Enhancing Personalization via Search Activity Attribution

Adish Singla[1,*], Ryen W. White[2], Ahmed Hassan[2], and Eric Horvitz[2]
[1]ETH Zurich, Universitätstrasse 6, 8092 Zürich, Switzerland
[2]Microsoft Research, One Microsoft Way, Redmond, WA 98052 USA
adish.singla@inf.ethz.ch, {ryenw,hassanam,horvitz}@microsoft.com

ABSTRACT
Online services rely on machine identifiers to tailor services such as personalized search and advertising to individual users. The assumption made is that each identifier comprises the behavior of a single person. However, shared machine usage is common, and in these cases, the activities of multiple users may be generated under a single identifier, creating a potentially noisy signal for applications such as search personalization. We propose enhancing Web search personalization with methods that can disambiguate among different users of a machine, thus connecting the current query with the appropriate search history. Using logs containing both person *and* machine identifiers, and logs from a popular commercial search engine, we learn models that accurately assign observed search behaviors to each of different users. This information is then used to augment existing personalization methods that are currently based only on machine identifiers. We show that this new capability to infer users can be used to improve the performance of existing personalization methods. The early findings of our research are promising and have implications for search personalization.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process*; *selection process*.

Keywords
Search activity attribution; Search personalization.

1. INTRODUCTION
Personalization of search results to individual users has been shown to be effective for improving search engine performance [2][8]. In online services, searchers are represented by a unique identifier associated with the machine that they are using to access the service, e.g., based on an IP address or a Web browser cookie. However, shared machine usage is common: 75% of U.S. households have a computer, and in most homes that machine is shared [4]. As a result, long-term histories of search behaviors of different people using the same machine may be interwoven, providing noisy signals for personalization methods designed to model the interests of individuals.

Recent research on *search activity attribution* [11] has shown that it is possible to (i) accurately determine whether a machine identifier comprises multiple users, and, if so, to (ii) assign observed search activity from a machine identifier to the correct individual, even if multiple people use a single machine. That study used logs comprising a census-representative sample of millions of users' search activities from the Internet analytics company comScore (comscore.com). Log entries contained both a person and a machine identifier. Searchers signed on to indicate that they were about to begin searching on the machine. These data, which are non-proprietary and available for purchase from comScore, facilitated the development and evaluation of activity attribution models to accurately assign observed queries to individuals.

In this paper, we describe research where we use an activity attribution model learned from comScore data and to integrate the attribution signal into long-term personalization. This allows us to perform feature generation at the level of individuals rather than machines (the current standard). As we will show, our results suggest that using this approach can lead to better search personalization. In addition, personalization can be useful when there is variation in user intent for the same query [10] and the value of long-term search histories can vary as function of query position in the session (i.e., they are more useful for initial queries) [2]. We therefore examined the effect of click entropy and query position on the performance of our personalization methods.

We make the following specific contributions:

- Demonstrate the need to consider shared machine effects, even when using only a limited duration profile (four weeks).
- Show that activity attribution methods can be effective in this context, and significantly outperform baseline methods.
- Apply the attribution signal for search personalization, and show that integrating this signal into existing methods can yield promising gains in retrieval performance.

We begin by describing the search activity attribution process.

2. SEARCH ACTIVITY ATTRIBUTION
We applied the methodology of [11] in personalization experiments. The main tasks performed in the search activity attribution process are as follows: (i) predict whether a machine identifier represents multiple individuals and estimate the number of individuals on the machine, (ii) cluster the search history on a machine to segregate the activity of individuals, and (iii) assign a newly arriving query to one of the clusters (i.e., the activity attribution task).

2.1 Dataset
We used a subset of the comScore data used in our previous work [11] for a five- week period in June-July 2013. Millions of panelists grant comScore permission to passively collect all of their online activity. Participants are offered incentives including computer security software, Internet data storage, virus scanning, and chances to win cash or prizes. The comScore data includes raw search queries, clicks on results, and the timestamp of each event. Importantly, each search action is tagged with a machine as well as a person identifier. The logs also contain a machine identifier (assigned to the machine) and a person identifier (assigned to each person who used the machine). An application is installed on participants' machines to record search activity and searchers are required to indicate to the software that they are searching at any given time.

In addition to comScore data, we also obtained logs from the Microsoft Bing search engine for the same five-week time period. The

* Research conducted during an internship at Microsoft Research.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609510

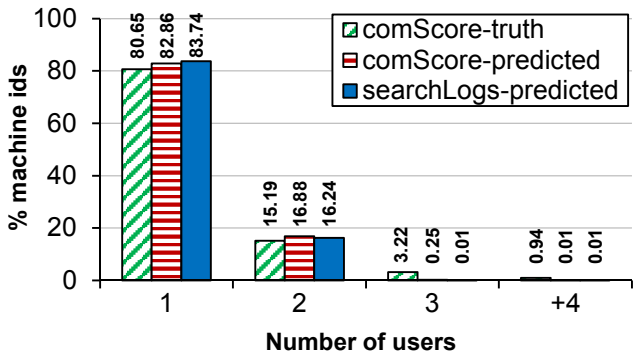

Figure 1. Percentage of machine identifiers comprised by number of users, from 1 to 4 or more. *comScore-truth* shows the numbers based on true person identifiers available in comScore datasets. The *comScore-predicted* and *searchLogs-predicted* shows the numbers based on prediction task.

search engine logs contain raw search queries, click events, and the top-10 results in the rank order to which they were presented to searchers. Each search action is also associated with machine identifier, but no person identifier is available. The challenge of the attribution process is to predict and assign a person identifier to the search actions in the search engine logs, using models learned from comScore data. We use the first four weeks of logs to construct logs for personalization and also prediction/clustering for user attribution. The logs from week five are used for the assignment, and for measuring the performance of web search personalization.

We limited both of the datasets to queries from the English-speaking United States locale, to minimize the effect of linguistic and cultural variability in our study. Furthermore, we filtered the comScore search activity to the search engine from which we obtained search interaction logs, so that attribution models learned from comScore logs can be applied to search logs without additional confounds from engine variations. The final data sets used for our analysis are comprised of search activities from 150K machines from comScore and two million machines from search engine logs. We segment the search activities into sessions, by introducing a session break whenever the searcher is idle for more than 30 minutes [11].

2.2 Predicting Multi-user Machines

The first step in the attribution challenge is to identify the machines with multiple individuals (a binary classification task) and then to estimate the number of individuals on a machine (a regression task). As the comScore data contains the information about person and machine identifiers, we use this data to train models using the robust multiple additive regression trees (MART) learning algorithm [6] for these classification and regression tasks. We used the first four weeks of comScore data, and used a total of 70 features for the learning algorithm, based on [11]. The features are grouped into classes including temporal, topical, behavioral, content, and referential features. The last two classes contain features such as the complexity of content pages and whether the queries from the machine contain pronouns that reference others who are likely to share a machine (e.g., spouse, child).

We first evaluate our models on comScore data itself using ten-fold cross-validation. We then use the first four weeks of the search engine logs to compute the features and use the learned models to predict the multi-user machines as well as estimate the number of individuals for machines that are predicted to be multi-user. Figure 1 illustrates the results of prediction task on search logs (*searchLogs-predicted*) and comScore data (*comScore-predicted*). Additionally, it shows the actual number of multi-user machines based on the person identifier information available in the data (*comScore-truth*). Table 1 reports the accuracy of classification task to predict that a machine is multi-user, compared to a baseline that always predicts any machine to be a single-user machine (the dominant class). In [11], we reported that 56% of machine identifiers comprise the Web search behavior of multiple searchers. In that study, we used two years of logs with activity from all major search engines, these numbers correspond to four weeks of logs with search activity filtered to only one engine (Microsoft Bing).

Table 1. Performance of various tasks in the search activity attribution challenge on comScore data. All the improvements (i.e., increase in accuracy, reduction in entropy, and increase in cluster purity) are significant using *t*-tests at $p < 0.001$.

Metric	*Attribution Methods*	*Baseline*
Prediction Task		
Classification Accuracy	0.818 (+1.5%)	0.806
Clustering Task		
Avg. Cluster Entropy	0.684 (–27.0%)	0.936
Avg. Cluster Purity	0.727 (+8.0%)	0.672
Assignment Task		
Accuracy	0.587 (+27.0%)	0.462
Cluster Purity	0.535 (+16.0%)	0.462

2.3 Clustering Search Activity

Given a multi-user machine as predicted by our models with number of estimated individuals to be k', the next task is to cluster the historic sessions (i.e., first four weeks) of search activity on the machine into k' clusters. Each cluster is then treated as an individual searcher, with their search activity belonging to that cluster.

One of the key components for the clustering algorithm is measuring pairwise similarity between sessions. We use MART-based regression [6], and the same methodology as [11]. Each session is first represented by feature vector, similar to those used in prediction task. Then, we compute a set of features between two sessions to capture their similarity or distance (for example, a binary feature indicating whether both sessions are on a weekend, or a real-valued feature computing the difference in number of number of queries issued in the session). We generate training data from comScore by randomly sampling 10% of the machine identifiers, then considering every pair of sessions on the multi-user machines, computing the pairwise features between these two sessions, and labeling them as 1 in case of belonging to same user, else 0. These data are then used to learn the regression model using MART.

Given that we have truth information available on comScore data, we can compute the performance of our clustering approach. This is shown in Table 1 and demonstrates significant improvements in the metrics of cluster entropy and purity, similar to [11]. Next, by using the regression model-based distance function learned from the comScore data, we apply it to the search engine logs to cluster the historic search activity (i.e., first four weeks) on multi-user predicted machines, into k' clusters. We assign each cluster a unique identifier, and use that to label all of the search sessions in the historic search logs, as a proxy for the person identifier.

2.4 Activity Assignment

As a final step of the user attribution process, we consider the task of assigning a newly arriving query to one of the clusters. We use the fifth week of comScore data and the search logs for this task. The method employs the regression-based session similarity function as used for the clustering. We consider assignment based on the first query of the search session, and all of the session-based features can be computed by limiting the session to a single query.

The assignment is done in two phases: (i) find the search session in the historic logs closest in distance to the newly arrived query, and (ii) assign the cluster identifier (as proxy for person identifier) of this most similar session to the newly arrived query.

We evaluated the performance of the task on comScore data by performing the assignment of first queries of sessions from the fifth week of data and evaluating using two metrics: (i) accuracy of assignment, and (ii) cluster purity. The performance is shown in Table 1. Assignment accuracy captures how often the most similar session that we assign indeed belongs to the same searcher. Cluster purity is defined as the proportion of true individual corresponding to the newly arrived query in the assigned cluster. We note that, based on our comScore data analysis, 97% sessions are performed by a single person (and the remaining 3% are noise given the 30-minute timeout method used for session identification). This means that performing attribution based on the first query in the session enables personalization for the full session thereafter.

We use the aforementioned methodology to assign the queries and sessions from the fifth week of search logs to the clusters formed from historic logs comprising first four weeks. At the end of this process, we have an identifier (proxy for person) associated with search activity in addition to the machine identifier present in logs.

3. SEARCH PERSONALIZATION

Given that we can attribute search behavior to individuals, we experimented with applying this method for personalization.

3.1 Features

The features used for the personalization include long-term click behavior and topical classifications of the clicked results, both similar to those shown to be effective in previous work on personalization [2][7][8]. As in the prior studies, we label the results visited by users across their long-term search histories using category labels from the Open Directory Project (ODP, dmoz.org). To do this automatically we use the content-based classifier described and evaluated in [1]. For each user, we calculate features such as the number of clicks on a URL by a user, the topic variation across all of their queries and result clicks, and the similarity between the retrieved results and the long-term profile of the user. These features are computed based on the first four weeks of search log data.

3.2 Datasets

We split the fifth week of search log data into three sets: training, validation, and testing. Table 2 shows the statistics for the datasets. Each machine identifier only appears in one of the three datasets. In building these datasets we focused on the machine identifiers in the top 10% of most likely to be multi-user given the confidence score of the multi-user prediction classifier. In practice, given the need for additional feature computation, search engines would likely only apply this method for machines thought to be multi-user.

3.3 Re-Ranking Models

The personalization models perform re-ranking of the top-10 results from a commercial search engine. For this experiment we used a variant of the ranker with no personalization, which was enabled as a control flight on the search engine for data gathering purposes.

Using our datasets, we train ranking models using the performant Lambda-MART learning algorithm [12] for re-ranking the top ten results of the query. We studied three personalization variants:

- **Machine:** Personalization features are generated using the machine identifier assigned by the search engine.
- **Person:** Personalization features are generated using the person identifier assigned by the activity attribution methods.

Table 2. Datasets used for personalization experiments.

Metric	*Training*	*Validation*	*Testing*
Dates	07/08-07/12	07/13	07/14-07/16
# Impressions	52,650	16,227	41,057
# Machine ids	5,233	3,441	4,652
# Sessions	28,280	9,006	22,443

Table 3. Gain in MAP over non-personalized baseline for each of the three personalization variants. Differences with *Machine* using Tukey post-hoc testing: * $p < 0.05$.

Metric	*Machine*	*Person*	*Machine+Person*
ΔMAP	+1.02	+0.99	+1.09*

- **Machine+Person:** Personalization features are generated using the union of the features for *Machine* and *Person*.

We generate the long-term features as described earlier and applied each of these models to re-rank the top-10 retrieved results. To determine the effectiveness of the methods for personalization, we use behavior-based judgments, tailored to individual searchers.

3.4 Judgments and Metrics

Evaluating personalization at scale is challenging. Searchers can have different intentions for the same query, meaning that third-party relevance labels may be insufficient. To address this concern, we exploit user clicks to obtain personalized relevance judgments for each query-document pair. Clicks can be classified into various types using the dwell time on the landing page. If the dwell time is too short, the searcher may be dissatisfied with the result. In this study, we label URLs with satisfied clicks (dwells $\geq$ 30 seconds [5]) positively, and others negatively. This method has been used for click-based judgments in prior personalization studies [2][8].

We measure performance using mean average precision (MAP), i.e., the mean of the average precision for each of our test queries:

$$MAP = \frac{1}{N}\frac{\sum_{i=1}^{n} Precision(i)Rel(i)}{\sum_{k=1}^{n} Rel(i)} \times 100 \qquad (1)$$

where n is the number of URLs in the impression, ranging from 4 to 10, depending on how many were shown. $Rel(i)$ is an indicator function returning one if the URL at rank i is relevant (positive), zero otherwise. $Precision(i)$ is the precision at cut-off i in the list.

3.5 Findings

We present our findings overall across all queries. To better understand search engine performance, in addition to all queries, we performed two additional experiments: broken out by the position of the query in the session and the click entropy of the queries. Since the performance of the baseline is proprietary we cannot report the absolute performance statistics. However, we can report the gain in MAP over the baseline from our personalization variants—results that let us more directly compare the different approaches.

3.5.1 All Queries

The results across all of the queries are presented in Table 3. All MAP gains are significant over the non-personalized baseline using paired *t*-tests (all $p < 0.001$). More importantly, the combination of machine plus person information yields modest (+7%), but significant, gains in performance over the machine-only method (one-way analysis of variance: $F(2,82112) = 3.82$; Tukey post-hoc test (*Machine+Person* vs. *Machine*): $p = 0.02$). Reasons that *Person* performs less well include an incorrect match in activity attribution or there being less historic data in *Person* from which to learn search preferences (worst case, both could apply). Since combining person

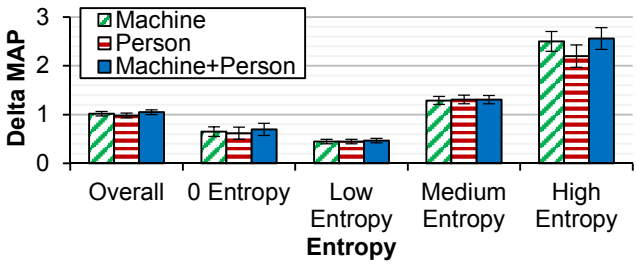

Figure 2. Gain in MAP over non-personalized baseline for each personalization variant split by bucketed entropy values.

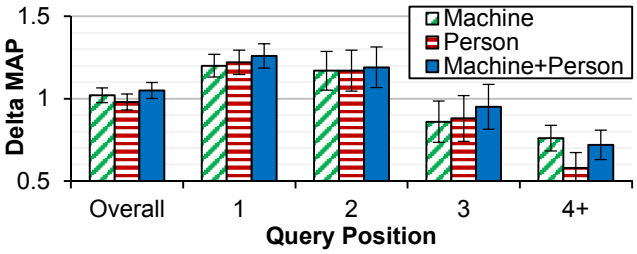

Figure 3. Gain in MAP over non-personalized baseline for each personalization variant at different query positions.

and machine features leads to an overall relevance gain, the combination may provide some protection against attribution errors, and *Person* may emphasize aspects of long-term interests not apparent with machine identifiers. In addition, we also sought to understand how personalization accuracy varied with attributes of the queries.

3.5.2 Click Entropy

Variability in clicks over all users for the same query may correlate with variations within a single machine identifier if it is comprised of multiple users. We computed the click entropy [3] for the queries in the test set using a set of logs from a one-year period that did not overlap with the timeframe for these experiments. We split the MAP gains by bucketed entropy values as shown in Figure 2. We divided queries into four different buckets. The first one is for queries with zero click entropy. These are largely rare queries that received no clicks or only one click. The rest of the queries were split into three equally sized buckets and were labeled as low entropy (entropy ≤ 0.69), medium entropy ($0.69 <$ entropy ≤ 1.4) and high entropy (entropy > 1.4). The figure shows that the combination of machine and person information is always in some way better than machine only or person only. Contrary to our expectations, machine information is also more useful than person information in high-entropy queries. One explanation is that there are relatively few of such queries for which the benefit of specific per-person targeting justifies the reduced quantity of training data.

3.5.3 Query Position in Session

The effectiveness of long-term personalization varies by position in session [2]. We examined MAP gains at query positions {1,2,3,4+} and report the gains for each method in Figure 3. The figure shows that person information is especially useful at the outset of the session where very limited information is available. This results in better performance (vs. machine information only) when we use either person information only or use both sources. In contrast, for the small number of queries at position 4 or more, person information is less useful and machine information yields the best performance.

4. DISCUSSION AND CONCLUSIONS

We explored the application of search activity attribution methods for search personalization and obtained promising early results (7% gain). Analysis of click entropy and query position in session showed particular benefit. For example, adding person information is effective for the first query in the session, where long-term features can most influence personalization [2].

User activity attribution is an important and largely unexplored aspect of personalization. We tested our attribution methods on the subset of comScore users who search with Bing. We need to better understand potential biases in this cohort and to explore ways to obtain this truth data with consent from search engine users directly, without requiring a separate dataset. Second, we need to more fully understand whether attributing activity to a person is important in personalization, or whether the subset of activity from a machine identifier focused around a particular interest would suffice. There may also be cases where users of a single machine share general interests (e.g., politics), but have different sub-interests or beliefs that would be amenable to targeted personalization; more analysis is needed. Finally, the history length in this study (four weeks) was limited by data availability. We focused on long-term personalization since search attribution is likely more valuable in that setting than short-term personalization; 97% of our search sessions involved only one person. It is important to study attribution for longer-term personalization, where shared machine usage may be more evident in search behavior. For example, earlier [11] we showed that 56% of machine identifiers over two years comprised the behavior of multiple searchers.

Future work involves improving the accuracy of the activity attribution, obtaining more insight into benefits from the attribution (e.g., over task-based alternatives [9]), as well as better understanding the types of queries and people where the methods perform best (e.g., dominant vs. non-dominant searchers on the same machine). We believe that it is also important to understand how best to integrate the attribution signal into personalization, including smart blending of results from both machine and person identifiers, and the selective application of individualized models for queries and/searchers—all with the goal of further enhancing the search experience for individual searchers.

REFERENCES

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

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

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

[4] File, T. (2013). *Computer and Internet Use in the United States*. http://www.census.gov/prod/2013pubs/p20-569.pdf

[5] Fox, S. et al. (2005). Evaluating implicit measures to improve Web search. *ACM TOIS,* 23(2): 147–168.

[6] Friedman, J.H., Hastie, T., and Tibshirani, R. (2008). *Additive Logistic Regression*. Tech. Report, Stanford University.

[7] Gauch, S., Cheffee, J., and Pretschner, A. (2003). Ontology-based user profiles for search and browsing. *WIAS*, 219–234.

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

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

[10] 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.

[11] White, R.W., Hassan, A., Singla, A., and Horvitz, E. (2014). From devices to people: Attribution of search activity in multi-user settings. *WWW*, 431–442.

[12] Wu, Q. et al. (2008). Ranking, boosting and model adaptation. *Microsoft Research Tech. Report MSR-TR-2008-10.*