Understanding How People Interact with Web Search Results that Change in Real-Time Using Implicit Feedback

Jin Young Kim
Microsoft Bing
Bellevue, WA USA
jink@microsoft.com

Mark Cramer
Surf Canyon
San Francisco, CA USA
mcramer@surfcanyon.com

Jaime Teevan
Microsoft Research
Redmond, WA USA
teevan@microsoft.com

Dmitry Lagun
Emory University
Atlanta, GA USA
dlagun@mathcs.emory.edu

ABSTRACT
The way a searcher interacts with query results can reveal a lot about what is being sought. Considerable research has gone into using implicit relevance feedback to identify relevant content in real-time, but little is known about how to best present this newly identified relevant content to users. In this paper we compare a traditional search interface with one that dynamically re-ranks and recommends search results as the user interacts with it in order to build a picture of how and when users should be offered dynamically identified relevant content. We present several studies that compare logged behavior for hundreds of thousands of users and millions of queries as well as self-reported measures of success across the two interaction models. Compared to traditional web search, users presented with dynamically ranked results exhibit higher engagement and find information faster, particularly during exploratory tasks. These findings have implications for how search engines might best exploit implicit feedback in real-time in order to help users identify the most relevant results as quickly as possible.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval---search process

Keywords
Interactive information retrieval, query log analysis, web search, dynamic ranked retrieval, implicit relevance feedback.

1. INTRODUCTION
Searchers do not always find what they seek after a single query. Instead, they often issue multiple queries, incorporating what they learn from the results to iterate and refine how they express their information needs. While many search engines try to make this easier by providing reformulation suggestions and some searchers are able to quickly incorporate this feedback, others explore each set of search results exhaustively and fail to benefit from any new information identified during the search process [3].

To address this, search engines have begun to incorporate session context into the search results they return for subsequent queries in a session [20]. For example, if a person issues the query *apples* and then clicks on a website about fruit, future queries in that session can be biased to return more fruit-related results versus company-related results. Different types of session context have been explored, including past queries [18], the topic or reading level [12] of clicked results and the snippet text [26]. The onus of reformulating the query, however, continues to rest upon the user. In cases where users' search skills are limited [30], their knowledge of the subject is weak [29], or the search topic is difficult or exploratory [31], reformulation can be daunting and sometimes overwhelming.

There is, however, an opportunity for search engines to exploit the implicit feedback users provide following a query to present the most relevant results immediately, without requiring a subsequent query. This can be done by dynamically altering the result set, in response to real-time implicit relevance feedback, as the user interacts with it. Little is currently known, however, about how people might best encounter and use newly identified content in the course of a single query. To provide insight into dynamic result set interaction, we explore how and when new content is useful to searchers in the course of a single query. We present several large-scale studies of people's interactions with Surf Canyon, a popular commercial search system that re-ranks search results in real-time following each user action.

Contextually relevant results that initially might have been ranked deep within the result set have the opportunity to be promoted based on the user's implicit signals. An example for the query *apples* is shown in Figure 1. Upon returning to the result page after clicking on the second result (a health food website) the user is presented with a "real-time recommendation" based on this action, including a result about the dental benefits of apples.

Because this search system is used daily by hundreds of thousands of users for millions of queries, the usage data paints a picture of how real-time, dynamically ranked results impact people's behavior. We present the results of two controlled studies designed to compare dynamic ranking with traditional, static web search. Our analysis reveals that users have a higher engagement in the search process when provided with a dynamic feedback experience, click-

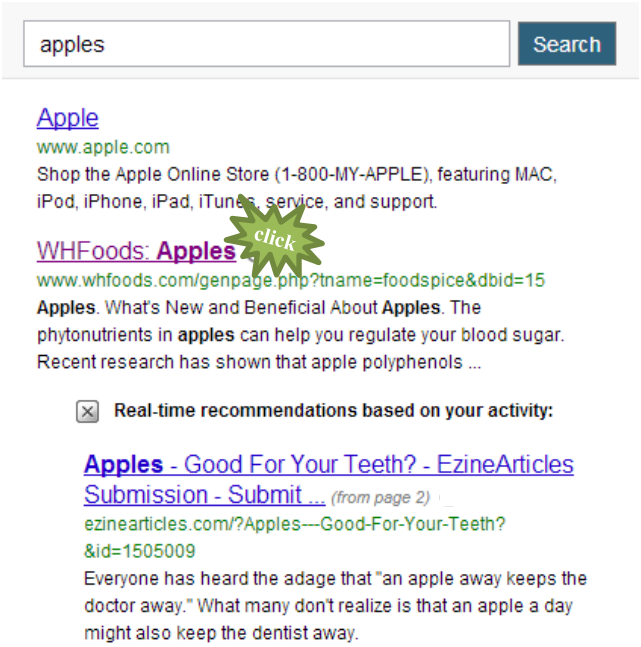

Figure 1. Real-time recommendations are presented inline based on the results the user clicks.

CIKM '13, October 27–November 1, 2013, San Francisco, CA.

ing on more results and spending a longer time searching. Participants aregiven exploratory tasks to complete on the two systems found more results while expending less time when given dynamic results as compared to a traditional search interface.

2. RELATED WORK

There is growing interest in the IR community around understanding and supporting personalization [18] [26]. Researchers have explored short-term personalization based on session context, including previous searches and clicked results, to identify relevant results [21,9]. Teevan et al. [27] present a framework to identify queries that would benefit the most from personalization. In this paper we explore which queries benefit most from the real-time personalized results.

Relevance feedback is the primary post-query method for automatically improving system representation of a searcher's information need and has been studied extensively. Explicit relevance feedback [13] allows users to select documents or terms to be used for query expansion. Explicit relevance feedback is rarely used, however, as it requires direct interaction, placing cognitive load on the user, and often results in the identification of irrelevant content [10].

To overcome some of the challenges with explicit relevance feedback, researchers have explored the use of implicit relevance feedback. With implicit relevance feedback, user behavior, such as clicking documents or scrolling, is unobtrusively monitored and used to expand the understanding of users' information needs beyond the query [32].

In addition to both explicit and implicit relevance feedback, other adaptive search systems have been explored. Kaplan et al. [11], for example, describe a navigation scheme that adapts to user behavior by using associative matrices that encode user preferences. The term *dynamic ranking* and a theoretical justification for the approach was formally introduced by Brandt et al. [7] where they proposed a retrieval model that optimizes relevance of dynamic recommendations based on the user's choice of documents. Cramer et al. [17] present a preliminary study of user behavior with dynamic ranking based on log data. They find that dynamic ranking increases the click through rate of recommended URLs compared to a baseline system. We extend this line of research by conducting user centric evaluation and presenting additional analysis of searcher behavior during the interaction with a dynamic ranking enabled system.

Search systems that rely on context and personalization are difficult to evaluate because they depart from the notion of global relevance that can be effectively assessed by judges to a notion of personalized relevance that is hard to measure by anyone other than the particular user. As such, Bennett et al. [4] suggest evaluating personalized search algorithms online by automatically identifying search result click behaviors that suggests satisfaction. Ageev et al. [1] propose a scalable, game-like framework for conducting remote user studies of searchers' successes. We compare a system that dynamically ranks results to a baseline system by conducting a recall-based user study with similar tasks to those used by Ageev et al. [1]. Online controlled experiments (e.g., [14]) and crowdsourcing evaluation (e.g., [6]) employ similar approaches.

To summarize, previous work into personalization, relevance feedback and adaptive search has resulted in promising ways to use context and implicit feedback to identify relevant results. Evaluating such systems, however, is difficult and little is known about how people actually interact with relevant content that is identified and presented to users in real-time. In this paper we conduct user-centric evaluations of a dynamic ranking system that provides helpful insights for researchers and practitioners. We begin by discussing the details of the dynamic ranking system we studied.

3. DYNAMIC RANKING SYSTEM

To understand how people interact with dynamically ranked results, we study Surf Canyon, a popular commercial search system. Surf Canyon provides a browser extension (available for IE, Firefox, and Chrome) that applies an interactive dynamic layer on top the results returned by a number of existing popular search engines (e.g., Google, Bing, and Yahoo!) which re-ranks search results as the users interact with them. In this section we briefly describe how dynamically ranked results are identified and look more closely at how these results are then displayed.

3.1 Identifying New Relevant Results

When a query is issued to a major search engine using a browser that has the Surf Canyon extension running, the system begins by fetching, in the background, the top 50-100 results from the underlying search engine. The user is initially presented with the top 10 results, as identified by the underlying search engine, in a typical fashion. However, by then monitoring user actions, including link clicks, web page visits, scrolling, and back button clicks, Surf Canyon starts generating a real-time model of inferred intent. Intent is derived from the titles, snippets and URLs presented to the user, and, in some cases, the page content and metadata information of clicked and skipped results. All activity in the current information session is used to build the model. This model is used to expand upon the initial understanding of the user's information need which was derived only from the query and other data available prior to the user submitting the query. The real-time inferred intent model is then exploited *immediately* after each user action in order to re-rank the result set.

3.2 Displaying Relevant Results to the User

Surf Canyon presents users with newly identified results by re-ranking the result set and then displaying content that has not yet been viewed. This is done in two ways: 1) as indented recommendations following a search result click, and 2) as an additional page of results following a request for more results. Only unseen content is displayed as previous research has shown that results that change as users interact with the search page can interfere with the ability to find information because results no longer appear where expected [25].

3.2.1 Recommendations Following a Click

After viewing a result, should the user's information need remains unsatisfied, for whatever reason, and the user clicks "back" to return to the search page, this is an opportunity to display re-ranked results. When users return to a result page while using Surf Canyon, newly identified results are displayed beneath the selected document. These results are drawn from a large set of "unseen" results that have been fetched from subsequent pages of results returned by the underlying search engine. Because changes to a search page that is actively being used can be disorienting [24], new results are indicated by indenting.

Each time a user clicks a search result the real-time inferred intent model is updated, the entire result set is anked and the most relevant previously unseen results are brought forward to the current search results page. This happens even when an indented re-ranked result is selected, in which case the recommendations are nested, which goes to a maximum of three levels deep.

The Surf Canyon click-based recommendations are intended to add a new element to existing search interfaces while minimizing the disturbance to a user's normal workflow. In practice users are typically able use the re-ranking feature without instruction or guidance. In this paper we focus on how people interact with dynamically re-ranking search results, but the notion of adding dynamically identified new content based on implicit behavior is quite gen-

eral and may easily be extended to other types of information, such as relevant entities and contextual advertisement.

3.2.2 Additional Requested New Content

With Surf Canyon, users are also able to interact with new, contextually relevant content when navigating to subsequent pages of results. Upon clicking "More Results", the real-time inferred intent model is once again updated. Rather than show results 11 to 20 as initially computed at the beginning of the session, the system displays the ten most relevant results as computed by the model. Because this content has not been previously viewed by the user, there is no need to visually indicate that this content is new by indenting.

4. METHODOLOGY

We explore how users interact with dynamically ranked results by studying the Surf Canyon logs. To compare the traditional static search experience with users' experiences with dynamically-ranked content, we directed a portion of Surf Canyon's traffic to a traditional static search experience. Although this method allowed us to compare the behavior of users with and without the dynamic ranking, even in a controlled experimental log study like this it is not easy to control for tasks. For this reason we conducted an additional user study with controlled tasks with self-reported success.

4.1 Log Data with Controlled Traffic

We studied user behaviors in a set of controlled traffic experiments in which different user groups were exposed to different configurations of the service. The data set we used consists of log data for major search engines collected from the Surf Canyon browser extension. The extension captures users' interactions with the search result pages, including queries, clicks on results and various other pieces of information. The data sets also contained session IDs, defined using a period of inactivity for 30 minutes as a session boundary [28]. Anonymous user IDs were used to group queries into information sessions. Since many of the clicks were on dynamically ranked results or re-ranked results on pages beyond the first, we utilized metadata associated with each click to distinguish them from regular result clicks. Since the results on subsequent pages were all re-ranked, all clicks on the second page and beyond were considered clicks on dynamically ranked results.

To quantify the differences in search behaviors caused by the two different ways dynamic results are displayed to the searcher, we ran two experiments for a large fraction of user traffic where we turned each off separately for 24 hours. To study the impact of dynamically ranked results, we turned off dynamic ranking for 20% of users. The other 80% of users received the standard dynamic ranking experience. To study the impact of dynamically ranked subsequent-page content, we turned off re-ranking for 50% of the users. The net result was 387,347 queries available to study the impact of click-based recommendations and 1,560,996 queries to study the impact of dynamically re-ranked next-page content.

Although traffic in the controlled online experiment was randomly split, there could be some task-based variation. Previous research has shown that search behavior can vary significantly by task [2]. For this reason, we further filtered to only look at overlapping queries by first aggregating the data from each group by query and then taking the intersection of the traffic based on the query string. This yielded 11,655 unique queries in the case of the dynamically ranked results. As there were a limited number of overlapping queries in the next-page case, we did not do this additional analysis for this experiment.

Navigational queries (targeted at navigating to a specific website) are common but search behavior surrounding these queries is known to be particularly different from other types of search behavior [2]. Conversely, people are more inclined to click many results following a query when their information need is exploratory in nature [31]. We hypothesized that dynamically ranked content is particularly likely to be useful for exploratory, open-ended searches and thus separate the data by task type in our analysis. To study the log data by task, we built a classifier to identify naviga-

Table 1: Query-level statistics of the controlled traffic for all queries and information queries only (bottom two rows) when click-based recommendations were selectively turned off.

Experimental Condition		*N*	Query Duration	Average Number of Result Clicks			
				Total	Original	Dynamic	Next Page
All Queries	**Dynamic**	11,655	****244.36**	****1.21**	1.16	*****0.04**	0.01
	Static	11,655	239.05	1.19	***1.18**	n/a	0.01
Non-Navigational Queries	**Dynamic**	5,545	**251.91**	***1.32**	1.24	*****0.07**	***0.02**
	Static	5,545	244.44	1.28	***1.27**	n/a	0.01

Table 2: Query-level statistics of the controlled traffic when subsequent page re-ranking was selectively turned off.

Experimental Condition	*n*	Average Number of Result Clicks				% of All Clicks	
		Total	Original	Dynamic	Next Page	Dynamic	Next Page
Dynamic	198,752	0.72	0.67	0.02	*****0.03**	2.56%	*****3.83%**
Static	188,595	0.72	0.67	0.02	0.02	2.56%	3.39%

Table 3: Session-level statistics for the controlled traffic when click-based recommendations (top two rows) and subsequent page re-ranking (bottom two rows) were selectively turned off.

Experimental Condition		Queries in a Session	Session Duration	Average Number of Result Clicks			
				Total	Original	Dynamic	Next Page
Click-Based Recommendations	**Dynamic**	2.86	****462.30**	*****1.65**	*****1.57**	*****0.07**	0.01
	Static	*****2.91**	452.10	1.63	1.62	n/a	0.01
Next Page Re-Ranking	**Dynamic**	*****1.87**	***424.54**	*****1.35**	*****1.26**	0.03	*****0.05**
	Static	1.85	418.56	1.32	1.24	0.03	0.04

tional queries based on methodologies in previous work [15].

4.2 User Study

To further control for task, we supplemented the log data with smaller-scale data where we provided users with tasks and asked for explicit feedback of success. To do this, we built upon previous work using an information search game for modeling success to evaluate different variants of interaction models. We used the UFindIt framework [1] to collect search behavior data from paid Amazon Mechanical Turk users. As in the original UFindIt game, we used Apache web server proxies to log all pages visited by users during the game. The searchers were given task descriptions as well as initial queries. While we pre-populated the search box with the initial query, searchers were allowed to change to whatever query terms they thought reasonable.

The study was a 2x2 design where half of the participants interacted with dynamically ranked results while the other half did not. In each case, half of the participants were given fact finding questions and asked to find a single answer, while the other half were given search topics that were intended to be more exploratory and were asked to find five different, relevant URLs. The fact finding questions were drawn from the 18 original UFindIt questions, and included, for example, "What were the deadliest tornadoes in history?" The exploratory search topics were drawn from Web track of TREC 2010 [8]. We randomly selected 10 topics and used their subtopics as intent specific task descriptions while providing topic names as initial queries.

In our analysis we used two definitions of search success: 1) self-reported success (i.e., whether any answer URL was submitted by the player), and 2) the correctness of the submitted URLs. Answer correctness was determined by examining the submitted URLs and checking if they contained correct answers for the task question. For this definition of success we obtained labels for 260 URLs submitted in factoid tasks as well 939 URLs submitted in exploratory tasks. The labeling process was crowdsourced through CrowdFlower, where each submitted URL was checked by independent assessors. To quantify agreement among the assessors we calculated Fleiss' kappa. For exploratory question the kappa value was 0.6, which indicated reasonable agreement among the judges. These two definitions of success allowed us to evaluate the quantity and quality of the interaction. Users who completed less than two tasks were dropped to ensure trustworthiness. The final dataset included the results of 826 tasks (417 fact finding, 409 exploratory) by 91 users.

5. COMPARING INTERACTION MODELS

In this section we compare people's interactive experiences with dynamic ranking versus the traditional search experience based on controlled log analysis and user study.

5.1 Controlled Traffic Analysis

We analyzed the controlled traffic data to understand how search behavior differed when people were presented with dynamically ranked results or re-ranked content on subsequent result pages. Specifically, we looked at the average amount of time people spent on a query and the average number of clicks people made on the search result content, broken down by whether they clicked on the original results returned prior to dynamic ranking, the dynamically recommended results or results on subsequent result pages. A summary of these behaviors are shown in Tables 1 through 3. Statistically significant differences (based on t-tests) are marked with asterisks (* for $p < .05$, ** for $p < .01$, and *** for $p < .001$).

Table 1 shows query-level results. We observe that users presented with dynamically ranked results spent more time searching and clicked on more results in total than the static results group. Users spent over 5 seconds more interacting with the result page when the results contained dynamic content (244.36 seconds) compared to when they did not (239.05 seconds). They also clicked more results (1.21 clicks compared with 1.19 clicks). Although people were less likely to click on results that were part of the original result set when presented with dynamic content, they more than made up for this by clicking on the dynamically ranked results.

We hypothesized that dynamic content would be particularly useful for tasks that are more exploratory in nature. The bottom two rows of Table 1 show the behavior observed for non-navigational queries, where people tend to search longer and click more. In particular, we observed that they are more likely to interact with dynamically ranked results. (We will explore these differences in greater depth in Section 6 when we look at how the use of dynamic results is correlated with characteristics of the queries, sessions and users.) The differences between the non-navigational behavior with and without dynamically ranked results echo what was seen for general query traffic, except that they are uniformly larger in magnitude and percentage.

Table 2 shows the results for the controlled study where subsequent page re-ranking was turned off for the control group. Behavior on subsequent pages is different when that content was generated dynamically during the course of the search. The percentage of clicks on subsequent page content is significantly higher when those results were re-ranked based on within-query user behavior than when they were not. Since users were exposed to the same interface, the difference in the percent of clicks may be attributed to a difference in the quality or content of the results. Note that the average number of results clicks is much higher in Table 1 than Table 2.

We also looked at session-level behavior for both experimental conditions, with summary statistics presented in Table 3. Session-level behavior is generally similar to query-level behavior. As was the case for the query-level statics, we observed that users who received click-based recommendations spent longer (by 10 seconds) searching than other users. They also clicked on significantly more results in total, with some of those clicks going to dynamic results at the expense of clicks on the originally presented results.

Notably, people who received click-based recommendations issued fewer queries (2.86 vs. 2.91) during a session than people who did not. This may be because the dynamic nature of the page surfaced results that mitigated the need to re-query. While at a query level we only saw an interaction between the presence or absence of next-page re-ranking with people's next-page behavior in Table 2, at a session level we see significant differences in all types of behavior as a function of the experimental condition. People with dynamic subsequent page content not only clicked on more next-page results, but also clicked on more results overall.

The controlled log data suggests that the inclusion of dynamically ranked content increases people's engagement with the results and the amount of time they spend searching, while decreasing the number of queries they need to issue. While previous research suggests that increased engagement tends to indicate a positive change for users [3], we wanted to better understand the impact of these changes on people's ability to find what they are looking for. We also wanted to further explore the particularly large changes observed for non-navigational queries. For this reason, we look at the results of the controlled user study.

5.2 User Study Results

As with the controlled log study, in the user study people were directed to a version of the search engine that either included dynamically ranked results or did not. In this study, however, users were performing directed tasks and asked to provide information

about what they found so that their success could be measured as described in Section 4.2.

Results are summarized in Table 4, where we report task-level averages for completion time, success rate, number of clicks on original results, number of clicks on dynamically ranked results, number of clicks on subsequent page results and number of queries. Note that we have two measures of success which focuses on the quantity (self-reported success) and quality (correctness of the URLs) of the interaction.

We begin our analysis of the results by looking at how successful people were at completing the two different types of search tasks: factoid and exploratory. Participants generally reported that they were very successful: 87% to 99% of all tasks were believed to be successfully completed. For factoid tasks there was no difference in reported success, but for exploratory tasks the users with dynamic recommendations felt significantly more successful (94% vs. 99%; $p < 0.01$ using a Proportions test).

In actuality, however, people were much less successful than they believed, with questions being answered correctly only 38% to 41% of the time. For factoid tasks, participants in the dynamic ranking group completed 41.4% of the tasks correctly, while participants in the static ranking group had a very similar 41.0%. Likewise, for exploratory tasks the mean success rates were 38.7% and 38.9% respectively. In both cases, any observed differences were not statistically significant ($p > .5$ using a proportions test).

The results indicate that dynamic ranking helps increase the quantity of the results although not necessarily the quality. This is understandable in that click-based recommendations allow users to explore more results than static search interface without query reformulation. It is also interesting to note that although people thought they were more successful for exploratory tasks than factoid tasks, they were less likely to answer correctly.

Task completion time is another important metric that has been used in prior work [33] to provide an intuitive criterion for assessing utility of a search interface. Figure 2(a) shows task completion times broken down by the availability of dynamically ranked results for factoid tasks. On average, players in the dynamic ranking group spent about the same time searching for answers (107.7 seconds versus 109.3 seconds) as the static ranking group. Differences between the means of the dynamic and static rankings are not statistically significant with a Student t-test ($p = .94$).

Similarly, Figure 2(b) shows task completion times for exploratory tasks. In contrast to what we observed with factoid tasks, completion time varied significantly between the dynamic and static ranking groups for exploratory tasks. Participants in the dynamic ranking group were able to complete tasks in 123.7 seconds on average whereas participants with static ranking had to spend 173.3 seconds, or 40% more time. The difference was significant ($p < .01$).

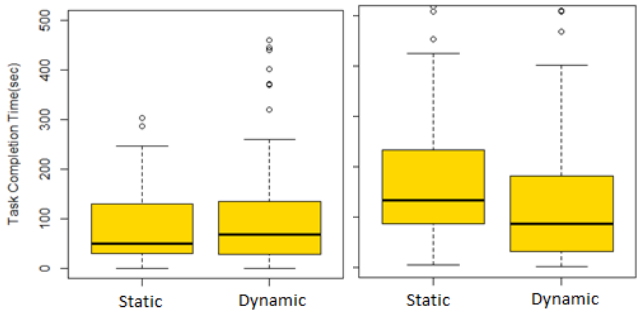

Figure 2: Task completion time box plots for factoid tasks (left) and exploratory tasks (right).

Another striking difference between factoid and exploratory tasks can be observed in the number of dynamically ranked results clicked for factoid (0.05) and exploratory (0.44) questions. This is consistent with the differences we observed in the logs for navigational and non-navigational queries (Table 1 and 5), again confirming our intuition that an interactive experience is best suited for more complex tasks.

Looking at the impact of the dynamic ranking on the click and query count, we found that for factoid questions the average number of clicks for the dynamic ranking and static ranking groups was 1.79 and 2.48 respectively. For exploratory tasks it was 4.26 and 6.12 respectively. Differences in the number of clicks were not significant ($p > .2$, Wilcoxon-Mann-Whitney test). The number of queries remained relatively similar for groups and task types, although the dynamic ranking group issued slightly fewer queries.

When comparing traditional search behavior with people's interactions with dynamically ranked content in the logs, we observed that people engaged more with search results when they were given dynamically ranked results as they searched, clicking more and searching longer. In contrast, when we controlled the tasks that users completed, we found that their searches were faster and they clicked less frequently.

Given the levels of success people achieved with dynamic content, it may be that for fixed tasks the dynamically ranked results reduce the effort required to complete the task, but for real world tasks

Table 4. Task-level statistics for user study participants, broken down into fact finding and exploratory tasks.

		n	Report Success	Answer Correctly	Queries per Task	Task Duration	Average Number of Result Clicks		
							Original	Dynamic	Next Page
Factoid Tasks	**Dynamic**	209	88%	41.4%	1.29	109.3	1.79	0.05	0.05
	Static	208	87%	41.0%	1.31	107.7	2.48	n/a	0.12
Exploratory Task	**Dynamic**	204	****99%**	38.7%	1.30	123.7	4.26	0.44	0.08
	Static	205	94%	38.9%	1.37	****171.3**	6.12	n/a	0.14

Table 5 Overall statistics of how people interacted with the dynamically ranked results, broken down by query type.

	n	Average Number of Result Clicks				Percent of Clicks	
		Total	Original	Dynamic	Next Page	Dynamic	Next Page
All Queries	831,853	0.94	0.88	0.04	0.03	3.73%	3.42%
Navigational	646,487	0.78	0.77	0.01	0.01	1.11%	0.89%
Non-Navigational	185,366	1.50	1.26	0.13	0.12	8.47%	8.02%

that can evolve and grow beyond the initial target, the new content encourages additional interaction and exploration. The fact that we observed the opposite trend in task duration between controlled traffic analysis and user study stresses the importance of controlling the search task in IR research.

6. CONCLUSIONS

We have compared dynamically ranking search results with traditional static web search. By studying user behaviors in a set of controlled traffic experiments, we found that dynamic content leads to higher user engagement in the search process. To ensure complete control of search task, we also conducted a user study with 91 participants and two types of search tasks, where the results showed that the click-based recommendation feature of dynamic ranking improves the user performance of exploratory search task significantly measured in task completion time and the number of the self-reported URL answers.

While our study was based on the interaction model employed by Surf Canyon, we believe that many of our findings have implications in enabling rich user interactions for web search in general. Future work includes the extension of dynamic ranking by removing some of its restrictions.

7. REFERENCES

[1] Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein, "Find it if you can: a game for modeling different types of web search success using interaction data," in *SIGIR*, New York, NY, USA, 2011, pp. 345-354.

[2] Azin Ashkan, Charles L. Clarke, Eugene Agichtein, and Qi Guo, "Classifying and Characterizing Query Intent," in *ECIR*, Berlin, Heidelberg, 2009, pp. 578-586.

[3] Anne Aula, Paivi Majaranta, and Kari-Jouko Raiha, "Eye-tracking reveals the personal styles for search result evaluation," in *Proceedings of the 2005 IFIP TC13*, Berlin, Heidelberg, 2005, pp. 1058-1061.

[4] Paul N. Bennett, Filip Radlinski, Ryen W. White, and Emine Yilmaz, "Inferring and using location metadata to personalize web search," in *SIGIR*, New York, NY, USA, 2011, pp. 135-144.

[6] Roi Blanco et al., "Repeatable and reliable search system evaluation using crowdsourcing," in *SIGIR*, New York, NY, USA, 2011, pp. 923-932.

[7] Christina Brandt, Thorsten Joachims, Yisong Yue, and Jacob Bank, "Dynamic ranked retrieval," in *WSDM*, New York, NY, USA, 2011, pp. 247-256.

[8] Nick Craswell, Dennis Fetterly, and Marc Najork, "Microsoft Research at TREC 2010 Web Track," 2010.

[9] Mariam Daoud, Lynda Tamine-Lechani, and Mohand Boughanem, "Towards a graph-based user profile modeling for a session-based personalized search," *Knowl. Inf. Syst.*, vol. 21, no. 3, pp. 365-398, 2009.

[10] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic, "The Use of Relevance Feedback on the Web: Implications for Web IR System Design," in *WebNet*, 1999, pp. 550-555.

[11] Craig A. Kaplan, Justine Fenwick, and James Chen, "Adaptive Hypertext Navigation Based On User Goals and Context," *User Modeling and User-adapted Interaction*, vol. 3, pp. 193-220, 1993.

[12] Jin Young Kim, Kevyn Collins-Thompson, Paul N. Bennett, and Susan T. Dumais, "Characterizing web content, user interests, and search behavior by reading level and topic," pp. 213-222, 2012.

[13] Jurgen Koenemann and Nicholas J. Belkin, "A Case For Interaction: A Study Of Interactive Information Retrieval Behavior And Effectiveness," , 1996, pp. 205-212.

[14] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne, "Controlled experiments on the web: survey and practical guide," *Data Min. Knowl. Discov.*, vol. 18, no. 1, pp. 140-181, Feb 2009.

[15] Uichin Lee, Zhenyu Liu, and Junghoo Cho, "Automatic identification of user goals in Web search," in *WWW*, New York, NY, USA, 2005, pp. 391-400.

[17] M. Wertheim, and D. Hardtke M. Cramer, "Demonstration of improved search result relevancy using real-time implicit relevance feedback," in *SIGIR Workshop*, 2009.

[18] Nicolaas Matthijs and Filip Radlinski, "Personalizing web search using long term browsing history," in *Web Search and Data Mining*, 2011, pp. 25-34.

[20] Barry Schwartz. (2012) Surviving Personalization With Bing & Google. [Online]. http://www.seroundtable.com/smx12-personal-bing-google-15250.html

[21] Xuehua Shen, Bin Tan, and ChengXiang Zhai, "Context-sensitive information retrieval using implicit feedback," in *SIGIR*, New York, NY, USA, 2005, pp. 43-50.

[24] Jaime Teevan, "How people recall, recognize, and reuse search results," *ACM Transactions on Information Systems*, vol. 26, pp. 1-27, 2008.

[25] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael A. S. Potts, "Information re-retrieval: repeat queries in Yahoo's logs," in *SIGIR*, New York, NY, USA, 2007, pp. 151-158.

[26] Jaime Teevan, Susan T. Dumais, and Eric Horvitz, "Personalizing search via automated analysis of interests and activities," in *SIGIR*, New York, NY, USA, 2005, pp. 449-456.

[27] Jaime Teevan, Susan T. Dumais, and Daniel J. Liebling, "To personalize or not to personalize: modeling queries with variation in user intent," in *SIGIR*, New York, NY, USA, 2008, pp. 163-170.

[28] Ryen W. White, Paul N. Bennett, and Susan T. Dumais, "Predicting short-term interests using activity-based search context," in *CIKM*, New York, NY, USA, 2010, pp. 1009-1018.

[29] Ryen W. White, Susan T. Dumais, and Jaime Teevan, "Characterizing the influence of domain expertise on web search behavior," in *Web Search and Data Mining*, 2009, pp. 132-141.

[30] Ryen W. White and Dan Morris, "Investigating the querying and browsing behavior of advanced search engine users," in *Research and Development in Information Retrieval*, 2007, pp. 255-262.

[31] Ryen W. White and Resa A. Roth, *Exploratory Search: Beyond the Query-Response Paradigm.*: Morgan \& Claypool, 2009.

[32] Ryen W. White, Ian Ruthven, and Joemon M. Jose, "Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes," in *SIGIR*, New York, NY, USA, 2002, pp. 57-64.

[33] Ya Xu and David Mease, "Evaluating web search using task completion time," in *SIGIR*, New York, NY, USA, 2009, pp. 676-677.