Designing Human-Readable User Profiles for Search Evaluation

Carsten Eickhoff[1], Kevyn Collins-Thompson[2], Paul Bennett[2], and Susan Dumais[2]

[1] Delft University of Technology, Delft, The Netherlands,
c.eickhoff@tudelft.nl
[2] Microsoft Research, Redmond, USA,
{kevynct,pauben,sdumais}@microsoft.com

Abstract. Forming an accurate mental model of a user is crucial for the qualitative design and evaluation steps of many information-centric applications such as web search, content recommendation, or advertising. This process can often be time-consuming as search and interaction histories become verbose. In this work, we present and analyze the usefulness of concise human-readable user profiles in order to enhance system tuning and evaluation by means of user studies.

1 Introduction

The value of information has been long argued to depend on the individual preferences and context of each person [3]. To account for this, state-of-the-art information services may rely heavily on personalisation techniques in order to incorporate knowledge about the user into the retrieval process [7]. Such user-centric applications are often evaluated quantitatively by means of large-scale query log analyses, trying to maximise ranking quality expressed by a number of performance scores. However, especially in early design stages, manual qualitative analysis of search rankings is often crucial for obtaining high-quality data for training and evaluation. Ideally, the actual users who are being targeted for personalization would make the judgments. In practice, however, individual users are rarely available for collaboration or discussion. Instead, the research community typically relies on external annotators who first need to form a mental image of the user before being able to judge the quality of personalised rankings. This step, however, can be difficult and time-consuming as it requires an in-depth inspection of the user's entire search and browsing history in order to accurately account for their interests and preferences.

In previous work, Amato et al. [1] use topical user modelling for content selection in digital libraries. Their profiles focus on users' preferences in a number of domains such as document content or structure. Nanas et al. [5] propose a hierarchical profile based on terms extracted from clicked documents. However, previous work has not deeply explored how to generate compact, human-readable user profile representations.

In this work, we present and analyze a means of summarizing a user's web search history into a compact, yet meaningful profile. Our profiles combine features that indicate topics of interest, representative queries, search context, and content complexity, to enable external judges to quickly form an accurate model of a user's interests and expertise. We apply our profiles in session judging tasks and analyze the correlation of profile features with inter-rater reliability and judging time.

2 Profile Design

Previous work in personalized search motivates the attributes to include in profiles (specific queries, general topics and content complexity), and work in human-computer interaction guides the presentation. Profiles include:

1. A user's interests can be summarized by a set of **topics** - but the topics must have clear and consistent definition, and not be too broad or too specific [1]. Additionally, the **most dominant** topics of a user's interests should be clearly recognisable.
2. Past **queries** should be included in order to provide concrete examples of common information needs [7].
3. The session **context** should be available in order to better understand the intention that motivated a sequence of queries [3].
4. User profiles should be **concise** in order to enable efficient work flows. Additionally, the variation in length between profiles should be limited in order to make the required work load predictable [6].
5. Content **complexity** has recently been shown to be a strong signal for search personalisation [4]. User profiles should reflect the general complexity of content consumed by the user.
6. **Consistency** in how profiles and sessions are shown enables more efficient processing [6].

We aimed to accommodate all of these considerations into the design of our user profile representation. Figure 1 shows an example of the resulting user profile. To obtain topics, we classify each clicked web search result into topical categories based on the Open Directory project hierarchy (ODP), as described by [2]. We use categories at the second level of the ODP tree (e.g. Science/Biology, Computers/Hardware) since this provides a consistent, sufficient level of specificity. A profile consists of one line per frequently-observed topic in the user's previous search history. We include each category that accounts for at least 5% of the overall amount of clicked pages. In this way, we ensure all profiles have a predictable length of 1-20 lines of text, regardless of how active the user was in the past. For each topic, we also show the 3 most frequent previously issued queries associated with that topic. To assign a topic to a query, we aggregate the topical classification of all clicked search results for that query. For example, for the query "Apple", if a user visited two pages classified as "Computers/Hardware", we would assign that topic to the query. We then display the queries that were most frequently associated with that topic in order to represent typical search patterns given a user and a topic. To further help the annotator form a model of the searcher, all queries are formatted as hyperlinks leading to the search engine result page for that particular query so that the annotator can see the topical spread of results. Finally, we include an estimate of the complexity of textual content in the form of a heat map of resource reading level. We estimate the reading level for each clicked result on a 12-point scale according to [4] and average the scores of clicked results for each query. We then highlight the query in green if the average reading level is less than or equal to 4, in red if the estimate is greater or equal to 9, and in blue if it is between these two levels. The resulting profiles have the added benefit that they can be applied to any profiling duration, ranging from a single query to months of search activity. This ensures conceptual conformity when, for example, comparing a single session with an extended period of previous activity.

55%	Sports/Soccer ("Messi vs Ronaldo", "real madrid wiki", "soccer odds")
14%	Recreation/Outdoors ("alps hiking", "REI store", "camp site protection")
8%	Business/Real Estate ("rent DC", "tenant rights DC", "craigslist DC")
5%	Health/Fitness ("60 day abs workout", "low fat diet", "nutrition table")

Fig. 1. An example of a condensed topical user profile.

3 Experimentation

We used the concise profiles we developed for assessing how typical an anonymized user session was with respect to that user's historical activity. Each assessment unit consisted of a compact profile (as in Fig. 1), followed by the list of queries comprising a search session generated by that user. A set of 100 sessions was sampled from anonymized logs from Microsoft Bing gathered during January 2012. To reduce variability in search behavior due to geographic and linguistic factors, we included only log entries generated in the English-speaking US locale. Three expert judges each evaluated all 100 sessions, making a 'typicality' judgment for each session on a five-point scale, with '1' being highly atypical for a user, and '5' being 'highly typical'. The degree of agreement between the three judges was computed using the variance across the typicality judgments. The time that each assessor took to judge each session was also recorded.

We computed several profile-based features for each assessed session (left column in Fig. 1): the number of queries in a given session (sessionQueryCount); the entropy of the profile's topic distribution (userProfileEntropy); and five similarity features based on query overlap (both whole query, and query terms): full user history vs. session (overlapH-S, overlapH-S-Terms), summary user profile vs. session (overlapP-S, overlapP-S-Terms), and summary user profile vs. full user history, filtered by session (overlapP-H-Terms).

Table 1 summarizes the Spearman rank correlations observed between these profile features and judging features. All overlap features had positive correlation with average typicality rating, the highest being profile-session overlap using query terms (overlapP-S-Terms, +0.39). In addition increasing the profile-session query overlap improved interrater agreement (overlapP-S-Terms is positively correlated with interrater agreement +0.24). High-overlap sessions were evaluated faster (-0.24 correlation of overlapP-S-Terms vs. time). In general, user profile-based features had a stronger influence on typicality scores and rating efficiency than their counterparts based on the full history.

We also found that sessions from highly-focused users, whose profiles were dominated by just a few topics (low userProfileEntropy) were evaluated faster, with higher typicality scores and agreement. That is, the entropy of a user's profile was positively correlated with time spent judging (+0.25), negatively

Profile features	Judging features		
	Typicality Average	Typicality Agreement	Average Time Spent Judging
overlapH-S	+0.10	+0.09	-0.14
overlapH-S-Terms	+0.32	+0.28	-0.16
overlapP-S	+0.24	+0.10	-0.17
overlapP-S-Terms	+0.39	+0.24	-0.24
overlapP-H	+0.37	+0.24	-0.19
sessionQueryCount	-0.07	-0.10	+0.41
userProfileEntropy	-0.29	-0.30	+0.25

Table 1. Spearman rank correlation of user profile/session features (rows) with judging features (columns). Judging features included (L to R) average typicality score, agreement on typicality, and average time to judge.

correlated with interrater agreement (-0.30), and negatively correlated with typicality (-0.29). Perhaps not surprisingly, the number of queries in a session (sessionQueryCount) was positively correlated (+0.41) with time spent judging.

4 Conclusion

In this work, we introduced a novel way of representing searchers' previous search history in the form of concise human-readable topical profiles. Benefits of the representation include its brevity and conformity across different time ranges while retaining comparable descriptive power to the information offered in the full log files in our typicality assessment task. In the future, we would like to focus on a stronger integration of interaction information from the original sessions, e.g., by offering a detail view on which clicked results, click order and dwell times are available to assessors. It would also be interesting to investigate our method's applicability in different domains, such as the manual evaluation of personalization performance.

References

1. G. Amato and U. Straccia. User profile modeling and applications to digital libraries. *Research and Advanced Technology for Digital Libraries*, 1999.
2. P.N. Bennett, K. Svore, and S.T. Dumais. Classification-enhanced ranking. In *WWW 2010*.
3. P. Borlund and P. Ingwersen. Measures of relative relevance and ranked half-life: performance indicators for interactive IR. In *SIGIR 1998*.
4. K. Collins-Thompson, P.N. Bennett, R.W. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In *CIKM 2011*.
5. N. Nanas, V. Uren, and A. De Roeck. Building and applying a concept hierarchy representation of a user profile. In *SIGIR 2003*.
6. S.B. Shneiderman and C. Plaisant. *Designing the user interface 4 th edition*. Pearson Addison Wesley, USA, 2005.
7. J. Teevan, S.T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR 2005*.