Table 1. Features used to compute the similarity between two tasks, t and t', in the computation of $k(t, t')$.

Feature name	Definition
FullQueryOverlap	The fraction of all queries in the union of t and t' that the two tasks have in common
QueryTermOverlap	The fraction of all unique query terms in the union of t and t' that the two tasks have in common
QueryTranslation	Semantic similarity between the queries in t and the queries in t' ($P(t \mid t', \theta)$ as defined earlier)
ClickedURLOverlap	The fraction of clicked URLs in the union of t and t' that the two tasks have in common
ClickedDomainOverlap	The fraction of clicked domains in the union of t and t' that the two tasks have in common
CategorySimilarityKL	The Kullback-Leibler divergence between the ODP category distribution from result clicks in t versus the same distribution from t'.
CategorySimilarityCosine	The cosine similarity between the ODP category distribution from result clicks in t versus the same distribution from t'.

and estimates of topic expertise. The first two are based on information that is readily available in the logs that we used for this study. The latter could be estimated based on patterns of activity, interest in a topic, and success within a topic over time. Queries and similar search tasks in this case are only drawn from the particular cohort (e.g., only from users in the same location as the current searcher rather than all searchers.

4.2.3.1 *Local Cohort*

It has been shown that interests can be location specific, e.g., a user querying for [msg] in New York City, NY may be more likely to mean Madison Square Garden than monosodium glutamate [3], and local experts may have better knowledge about the places to select [47]. In this case, we learn our features from users querying from the same location. To identify the user location we use the user's IP address to determine the city and state for every user. We could not use each {city, state} pair as its own cohort because the population of many locations is insufficient. To address this, we use the city for the most populated locations and back-off to state for the less populated ones. Specifically, the location of a given user is his city if they are in one of the largest 200 U.S. cities by population. Otherwise, the location is the state. For example, a user querying from Austin, Texas would be in the "Austin" cohort, whereas a user querying from College Station, TX would be in the "Texas" cohort.

4.2.3.2 *Web Browser / Search Entry Point Cohort*

We also created groups of users based on the combination of the Web browser(s) that they use (e.g., Internet Explorer, Firefox, Chrome, multiple browsers, etc.) and the entry point(s) that they use to reach Bing (e.g., Bing homepage, MSN.com, browser search box, multiple entry points, etc.). The determination for each user is based on a held out set of log data from before the time period examined for this paper. Our hypothesis was that users using the same Web browser and entry point may have similar search preferences or be similar demographically (as a recent report by ComScore suggests[1]), and demographics can influence search behavior [44].

4.2.3.3 *Topic Cohort*

Previous work has shown that people with topic knowledge are more efficient and effective in completing their search tasks [50]. We hypothesized that by focusing on the behavior of experts, we could help users target better quality content. As such, in defining the cohorts we limit the tasks to those from users with significant expertise in the topic of interest. This allows us to learn from expert users in particular, versus learning from one's personal history or the set of all users, comprising users of all domain expertise levels.

To identify users with significant expertise in different topics, we had to assign topic labels to different queries. We use one of 25 topics to describe any given query. For each such topic, a set of manually-labeled queries are collected from trained judges and a proprietary text classifier is trained using the labeled data. The classifier is then used to assign topics to other queries. We used a set of binary classifiers, one for each topic, allowing queries to belong to multiple topics. Example topics include: *Entertainment*, *Names*, *Commerce*, *Navigational*, *Travel, Technology* and *Sports*.

We use first week of data described in Section 3.1, corresponding to the feature-generation week. A user U is deemed to be an expert in topic P if the following three conditions are satisfied:

1. **Activity:** The number of queries submitted by U is more than the average number of queries per user.
2. **Topic Interest:** The percentage of queries ϵ P submitted by U exceeds the average percentage of queries ϵ P across all users.
3. **Success:** The task success rate of U on tasks ϵ P is greater than the average task success rate of all users on tasks ϵ P. Task success is predicted using the method in [18].

Unlike the location and entry-point cohorts, the expertise cohort does not use information about the current user. The intuition here is that experts will select better resources and being pointed to those resources will help all users irrespective of expertise level. Later we show that there is benefit from leveraging particular user cohorts.

4.3 Summary

In this section we have defined methods for computing inter-task similarity, defined the feature generation procedure and the particular features that are assigned to the URLs, and defined the groups from which similar tasks are drawn. We also described each of the cohorts that we investigate. In the next section we describe our experiments to measure the effectiveness of task-based models for personalization, including comparisons with personalization methods and query-based (not task-based) similarity.

5. EXPERIMENTS

Our log-based evaluation method focuses on a re-ranking task, assessing the extent to which the models promote clicked results.

5.1 Baselines

The original ranking from the Bing search engine is our primary baseline. We also setup competitive query-centric baselines:

1. **Query-based Global (QG; same query, all users):** This is a non-personalized approach that finds clicked URLs by matching the current query against previous queries over all users.

[1] http://www-cs-faculty.stanford.edu/~eroberts/cs181/projects/firefox-market-dynamics/present.html