Personalizing Search on Shared Devices

Ryen W. White
Microsoft Research
Redmond, WA 98052 USA
ryenw@microsoft.com

Ahmed Hassan Awadallah
Microsoft Research
Redmond, WA 98052 USA
hassanam@microsoft.com

ABSTRACT

Search personalization tailors the search experience to individual searchers. To do this, search engines construct interest models comprising signals from observed behavior associated with machines, often via Web browser cookies or other user identifiers. However, shared device usage is common, meaning that the activities of multiple searchers may be interwoven in the interest models generated. Recent research on activity attribution has led to methods to automatically disentangle the histories of multiple searchers and correctly ascribe newly-observed search activity to the correct person. Building on this, we introduce *attribution-based personalization* (ABP), a procedure that extends traditional personalization to target individual searchers on shared devices. Activity attribution may improve personalization, but its benefits are not yet fully understood. We present an oracle study (with perfect knowledge of which searchers perform each action on each machine) to understand the effectiveness of ABP in predicting searchers' future interests. We utilize a large Web search log dataset containing both person identifiers and machine identifiers to quantify the gain in personalization performance from ABP, identify the circumstances under which ABP is most effective, and develop a classifier to determine when to apply it that yields sizable gains in personalization performance. ABP allows search providers to personalize experiences for individuals rather than targeting all users of a device collectively.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process*; *selection process*.

Keywords

Attribution-based personalization; Interest model; Personalization.

1. INTRODUCTION

Personalization can be a powerful mechanism to individualize the search experience [5][19][35][38][39]. The application of personalization at scale relies on unique identifiers assigned to machines based on browser cookies or client-side software such as browser toolbars. Individualizing the search experience relies on a one-to-one correspondence between machine identifiers and individuals. Previous work has shown that machine identifiers are frequently associated with the search activity of multiple people (e.g., on over half of devices [47]), potentially impacting personalization effectiveness. We conjecture that by attributing observed search behavior to the correct *person* during the construction of interest models (profiles) and when processing received queries, the performance of personalization algorithms for applications such as predicting future interests or re-ranking search results will improve. We refer to this new approach as *attribution-based personalization*, defined as:

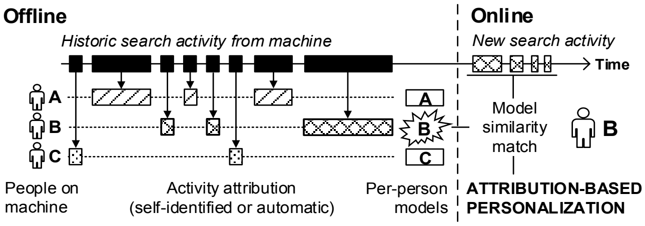
Offline
Online
Historic search activity from machine
New search activity
Time
A
B
C
A
B
C
Model similarity match
B
People on machine
Activity attribution (self-identified or automatic)
Per-person models
ATTRIBUTION-BASED PERSONALIZATION

Figure 1. Activity attribution and search personalization. Interest model construction happens offline (e.g., as part of a periodic process) and matching of new activity happens online (i.e., at query time). Historic and new data assigned to people using activity attribution (oracle or [47]). Models are built per person not per machine. New activity is assigned to searcher *B*.

DEFINITION: Attribution-based personalization (ABP) is a procedure to tailor search experiences to individuals. The three phases are (1) activity attribution and interest model construction for individual searchers from historic machine activity, (2) attribution of newly-observed activity to the correct searcher, and (3) application of that searcher's specific interest model for personalization.

Figure 1 illustrates the ABP process. Methods exist for phases 1 and 2 [47]. We focus herein on the value of attribution for phase 3.

Studies of personalization have examined its performance in controlled settings, given a clear mapping between people and profiles [39]. For the application of personalization in practice, e.g., at large scale in Web search engines, such mappings cannot be guaranteed. Shared used of a single machine is common and we have shown in previous work that we can build models to accurately attribute observed search activity to the correct person [47]. However, little is known about the value of attribution for personalization. To estimate this we perform an oracle study. The study compares interest models constructed given perfect (oracle) assignment of activity to people against models built from all machine activity, the standard practice in personalization [5][35]. The development of new methods for ABP is an exciting new frontier for Information Retrieval research. Our goal is to frame the ABP challenge and estimate the gains attainable from applying ABP in practice. To estimate its utility, we target future interest prediction, an important task in settings such as search, advertising, and recommendation [26][30][45].

We make the following contributions with this paper:

- Introduce attribution-based personalization as an important new area of personalization research and estimate its value for personalizing search experiences for people who share devices.
- Characterize differences in the searcher interest models constructed from person- and machine-based activity. We do so via a search log containing machine and person identifiers for each query. We demonstrate differences in the interest models built by attributing activity to specific people rather than machines.
- Show that these differences between identifier type (machine or person) are meaningful for an important application: predicting

SIGIR '15, August 09 - 13, 2015, Santiago, Chile
© 2015 ACM. ISBN 978-1-4503-3621-5/15/08…$15.00
DOI: http://dx.doi.org/10.1145/2766462.2767736

searchers' future interests. We experiment with using different sources for interest model construction (i.e., all historic activity linked to the identifier versus only activity for related queries).

- Controlling for search task effects, we identify properties of interest models (e.g., the number of searchers on machine) and queries (e.g., popularity, topic) for which ABP performs best.
- Learn a model from properties of the query and the constructed interest models to accurately predict *when* to apply ABP on a per-query basis. We show that applying this model for our task of future interest prediction yields significant gains in personalization performance over models that use all search activity from the current machine or simply apply ABP to all queries.

The remainder of this paper is structured as follows. Section 2 presents related work in areas such as personalization, activity attribution, and interest modeling. Section 3 describes our study, including research questions and methods. Section 4 characterizes multi-person searching. Section 5 describes a comparison of machine- and person-based interest models for future interest prediction. Section 6 explores the effect of query and model properties on ABP performance, as well as introducing a classifier to enable the selective application of ABP on a per-query basis. We discuss our findings and their implications in Section 7, and conclude in Section 8.

2. RELATED WORK

Research in a number of areas relates to our work here: (1) improving search using behavioral data, (2) search personalization, (3) attribution of search activity to individuals, (4) the construction of interest models, and (5) applications of these models for tasks such as recommending resources or predicting future searcher interests.

Improving Search using Behavioral Data: Behavioral data from search engines has been mined extensively to improve search relevance [1][21]. Radlinski and Joachims [28] proposed the use of query chains comprising connected sequences of queries to learn richer models of relevance that can capitalize on session behavior. Moving beyond individual queries, Radlinski et al. [29] model intent from queries and clicks for direct consumption by Web search engines. Similarly, Downey et al. [15] examined relationships between search queries and search goals estimated from the terminal page within the search session. Bilenko and White [7] used signals from aggregate post-query navigation trails to learn result rankings.

Search Personalization: The personalization of the search experience has received significant attention [5][19][23][33][39]. The potential for personalization [41] quantifies the improvement in retrieval performance that is attainable via personalization versus the performance that is obtained from trying to meet many searchers' needs. Long-term behavior has been applied to personalize search [35], focused on previous queries associated with the similar tasks over multiple search sessions [38]. Teevan et al. [39] showed that personalization improves as more data is available about the current searcher. Other signals such as short-term activity (e.g., in the same session) are also effective personalization signals [8][44][46]. Backing off from the activity of individual searchers, other signals such as physical location [4], the search activity of those engaged in similar tasks [48], or searchers similar along one or more other dimensions [43][50] have also been shown to be useful for personalization, especially when addressing "cold-start" scenarios where limited data can affect system performance [31].

Attribution of Search Activity: Truly personalizing the search experience relies on assigning observed search activity to an individual. At scale in search settings this is usually performed via an automatically-assigned unique identifier based on Web browser cookies, or to a toolbar or browser instance. Dasgupta et al. [13] studied the challenge of connecting different cookies to track the same user over time. This is important since cookie based identifiers are subject to cookie churn. Similarly, there may also be user variations within the machine identifier. White et al. [47] used the same dataset as employed in this study and showed that observing multiple people searching on the same machine is common (i.e., 56% of machine identifiers comprised activity from multiple searchers). They presented models to accurately estimate the number of searchers on a machine and attribute new search activity to specific searchers. The search activity attribution challenge that White et al. identified shares characteristics with other related domains such as fraud detection (to detect anomalous behaviors) [16] and signal processing (e.g., in blind signal separation, used to distinguish interleaved signal sources) [2][9]. However, there has been little attention (including in [47]) on the *application* of attribution for personalization. To our knowledge, the only exception is an early study by Singla et al. [34], who experimented with attribution methods for personalization, yielding some promising results. Since that study used a specific attribution method for a specific task (re-ranking results), the findings offer limited general insight about the potential utility of ABP. We address that shortcoming with the research in this paper.

Building Interest Models: To personalize the search experience, systems need to build and apply models that represent searcher interests. Interest models are common in user modeling and recommendation. Models can be constructed based on online activity only (as is the case with the personalization that we target in this paper), or based on client-side data, which can provide more complete insight into searcher preferences than is available from remote monitoring alone [24][39]. Interest models can assume a number of forms, including the text of search queries [33][36], specific URLs or Web domains visited in search [20][42] or browsing [24][44], or topical representations [19][35] (including topics specified explicitly [11][23]). Searcher interest models are often stored remotely, although client-side storage has been adopted if there is a need to access local data [39] or to address privacy concerns [6].

To reduce noise, personalization methods can target on-task activity only [5][38]. One extreme example is revisitation to the same URL for the same query. So called *personal navigation* [42], is also a powerful signal for search personalization and requires models comprising query-URL pairs mined from historic data. Personalization methods traditionally use overlap between queries to detect activity on the same task [38]. Such on-task interest models can be constructed from activity at different timeframes (short-term or long-term), and there are noteworthy tradeoffs in personalization performance given these timeframes (e.g., long-term models typically perform better for the first query in a session, where there is no session history) [5]. Other factors, such as the click entropy of the query, can also impact personalization performance [14].

Recommendation and Prediction: Once searcher interest models have been constructed, there are different ways in which they can be applied for personalization purposes. The traditional application is to adjust the ranking of search results, either through generating a new result set tailored to the searcher [27] or through re-ranking the initial results returned by a search engine [5][35][39]. Other applications of interest models include predicting which search results searchers will click [26], which display advertisements they will select [30], or which Web pages or topics they will focus on next [45][46]. For this study, we target the challenge of predicting future interests given historic interest models. The log data used in our study only contained data about queries and clicks (i.e., no result lists). Therefore, we could not reliably reconstruct the top-ranked search results presented at query time, as has been possible in prior personalization studies, where top retrieved results appeared in the logs used [4][5]. We therefore focus on future interest prediction.

Table 1. Descriptive statistics of dataset. Model building based on six months of logs. Evaluation uses one month of logs.

Statistic	*Dataset*	
	Model building	*Evaluation*
Total # machines	28,003 (of 490,754*)	
Total # searchers	68,908	49,869
Total # queries	19,145,916	3,619,913
Total # clicks	9,339,421	1,778,824
Avg. # queries/machine	683.3 (stdev=756.2)	129.2 (stdev=157.9)
Avg. # clicks/machine	333.3 (stdev=350.8)	63.5 (stdev=73.3)

*Number of machines with queries in train/test period for number of clicks ≥ 0.

Contributions: We extend prior work in many ways. First, through ABP, we focus on personalization at the sub-machine level, rather than assuming an exact correspondence between machine and individual, as has traditionally been the case in personalization. Second, we perform an *oracle* study highlighting differences in interest models depending the identifier type (machine or person) associated with search activity. This differs from the work of Singla et al. [34], where the focus was on applying a particular automatic attribution method, limiting the generalizability of derived claims about the effectiveness of ABP. Third, we show that observed differences between machine- and person-based models are meaningful, in our case for the task of future interest prediction. Fourth, we identify scenarios where ABP is most effective. As part of that analysis, we built a classifier to accurately determine when to apply ABP for each query, yielding strong gains in personalization performance.

3. EVALUATING ABP PERFORMANCE

We now describe our oracle study to understand the potential value from ABP. We describe our research questions, the data used, the models created, and their application in predicting future interests.

3.1 Research Questions

We answer the following four questions with our research:

RQ1 (Characterize interest models): What are some salient differences in interest models constructed based on the search activity from different identifier types (machine and person)?

RQ2 (Predict future interests): How accurately can we predict searchers' future interests by applying interest models based on search activity associated with the different identifier types?

RQ3 (Effect of interest model and query properties): What is the impact of properties of queries and interest models on the performance of models built from historic person or machine activity?

RQ4 (Automatically determine model source): Can we automatically determine which model source to apply per query? If so, what is the impact of taking this action on future interest prediction?

Answers to these questions help understand the value of ABP and useful scenarios for it. The answers can help search providers decide whether to invest in ABP and when to apply it in practice.

3.2 Data

The data that we used for our study was provided under contract by the Internet analytics company comScore. They recruited an opt-in consumer panel that has been validated to be representative of the online population and projectable to the United States population [18]. Millions of panelists provide comScore with explicit permission to passively measure all of their online activities using monitoring software installed on their computers. In exchange for joining the panel and providing search data, participants are offered a variety of benefits, including computer security software, Internet data storage, virus scanning, and chances to win cash or prizes.

The full dataset comprised unfiltered search queries on major Web search engines such as Google, Bing, and Yahoo!, collected over a two-year period from mid-2011 to mid-2013. The logs contained the text of queries, search result clicks, and the time that the events occurred (in the searcher's local time). Importantly for our study, the comScore search logs also contained a machine identifier (assigned to the machine) and a person identifier (assigned to each person who used the machine). An application is installed on the machine to record search activity and searchers are required to indicate to the logging software that they are searching at any given time. Machine-based identifiers are used in a range of online applications, either through Web browser cookies or other mechanisms such as search-provider toolbars; so their use in this study reflects the current state of the art. To remove variability caused by cultural and linguistic variation in search behavior, we only include log entries from the English-speaking United States locale. An advantage of using these data beyond the availability of both person and machine identifiers, is that they can be purchased from comScore to replicate and extend many of our findings (although the costs of data access may be prohibitive to some). The use of these logs is an important distinction from many log-based studies reported in the research literature, which rely on proprietary search logs, analyzed only by employees of commercial search providers. A disadvantage of using these data is that it limits the applications that can be studied to those not requiring result lists, e.g., we target interest prediction and not re-ranking since result lists are not available in this set.

For our study, we divided our dataset into two subsets: (1) *model building*: six months of comScore search logs for use in model construction (January 2013 to June 2013 inclusive)[1], and (2) *evaluation*: one month immediately following for use in evaluating the performance of the interest models at predicting future interests (July 2013). Activity (result clicks) from each person is used to construct the person-based models. All individuals associated with a machine are used in building the machine models. To ensure that we had sufficient data for our analysis, we required that there were at least 100 clicks from the machine during the model building period and at least 15 clicks from the same machine during the evaluation period. Note that searchers in our dataset only used a single machine; there was no movement of searchers between machines.

Summary statistics on the two datasets are shown in Table 1. These include some basic descriptive statistics such as the total number of queries and clicks in each set. As part of the data cleaning process, we removed machine identifiers that exhibited signs of being automated traffic by issuing more than 1000 queries on any given day.

3.3 Building Interest Models

We constructed interest models for each of the identifier types (machine and person) based on topical categorization of the clicked URLs in the *model building* dataset. We represented interest models as a distribution across categories in the Open Directory Project (ODP, dmoz.org) topical hierarchy (as in [45]). This provides a consistent topical representation of page visits from which to build models. ODP categories can also be effective for reflecting topical differences in the search results for a query [3] or a search context [46]. Given the large number of pages in our log data, we used automatic classification techniques to assign an ODP category label to each page. Our classifier assigned one or more labels to the pages based on a lookup into the ODP hierarchy using a similar approach

[1] Although logs spanned two years, we only used six months of log data for this task since (i) it captures searchers' recent interests, and (ii) we wanted to retain sufficient data for the set of prediction experiments described in Section 6.3.

Table 2. ODP-category-based interest models from all historic activity from the machine identifier and for each individual (*A*, *B*, *C*) whose activity comprises the machine history. Highlighting denotes top category in each model (column).

ODP Category	*Machine*	*Individual searchers*		
		A	*B*	*C*
Sports/Football	0.25	**0.75**	–	–
Sports/Tennis	**0.33**	–	**0.70**	0.30
Shopping/Clothing	0.07	–	0.10	0.10
Arts/Movies	0.18	0.25	0.10	0.20
Society/Issues	0.17	–	0.10	**0.40**

to [32]. In this approach, classification begins with URLs in the ODP and incrementally prunes non-present URLs until a match is found or miss declared. Similar to [32], we excluded pages labeled with the *Regional* and *World* top-level ODP categories, since they are location-based and are typically uninformative for constructing models of searcher interests. Since a Web page can appear in multiple categories in ODP, there may be more than one category assigned for each page. For URLs that were not present via direct lookup into the ODP, we instead used the output of a content-based ODP classifier [3] applied by the Microsoft Bing search engine during indexing. This combined approach resulted in coverage of 88% of all distinct clicked URLs in our dataset. The remaining 12% of URLs were not indexed by the search engine. Focusing on topics rather than URLs allowed us to build models with better coverage and were more semantically meaningful than might have been possible for models constructed from URLs or Web domains alone.

Table 2 presents fictitious but representative interest models comprising ODP categories, based on patterns observed in our dataset. The search activity on this machine comprised the queries and clicks from three searchers: *A*, *B*, and *C*. The table presents the normalized distributions for ODP categories for each individual. The table shows that the category distributions can be quite different between individual searchers and the machine, as well as between searchers. For example, *Sports/Tennis* receives most weight in the machine model, but is only highest weighted for one searcher (*B*) and is not an interest of searcher *A*. Some topics receive broad interest, e.g., all searchers are interested in movies. Although this is only one example, it is representative of the types of differences that we observe in our data and provides motivation for targeting interest models to specific searchers. Such individualized models may more accurately capture searchers' interests than combining all search activity associated with the machine into a single model.

3.4 Applying Interest Models

In this study, our chosen application of the interest models constructed during model building is in predicting future interests of searchers. Future interest prediction is employed in a variety of scenarios, including advertising [30], the development of relevance models [10], and recommending future content [45]. Given each of the queries in the one-month *evaluation* dataset, we focus on predicting the assigned ODP topic of the clicked URL given the interest models as described in the previous section. We use the ODP category label (pl_1) with the highest weight as the model prediction. For example, returning to Table 2, highlighted in bold are the predicted ODP labels for all activity in the models associated with the machine identifier and with each of the three individuals.

Task-relevant search activity: In addition to varying the *identifier type* (machine or person), we also varied the mechanism used for matching evaluation queries with those in interest model construction (i.e., the *match type*). We either used all historic activity or on-task activity only. Given a query, we define the *on-task* historic activity as clicks associated with queries with at least one non-stopword term in common with the provided query. Such a query matching methodology has been used in personalization to build focused, relevant models of searchers' historic interests [5][38].

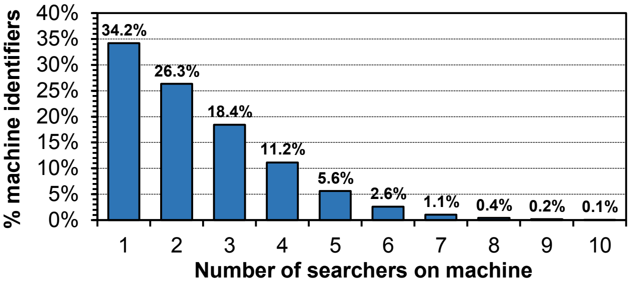

Figure 2. Fraction of machine identifiers in model building phase (6 months) with different numbers of searchers (*k*).

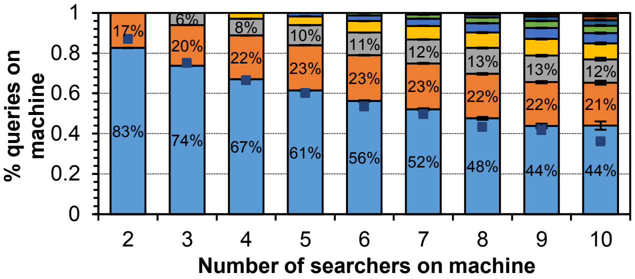

Figure 3. Distributions of queries per person across machines with different numbers of searchers (*k*) (±SEM).

Prior to analyzing the predictive performance of the models constructed by grouping based on the four combinations of identifier type and match type, we characterize shared device searching.

4. SHARED DEVICE SEARCHING

For the task of characterizing shared device searching we analyze data in the *model building* set. We examine the prevalence of shared device searching, how queries are distributed within machines, and the impact of multi-person searching on the general diversity of topical interests captured in models constructed from these data.

4.1 Prevalence

Figure 2 shows the fraction of machine identifiers that comprised different numbers of searchers. To construct this plot, we use data from the *model building* dataset (i.e., with at least 100 search-result clicks from the machine identifier over the six-month duration of that set). The figure shows that a significant fraction (65.8%) of machine identifiers comprise the search queries of more than one person. To understand the sensitivity of these trends to the duration of the activity used in model building, we also computed the distribution over one and three months of historic activity. In that data, 44.3% and 56.7% of machine identifiers comprised the search activity of multiple people for one and three months respectively. A core assumption in personalization is that all historic activity is associated with one person (the individual consumer of the service). The prevalence of shared device searching underscores the need to attribute search activity during model building and query handling.

4.2 Distribution between Individuals

Another important question relates to the distribution of queries between the individuals using a machine. For shared machines with more than one searcher (i.e., $k \geq 2$), we ranked searchers by the fraction of clicks and computed the average fraction of the total

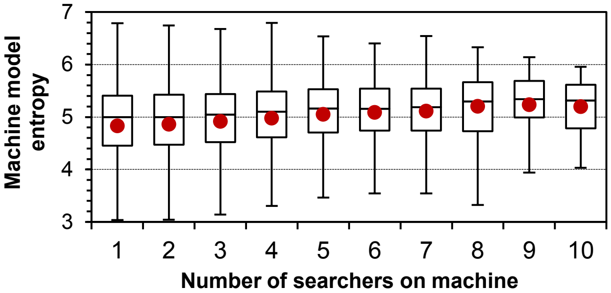

Figure 4. Box-and-whisker plot of machine model entropy for machines with diff. *k*. Mean is dot. Median is horizontal line.

query volume associated with each person. Figure 3 reports the distribution of queries per person across machines with different numbers of searchers. Each individual is allotted a region on the stacked bar (searcher #1 is first in blue, searcher #2 is second in orange, etc.). Percentage values are shown for the three most active searchers for each k. Other percentage values are excluded given space constraints and to simplify the graph. Also shown in the figure are error bars denoting the standard error of the mean (SEM). Given concerns about data skew, we also report the median percentages for the dominant searcher (squares in Figure 3), with similar trends.

The results show that for all values of $k \geq 2$ there is clearly a dominant searcher on each machine. For example, when k=2, there is usually an 80/20 division between click volumes on the machine between the primary and the secondary searcher. Although these are shared machines, the sharing is clearly not happening evenly. As the value of k increases, the query volume attributed to the dominant individual decreases in relation to the others (decreasing to around 40% at k=9). It is interesting to note that we only observe large changes in the search query volume associated with the primary/dominant searcher; the contributions of all other searchers on the machine remains largely unchanged and in some cases even increase with changes in k. The decrease in the contribution for the dominant searcher is also observed, even more markedly in fact, when we consider the median percentage of the query volume rather than the mean. These trends have implications for the effectiveness of ABP. At lower k, much of the activity associated with a machine identifier is already connected to an individual, meaning that ABP may be less effective. Also, since query volumes for non-dominant searchers are static or increase with k, there may be sufficient activity from at least some non-dominant searchers, irrespective of k, to build reliable interest models. We study the impact of k in the future interest prediction experiments reported later.

4.3 Diversity of Topical Interests

We were also interested in how k affected the breadth of resources accessed. Personalization performance would largely be unaffected by activity attribution, perhaps even hindered given sparser profiling data, if search intents were the same across different people. We computed the model entropies of the machine-based interest models, defined as $-\sum_{c \in C} p(c|m) \log(p(c|m))$, where c is a category drawn from the set of all ODP categories (C) assigned to result URLs clicked on machine m. Figure 4 presents a box-and-whisker plot for of this value across different k values. The horizontal segments inside the boxes represent the median, the top and bottom of the boxes denote the first and third quartiles, and the whiskers denote the maximum and minimum. The mean is denoted as a circle.

Figure 4 shows that there is an increasing trend in the topical diversity of interest models as the number of individuals associated with the machine increases (mean entropy = 4.83 at k=1 and 5.20 at k=10, Pearson's r = 0.981, p < 0.001). Interest models built based on historic data associated with the machine could therefore be more diverse and noisier if multiple people search on the machine.

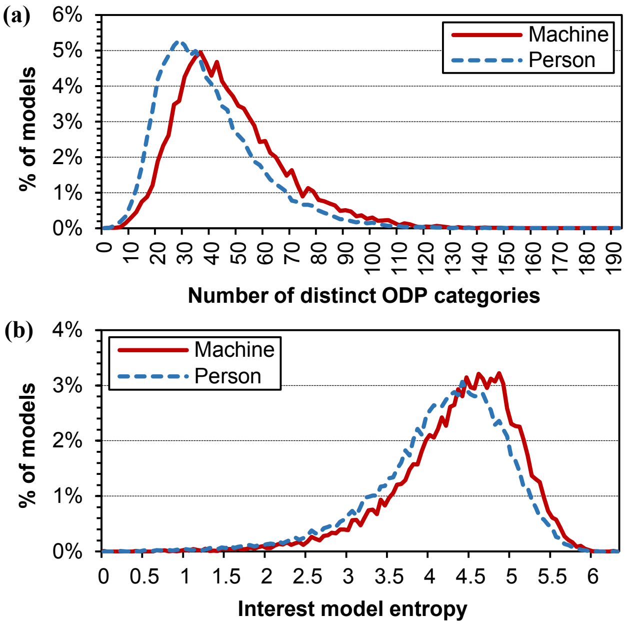

Figure 5. Comparison of person- and machine-based interest models in terms of: (a) number of distinct ODP categories comprising interest models, and (b) entropy of the models.

Turning our attention to comparing the person- and machine-based models, we examine the number of unique ODP categories and the interest model entropy. Figure 5 shows the distributions for each model type and highlights the differences between interest models constructed from machines or individuals. Models built from machine activity are generally more diverse, both in terms of the number of ODP categories assigned to clicks and the entropy of the normalized distribution of ODP categories. The differences are apparent visually and statistically (independent measures t-tests, both p < 0.001). We now explore whether the model differences are meaningful for an important application: predicting future interests.

5. PREDICTING FUTURE INTERESTS

We now describe the prediction task, the method, and the results.

5.1 Prediction Task

The prediction task was to predict the ODP categories of the clicked results for queries in our evaluation set. For each query, a distribution of ODP categories was computed, at the same hierarchy level as used for model construction (see Section 5.4.1). The analysis is performed at the query level, with the clicks from the person issuing the query as ground truth to evaluate the model. To avoid skewing our metrics toward highly-active searchers, we filtered the evaluation dataset to a maximum of 10 queries with result clicks per person. The exact number of queries per person in the evaluation set varies based on additional filters described at the end of Section 5.3, including the requirement that result clicks be classifiable in ODP.

5.2 Methodology

Given our evaluation set of queries and clicks (Q) comprising one month of logs, we generate tuples comprising {timestamp, machine identifier, person identifier, query, {result clicks}} for each query (q) in Q. Given this set of test queries and associated result clicks, we then took the following steps:

For each *identifier type* in {*machine*, *person*}:

- For each *match type* in {*all*, *on-task*}:

Table 3. Avg. metric values and percentage change for each combination of *match type* and *identifier type*. Bold = best.

Match type	*Identifier type*	*Metric*			
		P	*R*	*F1*	*RR*
All activity	*Machine*	0.179	**0.820**	0.294	0.307
	Person	**0.206**	0.781	**0.326**	**0.343**
	% Δ	+15.1	–4.8	+10.9	+11.6
On-task activity	*Machine*	0.625	**0.736**	0.676	0.695
	Person	**0.892**	0.732	**0.804**	**0.865**
	% Δ	+42.7	–0.5	+19.0	+24.5

- For each $q \in Q$:
  - If *identifier type* == *machine*:
    - If *match type* == *all*: Obtain all historic queries from the machine from the model building dataset.
    - Else if *match type* == *on-task*: Find all historic queries from the machine with ≥ 1 non-stopword terms in common with q in the model building dataset.
  - Else if *identifier type* == *person*:
    - If *match type* == *all*: Obtain all historic queries from the searcher from the model building dataset.
    - Else if *match type* == *on-task*: Find all historic queries from the searcher with ≥ 1 non-stopword terms in common with q in the model building dataset.
  - Obtain the clicked results for each of the queries.
  - Assign ODP categories to the clicked results using the methods described in Section 3.3.
  - Build an interest model (u) comprising the normalized distribution of ODP categories from the assignment.
  - Select top-weighted predicted label in u, denoted pl_1.
  - Compute the effectiveness of the method in relation to the ground truth (the normalized distribution of ODP categories across the URLs selected by the searcher).
- Average metric values for matchtype across all $q \in Q$ to compute the overall performance metrics for each of the four combinations of *identifier type* and *match type*.

Since we focus on shared device usage, we study devices with $k \geq 2$ in the rest of our analysis. This is a reasonable assumption practically. ABP will only apply when $k \geq 2$ and prior work has shown that shared device searching can be predicted with high accuracy (86%) given access to long-term search activity on a machine [47].

5.3 Metrics

We computed a number of metrics to assess the nature of the interest models that were created and their performance in predicting future activity in the form of a normalized distribution over ODP categories of the results that were accessed. Specifically, the metrics were associated with prediction correctness (i.e., the match between the predicted and actual category label(s)), measured as:

Precision (*P*): Precision was computed based on whether the top predicted label pl_1 equaled the actual label l_1 (1 or 0). For pages with multiple categories or queries with multiple clicks, a match was determined if pl_1 equaled *any* of the true categories. This strategy was adopted for the computation of the other metrics described in this section. Precision was averaged over all queries in the evaluation set to compute the final average precision value.

Recall (*R*): Recall is 1 or 0 depending on whether l_1 appears in the model, i.e., $\{pl_1, \dots, pl_p\}$, where the recall depth (p) is computed based on the number of predicted categories in the model or 10 (the median number of predicted categories across all queries and interest models), whichever is smaller. Limiting the depth of p is important in ensuring that obtaining a recall of 1 is a non-trivial task.

F1 score: This is the harmonic mean of P and R (i.e., $2 \times (P \times R) / (P + R)$). F1 has been used in a range of similar settings including the KDD Cup [22], as well as in related studies on the efficacy of models to predict future interests in search and browsing [45][46].

Reciprocal rank (*RR*): This is a commonly-used measure in Web search evaluation tasks, e.g., [12]. To compute this measure, the l_1 for the actual category label was compared progressively down the ranked list of category label predictions. If l_1 matched pl_i, the score assigned was the reciprocal of the prediction rank position $1/i$, and 0 otherwise. This was averaged across all queries evaluation dataset to compute the mean reciprocal rank (MRR) score.

Many of these metrics have been used previously in the assessment of search interest models [45]. They are computed for queries in the evaluation set for which we have: (1) clicks that can be classified in ODP, (2) a machine-based interest model from the search history, and (3) a person-based interest model from the search history. For the on-task variant, for (2) and (3) there needed to be at least one matching query in the search history from which to create interest models. There were a total of 442,690 queries that met all of these criteria for the all-activity analysis and 135,075 queries for the on-task portion of the analysis. The queries in both sets were used to evaluate predictive performance, although the on-task set is used for more analysis since it allows us to control for task effects, which have been shown to significantly affect search behavior [15].

5.4 Results: Person vs. Machine Models

We computed the range of metrics described in the previous section to compare the person and machine models. Table 3 reports the average metric scores for this comparison, for models built from all activity and on-task activity. The on-task models more accurately reflect the state-of-the-art in personalization [5][42]. The results reported suggest that there is a large benefit from ABP. We observe significant gains in precision, F1, and RR for person-based models over machine-based models of 11-15% for all activity and 19-43% for on-task search activity (two-way multivariate analysis of variance (MANOVA) over the four metrics, with *match type* and *identifier type* as factors: $F(4,2311056) = 10.22$, $p < 0.001$). All paired differences between machine and person were significant at $p < 0.001$ using Tukey post-hoc tests, with the exception of the recall comparison for on-task. Overall, the gain is higher from on-task activity. This reflects interests related to the current task, resulting in more focused interest models and better quality predictions. Recall is slightly higher for machine-based models for both match types, likely because these models are a superset of person-based models. Given the similar trends in our findings with the different metrics, we elect to focus on F1 for the remainder of our analysis. Focusing on F1 is useful since it considers both precision and recall, which are both important depending on the application.

5.4.1 Handling Near Misses

Since we rely on the structure of the ODP for our models, the metrics are sensitive to the match granularity. In particular, the metrics presented thus far rely on matching the full ODP labels in the interest models with the full label in the ground truth. We penalize the interest models for *any* mismatch between the predicted and actual label. However, small differences in estimates of search interests may be unimportant to search and recommendation systems. For example, interests represented by the ODP category *Sports/Football/NFL* could also be represented by *Sports/Football* with only a slight loss in precision. Using the matching approach described thus far, this would be regarded as a total miss, whereas it is actually a *near* miss. An amelioration strategy involves backing-off on all labels in the ground truth and the predictions to a specified level. One-level back-off means convert all ODP category labels to their

Table 4. Percentage change in F1-score at different hierarchy levels for person-based versus machine-based models (e.g., 1 = top level in ODP). Paired t-tests: ** $p < 0.001$, *** $p < 0.0001$.

ODP hierarchy level	*Match type*	
	All activity	*On-task activity*
1	+6.71%**	+10.25%***
2	+8.02%***	+13.48%***
3	+9.56%***	+17.87%***
All	+10.92%***	+19.02%***

top level (e.g., *Sports*), and two-level back-off means convert all labels to their top two levels (e.g., *Sports/Football*).

To understand the impact of such near misses on our experimental findings, we analyzed the percentage gain in F1 over the machine-based models from restricting the ODP category labels in both model building and evaluation to the top one, two, or three levels in the ODP hierarchy. The results are shown in Table 4. The table show that the gains still persist across all levels of category back-off, but the gains reduce in magnitude as we move up the hierarchy. As noted in the table, all differences remain significantly different (at $p < 0.001$) at all levels of back-off. The reduction in performance with higher ODP levels is interesting, and suggests that nuances in searchers' topical interests are important. These subtleties may be lost by backing off to higher levels in the hierarchy. This is also evident in Table 2, where the category *Sports* is insufficient to distinguish the preferences of person A for football from those of persons B and C, who favor tennis. For this reason, we utilize all hierarchy levels for the experiments in the remainder of the paper.

The results presented in this section are promising. They show that ABP can improve the performance of search personalization. We hypothesized that there may be other characteristics of the models and queries that may contribute to the performance of ABP.

6. IMPACT OF ADDITIONAL FACTORS

We now present additional analysis of the performance of the interest models for predicting future interests, conditioned on characteristics of the machine-based models and queries. To control for task effects, which influences behavior and hence experimental outcomes, we focus on the on-task variant of the models. We report comparative performance against the machine-based baseline, the state-of-the-art in model construction for personalization.

6.1 Effect of Model Properties

We study the effect of model properties on the performance of the person-based models. We focus on three aspects: (1) the entropy of the machine models, (2) the relative size of the person-based models as a fraction of the machine-based models, and (3) the number of searchers on the machine. We consider each property in turn.

6.1.1 Model Entropy

We were interested in whether there were any effects from the entropy of the machine models (defined in Section 4.3) on the relative performance of ABP. We hypothesized that if machine-based interests were diverse, then ABP would improve performance. We split machine entropy values into three equally-sized buckets (high, medium, low) and computed the average percentage gain in performance for each of the entropy groups. The results are reported in Figure 6a. The performance across all queries is shown with a dotted line in the figure. Error bars denote standard error of the mean (SEM). Given the large sample sizes, the error bars in this and other figures in this section are frequently small. The results show clear differences in the predictive performance as a function of the entropy of the machine model (one-way ANOVA, $F(2,135072) = 8.53$, $p < 0.001$). When the entropy is high, it is more likely that multiple searchers' interests are reflected in the machine activity (as seen in Figure 4), and applying ABP will be more beneficial.

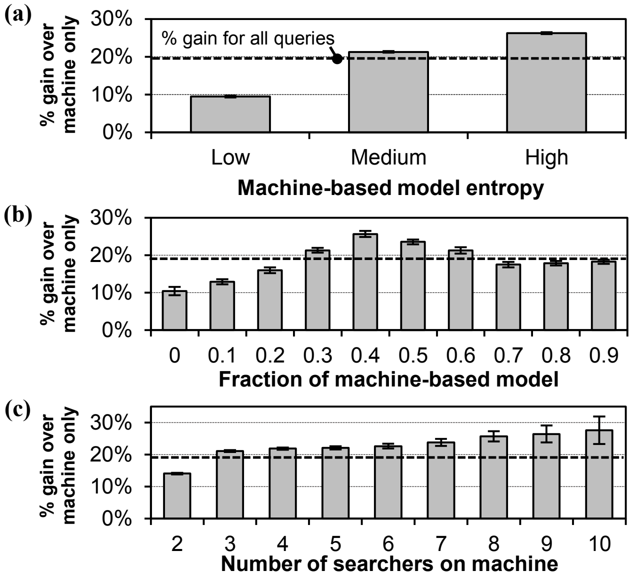

Figure 6. Effect of *model properties* on relative performance of machine-based vs. person-based models (±SEM). Effect of: (a) machine model entropy, (b) fraction of machine model, and (c) number of searchers. Dotted line denotes the predictive performance across all queries (i.e., % gain: 19.02).

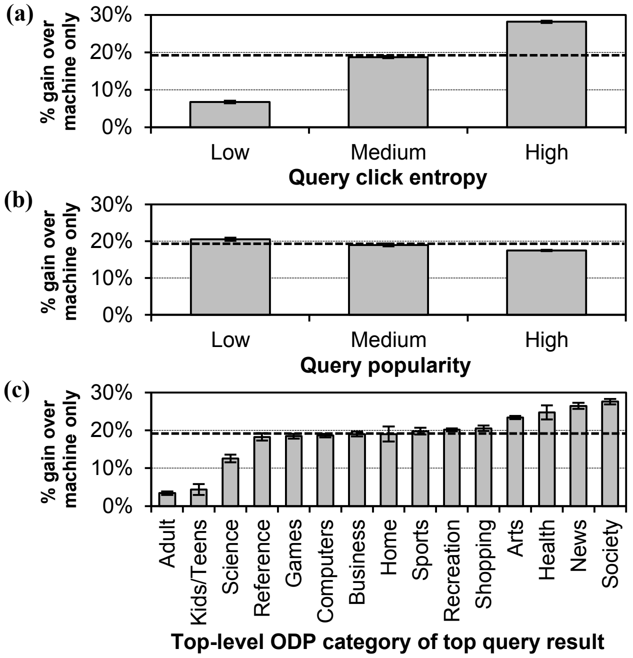

Figure 7. Impact of *query properties* on relative performance of machine- vs. person-based models (±SEM). Effect of: (a) query click entropy, (b) query popularity, (c) ODP category. Dotted line denotes the gain across all queries (i.e., 19.02%).

6.1.2 Relative Model Size

Another factor that may contribute to the relative differences in the performance of the person-based models is the relative model size. That is, the fraction of the number of clicks used to compute the person-based models compared to the total number of clicks for the machine models. Since we focus on machines with multiple searchers and all searchers contribute clicks, this fraction is always in the range (0,1). We bucket these fractions by rounding down to the nearest decile and compute the percentage gain over the machine-based models for each fraction. Figure 6b reports the percentage gain over the machine-based models for each of $\{0.0,\ldots,0.9\}$. There are clear differences in the gains depending on the fraction of machine-based model that the person occupies ($F(9, 135072) = 3.55$, $p < 0.001$). The gains from ABP are most apparent when there is no highly dominant searcher (in the range 0.4-0.6). In this region there may be sufficient data on each searcher's interests—which may not be the case when there is a highly-dominant searcher and one or more non-dominant counterparts with less data. The performance at 0.8 or above is similar to the baseline, likely because the machine and person models are more similar at high fractions.

6.1.3 Number of Searchers

Also related to the relative model size is the number of searchers on the machine, and in turn the number of searchers represented in the interest model. We speculated that more searchers lead to noisier and more diverse machine models, and greater benefit from ABP. To estimate the impact of the number of searchers, we computed the percentage gain in prediction performance of the person-based model across the range of different numbers of searchers on the machine, from 2-10. Figure 6c visualizes the performance across this range. The findings show an increase of predictive performance with a growth in the number of searchers (Pearson's $r = 0.913$, $p < 0.01$). The more searchers there are on the machine, the more likely the interests represented in the machine-based models are to be diverse, and the greater the benefit obtained from ABP.

6.2 Effect of Query Properties

In addition to properties of the interest models, we also examined the effect of query properties, namely: (1) click entropy, (2) popularity, and (3) topic. We consider each of these properties in turn.

6.2.1 Click Entropy

We examine the predictive performance of the machine and person models by considering the click entropy for the query. Click entropy measures the diversity in the clicks for the query [14]. To compute that statistic for the queries in our dataset, we used a separate set of query-click logs from the Microsoft Bing Web search engine, collected over a time period of two years overlapping fully the timeframe from the comScore logs. We did this since the search engine logs were much larger, enabling more reliable computations of click entropy and better query coverage than if we had used comScore logs alone. Indeed, 89% of the queries in our evaluation set appeared in the Bing logs, while only 65% appeared in comScore. Click entropy (CE) was divided into three buckets: (1) *high* ($CE \geq 2$), (2) *medium* ($1 \leq CE < 2$), and (3) *low* ($CE < 1$), based on the click entropy thresholds used in a prior personalization study [40]. The average percentage gains over the machine-based models for each of the three query-click entropy buckets are shown in Figure 7a. There are differences in click prediction performance between the three buckets (higher entropy is associated with higher gains for the person-based models, $F(2,120079) = 9.21$, $p < 0.001$). Higher click entropy queries represent more diverse interests are more amenable to personalization [40]. For these high CE queries, ABP is over 20% more effective than the machine-based baseline.

6.2.2 Popularity

To obtain information about query popularity, we used the same set of queries from the Bing search engine used in the calculation of click entropy. The queries in this dataset were grouped into three popularity buckets: (1) *low popularity-queries* that occur fewer than 10 times, (2) *medium popularity-queries* that occur 10-10,000 times, and (3) *high popularity-queries* that occur more than 10,000 times. The average percentage gains over the machine-based models for each of the three query popularity buckets are shown in Figure 7b. Although there are significant differences between the groups ($F(2, 120079) = 4.62$, $p = 0.01$), perhaps expected given the sample size), the effect size was small (partial eta squared (η^2) = 0.01) suggesting that differences are less meaningful. One explanation is the bucketing criteria (although our experiments with variations in the thresholds did not dramatically affect the results), or that raw popularity is not a good discriminator of queries for which person-based models may help (e.g., there may still be individual preferences for search results even within popular queries [42]).

6.2.3 Topic

As a final aspect of the analysis of query properties, we also examined the effect of the query topic on the performance of person-based models vs. machine-based models. For our analysis, the topic of the query was taken from the ODP category assigned to the top result returned for that query by the search engine whose logs were used to obtain entropy and popularity estimates. To simplify the analysis, we focus on the 15 top-level categories (again *World* and *Regional* were removed). Figure 7c shows the percentage gain for each of the topics. The topics are ordered on the x-axis in ascending order of gain with respect to the machine-based model. We see that although performance for many of topics was similar to the baseline, there were some topics that performed much better or worse than the machine-based baseline ($F(14,1004008) = 3.20$, $p < 0.001$). In the cases where the benefit was lowest (*Adult* and *Kids & Teens*) it is likely that only a subset of the household is interested in that content, so the machine-based approach is sufficient. For topics where the gain from ABP was highest (*News* and *Society*), there may be more general interest across searchers, and also more diversity in searcher preferences.

Although we examined the properties separately, there are likely to be *interactions* between these and other factors that influence the effectiveness of ABP. One way to understand this is to learn and inspect a classifier to predict when to use ABP on a per-query basis.

6.3 Applying Model and Query Properties

We used a separate set of 132,887 evaluation queries from 2,549 people (1,000 machines) taken from one month of comScore logs. We constructed on-task interest models from the six months prior (as before). These sets did not overlap in machines or time with the sets used in Sections 4 and 5. We featurized each model and query property described in the previous two subsections, namely:

- *MachineModelEntropy*: Entropy of the interest model constructed from activity on the current machine.
- *RelativeModelSize*: Fraction of machine interest model occupied by classified historic clicks from the current searcher.
- *NumberOfSearchers*: Number of distinct searchers whose activity is used in building the machine-based interest model.
- *QueryClickEntropy*: Click entropy for the query, computed based on the held-out Bing search log data (see Section 6.2.1).
- *QueryPopularity*: Popularity of the query, computed based on the number of instances in the same held-out search log.
- *QueryTopic*: Top-level ODP category of the top result for the query from the same search engine for the held-out data.

In practice, these features could be computed automatically for the query. Individual interest models for each searcher and the number of searchers can already be estimated from search histories [47].

For each of the queries in the 1000-machine dataset, we compared the F1 score of predictions using the machine- and person-based models. The prediction task is defined as predicting when to apply

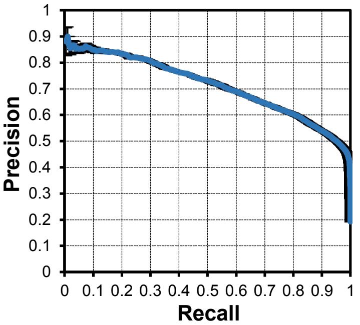

Figure 8. Precision-recall curve for the prediction of when to selectively apply attribution-based personalization. Values in the curve are averaged across ten folds for one run (±SEM).

ABP (vs. sticking with the machine-based model). In the dataset used for training (six months) and testing (one month), if the prediction based on the person-based model was more accurate (higher F1), then that was a positive example; if the prediction based on the machine-based model was more accurate, then it constituted a negative. Ties were treated as negative examples since in practice this would mean that ABP would not be applied by the search engine.

We used the featurized properties and these positive and negative labels to learn a binary classifier to determine when to apply ABP on a per-query basis. We applied a Multiple Additive Regression Trees (MART) [17] classifier to perform the prediction. MART uses gradient tree boosting methods for regression and classification. Advantages of MART include model interpretability (e.g., a ranked list of features is generated), facility for rapid training and testing, and robustness against noisy labels and missing values.

We employed ten-fold cross validation, stratified on the person identifier, so that queries from a person were either in train or test, but not both at the same time. We ran the experiment 100 times, each with a random assignment of searchers to folds, and compute the average F1 across all 100 runs. The findings of our experiments show that we can predict when to apply ABP with good accuracy. The average accuracy of the classifier (0.9179) exceeds the marginal baseline (0.7912) of always predicting the application of the machine-based interest models (the dominant class since in only 20.88% of queries ABP performs best; in 8.78% of queries the machine-based interest models perform best, and in the remaining 70.34% of queries the performance is tied). The gains over the marginal model are significant using a paired t-test ($t(999) = 4.73$, $p < 0.001$). Figure 8 shows the precision-recall curve generated by the model averaged across 10 folds for one experimental run. Examining the model in more detail, the most important features ranked based on their evidential weight are *MachineModelEntropy* (max), *RelativeModelSize* (0.699 of max), and *QueryTopic* (0.441 of max). The breadth of the topical interests on the machine, the contribution of the individual to the overall machine activity using in interest model construction, and the search topic of the query most affects the reliability of predictions about when to apply ABP.

While the strong classification performance is welcome, the test of true utility lies whether this will lead to gains in the accuracy of future interest prediction when the classifier is applied. To test this, we re-ran the prediction experiment described in Section 5, using the output of our classifier to determine whether we should apply ABP for each query. The results are reported in Table 5. The table shows clear gains from selective application over both models for all activity and on-task activity (all gains significant, $p < 0.001$). To help frame the magnitude of the observed gains, Table 5 also reports the performance of an oracle model (upper bound) that selects the best model source (machine or person) for each query in the evaluation set based on F1. We see that applying our classifier leads to ABP performance that is 88-96% of the oracle, a significant increase over always applying the person-based model and demonstrating the benefits of intelligently applying ABP for each query.

Table 5. Percentage change in F1 for selectively applying ABP models vs. always applying the machine-based model (all diffs sig. at $p < 0.001$, using paired t-tests).

Attribution Method	*Match type*	
	All activity	*On-task activity*
Always ABP	+10.90%	+19.02%
Learned	+14.13%	+23.00%
Oracle	+18.64%	+26.11%

7. DISCUSSION AND IMPLICATIONS

Shared searching on devices is common. Our oracle study, with perfect knowledge of the identifier (machines and people) to which actions were attributed, demonstrated clear utility from ABP. Conditioning on query and interest model properties showed that there are particular attributes that can help determine when to apply ABP for each query. We used the insights from our analysis to learn a model to accurately predict this type, and from its application, strong gains in personalization were realized. The effectiveness of ABP has significant implications for personalization, and heralds a new frontier for research in individualizing the search experience.

Before proceeding, we should discuss some limitations of the work as presented. This is a log-based analysis, meaning that we had limited insight into the intentions and interests of the searchers whose activity was examined. Interest prediction has been the focus of this and previous studies [45][46], but other applications such as re-ranking need to be studied (but could not be studied here given the lack of results lists in our logs). Although our study demonstrated the potential of attribution-based personalization, we only explored one implementation (via distributions of ODP categories). Experiments with other representations (e.g., page URLs or domains) and sources (e.g., query text or browsing behavior) are also needed. Finally, our focus is on long-term personalization and for queries with non-empty historic models for machine and person (i.e., only queries with some related history, a limitation of personalization generally [5]). Prior work has also employed within-session activity [8][44]. However, within session usage by multiple individuals is much less common: 97% of the search sessions in our logs are from one person. The remaining 3% are related to noise in using an inactivity timeout (30 mins, as in [46]) to demarcate search sessions.

Performing this oracle study required ground truth with both person and machine identifiers. This was needed to measure the value of ABP independent of the specific attribution method. Previous work has shown that the accurate automatic attribution to individuals is quite feasible [47]. We need to understand how ABP performance changes with such automated attribution. In addition, self-identification approaches (e.g., sign-in) need to be explored in this context, as do trends toward closer associations between people and their devices, which affects the prevalence of shared device usage.

8. CONCLUSIONS

We introduced attribution-based personalization and performed an oracle study to quantify the potential benefit of using it for future interest prediction. We show that there are significant opportunities to enhance search personalization via models tailored to individuals. We observe an increased accuracy in future interest predictions (11-19% in F1-score, depending on the match type) by applying this approach. The gains vary when we study particular properties of the interest models and the queries. We also show that further gains are obtainable by selectively applying ABP. Our findings show that ABP has strong potential, and we hope that they will inspire future work in this area. Our own future work will explore the application of automated attribution methods (building on [47]) to develop individualized interest models and combine these methods with our models to selectively apply ABP. Overall, it is clear that ABP could facilitate the delivery of more personalized results and recommendations to those searching on shared devices.

REFERENCES

[1] Agichtein, E., Brill, E., and Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. *SIGIR*, 19–26.

[2] Amari, S.I., Cichocki, A., and Yang, H.H. (1996). A new learning algorithm for blind signal separation. *NIPS*, 757–763.

[3] Bennett, P., Svore, K., and Dumais, S. (2010). Classification-enhanced ranking. *WWW*, 111–120.

[4] Bennett, P.N. *et al.* (2011). Inferring and using location metadata to personalize web search. *SIGIR*, 135–144.

[5] Bennett, P.N. *et al*. (2012). Modeling the impact of short and long-term behavior on search personalization. *SIGIR*, 185–194.

[6] Bilenko, M. and Richardson, M. (2011). Predictive client-side models for personalized advertising. *SIGKDD*, 413–421.

[7] Bilenko, M. and White, R.W. (2008). Mining the search trails of surfing crowds: Identifying relevant websites from user activity. *WWW*, 51–60.

[8] Cao, H. *et al*. (2009). Context-aware query classification. *SIGIR*, 3–10.

[9] Cardoso, J.F. (1998). Blind signal separation: Statistical principles. *IEEE*, 86(10): 2009–2025.

[10] Chapelle, O. and Zhang, Y. (2009). A dynamic bayesian network click model for web search ranking. *WWW*, 1–10.

[11] Chirita, P. *et al*. (2005). Using ODP metadata to personalize search. *SIGIR*, 178–185.

[12] Chowdhury, A. and Soboroff, I. (2002). Automatic evaluation of World Wide Web search services. *SIGIR*, 421–422.

[13] Dasgupta, A. *et al*. (2012). Overcoming browser cookie churn with clustering. *WSDM*, 83–92.

[14] Dou, Z., Song, R., and Wen, J.R. (2007). A large-scale evaluation and analysis of personalized search strategies. *WWW*, 581–590.

[15] Downey, D., Dumais, S., Liebling, D., and Horvitz, E. (2008). Understanding the relationship between searchers' queries and information goals. *CIKM*, 449–458.

[16] Fawcett, T. and Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3): 291–316.

[17] Friedman, J.H., Hastie, T., and Tibshirani, R. (1998). Additive logistic regression: a statistical view of boosting. Technical Report, Department of Statistics, Stanford University.

[18] Fulgoni, G.M. (2005). The "Professional Respondent" Problem in Online Survey Panels Today. Slides online at: http://www.sigmavalidation.com/tips/05_06_02_Online_Survey_Panels.ppt (Downloaded on May 15, 2015).

[19] Gauch, S., Chaffee, J. and Pretschner, A. (2003). Ontology-based interest models for search and browsing. *WIAS*, 219.

[20] Ieong, S., Mishra, N., Sadikov, E., and Zhang, L. (2012). Domain bias in web search. *WSDM*, 413–422.

[21] Joachims, T. (2002). Optimizing search engines using click-through data. *KDD*, 133–142.

[22] Li, Y., Zheng, Z., and Dai, H.K. (2005). KDD CUP-2005 report: facing a great challenge. *SIGKDD Expl.*, 7(2): 91–99.

[23] Ma, Z., Pant, G., and Sheng, O. (2007). Interest-based personalized search. *ACM TOIS*, 25(1): 5.

[24] Matthijs, N. and Radlinski, F. (2011). Personalizing web search using long term browsing history. *WSDM*, 25–34.

[25] Pitkow, J. *et al*. (2002). Personalized search. *CACM*, 45(9).

[26] Piwowarski, B. and Zaragoza, H. (2007). Predictive user click models based on click-through history. *CIKM*, 175–182

[27] Radlinski, F. and Dumais, S. (2006). Improving personalized web search using result diversification. *SIGIR*, 691–692.

[28] Radlinski, F. and Joachims, T. (2005). Query chains: Learning to rank from implicit feedback. *KDD*, 239–248.

[29] Radlinski, F., Szummer, M., and Craswell, N. (2010). Inferring query intent from reformulations and clicks. *WWW*, 1171–1172.

[30] Richardson, M., Dominowska, E., and Ragno, R. (2007). Predicting clicks: Estimating the click-through rate for new ads. *WWW*, 521–530.

[31] Schein, A.I. *et al*. (2002). Methods and metrics for cold-start recommendations. *SIGIR*, 253–260.

[32] Shen, X., Dumais, S.T., and Horvitz, E. (2005). Analysis of topic dynamics in web search. *WWW*, 1102–1103.

[33] Shen, X., Tan, B., and Zhai, C.X. (2005). Implicit user modeling for personalized search. *CIKM*, 824–831

[34] Singla, A. *et al*. (2014). Enhancing personalization via activity attribution. *SIGIR*, 1063–1066.

[35] Sontag, D. *et al*. (2012). Probabilistic models for personalizing web search. *WSDM*, 433–442.

[36] Speretta, M. and Gauch, S. (2005). Personalizing search based on user search histories. *WI*, 622–628.

[37] Sugiyama, K., Hatano, K., and Yoshikawa, M. (2004). Adaptive web search based on interest model constructed without any effort from users. *WWW*, 675–684.

[38] Tan, B., Shen, X., and Zhai, C. (2006). Mining long-term search history to improve search accuracy. *KDD*, 718–723.

[39] Teevan, J., Dumais, S.T., and Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. *SIGIR*, 449–456.

[40] Teevan, J., Dumais, S.T., and Liebling, D.J. (2008). To personalize or not to personalize: Modeling queries with variation in user intent. *SIGIR*, 163–170.

[41] Teevan, J., Dumais, S.T., and Horvitz, E. (2010). Potential for personalization. *TOCHI*, 17(1): 4.

[42] Teevan, J., Liebling, D.J., and Ravichandran, G.G. (2011). Understanding and predicting personal navigation. *WSDM*, 85–94.

[43] Teevan, J., Morris, M.R., and Bush, S. (2009). Discovering and using groups to improve personalized search. *WSDM*, 15–24.

[44] Ustinovskiy, Y. and Serdyukov, P. (2013). Personalization of web-search using short-term browsing context. *CIKM*, 1979–1988.

[45] White, R.W., Bailey, P., and He, L. (2009). Predicting user interests from contextual information. *SIGIR*, 363–370.

[46] White, R.W., Bennett, P.N., and Dumais, S.T. (2010). Predicting short-term interests using activity-based search context. *CIKM*, 1009–1018.

[47] White, R.W. *et al*. (2014). From devices to people: Attribution of search activity in multi-user settings. *WWW*, 431–442.

[48] White, R.W. *et al*. (2013). Enhancing personalized search by mining and modeling task behavior. *WWW*, 1411–1420.

[49] Xiang, B. *et al*. (2010). Context-aware ranking in web search. *SIGIR*, 451–458.

[50] Yan, J., Chu, W., and White, R.W. (2014). Cohort modeling for enhanced personalized search. *SIGIR*, 505–514.