Query Reformulation in E-Commerce Search

Sharon Hirsch[1], Ido Guy[2], Alexander Nus[2], Arnon Dagan[2], Oren Kurland[1]
[1]Technion - Israel Institute of Technology, Haifa, Israel [2]eBay Research, Netanya, Israel
sharonhi@campus.technion.ac.il,idoguy@acm.org,{alnus,ardagan}@ebay.com,kurland@ie.technion.ac.il

ABSTRACT

The importance of e-commerce platforms has driven forward a growing body of research work on e-commerce search. We present the first large-scale and in-depth study of query reformulations performed by users of e-commerce search; the study is based on the query logs of eBay's search engine. We analyze various factors including the distribution of different types of reformulations, changes of search result pages retrieved for the reformulations, and clicks and purchases performed upon the retrieved results. We then turn to address a novel challenge in the e-commerce search realm: predicting whether a user will reformulate her query before presenting her the search results. Using a suite of prediction features, most of which are novel to this study, we attain high prediction quality. Some of the features operate prior to retrieval time, whereas others rely on the retrieved results. While the latter are substantially more effective than the former, we show that the integration of these two types of features is of merit. We also show that high prediction quality can be obtained without considering information from the past about the user or the query she posted. Nevertheless, using these types of information can further improve prediction quality.

ACM Reference Format:
Sharon Hirsch, Ido Guy, Alexander Nus, Arnon Dagan, Oren Kurland. 2020. Query Reformulation in E-Commerce Search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397271.3401065

1 INTRODUCTION

Search over e-commerce platforms (e.g., Alibaba, Amazon, and eBay) is a very important and challenging task, which has been attracting a growing body of research work. Various aspects of e-commerce search have been explored. For example, ranking models [8, 21, 23, 42, 54, 56], query suggestion [18], intent identification [33, 48, 49], automatic query reformulation methods [19, 33, 50], personalization [54], and approaches to predicting purchase intent [29, 44]. In this paper, we focus on an important aspect of e-commerce search that has attracted very little research attention [47]: query reformulations performed by users. In contrast to the state-of-affairs in e-commerce search, there is a large body of work on analyzing and characterizing users' query reformulations in Web search [2–4, 24, 25, 36].

We present the first — to the best of our knowledge — large-scale and in-depth study of users' query reformulations in e-commerce search.[1] The study is based on the query logs of eBay's search engine — one of the largest platforms for e-commerce search. We analyze numerous aspects of search sessions composed of query reformulations; e.g., the number of reformulations and the distribution of their types, changes of search results pages (SERPs) as a result of the reformulations, clicks and purchases, and more. We contrast several of our findings with those reported in literature on query reformulations in Web search; e.g., the relative effectiveness of reformulations and the changes of SERPs.

In addition to the analysis of query reformulations, we address a novel challenge in e-commerce search: predicting whether a query will be reformulated *before* the retrieved results are presented to the user. High-quality prediction can potentially be of much merit. A case in point, a different ranking function can be applied. Similar rationale was used to motivate work on query performance prediction [9]; that is, predicting search effectiveness in lieu of relevance judgments. However, we are not aware of work on predicting whether a query will be reformulated, except for that of Awadallah et al. [2] on voice search in the Web domain. Given the fundamental differences between voice-based Web search and (text-based) e-commerce search, most of the information sources used for prediction by Awadallah et al. [2] cannot be utilized in our setting. Furthermore, since prediction takes place before the retrieved results are shown to the user, information about user engagement with the SERP cannot be utilized. This further sets apart the reformulation task we pursue here from work on other prediction tasks in e-commerce search, mainly purchase intent prediction [29, 44] and user satisfaction estimation (prediction) [31].

Our reformulation prediction approach utilizes various features, many of which are novel to this study. Some of the features operate prior to retrieval time. Others are based on analysis of the SERP before it is presented to the user. Large-scale extensive empirical evaluation, performed using the query logs of eBay's search engine, demonstrates the merits of our prediction approach. While both pre- and post- retrieval features are effective, the latter are substantially more effective than the former. Furthermore, their integration is of merit and yields high prediction quality. In addition, we show that high-quality prediction can be obtained without utilizing historical information about the user or the query at hand. Yet, using such information helps to further improve prediction quality.

Our focus in the evaluation is on queries which are not reformulations of other queries. The reason is that these queries are the separating points between ad hoc (zero shot) retrieval and interactive retrieval with query sessions. To summarize, our contributions are as follows:

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401065

[1]Singh et al. [47] studied reformulations only for queries for which no results were retrieved. Further comparison with their work is presented below.

- The first large scale and in-depth study of users' query reformulations in e-commerce search.
- The first approach in the e-commerce search realm to predicting whether a query will be reformulated by the user. Most of the features on which the approach relies are novel to this study. Large scale evaluation using eBay's query logs demonstrates the high prediction quality of the approach.
- In-depth analysis of the merits of using and integrating pre- and post- retrieval features, as well as those which utilize past information about the user and the query at hand.

2 RELATED WORK

The lines of work most related to ours are (i) query reformulation and various prediction tasks in e-commerce search; (ii) session search on the Web, as most previous work on analyzing search sessions that consist of query reformulations was in the Web retrieval domain; and, (iii) query performance prediction, because one of the tasks we pursue is predicting whether a query will be reformulated, which could be viewed as a specific query performance prediction task, as we discuss below.

E-commerce search. A few methods for automatic query reformulation in e-commerce search have been proposed [19, 33, 50]. In contrast, we analyze user query reformulations and set as a goal to predict whether they happen.

The vast majority of prediction challenges addressed in work on e-commerce search were focused on how the search session ends, and more specifically, whether a purchase will occur [44, 49, 52]. In contrast, we predict whether a query will be reformulated. Furthermore, features based on user engagement with the retrieved results, which are not available for our prediction task, play an important role in this line of work [44, 49, 52].

The only query-log-based study of user query reformulations in e-commerce search that we are aware of is that of Singh et al. [47]. They studied the way users reformulate (once) queries for which no results are returned. In contrast, we study reformulations for queries regardless of the number of retrieved results, and we analyze more than a single reformulation. Our log-based analysis of reformulations is based on factors (e.g., clicks, purchases and changes of SERPs) different than those analyzed by Singh et al. [47], who focused on user-based segmentation and corresponding features. In addition, Singh et al. [47] did not address the reformulation prediction challenge we pursue here.

There is work on estimating user satisfaction in e-commerce search with respect to the results retrieved for a query [31]. The main features used were based on user engagement with the retrieved results; these are not available for the reformulation prediction task we tackle. Furthermore, a reformulation event is used as a feature in this estimation task [31]; and to this end, estimates of whether the current query is a reformulation of a previous one are used. In contrast, we predict whether a reformulation will take place for the query at hand.

Web session search. There are various methods of determining search session boundaries in Web search; e.g., [1, 3, 6, 28]. We use a simple term-overlap-based approach (cf. [3, 28]) and show that it is highly correlated with both human annotations and a semantics-based approach. There is some work on evaluating user satisfaction with the results retrieved for a given query by estimating whether the known next query posted by the user is a reformulation of the given query [3]. We predict whether a query in e-commerce search will be reformulated.

The basic types of query reformulation we focus on (adding, deleting, or replacing a term) have been long studied for Web session retrieval; specifically, in studies where the goal was to characterize reformulation types/actions [24, 25]. There is also work on categorizing which type of reformulation was applied to a given query assuming that the query was reformulated [24, 36]. The tasks of developing finer-grained taxonomies of reformulation types/actions for e-commerce search and classifying reformulations based on these taxonomies are left for future work.

Predicting session search satisfaction is another task that has attracted some research attention in work on Web search [26]. We do not predict search satisfaction for a session, but rather predict whether a specific query will be reformulated.

As noted in Section 1, there is work on predicting whether a *voice* query will be reformulated in Web search [2]. The vast majority of features are different from those we use and are based on voice analysis. This is the only work we are aware of that tackled the same prediction task we address in this paper.

Query performance prediction. The task of query performance prediction (QPP) is to estimate search effectiveness with no relevance judgments. Pre-retrieval predictors analyze the query and utilize corpus-based term statistics; e.g., [13, 22, 57, 58]. Post-retrieval predictors also analyze the result list of top-retrieved documents; e.g., [9, 38, 41, 45, 51]. Several features that we use here to predict whether a query will be reformulated operate prior to retrieval time; some of them are effective pre-retrieval QPP methods [13, 57]. Other features that we utilize analyze the result list of the most highly-ranked results; some of these are effective post-retrieval predictors [14, 51, 58]. As in work on QPP, we found that pre-retrieval features are less effective than post-retrieval features [9] and that the integration of both is of merit [22, 38, 58].

Interestingly, one of the basic arguments for the motivation to engage in the QPP task was that the search system could change a ranking function if the search was predicted to be ineffective [9]. We used this example above to support the potential merits of effectively predicting whether a query in e-commerce search will be reformulated. Indeed, the reformulation prediction task can be viewed as a special *operational* case of the QPP task, where the reformulation is intended to improve the information need representation or to further exploration [4]. Although there has been an incredible progress in developing highly effective QPP methods, in most cases this progress was not translated to progress in addressing tasks using the predictors; in the vast majority of cases, the predictors were evaluated by the correlation with ground truth performance [38]. In contrast, here we clearly demonstrate the ability to effectively predict whether a query in e-commerce search will be reformulated.

3 DATASET AND DEFINITIONS

3.1 Datasets

For our research, we sampled two datasets from the query log of one of the world's largest e-commerce websites, eBay. The query

log is partitioned into *eBay sessions*, based on the commonly used definition: a sequence of queries by the same user, without an idle time longer than 30 minutes between each pair of consecutive queries in the sequence [3, 4, 28]. We sampled, uniformly at random, along a time period of exactly one week, 400,000 eBay sessions from the eBay US website, desktop devices only (PC, as opposed to all types of mobile devices), without any use of filters, such as item's condition (used vs. new), delivery options (free vs. any), listing type ('buy it now' vs. auction), or category-specific filters. We focused on sessions that included queries whose results were sorted by "best match", which is the default sorting and thus the most popular by a large margin. To avoid handling extreme cases, we disregarded eBay sessions that included more than 20 queries, which accounted for 0.37% of all eBay sessions in our sampled population.

As mentioned, we examined two samples: one based on a week in January 2019 (referred to as the *January* dataset) and the other based on a week in May 2019 (referred to as the *May* dataset). The sampled sessions were performed by 376,477 and 374,654 unique users, in January and May, respectively.

Each query in the datasets included, in addition to the query text itself, a timestamp and the list of retrieved results presented to the user on the SERP. Each returned result is a listed offer, or *listing* in short, by a specific seller. In other words, the same product may appear multiple times on the SERP, with different sellers, prices, delivery options, and so forth. By default, the eBay SERP presents 50 results. Our dataset included, for each result, its rank on the SERP[2] and a unique listing URL. In addition, for each query we had information about its associated clicks and purchases, if any were performed, including their ranks and corresponding listing URLs.

Each listing on eBay is associated with a *leaf category (LC)*, which is the most specific type of node in the eBay's taxonomy. The taxonomy includes tens of thousands of LCs, such as Home Plumbing Pipes, Prepaid Gaming Cards, or Engagement Rings. Each listing is also associated with one out of 43 *meta-categories (MCs)*, such as Home Improvement, Video Games, or Watches. For each result on the SERP, we had information about the LC and MC it belonged to. In addition, we associated all queries in our dataset with an MC, using an internal tool, which considers the distribution of MCs over the retrieved results, as well as past user interaction (clicks and purchases) with similar queries.

3.2 Definitions

We use the term *query* to refer to an instance of a specific query that was submitted by a given user in a given timestamp. According to this definition, different queries can have the exact same textual expression. For example, the most popular query text in our datasets is "iphone x", represented by over 3,000 different queries. Table 1 summarizes the definitions of different types of sessions and queries we examine throughout the rest of this paper. Several of the definitions are based on a *reformulation* relation between two queries. We next describe how this relation is derived.

3.3 Inferring Reformulations

In Table 1, we defined an eBay session. Such a session can potentially contain one or more *reformulation sessions*. We use the term *reformulation session* to refer to a sequence of queries which, broadly speaking, reflect the same intent, its narrowing/focusing (e.g., moving from "sedan cars" to "chevrolet sedan"), or broadening/generalizing (e.g., moving from "chevrolet sedan" to "sedan cars"). Accordingly, a reformulation session ends when the user stopped querying or her intent has completely changed.

Thus, the definition of a reformulation session relies on a basic notion of reformulation relation between two consecutive queries. There has been quite a lot of work in Web search on inferring whether a query is a reformulation of a previous query; i.e., marking boundaries of reformulation sessions — e.g., [1–3, 6, 28]. Since this task is not our main focus here, but is still highly important for the analysis we present and for the prediction challenge we address, we opted for a robust reformulation definition that satisfies a few desiderata: (i) highly correlated with human annotations, (ii) highly correlated with an alternative approach of defining a reformulation, (iii) fast to compute, (iv) complies with (some) practice in work on Web session search.

Inspired by some work on Web query session analysis [2, 3, 28], we have used the similarity between two consecutive queries as an indicator for a reformulation relation. To measure the similarity, we use the Jaccard similarity coefficient over their lower-cased tokens (cf., [2, 3, 28]), after spelling correction was automatically performed by the search engine. We examined a variety of thresholds for the Jaccard similarity and found, as detailed below, that a 0 threshold results in high adherance to the four desiderata described above. This means that we deem two queries (or more precisely, their

Table 1: Definitions of session types and query types.

Name	Definition
eBay session	A sequence of queries by the same user, without an idle time longer than 30 minutes between each pair of consecutive queries in the sequence. We consider the longest possible eBay sessions according to this definition, i.e., an eBay session cannot be a sub-sequence of another longer eBay session.
Reformulation session	A sub-sequence of an eBay session that contains at least two queries, with each query reformulating the preceding query in the sequence, when exists. We consider the longest possible reformulation sessions according to this definition, i.e., a reformulation session cannot be a sub-sequence of another longer reformulation session.
Non-reformulation session	An eBay session that does not contain any reformulation sessions.
Singleton session	An eBay session that contains a single query.
Reformulation query	A query that is part of a reformulation session. For each pair of consecutive queries in a reformulation session, we refer to the first as the *reformulated* query and the second as the *reformulating* query.
Fresh query	A query that is not a reformulating query in any reformulation session. Notice that fresh queries include the first queries in all reformulation sessions and all queries that do not belong to a reformulation session.
Final query	A query that is not a reformulated query in a reformulation session. Notice that final queries include the last queries in all reformulation sessions and all queries that do not belong to a reformulation session.
Non-reformulation query	A query that is part of a non-reformulation session. Notice that all non-reformulation queries are both fresh and final, but not all queries that are both fresh and final are necessarily non-reformulation queries.
First query	A query that opens an eBay session.
Last query	A query that ends an eBay session.
Singleton query	A query that comprises a singleton eBay session, i.e., a query that is both first and last.
Reformulation-first query	A query that opens a reformulation session; i.e., a query that is both a reformulation query and a fresh query.
Reformulation-last query	A query that ends a reformulation session; i.e., a query that is both a reformulation query and a final query.

[2]The top result is at rank 1.

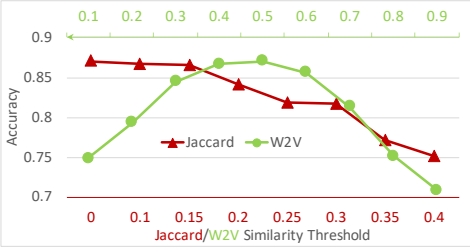

Figure 1: Accuracy w.r.t to human annotations of using Jaccard and Word2Vec with different thresholds to determine boundaries of reformulation sessions.

textual expressions) as similar if they share a token; if the queries are not identical, we assume a reformulation relation holds.

We asked four in-house annotators to tag the reformulation boundaries (i.e., which queries reformulated the previous ones and which did not) for 807 queries in 210 eBay sessions. These were sampled uniformly at random from our May dataset. The queries were presented in a sequence, eBay session by eBay session. For each query, starting from the second in each eBay session, annotators were asked to decide whether the query reformulates the previous one in the eBay session. The annotators' guidelines explained what a reformulation means (a change in the query for the same and/or more general/specific user intent); multiple examples of reformulations and lack thereof were provided to the annotators. The Fleiss Kappa [16] among the four annotators was 0.805.

We then used the human annotations to evaluate two inter-query similarity measures: lexical similarity using the Jaccard coefficient and semantic similarity using Word2Vec [34]. The latter was trained over 358M e-commerce queries sampled using the same criteria as our datasets, with embedding size of 300 and window size of 5. The query similarity was then measured using cosine similarity between the centroid of the word vectors. Figure 1 shows the accuracy, w.r.t. the manual annotations, of using the two similarity measures for determining whether a reformulation has taken place (i.e., marking reformulation session boundaries) .

We see in Figure 1 that Jaccard similarity peaks when the threshold is 0 and declines as the threshold grows, requiring higher portion of the queries' words in the intersection between the two queries. The Word2Vec similarity peaks at a threshold of 0.5. The peaks are of the same accuracy value (0.870); the Pearson correlation with the manual annotation was 0.738. In addition, the two measures — Jaccard with a 0 threshold and Word2Vec with a 0.5 threshold — are highly correlated with each other: the agreement rate on session boundaries is 93.8% and the Pearson correlation is 0.874.

Given that the Jaccard-based inter-query similarity measure with a 0 threshold as a means to inferring reformulations is (i) highly correlated with human annotations; (ii) highly correlated with the optimal (with respect to agreement with human annotations) Word2Vec-based semantic measure; (iii) extremely fast to compute, and hence allows online session boundary markup at high scale; and (iv) corresponds to some session definitions in work on Web session search, we have used it to infer reformulations, and accordingly, determine reformulation session boundaries.

To conclude this section, Table 2 shows an eBay session from our May dataset, with 13 queries in total, which demonstrate many of the query types defined in Table 1. The eBay session includes three different reformulation sessions of lengths 3, 2, and 5, respectively.

Table 2: Example eBay session from our May dataset, with different types of queries and their number within the reformulation session ('#', where '0' stands for a query that does not belong to any reformulation session).

#	Query	MC	Query Types
1	maggi4e barnes blouse 4x	Clothing	First, Fresh, Reformulation, Ref-first
2	maggie barnes blouse 4x	Clothing	Reformulation
3	maggie barnes blouse 4x pink	Clothing	Reformulation, Ref-last, Final
0	nancy drew flashlight series	Books	Fresh, Final
0	antique fishing floats	Antiques	Fresh, Final
1	fuzzy barbie doll boots	Dolls	Fresh, Reformulation, Ref-first
2	fuzzy barbie doll shoes	Dolls	Final, Reformulation, Ref-last
1	apple watch series 4	Watches	Fresh, Reformulation, Ref-first
2	apple watch series 4 44mm	Watches	Reformulation
3	apple watch band series 4 44mm	Watches	Reformulation
4	apple watch band series 4 44mm genuine	Watches	Reformulation
5	44mm milanese loop apple watch band	Watches	Final, Reformulation, Ref-last
0	base ball matt bases	Sports	Last, Fresh, Final

Table 3: Prevalence of query types in our datasets. Recall that the types are not disjoint.

	January 2019		May 2019	
	Total number	% of all queries	Total number	% of all queries
Reformulation	529,486	56.94%	525,706	56.88%
Fresh	584,245	62.83%	581,550	62.92%
First	400,000	43.01%	400,000	43.28%
Reformulation-first	184,177	19.81%	183,364	19.84%
Singleton	203,306	21.88%	203,710	22.04%
All	929,949		924,317	

4 REFORMULATION CHARACTERISTICS

4.1 Basic Analysis

The eBay sessions we analyzed included a total of 929,949 queries in the January dataset and 924,317 in May. Table 3 shows the portion of different query types, as defined in Table 1, out of all queries, in both our January and May datasets. It can be seen that queries that belong to a reformulation session (reformulation queries) account for nearly 57% of all queries in both datasets (and 72.5% of all non-singleton queries), reinforcing the motivation to further understand and predict query reformation in e-commerce search. The portions are very similar between the two datasets.

The average number of queries per eBay session in the May dataset was **2.31** (stdev: 2.21, median: 1, 75th percentile: 3, 90th percentile: 5), while the average number of queries per reformulation session was **2.87** (stdev: 1.53, median: 2, 75th percentile: 3, 90th percentile: 5). Note that an eBay session can be composed of a single query and hence, eBay sessions can be, on average, shorter than reformulation sessions, which are always part of eBay sessions. The average length of a query (in words[3]) was **3.32** (stdev: 1.85, median: 3, 75th percentile: 4, 90th percentile: 5).

4.2 Reformulation Types

In accordance with previous work on query reformulation in Web search [24, 25], we refer to three types of reformulations: (1) *Add*:

[3]Throughout this work, we use white-space tokenization for queries, after standard normalization, such as lower casing and redundant white-space removal [4].

Table 4: Distribution of reformulating queries by type.

January 2019			May 2019		
Add	Remove	Replace	Add	Remove	Replace
34.65%	17.88%	47.47%	34.66%	17.86%	47.48%

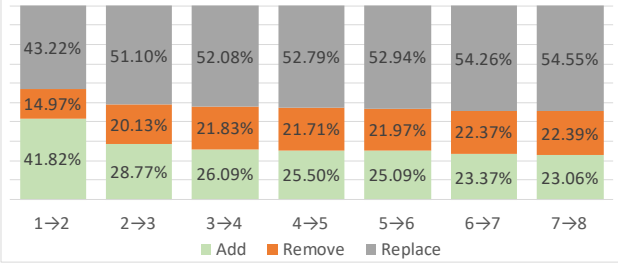

Figure 2: Reformulation type distribution by position within the reformulation session: reformulated query → reformulating query.

addition of one or more words to the reformulated query; (2) *Remove*: removal of one word or more from the reformulated query; (3) *Replace*: replacement of one word or more in the reformulated query. Previous research of query reformulation on Web search associated 'Add' reformulation with specialization, i.e., refinement to a narrower information need, and 'Remove' reformulation with generalization to a broader information need [25, 27]. Table 4 shows the distribution of types for each of the two datasets. It can be seen that both distributions are very similar. 'Replace' is the most common type, followed by 'Add' and finally 'Remove'. The same order of type frequency has been previously reported for Web search [25].

For the remainder of this work, we present the results for the May dataset only. We conducted our analysis across both datasets, but as the results were very similar, we focus on the latter for clarity of the presentation and space considerations.

Figure 2 shows the distribution of reformulation types according to their position within the reformulation session (the first reformulation in the session is marked '1→2' and so forth). It can be seen that for the first reformulation, the portion of 'Add' is especially high, at over 40%, while sharply decreasing for further reformulations along the reformulation session. The portion of 'Remove' reformulations increases along the session and so does the portion of 'Replace'. The portion of 'Add' reformulations, however, remains above the portion of 'Remove', even for higher positions. We therefore expect the query length to increase along the reformulation session, with a sharper increase in its beginning. Figure 3 confirms and quantifies this premise.

4.3 Changes of SERPs

We now focus on the SERP changes for reformulation queries along the reformulation session. To this end, we define the *overlap@k* between two queries as the size of the intersection of their top k results, normalized by k. The upper section of Table 5 presents the *item overlap@k* for all pairs of reformulated and reformulating queries in reformulation sessions and, for reference, for all fresh queries with the preceding query submitted by the same user. Recall that fresh queries include all queries that do not reformulate their

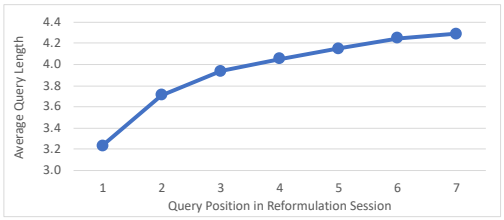

Figure 3: Query length distribution by position within the reformulation session.

Table 5: Overlap@k for $k \in \{10, 50\}$ across reformulation queries and fresh queries considering retrieved results (items), leaf categories, and meta-categories. Mean and median overlap, as well as percentage of reformulated and reformulating query pairs with 100% overlap are presented.

	Reformulation Queries			Fresh Queries		
	Mean	Median	% Full Overlap	Mean	Median	% Full Overlap
Item overlap@10	0.118	0	0.75%	0.012	0	0.07%
Item overlap@50	0.122	0.02	0.12%	0.001	0	0.02%
LC overlap@10	0.507	0.50	19.34%	0.197	0	4.22%
LC overlap@50	0.429	0.38	4.51%	0.156	0	1.01%
MC overlap@10	0.733	1	40.18%	0.480	0.40	15.44%
MC overlap@50	0.611	0.78	12.46%	0.384	0.22	4.92%

preceding query, as described in Table 1. We measure the overlap at k=10 and k=50, to reflect the top results and the default number of results displayed on the SERP, respectively.

Table 5 shows that while the item overlap is noticeably higher within reformulation sessions than for fresh queries, it is still rather low: around 12% overlap, on average, for both the top 10 and top 50 results, with over 50% of the query pairs having no overlap between the top 10 results, and very small portion with full overlap (10 of 10 or 50 of 50, respectively). This indicates that reformulating queries lead to a substantial change in the SERP, often replacing the entire result list, or almost all of it. Past work required an overlap@10 of at least one result for two queries to belong to the same "topical" search session on Web search [4]. As we see, in e-commerce search such overlap does not exist even when the topic/category remains the same. An explanation to this difference from Web search can be that e-commerce search is performed over a more structured corpus, while query terms often refer to an attribute value, such as brand or color [19, 43]. A reformulation often changes the set of item properties over a huge inventory, in the case of eBay [20, 53], leading to a radical change of the SERP.

As we witnesses a low item overlap between SERPs, we also set out to explore the category overlap. To this end, we represent each listing result on the SERP via its LC and MC, to compute the *LC overlap@k* and *MC overlap@k*, similarly to item overlap@k, as described above. The lower sections of Table 5 present the category overlap results. It can be seen that the category overlap within a reformulation session is rather high, both with respect to the item overlap and to the category overlap of fresh queries with their preceding ones. We conjecture that while the reformulation substantially changes, refines, or broadens the attributes of the searched listings, thereby leading to a considerable change in the results, the scope of intent remains the same, reflected through a much milder change in the retrieved listings' categories.

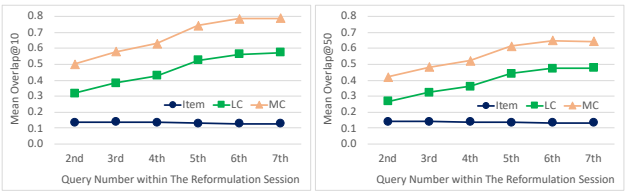

Figure 4: Average Item, LC, and MC overlap@10 (left) and overlap@50 (right) between two consecutive queries in a reformulation session by position of the reformulating query.

Figure 4 presents the item, LC, and MC overlap along reformulation sessions. While both LC and MC overlap increase substantially as the query position is higher, the item overlap remains low along the session (~13% for both k=10 and k=50). This indicates that as reformulation sessions become longer, the category distribution on the SERP becomes even more steady, but the SERPs still include very different listings with each reformulation.

4.4 Clicks and Purchases

One of the two most fundamental user interaction signals in e-commerce search are clicks on retrieved results and purchases of listings following these clicks. They have also been found to be the most robust training signals for learning-to-rank in e-commerce search [42]. For our analysis of user interaction signals, we consider click-through and purchase-through rates, as well as the ranks of the clicked and purchased results. Specifically, we examine the following measures:

- *Click-through rate (CTR)*: the portion of queries for which at least one click was performed on one of the retrieved results.
- *Purchase-through rate (PTR)*: the portion of queries for which at least one of the retrieved results was purchased through a click on the retrieved results page.
- *Median click rank*: the median rank of a click across all clicked results.
- *Median first click rank*: the median rank of the chronologically-first click performed for each query (only considering queries for which at least one click was performed).
- *Median purchase rank*: the median rank of a purchase across all purchased results.

Table 6 presents the results of these measures for different types of queries, as defined in Table 1. Inspecting the upper section of the table, it can be seen that reformulation queries, in general, have somewhat lower rates of clicks and purchases compared to the rest of the queries: lower CTR and PTR and lower ranks of the clicks and purchases. It could be that as users reformulate, clicks and purchases spread across a higher number of queries and are thus lower when inspected per query. Indeed, for non-reformulation queries and singleton queries, both reflecting query types with a different intent per single query, the rates are higher. These results are generally aligned with previously reported results for Web search, where reformulation query pairs (reformulated and reformulating) had a CTR lower by 21% than the average CTR [3].

The lower sections of Table 6 indicate that for fresh, first, and reformulation-first queries, the CTR and PTR are lower than for final, last, and reformulation-last queries, respectively; e.g., the PTR almost triples for reformulation-last w.r.t reformulation-first.

Table 6: Clicks and purchases for different types of queries. The rates and ranks over all non-singleton queries (bottom row) serve as the reference point for all measures.[4]

Query type	Click through rate	Purchase through rate	Median click rank	Median first click rank	Median purchase rank
Singleton	$1.14c$	$1.66p$	r	s	t-1
Reformulation	$0.98c$	$0.92p$	$r+1$	$s+1$	$t+1$
Non-reformulation	$1.03c$	$1.28p$	$r-2$	$s-1$	$t-2$
Fresh	$0.93c$	$0.9p$	$r-1$	s	$t-1$
Final	$1.14c$	$1.38p$	$r-1$	s	$t-1$
First	$0.9c$	$0.87p$	r	s	t
Last	$1.23c$	$1.56p$	r	s	t
Reformulation-First	$0.81c$	$0.56p$	$r+1$	$s+1$	$t+1$
Reformulation-Last	$1.24c$	$1.57p$	r	$s+1$	t
All non-singleton	c	p	r	s	t

To further explore the latter finding, Table 7 shows the CTR and PTR for reformulation and eBay sessions by session length, from 2 to 7 queries. Across all session lengths, and both types of sessions, it can be seen that the CTR soars by 30% to 40% for the last query in the session, while remaining rather stable along the previous queries in the session, starting from the first one. For PTR, these trends are even sharper, and especially for reformulation sessions. These results reflect the two principal reasons previously found for users to end a session in the Web search literature [4, 36]: (1) the user is satisfied: in the case of e-commerce search, this can be reflected in a purchase; (2) the user is not satisfied and gives up on finding (what to purchase), yet clicked on the way to this decision. In addition, our findings are aligned with past research on reformulation of voice queries in Web search, which found that reformulating queries have higher CTR than the rest of the queries [2].

5 REFORMULATION PREDICTION

Our prediction task is to decide whether a query will be reformulated or not. We focus our task on first queries, i.e., queries that open a new eBay session (see Table 1). High prediction quality for this task could allow an e-commerce search engine to improve its serving: a query that will not be reformulated is likely best handled using a zero-shot retrieval (for a target-finding [49] or target purchasing [48] intent on e-commerce search), while a query likely to be reformulated, implies either a struggle or deep exploration on the part of the user (decision making or shopping intent [48, 49]). We additionally examine a broader task, which is focused on fresh queries (Table 1). This set represents all queries that are not within a reformulation session. In other words, in addition to first queries, all queries inside an eBay session that have not reformulated their predecessor. While we experimented with both the January and May datasets, we report only the results for the latter, as in the previous section, since the results for both were very similar. As shown in Table 3, first and fresh queries account for 43.3% and 62.9% of all queries in the May dataset, respectively.

5.1 Training and Evaluation

We used 5-fold cross-validation to train, tune the hyper-parameters, and evaluate the classifiers. In each iteration, 3 folds were used for

[4] Actual values are not disclosed due to business sensitivity.

Table 7: Clicks and purchases for different query positions by session length, for reformulation sessions ('R') and eBay sessions ('S'). Rates over non-singleton queries serve as the reference point for all measures (see bottom row of Table 6).

Session Length	Query Number	Click-through Rate		Purchase Rate	
		R	S	R	S
2	1st	0.81c	0.89c	0.56p	0.88p
	2nd	1.22c	1.23c	1.60p	1.67p
3	1st	0.81c	0.89c	0.56p	0.88p
	2nd	0.88c	0.94c	0.58p	0.89p
	3rd	1.25c	1.23c	1.52p	1.57p
4	1st	0.81c	0.90c	0.55p	0.89p
	2nd	0.86c	0.92c	0.62p	0.88p
	3rd	0.90c	0.93c	0.60p	0.74p
	4th	1.27c	1.24c	1.59p	1.50p
5	1st	0.83c	0.93c	0.54p	0.83p
	2nd	0.87c	0.91c	0.52p	0.83p
	3rd	0.86c	0.92c	0.55p	0.78p
	4th	0.89c	0.95c	0.53p	0.71p
	5th	1.26c	1.22c	1.67p	1.48p
6	1st	0.85c	0.91c	0.53p	0.89p
	2nd	0.90c	0.91c	0.64p	0.78p
	3th	0.89c	0.93c	0.52p	0.71p
	4th	0.86c	0.90c	0.45p	0.79p
	5th	0.91c	0.95c	0.52p	0.79p
	6th	1.30c	1.25c	1.43p	1.44p
7	1st	0.89c	0.92c	0.54p	0.90p
	2nd	0.87c	0.94c	0.60p	0.88p
	3rd	0.91c	0.91c	0.60p	0.72p
	4th	0.89c	0.91c	0.76p	0.88p
	5th	0.83c	0.90c	0.57p	0.58p
	6th	0.96c	0.94c	0.67p	0.65p
	7th	1.26c	1.23c	1.49p	1.31p

training, one fold for validation, and one for test. We report the average results over the 5 test folds. As evaluation metrics, we used accuracy, area under the ROC curve (AUC), and F1 for the positive (reformulation) class. Statistically significant differences of accuracy, AUC and F1 were determined using the two-tailed approximate randomization test [35] test at a 95% confidence level, with a random sample of 10,000 permutations, following the recommendations in [15]. We applied Bonferroni correction for multiple comparisons. Hyper-parameter tuning was performed over the validation set, optimizing for the accuracy metric.

We experimented with four different classifiers: Random Forest [7], Gradient Boosting Decision Tree (GBDT) [17] using the implementation of XGBoost [11], Support Vector Machines (SVM) [12] with both linear and RBF kernels using LIBSVM [10], and a two-layer fully-connected Neural Network [5]. For Random Forest, we tuned the number of trees, max depth of trees, minimum sample split, and minimum sample leaf. The values were selected from {50, 100, ..., 800}, {5, 10,..., 200}, {2, ..., 10}, and {1, 2, 4}, respectively. For GBDT, we tuned the number of trees, max depth of trees, subsampling of columns in a tree, sub-sample ratio of the training instances, and minimum child weight. The examined values spanned {100, ..., 1000}, {10, ..., 500}, {0.5, 0.6,..., 1}, {0.8, 0.9, 1}, and {2, 3,..., 7}, respectively. For SVM, we tuned the penalty parameter C across the range of {0.01, 0.1, 1, 10, 100}, and for the RBF kernel also the similarity parameter Γ across {0.001, 0.01, 0.1, 1, 10, 100}. For the Neural Network, we tuned the learning rate, number of epochs, hidden layer size, dropout rate, and batch size. Values were selected from {0.01, 0.001, 0.0001}, {10, 15, ...,30}, {300, ..., 600}, {0, 0.1, ..., 0.4}, and {16, 32, 64, 128}, respectively.

5.2 Features

Our feature set includes pre-retrieval and post-retrieval features. The former can be produced before the e-commerce search engine retrieves the results for the query. These features typically make use of the query text itself, the corpus as a whole, and historical data from the query log. Post-retrieval features make use of the results retrieved by the search engine for the query in question, typically the top ones. They can induce information from the characteristics of retrieved results (e-commerce listings, in our case) and their relationship (e.g., similarity) with the query in question [9]. From a practical perspective, using only pre-retrieval features allows to predict if a reformulation will occur immediately after the query is submitted by the user. On the other hand, using both pre-retrieval and post-retrieval features requires waiting until the results are retrieved by the search engine, before applying the predictive model. In both cases, however, as mentioned in Section 1, the prediction can still take place before the user is presented with the results. The difference is only with regards to the response time, which includes the retrieval time in case of using post-retrieval features.

As opposed to past work, which focused on estimating a reformulation in an archived query log, against a manually-annotated ground truth [3], our prediction task is aimed for real-time. We therefore do not consider user interaction features, such as clicks and purchases, for the query in question, as our goal is to take advantage of the prediction before the user interacts with the SERP.

Table 8 presents a detailed description of all the features. They span six different "*families*": the first two relate to e-commerce qualities: Category features relate to the taxonomy behind the inventory and involve both LCs and MCs. Features from the Attributes family relate to the structured nature of e-commerce listings and their name-value pairs, which are either explicitly attached to the listing or extracted from the query using named entity recognition [55].

The next two families involve query performance predictors (QPPs) [9], which, as already mentioned, closely relate to the reformulation prediction task. The first family includes primary QPPs previously defined in the literature. For pre-retrieval predictors, we select those shown to be effective for document search in large-scale studies [22, 46]; namely, predictors based on IDF values of query terms [13] and the variance across documents in the corpus of TF.IDF values of the query terms [57]. For post-retrieval predictors, we use standard deviation [14], WIG [58], and SMV [51], which could be viewed as integrating the first two; the three were shown to be highly effective for document retrieval in various studies [9, 41, 46]. These predictors are unsupervised, fast to compute in our setting — based on statistics of surface-level similarity scores between the query text and listings' titles; furthermore, similar predictors were found effective for a number of retrieval tasks [30, 32, 39]. The second family includes additional QPPs we defined, to capture further similarity aspects between the query and the retrieved results.

Finally, the last two feature families take advantage of having a query log at hand. They relate to the user's historical activity and the query text's past occurrences in the log, respectively, both over a period of the preceding 5 months. For users, we had historical data for 93.5% of the queries in our dataset. We completed the missing feature values for users with no history with their average values across the training set. Historical data was available for 63.1% of the

Table 8: Pre- and Post-retrieval features used for query reformulation prediction.

Category	Pre	– MC of the query as determined by the predictive tool described in Section 3.1.
	Post	– Most common MC across all top k retrieved results for $k \in \{10, 50\}$. – Number of different MCs within the top k retrieved results for $k \in \{10, 50\}$. – Variance and entropy of the MC and LC distributions across the top k retrieved results for $k \in \{10, 50\}$.
Attributes	Pre	– Binary predicates and counts for each of the attributes extracted from the query using NER [55]. We included features for the most common attributes in our dataset: type, brand, product identifier, unit of measure, material, color, demographics, location, and size.
	Post	– Number of times that query terms match an attribute value of a listing on the SERP (each term can match more than one attribute, e.g., 'crystal' is both a brand and material). – Number of times that query terms match MC attribute values, per each of the 43 MCs (for each MC, we created a list of values across its most popular attributes, based on all listings that belong to this MC in our dataset). – Percentage of query terms that match at least x attribute values, for $x \in [1, 5]$.
Previously Reported QPPs	Pre	– Query length (in words) [9]. – Minimum, maximum, and sum of the IDF values of the query terms [37]. – Minimum, maximum, and sum of the variance of TF.IDF values of the query terms across documents in the corpus [57].
	Post	– Total number of retrieved results [9]. – The thresholded standard deviation of retrieval scores prediction value [14], with a 50% threshold, computed for the titles of listings in the top k retrieved results for $k \in \{10, 50\}$ with respect to the query text. We use both Okapi-BM25 [40] and standard TF.IDF retrieval scores using cosine similarity as the similarity measure. – The weighted information gain (WIG) prediction value [58] without corpus-based normalization [46] of the Okapi-BM25 scores computed for the titles of listings in the top k retrieved results for $k \in \{10, 50\}$ with respect to the query text. – The score magnitude and variance (SMV) prediction value [51] computed for the Okapi-BM25 scores of titles of listings in the top k retrieved results for $k \in \{10, 50\}$ with respect to the query text [51]. Inspired by [58], we use the squared root of the query length, rather than the average retrieval score in the corpus, as a normalizer.
Extended QPPs	Post	– Average, standard deviation, median, minimum, and maximum of Jaccard coefficient between the query terms and title terms of top k retrieved results for $k \in \{10, 50\}$. – Average, standard deviation, median, minimum, and maximum of the portion of query terms appearing in at least one of the top k retrieved result titles for $k \in \{10, 50\}$. – Average, standard deviation, median, minimum, and maximum of the portion of title terms appearing in the query out of all title terms across the top k retrieved results for $k \in \{10, 50\}$. – The portion of top k retrieved results that have at least one query term in their title, for $k \in \{10, 50\}$. – The portion of top k retrieved results that have all the query terms in their title, for $k \in \{10, 50\}$.
User History	Pre	– The total number of queries submitted and the total number of eBay sessions performed by the user. – The average, standard deviation, and median eBay session length (number of queries). – The portion of reformulation queries out of all of the user's submitted queries. – The portion of add, remove, and replace reformulating queries out of all of the user's reformulating queries. – User's CTR: the portion of SERPs for which the user performed at least one click over one of the retrieved results. – User's PTR: the portion of SERPs for which the user performed at least one purchase of one of the retrieved results.
Query History	Pre	– The total number of times the query was submitted. – The average, standard deviation, and median eBay session length across eBay sessions containing the query. – The portion of times the query was reformulated, i.e., the portion of occurrences as a reformulated query. – The portion of add, remove, and replace out of all the query's reformulations. – Query's CTR: the portion of SERPs for which at least one click was performed over one of the retrieved results. – Query's PTR: the portion of SERPs for which at least one purchase was performed of one of the retrieved results.

queries in our dataset (considering exact match, after normalization as described in Section 4.1). Since the portion of queries without any history is considerably large, we split the dataset and trained two separate models: one for queries with history and the other for queries without any history. The presented results where Query History features are involved are based on running the adequate model of the two for each query in the test set.

6 PREDICTION RESULTS

In this section, we present the results for the main tasks of predicting reformulation for first and fresh queries. We then delve deeper into feature importance by performing different types of ablation tests.

Table 9 shows the performance results of the different classifiers for the reformulation prediction task. We use the majority class (no reformulation) as the naïve baseline. For first queries, the GBDT classifier achieves the best results, reaching an accuracy of 84.48%, an improvement of +24.16% over the majority class, and an AUC of 0.85. For fresh queries, the results are generally lower, with GBDT again achieving the highest performance at 81.57% accuracy (+18.14%) and 0.79 AUC. This implies that the reformulation prediction becomes harder for queries that do not open an eBay session. Our initial experiments with other types of queries beyond first and fresh, i.e. those that include queries that already reformulate their predecessor, indicated that the prediction quality using the proposed features is even lower (an overall accuracy of 74.05% across all queries, a +17.54% gain over the majority class). We leave further investigation of this extended prediction task, including the use of additional session features, to future work. The remainder

Table 9: Performance results of different classifiers for the main task: predicting whether a query will be reformulated. 'm' and 'g' mark statistically significant differences with Majority Class and GBDT, respectively. The best result in a column is boldfaced.

Classifier	First Queries			Fresh Queries		
	Accuracy	AUC	F1	Accuracy	AUC	F1
Majority Class	67.99^{g}	0.50^{g}	–	68.52^{g}	0.50^{g}	–
SVM (RBF Kernel)	82.82^{g}_{m}	0.82^{g}_{m}	74.51^{g}	77.70^{g}_{m}	0.71^{g}_{m}	60.47^{g}
SVM (Linear Kernel)	83.14^{g}_{m}	0.82^{g}_{m}	75.27^{g}	78.36^{g}_{m}	0.74^{g}_{m}	63.79^{g}
Two-Layer FC NN	83.32^{g}_{m}	0.83^{g}_{m}	76.15^{g}	79.79^{g}_{m}	0.77^{g}_{m}	68.90^{g}
Random Forest	83.85^{g}_{m}	0.83^{g}_{m}	75.91^{g}	80.46^{g}_{m}	0.75^{g}_{m}	66.52^{g}
GBDT	$\mathbf{84.48}_{m}$	$\mathbf{0.85}_{m}$	**77.76**	$\mathbf{81.57}_{m}$	$\mathbf{0.79}_{m}$	**71.00**

of the results in this section are reported for the GBDT classifier, as it yielded the best results for both of our main tasks.

Table 10 presents the results when using only pre-retrieval features and only post-retrieval features, compared to using both. For both first and fresh queries, both pre-retrieval and post-retrieval features yield a statistically significant gain over the majority class. In addition, post-retrieval features yield a statistically significantly higher performance than pre-retrieval features. In the case of first queries, the gap between the performance using these two types of features is especially large. The combination of using both pre-retrieval and post-retrieval features yields statistically significantly higher performance than using only pre-retrieval or post-retrieval features for both first and fresh queries. These findings indicate,

Table 10: Performance of the GBDT classifier when using only pre-retrieval features, only post-retrieval features, and both. 'm', 'r', and 'o' mark statistically significant differences with Majority Class, Pre-retrieval, and Post-retrieval, respectively. The best result in a column is boldfaced.

Feature Group	First			Fresh		
	Accuracy	AUC	F1	Accuracy	AUC	F1
Majority Class	67.99	0.50	–	68.52	0.50	–
Pre-retrieval	68.65_m	0.53_m	16.29	70.93_m	0.57_m	31.29
Post-retrieval	83.64^r_m	0.84^r_m	77.34^r	73.89^r_m	0.71^r_m	60.22^r
All Features	$\mathbf{84.48}^{ro}_m$	$\mathbf{0.85}^{ro}_m$	$\mathbf{77.76}^{ro}$	$\mathbf{81.57}^{ro}_m$	$\mathbf{0.79}^{ro}_m$	$\mathbf{71.00}^{ro}$

Table 11: Family-level ablation tests: performance of the GBDT classifier when excluding feature families. The top row presents the majority class baseline and the bottom row presents the results when using all features. 'm' and 'g' mark statistically significant differences with Majority Class and GBDT with all features, respectively. The best result in a column is boldfaced.

Excluded Family	First			Fresh		
	Accuracy	AUC	F1	Accuracy	AUC	F1
Majority Class	67.99^g	0.50^g	–	68.52^g	0.50^g	–
Category	84.42^g_m	0.84^g_m	77.58^g	81.55_m	0.78^g_m	70.82^g
Attributes	84.43^g_m	0.84^g_m	77.53^g	81.52^g_m	0.78^g_m	70.71^g
Previously-Proposed QPPs	68.88^g_m	0.55^g_m	24.09^g	70.94^g_m	0.58^g_m	32.85^g
Extended QPPs	82.39^g_m	0.81^g_m	73.66^g	79.93^g_m	0.76^g_m	67.70^g
User History	84.12^g_m	0.84^g_m	77.43^g	81.15^g_m	0.7^g_m	70.54^g
Query History	84.01^g_m	0.84^g_m	77.12^g	81.36^g_m	0.78^g_m	70.72^g
All features	$\mathbf{84.48}_m$	$\mathbf{0.85}_m$	**77.76**	$\mathbf{81.57}_m$	$\mathbf{0.79}_m$	**71.00**

overall, that post-retrieval features are highly effective for achieving good predictive performance for our task and are worth the actual retrieval time required to calculate them. As previously noted, the results presented to the user can still be adapted based on the prediction, even when using post-retrieval features.

To further understand the contribution of each of the feature families presented in Table 8, we performed ablation tests: for each family, we trained and tuned a model based on all features, excluding those that belong to that family. Table 11 presents the performance results for both first and fresh queries. QPP features clearly contribute the most to the overall performance, as their exclusion leads to the largest decrease in all metrics. Previously-Proposed QPPs are playing an especially critical role and their removal yields the largest performance drop, by a large margin. Extended QPPs are the second most important. User History and Query History features make a much more modest contribution to performance, but their removal still leads to a statistically significant performance drop. These features require access to a historical query log (5 months in our case). Category and Attributes features make the smallest contribution, yet still significant in most cases. The combination of all six feature families yields the highest performance for both prediction tasks: first queries and fresh queries. Moreover, the removal of each of the feature families leads to a statistically significant decrease in all performance metrics, aside from accuracy for Category features in the case of fresh queries.

We also performed ablation tests at the single-feature level. Table 12 lists the features that led to the highest accuracy decrease when removed from the set of all features for first queries. The list is topped by the portion of reformulation queries out of all of the user's past queries – a feature that belongs to the User History family. Altogether, the features in the list represent most families, during pre- and post-retrieval, and indicate, again, that the mix of features and families is productive.

Table 12: Feature-level ablation tests: performance of the GBDT classifier when excluding single features, for the first queries prediction task. The top eight features w.r.t accuracy drop are presented. The top row presents the majority class baseline and the bottom row presents the results when using all features. 'm' and 'g' mark statistically significant differences with Majority Class and GBDT with all features, respectively. The best result in a column is boldfaced.

Excluded Feature	Family		Accuracy	AUC	F1
Majority Class			67.99^g	0.50^g	–
Portion of reformulation queries out of all user's submitted queries	User History	Pre	83.67^g_m	0.84^g_m	77.07^g
Avg % of query terms appearing in at least one of the top 50 result titles	Extended QPPs	Post	83.80^g_m	0.84^g_m	76.60^g
Binary NER Brand	Attributes	Pre	83.82^g_m	0.84^g_m	76.80^g
Binary NER product identifier	Attributes	Pre	83.84^g_m	0.84^g_m	76.74^g
Minimum IDF of query terms	Previously-Proposed QPPs	Pre	83.90^g_m	0.84^g_m	77.00^g
Okapi-BM25 WIG for top 10 results	Previously-Proposed QPPs	Post	83.91^g_m	0.84^g_m	76.89^g
LC entropy for top 10 results	Category	Post	83.93^g_m	0.84^g_m	76.91^g
All features			$\mathbf{84.48}_m$	$\mathbf{0.85}_m$	**77.76**

7 CONCLUSION AND FUTURE WORK

We presented a first comprehensive study of query reformulation in e-commerce search. Our log analysis shows that well over 50% of the queries take part in a reformulation session. Reformulations gradually increase the query length and lead to a rise in both click and purchase rates for the last query, for all reformulation session lengths. Each reformulation leads to a thorough change in the results presented on the SERP, in many cases without any overlap with the SERP of the previous query. Overall, our analysis demonstrates the important and unique role reformulation queries play in e-commerce search.

We presented, for the first time in e-commerce search, the task of predicting reformulation for queries that are not already part of a reformulation session, before the results are presented to the user. This kind of prediction allows to adapt the retrieved results presented to the user, and account for the anticipated intent [48, 49]. Using a basic set of features, our prediction reaches high performance, significantly over the majority baseline. Post-retrieval features and query performance predictors, especially a few of the fundamental ones, contribute the most to the prediction performance. Yet, mixing them with user and query history, as well as category and attribute features of both the query and the retrieved results, yields the most effective prediction.

Our work leaves room for many future directions. Among these are the extension of the prediction task to all queries, including those that are already part of a reformulation session; from another viewpoint, this amounts to predicting if a query will end the reformulation session. Using session characteristics, which have not

been used at all in this work, is likely to play a factor in this prediction task. For our own task(s), extending to more features, such as additional QPPs, can further enhance the prediction performance. Further ahead, predicting the type and semantics of the reformulation can help satisfy consumers' needs more rapidly and effectively, as reformulation plays such a central role in e-commerce search.

Acknowledgements This paper is based upon work supported in part by the Israel Science Foundation under grant no. 433/12, the German Research Foundation (DFG) via the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1), and an eBay grant.

REFERENCES

[1] Martin Arlitt. 2000. Characterizing web user sessions. *SIGMETRICS Perform. Eval. Rev.* 28, 2 (2000), 50–63.
[2] Ahmed Hassan Awadallah, Ranjitha Gurunath Kulkarni, Umut Ozertem, and Rosie Jones. 2015. Characterizing and predicting voice query reformulation. In *Proc. of CIKM*. 543–552.
[3] Ahmed Hassan Awadallah, Xiaolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proc. of CIKM*. 2019–2028.
[4] Ahmed Hassan Awadallah, Ryen W. White, Susan T. Dumais, and Yi-Min Wang. 2014. Struggling or exploring?: disambiguating long search sessions. In *Proc. of WSDM*. 53–62.
[5] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. 2017. *Deep learning*.
[6] Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, Aristides Gionis, and Sebastiano Vigna. 2008. The query-flow graph: model and applications. In *Proc. of CIKM*. 609–618.
[7] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[8] Eliot Brenner, Jun Zhao, Aliasgar Kutiyanawala, and Zheng Yan. 2018. End-to-end neural ranking for eCommerce product search: An application of task models and textual embeddings. In *The SIGIR 2018 Workshop On eCommerce*.
[9] David Carmel and Elad Yom-Tov. 2010. *Estimating the query difficulty for information retrieval*. Morgan & Claypool Publishers.
[10] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM TIST* 2, 3 (2011), 27.
[11] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proc. of KDD*. 785–794.
[12] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297.
[13] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting query performance. In *Proc. of SIGIR*. 299–306.
[14] Ronan Cummins, Joemon M. Jose, and Colm O'Riordan. 2011. Improved query performance prediction using standard deviation. In *Proc. of SIGIR*. 1089–1090.
[15] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proc. of ACL*. 1383–1392.
[16] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378–382.
[17] Jerome H Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of statistics* (2001), 1189–1232.
[18] Katharina C Furtner, Thomas Mandl, and Christa Womser-Hacker. 2015. Effects of Auto-Suggest on the Usability of Search in eCommerce.. In *ISI*. 178–190.
[19] Sreenivas Gollapudi, Samuel Ieong, and Anitha Kannan. 2012. Structured query reformulations in commerce search. In *Proc. of CIKM*. 1890–1894.
[20] Ido Guy and Kira Radinsky. 2017. Structuring the unstructured: From startup to making sense of eBay's huge eCommerce inventory. In *Proc. of SIGIR*. 1351.
[21] Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2018. Mix 'n match: Integrating text matching and product substitutability within product search. In *Proc. of CIKM*. 1373–1382.
[22] Claudia Hauff, Leif Azzopardi, and Djoerd Hiemstra. 2009. The combination and evaluation of query performance prediction methods. In *Proc. of ECIR*. 301–312.
[23] Yujing Hu, Qing Da, Anxiang Zeng, Yang Yu, and Yinghui Xu. 2018. Reinforcement learning to rank in e-Commerce search engine: Formalization, analysis, and application. In *Proc. of SIGKDD*. 368–377.
[24] Jeff Huang and Efthimis N. Efthimiadis. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In *Proc. of CIKM*. 77–86.
[25] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. 2009. Patterns of query reformulation during Web searching. *JASIST* 60, 7 (2009), 1358–1371.
[26] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W. White. 2015. Understanding and predicting graded search satisfaction. In *Proc. of WSDM*. 57–66.
[27] Jyun-Yu Jiang, Yen-Yu Ke, Pao-Yu Chien, and Pu-Jen Cheng. 2014. Learning user reformulation behavior for query auto-completion. In *Proc. of SIGIR*. 445–454.
[28] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proc. of CIKM*. 699–708.
[29] Li-Jen Kao and Yo-Ping Huang. 2017. Predicting purchase intention according to fan page users' sentiment. In *Proc. of SMC*. 831–835.
[30] Elad Kravi, Ido Guy, Avihai Mejer, David Carmel, Yoelle Maarek, Dan Pelleg, and Gilad Tsur. 2016. One Query, Many Clicks: Analysis of Queries with Multiple Clicks by the Same User. In *Proc. of CIKM*. 1423–1432.
[31] Rohan Kumar, Mohit Kumar, Neil Shah, and Christos Faloutsos. 2018. Did we get it right? predicting query performance in e-Commerce search. In *The SIGIR 2018 Workshop On eCommerce*.
[32] Or Levi, Ido Guy, Fiana Raiber, and Oren Kurland. 2018. Selective Cluster Presentation on the Search Results Page. *ACM TOIS* 36, 3, Article 28 (2018), 42 pages.
[33] Saurav Manchanda, Mohit Sharma, and George Karypis. 2019. Intent term selection and refinement in e-commerce queries. *CoRR* abs/1908.08564 (2019).
[34] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. of NIPS*. 3111–3119.
[35] Eric W Noreen. 1989. *Computer-intensive methods for testing hypotheses*. Wiley New York.
[36] Daan Odijk, Ryen W. White, Ahmed Hassan Awadallah, and Susan T. Dumais. 2015. Struggling and success in web search. In *Proc. of CIKM*. 1551–1560.
[37] Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proc. of SIGIR*. 275–281.
[38] Fiana Raiber and Oren Kurland. 2014. Query-performance prediction: setting the expectations straight. In *Proc. of SIGIR*. 13–22.
[39] Hadas Raviv, Oren Kurland, and David Carmel. 2014. Query performance prediction for entity retrieval. In *Proc. of SIGIR*. 1099–1102.
[40] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proc. of TREC-3*.
[41] Haggai Roitman, Shai Erera, Oren Sar Shalom, and Bar Weiner. 2017. Enhanced mean retrieval score estimation for query performance prediction. In *Proc. of ICTIR*. 35–42.
[42] Shubhra Kanti Karmaker Santu, Parikshit Sondhi, and ChengXiang Zhai. 2017. On application of learning to rank for e-Commerce search. In *Proc. of SIGIR*. 475–484.
[43] Nikos Sarkas, Stelios Paparizos, and Panayiotis Tsaparas. 2010. Structured annotations of web queries. In *Proc. of SIGMOD*. 771–782.
[44] Humphrey Sheil, Omer Rana, and Ronan Reilly. 2018. Predicting purchasing intent: Automatic feature learning using recurrent neural networks. In *The SIGIR 2018 Workshop On eCommerce*.
[45] Anna Shtok, Oren Kurland, and David Carmel. 2016. Query performance prediction using reference lists. *ACM Trans. Inf. Syst.* 34, 4 (2016), 19:1–19:34.
[46] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting query performance by query-drift estimation. *ACM Transactions on Information Systems* 30, 2 (2012), 11.
[47] Gyanit Singh, Nish Parikh, and Neel Sundaresn. 2011. User behavior in zero-recall ecommerce queries. In *Proc. of SIGIR*. 75–84.
[48] Parikshit Sondhi, Mohit Sharma, Pranam Kolari, and ChengXiang Zhai. 2018. A taxonomy of queries for e-Commerce search. In *Proc. of SIGIR*. 1245—-1248.
[49] Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. 2018. User Intent, behaviour, and perceived satisfaction in product search. In *Proc. of WSDM*. 547–555.
[50] Zehong Tan, Canran Xu, Mengjie Jiang, Hua Yang, and Xiaoyuan Wu. 2017. Query rewrite for null and low search results in eCommerce. In *The SIGIR 2017 Workshop On eCommerce*.
[51] Yongquan Tao and Shengli Wu. 2014. Query performance prediction by considering score magnitude and variance together. In *Proc. of CIKM*. 1891–1894.
[52] Arthur Toth, Louis Tan, Giuseppe Di Fabbrizio, and Ankur Datta. 2017. Predicting shopping behavior with mixture of RNNs. In *The SIGIR 2017 Workshop On eCommerce*.
[53] Hen Tzaban, Ido Guy, Asnat Greenstein-Messica, Arnon Dagan, Lior Rokach, and Bracha Shapira. 2020. Product Bundle Identification using Semi-Supervised Learning. In *Proc. of SIGIR*.
[54] Teng Xiao, Jiaxin Ren, Zaiqiao Meng, Huan Sun, and Shangsong Liang. 2019. Dynamic bayesian metric learning for personalized product search. In *Proc. of CIKM*. 1693–1702.
[55] Yingwei Xin, Ethan Hart, Vibhuti Mahajan, and Jean-David Ruvini. 2018. Learning better internal structure of words for sequence labeling. In *Proc. of EMNLP*. 2584–2593.
[56] Yuan Zhang, Dong Wang, and Yan Zhang. 2019. Neural IR meets graph embedding: A ranking model for product search. In *Proc. of WWW*. 2390–2400.
[57] Ying Zhao, Falk Scholer, and Yohannes Tsegay. 2008. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Proc. of ECIR*. 52–64.
[58] Yun Zhou and W. Bruce Croft. 2007. Query performance prediction in web search environments. In *Proc. of SIGIR*. 543–550.