Re-finding Behaviour in Vertical Domains

SEYEDEH SARGOL SADEGHI, RMIT University
ROI BLANCO, Yahoo Research
PETER MIKA, Yahoo Research
MARK SANDERSON, RMIT University
FALK SCHOLER, RMIT University
DAVID VALLET, Google

Re-finding is the process of searching for information that a user has previously encountered, and is a common activity carried out with information retrieval systems. In this work, we investigate re-finding in the context of vertical search, differentiating and modeling user re-finding behavior within different media and topic domains, including images, news, reference material, and movies. We distinguish the re-finding behavior in vertical domains from re-finding in a general search context, and engineer features that are effective in differentiating re-finding across the domains. The features are then used to build machine-learned models, achieving an accuracy of re-finding detection in verticals of 85.7% on average. Our results demonstrate that detecting re-finding in specific verticals is more difficult than examining re-finding for general search tasks. We then investigate the effectiveness of differentiating re-finding behavior in two restricted contexts: We consider the case where the history of a searcher's interactions with the search system is not available. In this scenario, our features and models achieve an average accuracy of 77.5% across the domains. We then examine the detection of re-finding during the early part of a search session. Both of these restrictions represent potential real-world search scenarios, where a system is attempting to learn about a user but may have limited information available. Finally, we investigate in which types of domains is re-finding most difficult. Here, it would appear that re-finding images is particularly challenging for users. This research has implications for search engine design, in terms of adapting search results by predicting the type of user tasks, and potentially enabling the presentation of vertical-specific results when re-finding is identified. To the best of our knowledge, this is the first work to investigate the issue of vertical re-finding.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms: Performance, Measurement, Experimentation, Human Factors

Additional Key Words and Phrases: Re-finding Behavior, Search Feature, Vertical, Predictive Models, Difficulty

1. INTRODUCTION

Current search engines extend traditional search results to incorporate answers from a variety of media types and domains (including, but not limited to, videos, images, and news), which are referred to as *verticals* [Arguello et al. 2009]. Modern search engines dynamically merge the results from different verticals on the main search results page to better satisfy users' information needs.

One type of search task that is conducted by users is a process known as *re-finding*: locating an information item that searchers previously encountered. This is a search task that is both common [Teevan et al. 2007] and at times challenging for users [Capra III 2006; Elsweiler and Ruthven 2007; Teevan 2006], particularly, if the user cannot remember how they previously encountered the information item. Detecting re-finding activity while a user is searching might allow a search engine to adapt results to help the user locate the item they seek.

The fundamental features and re-finding extraction in this paper are based on our previous work on predicting re-finding tasks and search difficulty [Sadeghi et al. 2015]; however, in this work, new vertical identification approach was proposed and the features were studied across different verticals and also customized for different explorations such as re-finding predictions at the early stage of search tasks.

DOI: http://dx.doi.org/10.1145/0000000.0000000

Original Goal

Q: most visited websites T: 2
C(1): alexa.com/topsites T: 25
C(2): ebizmba.com/articles/most-popular-websites T:17
C(6): youtube.com/watch?v=8g5D_Ks3ago T:27
C(8): computerweekly.com/news/one-billion-visits T:30
C(4): en.wikipedia.org/wiki/most_popular_websites T: 35
C(5): answers.yahoo.com/question/index?qid=200830AAF0

Re-finding Goal

Q: which sites people see more T: 45
Q: frequently seen sites T: 40
Q: web sites frequently seen T: 15
C(4): en.wikipedia.org/wiki/Wikipedia:Popular_pages T: 28
C(5): en.wikipedia.org/wiki/most_popular_websites T: 40
Q: popular websites T: 5
C(4): answers.yahoo.com/question/index?qid=200830AAF0

Fig. 1. Each line presents a query (Q) or a click (C). C(n) indicates the rank of the clicked URL. T: n represents the dwell time in seconds between queries and clicks. Different documents in different orders might be retrieved in the later visit of the user (i.e. re-finding goal).

Figure 1 illustrates an example where the user has clicked documents coming from different verticals in the *original* search task; however, later on the user is looking for a particular reference link (*re-finding* task), but is unable to remember the previous query and where that particular link was seen (e.g. through the links in a Wikipedia page or Yahoo! answers page, or somewhere else). If these cases are properly identified, a search engine could guide the user to the particular vertical reference documents, as the user is not interested in results coming from diverse verticals.

While re-finding has been studied for some time [Teevan et al. 2007], re-finding *behavior* for particular topical domains or media types remains a largely unexplored topic. We investigate whether there are differences between re-finding across these vertical domains and also in comparison to general web search. We define *general search* as a task when the user is not aware of the type of the document that could address their information need and therefore multiple verticals can be applicable to the user's task, while *vertical search* is viewed as retrieval where the domain or media type of the document is part of the user's information need. In this study a *"vertical"* refers to the topic domain of documents that are listed in the main search result page (regardless of whether the documents come from specialized collections, or ranked general web results).

The three research questions addressed in this work are:

— RQ 1: What features are effective in distinguishing re-finding tasks in different verticals?
— RQ 2: How predictable is re-finding within each vertical in contrast to searches that are not re-finding?
— RQ 3: What are the types of vertical documents that users have more difficulty in re-finding?

To investigate if such differences between general and re-finding in verticals can be modeled and predicted, we examine a set of behavioral search features that distinguish between verticals when users are re-finding. Then, we evaluate the predictive power of those features in identifying re-finding tasks from different verticals and general web search tasks. Additionally, user difficulty when re-finding within each vertical is studied, to determine if more effort is required for re-finding documents in some verticals than others.

This paper has several contributions including:

— identifying features effective in distinguishing re-finding tasks in different verticals;
— proposing a semi-automated way of vertical identification and minimal labeling exercise;
— predicting re-finding in verticals and distinguishing from general web search, which could consequently lead to adaptability of search results;
— exploring early predictions and different feature group performance in predicting re-finding in verticals;

— investigating user's difficulty in re-finding tasks in verticals, and reporting those verticals that might be more challenging in re-finding and therefore require improvements for the search experience of users.

Note that the purpose of this study was not to introduce new features for verticals; instead the aim was to identify behavioral differences across verticals. In this paper, the fundamental features and re-finding extraction were from our previous work and are used as a starting point. However, in this study the problem is different from our previous work, and requires a method for identifying verticals and multi-class classifications. Moreover, the features need to be customized for explorations in this work. For example we need to compute features to present the early stage of searches for early predictions.

The rest of this paper is organized as follows. In Section 2 we describe the past research for both re-finding and vertical prediction. Section 3 explains the datasets and experimental process of this study, in particular how we identify re-finding in verticals and collect ground truth data. Section 4 investigates some key features for the identification of re-finding behavior in verticals. Those features are used to construct predictive models described in Section 5. Section 6 discusses the real-time applicability of several vertical classification models. The last research question, regarding the interplay between vertical domain and difficulty, is addressed in Section 7. The paper concludes with a discussion and future work Section.

2. RELATED WORK

The following section outlines past studies related to this work under three main categories: re-finding identification, re-finding difficulty, and vertical selection.

2.1. Re-finding versus General Search Tasks

Re-finding information that the user has seen is defined as the dominant search task in a *personal search* context (e.g. as mentioned by Elsweiler and Ruthven [2007], Kim and Croft [2009], and Capra III [2006]). Personal search covers a range of areas including: desktop search for retrieving documents on personal computers (e.g. studies by Cutrell et al. [2006] and Dumais et al. [2003]) or messages from email box (e.g. studies by Elsweiler et al. [2011a,b]; Harvey and Elsweiler [2012]); and known-item search when the document to be searched has been seen before, even on public sources like the web (e.g. studies by Elsweiler and Ruthven [2007], Kim and Croft [2009]) or social media (e.g. studies by Meier and Elsweiler [2014]). Moreover, in the latter studies, personal search is highlighted as one of the key components for Personal Information Management (PIM) systems. PIM systems are focused on the methods that individuals use to manage their personal information including gathering, organizing, maintaining and retrieving on a daily basis [Bergman et al. 2008a,b, 2010; Jones 2007; Lansdale 1988; Rodden and Wood 2003; Teevan et al. 2006; Whittaker et al. 2006, 2010]; where personal search is concentrated on the retrieval part.

Note that the general concept of *task* is referred to as an atomic information need of the user. To contrast with re-finding, all other types of search tasks are referred to as *general search* in this study.

In terms of differences between re-finding and general search, some main points have been discussed by [Jones and Bruce 2007]. The differences come from the *prior experience* of the user in re-finding. This personal experience can bring different levels of knowledge in relation to the target information (*known item*), and different levels of expectation to satisfy the need (*exact match*). Jones and Bruce argued that relevant information can be recognized more easily in re-finding, since the target has been seen before. However, sometimes recognizing relevant information in general finding is easier; for example when users do not have the knowledge of the best existing match, and they will be satisfied with a result appeared earlier in top results [Jones and Bruce 2007]. Similar to the prior experience, the *prior frequency* has been highlighted in a study by Capra III [2006] in distinguishing re-finding tasks. In other words, re-finding is not only distinguishable from general finding by occurring in different search sessions, but also sometimes by the frequency of the task in the same session.

On the other hand, some points of similarity between re-finding and general search have been proposed, for example where both search needs should be addressed by the same tools [Jones and Bruce 2007]. In the third part of a book by Ruthven and Kelly [2013], re-finding has been defined as a condition that users apply on a general finding task, where a re-finding task of one user can be viewed as a general finding for another user. The same view has been highlighted by Capra III [2006] as well.

From discussed studies in this section, the typical retrieval request in the personal context has been mentioned as re-finding known information items. The prior experience to the information makes re-finding different from general search tasks. However, in some conditions re-finding can be seen the same as a general finding. In this work, we will explore how re-finding can be different or similar to general search tasks.

2.2. Identified Gaps in Re-finding Studies

Past research distinguished re-finding tasks from general web search mainly dependent on repeated occurrences of queries and clicks [Teevan et al. 2007]. However, it was shown that even after one hour, people mis-remember keywords around 30% of the time, and therefore might not be able to repeat the previous search tasks [Teevan et al. 2007]. On the other hand, repeating clicks on search results might be achieved at the end of search tasks, while users have been struggling from the beginning of the search. This requires tools for more accurate and possibly earlier identification of re-finding tasks.

There are studies where re-finding tasks have been distinguished based on some common underlying features of recorded tasks that are not content-based (such as the granularity of information to be re-found [Elsweiler and Ruthven 2007]). However, current examined features in the re-finding context are limited such as user's self-reported features (e.g. topic familiarity) [Capra III 2006],or they are related to a specific search context (e.g. email opening folder) [Elsweiler et al. 2011b].

One strong indication for differentiating tasks is the level of task difficulty [Liu et al. 2010, 2012]. In general web search, behavioural features in search engine use have been highlighted as the main indication for task difficulties in comparison to the other features such as topic or search experience [Liu et al. 2012]. In investigations of search engine use, other small scale user studies found it difficult to distinguish re-finding tasks from general web searches [Capra III 2006]. Larger-scale query log search features were limited to the level of equal queries and clicks to identify re-finding [Teevan et al. 2007]. Supervised learning is another technique that was used to identify re-finding; however, it is limited to the occurrence of equal queries, and difficulty behavioural features in re-finding tasks have not been studied [Teevan et al. 2007; Tyler and Teevan 2010].

Moreover, there has been research in the general search context, where user's search task and difficulty have been studied based on the underlying topic domain such as image retrieval (e.g. studies by Goodrum and Spink [2001], and Diaz [2009]). However, there is no such a distinction in the context of re-finding tasks.

In this research, we examine a broader range of behavioural features in a) distinguishing re-finding tasks from general web search, b) examining different levels of re-finding difficulty, and c) identifying various search behvaiour and difficulties across re-finding vertical documents. The following sections will highlight the gap in past research for the introduced problems.

2.3. Re-finding Identification

In one of the first studies on web-based re-finding, Teevan et al. [2007] 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 [2010] studied re-finding at the level of sessions, finding that queries change more across sessions than within. Later, Tyler et al. [2010] examined query features and the rank of the clicks to identify re-finding.

Capra III [2006], 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 [2007], 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 [Kotov et al. 2011; Wang et al. 2013b]. In a study by Kotov et al. [2011], 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 behavioural features that have been studied 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 would be useful to examine a broader range of features, specifically early in search. A recent work on identifying re-finding tasks could enable more accurate task distinctions incorporating more behavioral features [Sadeghi et al. 2015]. However, as yet, no distinctions were made between re-finding tasks in verticals and in general search. The underlying problem in this work is different from our previous work, as we need to identify the behavioural differences across different verticals and in different search stages, which introduces a multi-class problem with customizing behavioural features early at search.

2.4. Re-finding Difficulty

Re-finding difficulty has also been studied in past research: Capra III [2006] explored a set of features including the number of search URLs, task completion time, and the elapsed time between search tasks, to identify difficulty. He distinguished easy and difficult tasks based on a set of specific topics (e.g. finding yellow pages vs. flight information). The main indicators of re-finding difficulty was task frequency, topic familiarity, and determining that target information had been moved from the web page where it was originally found.

Information being relocated on the web, as well as changes in target document rank position, were highlighted as a cause of re-finding difficulty by Teevan [2004, 2006]. In observations reported by the researcher, changes in the path to reach the target information was a stronger indicator of user difficulty than temporal features such as elapsed time. Elsweiler and Ruthven [2007] studied the difficulty of re-finding in terms of two features: the granularity of information to be re-found; and also the elapsed time between re-finding. There were no significant differences in terms of the granularity feature for task difficulties. However, it appears that longer time gaps could indicate that users were having difficulties for some types of re-finding tasks.

Although some features were examined for identifying difficulties in re-finding, they are mainly limited to user's self-assessed features (e.g. topic familiarity) or target information (moved web page, or granularity of information) or related to a specific search context (e.g. email [Elsweiler et al. 2011a] or social network [Meier and Elsweiler 2014]). In general web search, large scale query log features have been extensively used to predict search difficulty [Liu et al. 2010, 2012], as well as user frustration, dissatisfaction, or success/failure [Ageev et al. 2011; Hassan et al. 2010, 2013, 2011]. Features ranged from temporal to user behavioural, 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. Moreover, for different stages at search (e.g. initial, middle, and end points) different features indicative of difficulty have been investigated in the general search context (e.g. a study by Liu et al. [2014]). These features can be developed for the re-finding context, and further incorporated into constructing predictive models, where search engines could adapt search results based on underlying search tasks. In recent work, Sadeghi et al. [2015] examined the effect of other behavioral features taken from general web search difficulty predictions. Although they obtained some level of accuracy to predict whether the user is struggling, yet there is a gap in the link between difficulties in re-finding and verticals.

2.5. Vertical Selection

The search tasks across different verticals have been studied in the general search context (e.g. studies by Goodrum and Spink [2001], and Diaz [2009]). In identifying relevant verticals corresponding to user queries, previous studies have focused on machine learning approaches [Arguello et al. 2009],[Arguello et al. 2009]. In general those approaches outperform traditional methods employed for resource selection in distributed information retrieval research [Arguello et al. 2009]. A machine learning model trained on a labeled dataset is used to predict verticals based on a range of features.

We categorise the features that have been considered so far under two major groups: content-based [Arguello et al. 2009; Hong and Si 2013], and related to users' search behavior [Diaz and Arguello 2009; Richardson et al. 2007]. The latter group allows classification to be generalized beyond the level of queries. Some other work has examined the portability of existing predictions from available datasets for verticals to which no training set is available [Arguello et al. 2010]. However, in the context of re-finding, we do not know of any research on predicting different verticals.

3. DATA AND METHODS

To study re-finding behavior in verticals, we analyze searcher behavior in query logs. In this section we discuss the datasets and experimental process.

3.1. Datasets

The analysis in this paper uses a sample of a query log from two months in June and October 2012, gathered from the Yahoo! search engine. The dataset included interactions of 7,380,610 unique users, described by an anonymous user id and a timestamp of when the user started searching. Logged events included submitted queries, the URL & rank position of clicked search results, and a timestamp for each event. The terms of service and privacy policies of Yahoo! were strictly followed.

Different log segmentations have been proposed in past research including sessions, goals, and missions as defined by Jones and Klinkner [2008]:

- A *session* is identified based on a fixed timeout in user activity.
- A *goal* is composed of a group of related queries and corresponding clicks submitted by a user.
- A *mission* segmentation includes related but multiple information needs.

We focus on *goal* segmentations as the representative of a task in this work, since a goal is related to an atomic search need; while missions identify multiple information needs. Previous work has also shown that goals are more accurate than session timeouts for identifying task boundaries.

To extract task boundaries for goal segmentations, Jones and Klinkner developed classifiers to identify the relatedness between queries to be considered under the same task. The developed classifiers are based on four types of features as follows:

- **Temporal features:** Examples of temporal features are inter-query time, and whether the queries are sequential in time. It was shown that the temporal features might not be effective in detecting task boundaries by themselves; however, they are helpful jointly with other features.
- **Edit-distance features:** The main idea for edit-distance features is that common words between queries could increase the chance of query relatedness for a task. This type of features is in two levels of character-edit distance for spell corrections and common stems (e.g. number of characters in common starting from the left), and word-level features (e.g. number of words in common).
- **Query log features:** The type of query log features is for identifying semantic relationships between queries, particularly useful when there is no syntactic commonalities. The researchers used the log-likelihood ratio scores to identify pairs of queries which happen together not by a chance.
- **Web search features:** This category of features measures the relatedness between queries based on the commonalities between the terms of their corresponding search results.

The goal identification approach discussed above has recently been used for a task discovery with an accuracy of 92% [Lucchese et al. 2013]. We used the same approach for identifying goals in this work.

Sequential goals from a user:
$\{G1, G2, ..., Gn\}$ n: Number of goals extracted for a user.

Examples of *paired* goals
$< G1, G2 >, < G1, G3 >, ..., < G1, Gn >,$
$< G2, G3 >, ..., < G2, Gn >, ..., < Gn - 1, Gn >$

Fig. 2. Pairing search goals from one user to identify potential re-finding goals. G: a search goal generated by the user.

As a basic constraint, re-finding happens over time for each user. Therefore, we ordered all goals from the same user by their timestamp, and all possible goals were *paired*, as shown in Figure 2. However, we did not consider paired goals that occurred less than thirty minutes apart, since we were not interested in short-term re-finding. Note that we did not apply this time constraint for the identification of a goal, since there are related queries belonging to a task that are interleaved with queries from the other tasks in a short time interval. First we identified the possibly interleaved tasks through the goal classification proposed by Jones and Klinkner. We then applied the time constraint of the short-term re-finding removal on the output of the goal classification.

This pair-wise approach has been previously used to identify potential re-finding goals [Sadeghi et al. 2015]. The first goal in a pair is referred to as the *original*, and the second as the *potential re-finding* goal. Next we discuss how to identify whether the latter goal indicates re-finding or not.

3.2. Method for Identifying Re-finding Goals in Verticals

The proposed method of this study for identifying re-finding goals in verticals employed supervised machine learning models, where labeled training sets are required. In order to generate a labeled training set, we needed two types of information: first, was the potential re-finding goal representative of a re-finding task or not; and second, which vertical should be associated with the re-finding target document. To generate sufficient labeled data, and in order to optimize manual assessment efforts, we employed signals of re-finding and vertical selections that can be applied automatically to our datasets.

Regarding the automatic identification of re-finding, a final last click on a common URL in two paired goals (referred to as *exact last clicks*) has been identified in previous work as a strong feature of re-finding activity. Specifically, on average, 94.1% of paired goals with the exact last clicks feature were identified as representing re-finding, and the item clicked at the end of the goal was the target of the re-finding [Sadeghi et al. 2014]. The illustration in Figure 1 is an example of a paired goal with exact last clicks (answers.yahoo.com/question/index?qid=200830AAF0).

When labelling, assessors were presented with the paired goals and were asked *"Do you think that in the second search the user is re-finding document(s) that were found in the first search?"* (Possible responses were "yes", "no", "not sure".) "Re-finding" was defined as repeat searching

Table I. Examples of detailed guidelines as seen by assessors for identifying re-finding.

Hint Title	Description
Exact repetitions in queries/clicks	— By comparing queries in the original search task with queries in the potential re-finding task, if the exact queries have been repeated, it could be a signal of re-finding. — The same comparison for the repetitions in clicks is applicable. If the intermediate clicks in the paired search tasks have been repeated, it could be an indication of re-finding.
Common or semantically related query terms	— By comparing queries in the original search task with queries in the potential re-finding task, if they are not exactly the same, but semantically related, it could be a signal of re-finding.
Common terms between queries and clicks	— Commonalities between the terms of queries of the potential re-finding and the terms in the clicks of the original search could be a signal of re-finding. (or vice versa: commonalities in the terms of the original queries with the terms in the clicks of the re-finding)

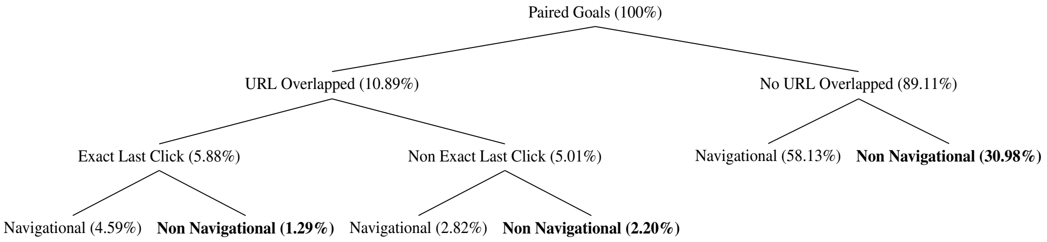

Fig. 3. The percentage of paired goals where the second goal consists of a single query and a single click (i.e. Navigational), or more than one query and one click (i.e. Non Navigational).

for a document that was previously found, and therefore, if the user seems to be interested in any other documents in the second goal, the search would be labeled as “Not Re-finding” regardless of whether it is a related sub-task to the original goal or not. Twelve assessors were recruited from RMIT university to participate in labelling query logs. They were post-graduate students from different disciplines in the school of Computer Science. Two levels of guidelines were provided in labelling experiments. In the initial level of guidelines, the only hints given to assessors were a set of examples of re-finding and not re-finding tasks. In the detailed level of guidelines, in addition to the examples provided in the initial guidelines, more detailed instructions were provided to the assessors. The examples of detailed instructions regarding the identification of re-finding are illustrated in Table I, where some hints are provided in terms of the content of queries and click URLs. Further details in regards with the labelling experiments are available from our previous work [Sadeghi et al. 2014].

In order to determine which vertical (topic domain) the commonly clicked item is associated with, the last item clicked was matched to web sites that would be searched by a vertical search engine. The web analytics site Alexa provides a categorisation of web pages into vertical categories. We used the top 50 websites for the major verticals in our analysis.[1] The rank of the websites in Alexa is calculated based on the number of daily users and pageviews over the past month. Note that although this vertical selection approach is focused on top websites from Alexa, it includes verticals from the specific search engine that we study even broader in scope. An in-depth analysis indicated that the vertical categories of “movie”, “news”, “reference”, and “image” were the most frequent in our dataset. Examples of these verticals in our dataset are illustrated in Table II. As the incidence of clicked pages from other verticals were low in number, we did not consider them in this study.

We manually verified the vertical selection process. A set of 200 paired goals (50 items per vertical) with more than one query and click were randomly selected and manually verified by a judge. The percentage of agreements between manual labels and identified verticals from Alexa was around 84%. This is in line with levels of inter-rater agreement for the labeling of paired goals in past research [Sadeghi et al. 2015], giving us high confidence in the approach.

3.3. Ground-truth Dataset

Using the vertical selection process, we generated a dataset consisting of instances from each vertical category labeled as *image re-finding*, *reference re-finding* (e.g. Wikipedia, Yellowpages, etc), *movie re-finding*, and *news re-finding*. An analysis of the vertical dataset showed that 82.3% of the paired goals consist of only one query and one click. This indicates that they are most likely to be navigational queries. As one focus of this work is on detecting re-finding difficulties, we require goals with a higher number of interactions. Therefore, we removed navigational goals from our dataset using a set of rules taken from previous work [Sadeghi et al. 2014]. For example, paired

[1] http://www.alexa.com/topsites/category categories accessed on 25/02/2014.

Table II. Examples of 10 random out of top 50 websites known for verticals.

Image
myfreecams.com, photobucket.com, mashable.com, tineye.com, istockphoto.com,everystockphoto.com, photo.net, images.yahoo.com, gettyimages.com, freeimages.com
Reference
whitepages.com, stackoverflow.com,wikipedia.org, urbandictionary.com, yellowpages.com, answers.yahoo.com, thefreedictionary.com, thesaurus.com,wiki.answers.com, wordreference.com
Movie
fandango.com, movieweb.com, movies.yahoo.com, comingsoon.net, topdocumentaryfilms.com, moviefone.com, rottentomatoes.com, filmaffinity.com, boxofficemojo.com, nextmovie.com
News
theguardian.com, news.com.au, news.yahoo.com, foxnews.com, nytimes.com, cnn.com, bbc.co.uk/news/, news.google.com/, nbcnews.com, washingtonpost.com

goals, where the re-finding goal includes queries with top domain names (e.g. “Youtube.com”) were removed. Examples of filtering rules are illustrated in Table III.

Among the remaining non-navigational paired goals across the four vertical groups, the “image” vertical has the minimum size, consisting of 297 pairs. To create a balanced dataset, this same number of pairs was randomly selected from the non-navigational instances of the other vertical categories.

We also require *not re-finding* instances to determine differences between re-finding in verticals and general web search tasks. In previous work [Sadeghi et al. 2015], examples of paired goals that were not necessarily ended with exact last clicks and were not re-finding were detected by a human assessor. We used the same approach and identified examples of general search tasks, and a random sample of 297 *not re-finding* instances were added into our dataset. This dataset, with a total size of 1,485 (5 × 297) paired goals, is used for constructing predictive models in this study. Note that we did not select the *not re-finding* examples per each vertical, as in this study they are used for comparing a re-finding task against a general search, and not comparing these two contexts given a particular vertical. In general search, it happens that the user does not target any specific vertical domain in advance; whereas, in re-finding it is more likely that the user is looking for a specific document in a specific vertical type from the beginning of the search. Studying general searches in specific verticals is not the purpose of this study and therefore, they are not considered in our dataset.

3.3.1. Limitations. Note that working with a relatively small dataset does not necessarily reflect that the re-finding problem we study is small; rather, applying our set of filtering rules gives us a

Table III. Filtering rules for identifying challenging re-finding goals.

Filtering Rule	Description	Example
Top domain signals in query	Excluding paired goals including top domain names in their queries in the re-finding goal (Top domains were identified through top 50 ranked websites from Alexa.com.)	“Youtube”, “Facebook”, ...
Navigational signals in query	Excluding paired goals with queries containing signals of URL addresses in the re-finding goal	“www”, “.com”, “.aero”, ...
Navigational signals in clicks	Excluding paired goals with clicks containing “login” or “signup” in the re-finding goal	https://login.yahoo.com/
Navigational signals in query and domain names	Excluding paired goals where in the re-finding goal there is a query equal to the domain name, or the domain name is the merge of words in query, or the domain name is the corrected spell of the query	query: “banana” with the click of “banana.com”, query: “bank net”, click: “netbank.com”, query: “youtibe” with the click of “youtube.com”

dataset where we will find a concentration of re-finding problems. Although focusing on the exact last click pairs is limited in the number of likely non-navigational goals, it enabled us to identify re-finding across verticals with reasonable accuracy while minimizing manual labeling efforts.

From Figure 3, we see there are other types of re-finding where overlaps between clicked URLs exist (i.e. *URL Overlapped*), but not necessarily in the last click of the goal (i.e. *Non Exact Last Click*). Moreover, in a re-finding identification study by Sadeghi et al. [2015], the authors mentioned that there are re-finding types with no overlapping in the clicked URLs of paired goals, which are referred to as *No URL Overlapped* paired goals, 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 when the user has failed to reach the same target document, thus having the same task but not resulting on overlapping URLs"*. While these cases might be more likely to include non-navigational re-finding with more number of interactions, identifying such cases is challenging from a query log study [Sadeghi et al. 2015], and it is left for future work.

Moreover, for the identification of verticals, we focused on the vertical domain list from the search engine and top known domains from Alexa, where paired goals with exact last clicks were classified into different verticals matching the domain of the last click with the vertical list.

Although query log approach has been extensively used in different IR studies on search behavior, the particular design of search interface might influence on the user behavior and consequent findings. As future work, we plan to study the impact of search interface on user behavior in re-finding vertical documents and identify behavioral patterns across different search engines.

4. FEATURE VARIATIONS IN VERTICALS

In this section we discuss features for identifying re-finding behavior in verticals and investigate their variability across the different vertical categories on the ground truth dataset.

4.1. Features

To distinguish re-finding behavior within vertical domains and also from general search, a set of 124 features were examined from a study on predicting re-finding tasks and search difficulty by Sadeghi et al. [2015]. The authors have explained the complete set of features, ranging from those that are related to the content of queries and documents, to search behavioral features. We turn to behavioral search features because the type of content and its richness would vary across verticals. These type of features are also potentially reflective of search difficulties. The features are calculated from the paired goals in our datasets, and we used the same terminology as proposed in Sadeghi et al.'s work; the first goal is referred to as the *original* goal, and the second is the *re-finding* goal. Some features are specific to the context of re-finding (e.g. *"days between paired goals"*), some are more representative of a general search behavior (e.g. *"goal length in number of queries"*), and others could be potential indicators of search difficulties (e.g. *"fraction of queries for which no click"*).

In terms of computation, some features only require information from the re-finding goal (e.g. *"re-finding goal length in number of all queries"*), some are calculated based only on the original goal (e.g. *"original number of engaged clicks"*, i.e. the number of clicks with dwell time greater than 30 seconds), and others require access to both goals (e.g. *"re-finding mean dwell time of common clicks"*, which requires the identification of common clicks between paired goals). The first category of features is referred to as *history-independent*, as they do not need information from the previous searches of the user and can be computed from the re-finding goal. However, the latter features that require access to the original goal, the previous search of the user, are named as *history-dependent* features. Note that these two feature categories have no intersection, and they are used in the next sections for different analyses considering the availability of past search information.

We also consider re-finding without knowledge of the original user goal, as past research has mentioned that re-finding could also occur without an original search: often users employ a search engine to find something they saw in another context (item found while browsing on the web or shown to them by a colleague or on social media) [Sadeghi et al. 2015].

Table IV. Features that significantly distinguish one vertical from the other three. Based on a Kruskal-Wallis test followed by Kruskalmc: Multiple comparison test after Kruskal-Wallis ($p - value < 0.05$).

image from movie, news, & reference
re-finding min inter-click time
original effective search time
original total dwell time after queries
movie from image, news, & reference
re-finding goal unique clicks count
original mean query length of all clicks
original mean query length of common clicks
news from image, movie, & reference
equal query elapsed time

4.2. Feature Performance

We use statistical analysis to distinguish between verticals. As the data may not meet the normality assumptions of ANOVA, we used the non-parametric Kruskal-Wallis and Kruskalmc as Multiple comparison test after Kruskal-Wallis [Siegel and Castellan 1988]. Features that significantly ($p < 0.05$) distinguish one vertical from the other three are listed in Table IV.

The distinctive features of the "image" vertical are more time-based. For example, the minimum time between clicks (*"re-finding min inter-click time"*) was a strong feature in image re-finding, as was (*"original effective search time"*), which measures dwell times in the original goal. Also, the total time spent between issuing queries and other search events (e.g. clicks or reformulated queries) was distinctive for the original goal of an image re-finding (*"original total dwell time after queries"*). The significance and higher average of the time-based features, suggest that users took longer to locate this type of information than the other verticals. This suggests the requirement of improving search result summaries in image search. The difficulty of image searching was also mentioned by Tseng [2012] in comparison to video type of documents. Although time-based features are highlighted for image re-finding, past research in non re-finding context showed that query-based features are distinctive for image searches [Goodrum and Spink 2001]. As an example, in a study by Jansen et al. [2000], image queries were longer in comparison to searches for the video type of documents, which was not distinctive in the re-finding context.

In "movie" re-finding, the number of unique (not repeated) clicks in the re-finding goals (*"re-finding goal unique clicks count"*) was lower than other verticals, suggesting that users are more successful in re-finding movie documents in comparison to other verticals; however, in non re-finding context, it appears that the number of clicks in targeting the video type of documents is higher in comparison to the news documents [Sushmita et al. 2010], which is not the same as the re-finding context. Moreover, the length of queries in the original goal of movie re-finding (*"original mean query length of all/common clicks"*) was longer than other verticals. This is in contrast with past research, where the queries for the video type of documents were shorter in comparison to the image searches, which can be due to long names associated with entities in movies.

The distinctive features of "news" re-finding was a longer time gap between the original and re-finding goal (*"equal query elapsed time"*). This suggests that users wish to re-find this item over longer time period than information covered by other verticals.

There are other features that are not uniquely distinctive. For example, for re-finding "reference" documents, there are features distinctive from "images" and "news", but not "movie". The examples of these features include *"re-finding mean time to the first query clicks"*, *"re-finding number of clicks per query"*, and *"fraction of queries with no click"*. Another feature that distinguishes the "news" from "movie" and "reference" is *"the number of engaged clicks"*, counting clicks with dwell time greater than 30 seconds. However, considering the total number of clicks, this number is lower for the news documents in comparison to the video type of documents in non re-finding context [Sushmita et al. 2010].

There are features that were significantly distinctive for more than one vertical. For example, the mean value of the reciprocal rank of the common clicks were distinctive in re-finding both “reference” and “movie” documents in comparison to all other verticals, which could be representative of particular system performance for these two verticals. Although there might be similarity between verticals, in this work the primary focus is on identifying distinctive behavior in single verticals.

In addition to such distinguishing features there could be other effects from interactions between features. Moreover, we have not yet considered how accurate the features are for predicting re-finding in different verticals. To address these questions, we build a set of predictive models from these features in the next section.

5. PREDICTION MODELS OF RE-FINDING IN VERTICALS

In this section we build classification models to assess whether the vertical to which a re-finding is targeting can be predicted and differentiated from general web search. We also discuss the performance of classifications in this section.

Distinguishing multiple verticals can be cast as a multi-class classification problem. We build multiple binary classifications for each vertical as suggested in the past work (one-versus-all method) [Diaz and Arguello 2009]. We used Support Vector Machines as our classification model, trained with the Sequential Minimal Optimization (SMO) algorithm with a default poly kernel setting, as this algorithm has been shown to work well in broadly similar classification scenarios [Teevan et al. 2007]. Five binary classifiers were trained using the dataset described in Section 3.3. We applied a set of mappings for generating corresponding binary training sets. As an example, for *image re-finding* predictions, all “image” labels in training sets were mapped as “yes”, and the labels for other verticals and *not re-finding* were mapped as “no”. The same approach was taken for generating binary training sets for other verticals. In addition to the verticals we build a training set for *generic re-finding*, where re-finding goals would be predicted regardless of the type of vertical. For creating a corresponding binary training set for generic predictions, we mapped all “not re-finding” labels to “no”, and the rest of vertical labels were mapped to “yes”, as they are all re-finding goals. In this section two different prediction problems were studied using available training sets. First differentiating re-finding in a particular vertical from other verticals and general searches, where the positive instances in training sets are re-finding searches in that specific vertical and negative instances come from re-finding in other verticals and also not re-finding searches. Second, with the aim of identifying the differences between re-finding and general searches, for the positive instances we include all the re-finding instances regardless of the type of the vertical and negative instances are from not re-finding searches.

We report the F-measure as a performance metric, calculated as an overall weighted average of F-measure scores per binary class. We also report precision and recall scores. We employed a 10 times 10-fold cross-validation approach, which repeats 10-fold cross- validation and measures the average of the obtained results, as recommended for comparing classification models [Nadeau and Bengio 2003]. We used a paired two-tailed t-test to test for statistically significant differences in effectiveness. Unless otherwise specified, p-values below the 0.05 level are interpreted as being statistically significant. Using the training set and features described in Sections 3.3 and 4.1, we constructed different classification models for predicting re-finding in verticals. We first report results from predictions using all features. Second, we analyse the effects of using different feature sets (both dependent and independent of original goals) on the prediction performance.

5.1. Overall Predictions

Using all features discussed in Section 4.1, we constructed five binary classification models with binary class labels for each model. We also replicated a model proposed by Teevan et al. [2007] as a state of the art baseline, which used features limited to the level of equal queries (e.g. elapsed time of occurring equal queries between original and re-finding, the length of equal queries, and the number of common clicks of equal queries between original and re-finding pairs). Note that the baseline classification model does not consider features at the level of goals, and only predicts

Table V. Accuracy of classifications for re-finding in verticals using P: Precision, R: Recall, and F: F-measure. Baseline is the re-finding classification proposed by Teevan et al. [Teevan et al. 2007]. Scores are reported in percentages.

	Image Re-finding	Reference Re-finding	Movie Re-finding	News Re-finding	Generic Re-finding	Baseline
Predictions	P: 86.6 R: 87.3 F: 86.9	P: 89.7 R: 89.3 F: 89.5	P: 84.6 R: 85.1 F: 84.8	P: 80.7 R: 82.6 F: 81.6	P: 97.5 R: 97.5 F: 97.5	P: 92.0 R: 92.2 F: 92.1

Table VI. Accuracy of classifications for re-finding in verticals using P: Precision, R: Recall, and F: F-measure. The first row of the result shows overall scores per binary class for each classifier based on only history-dependent features; whereas the second row of the results are based on history-independent features. Baseline is the re-finding classification proposed by Teevan et al. [Teevan et al. 2007]. Scores are reported in percentages.

	Image Re-finding	Reference Re-finding	Movie Re-finding	News Re-finding	Generic Re-finding	Baseline
History-dependent Predictions	P: 85.7 R: 86.5 F: 86.1	P: 89.5 R: 89.0 F: 89.2	P: 83.8 R: 83.8 F: 83.8	P: 75.5 R: 80.2 F: 77.8	P: 97.6 R: 97.6 F: 97.6	P: 92.0 R: 92.2 F: 92.1
History-independent Predictions	P: 82.1 R: 81.8 F: 81.9	P: 74.2 R: 79.9 F: 76.9	P: 71.4 R: 79.8 F: 75.4	P: 72.1 R: 79.9 F: 75.8	P: 83.5 R: 84.8 F: 84.1	not supported

re-finding when the user can repeat their queries at the re-finding time. However, we incorporated information about search goals, which is broader than the query level.

The results for classification models are reported in Table V. As the baseline classifies whether equal queries could lead to repetition in clicks regardless of the type of the vertical, we compared our "generic re-finding" classification against the baseline, where we could obtain a relative significant improvement of 5.9% in terms of F-score. The more accurate re-finding prediction in comparison to the state of the art can be due to the broader feature sets (at the level of goals), which we considered in our prediction models. In further analysis, we compared the performance of vertical predictions against the more accurate classification model from the "generic re-finding".

We investigated the prediction results from two viewpoints. First, we compared the prediction performance of re-finding in each vertical with the "generic re-finding" group. It can be seen that "generic re-finding" obtained the highest accuracy with an F-score of 97.5. Although it is reasonable to expect that providing more fine grained re-finding distinctions (detecting a corresponding vertical) would be more challenging than distinguishing re-finding from a general search task, we were interested in the effect size. Comparing the overall performance of each vertical classifier with the generic group in Table V, "reference" most closely approaches the performance of "generic re-finding" with an F-score of 89.5, while "news" is most different (F-score: 81.6). This indicates that search behavior in re-finding "reference" documents is more distinctive than other verticals, relative to the generic group. Although distinctive features for re-finding "reference" documents against all other verticals were not identified in Section 4.2, here "reference" prediction models obtained the highest accuracy among verticals, which could be due to the interactions between features that were not considered in Section 4.2.

As a second viewpoint, we investigated significant differences in re-finding predictions only within verticals, which are the first four columns in Table V. For pair-wise comparisons within vertical predictions, two-tailed t-tests were carried out on 10 times 10-fold cross-validation runs, as recommended in past research [Nadeau and Bengio 2003]. This showed statistically significant differences between each pair of the four verticals ($p < 0.05$). These results suggest that re-finding goals are indeed distinguishable across verticals.

5.2. History-dependent vs. Independent Predictions

Information from the original goals could make predictions easier for identifying re-finding in verticals, particularly when it can be directly established that the user is repeating previously submitted

Table VII. Accuracy of classifications for re-finding in verticals using P: Precision, R: Recall, and F: F-measure. The first row of the result shows overall scores per binary class for each classifier based on only query-based features; whereas the second row of the results are based on click-based features. Scores are reported in percentages.

	Image Re-finding	Reference Re-finding	Movie Re-finding	News Re-finding	Generic Re-finding
Query-based Predictions	P: 71.6 R: 79.9 F: 75.5	P: 64.0 R: 79.9 F: 71.1	P: 72.5 R: 79.6 F: 75.9	P: 64.0 R: 79.8 F: 71.0	P: 93.2 R: 93.2 F: 93.2
Click-based Predictions	P: 86.3 R: 87.0 F: 86.7	P: 89.7 R: 88.9 F: 89.3	P: 85.7 R: 83.4 F: 84.5	P: 76.9 R: 80.5 F: 78.7	P: 97.3 R: 97.3 F: 97.3

queries and clicks. However, this information might not be available for search engines at re-finding time, for example because of the user not being logged into the search engine, or due to having originally found the item by browsing rather than searching [Sadeghi et al. 2015]. In this section we investigate the accuracy of the vertical predictions with and without accessing information in the original of the paired goals. As explained in Section 4.1, a subset of features that require accessing original goals are referred to as *history-dependent features*, while features that can be computed based only on the re-finding part of paired goals are *history-independent*. Note that these two sets of features are separated and there is no intersection between them. The performance of classifications, categorised by these feature groups, is shown in Table VI.

Given that the baseline is limited to equal query features (see Section 5.1), and does not support history-independent features, we compare the performance of vertical predictions against the “generic re-finding” group. The lack of access to the original goals (the row labeled *History-independent* in Table VI) decreases the overall accuracy of the classifiers, either in comparison to history- dependent predictions, or predictions using all features (Table V). It can also be seen that despite the overall reduction in effectiveness, identifying “image re-finding” is less dependent on the original goals with an F-measure score of 81.9 compared to the other verticals, as it suffered the least amount of reduction. On the other hand, “reference re-finding” appeared to be more substantially affected by excluding history-dependent features with an F-score of 76.9. The performance of “movie” and “news” predictions independent of search history were similar with an accuracy of 75.4 and 75.8 respectively, whereas they were different using all features in Table V (84.8 vs. 81.6). This suggests that distinctive features for these two verticals are mainly from the original goals; examples of theses features were discussed in Section 4.2.

5.3. Query vs. Click-based Predictions

As another exploration on feature groups, we compared the performance of query versus click-based features for vertical predictions in Table VII. Examples of query-based features are: ‘inter-query time’, ‘queries per second’, and click-based features include ‘inter-click time’ and ‘time to the first click’. Features that are dependent on both queries and clicks, such as ‘effective search time’ were not considered in this exploration. This way of grouping features, which was motivated by general web search studies (e.g. a study by Hassan et al. [2013]), is particularly useful for predicting re-finding in verticals for non-clicking users.

It seems that in vertical predictions click-based features are more effective than query-based features, as the average F-score for only click-based features across verticals is 84.8% ,whereas for query-based features is 73.4% from Table VII. In comparing this table with the prediction results in using all features in Table V, the performance of click-based features almost approaches the overall performance using all features. This may suggest that it is more challenging to identify re-finding in verticals using only query-based indications of user behavior. Although this task of vertical identification using query-based features is challenging, it is more predictive for the ‘movie’ vertical in comparison to others. This may indicate more distinctive query behavior in movie searches in comparison to other verticals.

Table VIII. Top five highest and lowest ranked features by an SVM classifier.

Highest Ranked Features	Lowest Ranked Features
common click in relation to last click	re-finding advanced query syntax
original next page ranked clicks count	original ended with query
re-finding min goal position of common clicks	both goals ended with query
re-finding last click rank	original advanced query syntax
original last click rank	same rank common clicks

5.4. Feature Importance Analysis

To better illustrate the importance of underlying features in the predictions of verticals we employ the algorithm proposed by Guyon et al. [2002], which has been used for feature selections in multi-class problems. This algorithm ranks the importance of features by the square of the weight assigned by an SVM classifier, where features are ranked for each class using a one-versus-all method, and then from the top features of each class, a final ranking is generated.

Table VIII shows the top 5 highest and lowest ranked features. These features are different in comparison to the features discussed in Section 4.2 using Kruskal Wallis test, which could be due to considering the interactions between features in the feature selection algorithm. The top feature of *"common click in relation to last click"*, which indicates whether there is a common click repeated at the end of either original or re-finding goal, is particularly important in distinguishing re-finding from general search tasks. The number of clicks ranked beyond the first page in the original search (*"original next page ranked clicks count"*) is important in predictions across verticals. In previous work, position in the goal is defined in terms of the number of queries and clicks from the beginning of the search [Sadeghi et al. 2015]; in our experiments, the earliest position in the re-finding goal where a common click with the original goal occurs (i.e. *"re-finding min goal position of common clicks"*) is also important in predicting verticals. The rank of the last click (in either re-finding or the original goal) is another highly predictive feature across verticals.

Among the features with the lowest ranks, existing advanced syntax such as quotes in queries (*"re-finding/original advanced query syntax"*), or ending search tasks with query (*"original/both goals ended with query"*), or whether the common clicks between paired goals are in the same rank (*"same rank common clicks"*) are not as important as other features in predictions across verticals.

Overall, in building our predictive models, there are important features, which are specific for re-finding context and also essential for the prediction of verticals. These specific re-finding features make our models different from existing vertical predictions in the general web context.

To further investigate the importance of these features, we examined the performance of our prediction models by removing the features that are specific for re-finding, and also related to commonalities between original and re-finding goals particularly in terms of queries and clicks. In other words, if a feature can be computed on either re-finding or original goal in an isolated way, then the feature would remain in the feature set. The removed subset of features (as shown in Table IX) is indicative of differences between the re-finding and general vertical context; in the latter, the commonality between searches by the same user would not be applicable. This feature removal resulted in a significant decrease in the performance of our models, by 7.1%, 19.0%, 39.4%, and 42.9% for image, news, movie, and reference verticals respectively. This suggests that the re-finding-specific features are essential for the prediction of verticals in the re-finding context, while features that can be computed independently of search commonalities are not adequate for vertical re-finding prediction.

Moreover, despite the focus of past research on content and query-based features for predicting verticals [Arguello et al. 2009], it appears that click-based features can also be important in the prediction of verticals. As an example, for the prediction of image searches, we assumed that the query length might be effective in the performance of the model, as it has been suggested as a distinctive feature for image searches (e.g. [Jansen et al. 2000]); however, building image re-finding models on the query length feature was 3 times less effective than a model based on the position of common clicks between re-finding and original searches. This may suggest that for re-finding

Table IX. A subset of re-finding specific features related to the commonalities between the re-finding and original goals.

equal query class
equal query elapsed time
equal query length
equal query no. of common clicks
query overlap/ URL overlap
no. of common clicks
common click in relation to the last click
same rank common clicks
re-finding mean query length of common clicks
re-finding mean dwell time common clicks
re-finding mean relative dwell time common clicks
re-finding mean query click counts of common clicks
re-finding mean reciprocal rank of common clicks
re-finding next page ranked counts in common clicks
re-finding no. of repeated common URLs
re-finding total time to reach to the first common click
re-finding first common click rank
re-finding mean relative common clicks goal position label
re-finding min goal position of common clicks
re-finding max goal position of common clicks
re-finding mean relative goal position of common clicks

context, in addition to the search content, the contextual and behavioral features can also be essential in predicting verticals.

5.5. Prediction Error Analysis

In this section, we perform an error analysis in order to determine the reasons for classification errors in the different classifiers. Using an SVM classifier and 10-fold cross-validation test, we evaluated a multi-class classification model on our training set including all five class labels, to examine which instances are incorrectly identified and which wrong labels were associated. In total, there were 20 categories of errors (5×4), where each of the five class label can be misclassified by the other four labels. The most frequent errors with 12.8% of incorrectly identified instances were for “news” re-finding, which was mostly mis-classified with “movie” re-finding. We randomly sampled 20% of mis-classified instances from each category of the errors to study the underlying reason for such failures. Although for some errors it was difficult to detect underlying causes, we were able to identify two common types of errors due to misleading behavioral indications, and limitations in the labeling of vertical selections. These occur in 34.4% and 3.1% of the sampled instances, respectively.

5.5.1. Misleading Behavioral Indications. One type of error is because of the similar behavior of users in re-finding a vertical document to another type in terms of an indication that is distinctive for that vertical. For example, the rank of the last click in the original goal is an important feature, particularly in predicting “image” re-finding it appears that it is likely for the user to click on the suggestion results from the search engine. There were cases where the user is looking for another type of document (e.g. movie) but with a similar behavioral pattern, which seems to result in mis-classification (see Figure 4).

5.5.2. Limitations in Vertical Selection. The other type of errors that occur for a particular type of search task are due to the limitations in the current vertical selection approach, which is based on the domain type of the last click in search (as discussed in Section 3.2). There are types of search tasks where the user might need to re-find a particular type of document (e.g. a movie) through a news website as illustrated in Figure 5. Here due to our vertical identification approach, the “news” label is considered as the true label, whereas the underlying need of the user is a document in the “movie” type. Due to the limitations of the underlying training set, this is reported as a mis-classification. In future work, we plan to explore more complex labeling schemes that allow items to be in more than one category.

Original Goal
Q: harry potter and the deathly hallows T: 2
C(3): youtube.com/watch?v=_EC2tmFVNNE T: 20
C(): movies.yahoo.com/movie/1810004624/info
Re-finding Goal
Q: potter T: 11
Q: trailer harry potter T: 50
Q: trailer harry potter and the deathly hallows T: 5
C(2): movies.yahoo.com/movie/1810004624/info

Fig. 4. An example of mis-classification between image and movie documents. The click without rank, c() in the original goal means a suggestion click from a search engine.

Original Goal
Q: jumps on dead whale T: 2
C(1): dailymail.co.uk/news/article-2816706 T:25
C(3): youtube.com/watch?v=7A3M8X5RE2A T: 20
C(4): bbc.com/news/world-australia-29876477 T: 10
C(5): cnn.com/video/man-rides-dead-whale
Re-finding Goal
Q: whale climbing T: 35
Q: jumping on a whale T: 10
C(3): youtube.com/watch?v=7A3M8X5RE2A T: 5
Q: Australian man riding a dead whale T: 5
C(2): cnn.com/video/man-rides-dead-whale

Fig. 5. An example of mis-classification between news and movie documents.

In this section we focused on the potential errors in the classification techniques, as the phase of detecting re-finding and verticals in this study were mainly automatic. However, extending labeling schemes with collecting user self-assessed reports will introduce different types of errors, particularly in terms of factors influential in user behavior, which we plan to study in future work.

6. EARLY PREDICTIONS

The previous section shows how prediction models for re-finding goals in verticals could obtain a reasonable level of accuracy without accessing the corresponding original goals. In this section we examine the effectiveness of predictions across a developing search goal, using real-time interactions while the user is searching. For this examination, we used our training set explained in Section 3.3 focus on non-navigational paired goals. This class of task includes longer re-finding goals, which are therefore more amenable to time-based analysis, unlike the majority of navigational paired goals that consist of a single query and click.

To build classification models at the early stages of a goal, we need to construct appropriate training sets since our current datasets are based on features computed over information from the entire re-finding goals. Among the features described in Section 4.1, although some features can be measured at the early stages of a goal (such as *"time to first click"*), others naturally require accessing the whole goal to be computed (for example, *"total dwell time after all clicks"*). To customize our training sets for the current investigation, instead of only incorporating features based on whether they can be computed on a real-time basis or not, we computed all features but given only the partial interaction information up to the stages in the re-finding goal currently under consideration. We studied early stages in re-finding in terms of two aspects: a) time dedicated to search, and b) the number of issued queries and clicks.

6.1. Query/Click Time

We first examine effectiveness based on wall-clock time, to give an overall indication of how well predictive models can be expected to perform if required to give a real-time response as searches unfold. This comparison is of particular interest since a search engine analysing the behavior of a real user has no advance indication of how long the particular search goal will be. To construct predictive models over time (i.e. based on partial information rather than the whole goal), interactions from re-finding goals were incorporated into generating datasets at different time points. For exam-

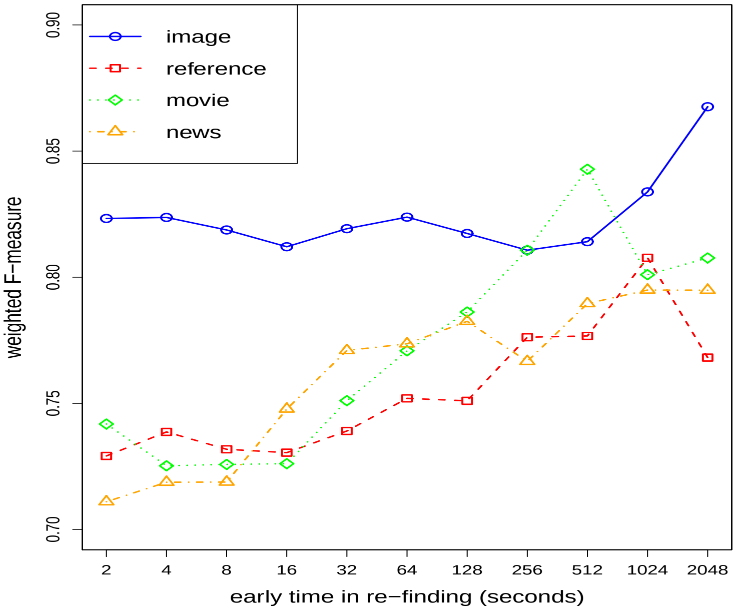

Fig. 6. Accuracy of real-time vertical predictions over time; the size of training sets is decreasing over time, as completed re-finding goals are discarded.

ple, given a time point of 2 seconds, all paired goals in the non-navigational dataset were truncated to only include interactions recorded before 2 seconds from the start of the search.

Over time, goals may reach a natural completion point where the final action of the goal has taken place. Completed goals are removed from the datasets at later time points. We limited the goal time to around half an hour (2,048 seconds consisting of 408 training items) from the beginning of the goal. Due to low number of cases close to the maximum completion time, we did not consider this time point in training predictions and we focused on early time points, which are more important for our real-time analysis. For the time-based datasets with partial information in the re-finding goals, all features listed in Section 4.1 were computed.

New classifiers were built using training sets corresponding to the different time points. The accuracies of these classifiers are shown in Figure 6. The x-axis shows the time lengths of re-finding interactions for each training set, and the y-axis shows the weighted F-measure scores of corresponding classifications. The overall weighted F-score is the the mean of F-score of classes, weighted by the proportion of items in each class.

From Figure 6, "image" predictions obtained the greatest accuracy, 82%, compared to other verticals early in re-finding goals. This vertical was also shown to be relatively independent from the original goal of the users in comparison to the other verticals, as discussed in 5.2.

Contrary to expectations, the performance of the classifiers for some vertical groups begins to decrease in performance at the higher end of the considered time spectrum. This could be due to a decreasing number of training items over time, as completed re-finding goals are removed. To better illustrate the effect of only partial information from re-finding goals on real-time predictions, we constructed new classifiers, where the size of training sets is held constant at all time points. The size of training set at the longest time point in our analysis, 2,048 seconds, had the smallest training set, consisting of 408 items. To construct training sets of equivalent size for the earlier time points,

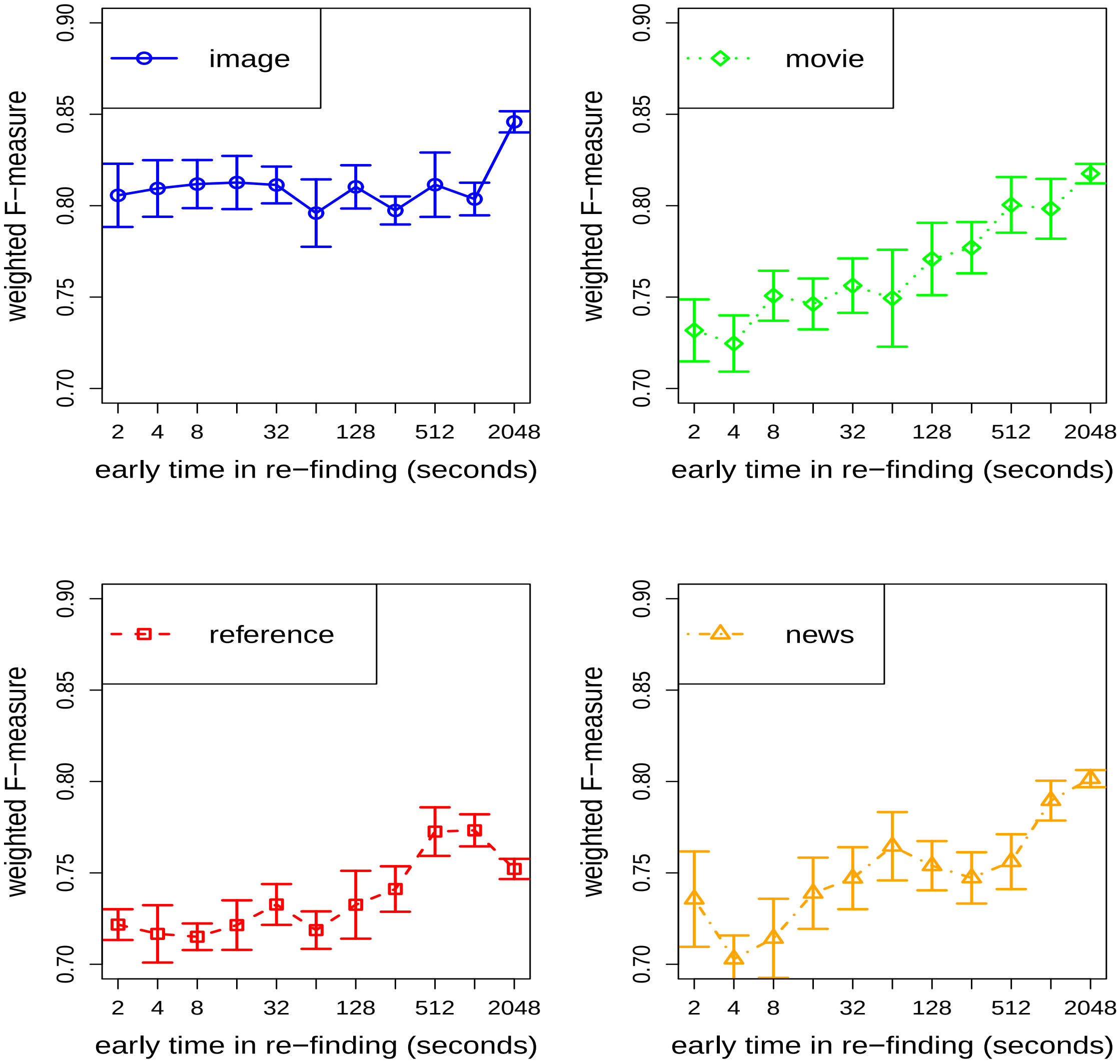

Fig. 7. Accuracy of real-time vertical predictions over time; the size of training sets are equal.

408 items were sampled from each of these; the sampling process was repeated 10 times for each time point before the last, and the mean results are reported.

On building new classifiers on the same size training sets, we used 10-fold cross-validation on each sampled training set. The mean and confidence interval values of weighted F-measure scores were computed for new classifiers and are shown in Figure 7. The confidence intervals reflect the variance in performance due to sampling and 10 times 10-fold cross-validation runs. As can be seen, "image" vertical is still highly predictable, even in the early stages of re-finding, and this performance is relatively constant over time. Other verticals demonstrate upward trends over time, suggesting a greater dependency of these verticals to entire information on re-finding goals.

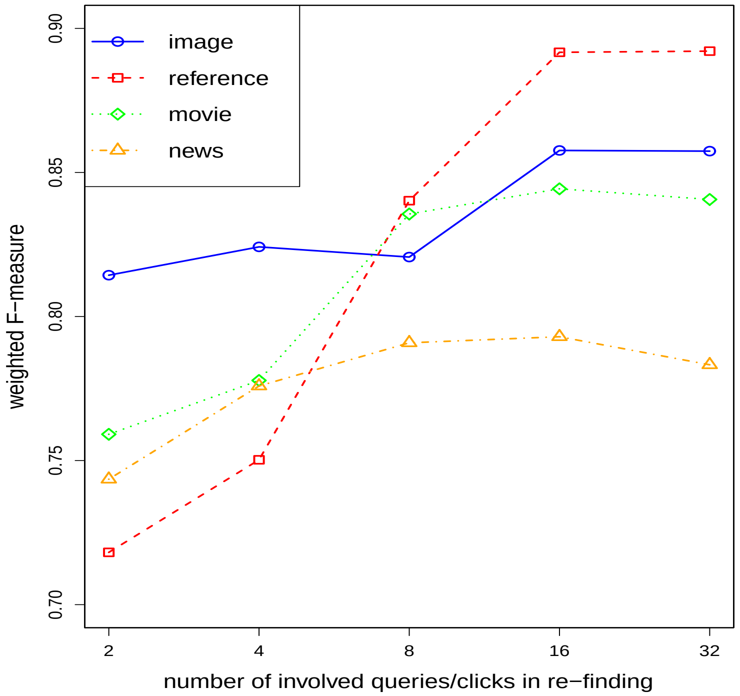

Fig. 8. Accuracy of real-time vertical predictions over the number of issued queries and clicks.

6.2. Query/Click Count

Overall, we found a general upward trend in the accuracy of predictions over time. However, wall-clock time may not always be reflective of the actual development of progress in a search goal, for example when the user interrupts their search activity to carry out an unrelated task. We therefore studied another way of measuring early stage in re-finding by counting the number of issued queries and clicks.

We employed the same approach as in Section 6.1 to build multiple classifiers over the number of queries and clicks in re-finding goals instead of the time dedicated to these interactions. On our non-navigational dataset, the minimum and maximum number of queries and clicks were 4 and 56 respectively. For training datasets at different interaction points, we computed all features but only using the subset of interactions up to the chosen point. As previously, those re-finding goals that have already been completed were discarded once their final number of queries/clicks was exceeded. In this exploration, the size of the training sets was quite balanced (from 1,485 instances for 2 queries/clicks to 1,471 instances for 32 queries/clicks).

The performance of the predictive models is illustrated in Figure 8, where predictions are built starting from only two queries and clicks in re-finding goals. Here, mostly upward trends can be observed in the performance of vertical predictions given a higher number of queries/clicks in re-finding goals. Similar to the results of Section 6.1, predictions of the “image” vertical appears to be less dependent on the entire information on re-finding (81.4 accuracy given one query and one click from the re-finding goal and information from the original goal). However, given more interactions of re-finding goals, the accuracy of predicting “reference” verticals outperform the “image” predictions. This trend does not appear in Section 6.1, because by increase in the count of queries/clicks we are further along in search than the corresponding clock-time covered in the previous section. As previously discussed in Section 5.1, when the full set of re-finding information is available, “reference” predictions are more accurate than “image” predictions.

Overall, we can conclude that although prediction accuracy can be improved given more interactions of the search, there is already a reasonably high level of predictive accuracy achievable in the early phases of search goals. This would benefit search engines for the adaptation of search results at the early stage of re-finding activities.

In practice to take advantage of these features, it is not necessarily required to either construct paired goals in advance, and the history-dependent features can also be useful during search particularly for logged in users. Using features that are distinctive for re-finding goals, the system can predict when the user is looking for a previously seen document early in search, and by further adapting search results, it can better help the user to reach to the target document. As an example for logged in users, in case of occurring ’common clicks’ between the current search and previous history of the user, the system can offer documents from other engaged clicks in the search history of the user in higher rank search results, where the user can reach to the target document with less amount of time and effort.

As the first study in distinguishing re-finding in verticals, we provided examples of grouping features based on their dependency to the search history of the user, which can be useful for early predictions, as one of the explorations in this work. Moreover, we examined how the contribution of query-based features might be different from click-based features into the prediction models. Other feature groups can also be explored in future work.

7. DIFFICULTIES OF RE-FINDING IN VERTICALS

Predictions of verticals could be particularly useful when users are struggling in re-finding documents. In these situations, search engines need to adapt the results, considering that there is a capability to predict re-finding in verticals early in the search, as established in the previous sections. We therefore further examine whether there are different levels of user difficulty across verticals. The verticals where users struggle the most when re-finding need to be identified for the attention of search engines to focus on possible improvements. In this exploration, we studied the effort of users as an indication of difficulty, as has previously been suggested in the context of general web search [Liu et al. 2010].

The effort of users (or the difficulties that they are having) can be approximated in terms of the number of submitted queries and clicks in the re-finding goals, as shown in previous work [Sadeghi et al. 2015]. We categorized user effort under four levels, as illustrated in Table X for each vertical. The first level includes likely easy navigational tasks, where re-finding goals consist of only one query and one click. Note that in this analysis we incorporated the likely navigational instances to be able to compare easy level tasks relatively across verticals as well. The next two levels of effort include either multiple queries (more likely search-based efforts) or multiple clicks (more likely indicates browse-based efforts). The final level contains the rest of re-finding goals, with multiple queries and multiple clicks.

The frequency of occurrence of each of the four effort classes is shown in Table X for the four vertical groups. Recall from Section 3.3.1, the focus of this study is on a limited type of re-finding tasks, and the frequencies are not representative of the entire scope of re-finding tasks. There are other types of re-finding where difficult search instances exist, such as when the user has failed to reach to the same documents as they clicked before (Figure 3).

A chi-square significance test over the counts of effort levels for all four verticals indicates significant differences ($p < 0.001$). All pair-wise comparisons between verticals were also statistically significant. The “news” vertical group showed the highest percentage of first level (i.e. easy) tasks compared to other verticals. Re-finding “image” documents appeared to be the most difficult for users with a higher proportion of re-finding in the fourth level of effort (last row of table). The other two vertical groups, “reference” and “movie”, appear to be highly similar. However, this could also be related to the nature of user behavior in re-finding documents, where the user submits more

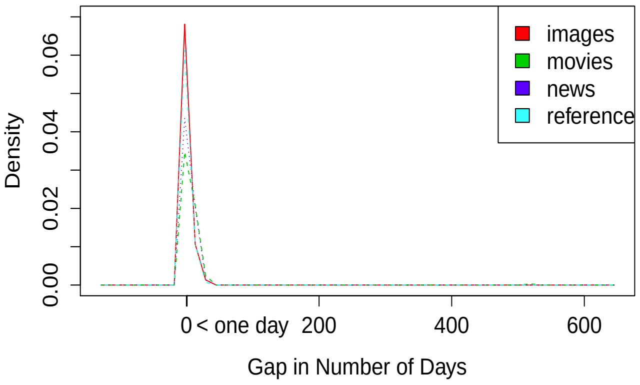

Fig. 9. The probability density of time gaps between re-finding and original goals in terms of the number of days across verticals.

queries and clicks for re-finding an “image” in comparison to the other verticals. As another reason for the difficulty of re-finding “image” documents could be the short length of queries and a small number of terms that could occur frequently in image retrieval as mentioned by Goodrum and Spink [2001].

We furthermore hypothesized that the number of queries and clicks should be correlated with the time gap between paired goals; however, only the “reference” vertical was significantly correlated (Pearson correlation $p < 0.0005$). Moreover, we examined whether the time gaps differ among verticals. In Figure 9, the probability density function for the time gaps are illustrated, which describes the relative likelihood for the time gaps to take on a given value. As can be seen from this figure, the variance of time gaps in terms of the number of days between re-finding and original goals are similar and long-tailed across media. We also calculated a Chi square test to assess whether the distribution of the time gaps were significantly different from each other. Among four verticals, the number of days between re-finding goals were significantly different for the “movie” vertical (Pearson chi-squared $p < 0.05$). For re-finding “movie” documents, about 50% of the time, re-finding happens in the gap of 4 or more number of days with the maximum of 1.5 years gap, while 50% of re-finding “image” and “reference” documents are within the same day with the maximum of one month gap. However, we could not observe higher number of queries/clicks in re-finding “movie” documents, although they might occur in longer time gaps in comparison to other verticals.

In relation to the time gap and search difficulty, *repeated* re-finding goals over time can also be studied and enriched by gathering users’ self-assessed reports on how their familiarity with the task would make difference in the difficulty of re-finding vertical documents. However, the focus of this study is on the behavioral indications of re-finding tasks and difficulties. Generating manual ground-truth data and investigating other indications of difficulty could be helpful to distinguish user difficulties across verticals, and we plan to consider this in future work.

Comparing the types of effort in the second and third categories, it can be seen that re-finding tasks across verticals tend to be more browse-based rather than search-based. This is also a natural feature for generic re-finding, which tends to be browse-based [Capra III 2006].

We also examined two time-based features, which were highlighted in past studies as being effective for predicting the satisfaction of users, which one could view as a concept inversely related to difficulty. Would such features be effective in a vertical re-finding context?

The two features examined are *“time to first click”* and *“number of engaged clicks”* (i.e. clicks with dwell time greater than 30 seconds) [Hassan et al. 2011]. From the Kruskall-Wallis and post-

Table X. Re-finding in verticals and the effort of the users in number of submitted queries and clicks.

	Image	Reference	Movie	News
one_query & one_click	8,115 (76.4%)	42,712 (81.1%)	707,536 (81.7%)	167,637 (85.4%)
one_query & multi_clicks	297 (2.8%)	3,036 (5.8%)	46,833 (5.4%)	8,107 (4.1%)
multi_queries & one_click	74 (0.7%)	938 (1.8%)	14,768 (1.7%)	1,330 (0.7%)
multi_queries & multi_clicks	2,134 (20.1%)	5,992 (11.4%)	96,326 (11.1%)	19,200 (9.8%)

hoc analysis in Section 4.2, the *"time to first click"* was not shown to be significantly different across verticals; however *"number of engaged clicks"* differentiates the "news" vertical from the "movie" and "reference" verticals.

The mean values for the *"number of engaged clicks"* in the latter verticals (i.e. 2.8, 3.2) were lower than for the "news" vertical (i.e. 3.5). However, this might not necessarily indicate that the users are less satisfied with "movie" and "reference" re-finding than "news" as a smaller number of engaged clicks exist. There are clicks with high dwell time in re-finding "news" documents when the user engaged with some other documents that are not related to the target document.

Apart from distractions, given the last click as the target known document, clicks with high dwell time in the middle of the search are not necessarily an indication of satisfaction. They might be more reflective of users struggling to recognize the target document rather than engage with relevant documents.

This shows that previously established features that indicate satisfaction or difficulty in the general web context may not necessarily be applicable for the re-finding context and they could vary on different verticals. The differences in difficulties could be either due to the behavioral variations in *re-finding* tasks, or various nature of vertical documents to be retrieved. As an example, people might have difficulty in formulating queries for documents in the type of images, which might not be the case for retrieving news documents with more textual content. We plan to study particular indications of difficulties for re-finding documents across verticals in future work.

8. DISCUSSIONS AND FUTURE WORK

This work focused on differences between re-finding behavior across verticals and examined comparisons with general web search. For the identification of verticals, we employed an automatic approach taking advantage of detecting re-finding goals based on common last clicks. In this automatic approach, a set of websites that have been already categorized under different verticals were matched with the last clicks in re-finding goals.

As this work is the first study in distinguishing re-finding across verticals, the aim was to investigate possible differences in a particular segment of the re-finding landscape. Moreover, we compared re-finding in verticals against general search, and we have not studied comparisons against not re-finding instances per each vertical. Due to the main goal of this study in differentiating the re-finding behavior and difficulties across verticals, our proposed vertical identification approach was designed for the re-finding context, which is not extendable to identify non re-finding vertical searches. In future work, we plan to generate training sets where non re-finding tasks can be selected per each vertical, which will provide insights about the differences between re-finding and general tasks specifically per each vertical.

Regardless of the particular type of re-finding or general tasks, from section 4.2, it can be seen that there are differences between vertical searches either in original or re-finding tasks, where search engines can be customized based on the type of the vertical. Examples of the implications of vertical differences for improving search engines are illustrated in Table XI. The search services in this table are suggested based on a set of differences between verticals, which can be operationalized by better understanding of searches across verticals.

Table XI. The implications of vertical search differences to be used for proposing search engine services.

Vertical	Feature Example	Service Suggestions
image	long search time and high number of clicks	improving the *summarization* of search results to help the user to recognize relevant results particularly if it can be addressable in an image.
reference	high rank in the first click	importance of the results in top ranks: it seems that for reference type of tasks, users rely on the top results, and therefore, it is important to *position* reference type of documents top in the search result page.
movie	queries formulated by named entities	enhancing *query suggestion* and *auto-completion* by identifying entity types in the search context.
news	high number of engaged clicks	providing same information from *diverse* sources (this behavior might be due to the fact that users would like to check news from different sources, and therefore, providing top documents from diverse sources might be useful.

There has been research on user's search behavior for different verticals in the general search context (e.g. studies by Goodrum and Spink [2001], and Diaz [2009]). Some of the results from this study are in line with previous findings. For example, we identified that re-finding "image" documents is challenging in comparison to other verticals. Similar findings were identified by Goodrum and Spink [2001], who compared the difficulty of retrieving (visual) "image" documents comparing to textual information. The researchers found out that a small number of terms can be repeated by users in image retrieval. This could be another reason for the difficulty of re-finding "image" documents, where we identified that users submit more number of queries and clicks, and re-finding "image" documents takes longer than other verticals.

There are past studies on vertical selection where their focus is on either query-based features or using click data for user's search modeling and their satisfaction in a heterogeneous environment. For example, in a study by Arguello et al. [2010] for vertical selection, features are mainly based on the content of queries, whether they are directly issued to a vertical search ('query-vertical features') or their contents are related to a vertical ('query features'). However, in our work, query-based features are focused on the behavioral aspect of issuing queries such as 'inter-query time', which makes it difficult for direct comparisons between features. There are other studies focusing on click behavior on vertical results but with a different purpose. As an example, in a study by Wang et al. [2013a], click distributions were incorporated into proposing a vertical-aware search model, where the likelihood of clicking on a vertical first in combination of different verticals were examined. Although the re-finding context is different from general search in a way that the user aims to find out a specific result in a particular vertical type, this will introduce a new problem on the likelihood that the user with a re-finding need might be satisfied with a different vertical result from the one that they have seen before. We plan to extend our current work to incorporate other types of re-finding, and also compare them against general searches per each vertical, where we need to generate manual ground truth data for vertical identifications.

Through generating ground truth data with a variety of verticals and also increasing the size of datasets, we also would be able to study re-finding in other verticals such as "shopping", etc. Moreover, in previous work by Arguello et al. [2010], it was shown that some features are portable across different verticals, and we can potentially explore this topic further. Considering topic-based features together with behavioral features could improve the accuracy of vertical predictions.

One of the applications in predicting re-finding in verticals could be narrowing the search results to a particular domain for re-finding in only one vertical, and if the search history of the user exists, previously seen documents in the particular domain can get higher ranks. However, the usability of this approach from the perspective of users requires user-based experiments. In particular, past research has shown that users tend to repeat the same actions [Capra III 2006], and that changing the ranks of documents in an unexpected way might in fact make re-finding more difficult [Teevan

2006]. This problem is referred to as change blindness. Therefore, changing search results to be more representative of a particular vertical, may need to be considered when the confidence of the prediction is high. In future work, we will examine the effectiveness of predictive models on improving the experience of users by conducting user experiments.

9. CONCLUSIONS

This study is, to the best of our knowledge, the first investigation of searcher behavior when re-finding the types of of documents associated with vertical domains. Our work aimed to a) identify potential distinguishing features of re-finding across verticals; b) predict re-finding within each vertical and investigate how different they are from searches that are not re-finding; and c) detect vertical documents to which users have more difficulty in re-finding.

A set of search behavioral features that were distinctive for re-finding tasks across verticals were identified. We also constructed classification models depending on the types of features (history-dependent vs. independent). On average, the accuracy of predictions across verticals is 85.7%. This compares to 97.5% for distinguishing re-finding from general search tasks. Prediction of re-finding "reference" documents acquired the highest accuracy among other verticals (89.5%).

When considering re-finding independent of the original search of the user, it seems that identifying "image" is the easiest to do (81.9%) with the accuracy of "movie" re-finding, being the hardest (75.4%). We hope that this type of prediction will enable the creation of re-finding services that operate independent of the search history of the user.

Further investigating the real-time prediction effectiveness of the models showed that predicting "image" document re-finding obtained the highest accuracy early in the search. Early predictions would benefit search engines with adaptation of search results during re-finding activities.

In studying difficulty in terms of user effort, re-finding in the "image" vertical appears to take more effort in number of issued queries and clicks than other investigated verticals, while re-finding "reference" documents seems to be more time consuming when there is a longer time gap between the re-finding and corresponding original search. Exploring other features suggests that there could be particular difficulty indications for the re-finding context and specific to each vertical, which we plan to investigate in our future work.

APPENDIX

In this appendix, we describe the features that were used for detecting re-finding and difficulties, which are summarized in Table XII.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

ACKNOWLEDGMENTS

This work was supported in part by the Australian Postgraduate Award (APA) scheme, and also by a Yahoo Research Award.

REFERENCES

Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. 2011. Find it if you can: a game for modeling different types of web search success using interaction data. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 345–354.

Jaime Arguello, Jamie Callan, and Fernando Diaz. 2009. Classification-based resource selection. In *Proc. CIKM*. ACM, 1277–1286.

Jaime Arguello, Fernando Diaz, Jamie Callan, and Jean-Francois Crespo. 2009. Sources of evidence for vertical selection. In *Proc. SIGIR*. ACM, 315–322.

Jaime Arguello, Fernando Diaz, and Jean-François Paiement. 2010. Vertical selection in the presence of unlabeled verticals. In *Proc. SIGIR*. ACM, 691–698.

Ofer Bergman, Ruth Beyth-Marom, and Rafi Nachmias. 2008a. The user-subjective approach to personal information management systems design: Evidence and implementations. *Journal of the American Society for Information Science and Technology* 59, 2 (2008), 235–246.

Ofer Bergman, Ruth Beyth-Marom, Rafi Nachmias, Noa Gradovitch, and Steve Whittaker. 2008b. Improved Search Engines and Navigation Preference in Personal Information Management. *ACM Trans. Inf. Syst.* 26 (2008), 20:1–20:24.

Ofer Bergman, Steve Whittaker, Mark Sanderson, Rafi Nachmias, and Anand Ramamoorthy. 2010. The effect of folder structure on personal file navigation. *Journal of the Association for Information Science & Technology* 12 (2010), 2426–2441.

Robert G Capra III. 2006. *An investigation of finding and refinding information on the web*. Ph.D. Dissertation. Virginia Polytechnic Institute and State University.

Edward Cutrell, Daniel Robbins, Susan Dumais, and Raman Sarin. 2006. Fast, flexible filtering with phlat. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 261–270.

Fernando Diaz. 2009. Integration of news content into web results. In *Proc. of WSDM*. ACM, 182–191.

Fernando Diaz and Jaime Arguello. 2009. Adaptation of offline vertical selection predictions in the presence of user feedback. In *Proc. SIGIR*. ACM, 323–330.

Susan Dumais, Edward Cutrell, Jonathan J Cadiz, Gavin Jancke, Raman Sarin, and Daniel C Robbins. 2003. Stuff I've seen: a system for personal information retrieval and re-use. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 72–79.

David Elsweiler, Mark Baillie, and Ian Ruthven. 2011a. What makes re-finding information difficult? a study of email re-finding. In *Advances in information retrieval*. Springer, 568–579.

David Elsweiler, Morgan Harvey, and Martin Hacker. 2011b. Understanding re-finding behavior in naturalistic email interaction logs. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 35–44.

David Elsweiler and Ian Ruthven. 2007. Towards task-based personal information management evaluations. In *Proc. SIGIR*. ACM, 23–30.

Abby Goodrum and Amanda Spink. 2001. Image Searching on the Excite Web Search Engine. *Inf. Process. Manage.* (2001), 295–311.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning* 46, 1-3 (2002), 389–422.

Morgan Harvey and David Elsweiler. 2012. Exploring Query Patterns in Email Search. In *Advances in Information Retrieval - 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012. Proceedings*. 25–36.

Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. 2010. Beyond DCG: user behavior as a predictor of a successful search. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 221–230.

Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2019–2028.

Ahmed Hassan, Yang Song, and Li-wei He. 2011. A task level metric for measuring web search satisfaction and its application on improving relevance estimation. In *Proc. CIKM*. ACM, 125–134.

Dzung Hong and Luo Si. 2013. Search result diversification in resource selection for federated search. In *Proc. SIGIR*. ACM, 613–622.

Bernard J. Jansen, Abby Goodrum, and Amanda Spink. 2000. Searching for multimedia: analysis of audio, video and image Web queries. *World Wide Web* 3, 4 (2000), 249–254.

Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proc. CIKM*. ACM, 699–708.

William Jones. 2007. Personal information management. *Annual review of information science and* *technology* 41, 1 (2007), 453–504.

William Jones and Harry Bruce. 2007. A report on the nsf-sponsored workshop on personal information management, seattle, wa, 2005. In *Report on the NSY PIM Workshop, January*. 27–29.

Jinyoung Kim and W Bruce Croft. 2009. Retrieval experiments using pseudo-desktop collections. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 1297–1306.

Alexander Kotov, Paul N Bennett, Ryen W White, Susan T Dumais, and Jaime Teevan. 2011. Modeling and analysis of cross-session search tasks. In *Proc. SIGIR*. ACM, 5–14.

Mark W Lansdale. 1988. The psychology of personal information management. *Applied ergonomics* 19, 1 (1988), 55–66.

Chang Liu, Jingjing Liu, and Nicholas J. Belkin. 2014. Predicting Search Task Difficulty at Different Search Stages. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 569–578.

Jingjing Liu, Jacek Gwizdka, Chang Liu, and Nicholas J Belkin. 2010. Predicting task difficulty for different task types. *Proc. ASIS&T* 47, 1 (2010), 1–10.

Jingjing Liu, Chang Liu, Michael Cole, Nicholas J Belkin, and Xiangmin Zhang. 2012. Exploring and predicting search task difficulty. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 1313–1322.

Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. 2013. Discovering Tasks from Search Engine Query Logs. *ACM Trans. Inf. Syst.* 31, 3 (Aug. 2013), 14:1–14:43.

Florian Meier and David Elsweiler. 2014. Tweets I'Ve Seen: Analysing Factors Influencing Refinding Frustration on Twitter. In *Proceedings of the 5th Information Interaction in Context Symposium (IIiX '14)*. ACM, 287–290.

Claude Nadeau and Yoshua Bengio. 2003. Inference for the generalization error. *Machine Learning* 52, 3 (2003), 239–281.

Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting clicks: estimating the click-through rate for new ads. In *Proc. WWW*. ACM, 521–530.

Kerry Rodden and Kenneth R. Wood. 2003. How Do People Manage Their Digital Photographs?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. ACM, 409–416.

Ian Ruthven and Diane Kelly. 2013. *Interactive information seeking, behaviour and retrieval*. Facet Publ.

Sargol Sadeghi, Roi Blanco, Peter Mika, Mark Sanderson, Falk Scholer, and David Vallet. 2014. Identifying Re-finding Difficulty from User Query Logs. In *Proc. of ADCS*. ACM, 105:108.

Sargol Sadeghi, Roi Blanco, Peter Mika, Mark Sanderson, Falk Scholer, and David Vallet. 2015. Predicting Re-fidning Activity and Difficulty. In *Proc. ECIR*. Springer.

Sidney Siegel and N. John Castellan. 1988. Nonparametric statistics for the behavioural sciences. *McGraw-Hill* (1988).

Shanu Sushmita, Hideo Joho, Mounia Lalmas, and Robert Villa. 2010. Factors affecting click-through behavior in aggregated search interfaces. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 519–528.

Jaime Teevan. 2004. *How people re-find information when the web changes*. Technical Report. MIT AI.

Jaime Teevan. 2006. *Supporting finding and re-finding through personalization*. Ph.D. Dissertation. Massachusetts Institute of Technology.

Jaime Teevan, Eytan Adar, Rosie Jones, and Michael AS Potts. 2007. Information re-retrieval: repeat queries in Yahoo's logs. In *Proc. SIGIR*. ACM, 151–158.

Jaime Teevan, William Jones, and Benjamin B Bederson. 2006. Personal information management. *Commun. ACM* 49, 1 (2006), 40–43.

Liang-Chun Tseng. 2012. *Modelling users' contextual querying behaviour for web image searching*. Ph.D. Dissertation.

Sarah K Tyler and Jaime Teevan. 2010. Large scale query log analysis of re-finding. In *Proc. WSDM*. ACM, 191–200.

Sarah K. Tyler, Jian Wang, and Yi Zhang. 2010. Utilizing Re-finding for Personalized Information Retrieval. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*. ACM, 1469–1472.

Chao Wang, Yiqun Liu, Min Zhang, Shaoping Ma, Meihong Zheng, Jing Qian, and Kuo Zhang. 2013a. Incorporating Vertical Results into Search Click Models. In *Proc. SIGIR (SIGIR '13)*. ACM, 503–512.

Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ryen W White, and Wei Chu. 2013b. Learning to extract cross-session search tasks. In *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1353–1364.

Steve Whittaker, Victoria Bellotti, and Jacek Gwizdka. 2006. Email in personal information management. *Commun. ACM* 49 (2006), 68–73.

Steve Whittaker, Ofer Bergman, and Paul Clough. 2010. Easy on That Trigger Dad: A Study of Long Term Family Photo Retrieval. *Personal Ubiquitous Comput.* (2010), 31–43.

Online Appendix to: Re-finding Behaviour in Vertical Domains

SEYEDEH SARGOL SADEGHI, RMIT University
ROI BLANCO, Yahoo Research
PETER MIKA, Yahoo Research
MARK SANDERSON, RMIT University
FALK SCHOLER, RMIT University
DAVID VALLET, Google

A. DESCRIPTION OF RE-FINDING AND DIFFICULTY PREDICTION FEATURES

The descriptions of features used to detect re-finding and difficulties are summarized in Table XII.

Table XII: The description of 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: ∗.

Baseline query level features (corresponding to the study proposed by [Teevan et al. 2007])	**Description**
equal query class ∗	The corresponding class of Teevan's classification for equal queries occurring in original and re-finding goals.
equal query elapsed time ∗	The elapsed time between equal queries.
equal query length ∗	The length of equal queries in terms of number of characters.
equal query no. of original clicks †	The number of clicks of the equal query in the original goal.
equal query no. of common clicks ∗	The number of common clicks of equal queries between the original and re-finding goals.
equal query no. of original uncommon clicks †	The number of clicks for the equal query in the original goal that are not common with the clicks of the equal query in the re-finding goal.
General web search (related) difficulty features	**Description**
goal length in no. of both queries and clicks ‡	The sum of both query and click counts in the goal.
goal length in no. of unique/all queries ‡	The number of unique (not repeated)/ all (including repeated) queries in the goal.
goal length in no. of unique/all clicks ‡	The number of unique (not repeated)/ all (including repeated) clicks in the goal.
mean no. of clicks across all queries ‡	The average number of clicks for all the queries in the goal.
time to the first click ‡	The spent time to the first click in the goal.
min/max/mean time to the first click of all queries ‡	The minimum, maximum, and average spent time to the first click for all queries in the goal.

DOI: http://dx.doi.org/10.1145/0000000.0000000

min/max/mean inter-query time ‡	The minimum, maximum and average spent time between queries in the goal.
min/max/mean inter-click time ‡	The minimum, maximum, and average spent time between clicks in the goal.
no. of engaged clicks (dwell time >30 seconds) ‡	The number of clicks in the goal with dwell time greater than 30 seconds.
no. of clicks on next page ‡	The number of clicks in the goal that are not in the first result page.
ended with query ‡	A boolean feature that indicates whether the goal has ended with a query or not.
exist advanced query syntax (e.g. quotes) ‡	This boolean feature indicates whether advanced options are used in issuing the query. The advanced options include quotes, +, and field operators.
queries per second ‡	The number of queries over total spent time in the goal.
clicks per query ‡	The number of clicks over the number of queries in the goal.
fraction of queries for which no click ‡	The number of queries in the goal for which there is no click over the number of queries with at least one click.
time span of goal ‡	The total spent time in the goal.
Extended re-finding features	**Description**
query overlap/URL overlap ∗	The corresponding class of the re-finding goal based on the query/click commonalities with the original goal. All classes have been explained in previous work by Sadeghi et al. [Sadeghi et al. 2015].
no. of common/uncommon/all clicks † ‡	The number of clicks in the goal in three main categories of common clicks, uncommon clicks, and all clicks.
mean query length of common/all clicks † ‡	The average length of queries corresponding to the common clicks and all clicks.
mean no. of query common/all clicks † ‡	The average number of queries corresponding to the common clicks and all clicks.
mean no. of uncommon clicks of all queries † ‡	The average number of uncommon clicks across all queries
mean no. of uncommon clicks of common click queries † ‡	The average number of uncommon clicks for queries with common click
days between paired goals ∗	The time gap between the paired goals in terms of the number of days.
effective search time † ‡ ∗	The total dwell time after queries and those clicks that have low dwell time (less than 30 seconds).
total dwell time after all queries † ‡	The total dwell time spent after submitting queries.
total dwell time after all clicks † ‡	The total dwell time spent after clicks.
total time to reach to the first common click † ‡	The total amount of time spent to reach to the first common click between the goals.
rank of the first reached common click † ‡	The rank position of the first reached common click between paired goals.
mean reciprocal rank of common clicks † ‡	Given each common click as the potential target document, the reciprocal rank of each common click is calculated and then averaged over all common clicks.
rank of the last click † ‡	The rank position of the last click of the goal.
no. of non-first-page ranked clicks in common/all clicks † ‡	Within common and all clicks, the number of clicks where they are not located in the first page result.

all common clicks skipped † ‡	A boolean feature which indicates whether whether there is a click at a lower rank, followed by the all common clicks at higher ranks.
exist jumped common clicks † ‡	This boolean feature investigates whether there is a common click, followed by a click at a higher rank.
exist non-sequential clicks † ‡	This feature investigates whether search results are clicked in a non-sequential way, rather than from top to bottom of the result page.
mean dwell time/relative dwell time of common clicks † ‡	The mean dwell time is based on the average time spent on common clicks; whereas, the relative dwell time is computed in terms of the fraction of click dwell time to the total time-span of the goal.
no. of repetitions of common clicks † ‡	The number of times that common clicks have been re-visited in the goal.
fraction of queries with no common clicks † ‡	The fraction of queries to which no common click exists in compariosn to teh corresponding goal.
re-finding is longer than original in length ∗	This feature investigates whether the re-finding goal is longer in terms of the sum of the number of queries and number of clicks.
re-finding is longer than original in no. of queries ∗	This feature investigates whether the re-finding goal is longer in terms of the number of queries.
re-finding is longer than original in no. of clicks ∗	This feature investigates whether the re-finding goal is longer in terms of the number of clicks.
re-finding missed engaged later clicks in original ∗	This feature is true if, after some common click, there are engaged clicks (with dwell time greater than 30 seconds) in the original goal that have not been clicked in the potential re-finding goal.
first query transformation type within pairs ∗	This feature measures the differences between the initial queries of original and likely re-finding goals (based on traditional query reformulation types: “exactly the same”, “error correction”, “specialization”, “generalization”, and non-trivial transitions considered as “other”).
exist common click in different ranks within pairs ∗	This feature investigates the existence of differences in the rank position of the common clicks between the paired goal.
common click in relation to the last click ∗	This feature examines whether a common click occurred in the last click of either the original or the potential re-finding goal.
mean relative goal position of common clicks † ‡	The position of common clicks measured over the total length of the goals, and then the average of the relative positions are computed.
min/max goal position of common clicks † ‡	The minimum and maximum of the positions of common clicks in the goal.
mean relative common clicks goal position (early, middle, late) † ‡	This feature is the categorized version of the “mean relative goal position of common clicks” in relative to the length of the goal, whether the position of the common click is in the initial of the goal (i.e. early), or in the middle of the goal (i.e. middle), or towards the end of the goal (i.e. late).
goal length in no. of both queries and clicks † ∗	The number of both queries and clicks in the goal.

goal length in no. of unique/all queries † ∗	The number of unique (not repeated), and all (might include repeated) queries in the goal.
goal length in no. of unique/all clicks † ∗	The number of unique (not repeated), and all (might include repeated) clicks in the goal.
mean no. of clicks across all queries †	The average number of clicks for all the queries in the goal.
time to the first click †	The spend time to the first click in the goal.
min/max/mean time to the first click of all queries †	The minimum, maximum, and average spent time to the first click for all queries in the goal.
min/max/mean inter-query time †	The minimum, maximum and average spent time between queries in the goal.
min/max/mean inter-click time †	The minimum, maximum, and average spent time between clicks in the goal.
no. of engaged clicks (dwell time >30 seconds) †	The number of clicks in the goal with dwell time greater than 30 seconds.
no. of clicks on next page †	The number of clicks where they are not located in the first page result.
ended with query † ∗	A boolean feature that indicates whether the goal has ended with a query or not.
exist advanced query syntax (e.g. quotes) † ∗	This boolean feature indicates whether advanced options are used in issuing the query. The advanced options include quotes, +, and field operators.
queries per second † ∗	The number of queries over total spent time in the goal.
clicks per query † ∗	The number of clicks over the number of queries in the goal.
fraction of queries for which no click † ∗	The number of queries in the goal for which there is no click over the number of queries with at least one click.
time span of goal †	The total dwell time spent in the goal.