain. ACM Press, 2009, pp. 132–141.