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 designed for application at industry scale 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 interactions is often crucial for obtaining high-quality data for training and evaluation. Ideally, such an investigation would involve the actual users who are being targeted for personalisation. In practice, however, they 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 original 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 account for their preferences as accurately as possible.

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. While our profiling method combines similar types of information, so far, related work has not deeply explored human-readable user profile representations.

In this work, we present and analyze a means of summarizing a user's (verbose) 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 mental image of a user's interests and expertise. We apply our profile in an online session judging task and analyze the interaction of profile features with interrater reliability and judging time. Other potential areas of application include lab-based studies or crowdsourcing campaigns but might extend to settings such as marketing or advertising.

2 Profile Design

Previous work motivates a number of profile criteria to enhance effective and efficient processing by annotators:

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 reflect examples of concrete information needs in the form in which they were originally issued [7].
3. Search **context** should be available in order to better understand the intention that drove a given session [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. Recently, content **complexity** has 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. **Conformity** in how profiles and sessions are shown helps limit context changes, which ultimately results in more efficient processing [6].

We aimed to accommodate all of the above into the design of our user profile representation. We will refer to the indices of the above requirements as we address our way of incorporating each of them. Figure 1 shows an example of the resulting user profile format. 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 **(1)**. A profile consists of one line per frequently-observed topic in the user's previous search history **(2)**. For each topic, we show a set of n most representative previously issued queries **(3)**. To do this, we aggregate the topical classification output of all clicked search results at the query level. For example, the query "Apple", for which the user visited two pages classified as "Computers/Hardware", would be assigned the same label. In this way, we capture the user's intention behind a query and account for different usage of the same literal query **(4)**. Among the queries affiliated to each category, we display those that were issued most frequently in order to represent typical search patterns given a user and a topic. To further help the annotator form their mental image of the searcher and their original intent, all queries are formatted as hyperlinks leading to a web search engine' result page for that particular query. In this way, the annotators can get an impression of the topical spread within the subset of relevant web documents. 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 **(5)**. Finally, we include an estimate of textual content complexity in the form of a heat map of resource reading level **(6)**. We estimate the reading level per clicked page on a 12-point scale according to [4] and average scores per shown query. We then highlight the query in green if the average reading level is less than or equal to 4, or in red, if the estimate greater or equal to 9. The resulting profiles have the added benefit that they can be applied to any scope of 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 **(7)**.

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 applied our concise profiles to an annotation task: assessing how typical an information need was, as expressed in an anonymized user's search session, 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 a proprietary dataset, consisting of 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 variance across label values. The time (in milliseconds) that each assessor took to judge each session was also recorded.

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

Table 1 summarizes the rank correlations observed between the above profile-based features and judging features. All overlap features had positive correlation with average typicality rating, the highest being session-profile query term overlap (overlapUPQ-S-Terms, +0.39). We hypothesized that increasing the profile-session query overlap would improve interrater agreement. Indeed, the degree of

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
overlapUPQ-S	+0.24	+0.10	-0.17
overlapUPQ-S-Terms	+0.39	+0.24	-0.24
overlapUPQ-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.

session-profile query overlap (overlapUPQ-S) is positively correlated with interrater agreement, especially for term-based overlap (+0.24). Such high-overlap sessions also tended to be evaluated faster (-0.24 correlation of overlapUPQ-S-Terms vs. time). To sum up, user profile-based features were generally observed to have 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 (lower topic entropy) were typically able to be evaluated faster, with higher agreement and typicality scores. That is, the entropy of a user's profile was positively correlated with time spent judging (+0.25), negatively correlated with interrater agreement (-0.30), and negatively correlated with typicality (-0.29). Perhaps not surprisingly, session query count was positively correlated (+0.41) with time spent judging.

4 Conclusion

In this work, we introduced a novel way of representing searchers' previous activity history in the form of concise human-readable topical profiles. The key benefits of the method include its brevity and conformity across different time ranges while retaining comparable descriptive power to the information offered in the full log files. Our analysis is based on real-world usage data from Microsoft Bing. 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 recommender systems or for representing high-level summaries of the skills and qualifications in professional resumes or employee records.

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. B. Shneiderman and S. Ben. *Designing the user interface*. Pearson, 1998.
7. J. Teevan, S.T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR 2005*.