Developing Evaluation Metrics for Instant Search Using Mixed Methods

Praveen Chandar, Jean Garcia-Gathright, Christine Hosey, Brian St. Thomas, Jennifer Thom
Spotify
Boston, MA, USA
praveenr|jean|chosey|brianstt|jennthom@spotify.com

ABSTRACT

Instant search has become a popular search paradigm in which users are shown a new result page in response to every keystroke triggered. Over recent years, the paradigm has been widely adopted in several domains including personal email search, e-commerce, and music search. However, the topic of evaluation and metrics for such systems has been less explored in the literature thus far.

In this work, we describe a mixed methods approach to understanding user expectations and evaluating an instant search system in the context of music search. Our methodology involves conducting a set of user interviews to gain a qualitative understanding of users' behaviors and their expectations. The hypotheses from user research are then extended and verified by a large-scale quantitative analysis of interaction logs. Using music search as a lens, we show that researchers and practitioners can interpret the behavior logs more effectively when accompanied by insights from qualitative research. We demonstrate that metrics identified using our approach are more sensitive than the commonly used click-through rate metric for instant search.

ACM Reference Format:
Praveen Chandar, Jean Garcia-Gathright, Christine Hosey, Brian St. Thomas, Jennifer Thom. 2019. Developing Evaluation Metrics for Instant Search Using Mixed Methods. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331293

1 INTRODUCTION

In 2010, Google introduced the *Instant Search* feature that updates the results displayed as users type in the search box. Since then, popular domain-specific search applications including LinkedIn [9], Kayak, and Spotify have adopted the feature, as it promises to save keystrokes and time while users are searching. Over the years, at least three different variants have been deployed in commercial search applications: (1) *Query Autocomplete*: the most popular variant in which users are provided with completions to their query as they type; (2) *Instant Result Search* where the entire result page updates for each keystroke; and, (3) *Hybrid Approach* which is a combination of query autocomplete and instant result search.

Several studies have been proposed to address various challenges in instant search, including efficient indexing strategies, and personalizing suggestions [2]. However, few studies in the past have focused on evaluation of instant search systems. Metrics such as mean reciprocal rank and success rate are commonly used in offline evaluations. Kharitonov et al. [8] introduced metrics that were inspired by the cascade family of user-model. Hofmann et al. [4] conducted an eye-tracking study to analyze users' interactions while using a query autocomplete system. While these studies are valuable in developing metrics for query autocomplete in the context of web search, they fail to explicitly incorporate the user's perceptions of a satisfactory search experience, particularly with respect to user goals and expectations in a domain-specific setting.

In this work, we take a holistic approach to developing metrics for domain-specific instant search applications by relying on insights from qualitative and quantitative studies. While mixed methods evaluation has been employed in the information retrieval community for some time [7], here we restrict our focus to *instant result search* in the context of music search. Our approach begins with a qualitative study in which users of a large music streaming platform are interviewed with the goal of understanding their needs, expectations, and behavior while using the search feature. The interviews followed a funnel structure that began with discussing users' broad interests and goals, then narrowed down to specific behaviors that could be tied to user satisfaction. The study enabled us to gain a deeper understanding of various patterns in user behavior, which was later verified and validated in both a large-scale analysis of logged user behavior and in an online experiment.

We show that augmenting semi-structured interviews with large-scale log analysis is an effective way to build metrics, especially for domain-specific search where user goals can be unclear. Our contribution in this work is to provide valuable insights for developing metrics for domain-specific instant search using music search as an example.

The rest of the paper is organized as follows. In Section 2, we provide a detailed description of our interview methodology and summarize the key insights from the study. Next, we describe our log analysis to validate and generalize the findings from our qualitative study in Section 3. We validate the proposed metrics in Section 4 and finally, conclude in Section 5.

2 USER INTERVIEWS

Our iterative mixed methods approach to developing user satisfaction metrics begins with a strong foundation of qualitative user

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331293

research that augments and informs each successive step of quantitative metrics development. We then use these insights to form hypotheses about user satisfaction. These hypotheses serve as guides for how to proceed with quantitative analysis; the resulting insights are used to measure performance and optimize search systems.

In particular, we employ semi-structured interviews with users to gather deep insight into user behaviors. Details on the interview methodology and analysis can be found in [5]; we provide a brief summary in the following sections.

2.1 Cohort Selection

We relied on logged user behavior to create a data-informed set of four cohorts for our qualitative interview sample. The cohorts varied along two dimensions: subscription type (free vs. paid) and account age (<1 month on the platform vs. >3 months on the platform). We chose these dimensions to ensure that both subscription levels were represented in data collection as free users have certain restrictions (e.g., free users cannot play specific tracks on demand), and we hypothesized that search habits may evolve over tenure with the platform. Furthermore, to ensure our sample represented a range of search behaviors, we selected participants with varying search frequency (7 - 167 searches in the month prior to the session) and varying average length of time per search (2.5 - 35 seconds).

We recruited 14 participants from a large city in the Northeastern United States via email. Participants were users of our platform and ages ranged from 18 to 40 and included 5 females and 9 males. Participants received a $100 gift card as compensation.

2.2 Interview Structure and Data Analysis

We conducted 60-minute semi-structured interviews with each of the 14 participants individually. The structure of the interview followed a funnel structure, where we begin by discussing music listening more generally; we then narrowed the conversation to the search function on the platform. The search-focused interview protocol moved from broader to more specific as well. We asked participants about their attitudes, expectations, and preferences regarding search on the platform. We then asked participants to describe from memory how and why they typically used search. Next, we asked participants to view their recent search history on the app and walk us through several specific searches. The funnel structure allows participants to guide the substance of the interview without biasing them by the more specific questions.

We then asked participants to describe experiences with search that went well for them, and experiences that did not go well. As they described these experiences, we probed around specific actions they took within the app as a function of the quality of their experience. Finally, we completed the interviews with a deep dive into 31 specific interactions with search, and asked participants if they had ever performed that action while searching.

To analyze the interview data, we took an inductive approach based on thematic analysis [1]. Though there is an extensive literature on user motivations and experiences with search [6], we wanted to allow for new themes and insights to emerge bottom-up from the interview data, especially in the specific context of searching in app for music. We took repeated passes through each participant's video and corresponding transcript to document relationships between the codes within each individual. We used an iterative open-coding system to capture emergent and evolving codes and then identified and refined meaningful themes by looking at the relationships between the codes. From there, we defined the themes that we identified and organized them into a coherent set that fit together to capture the experiences of the participants.

The design of our interview and analysis methodology served to inform subsequent analysis of logged behaviors. Our interview protocols asked participants to recreate and recollect past actions, based on pure recall and cued recall, so that they could think about the mundane interactions that comprise their user experience. We conducted a deep dive on specific behaviors (e.g., the 31 possible search interactions within the app) during our interviews so that we could pair how people talk about specific behaviors and integrate how they described those behaviors in isolation with how they described (or omitted) them in context. The final analysis was guided by what is possible to observe in log data. For instance, we looked for what behaviors are ambiguous or unambiguous signals of positive or negative user experience, and recommended unambiguous signals for operationalizing metrics quantitatively.

2.3 Insights

2.3.1 Success and Effort. We observed that users described their experiences with search within our platform on the dimensions of *success* and *effort*. A good experience was one where participants could find the content they wanted, ideally with little effort. A bad experience was expressed as not finding the desired content, or struggling to find what they were looking for.

Success was a higher priority to participants for a good search experience. The ideal search experience, however, required low effort with less typing, minimal reading, and less clicking and scrolling. Importantly, effortful searches were not necessarily perceived by participants as unsuccessful. In those instances, effort is either habitual, based on prior interactions, or expected, based on a more open-ended information need. Participants expressed frustration with effort when they had to take more actions than expected to reach their goal.

2.3.2 User Goals. Interviews revealed four overarching goals for using search on the streaming platform: listen, organize, share and fact check. *Listening* to content was the most common goal reported by participants in the study. The next most common goal was to *organize* content where users employed search to find content to create collections so that they could access them more easily in the future. Much less frequently reported by participants was *sharing* their content with others by looking up music to pass along to friends. Finally, the rarest search goal that participants described was *fact checking* or gathering additional information about content, such as finding out song featured on a TV show or movie.

2.3.3 User Journey. From our interview data, we identified three phases of the search user journey. First, the user communicates their intent by *typing* a query. The content of the query is not necessarily a perfect representation of the intent, but rather, provides an ambiguous signal. For example, participants searched for an artist when their underlying intent was to navigate to an artist, track,

playlist, genre or era. More generally, typing behaviors were an indication of effort, though often an expected (and therefore acceptable) form of effort. Participants reported that going backward (i.e., backspace, delete string, toggle) felt worse than going forward (i.e., character entry).

Next, the user *considers* the results and evaluates what the system has shown. All behaviors (e.g. scrolling, clicking on items varied on position) on the result page are indicative of effort, again much of which is expected. Participants anchored on the top result and often ignored the rest of the results page, choosing to continue typing or reformulating if the top result was incorrect rather than scroll. Behaviors that took participants away from result page (e.g., to an artist page) were also indicators of effort. Again, going backward (i.e.back button click) felt worse to participants than going forward (i.e., page view, short streams, previews).

Finally, the user *decides* and ends the search session. This phase of the user journey provides us with signals that can be used to indicate search success in our platform across each user goal. If the user successfully found something to listen to, then streaming and adding to queue were behaviors that participants reported. When organizing content, behaviors such as following artists and playlists, saving to library, adding to playlist, and downloading after a search would allow participants to meet that particular goal. If the participant's goal was to share music, following a user profile or using the share function was a signal of search success. There was no clear signal of success reported for the rarest goal, fact checking.

In summary, interview data suggests that users of the search function evaluate search based on success and effort and have four main goals: listen, organize, share and fact check. The search user journey has three phases: type, consider and decide. Behaviors in the type and consider phases of the user journey are indicative of effort and going backwards feels more effortful. The behaviors in the decide phase of the search user journey provide indicators of success for each of the goals.

3 LARGE-SCALE LOG ANALYSIS

Query logs are an extremely valuable resource for understanding user behavior; this topic has been heavily investigated in the context of web search. However, interpreting the logs for domain-specific search applications is tricky because user goals and expectations are often unclear. In this section, we incorporate insights obtained from user interviews to interpret the logs collected in the context of music search. Further, we validate and verify the hypotheses from Section 2.

3.1 Dataset

In order to analyze user behavior in an instant search session, we logged the search result pages returned for each keystroke along with users' interactions (such as clicks or taps) on the results. Query prefixes were grouped together into a session using a timeout threshold. We do not use a click on a search result to terminate a session because in our interface users are allowed to navigate back to the search result page and continue interacting with it.

We compiled a dataset by logging instant search sessions on a large music streaming platform over a period of two weeks in August 2018. The dataset consisted of 3,391,764 sessions from 331,980 users sampled randomly. Based on the insights from Section 2.3, we proposed a set of behavioral signals to be used as metrics, corresponding to the three phases of user journey: type, consider and decide. The signals we identified for each of the phases are listed below [1].

- ***Type Phase***
  - number of keystrokes
  - number of deletions
- ***Consider Phase***
  - clicks on results
  - clicked rank position (as a proxy for scrolling)
- ***Decide Phase***
  - number of streams
  - follow playlist/artist
  - add to personal collection or listening queue

3.2 Analysis

To gain a high-level understanding of the behavioral signals enumerated above and validate our logging infrastructure, we extracted various descriptive statistics. We observed that users' attention was mostly focused on the top result in the ranked list; about 80-85% of the clicks occurred on the top result. Users typically input 6 characters (i.e., average prefix length was around 6) before they either clicked or cleared the search box. Not surprisingly, these results align with studies on query autocomplete for web search [4].

In order to gain insights about the relationship between the proposed behavioral signals and validate the major themes that emerged from our qualitative study, we carried out a principal components analysis (PCA). PCA yields correlation patterns, allowing us to compare the correlation structure in log data to qualitative themes. Each instant search session described in Section 3.1 was augmented with the behavioral signals, producing scores aggregated for each user. PCA was conducted on the resultant $n \times m$ matrix, where n is the number of users and m is the number of behavioral signals considered.

Figure 1 shows the results of our analysis. The first two principal components explain 32.8% of the variance in the data set. We find that the first principal component separates the success and effort metrics very distinctly. Secondly we see that the second principal component separates the effort events into two groups. At one end of the second principal component, there are the number of deletions and number of keystrokes, while the number of clicks and click position is clustered together at the other end. Session duration and number of clicks end up between success and effort metrics. This suggests that these signals could provide mixed signals, i.e., more clicks might not always mean increased satisfaction.

This result is particularly interesting, because they are in line with two of the three major themes observed in our qualitative study — (1) users evaluate their experience in terms of success and effort; (2) user journey can be described through three phases: type, consider, and decide. Further, we observe that among the effort signals, there is clear separation between the typing and backwards effort (through deletions), and the forward clicking effort.

[1] We considered back button clicks as a metric for the *consider* phase, but these were not available for this analysis due to a logging issue.

Confirming this structure makes us more confident in both the qualitative findings and their generalizability.

In summary, combining the insights from user interviews and log analysis, we identify the following behavior to indicate *success*: stream, add to playlist, save to collection, follow artist, and follow playlist. Furthermore, we identified number of keystrokes, number of deletions, and click position to indicate *user effort*.

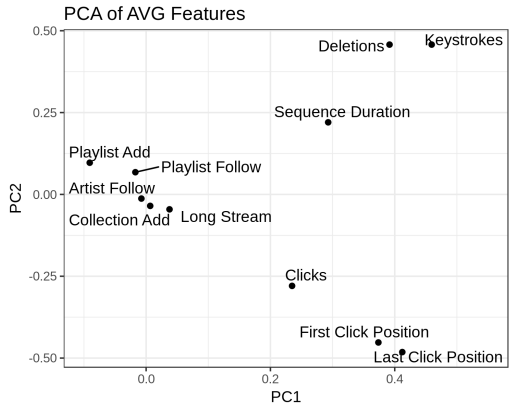

Figure 1: The first two principal components of a Principal Components Analysis. User behavior on these two directions cluster in a way that is consistent with the two main themes from thematic analysis. These components explain 32.8% of the variance.

4 METRICS VALIDATION

In this section, we demonstrate the effectiveness of our proposed metrics by comparing a composite metric to click-through rate and show that our proposed metric is more sensitive. We evaluate the two metrics based on directionality and sensitivity, two key qualities of a good metric for online experimentation [3]. We curated a dataset from a set of previously run online experiments where the treatment resulted in a positive user experience for the users; we relied on various engagement metrics to determine if the treatment effect was positive. The dataset consisted of 6 different experiments in which the treatment introduced changes to the search feature on the platform.

The composite metric combines all the success related behavior signals identified in Section 3 into a single binary metric (the sessions was considered successful if any success metric was non-zero). We refer to this metric as *Success Rate*. Figure 2 shows the difference in mean between control and treatment for a 14-day time period as measured by success rate and click-through rate. Both metrics agree on the directionality (i.e., they detect a positive treatment effect), but clearly, success rate is more sensitive than click-through rate. Finally, to verify that our metric is not overly sensitive, we repeated the procedure on an A/A experiment where control and treatment are the same. As expected, we found no significant difference between control and treatment.

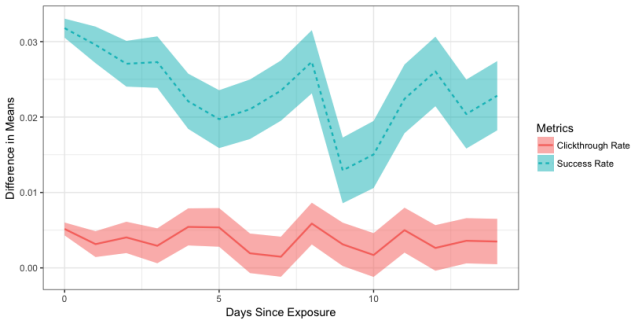

Figure 2: The outcome of an A/B test plotted by the difference in absolute mean between treatment and control as measured by *Success Rate* (dashed blue line) and click-through rate (solid red line) over a period of two weeks. Positive values indicate that the treatment outperforms control. Error bars indicate 99% confidence interval.

5 DISCUSSION AND CONCLUSION

Developing reliable evaluation metrics is essential for improving search applications. In this paper, we took a holistic approach to evaluate an instant search system in the context of music search. We learned that users' journey in instant search can be characterized by three phases: *type*, *consider*, and *decide*. And, users primarily describe their experiences along two dimensions: *success* and *effort*. We identified metrics in our logs that capture the above characteristics of user behavior and validated them using principal component analysis on millions of logged search sessions. Further, we demonstrated the effectiveness of our proposed metric in terms of directionality and sensitivity on online experiments.

Measuring user satisfaction gets increasingly complex in sophisticated interactive IR systems such as instant search. As observed by various previous studies [7], we found that qualitative studies was extremely effective for understanding the end-to-end experience of a user in a search session and these insights made it easier to interpret logged interaction data. For instance, our interpretation of the PCA analysis would have been uninformed with respect to the user experience without the insights from user interviews. We believe the approach presented in this work would benefit metric development for search applications in other domains.

REFERENCES

[1] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.
[2] F. Cai and M. de Rijke. *A Survey of Query Auto Completion in Information Retrieval*. Now Publishers Inc., Hanover, MA, USA, 2016.
[3] A. Deng and X. Shi. Data-driven metric development for online controlled experiments: Seven lessons learned. In *Proceedings of KDD '16*, pages 77–86, 2016.
[4] K. Hofmann, B. Mitra, F. Radlinski, and M. Shokouhi. An eye-tracking study of user interactions with query auto completion. 2014.
[5] C. Hosey, L. Vujović, B. St. Thomas, J. Garcia-Gathright, and J. Thom. Just give me what i want: How people use and evaluate music search. In *Proc. of SIGCHI '19*.
[6] P. Ingwersen. Interactive info. seeking, behaviour and retrieval. *Journal of the American Society for Information Science and Technology*, 63(10):2122–2125, 2012.
[7] D. Kelly et al. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, 3(1–2):1–224, 2009.
[8] E. Kharitonov, C. Macdonald, P. Serdyukov, and I. Ounis. User model-based metrics for offline query suggestion evaluation. In *Proceedings of SIGIR '13*, pages 633–642.
[9] G. Venkataraman, A. Lad, L. Guo, and S. Sinha. Fast, lenient and accurate: Building personalized instant search experience at linkedin. In *IEEE Conf. on Big Data*, '16.