Predicting Re-finding Activity and Difficulty

Sargol Sadeghi[†], Roi Blanco[‡], Peter Mika[‡],
Mark Sanderson[†], Falk Scholer[†], and David Vallet[*]

[†] RMIT University, Melbourne, Australia,
[‡] Yahoo! Research , Barcelona, Spain,
[*] Google, Sydney, Australia
{seyedeh.sadeghi,mark.sanderson,falk.scholer}@rmit.edu.com,
{roi,pmika}@yahoo-inc.com,
dvallet@google.com

Abstract. In this study, we address the problem of identifying if users are attempting to re-find information and estimating the level of difficulty of the re-finding task. We propose to consider the task information (e.g. multiple queries and click information) rather than only queries. Our resultant prediction models are shown to be significantly more accurate (by 2%) than the current state of the art. While past research assumes that previous search history of the user is available to the prediction model, we examine if re-finding detection is possible without access to this information. Our evaluation indicates that such detection is possible, but more challenging. We further describe the first predictive model in detecting re-finding difficulty, showing it to be significantly better than existing approaches for detecting general search difficulty.

Keywords: Re-finding Identification, Difficulty Detection, Behavioral Features

1 Introduction

Re-finding is a task where people seek information they have previously encountered. Examining a year of web search logs, Teevan et al. [18] determined that 40% of queries are attempts to *re-find*. While many such tasks are simple, such as searching for a home page, there are re-finding tasks that are more difficult, such as when only the broad sense of what was previously encountered can be recalled [17]. Current search engines are not optimised for re-finding [4, 16]. Being able to detect and estimate how *difficult* a re-finding task is proving to be, would enable a search engine to employ services to help the user, such as biasing results towards a searcher's history, or customizing snippets to include texts and images that might be more memorable.

Research on re-finding difficulty has focused on users coping with changes to web sites and search results [2, 16]. Difficulties have been studied for specific application areas, such as email search [3, 4]. Beyond re-finding, identifying user difficulties has been explored for different task types. For example, Liu et al. [12, 13] have shown that it is useful for IR systems to predict when a user is struggling, where systems could consequently adapt search results.

Current re-finding prediction is limited to the level of queries [18]. Because a re-finding user will likely engage in multiple searches, prediction of re-finding beyond a

single query is crucial. Past research has also emphasized the importance of *tasks* either in identifying re-finding behavior [2], or generally detecting difficulties [12]. Although task-level re-finding has been examined [2, 19], the work is limited in using behavioural features to predict re-finding tasks. As users can easily encounter information items through browsing, or receiving information via a social network, it is also important to examine how the identification of re-finding can be performed independent of the search history of the user using behavioural features.

Two research questions are explored: (1) Re-finding identification: How can re-finding tasks be differentiated from general web search tasks? (2) Re-finding difficulty: What features characterize user difficulties in completing a re-finding task?

We first describe past work, followed by a description of the experimental methodology. Next, we explain the features used in the predictive model. We then detail the setup of the prediction models, along with results from a range of experiments exploring different types of re-finding and feature sets.

2 Related Work

Re-finding Identification. In one of the first studies on web-based re-finding, Teevan et al. [18] used query log features to predict if the same result would be clicked on by a user given that they had re-submitted a previously entered query. Tyler and Teevan [19] studied re-finding at the level of sessions, finding that queries change more across sessions than within. Later, Tyler et al. [20] examined query features and the rank of the clicks to identify re-finding. Capra [2], studying 18 search tasks of users, found it difficult to distinguish between generic web search engine use and re-finding. From a diary study by Elsweiler and Ruthven [5], re-finding tasks were classified using the granularity of the information to be re-found (lookup, one-item, and multi-item).

Many search features were studied in the related area of predicting task continuation and cross-session tasks [11, 21]. In a study by Kotov et al. [11], session-based features (e.g. "number of queries since the beginning of the session"), history-based features (e.g. "whether the same query appeared in the user's search history"), and pair-wise features (e.g. "number of overlapping terms between two queries") were examined.

Overall, current studied behavioural features for the re-finding context are limited and dependent on the search history of the user. However, for identifying particularly difficult re-finding tasks, it is required to examine a broader range of features.

Re-finding Difficulty. Capra [2] explored features to detect user difficulty including the number of search URLs, task completion time, and the elapsed time between search tasks. The best features included task frequency, topic familiarity, and determining that target information had been moved from the page where it was originally found. Teevan highlighted [16] information being moved, as well as changes in target document rank position, as causes of re-finding difficulty. She found that changes in the path to reach target information was a stronger indicator of user difficulty than temporal features. Elsweiler and Ruthven [5] studied the granularity of information and found no significant influence of granularity on difficulty. However, they reported that longer time gaps could indicate that users were having difficulties for some re-finding.

In general web search, large-scale query log features have been used to predict search difficulty [12, 13], as well as user frustration, dissatisfaction, or success/failure [1, 7–9]. Features ranged from temporal to user behavioral, and search result ranks. Examples of studied features include time interval between queries, number of clicks with high dwell time, and mean reciprocal ranks of clicks for each query.

Overall, current examined features for detecting difficulties in re-finding are mainly limited to user's self assessed features (e.g. topic familiarity) or target information (e.g. moved web page), and the construction of predictive models using behavioural features has not been considered.

3 Experimental Methodology

Our prediction model is based on the analysis of query logs. In this section, we describe the explored data sets and the methodology for evaluation.

3.1 Dataset

Our data consists of a sample of logs taken from 30 days of interactions with the Yahoo search engine gathered from the $1^{st} - 30^{th}$ of October 2012. The interactions of 2,847,028 unique anonymised users were logged including submitted queries, the URL, the rank position of clicked search results, and a timestamp for each event. The terms of service and privacy policies of Yahoo were followed.

To identify task boundaries, the logs were segmented into *goals*, which is defined as a group of related queries and corresponding clicks submitted by a user to perform a task with an atomic search need. Goal segmentation was performed using the technique described by Jones and Klinkner [10], where classifiers are used to predict goal boundaries based on features indicative of relatedness between queries (e.g. number of words in common) with an accuracy of 92. Note that other log segmentation approaches are either less accurate (e.g. sessions), or consist of more than one information need (e.g. missions) [10], and therefore we considered the goal segmentation. All goals from the same user were extracted and ordered by their timestamp, and all possible goals were *paired*. As we were not interested in short-term re-finding, paired goals that occurred less than thirty minutes apart were not considered. In total, 39,683,301 paired search goals were extracted.

3.2 Potential Re-finding Goals

Teevan et al. [18] classified pairs of queries and clicks into different types of re-finding. They examined whether the paired queries were equal or not, and explored result click overlap. We extend the approach to the level of pairs of goals across multiple queries and clicks.

We measure queries and clicks equivalence using a 5-point scale, resulting in a total of 25 combined classes. For queries, this includes sharing a term, term stem, or term corrections (simple edits for the purpose of spelling correction). For clicks, equivalence levels include overlapping URLs as well as at what point in the goal the overlapping clicks occurred. For example, common clicks that occurred at the end of a goal

Query Overlap	URL Overlap
Query	Last URL + URL
Query Term	Last URL + URL Root
Term Correction	Last URL
Term Stem	URL
No Query Overlap	URL Root

Original Goal
Q: bleacher report college football T: 2
C(3): `www.cbssports.com/collegefootball` T: 15
C(10): `bleacherreport.com/college-football`
Re-finding Goal
Q: college fottball T: 2 *Query term overlap*
Q: college fottball T: 9 *Query term correction*
C(1): `espn.go.com/college-football/` T: 16
C(39): `www.cbssports.com/collegefootball` T: 20 *URL overlap*
C(43): `bleacherreport.com/college-football` *Last URL overlap*
Classification: Query term overlap, Last URL + URL overlap

Fig. 1. Left: Definitions of query and click overlaps used across paired goals. Right: An example paired goal from the logs, with its classification.

(*last URL*) are distinguished. We also considered whether two URLs matched fully or only partially (based on the server name or *URL root*). As an example the overlap between these two URLs is considered as the URL root overlap: `en.wikipedia.org/wiki/Doc_Martin` and `en.wikipedia.org/wiki/Dr._Martin`. The query and click levels with some examples are illustrated in Figure 1. If a paired goal could belong to more than one class, the most restrictive class was selected. Paired goals where there was no URL overlapping were eliminated, as some minimum level of click commonality was required [18, 19]. From the overlapping classes, 4,968,243 paired goals were extracted for our dataset. Note that the proposed classes are means to identify potential re-finding cases through the overlapping between parts of a paired goal; however, this does not mean that each overlapping is certainly a re-finding case. For example, users might repeat the same query but with a different search need, or clicks might have overlapping in their root URL, while referring to two different documents.

On the other hand, we note that there are other potential types of re-finding as shown in Figure 2. The paired goals might not always have overlapping in clicked URLs, such as cases where the URL has changed by the time that re-finding is attempted, but the corresponding web document is the same; or where the user failed to reach the same target document, thus having the same task but not resulting on overlapping URLs. We refer to this type of re-finding as *paired* but with *no URL overlapped*. Moreover, we made an assumption that there is a corresponding original search for each identified re-finding task (*paired* goals); whereas in some cases re-finding could occur in an *isolated* form. An example of the isolated re-finding is when the searcher cannot be identified (e.g. no login information, accessing from a different location), or the information being re-found may originally have been found by means other than searching (e.g. browsing, or social links). While, these cases might be more likely to include difficult re-finding, the identification of such cases is challenging from a query log study and is left for future work. However, we focused on those *URL overlapped paired* goals that are *non-navigational* and more likely include difficult cases.

Teevan et al. noted that much re-finding, such as navigational searches, are easy to detect. The navigational searches were identified based on equal query and single identical clicks. As the focus of our work was detecting more challenging forms of re-finding, we created a set of filters to remove easy cases. Paired goals where the queries

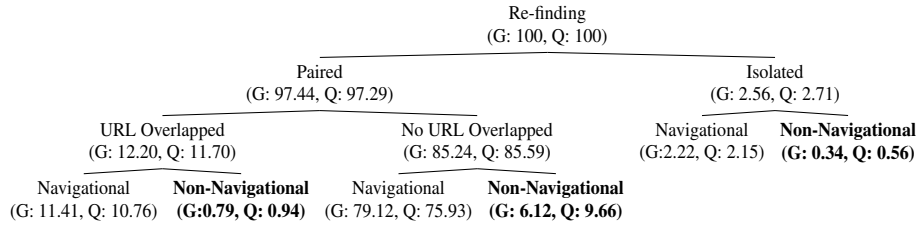

Fig. 2. The landscape of re-finding tasks. G: The percentage of goals, Q: The percentage of queries.

contained only popular domain names[1] or terms such as "www" and ".com" were removed. If the domain name of the clicked URL matched the corresponding submitted query, or was a spell-corrected version of the query, the paired goals were also removed. Only paired goals where each element of the pair contained more than one query or one click were considered. Filter accuracy was checked by manual investigation of a sample of paired goals. Our analysis of removed pairs showed that at worst only 1.6% were incorrectly removed.

After removing the easy paired goals, 322,639 pairs remained. The large reduction in size of data does not necessarily reflect that the re-finding problem we study is small; rather, applying our filtering rules is a way of giving us a dataset where we are confident we will find a concentration of challenging re-finding problems. Once a re-finding classification is constructed from this data set, other examples of re-finding can be explored in the full query logs. The summed percentages of non-navigational goals in Figure 2 is 7.25% (i.e. 0.79% + 6.12%+ 0.34%). Detecting and eventually helping users with their re-finding goals from this notable fraction of the query log has the potential to provide help on the most difficult re-finding tasks.

3.3 Ground Truth Dataset

We manually label re-finding activity and re-finding difficulty. To the best of our knowledge, this study is the first to include labeling of re-finding difficulty in web search logs.

We designed a labeling interface where paired goals were presented showing queries, clicked URLs plus their rank, the time gap between queries and clicks, as well as the gap between the paired search goals. Each assessor was asked to answer two questions: 1) *"Do you think that in the second search the user is re-finding document(s) that were found in the first search?"* (Responses: "yes", "no", and "not sure")? [2] 2) *"In terms of search difficulty, would you say the second search is..."* (Responses: "easy", "difficult", "not sure")? The notion of "difficulty" was defined for assessors in a broad sense of whether it seems that the user is struggling to find the target document. Specifically assessors were instructed to consider the effort of the user in a) reformulating queries,

[1] Identified through top 50 ranked websites from `Alexa.com` (e.g. "youtube").

[2] Note, we ask about re-finding the same *document(s)* not the same *information*, as there could be cases where it might not be possible to infer whether the user was searching for the same information (due to the dynamic content of web documents, for example news pages).

b) clicking relevant documents, and c) recognizing the target document. Examples of easy and difficult cases were shown to assessors.

All 25 combined classes of query overlap/URL overlap levels in Figure 1 were uniformly sampled (75 pairs from each class). However, eight were low in frequency (fewer than 25), and so were not considered in our sampling. In total, 1,275 paired goals were labeled by an *experienced* assessor, who had conducted the same labeling exercise on a separate dataset. The fraction of "not sure" labels was 8%, which reduced the size of our data to 1,167.

Examining ground-truth data reliability, we randomly sampled sixty instances and asked three other experienced assessors to assign labels once more. Mean pairwise Cohen's kappa (κ) for inter-assessor agreement was 0.89 and 0.47 for identifying re-finding and difficulty assignments respectively. By comparison, κ agreement scores in the range of 0.23-0.71 were achieved for relevance judgments for the TREC Legal Track [22]. This gave us confidence that the ground-truth data is sufficiently consistent.

We noticed a low frequency of goals labeled "difficult" making the data set imbalanced, which could be due to the limitation in the identification of re-finding based on query/click overlapping as discussed in Section 3.2; whereas in more difficult cases a fewer number of overlapping could occur, as the user might not be able to repeat queries and clicks from the original search. Consequently, we employed a form of active learning to increase the frequency of difficult instances in our training set. A classification model was learnt on our original labeled data and applied to unlabeled goals taken from the unlabelled data. The goals were ranked based on the estimated probability of them belonging to the "difficult" class. The top fifty, along with ten random instances from the rest of the predictions, were manually labeled and added to our data set. The procedure was repeated for ten iterations; at this point, a balanced number of "difficult" labels (48.3% of the identified paired goals) were obtained, and the procedure was stopped. After removing the "not sure" labels, the size of our final training set was 1,706 (with 74.4% re-finding case). This data was used for training and evaluating our classification models.

4 Features

This section explains the set of features that were used to construct predictive models for the identification and difficulty classification of re-finding.

4.1 Feature Categories

Features in three main groups are considered: (1) baseline query-level features from past research [18] ; (2) features from general web search related studies on detecting search difficulty and failure ; and (3) new features extended in our study for the re-finding context. All features considered are listed in Table 1. Most features are numerical, except for some Boolean features such as *"ended with query"*, *"exist advanced query syntax"*, *"all common clicks skipped"*, *"exist jumped common clicks"*, *"exist non-sequential clicks"*, and *"exist common clicks in different ranks between original and re-finding"* goals.[3]

[3] A detailed description of features: `http://tinyurl.com/feature-description`

Some features, indicated by '*' in the table, can be measured across the paired goals, in addition to being measured on each goal independently. For example, for the feature *"goal length in no. of queries"*, the pair-wise version of this feature would measure the relative difference of the goal length between the original and re-finding paired goal. For starred numerical features, we measure the difference between the paired goals; for Boolean features, we apply logical 'and' between the corresponding values of each goal. Given the defined notions in Table 1, the total number of features that could be calculated for a paired goal is 124.

We further separate the features into two broader groups: those requiring access to the original goal (*history-dependent*) and those that do not (*history-independent*, i.e. current goal only). This could be particularly useful for identifying *no URL overlapped* and *isolated* re-finding tasks illustrated in Figure 2.

4.2 Feature Discussion

The two features *"all common clicks skipped"* and *"exist jumped common clicks"* were inspired by a related study [15], which re-ranks repeated search results based on the behavior of users in clicking, skipping, or missing results. As our log data did not contain viewed results, we implemented a similar idea for clicked results in relation to their ranks. The first feature indicates whether there is a click at a lower rank, followed by the common clicks at higher ranks. The second feature indicates whether there is a common click, followed by a click at a higher rank. These assumptions are based on the fact that the user is likely to browse the result page from top to bottom. The feature *"exist common click in different ranks within pairs"* was inspired by Teevan's study [16], where changes in the rank of the clicks make re-finding difficult. Moreover, we added a condition that common clicks in following result pages could increase the difficulty of the re-finding task (*"no. of non-first-page ranked clicks"*). Some features considered the position of common clicks. For example, *"common click in relation to the last click"* examines whether a common click occurred in the last click of the original and re-finding paired goal. In terms of the importance of engaged clicks, we developed the feature, *"missed engaged later clicks in original"*, which is true if, after a common click, there are engaged clicks in the original goal that have not been clicked in the re-finding goal.

A dwell time of greater than 30 seconds has been highlighted as an indication of *engaged* and relevant clicks [9]. We added "relative dwell time", which is computed in terms of the fraction of click dwell time to the total time-span of the goal. Dwell time after clicks might not be entirely reflective of search time, as the user might spend time on acquiring knowledge, or inspecting a document. Therefore, we define *"effective search time"*: the total dwell time after queries and those clicks that have low dwell time (less than 30 seconds).

The feature *"query overlap/URL overlap"* is defined in terms of the classification between query and click commonalities of paired goals (see Figure 1). More commonality could increase the chance of re-finding. On the other hand, differences could be indicative of greater difficulties. As an example, *"first query transformation type within pairs"* measures the differences between the initial queries of the original and re-finding goals (based on query reformulation types: "exactly the same", "error correction", "specialization", "generalization", and non-trivial transitions considered as "other").

Table 1. Features used to detect re-finding and difficulties. Each feature could be related to either original goal: †, or re-finding goal: ‡, or a relative difference between both goals: ∗. Features signed by † and ∗ are *history-dependent*; whereas, ‡ features are *history-independent*.

Baseline query level features (from past re-finding work)
equal query class ∗
equal query elapsed time ∗
equal query length ∗
equal query no. of original clicks †
equal query no. of common clicks ∗
equal query no. of original uncommon clicks †
General web search (related) difficulty features
goal length in no. of both queries and clicks ‡
goal length in no. of unique/all queries ‡
goal length in no. of unique/all clicks ‡
mean no. of clicks across all queries ‡
time to the first click ‡
min/max/mean time to the first click of all queries ‡
min/max/mean inter-query time ‡
min/max/mean inter-click time ‡
no. of engaged clicks (dwell time >30 seconds) ‡
no. of clicks on next page ‡
ended with query ‡
exist advanced query syntax (e.g. quotes) ‡
queries per second ‡
clicks per query ‡
fraction of queries for which no click ‡
time span of goal ‡
Extended re-finding features
query overlap/URL overlap ∗
no. of common/uncommon/all clicks † ‡
mean query length of common/all clicks † ‡
mean no. of query common/all clicks † ‡
mean no. of uncommon clicks of all queries † ‡
mean no. of uncommon clicks of common click queries † ‡
days between paired goals ∗
effective search time † ‡ ∗
total dwell time after all queries † ‡
total dwell time after all clicks † ‡
total time to reach to the first common click † ‡
rank of the first reached common click † ‡
mean reciprocal rank of common clicks † ‡
rank of the last click † ‡
no. of non-first-page ranked clicks in common/all clicks † ‡
all common clicks skipped † ‡
exist jumped common clicks † ‡
exist non-sequential clicks † ‡
mean dwell time/relative dwell time of common clicks † ‡
no. of repetitions of common clicks † ‡
fraction of queries with no common clicks † ‡
re-finding is longer than original in length ∗
re-finding is longer than original in no. of queries ∗
re-finding is longer than original in no. of clicks ∗
re-finding missed engaged later clicks in original ∗
first query transformation type within pairs ∗
exist common click in different ranks within pairs ∗
common click in relation to the last click ∗
mean relative goal position of common clicks † ‡
min/max goal position of common clicks † ‡
mean relative common clicks goal position (early, middle, late) † ‡
goal length in no. of both queries and clicks † ∗
goal length in no. of unique/all queries † ∗
goal length in no. of unique/all clicks † ∗
mean no. of clicks across all queries †
time to the first click †
min/max/mean time to the first click of all queries †
min/max/mean inter-query time †
min/max/mean inter-click time †
no. of engaged clicks (dwell time >30 seconds) †
no. of clicks on next page †
ended with query † ∗
exist advanced query syntax (e.g. quotes) † ∗
queries per second † ∗
clicks per query † ∗
fraction of queries for which no click † ∗
time span of goal †

5 Prediction Models

We used Support Vector Machines as our classification model, trained with a Sequential Minimal Optimization (SMO) algorithm, as this has been shown to work well in similar classification scenarios [18]. We trained a binary classifier to classify a goal as re-finding or not; and the second one to predict re-finding difficulty (easy or difficult).

We employed a ten times ten-fold cross-validation approach, which repeats ten-fold cross-validation and measures the average of classification results [14]. We report precision, recall, and F-measure scores. A paired two-tailed t-test was used to test for statistically significant differences in effectiveness.

Table 2 reports the accuracy when using different groups of features (see Table 1). Considering the columns *all features* in Table 2 (using all features in Table 1), our SMO classifier achieves an F-measure of 91.6 on the identification problem (left table), and 82.7 on the difficulty prediction problem (right table). We replicated a model proposed by Teevan et al. [18] as a state of the art baseline, which used the "Baseline query level features" introduced in Table 1. It can be seen that re-finding identification improves from 89.8 to 91.6, a relative increase of 2.0%. Examining re-finding difficulty, we obtain

Table 2. Re-finding classification performance of feature sets measured using P: Precision, R: Recall, and F: F-measure.

	All features	History-dependent	History-independent
Baseline query level identification	P: 89.8 [1] R: 89.8 F: 89.8	P: 89.8 R: 89.8 F: 89.8	-
Re-finding identification	P: 91.6 R: 91.7 F: 91.6	P: 91.6 R: 91.7 F: 91.6	P: 67.6 R: 74.0 F: 70.7

[1] The same as history-dependent.

	All features	History-dependent	History-independent
General web search difficulty	P: 79.2 [2] R: 78.9 F: 79.0	-	P: 79.2 R: 78.9 F: 79.0
Re-finding difficulty	P: 82.8 R: 82.7 F: 82.7	P: 81.0 R: 80.9 F: 80.9	P: 79.3 R: 79.0 F: 79.1

[2] The same as history-independent.

4.7% relative improvements compared to the best found baseline, which was trained on "General web search difficulty features" in Table 1. The changes in F-measure scores are all statistically significant ($p < 0.05$) with the Cohen's effect size of 1.4 and 1.2 for re-finding identification and difficulty detection using all features.

The vast majority of re-finding research has focussed on re-finding where the information was originally found with a search engine and that finding activity was logged. We also consider the detection of re-finding without the information from the original (historical) goal. Using only history-independent features reduces re-finding accuracy (F-measure of 70.7); past work has not considered this type of identification, so there is no baseline to compare to (and the scores of the baseline using all features and history-dependent features are the same). We plan to improve the performance of this classification by studying history-independent features in future work, which enables the identification of more challenging re-finding tasks. Examining the history-independent column for the difficulty problem, similar accuracy was obtained for both re-finding and general search. However, features from the history-dependent group improve the performance of the classifier (i.e. 80.9).

6 Feature Importance Analysis

We calculated the information gain of each individual feature in order to assess their importance for the two prediction tasks. This measure estimates the amount of information that can be obtained about the class prediction from each feature [6]. The ten with the highest information gain are shown in Table 3. Some features are related to commonalities between paired goals (e.g. *"min goal position of common clicks"*), whilst others record measurements across a goal (e.g. *"effective search time"*). We start by analysing all features from paired goals.

All Features. Perhaps unsurprisingly, the most important feature was (*"query overlap/URL overlap"*) measuring the level of query and clicked URL overlap between the paired goals. This categorization ranks higher than all features used in past work [18]. Contextual features appeared to be important for identifying re-finding such as *"common click in relation to the last click"*, *"no. of common clicks"*, or *"mean query length of common clicks"*.

The first ranked feature for difficulty detection was *"effective search time"*, which was a stronger indicator than the length of the search measured in queries and clicks

(e.g. *"goal length in no. of both queries and clicks"*). The *"total dwell time after all queries"* was second. The corresponding feature for clicks (i.e. *"total dwell time after all clicks"*) did not appear in the top ten, suggesting that time spent after submitting queries is more likely to be representative of task difficulty than the time allocated after clicks.

Among other features in Section 4.2 that were not ranked in the top ten, but still ranked relatively strongly, *"all common clicks skipped"* and *"missed engaged later clicks in original"* appeared to be more effective in the identification of re-finding rather than difficulty detection. These features could provide signals that the user is not interested in previously seen documents, and therefore the underlying task is not re-finding. The information gain of *"no. of non-first-page ranked common clicks"* indicated that when the user navigates to the next result page, it is more indicative of search difficulty than re-finding. Similarly for the *"exist jumped common click"*, jumping to the previously seen document could be more indicative of an easy task in recognizing a target document rather than a particular re-finding behavior.

History-independent Features. We also ranked history-independent features as shown in Table 3. It appeared that time-based features are important in identifying re-finding tasks independent of the search history of the user. As an example, *"max inter-click time"* acquired the highest information gain. Here, the time spent between clicks seems to be more important than the time between queries (i.e. *"max inter-query time"*). Other features indicative of the goal length in terms of number of queries/clicks and also the length of the queries obtained the top ranks in the identification of re-finding.

The top features indicative of difficulty in re-finding (discussed above) are history-independent (*"effective search time"* and *"total dwell time after queries"*). Apart from the proposed features in this study, there are other features from past research, which are also indicative of difficulty in re-finding, such as time to the first click and the number of engaged clicks [9].

In comparing re-finding identification features with difficulty indications, it can be seen that "goal no. of all queries" is a stronger signal for the identification of re-finding; whereas, "goal no. of all clicks" is more important in detecting the difficulty of the task. Some features particularly indicative of re-finding difficulty were history-independent and some could be computed during the search (e.g. *"mean time to first clicks"*). The latter features are referred to as *real-time* in the literature [13], and search engines that make use of them could provide real-time predictions. Using all the developed features in this study, we measured the accuracy of predictions given partial information from the beginning of re-finding tasks (after 2, 4, 8, 16, 32, and 64 seconds). The average F-score of 83.7 and 74.3 were obtained for re-finding identification and difficulty detection respectively, which could indicate the predictability of these two tasks at real-time for an online user support that can be further explored in future work.

7 Conclusions and Future Work

This paper focuses on better understanding re-finding behavior by answering two questions: a) how can re-finding tasks be differentiated from general web search tasks; and b) what features characterize user difficulties in completing a re-finding task.

Table 3. Top 10 features for re-finding identification and difficulty detection ranked by information gain. A †, ‡, or ∗ indicate feature related to original, re-finding or both, respectively.

	All features	History-independent
Re-finding identification	1. query overlap/ URL overlap ∗	1. max inter-click time ‡
	2. common click in relation to the last click ∗	2. goal no. of all queries ‡
	3. no. of common clicks ∗	3. max inter-query time ‡
	4. equal query class ∗	4. total dwell time after clicks ‡
	5. mean no. of clicks for common click queries ‡	5. mean inter-click time ‡
	6. max goal position of common clicks ‡	6. mean inter-query time ‡
	7. min goal position of common clicks †	7. total dwell time ‡
	8. mean relative goal position of common clicks ‡	8. clicks per query ‡
	9. mean no. of clicks for common click queries †	9. mean no. of clicks across all queries ‡
	10. mean query length of common clicks ‡	10. mean query length of all clicks ‡
Re-finding difficulty	1. effective search time ‡	1. effective search time ‡
	2. total dwell time after all queries ‡	2. total dwell time after all queries ‡
	3. max goal position of common clicks ‡	3. goal no. of all clicks ‡
	4. goal length in no. of all clicks ‡	4. goal length in no. of both queries and clicks ‡
	5. goal length in no. of both queries and clicks ‡	5. max time to first clicks ‡
	6. max time to the first click of all queries ‡	6. mean time to first query clicks ‡
	7. mean time to the first click of all queries ‡	7. goal no. of unique clicks ‡
	8. goal length in no. of unique clicks ‡	8. no. of engaged clicks ‡
	9. no. of engaged clicks ‡	9. goal no. of all queries ‡
	10. goal length in no. of all queries ‡	10. no. of clicks on next page ‡

We proposed a set of features and constructed prediction models for both re-finding identification and difficulty detection. Classifiers built using our feature sets achieved an F-measure of 91.6 for identifying re-finding, and 82.7 for predicting re-finding difficulty. Our model significantly outperformed existing state of the art re-finding identification approaches, which are based on query repetitions and dependent on the search history of the user, with a 2.0% improvement in accuracy. To the best of our knowledge, our work is the first to investigate the re-finding difficulty classification problem; we therefore compared our results against an adaptation of general web task difficulty detection approaches, resulting in a significant improvement of 4.7% for difficulty detection. We examined the effectiveness of predictors based on features, which can be computed without identifying the user and their search history. In this case, we obtained F-measure scores of 70.7 and 79.1 for detecting re-finding and difficulty respectively. The history-independent analysis could enable the identification of more complex re-finding tasks, which was not addressed in past research.

An analysis of the effectiveness of individual features for the two re-finding classification problems demonstrated that our proposed features, such as *"query overlap/URL overlap"* and the use of the *"effective search time"*, are ranked highly in terms of their information gain impact. Our analysis showed that some top ranked features can be calculated as the search task progresses (e.g. *"time to first click"*), which means that search engines can potentially take advantage of real-time prediction, even if there is no access to the search history of the user.

In future work, we plan to investigate further improvements to our predictive models by incorporating more real-time and fewer history-dependent features, and identify more distinctive behavioural features from a general search task. Moreover, some basic hypotheses in this study can be extended and further examined. For instance, instead of pairing sequential goals from the same user, we could also take into consideration chains of goals (due to the repeated nature of re-finding tasks). Furthermore, it would be

interesting to carry out controlled user experiments to identify and incorporate user-side factors that cannot be derived from query log analysis.

References

1. Ageev, M., Guo, Q., Lagun, D., Agichtein, E.: Find it if you can: A game for modeling different types of web search success using interaction data. In: Proc. SIGIR. pp. 345–354. ACM (2011)
2. Capra III, R.G.: An investigation of finding and refinding information on the web. Ph.D. thesis, Virginia Polytechnic Institute and State University (2006)
3. Elsweiler, D., Baillie, M., Ruthven, I.: What makes re-finding information difficult? a study of email re-finding. In: Advances in information retrieval, pp. 568–579. Springer (2011)
4. Elsweiler, D., Harvey, M., Hacker, M.: Understanding re-finding behavior in naturalistic email interaction logs. In: Proc. SIGIR. pp. 35–44. ACM (2011)
5. Elsweiler, D., Ruthven, I.: Towards task-based personal information management evaluations. In: Proc. SIGIR. pp. 23–30. ACM (2007)
6. Hall, M.A.: Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato (1999)
7. Hassan, A., Jones, R., Klinkner, K.L.: Beyond DCG: User behavior as a predictor of a successful search. In: Proc. WSDM. pp. 221–230. ACM (2010)
8. Hassan, A., Shi, X., Craswell, N., Ramsey, B.: Beyond clicks: Query reformulation as a predictor of search satisfaction. In: Proc. CIKM. pp. 2019–2028. ACM (2013)
9. Hassan, A., Song, Y., He, L.w.: A task level metric for measuring web search satisfaction and its application on improving relevance estimation. In: Proc. CIKM. pp. 125–134. ACM (2011)
10. Jones, R., Klinkner, K.L.: Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In: Proc. CIKM. pp. 699–708. ACM (2008)
11. Kotov, A., Bennett, P.N., White, R.W., Dumais, S.T., Teevan, J.: Modeling and analysis of cross-session search tasks. In: Proc. SIGIR. pp. 5–14. ACM (2011)
12. Liu, J., Gwizdka, J., Liu, C., Belkin, N.J.: Predicting task difficulty for different task types. Proc. ASIST 47(1), 1–10 (2010)
13. Liu, J., Liu, C., Cole, M., Belkin, N.J., Zhang, X.: Exploring and predicting search task difficulty. In: Proc. CIKM. pp. 1313–1322. ACM (2012)
14. Nadeau, C., Bengio, Y.: Inference for the generalization error. Machine Learning 52(3), 239–281 (2003)
15. Shokouhi, M., White, R.W., Bennett, P., Radlinski, F.: Fighting search engine amnesia: Reranking repeated results. In: Proc. SIGIR. pp. 273–282. ACM (2013)
16. Teevan, J.: Supporting finding and re-finding through personalization. Ph.D. thesis, Massachusetts Institute of Technology (2006)
17. Teevan, J.: How people recall, recognize, and reuse search results. TOIS 26(4), 19 (2008)
18. Teevan, J., Adar, E., Jones, R., Potts, M.A.: Information re-retrieval: Repeat queries in yahoo's logs. In: Proc. SIGIR. pp. 151–158. ACM (2007)
19. Tyler, S.K., Teevan, J.: Large scale query log analysis of re-finding. In: Proc. WSDM. pp. 191–200. ACM (2010)
20. Tyler, S.K., Wang, J., Zhang, Y.: Utilizing re-finding for personalized information retrieval. In: Proc. CIKM. pp. 1469–1472. ACM (2010)
21. Wang, H., Song, Y., Chang, M.W., He, X., White, R.W., Chu, W.: Learning to extract cross-session search tasks. In: Proc. WWW. pp. 1353–1364. International World Wide Web Conferences Steering Committee (2013)
22. Webber, W., Toth, B., Desamito, M.: Effect of written instructions on assessor agreement. In: Proc. SIGIR. pp. 1053–1054. ACM (2012)