One Query, Many Clicks: Analysis of Queries with Multiple Clicks by the Same User

Elad Kravi[†‡], Ido Guy[†§], Avihai Mejer[†], David Carmel[†], Yoelle Maarek[†], Dan Pelleg[†], Gilad Tsur[†]
[†]Yahoo Research, Haifa, Israel
[‡]Technion – Israel Institute of Technology, Haifa, Israel
[§]Ben-Gurion University of the Negev, Beer Sheva, Israel
ekravi@cs.technion.ac.il idoguy@acm.org

ABSTRACT

In this paper, we study multi-click queries – queries for which more than one click is performed by the same user within the same query session. Such queries may reflect a more complex information need, which leads the user to examine a variety of results. We present a comprehensive analysis that reveals unique characteristics of multi-click queries, in terms of their syntax, lexical domains, contextual properties, and returned search results page. We also show that a basic classifier for predicting multi-click queries can reach an accuracy of 75% over a balanced dataset. We discuss the implications of our findings for the design of Web search tools.

Keywords: exploratory search; multiple click queries; query log analysis; query session

1. INTRODUCTION

Leading commercial Web search engines take advantage of large-scale user interaction data in order to enhance their result ranking and presentation. Click-through information plays an important role in this process, as a source of implicit relevance feedback from the user [1, 25]. For some queries, no click is performed, e.g., due to user dissatisfaction with the results or, by sharp contrast, to complete satisfaction with a correct result presented directly on the search engine results page (SERP) [28]. For other queries, one click is performed, for example to navigate to a searched website, or to find a previously visited page. Yet for other queries, clicks on multiple results are performed, e.g., because the first clicked result was not completely satisfying, or because the user is looking for a variety of opinions, for example when performing a market research, looking for the symptoms of a disease, or reviewing scholarly literature.

In this work, we focus on the third type of queries, to which we refer as *multi-click queries*. Despite being the most infrequent type out of the three click behaviors, we believe it is of particular interest. Queries with multiple clicks are likely to represent complex information needs, which cannot be satisfied by a single result page. Users with such needs, and the right state of mind, are likely to deeply engage with the search interface. Moreover, as our analysis demonstrates, multi-click queries are more correlated with tail queries and the notion of difficult queries [5] than single- or no-click queries. For search services, it is important to satisfy such queries, and not only head queries, in order to avoid a "radical variance in performance" [5]. It has indeed been shown, in multiple Web domains, that even ordinary people exhibit extraordinary tastes, which should not be overlooked, but rather treated with care [14, 32].

To the best of our knowledge, multi-click queries have not been comprehensively studied. While many papers explored click models and leveraged click data, they did not make the explicit distinction between multi-click queries and queries with one or zero clicks. On the other hand, studies of exploratory search examined a more complex notion of a "task" or "mission", which involves sessions of multiple queries. The definition of a session in this case is challenging, and sessions are often interleaved or hierarchically organized [27]. The proposed tools for such tasks typically work at the granularity of multiple queries, for example, by aggregating results across several queries or logging user interaction with previous queries in the session [11, 21, 37]. In contrast, we focus on the granularity of a single query session, which encapsulates a direct interaction with the search engine, i.e., the user query, the retrieved search results, and the user click(s), without a dependency in more complex sessions or tasks[1]. We formally define a model for identifying multi-click queries, based on the corresponding query session(s). The ability to distinguish multi-click queries can potentially help the search service to instantly adapt its search response, for example by clustering of search results or presentation of a "digest" of opinions.

Our analysis is based on over 30 million query sessions, sampled from the query log of a commercial Web search engine over a period of two weeks. The log includes English queries, submitted from the United States. Based on our model, we identify multi-click queries in the log, which account for 6.5% of all query sessions and 11.4% of all unique queries. We compare the multi-click queries with the rest of the queries by various characteristics, including the context (e.g., time-of-day, device type, user's age and gender), the SERP (e.g., result scores, number of unique domains), the

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983856

[1]We consider each such interaction an independent session; as opposed to a multi-query session, a sequence of interactions by the same user, in a short time interval, is not joined together.

clicks, and, primarily, the query text itself, by both syntactic and lexical analysis. Our comparison reveals a variety of unique features characterizing multi-click queries. For example, we discover that the plural form of nouns is used substantially more frequently on multi-click queries, either intentionally or sub-consciously. To further examine which characteristics can contribute to a predictive performance, we develop a simple classifier, trained and tested over a balanced dataset (50% multi-click queries, 50% other queries). The classifier achieves 75% accuracy, with over 80% recall and 70% precision for the multi-click class, providing an initial indication that multi-click query prediction is plausible. Query and SERP features are shown to have a promising predictive performance, while context features do not show such a potential.

Overall, our work offers the following key contributions:

- We introduce a simple formalization of the notion of multi-click queries.
- We present a comprehensive analysis distinguishing multi-click queries from the rest of the queries by different characteristics, spanning the query's text, context, and SERP.
- We show that some of these characteristics can be used to predict multi-click queries, over a balanced dataset, with 75% accuracy.

Our findings suggest that search systems can enhance their support and take advantage of the unique features of multi-click queries. We conclude the paper by summarizing the key findings, discussing their implications, and offering directions for future work.

2. RELATED WORK

Multi-click queries represent complex search tasks for which the user needs cannot be satisfied by one Web page. Many works studied complex search tasks in information retrieval (e.g. [29, 37, 27, 21]). These tasks often involve multiple search queries that span multiple sessions. The general approach for handling such tasks is providing better search tools to the user, e.g., letting the user refine her query throughout the search process, or classify the search results into various facets [34].

Radlinski and Joachims [29] introduced the notion of "query chain" as a sequence of reformulated queries that express the same information need. Jones et al. [27] discussed how a sequence of queries can be mapped into the same "mission", while Donato et al. [11] defined the notion of "research missions", automatically found by tracking multi-query sessions. The on-the-fly identification of research missions has been implemented in a "search pad" that helps users keep track of results they have consulted.

Hassan and White [20] studied complex search tasks such as planning a vacation and proposed "task tours" for helping users understand the required steps to complete a task. Similarly, Raman et al. [30] improved the support for complex search tasks involving several queries having the same context, such as "vacation planning", "comparative shopping", and "literature surveys". All these approaches focused on exploration through multi-query sessions, while our work is focused on exploration through a multi-click single query session.

Several works modeled the user clicks on the search results page, aiming at capturing user behavior patterns. Joachims et al. [26] showed that the click probability depends on the rank position of the document on the search results page. Craswell et al. [10] investigated how to model click behavior assuming (1) each Web document in the search result is examined regardless of where it appears, and (2) the click probabilities at different positions are independent. Guo et al. [15, 16] expanded this model by incorporating dependencies between the user clicks. A set of position-dependent parameters were added to model the probabilities that the user returns to the search result page and resumes the examination after a click. Wang et al. [35] investigated how to properly incorporate non-sequential behavior (both examination and click) into the click models. Huang et al. [24] developed a click model for users who 'branch' to new browser tabs or windows when exploring search results. They found that users typically continue inspecting search results before looking into a branched result, which helped improve the click model for branching queries. A comprehensive overview of user click models in Web search can be found in [8]. While these works model the user click behavior, they mostly focus on estimating the click probability of a specific result, while we are interested in identifying the type of queries that lead to the multi-click behavior.

Identifying the user intent, as represented by the query, is another important research area that relates to our work [4, 36, 12, 38, 33]. Wang and Agichtein [36] measured the entropy of the click distributions of individual searchers per query to distinguish between informational and ambiguous queries. Duan et al. [12] argued that the multi-clicks on the search results page represent complex query intent that cannot be captured by considering each click separately. They used "click patterns" to capture the relationship among clicks by treating the set of clicks as a single unit. The click patterns were clustered to create a rich representation of multiple navigational and informational intents.

Another typical approach for query intent classification is using a specific language model for each domain and computing the domain's query-likelihood of the query as a selective criterion. This works quite well for many domains, however, multi-click queries are harder to distinguish by modeling the query text alone, since they cover many different topics that spread many domains. Following Tsur et al. [33], who identified Web queries with question intent using a rich set of syntactic features, we also consider syntactic features for classifying multi-click queries.

3. MULTI-CLICK MODEL

In this section, we describe our model for multi-click queries, i.e., queries typically followed by multiple clicks of the searcher on more than one search result. We denote a basic interaction of a user with the search engine as a *query session*. The query session is defined by the tuple $QS=<u, q, D, C>$, where u is the searcher-id; $q=<text, time>$ is the user query containing the query's raw text and the time-stamp of submission; D is the list of ranked results returned by the search engine, associated with their relevance score; and C is the set of clicked results, where each clicked result $c \in C$ is the tuple $c=<d, pos, time>$, where $d \in D$ is the returned result, pos is its rank position in D, and $time$ is the time-stamp of the click.

A *Multi-click Query Session* is a query session with $|C|>1$.[2] Note that according to our definition, a query may be

[2]In our analysis, we only consider clicks on organic search results ("blue links"), and discard ads, images, direct-displays,

Query (q)	$\|QS(q)\|$	0	1	2	3	4	5	6	7+
stephanie mcmahon nude	23	5	6	4	2	2	2	2	0
decades tv network schedule	19	3	6	6	3	1	0	0	0
playboy swinger videos	16	2	2	2	2	3	2	1	2
crossdressing stories	15	5	1	5	2	0	0	0	0
unfriended torrent	14	4	2	3	3	1	1	0	0

Table 1: Popular *MCQ*s and number of clicks distributed across related query sessions.

associated with many query sessions, as the same query may be submitted by different users, in different times, and the search results may vary according to the user's characteristics. For defining the notion of a multi-click query, it is therefore reasonable to assert that a query reflects a multi-click behavior if a substantial portion of its associated query sessions are multi-click query sessions. Formally, we define multi-click queries as follows:

DEFINITION 3.1 (MCQ). *Given a query q and a value $0 \leq p \leq 1$, we define $QS(q)$ to be the set of all query sessions associated with q, and $MCQS(q)$ as the set of all multi-click query sessions associated with q. The query q is a* Multi-Click Query (MCQ) *w.r.t p, if its fraction of associated query sessions that are multi-click query sessions is at least p, i.e., $\frac{|MCQS(q)|}{|QS(q)|} \geq p$.*

After experimenting with various values of p, we opted to set it to 0.5, meaning a query q will be considered *MCQ* if at least half of $QS(q)$ are multi-click query sessions. Note that according to this definition head queries, such as *facebook* or *youtube*, will not be deemed as *MCQ*s, even though they have associated multi-click query sessions, e.g., due to clicks performed carelessly or by mistake. While the absolute number of multi-click query sessions related to such head queries may be high, due to their popularity, the portion out of all related query sessions is low. Table 1 demonstrates some popular multi-click queries, and the distribution of number of clicks across their related query sessions.

In the remainder of this paper, we denote *MCQ* as the set of multi-click queries, and *SCQ (Sparse Click Queries)* as the complementary set, including all other queries, i.e., queries such that the fraction of their associated query sessions having only zero or one clicks, is larger than 0.5.

3.1 Dataset

Our dataset includes 31.42 million query sessions, with English queries only, sent to a popular Web search engine in the United States, which have been sampled at random between May 1^{st} and May 14^{st}, 2015. For logged-in users (about a third of all query sessions), age and gender are given. In addition, information about the search results is included, in particular the URL, page title, and the relevance score for each of the top 10 results.

The choice of the threshold for defining *MCQ* to be p=0.5 allowed us to retain 87% of the multi-click query sessions as belonging to *MCQ* and, on the other hand, considerably reduce the amount of noise: only 7.3% of the query sessions associated with *MCQ*s have zero or one clicks. Overall, the portion of query sessions associated with *MCQ*s in our dataset is 6.5%, while the portion of unique queries deemed as *MCQ*s is 11.4%.

and other types of non-organic results. We also discard repeated clicks on the same rank position within the same query session, and only consider multiple clicks when they are performed on different results.

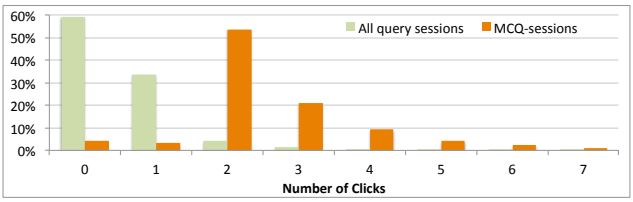

Figure 1: The distribution of number of clicks for all query sessions and for query sessions associated with *MCQ*s.

	All	MCQ
% Unique queries with 1 query session	88.7%	97.7%
% Unique queries with 5+ query sessions	2.28%	0.07%
Maximum query sessions per unique query	621,535	23
% Query sessions s.t. $\|QS(q)\|=1$	46.7%	94.7%
% Query sessions s.t. $\|QS(q)\|\geq 5$	42.1%	0.45%
% Query sessions s.t. $\|QS(q)\|\geq 1000$	17.4%	0

Table 2: Search results characteristics. (Top:) Distribution of query sessions over unique queries. (Bottom:) Distribution of query sessions over all queries.

Figure 1 presents the distribution of the number of clicks per query session versus the distribution of the number of clicks over query sessions associated with *MCQ*s (*MCQ*-sessions). It can be clearly seen that the separation works quite well: while the vast majority of all query sessions (92.9%) have zero or one clicks, the majority of *MCQ*-sessions (92.6%) have multiple clicks.

Table 2 (upper section) presents statistics about the distribution of query sessions across unique queries in our dataset, for all queries and for *MCQ*s only. Of all unique queries, 88.7% occurred only once, i.e., they are associated with only one query session. Only 2.28% of the unique queries relate to five query sessions or more. This gives a strong indication of the very long tail of user queries. On the other hand, the most popular query is *facebook*, and is associated with nearly 2% of all query sessions in the dataset (over 620K). This distribution is even more skewed for *MCQ*s: 97.7% are associated with only one query session (occurred only once).

The lower section of Table 2 presents the distribution of query sessions over all (non-unique) queries, which is much flatter. While 46.7% of the query sessions account for queries that occurred only once ($|QS(q)|=1$), 42.1% account for queries that occurred at least 5 times ($|QS(q)|\geq 5$), and 17.4% account for queries that occurred 1000 times or more ($|QS(q)|\geq 1000$). The situation across *MCQ*s is very different in this case: the portion of query sessions with $|QS(q)|=1$ is very high at nearly 95%, while only a few are associated with queries that occurred 5 times or more. This demonstrates that *MCQ*s account for long-tail queries, whose number of associated query sessions is low.

4. MCQ CHARACTERISTICS

This section characterizes *MCQ*s in comparison with *SCQ*s. The analysis examines the queries themselves, both syntactically and lexically; the clicks, e.g., their rank positions and popular domains; contextual properties, such as time-of-day, device type, and users' age and gender; and SERP signals, such as result scores and number of unique domains.

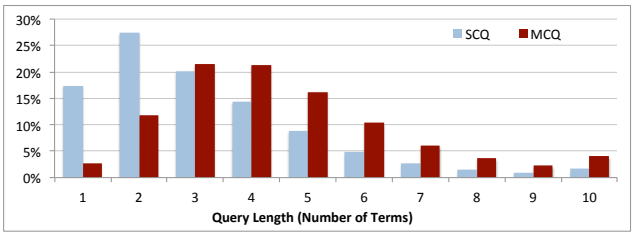

Figure 2: Distribution by query length.

Query length		1	2	3	4	5	6	7	8	9	10+
%WH	SCQ	0.01	0.03	0.44	1.8	5.1	11.7	19.0	27.8	33.3	34.4
	MCQ	0.00	0.01	0.32	1.5	4.5	10.1	17.8	26.1	32.1	36.8
%WH+ Y/N	SCQ	0.16	0.07	0.58	2.4	6.3	13.9	23.1	33.5	40.1	42.6
	MCQ	0.12	0.04	0.48	2.1	5.9	12.7	22.1	32.2	40.2	46.7

Table 3: Percentage of question queries by query length.

4.1 Query Analysis

A major part of our analysis focuses on the queries. We examine both syntactic aspects, such as query length, question-phrased queries, and part-of-speech, and lexical aspects, including characterizing terms and query categories.

4.1.1 Query Syntax

Query Length. Figure 2 presents the query length distribution (number of terms, based on white-space tokenization) for *MCQ*s versus *SCQ*s across all query sessions in our dataset. It can be seen that the two distributions are quite different. For example, for *MCQ*s, the portion of one-term queries is very low: 2.7%, compared to 17.4% for *SCQ*s. On the other hand, the portion of long queries is substantially higher for *MCQ*s: 42.8% of the *MCQ*s are verbose queries (having 5 or more terms [17]), compared to only 20.8% of the *SCQ*s. Overall, the average query length for *MCQ*s is 4.8 (stdev: 1.8, median: 4) versus 3.3 for *SCQ*s (stdev: 1.4, median: 3). This substantial gap can be explained by the fact that verbose queries often indicate complex information needs [17], which are more likely to yield exploratory user behavior reflected by multiple clicks.

Question Queries. As we found that *MCQ*s are substantially longer than average queries, we set out to explore how many of them are phrased as questions, taking a similar approach to a recent study on "question queries" [38]. Overall, we found that **6.5%** of the *MCQ*s start with a WH-question word[3], compared to only **3.3%** of the *SCQ*s. Considering also yes/no queries (start with "is", "do", "can", etc. [38]), the portions increase to **8.3%** versus **4%**, respectively. Table 3 presents this comparison by query length, which indicates that the difference between *MCQ*s and *SCQ*s is due to their length differences: for a fixed query length, the portions of WH-questions and yes/no questions are rather similar. Another indication, albeit sparse, of the association of *MCQ*s to questions, is the existence of a question mark: 0.8% of the MCQs contain a question mark, compared to only 0.37% of the *SCQ*s.

Part-of-Speech Distribution. Part-of-speech (POS) tagging associates each word in a given text with a corresponding part of speech, such as a noun, verb, or adjective. For analyzing POS tag distribution across queries, we used the OpenNLP[4] toolkit to train a specialized model for Web queries, following the method described in [13]. Figure 3 shows the portion of queries that include each of the seven main POS tags. It can be seen that almost every query includes a noun, for both *MCQ*s and *SCQ*s. Looking more closely into the three main types of nouns, we observe different behaviors: common nouns in their singular (or mass) form (NN) are included in more *MCQ*s than *SCQ*s – **72.3%** versus **61.7%**, respectively. For common nouns in plural form (NNS), the gap is even more substantial at **36.7%** versus **24.1%**. Proper nouns (NNP), on the other hand, were less common on *MCQ*s than on *SCQ*s – **47.1%** versus **50.9%**, respectively.

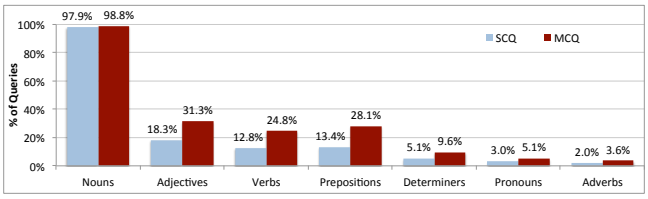

Figure 3: Percentage of queries containing main part-of-speech tags

The other six POS tags, as can be seen in Figure 3, were included in substantially higher portions of the *MCQ*s than the *SCQ*s, indicating a richer language used in the former, which should come as no surprise given their higher average length. In spite of length being an inherent characteristic of *MCQ*s, we also set out to explore POS tag distribution of *MCQ*s versus *SCQ*s of the same length. Table 5 compares the occurrence of POS tags in *MCQ*s versus *SCQ*s when controlling for the query length. The first section refers to common nouns in their singular (or mass) form (NN). It can be seen that for a given query length, the portion of queries containing a term POS-tagged as NN is similar for *MCQ*s and *SCQ*s. This indicates that the general difference we have seen between the two can be mostly explained by the difference in query length. For common nouns in plural form (NNS), a consistent difference can be observed: even for a fixed query length, higher portions of the *MCQ*s contain a plural noun. This suggests that users tend to express their need for multiple results by using plural rather than singular form, as in the query *senior people jokes* or *travel tips florence*. Since we observed that the plural noun often appears at the end of the query, we also compared the portion of *MCQ*s versus *SCQ*s ending with an NNS. A notable difference was found, at **19.6%** for *MCQ*s versus **14.2%** for *SCQ*s.

The third section of Table 5 refers to proper nouns (NNP). Interestingly, as opposed to other POS tags, the portion of queries containing NNP does not have a clear growth trend with the number of terms. It can be seen, both for *SCQ*s and *MCQ*s, that the portion mildly increases and reaches a maximum at 5 terms, and then decreases again. This implies that the likelihood of a query to involve a named entity (e.g., a place, a person, or a product name) does not sharply increase with its length. As for the comparison of *MCQ*s and *SCQ*s, a consistent difference across all query lengths can be observed in favor of *SCQ*s. A possible explanation is that queries that involve a proper noun have a more focused information need, which can more often be addressed in a

[3]Who, where, why, when, how, what, and which.

[4]http://opennlp.apache.org

Unigrams		Bigrams		Opening		Ending	
SCQ	**MCQ**	**SCQ**	**MCQ**	**SCQ**	**MCQ**	**SCQ**	**MCQ**
facebook	how	yahoo mail	for sale	facebook	how	com	sale
yahoo	sale	facebook login	how to	yahoo	what	facebook	reviews
google	what	facebook com	in the	google	can	google	videos
login	sex	yahoo com	for a	youtube	why	mail	tumblr
mail	can	google com	sale in	craigslist	free	login	tube
youtube	is	facebook sign	can you	gmail	best	youtube	sex
craiglist	does	google maps	how much	www	is	yahoo	pics
gmail	do	wells fargo	how long	amazon	does	craigslist	pdf
bank	i	sign up	of the	ebay	mature	gmail	recipe
amazon	free	com login	is the	mapquest	wife	amazon	download
ebay	best	bank of	on a	chase	gay	bank	torrent
airlines	women	www facebook	do i	hotmail	where	airlines	problems

Table 4: Most distinctive terms by KL divergence.

Query length		1	2	3	4	5	6	7	8	9	10+
%NN	SCQ	42.5	53.2	66.8	70.1	74.5	78.7	83.3	86.5	89.6	93.5
	MCQ	35.6	55.2	66.8	71.9	76.4	81.1	84.9	88.4	91.0	94.9
%NNS	SCQ	7.2	16.4	27.1	33.1	36.8	39.6	40.2	41.4	40.5	47.6
	MCQ	14.4	25.2	32.5	37.6	40.7	42.9	44.4	44.8	44.9	48.2
%NNP	SCQ	43.1	54.5	49.4	54.3	54.8	52.8	50.7	46.5	45.4	42.5
	MCQ	33.0	46.9	46.6	49.6	50.4	48.9	46.3	43.1	41.3	38.1
%JJ*	SCQ	3.0	10.4	18.4	24.6	29.0	33.0	35.5	38.7	40.4	52.5
	MCQ	5.5	14.5	23.4	28.6	32.6	36.2	39.7	42.3	45.6	53.4
%VB*	SCQ	0.27	0.57	1.5	2.4	3.7	5.7	7.9	9.9	11.6	12.6
	MCQ	0.38	0.94	1.7	2.6	3.9	5.7	7.8	9.7	11.2	12.5

Table 5: Percentage of queries containing different POS tags by query length.

single or no click. For example, consider a search regarding a celebrity, a specific movie, or even navigational queries where the site name represents a company name.

The fourth section of Table 5 presents the results for adjectives, in all their forms (JJ*). It can be observed, across all query lengths, that higher portions of the *MCQ*s involve an adjective. This may indicate a more subjective search [22] and also a more generic intent (e.g., running shoes versus Nike shoes).

For all other POS tags, including verbs, prepositions, determiners, pronouns, adverbs, cardinal digits, and punctuation marks, the portion of containing queries was very similar for *MCQ*s and *SCQ*s of the same length. The bottom section of Table 5 shows the results for verbs (VB*), as an example.

4.1.2 Lexical Analysis

Distinctive Terms. In order to inspect the lexical differences between *MCQ*s and *SCQ*s, we set out to explore which terms mostly characterize each of them when compared to the other. To this end, we used Kullback-Leibler (KL) divergence, which is an asymmetric distance measure between two given distributions [2, 18]. Specifically, we calculated the terms that contribute the most to the KL divergence between the *MCQ* and *SCQ* language models, for unigrams and bigrams[5]. Table 4 reports the terms with the highest KL divergence for *SCQ*s (w.r.t *MCQ*s) and for *MCQ*s (w.r.t *SCQ*s). The list of unigrams for *SCQ*s includes popular website names, used for navigational queries, while the *MCQ* unigram list includes question words (both WH and yes/no), the nouns "sale", "sex", and "women", the adjective "best", and the pronoun "i". For bigrams, the *SCQ* list again contains navigational phrases, including site names (e.g., "google maps" or "wells fargo"), login or sign-up, or URL parts. On the other hand, the *MCQ* list includes many natural language phrases ("in the", "for a"), question phrases ("how much", "can you"), and shopping phrases ("for sale", "sale in").

[5]For unknown terms, we used standard Laplace smoothing [7] .

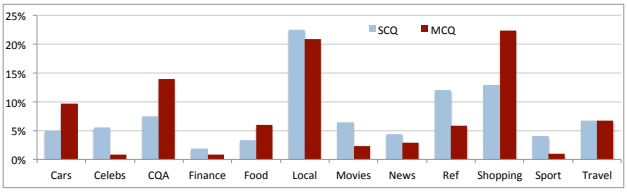

Figure 4: Distribution by query categories.

The right columns of Table 4 report the most distinctive unigrams to appear in the beginning ('opening') and end ('ending') of queries. For *SCQ*s, the most distinctive opening term was "facebook" (with other website names further down the list as well as the prefix "www"), while for *MCQ*s it was "how" (with "what" at close second; and "free" and "best" also high on the list). The distinctive terms for ending queries are particularly interesting: For *SCQ*s, they clearly reflect navigational needs, while for *MCQ*s, they reflect a diverse set of needs that includes shopping, rich media (photos, videos, and PDF documents), downloads, opinions (reviews, solutions to problems), recipes, and adult content. The use of the plural form is common in these *MCQ* terms, demonstrating the findings from our syntactic analysis.

Query Categories. To further examine the lexical differences between *MCQ*s and *SCQ*s at a higher level, we used an in-house query classification tool, which is based on named entity recognition and supervised machine learning, trained over a huge corpus of queries, categorized according to their associated clicked websites. We only considered the category with highest confidence for each query; in case the overall confidence was too low (below 50%), we discarded the query in our analysis. We further discarded queries categorized into two very broad categories ("Web" and "Search"). Overall, 54.7% of the *MCQ*s and 54.9% of the *SCQ*s could be classified according to this scheme. Their distribution across the categories is presented in Figure 4.

It can be observed that *MCQ*s were more commonly classified as shopping, community question answering (CQA), cars, and food. On the other hand, the categories Reference

	1	2	3	4	5
Max rank position	1.85	3.86	5.55	6.73	8.18
Avg rank position	1.85	2.91	3.63	4.10	4.65
Min rank position	1.85	1.96	1.85	1.71	1.48

Table 6: Mean of the average, minimum, and maximum click rank position for different number of clicks in query sessions.

(marked 'ref'; includes querying for information in encyclopedias, dictionaries, museums, and similar), sports, news, movies, finance, and celebrities (with a particularly sharp ratio), were more common in *SCQ*s. The categories local (maps, directions, and similar) and travel had similar representation in *MCQ*s and *SCQ*s. The difference between CQA and Reference queries is interesting (the former are more common with *MCQ*s and the latter with *SCQ*s). One interpretation may be that when users ask the community or the crowd, they more often look for multiple opinions, whereas in reference search they more often look for one authoritative answer. It is also possible that CQA content is of lower quality and requires more clicks from users to find what they need.

4.2 Click Analysis

In this section, we focus on click analysis. We first examine the distribution of click positions on the SERP for multi-click query sessions and then examine the clicked domains, reflecting on the query lexical analysis from the previous section.

Click Position. Table 6 presents the mean of the average, maximum, and minimum click rank position on the SERP across query sessions, split by the number of clicks. As expected, the maximum rank of a clicked result increases as the total number of clicks grows. The average position of a clicked result also increases, albeit more moderately, indicating that the rank distribution for query sessions with more clicks is less biased towards highly ranked results. The minimum rank position increases from 1-click query sessions to 2-click sessions, meaning that the top click among the two is still ranked lower, on average, than a single click. Starting at three clicks, the minimum starts to decrease.

Table 7 presents a more detailed analysis that considers the chronological order of the clicks, i.e., the 1^{st} click refers to the earliest click in time, and so forth. For queries with $i \in \{1, \ldots, 5\}$ clicks, each column shows the average position of the j^{th} click, $j \in \{1, \ldots, i\}$, while the rightmost column shows the percentage of *non-sequential query sessions* – query sessions for which the chronological order of clicks does not match the order of their rank positions (i.e., there exists at least one pair of clicks, $<c1, c2>$, where $c1$ was performed before $c2$, but is ranked lower on the SERP). Observing the average rank of the first click, it substantially increases from 1-click to 2-click query sessions, and slightly further for 3-click sessions, probably as users are less satisfied with the top results. As the number of clicks grows beyond three, the rank position of the first decreases back to some extent. A similar trend can be observed for the second click and so forth.

The portion of non-sequential query sessions naturally increases with the number of clicks, from 16% for 2-click sessions to over 50% for 5-click sessions. This indicates that users do not always click according to the presented order of the results, in agreement with the click analyses presented in [24, 35], suggesting there is a room for improvement in result ranking and presentation for *MCQ*s.

#clicks	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	%non-sequential
1	1.85					0
2	2.30	3.52				16.1%
3	2.35	3.66	4.89			31.7%
4	2.29	3.53	4.75	5.81		43.1%
5	2.14	3.21	4.29	5.28	6.13	55.3%

Table 7: Average rank position of the 1^{st} to 5^{th} clicks, and the percentage of non-sequential query sessions.

	SCQ	MCQ	
	Domain	Domain	Ratio
1	www.facebook.com	en.wikipedia.org	0.55
2	en.wikipedia.org	answers.yahoo.com	2.33
3	www.google.com	www.amazon.com	1.27
4	www.youtube.com	www.pornhub.com	1.57
5	mail.yahoo.com	www.facebook.com	0.15
6	www.ebay.com	www.youtube.com	0.46
7	www.amazon.com	www.ebay.com	1.01
8	www.yahoo.com	xhamster.com	2.43
9	www.pornhub.com	www.answers.com	2.03
10	answers.yahoo.com	www.yellowpages.com	1.77

Table 8: Most clicked domains.

Clicked Domains. Table 8 presents the top clicked domains for *SCQ*s versus *MCQ*s (a domain is determined by the 'host' part of the result's URL). The rightmost column shows, for each of the top *MCQ* domains, the ratio between its percentage of clicks in the *MCQ* sample and the percentage of clicks in the *SCQ* sample (a ratio higher than 1 indicates a higher percentage of clicks on the respective domain in the *MCQ* sample). It can be seen that while *facebook.com* is the most clicked domain on *SCQ*s, it is only 5th on *MCQ*s, with a particularly low ratio of 0.15. *En.wikipedia.org*, the second most clicked on *SCQ*s, is the first on *MCQ*s, yet with a ratio of 0.55. It should be noted that the *MCQ* domain distribution is flatter: the top 4 domains in the *SCQ* list account for 12.9% of all clicks, while the top 4 domains in the *MCQ* list cover only 4.5% of all clicks, again indicating the long-tail nature of *MCQ*s.

Two types of sites emerge as significantly more common with *MCQ*s. The first are CQA sites, coinciding with our lexical analysis: *answers.yahoo.com* is 2nd on the *MCQ* list with a ratio of over 2, and *answers.com* is 9th with a high ratio. The second is adult sites: *Pornhub* and *Xhamster* are at 4th and 8th, respectively, both with a high ratio.

Inspecting further down the lists reveals that *MCQ*s led to more clicks on food related sites (e.g., *food.com* with a ratio of 2.37, *allrecipes* 1.39), different health sites (e.g., *webmd* 1.38, *nih.gov* 1.56, *medhelp* 2.53), travel (e.g., *tripadvisor* 1.58), real-estate (*trulia* 1.9), and CQA sites (*ehow* 2.61, *wikihow* 1.83). On the other hand, *SCQ* clicks were more common for finance and banking (e.g., *bankofamerica* 0.05), news (*cnn* 0.36), sports (*espn.go* 0.29), maps (*mapquest* 0.28), retail (*walmart* 0.49), and other focused intents such as weather (*weather.com* 0.22) or dictionary (*dictionary.reference.com* 0.51). Many of these domains coincide with the categories found most common for *SCQ*s in Section 4.1.2. As opposed to *Facebook*, clicks on *LinkedIn* were more common on *MCQ*s (ratio 1.48), perhaps due to the ambiguity when looking for "ordinary" professionals.

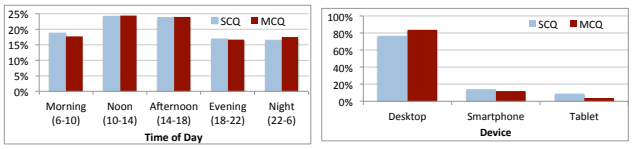

Figure 5: Distribution by (left:) time of day and (right:) device type.

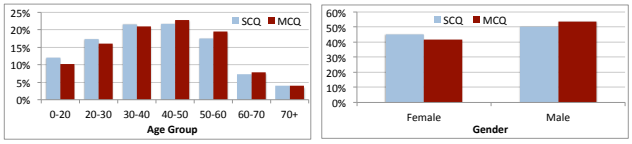

Figure 6: Distribution by (left:) age and (right:) gender.

4.3 Contextual Properties

In this section, we compare the query sessions associated with *MCQ*s and *SCQ*s across various contextual characteristics.

Temporal Aspects. Inspecting the day-of-week, we found a slightly higher portion of the *MCQ*s were performed on weekends compared to *SCQ*s (up 3%). Figure 5 (left plot) shows the distribution of *MCQ*s and *SCQ*s across time-of-day. It can be seen that the *MCQ* portions are somewhat lower in the morning and higher at night. We conjecture that at night users may have more time to perform exploratory searches that require more attention and interaction (e.g., CQA or adult queries, as we have seen in the lexical analysis), while in the morning users are busier and tend to perform more ad-hoc searches (e.g., navigational queries or queries with direct answers).

Device type. Our sample was taken from Web search logs and does not include searches from native mobile applications. Therefore, the majority of searches were performed from desktop devices. Figure 5 (right plot) shows the distribution of queries by device type. It should come by no surprise that *MCQ*s are more common on desktop devices, which allow viewing more results at a time and easier exploration. For smartphones, and even more so tablets, the *SCQ* potions are higher.

User attributes. Figure 6 (left plot) shows the query distribution based on different age groups (for logged in users only). The portions of *SCQ*s are higher at younger ages (below 40), while for older ages *MCQ*s have higher portions, with a difference peak at ages 50-60. One possible explanation may be that people at these ages have more free time to perform exploratory searches.

The right plot of Figure 6 shows the query distribution by gender. It can be seen that men have somewhat higher *MCQ* portions. This difference can be explained by the use of adult queries, which are more common with men (our data shows a nearly 1:3 ratio). Further inspection of the gender differences reveals that they emerge on desktop only and are particularly prevalent at ages 50-80 (for smartphones and for ages younger than 20 there are even slightly higher *MCQ* portions for women). For instance, a particularly high difference between men and women was observed for desktop users at the ages 60-70 (for men, 8.1% of their queries were *MCQ*s, while for women only 6.2%).

	Average (Stdev)		Median	
	SCQ	MCQ	SCQ	MCQ
1st result score	16.54 (16.78)	7.76 (14.45)	15.46	6.59
Avg result score	1.47 (10.43)	-1.83 (10.15)	0.69	-2.25
NQC	4.44 (3.28)	2.65 (2.29)	3.6	2.07
1st result sim(q,title)	0.37 (0.2)	0.33 (0.19)	0.33	0.3
Avg result sim(q,title)	0.27 (0.13)	0.24 (0.12)	0.27	0.22

Table 9: Statistics of search result scores and query-title similarity.

4.4 Search Results Page

In this section, we compare various characteristics of the search engine results page (SERP) between *MCQ*-sessions and *SCQ*-sessions. SERP characteristics are also commonly referred to as post-retrieval parameters, as they build on the search engine's retrieval technology. They are often used for other types of query analysis, such as query performance prediction [5].

Table 9 presents a comparison of two search result characteristics between *MCQ*s and *SCQ*s: the score of the results and their title similarity to the query. The upper section compares properties of the search result scores. The search result score is a real number (positive or negative) assigned by the search engine to each result. While we have no information about the intrinsic formula, it can be safely assumed that a higher score reflects a "better" result, with a higher confidence in its relevance to the user. It can be seen that both the score of the top result and the average score across all 10 results are substantially lower for *MCQ*s. We also report the *Normalized Query Commitment* (NQC) – a query performance predictor, calculated as the standard deviation of the result scores, normalized by the score of a document representing the corpus [31]. It can be seen that the NQC is substantially lower for *MCQ*s, as lower NQC reflects lower quality of the search results, which coincides with more difficult queries.

The lower part of the table compares the textual similarity of the result's Web page title and the query. To this end, we applied a simple Jaccard-based word similarity. It can be seen that for *MCQ*s, the query-title similarity is slightly lower, both for the top result and when averaged across all 10 SERP results.

Number of Domains. Figure 7 shows the query distribution by the number of unique URL domains within the SERP. Interestingly, it can be seen that multi-clicks are more common when the SERP includes 7-9 unique domains. We conjecture that a repeated domain within the top results may lead to multiple clicks on the results from that domain. Also, a repeated domain may be more common for more complex or specific queries. For example, a technical query may yield multiple results from the CQA site StackOverflow. Inspecting query sessions with 10 unique domains, we observed that the percentage of 1-term queries, which are often navigational, is particularly high, at 30.6%, compared to 16.5% across all query sessions. On the other hand, a particularly low number of unique domains may often be caused by an explicit mention of the desired site in the user's query, which often coincides with a navigational intent. A few examples from our dataset are: *www.bankofamerica.com*, *allrecipes.com recipe search*, and *ebay.com usa*). Inspecting query sessions with 5 or fewer domains, we indeed found

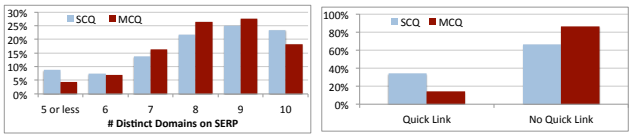

Figure 7: Distribution by (left:) number of domains and (right:) the presence of a quick link on the SERP.

Classifier	Accuracy	MCQ Precision	MCQ Recall
AROW	**75.2%**	**72.4%**	81.2%
Logistic Regression	70.8%	71.0%	70.2%
Random Forest	74.8%	70.3%	**86.0%**

Table 10: Performance of AROW, Logistic Regression, and Random Forest over a balanced dataset.

that the portion of queries containing the suffix "com" and the prefix "www" is particularly high, at 3% and 12.8%, respectively, compared to 1.7% and 7.2%, across all queries.

Quick Link. Search engines improve their user experience by features allowing the user to get a direct or more detailed response. One example is the *quicklink*: a set of shortcuts displayed below the website homepage on a SERP, which let the users directly jump to selected points within the website [6]. Figure 7 shows the query distribution by the occurrence of a quicklink on the SERP. As expected, since a quicklink is a strong characteristic of navigational queries, a substantially higher portion of the *SCQ*s have a quicklink on their SERP. For *MCQ*s, the portion is smaller, but still greater than zero. For example, a query for a hotel name may present a quicklink for the hotel's website, but the user may also inspect the hotel's page on travel sites, such as TripAdvisor or Booking.com.

5. MCQ PREDICTION

In order to examine the practical implications of the differences revealed in the previous section, we conducted initial experimentation with predicting *MCQ*s. We approach *MCQ* prediction as a binary classification task. To this end, we experimented with three common classifiers: *Logistic Regression*, *Random Forest* (both implemented under the Weka workbench [19]), and *AROW (Adaptive Regularization of Weights)* [9], an online variant of linear SVM (in-house implementation).

As previously reported, query sessions that involve *MCQ*s are relatively sparse – a random sample of query sessions is expected to include about 6.5% *MCQ*-sessions. Common classifiers, as well as common evaluation metrics for binary classification, do not typically work well with such imbalanced data, and there are various techniques to approach classification in such cases [23]. While building a full-fledged classifier is beyond the scope of this paper, we set out to examine which of the *MCQ* properties are useful for prediction. To this end, we opted to under-sample the data and create a balanced dataset. We randomly sampled 500,000 *MCQ*-sessions and 500,000 *SCQ*-sessions that occurred between May 1^{st} and May 14^{th}, as a training set, and a similar set of 1 million query sessions that took place between May 15^{th} and May 21^{st}, as a test set.

5.1 Feature Representation

Each query session is represented as a feature vector. Our features correspond to the analysis presented in Section 4, and span the three categories: query, context, and SERP. Overall, we extracted 1446 features, mapped to these three categories, and several families within each category, as follows (see Table 11):

- **Query**. Features derived from the query's text: (a) surface descriptors: the number of terms, characters, stopwords, punctuation marks, and binary indicators for the use of each of the WH question words and a question mark; (b) POS tags: a set of binary indicators for the presence of specific POS tags and coarse POS tags in the query (e.g., VBN and VB*, respectively). Additionally, we included binary indicators for the presence of POS tags at a specific position (e.g., NNS_1 or NNS_{last}, indicating NNS in the first or last position of the query, respectively). This feature family is by far the largest and contains 1352 features; (c) language model: we trained two models based on independent sets of 1M *MCQ*s and 1M random queries, respectively, and assigned three scores per language model: the length-normalized log probability of the whole query text; the maximum log probability of a term in the query; and the minimum log probability of a term in the query. We also included the difference between the two language model scores for the whole query[6].
- **Context.** Context-based features, all represented as binary indicators per each categorical value: (a) time: day-of-week and time-of-day; (b) user: device type, age, and gender.
- **SERP**. Post-retrieval features, derived from the output of the search engine: (a) the scores assigned to the results by the search engine (various statistics such as maximum, minimum, average, standard deviation, and NQC); (b) the textual similarity between the result titles and the query's text (various statistics); (c) the number of unique domains on the SERP (binary indicators from 5-or-less to 10); and (d) the presence of a quick link.

5.2 Results

Table 10 presents the accuracy, as well as the precision and recall for the *MCQ* class, achieved by the three classifiers. AROW and Random Forest reached a similar accuracy of about 75%, while Logistic Regression achieved nearly 71%. In terms of precision and recall, AROW achieved the highest precision at 72.4%, while Random Forest reached the highest recall at 86%.

In order to examine the contribution of each feature category and its corresponding families to the classifier, we conducted two more experiments. In the first, a single feature family (or category) was used for representing the query session and in the second, a single family (or category) was excluded from the query session representation (ablation

[6]In another attempt to capture lexical characteristics of the query text, we experimented with lexical features representing the occurrence of each term in the query (i.e., a binary indicator for each term in the vocabulary, which appears in at least k queries; we experimented with different values of k). Since these did not yield a substantial performance improvement, but produced a large number of sparse features, we report the results of the classifiers that did not make use of such lexical features.

Feature	Count	Accuracy	Ablation
Query - All	1,376	67.1%	72.7%
POS tags	1,352	67.4%	74.6%
Surface	17	66.8%	74.1%
Language model	7	64.1%	74.3%
Context - All	37	54.7%	75.1%
User	22	54.2%	75.2%
Time	15	51.0%	75.0%
SERP - All	33	**72.5%**	**68.6%**
Result scores	16	64.8%	73.2%
Number of domains	6	62.0%	74.3%
Quick link	1	60.9%	74.2%
Textual similarity	10	55.4%	74.6%

Table 11: Performance results using or excluding specific feature families, with AROW classifier over a balanced dataset.

test). The first experiment evaluates the contribution of each family/category on its own, while the second evaluates its contribution on top of all other features. In both cases, the classifier was re-trained and re-tested using the modified representation. For these experiments, we used the AROW classifier, which achieved the highest accuracy and precision at the overall classification task.

Table 11 presents the results. The category of SERP features shows the greatest contribution to the classification accuracy, both when used alone, reaching an accuracy of 72.5%, and in the ablation test, where we can see a relative drop of nearly 9% when SERP features are excluded. The query features are the next in importance – relying only on the query's text allows reaching over 67% accuracy. Lastly, the context features show little contribution – on their own they reach accuracy only moderately higher than 50%, and their exclusion hardly affects performance. Indeed, we have seen in Section 4.3 that the differences between *MCQ*s and *SCQ*s for these features were relatively small.

Inspecting the feature families, we see that all three feature families of the query category reach a relatively high accuracy on their own, with POS tags reaching the highest, followed by the surface features. Apparently, using a basic set of text characteristics (length, punctuation marks, question words, etc.) already yields accuracy of just over two thirds. The accuracy drop in the ablation test is not large for each family separately, probably due to the overlap among the query feature families (e.g., query length and question words are also covered by POS tags, language models also reflect some surface descriptors and POS tags). For SERP features, the result score family emerges as most important, followed by the number of domains and quick link. The latter on its own allows reaching nearly 61% accuracy, as it effectively captures navigational queries. Again, ablation tests per family show only a mild decrease, implying some dependency among the SERP features (e.g., result score and query-title text similarity).

6. DISCUSSION AND FUTURE WORK

In this section, we summarize our findings and discuss implications, limitations, and directions for future work.

Our analysis revealed a variety of unique characteristics of multi-click queries, as compared to the rest of the queries. *MCQ*s tend to be substantially longer, with an average length of nearly five words. As such, their language is richer, with higher diversity across part-of-speech tags. Proper nouns, which may indicate a more specific intent, are less common on *MCQ*s, while the use of plural nouns, implying a desire for a variety of results, is substantially more common, in particular at the end of the query. We also found evidence, by analyzing both the queries and the clicks themselves, that *MCQ*s tend to focus on domains that often require research and exploration, such as recipes, health, shopping, community question-answering, and adult content. On the other hand, *MCQ*s are less likely to aim for news, sports, movies, celebrities, as well as to be used for navigational search, ad-hoc search (e.g., weather, definitions), or search for authoritative answers (e.g., from Wikipedia). In terms of context, *MCQ*s showed a slight tendency towards search over the weekend and at night, and search by men and by 40–70 year-olds. As could be expected, *MCQ*s are more common on desktop; yet, they are not absent from smartphones. From a SERP's perspective, the set of results for *MCQ*s tend to be more diverse, without a clear "winner", which implies more difficulty in addressing the query [5]. In particular, the substantial gap in normalized query commitment tightens the association of *MCQ*s with the notion of difficult queries [31].

Our feature analysis, as part of a basic *MCQ* classification task, is generally in agreement with the descriptive findings. The differences in query syntax and language model, as well as in SERP characteristics, show a potential predictive performance, while the contextual differences (time, gender, age) are too minor to contribute.

Commercial Web search engines put most of their efforts on effectively addressing popular information needs, which often represent simple look-ups, rather than on tail queries, which often represent more complex and specialized needs [32]. Yet, it is those more complex and specialized searches in which users are likely to engage the most, remember the most, and get the most value from when successful [32]. Our analysis shows that multi-click queries all account for the long tail – the most popular *MCQ* repeated only 23 times in a log of over 30 million records. Multi-click queries therefore provide search engines an opportunity to identify a substantial set of tail queries, which may be better treated by means such as enhanced ranking (e.g., by considering diversity differently; our analysis shows many of the clicks are currently performed non-sequentially), clustering of search results, and summary of key aspects in the result set.

As one concrete example, consider the case of queries with question intent, typically satisfied by a CQA vertical, which have recently been shown to account for roughly 10% of Web search traffic [33, 38]. For such queries, it may be desirable to distinguish between those that reflect an information need for one authoritative answer, such as "*what is the purpose of the cuticle*", "*where do grizzly bears live*", or "*how to delete apps on iphone*" and queries that reflect a need for a variety of opinions, such as "*birthday gift ideas for 10 years old boy*", "*why does my heel hurt when i run*", or "*how to spend 12 hours in Rome*"[7]. For the former, a single direct answer, which may satisfy the searcher's need, can be presented at the top of the SERP [3], while for the latter, a digest of ideas or opinions, or a summary of key aspects, might be more useful for the searcher.

[7]Actual examples from our query log.

Our *MCQ* classifier was trained and tested over a balanced dataset. In practice, however, we saw that the portion of *MCQ*s is substantially lower than the rest of the queries, which might make the prediction task harder. Future research should further explore this challenge, by means such as using more advanced machine learning or extending the set of features (e.g., by considering the previous queries and clicks in the multi-query session). In addition, the specific application in which the *MCQ* classifier is used may affect the balance between the classes. Going back to our example from the previous paragraph, if the classifier is intended to focus on queries with question intent, typically known to be verbose tail queries [33], the ratio between multi-click queries and their complementary class is expected to become more even.

As discussed in Section 2, previous work has examined complex search tasks, which involve multiple queries; *MCQ*s may often take part in such complex tasks, as smaller "building blocks", and can help identify them more easily. Future research should further explore the connection between *MCQ*s and the broader notion of exploratory search tasks, and address research questions such as: how many of the tasks include *MCQ*s and in which stage? in what reformulation patterns are *MCQ* involved? and how do click patterns on *MCQ*s affect the success of their corresponding tasks?

Our analysis did not consider dwell time – the amount of time a user spends on a clicked page. Going forward, dwell time analysis may help refine the definition of *MCQ*s and their distinction from the rest of the queries. Future *MCQ* research should examine dwell time, as well as other types of user behavior information, such as scroll depth and mouse movement [1]. Furthermore, our research method focused on analyzing a large-scale Web search query log; future research may refine our findings by conducting other types of studies, with additional scientific tools, such as eye tracking or user interviews.

7. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of SIGIR*, pages 19–26. ACM, 2006.

[2] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proceedings of SIGIR*, pages 222–229, 1999.

[3] M. S. Bernstein, J. Teevan, S. Dumais, D. Liebling, and E. Horvitz. Direct answers for search queries in the long tail. In *Proceedings of CHI*, pages 237–246. ACM, 2012.

[4] D. J. Brenes, D. Gayo-Avello, and K. Pérez-González. Survey and evaluation of query intent detection methods. In *Proceedings of WSCD*, pages 1–7. ACM, 2009.

[5] D. Carmel and E. Yom-Tov. Estimating the query difficulty for information retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2(1):1–89, 2010.

[6] D. Chakrabarti, R. Kumar, and K. Punera. Quicklink selection for navigational query results. In *Proceedings of WWW*, pages 391–400. WWW, 2009.

[7] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, 1999.

[8] A. Chuklin, I. Markov, and M. d. Rijke. Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 7(3):1–115, 2015.

[9] K. Crammer, A. Kulesza, and M. Dredze. Adaptive regularization of weight vectors. *MLJ*, 91(2):155–187, 2013.

[10] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of WSDM*, pages 87–94. ACM, 2008.

[11] D. Donato, F. Bonchi, T. Chi, and Y. Maarek. Do you want to take notes?: Identifying research missions in Yahoo! search pad. In *Proceedings of WWW*, pages 321–330. WWW, 2010.

[12] H. Duan, E. Kiciman, and C. Zhai. Click patterns: An empirical representation of complex query intents. In *Proceedings of CIKM*, pages 1035–1044. ACM, 2012.

[13] K. Ganchev, K. Hall, R. McDonald, and S. Petrov. Using search-logs to improve query tagging. In *Proceedings of ACL*, pages 238–242. ACL, 2012.

[14] S. Goel, A. Broder, E. Gabrilovich, and B. Pang. Anatomy of the long tail: Ordinary people with extraordinary tastes. In *Proceeding of WSDM*, pages 201–210. ACM, 2010.

[15] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y.-M. Wang, and C. Faloutsos. Click chain model in web search. In *Proceedings of WWW*, pages 11–20. ACM, 2009.

[16] F. Guo, C. Liu, and Y. M. Wang. Efficient multiple-click models in web search. In *Proceedings of WSDM*, pages 124–131. ACM, 2009.

[17] M. Gupta and M. Bendersky. Information retrieval with verbose queries. *Foundations and Trends in Information Retrieval*, 9(3-4):209–354, 2015.

[18] I. Guy. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of SIGIR*, pages 35–44. ACM, 2016.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[20] A. Hassan and R. W. White. Task tours: Helping users tackle complex search tasks. In *Proceedings of CIKM*, pages 1885–1889. ACM, 2012.

[21] A. Hassan Awadallah, R. W. White, P. Pantel, S. T. Dumais, and Y.-M. Wang. Supporting complex search tasks. In *Proceedings of CIKM*, pages 829–838. ACM, 2014.

[22] V. Hatzivassiloglou and J. M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of COLING*, pages 299–305. ACL, 2000.

[23] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, 21(9):1263–1284, Sept. 2009.

[24] J. Huang, T. Lin, and R. W. White. No search result left behind: Branching behavior with browser tabs. In *Proceedings of WSDM*, pages 203–212. ACM, 2012.

[25] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of KDD*, pages 133–142. ACM, 2002.

[26] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR*, pages 154–161. ACM, 2005.

[27] R. Jones and K. L. Klinkner. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proceedings of CIKM*, pages 699–708. ACM, 2008.

[28] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *Proceedings of SIGIR*, pages 43–50. ACM, 2009.

[29] F. Radlinski and T. Joachims. Query chains: Learning to rank from implicit feedback. In *Proceedings of KDD*, pages 239–248. ACM, 2005.

[30] K. Raman, P. N. Bennett, and K. Collins-Thompson. Toward whole-session relevance: Exploring intrinsic diversity in web search. In *Proceedings of SIGIR*, pages 463–472. ACM, 2013.

[31] A. Shtok, O. Kurland, D. Carmel, F. Raiber, and G. Markovits. Predicting query performance by query-drift estimation. *ACM Trans. Inf. Syst.*, 30(2):11:1–11:35, May 2012.

[32] G. Singer, U. Norbisrath, and D. Lewandowski. Ordinary search engine users carrying out complex search tasks. *Journal of Information Science*, 2012.

[33] G. Tsur, Y. Pinter, I. Szpektor, and D. Carmel. Identifying web queries with question intent. In *Proceedings of WWW*, pages 783–793. WWW, 2016.

[34] D. Tunkelang. Faceted search. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–80, 2009.

[35] C. Wang, Y. Liu, M. Wang, K. Zhou, J.-y. Nie, and S. Ma. Incorporating non-sequential behavior into click models. In *Proceedings of SIGIR*, pages 283–292. ACM, 2015.

[36] Y. Wang and E. Agichtein. Query ambiguity revisited: Clickthrough measures for distinguishing informational and ambiguous queries. In *HLT*, pages 361–364. Association for Computational Linguistics, 2010.

[37] R. W. White, S. M. Drucker, G. Marchionini, M. Hearst, and m. c. schraefel. Exploratory search and hci: Designing and evaluating interfaces to support exploratory search interaction. In *Proceedings of CHI EA '07*, pages 2877–2880. ACM, 2007.

[38] R. W. White, M. Richardson, and W.-t. Yih. Questions vs. queries in informational search tasks. In *Proceedings of WWW Companion*, pages 135–136. WWW, 2015.