Find It If You Can: A Game for Modeling Different Types of Web Search Success Using Interaction Data

Mikhail Ageev*
Research Computing Center
Moscow State University
ageev@mail.cir.ru

Qi Guo Dmitry Lagun Eugene Agichtein
Mathematics and Computer Science Dept.
Emory University
{qguo3,dlagun,eugene}@mathcs.emory.edu

ABSTRACT

A better understanding of strategies and behavior of successful searchers is crucial for improving the experience of all searchers. However, research of search behavior has been struggling with the tension between the relatively small-scale, but controlled lab studies, and the large-scale log-based studies where the searcher intent and many other important factors have to be inferred. We present our solution for performing controlled, yet realistic, scalable, and reproducible studies of searcher behavior. We focus on difficult informational tasks, which tend to frustrate many users of the current web search technology. First, we propose a principled formalization of different types of "success" for informational search, which encapsulate and sharpen previously proposed models. Second, we present a scalable game-like infrastructure for crowdsourcing search behavior studies, specifically targeted towards capturing and evaluating successful search strategies on informational tasks with known intent. Third, we report our analysis of search success using these data, which confirm and extends previous findings. Finally, we demonstrate that our model can predict search success more effectively than the existing state-of-the-art methods, on both our data and on a different set of log data collected from regular search engine sessions. Together, our search success models, the data collection infrastructure, and the associated behavior analysis techniques, significantly advance the study of success in web search.

Primary Classification: H.3.3 [**Information Storage and Retrieval Languages**]: Information Search and Retrieval; *Search process*

General Terms: Experimentation

Keywords: User studies, query log analysis, web search success.

* work done while visiting Emory University.

SIGIR'11, July 24–28, 2011, Beijing, China.

1. INTRODUCTION

"Knowledge must come through action." - Sophocles

Searching the web is never a goal in itself: the searchers typically want to obtain the required information quickly and with minimum effort. For many common search tasks, search engines excel at returning the needed information in the first few results, and sometimes directly in the search engine result page (SERP) itself. However, there are difficult search tasks that require significant effort and ingenuity – where only some of the searchers succeed while the majority fails. What makes some searchers successful? What are the key failure modes that tend to result in failed searches? What patterns of searcher behavior can reliably identify a successful search session?

We propose a principled search framework designed to investigate these questions, and present a novel competition-based data collection methodology, to enable empirical study of the different paths to search success. Our approach is based on enticing participants to compete in a game-like setting, to find answers to real informational questions, while tracking the resulting search behavior.

This approach has a number of advantages over previously reported search evaluation methods: i) the information needs are real, well-defined questions selected from community forums such as wiki.answers.com and Yahoo! Answers. ii) the goals (needed information) are known to both the searchers and the assessors, and are well-defined, allowing for objective measures of success; and iii) sufficient amount and diversity of search behavior can be acquired for difficult queries (which are relatively rare among all queries submitted to search engines), thus enabling in-depth study of the behavior characteristics for these difficult and rare queries that were not previously available through passive log analysis.

In summary, our contributions include:

- A flexible and general informational search success model for in-depth analysis of search success and failure for different definitions of success (Section 2).
- A scalable infrastructure for collecting realistic search interaction data with objective success criteria (Section 3).
- Effective machine learning-based techniques for predicting and analyzing different types of search success (Sections 4-6).
- Source code and data are shared with the research community.

Next, we present our search success model, which allows us to formulate hypotheses that motivate our study design, and the experiments described in the rest of the paper.

2. SEARCH SUCCESS MODEL

To better analyze the search process, we propose a simple, yet powerful four-stage *QRAV* (*Q*uery-*R*esult-*A*nswer-*V*erification) model of an informational search success. This model is primarily geared towards analyzing and describing searches with specific, direct information needs, such as those phrased as factoid questions.

We conceptually divide the process of successfully answering factual queries into several parts. First, the user should correctly understand the question and issue a relevant query (Q for *Query formulation*). If the query retrieves at least one target document in the top 10 results, we call it a *Good Query*. Then, the user has to find the correct result on a search engine result page (SERP), and click on it to examine the document (R for *Result identification*). If a result document contains the correct answer, we call it a *Good URL*, indicated by R^+. Furthermore, we can consider how many clicks away this document was from the SERP, indicated by the subscript. For example, if a document containing a correct answer was the last in the search session, we indicate it as R_L^+. The next step is *extracting* an answer from a document (A for *Answer extraction*). Finally, the answer is *verified* that it correctly answers the question and is in fact supported by the document (V for *Verification of the answer*).

Thus, the final success of a user in finding an answer to a question depends on successfully performing each stage in the *QRAV* model, represented as $Q^+R^+A^+V^+$. In a case where a user issues a good query and clicks on a good document, but submits an incorrect answer, the outcome would be represented as $Q^+R^+A^+V^-$. Finally, if the success in a particular stage is unknown (or not considered), the "?" mark is used. The model is illustrated graphically in Figure 1, where the states in the process are represented by circles and the arrows represent possible state transitions. Note that some transitions in the model are not possible, e.g., by definition it is not possible to directly go from a bad query (Q^-) to a good result (R^+).

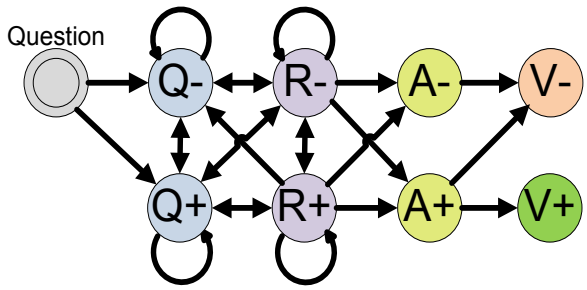

Figure 1: Possible state transitions in QRAV model

Using our *QRAV* model we can describe (and estimate) the success factors at each stage of the search process, and naturally represent previously posed definitions of search success:

- $Q^+R_*^+A^+V^+$: The correct answer was found and validated to be supported by a good document — it is a search success in the strictest sense, most similar to the definition of a correct answer in a TREC question answering track [27].
- $Q^+R_*^+A^+V^?$: An answer was found on a good result, and submitted — which means that the participant was satisfied with the session and believed that she found an answer (but the answer could still be incorrect). This definition of success matches the definition of Aula et al. [2], where the users submitted the answer to a difficult search task, but the answer was not validated.
- $Q^+R_*^+A^?V^?$: A good URL was found — the user visited a relevant page, but did not necessary extract the correct answer. This definition matches the model based on analysis of search engine click logs by independent assessors as in [11].
- $Q^?R_L^+A^?V^?$: A good URL was found *and it was the last in the search session*. This follows the ideas of marginal document *utility* in a search session [9] - after viewing the last document in the search session, the user is satisfied and stops searching.

Having defined our measures of success, we can now analyze the sets of actions (behavior) that correlate with each of the above definitions of success. In other words, at each stage of the process, searchers perform actions, some of which are indicative of a success or failure, or a continuation at one of the *QRAV* stages above. Our study, described next, was designed with the goal of being able to validate, and isolate the success or failure of the searcher at each stage of the process. Using this model we can also analyze the corresponding behavioral clues for *predicting* the success or failure at each stage of the search process.

3. ACQUIRING SEARCH BEHAVIOR DATA

The overall design of the study was modeled on a game, more precisely as a *search competition*: the participants played a game consisting of 10 search tasks (questions) to solve, with a timer displaying the number of seconds remaining, shown to the subject (see Figure 2). The stated goal of the game was to submit the highest possible number of correct answers within the allotted time. Overall, four game rounds were used in this study, with participants recruited and scored separately for each round. The top "players" in each round (typically, those successfully posting a correct answer to 7 or 8 out of the 10 questions) received a bonus payment.

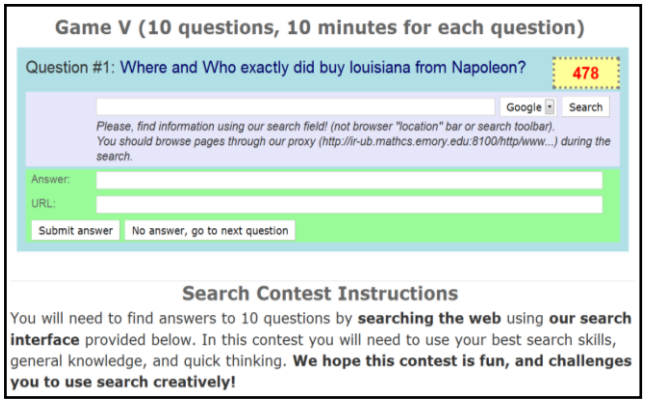

Figure 2. An example search game interface, which has the question, the search query window, and a dropdown box for choosing a search engine to use (Google, Yahoo! Search, or Bing). When the answer is found, the participant submits it together with the supporting URL. The query result page is opened in new tab, allowing natural querying and browsing.

3.1 Search Tasks: Informational Questions

The search tasks were selected from community question answering sites such as wiki.answers.com and Yahoo! Answers by the researchers. The criteria used were that the question should be clearly stated, had a clear answer, and that finding this answer was not a trivial task, that is, the answer was not retrieved simply by submitting the question verbatim to Google, Bing, or Yahoo! Search engines. Overall, 40 such questions were selected, posing (as it turned out) greatly varying levels of difficulty for participants. These questions were randomly split into four game rounds of 10 questions each. A list of questions for an example game is provided in Figure 3.

How many Swedes speak English as a percentage?
When the jominy test was invented?
Which metals float on water?
What is oxygen partial pressure at 5000 feet?
How many Argentine soldiers died in falklands islands war?
What ingredients in cough medicine make you hallucinate?
How do you say welcome in kashmiri?
Am I allowed to carry a parachute onboard as a hand luggage?
What animal is smaller than a bear but it eats a plant called bearberry?
What is the highest peak in western hemisphere?

Figure 3. An list of questions for an example game (the original grammar and spelling are preserved).

3.2 Participants and Study Procedure

Participants were recruited through the Amazon Mechanical Turk (MTurk) website. As a first step, the workers had to solve a *ReCaptcha*[1] puzzle to verify that they are human and not an automated "bot". Then, each participant was directed to the task instructions (describing the game rules), and then clicked on a link to start a game. At that point, a search interface as in Figure 2 appeared, which was to be used for submitting all search queries. The search results appeared either below the search box, or in a different tab (depending on user's normal search preferences), in the original search engine result format. To generate these search results, the queries were submitted (and logged) through our proxy server, which then retrieved and logged the search engine responses and displayed them to the user in the original format.

In order to capture all of the participants' search actions, they were instructed to use only our search interface (and not a separate browser window). After the searcher decided that they found the answer, they were instructed to copy and paste the answer text from the document, together with the supporting URL, into the corresponding fields in the game interface. Each search session (for one question) was completed by either submitting an answer or clicking the "skip question" button to pass to the next question.

Every participant who followed the directions (i.e., used our search interface and submitted at least one answer, regardless of whether the answer was correct or not) was guaranteed the base participation payment of $1. Additionally, the "best" players (i.e., those with the most correct answers found as assessed by the researchers) were promised an additional $1 bonus to stimulate a competitive "feel" of the game. Based on a pilot study used to improve and validate the game design, we found that the bonus feature was crucial to motivate the workers to persist through difficult search tasks. We also found that providing the bonus option dramatically increased the rate of task completion and reduced the rate of "junk" submissions.

[1] www.recaptcha.com

3.3 System Implementation

Our UFindIt game interface was built using Python, Django framework and Apache Web server. The Apache proxy functionality was used by configuring the modules *mod_proxy*, *mod_proxy_html*, and *mod_sed* so that the users could search and browse the Web in a usual way, while the URLs in the html links were automatically replaced. For example, instead of http://site.com/some?params, the HTML links were automatically re-written to request the URL through our proxy.

The behavioral (interaction) events were logged by our proxy and written in the server log. For each user-question session we extracted all the queries, URLs and contents of the visited pages, clicks on search engine result pages, click positions, browsing trails, and the times of all the events.

Our experimental infrastructure and the data used in this study are available for the research community at: http://ir-ub.mathcs.emory.edu/uFindIt/.

4. PREDICTING SEARCH SUCCESS

This section describes the models, features, and algorithms we use for analyzing and predicting the success of searchers.

4.1 Algorithms

Markov Model (MML+Time): As our baseline state-of-the-art model we adapt the Markov Model approach introduced in the reference [11]. As in the original model, the states are the types of visited page — "Q" for SERPs, "R_1" for pages clicked from SERP — or "E" for end-of-session. The only difference from original model is that our data does not contain sponsored-search clicks and search engine-specific links like "related search", but we have information about pages visited from hyperlinks — those pages correspond to additional model state "$R_{>1}$".

The transitions of Markov Model are events of new page visit. Given the session success $B \in \{1,0\}$ (success or fail), the transition probability between two states $s_i, s_i \in \{Q, R_1, R_{>1}, E\}$, is estimated on the training set:

$$P(s_i \to s_j, \Delta t|B) = \frac{N_{s_i,s_j,B}}{N_{s_i,B}} \cdot \Gamma(\Delta t, k, \theta) \quad (1)$$

where $N_{s_i,s_j,B}$ is a frequency of transitions $s_i \to s_j$ in sessions with a given result B, $N_{s_i,B}$ is a frequency of state s_i in those sessions, and Δt is a time delta between events s_i and s_j. The model in reference [11] assumes that Δt has the Gamma distribution $\Gamma(\Delta t, k, \theta)$ with parameters k and θ, estimated from the training set. Parameters k and θ also depend on s_i, s_j, and B.

The trained Markov Model is used to predict session success from the search behavior data. For the sequence of states with known time deltas $S=s_0 \xrightarrow{\Delta t_1} s_1 \xrightarrow{\Delta t_2} \dots \xrightarrow{\Delta t_n} s_n$ the log likelihood of success and failure is estimated as:

$$LL_B(S) = \sum_{i=1}^{n} \log P(s_i \to s_j, \Delta t|B) \quad (2)$$

and the session success is defined as:

$$Pred(B) = \begin{cases} 1 & \text{if } LL_1(S) \geq LL_0(S) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

We tested the performance of the Markov Model, with and without the time delta distribution features, and the experiments confirmed that incorporating time delta distribution indeed improves performance. This agrees with the results described in reference [11] and validates our implementation.

Conditional Random Fields (CRF): we use an extension of the Markov Model approach above, by adapting the CRF model [20] for our task. The benefit of CRF is that it allows us to augment the Markov Model with additional *search behavior features*, derived from previous works in references [10, 11, 29], and described next. We use the Mallet[2] implementation of CRF, freely available for research.

A CRF allows us to define a conditional probability $P(y|x,\lambda)$ over the hidden state sequences $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ given a particular observation sequence of n page views $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, and the CRF parameters (λ) that are estimated at training. At training time, the hidden state of an observation (i.e., a page view) can be assigned a "+" or "-" label, depending on whether the user was successful in the task. Alternatively, we can also assign "Q^+", "Q^-", "R^+", "R^-" labels to the intermediate stages in the session to allow more fine-grained modeling. At test time, given an observed page view sequence $\boldsymbol{x'}$, the most likely state sequence $\boldsymbol{y'}$ can be inferred by maximizing the conditional probability $P(\boldsymbol{y'} \mid \boldsymbol{x'}, \boldsymbol{\lambda})$ using the formula:

$$P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\lambda}) = \frac{1}{Z(x)} \exp\left(\sum_j \lambda_j F_j(\boldsymbol{y},\boldsymbol{x})\right) \quad (4)$$

where $Z(\boldsymbol{x})$ is a normalization factor and $F_j(\boldsymbol{y},\boldsymbol{x})$ is the j^{th} feature function, which could be either a state feature function or a transition feature function.

An example configuration is shown in Figure 4. Each observation corresponds to a page view (i.e., either a search engine result page, $Query_i$; or the landing page of a clicked result, $Result_i$), and is represented by a *vector of features* introduced next, (Table 1) such as dwell time on the page, query length in words, and number of queries in a session.

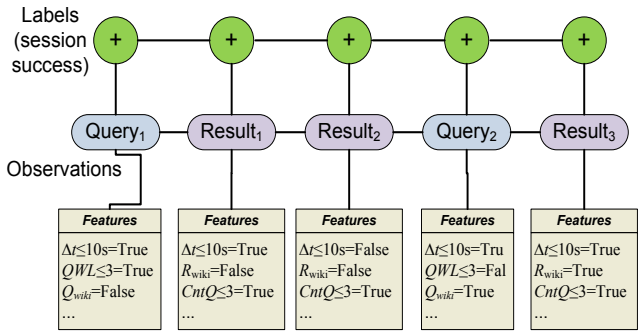

Figure 4. CRF implementation of session-level model. The labels represent overall session success; the observations at each step in the sequence are the features in Table 1.

To experiment with the tradeoff between the precision and recall, we use the *marginal probability* of the last hidden state y_n="+" as the classification confidence, since the last state indicates whether a searcher is successful or not across all potential CRF configurations. The marginal probability is computed by summing over the probabilities of all labeled sequences Y^+ that end with label "+" in their last states, according to the following formula:

$$conf = \sum_{y \in Y^+} P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\lambda}) \quad (5)$$

For our experiments we used the Mallet implementation of CRF, which allows only nominal features. Therefore, we discretize the numeric features using increasingly large thresholds. For numeric features, the used discretization thresholds are shown in the column "Bins". The complete set of behavioral features used for the CRF model is reported in Table 1. We also used aggregated behavioral features for analysis of user success, as described in the next section. The aggregation function is shown in the column "User" for the features that have a reasonable interpretation for describing an individual searcher.

4.2 Search Behavior Representation: Features

We represent searcher behavior by adapting and extending the features introduced in previous studies. Specifically, we use the browsing features from [29]. Additional features, such as session duration and number of viewed pages are adapted from [10]. We also added a feature for the average page trail length.

Feature	*Description*	*CRF Bins*	*User-level Aggregation*
state	Type of visited page $s \in \{Q, R_1, R_{>1}, E\}$, see the Markov		
Δt	Time delta between previous state and current state	≤3s ≤10s ≤30s	$\sum \Delta t$
Q_{engine}	One of {google, bing, yahoo}		
$Q_{abandoned}$	True if there no clicks for the query		
QWL	Query word length	≤3	Avg
Q_{wiki}	True if wikipedia.org is on SERP		
Q_{ADV}	True if the query use advanced query syntax. (i.e., queries that use search operators – quotes, "+" operator, and field operators like "site:" and "allintext:".)		Avg Count
Q_{DT}	Query Deliberation Time - minimum time delta between query and first click		Avg
R_{wiki}	True if visited page is on wikipedia.org		
$R_{Q\_serp\_pos}$	Position of SERP click	≤2 <5	Avg
R_{trail}	Length of trail from search engine result page, defined as the number of clicks from SERP	≤1	Avg
ref_{serp} ref_{start}	True if visited page was clicked from the SERP or from the start of a game (these features are extracted from HTTP Referer header, and could catch some patterns of non-linear browsing, when user uses several browser tabs)		
Session-level aggregates			
CntQ *CntR*	Count of queries and pages in the session	≤1 ≤3	Avg
QPS	$QPS = \frac{CntQ}{\sum \Delta t}$ — average number of Queries submitted by a user Per Second		Avg
CPQ	$CPQ = \frac{CntR}{CntQ}$ — average number of result Clicks Per Query		Avg

Table 1: Behavior features used for CRF. "Q*" features are defined only for SERPs (if the state=Q), "R*" features are defined only for non-SERP pages. Discretization thresholds are shown in the "Bins" column. For the features used in the searcher behavior analysis, the aggregation function is shown in the last column.

[2] Available at: http://mallet.cs.umass.edu/

For the analysis we used only the features that could be reasonably matched to user's search skills or expertise. For example, we hypothesized that highly successful searchers view more results (i.e., *CntR* is higher), use more advanced syntax (i.e., Q_{ADV} is higher), and perform search faster (i.e., the duration of session in seconds — $\sum \Delta t$ is lower). As will be shown in section 5.2, the first two hypotheses are supported by our data, but the third is not. To compute these features, the base feature values in Table 1 are aggregated for each participant according to the rules presented in the last column, and then averaged over all the sessions that the user has completed.

5. RESULTS AND DISCUSSION

This section presents the results of analyzing and predicting search success. First, we describe the participant data and provide descriptive statistics. Then, we contrast the behavior patterns of successful vs. unsuccessful searchers. Then, we analyze the behavior patterns associated with search task difficulty in order to automatically predict search success for the various definitions of success in our model.

5.1 Participant and Search Behavior Data

A total of 200 MTurk participants finished at least one of the game rounds. The user sessions were both automatically and manually checked to detect violations of game rules. For example, some users did not use our search interface, or used unsupported browsers, despite being warned not to do so in the task instructions. We also filtered out the users who did not answer even the easy, effectively trivial questions, as it indicated either poor understanding of the game rules, or an attempt to make a quick buck without effort. After this filtering, 159 users (79.5%) remained in the dataset. Our data for these users consists of 1487[3] search sessions, distributed among 40 distinct questions in 4 game rounds. All these 159 users were paid the base $1 payment. The top 25% of the searchers (ranked by the number of correct answers submitted) were paid an additional $1 bonus. *In total, the user payments cost less than $250.*

Overall, the study participants enjoyed the task, and provided positive comments indicating that the game was interesting (at least, in comparison to the other MTurk tasks), and challenging enough. A sample of users' comments is provided in Figure 5.

"The game was somewhat hard, but I tried my best"

"I enjoyed the search task. But in a few cases defining the proper search term took some time as it had to be refined after the first search. And for that count I may have missed one or two."

"That was pretty interesting and worked well. I felt like I was able to get answers to most of the questions pretty easily."

"Only one was really hard, the Larry Bird one!"

"Little confusing at first… search engines were not very helpful on most without some in depth searching"

Figure 5: Optional feedback submitted by MTurk workers.

Overall, there were from 30 to 50 valid search sessions collected for each question. For each session we extracted the browsing events from our proxy log. Each session was finished by either submitting an answer (87% of sessions), or by passing to the next question. The submitted answers were manually marked as either correct (65% of all sessions) or incorrect. An answer was considered correct if the page of the submitted URL indeed contained the submitted answer. There were 4382 search engine queries in our log data, and 14676 page visits. This data is available at (http://ir-ub.mathcs.emory.edu/uFindIt/).

Additional data statistics, including the overall average session times, number of actions, average query length, and others are reported in Table 2. These statistics largely agree with the published statistics from a similar study of web search success reported by Aula et al. [2]. The small quantitative differences can be expected, as our study focuses on relatively difficult information gathering tasks.

	All sessions	*Successful sessions*	*Unsuccessful sessions*
Count	1487	971	518
Average duration, sec. ($\sum \Delta t$)	215 (223)	182 (176)	276 (384)
Average number of query terms/query (*QWL*)	6.0 (4.8)	5.8 (4.7)	6.6 (5.1)
Average number of queries per session (*CntQ*)	2.9 (6.7)	2.3 (5.0)	4.0 (12.4)
Ratio of queries with operators (Q_{ADV})	0.05 (0.07)	0.05 (0.06)	0.05 (0.13)

Table 2: Descriptive statistics for search sessions collected with the UFindIt game. The corresponding statistics from reference [2] are shown in parentheses.

5.2 Analysis of Successful Searchers

5.2.1 Analysis of Behavioral Features

We use the number of correct answers as a measure of a user's success level. We divide users into three groups, according to different levels of success: HIGH level of success, if the user submitted correct answers to eight or more questions; MEDIUM level for those able to submit correct answers to five to seven questions; and LOW for those submitting correct answers to fewer than five questions.

Feat	User Success Level			LOW vs. HIGH	
	LOW	MEDIUM	HIGH	Effect size d	Power (1–β)
count	34	66	59		
$\sum \Delta t$	216	223	205	0.14	0.15
CntQ	**2.33**	**2.88**	**3.15**	**0.48**	**0.70**
QPS	0.013	0.015	0.016	0.39	0.54
QWL	**7.07**	**5.90**	**5.23**	**1.04**	**1.00**
$R_{Q\_serp\_pos}$	3.58	3.62	3.38	0.19	0.22
CPQ	**0.87**	**1.03**	**1.09**	**0.49**	**0.71**
Q_{DT}	**25**	**15**	**11**	**0.89**	**0.99**
Q_{ADV}	**0.001**	**0.036**	**0.050**	**0.64**	**0.90**
CntR	**2.80**	**3.81**	**3.91**	**0.73**	**0.95**
R_{trail}	**1.13**	**1.20**	**1.28**	**0.55**	**0.79**

Table 3: Comparisons of user behavioral features for different user success levels. Features with significant effect size are shown in bold; observations supported by high statistical power are shaded.

[3] After filtering a small number of empty or broken sessions due to proxy failure or other logging problems.

Specifically, we explore the relationship between searcher success level and the different features of searching behavior. First, we compare the search behavioral features for different success level groups by the mean values. Secondly, we perform independent measures Wilcoxon-Mann-Whitney test for assessing whether the means for LOW and HIGH success-level groups of users differ significantly. We applied Cohen's d-test to determine the effect size for each feature, and evaluate statistical power (1 - β) using at (α = 0.05). Table 3 reports the mean, effect size, and statistical power for browsing features.

Our analysis shows that, as compared to users with LOW success, users with HIGH differ as follows:

- Issue more queries for each question, view more pages, analyze more documents per query, and browse deeper from search engine result pages.
- Issue shorter queries.
- Analyze search result page faster and click faster.
- Use advanced search syntax more frequently.

All three groups of users made a similar effort in terms of amount of working time ($\sum \Delta t$), but successful users exhibited better web search skills, indicated by greater speed.

Our study provides an interesting complement to a larger, but passive log-based study performed by White and Morris [29]. We confirmed empirically their hypothesis that "advanced" users (who use advanced query syntax) indeed are more successful in Web search. Our data agreed with their findings in that HIGH users issued more queries (*CntQ*), but disagreed with the data in [29] on the difference in query length (*QWL*), and the ratio of clicks to queries (*CPQ*). The advanced users analyzed in reference [29] issue longer queries, and perform fewer clicks than LOW users, whereas this difference was not observed in our data. We hypothesize that the disagreements are due to more difficult search tasks in our data, compared to the more common tasks found in a regular search engine log. Other features (such as query deliberation time Q_{DT}, and the number of queries per second QPS) studied both in our research and in [29] are not statistically significant according to both studies.

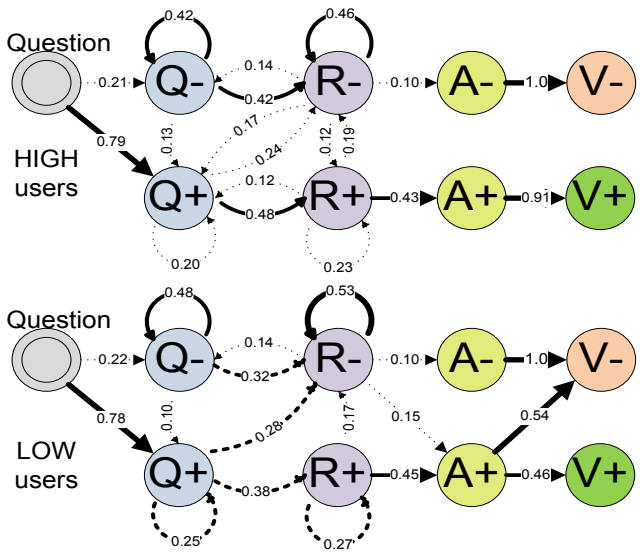

Figure 6: State transition probabilities estimated for users with HIGH and LOW search success ratings. The transition probabilities are indicated by the line weight; transitions with probabilities of less than 0.1 are not shown.

5.3 Analysis of Search Paths

A similar statistical analysis was performed for the search session success, with different meanings of success following our QRAV model. For this study we approximate the definition of *Good URL* (R^+) and *Good Query* (Q^+) by the partial assessments extracted from our data in the following way: if a URL was submitted by any user as a correct answer, than we consider it to be a *Good URL*, and if a query contains any *Good URL*s on the SERP, than it is considered to be a Good Query. As we have many answers for each question (from 30-50 users), we believe this approach results in a good approximation.

Our data shows that HIGH users, compared to LOW users, had higher scores on different stages of QRAV: they are more likely to issue good query during the session (Q^+), click good documents on result pages (R^+), and submit an answer (A^+).

But the significance of different success levels is non-uniform for different levels of success. LOW users issued good queries in 87% of sessions versus 95% for HIGH users; however, for the LOW users only 42% of the good queries resulted in correct answers as compared to 89% for the HIGH users.

To gain an intuition for the differences in behavior between highly successful searchers (HIGH) and the less successful ones (LOW), consider Figure 6, which summarizes the most likely search paths taken by these user groups, and omits transitions with probability less than 0.1. Interestingly, both HIGH and LOW users tend to submit good queries on the first try (with probabilities of 0.79 and 0.78, respectively). However, LOW users are less likely than HIGH users to *recognize* relevant results (0.38 vs. 0.48). LOW are also more likely to be "stuck" examining bad results (0.53 vs. 0.46), and are very unlikely to transition from bad results to good results ($P<0.1$) or to a good query reformulation ($P<0.1$). In contrast, HIGH users eventually move on from bad results to a good query reformulation ($P=0.17$) or directly to good results ($P=0.12$).

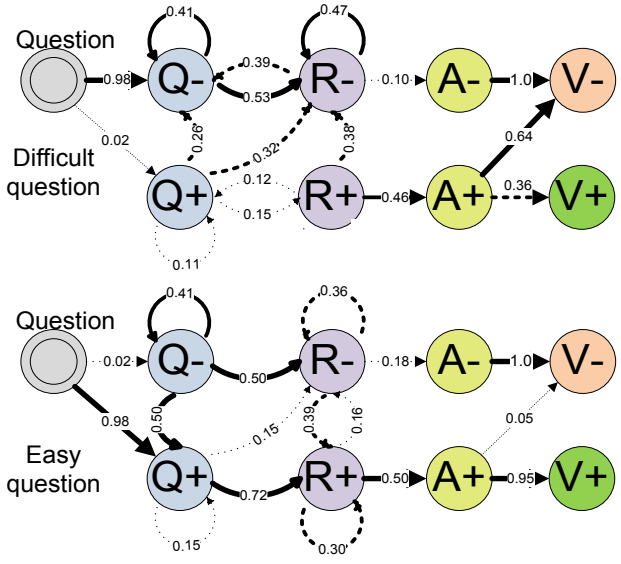

Figure 7: State transition probabilities estimated for all users, for an example difficult and easy question, respectively. The transition probability is indicated by the thickness and number on the transition arcs.

Interesting behavioral differences also exist for "easy" and "difficult" questions (those answered by > 75% and < 25% of participants, respectively). An example of a difficult question is "*When was the jominy test invented?*" For this question, only 8 participants submitted correct answers out of the 44 who attempted this question. The state transitions diagram for this question is shown in Figure 7. It appears that the most difficult aspect of this question is formulating a good search query, with only 2% of the searchers able to formulate a successful query on the first try, and only 29% eventually able to generate a good query reformulation. Furthermore, even after submitting a good query, only 15% of the participants were able to find a relevant document. Once a relevant document is found, 46% of the time a plausible answer was submitted, and 36% of these answers turned out to be valid[4].

In contrast, consider the "easy" question "*What is the highest peak in western hemisphere?*", for which 39 users submitted correct answers out of the 41 who attempted it. The state transitions diagram is also shown in Figure 7. The most probable path for this easy question is the shortest path: $Q^+ \rightarrow R^+ \rightarrow A^+V^+$, which occurs in 37% of the sessions for this question. As another example, for the question "Where and who exactly bought Louisiana from Napoleon?", any reasonable query immediately leads to a document about the history of Louisiana purchase[5]. While the page is definitely relevant, but is relatively long, and it is difficult to extract the correct answer. Some common incorrect answers included "Spain" and "President Jefferson bought Louisiana". In this case, most users found the relevant document (Q^+R^+), but submitted incorrect answers or no answer at all (or $Q^+R^+A^+V^-$ or $Q^+R^+A^-$, respectively).

5.4 Prediction of Search Session Success

In this section we report results on the prediction of session success by using the behavioral features as input. We compare our Conditional Random Fields (CRF) algorithm described in section 4.1 to both the naïve baseline algorithm that always predicts "success" (the majority class), as well as to the state-of-the-art Markov Model method that incorporates time distribution between actions (MML+Time) described in [11] and summarized in section 4.1.

Success Definition	*Baseline*		*MML+Time*		*CRF (All)*		*CRF (Selected)*	
	Acc	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*
$Q^?R^+A^+V^+$	0.65	0.40	0.61	0.58	0.68	0.60 (+4%)	0.68	0.62
$Q^?R^?A^+V^?$	0.87	0.47	0.72	0.55	0.86	0.64 (+17%)	0.88	0.66
$Q^?R^+A^?V^?$	0.83	0.45	0.66	0.53	0.80	0.57 (+8%)	0.81	0.59
$Q^?R_L^+A^?V^?$	0.60	0.38	0.59	0.53	0.68	0.66 (+26%)	0.69	0.67

Table 4: Prediction of search session success for different levels of success in QRAV model. Relative improvement against MML+Time model is shown in parenthesis.

We use 4-fold cross-validation in the following manner: for each of the four game rounds we train our model on all sessions from the *other three games*, and apply the trained model to predict the searcher success of the current game. Thus, we have four folds of roughly equal size. For each fold, neither the users nor the questions intersect. We compare algorithms by accuracy and F-measure, macro-averaged over the positive (successful) and negative (unsuccessful) classes. The results are presented in Table 4, which shows that our CRF model exhibits significant improvement over both the baseline and MML+Time models proposed in [11] for all definitions of success except the first one.

5.4.1 Feature significance

We now explore the relative significance of behavioral features used for training CRF by using a subset of the features, starting from one feature, and use a greedy search to extend the used subset, one feature at a time, by adding the best of the remaining features. In each step we choose the feature that gives the highest F1 performance of the CRF algorithm for predicting $Q^?R_L^+A^?V^?$ success. The results are shown in Table 5, and the best results obtained by this greedy feature selection for each definition of success are shown in table 4. As we can see, the first, most significant feature (*State*) is the same one as reported in [11] —the search action itself. Interestingly, the time interval between the actions and the choice of the web search engine are the two next most useful features. This makes sense, as the faster searchers are also likely to be more advanced or experienced, and are also more likely to experiment with switching search engines (encouraged by a drop-down box in our search interface). Finally, the position of the search result clicks provides additional indication of the search result quality – which in turn indicates the presence of a good query.

Feature	*F1*	*Accuracy*
State	0.624	0.675
$+\Delta t_{\leq 10}$	0.655 (+5%)	0.680
$+Q\_engine$	0.666 (+1.7%)	0.680
$+R_1\text{serp\_pos}_{\leq 2}$	0.670 (+0.6%)	0.687
$+R_1\text{serp\_pos}_{\leq 5}$	0.671 (+0.1%)	0.686

Table 5: Prediction of search success by the CRF model, when adding one best-performing feature at a time.

We now explore the differences in the performance of the MML variants and our CRF system, for different definitions of search success (Section 2). Figures 8 (a-d) report the precision vs. recall plots of identifying the Successful class. CRF performs best, and significantly better than MML for the definitions proposed in [11] (b) and for the definition proposed in [9] (d). For the other definitions of success (e.g., the most strict one (a)), the improvement of CRF is less striking, while MML variants exhibit performance comparable to the reports in the original study [11].

6. REAL WORLD SUCCESS PREDICTION: A LOG-BASED STUDY

We have shown that our model can successfully predict success of over a hundred participants in the tournament-like setting. Can we use the resulting model, trained on our contest data, to predict search success in a real-world search engine log? To answer this question, we use a large log of web searches performed from hundreds of shared-access workstations in a major university library, and attempt to predict search success using the models trained on our contest data.

[4] Any answer that correctly stated at least the year the Jominy hardness test was invited (1938) was accepted as valid.

[5] http://www.nps.gov/archive/jeff/lewisclark2/circa1804/heritage/louisianapurchase/louisianapurchase.htm

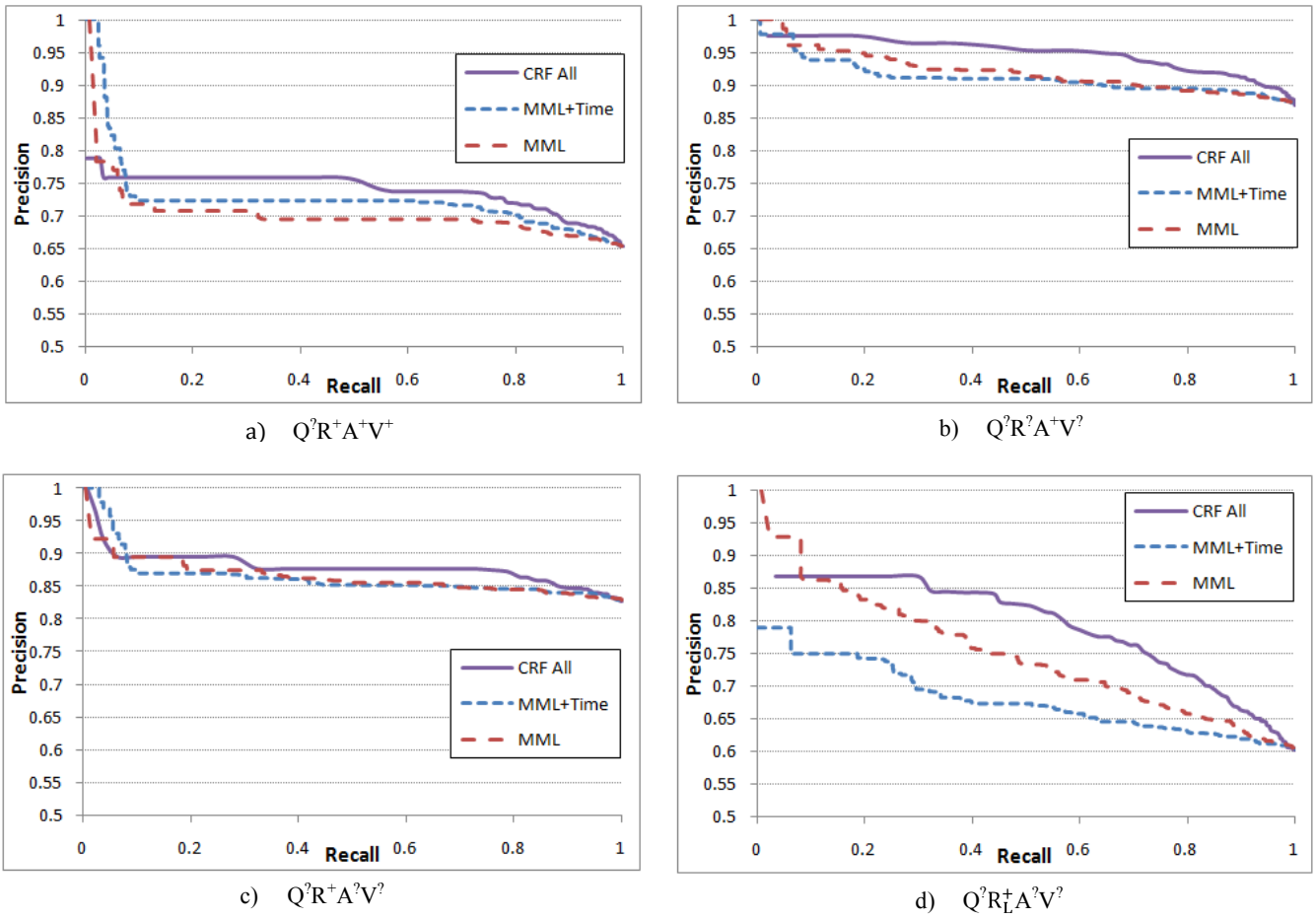

Figure 8. Recall-precision curves for compared algorithms, for different definitions of session success in QRAV model.

6.1 Experimental Setup

The data was collected by instrumenting over 100 shared-use workstations at the Emory University library using the EMU toolkit[6], with participants explicitly opting in to allow the searches to be tracked for library improvements (with roughly 60% opt-in rate). 16,693 search sessions (using almost primarily the Google search engine) were collected over a period of 6 months.

A sample of 175 search sessions consisting of more than one query (and thus more likely to be a non-trivial search) were manually labeled by the researchers to be successful or not, using the methodology and criteria outlined in Hassan et al. [11]. Specifically, the assessors used their best guess of the searcher intent based on the sequence of queries submitted, clicks on the results, and by manual examination of the visited result pages (recorded in the proxy log). 29% of the sessions were labeled as *successful*, 28% as *unsuccessful*, and 43% as *unknown* – where the assessors were not able to infer the searcher intent or determine whether the visited results satisfy the search.

6.2 Methods Compared

We used the CRF trained on all our contest data, using all of the features in Table 2. We trained the algorithms using the two most successful definitions of search success in *QRAV* model, namely $Q^?R^+A^?V^?$ (finding a good document) and $Q^?R_L^+A^?V^?$ (finding a good document as last in the search session). Then, the CRF model trained on the UFindIt game data was applied to log data.

6.3 Results and Discussion

Table 7 reports the results of predicting search session success. While the absolute accuracy and F1 values are lower than those on the original search contest data, the predictions significantly and substantially outperform the baseline. This experiment demonstrates that training a success model on search contest data can have significant practical applications, by directly applying the trained models to estimate search success of users of a production search engine.

Training Model: Success Definition	*Baseline*		*CRF (All)*	
	Acc	*F1*	*Acc*	*F1*
$Q^?R^+A^?V^?$	0.51	0.34	**0.55 (+8%)**	**0.52 (+53%)**
$Q^?R_L^+A^?V^?$	0.51	0.34	0.53 (+4%)	0.44 (+29%)

Table 7: Prediction of search success for real-world log using CRF trained on contest data, for success definitions in [11] and [9] respectively.

7. RELATED WORK

Our work spans several areas of modeling searcher behavior, including analyzing search log to understand variances in user behavior, evaluating search engine performance, conducting online study using crowd-sourcing approach, and predicting search success and frustration.

[6] http://ir.mathcs.emory.edu/EMU/

Search behavior using general search engine logs has been researched extensively. Jansen et al [16, 17] studied query and search distributions; Beitzel et al. [3], studied topic distribution and other fine-grained properties of a search engine log. Additionally, see [14] for a great overview of search engine log mining in general.

A number of web search evaluation metrics have been proposed. Clarke et al. [4] proposed an evaluation metric based on providing the necessary coverage for the information need. While these metrics provide a good estimate of the quality of the search results, and in turn have been shown to correspond to search effectiveness of users, these do not take into account the search success of a specific user for a session.

There were previous efforts to run a search engine contest to acquire search data. Most notably, the Yandex search engine[7] (a leading web search engine in Russia and Eastern Europe) ran a search contest to evaluate search engine performance, and characteristics of the winner strategies were studied [7, 31]. However, we perform a significantly more in-depth analysis of the search success and behavior and provide a re-usable infrastructure for a very different participant group and higher level of instrumentation detail. Microsoft's Bing search engine recently introduced a search game to improve search engine performance [22], which performed page-centric analysis on which pages were easier or more difficult to find using the Bing search engine.

The previous research closest to the first half of our work is the study of behavioral differences between "expert" users and "novice" users (White and Morris [28]), and the study of behavioral change in difficult search tasks (Aula et al. [2]). White and Morris [28] analyzed search logs of different search engines and defined "expert" searchers as those who use advanced operators in their queries. The authors found several differences in the behavior of expert searchers compared to those who do not use advanced operators. However, the lack of information about user goals and success prevented a more in-depth analysis of the reasons behind the behavioral differences. In our study, we adapt the same criteria to identify advanced users in our data, but, unlike in previous studies, we have well-defined search goals and success definitions, and find interesting similarities and differences. Aula et al. [2] examined the changes of searcher behavior associated with difficult search tasks using data collected from a usability lab study and an online study. To some extent, their approach of online study is similar to ours. However, the users in the previous online study were relatively passive while in our study we provide a competition-like setting to encourage more active search to let the users find best answers possible, which is potentially valuable to understand how advanced searcher behavior can help "novice" users in such difficult tasks and at the same time test the limit of search engine in those difficult tasks. Also, instead of using self-reported success (which could be misleading – as we show in our analysis, some users submit wrong answers that they thought were correct), we defined different types of success in principled way that could be potentially valuable for various applications.

Closest to the second half of our work is the recent work by Hassan et al. [11], which studied the prediction of searcher success on session-level. In their work, the success of a search task is judged by a human assessor using log information, where real search goals are not available, making the judgment potentially noisy. Another related work is by Feild et al. [10] that studied behavioral clues to detect search frustration. The real search goals are known since the data was collected from a user study, but the scale of the data is limited. In our study, we addressed such tension between the relatively small-scale, but controlled lab studies, and the large-scale log-based studies where the searcher goals unknown, by crowd-sourcing the data analysis using a game-like setting. In addition to the relative large scale of our study (hundreds of users) and the known search goals for each task, our study also benefits from the objective and well-defined *search success* metrics that can be used to evaluate whether the searcher goal was actually achieved.

To summarize, the key distinctions of our work compared to previous efforts are: a clear, well-defined QRAV model of search success for informational tasks; an online game-like study framework that enables large-scale user study with clear search goals and realistic search behavior; and an in-depth analysis of searcher behavior that results in more accurate prediction of search success than a previous state-of-the-art model.

8. CONCLUSIONS

We presented a novel methodology for analyzing searcher success in relation to the searcher behavior. What sets our work apart from previous studies is the emphasis on realistic search tasks (extracted from the real user questions on popular community question answering sites), yet allowing for well-defined, objective success criteria. This allowed us to develop a principled fine-grained model of web search success that naturally encompasses previously proposed models. We also developed and validated a competition-based framework for acquiring search behavior data at sufficient scale to allow analysis of successful and unsuccessful searcher behavior, and to train effective models for predicting search success.

Additionally, our study identified important search behavior characteristics that relate to search success. We showed that more successful users are faster and more active in their effort — they issue more queries, clicks, and browse more pages for each question, issue shorter queries, and more actively use query reformulations and advanced query syntax. The key characteristics that distinguish highly successful searchers were the ability to consistently identify relevant documents from a search engine result page, and to extract correct answer from the actual documents. Remarkably, the models trained on our acquired behavior data performed adequately on predicting searcher success based on the regular search engine log.

These findings suggest future directions for collecting real search behavior data in the academia, evaluating search engine performance, and improving search engine result presentation. Other areas for future exploration include personalization of the search success models, developing success metrics for other search intents (e.g., for exploratory search), and extending the contest-based approach to an even larger group of participants. Potential future applications of this work include applying models trained for different success definitions to provide labels for learning-to-rank improvements, and for developing evaluation metrics based on the behavioral metrics developed in this paper.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation grant IIS-1018321, and by the Yahoo! Faculty Research Engagement Program.

[7] http://www.yandex.com

9. REFERENCES

[1] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. In *Proc*. of SIGIR 2007

[2] Anne Aula, Rehan M. Khan, and Zhiwei Guan. How does search behavior change as search becomes more difficult? In Proc. of CHI 2010

[3] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, and Ophir Frieder. 2004. Hourly analysis of a very large topically categorized web query log. In *Proc. of* SIGIR 2004

[4] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proc. SIGIR'08*, 659-666, 2008.

[5] Honghua (Kathy) Dai, Lingzhi Zhao, Zaiqing Nie, Ji-Rong Wen, Lee Wang, and Ying Li. 2006. Detecting online commercial intention (OCI). In *Proc.* of WWW 2006

[6] Ali Dasdan, Kostas Tsioutsiouliklis, and Emre Velipasaoglu. 2010. Web search engine metrics: (direct metrics to measure user satisfaction). In *Proc*. of WWW 2010

[7] Dobrov B., Loukachevitch N., Dobrov G., Reznikov Y., Shternov S. Study of Query Transformations in the First Round of Yandex's Cup. In Proc. of "Internet-Mathematics-2005", (in Russian)

[8] D. Downey, S. Dumais, and E. Horvitz. Models of searching and browsing: Languages, studies, and applications. Journal of the American Society for Information Science and Technology (JASIST), 58(6):862--871, 2007

[9] Dupret, G. and Liao, C. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. *In Proc. of* WSDM 2010, 181-190

[10] Henry A. Feild, James Allan, and Rosie Jones. Predicting searcher frustration. In Proc. of SIGIR 2010

[11] Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. Beyond DCG: user behavior as a predictor of a successful search. In Proc. of WSDM 2010

[12] Jeff Huang and Efthimis N. Efthimiadis. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In *Proc.* of SIGIR 2009

[13] Scott B. Huffman, Michael Hochster, How well does result relevance predict session satisfaction? In *Proc.* of SIGIR 2007

[14] Bernard J. Jansen, Web Analytics. In *Synthesis Lectures on Information Concepts, Retrieval, and Services*), Morgan & Claypool, 2009

[15] Bernard J. Jansen, Danielle Booth, and Amanda Spink. 2009. Predicting query reformulation during web searching. In *Proc.* of CHI 2009 Extended Abstracts

[16] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.* 36, 2 (January 2000), 207-227

[17] Bernard J. Jansen and Amanda Spink. 2006. How are we searching the world wide web?: a comparison of nine search engine transaction logs. *Inf. Process. Manage.* 42, 1 (January 2006), 248-263.

[18] Kalervo Järvelin. 2009. Explaining User Performance in Information Retrieval: Challenges to IR Evaluation. In *Proc. of* ICTIR '09

[19] Diane Kelly, Methods for Evaluating Interactive Information Retrieval Systems with Users, Foundations and Trends in Information Retrieval, v.3 n.1—2, p.1-224, January 2009

[20] J. Lafferty, A. McCallum, and F. Pereira. Conditional random Fields: Probabilistic models for segmenting and labeling sequence data. *In Proc.* ICML-01, pages 282-289, 2001.

[21] Chang Liu, Jacek Gwizdka, and Jingjing Liu. 2010. Helping identify when users find useful documents: examination of query reformulation intervals. In *Proceeding of the third symposium on Information interaction in context* (IIiX 2010).

[22] Ma, H. and Chandrasekar, R. and Quirk, C. and Gupta, A. Page hunt: improving search engines using human computation games. *In Proc. of* SIGIR 2009, 746-747

[23] Benjamin Piwowarski, Georges Dupret, and Rosie Jones. 2009. Mining user web search activity with layered bayesian networks or how to capture a click in its context. In *Proc. of* WSDM 2009

[24] Mark Sanderso*n, Monica L*estari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *Proc.* of SIGIR 2010

[25] Catherine L. Smith and Paul B. Kantor. 2008. User adaptation: good results from poor systems. In *Proc. of* SIGIR 2008.

[26] Andrew Turpin and Falk Scholer. 2006. User performance versus precision measures for simple search tasks. In *Proc.* of SIGIR 2006

[27] E.M. Voorhees, H.T. Dang. Overview of the TREC 2005 Question Answering Track, NIST.

[28] Ryen W. White, Paul N. Bennett, and Susan T. Dumais. 2010. Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (CIKM 2010)

[29] Ryen W. White and Dan Morris. Investigating the querying and browsing behavior of advanced search engine users. In Proc. of SIGIR 2007.

[30] Ya Xu and David Mease. 2009. Evaluating web search using task completion time. In *Proc.* of SIGIR 2009.

[31] Yandex Internet Search Competitions http://kubok.yandex.ru/organize_eng.html