Impact of Domain and User's Learning Phase on Task and Session Identification in Smart Speaker Intelligent Assistants

Seyyed Hadi Hashemi*
University of Amsterdam
Amsterdam, The Netherlands
hashemi@uva.nl

Kyle Williams
Microsoft
Redmond, USA
Kyle.Williams@microsoft.com

Ahmed El Kholy
Microsoft
Redmond, USA
Ahmed.ElKholy@microsoft.com

Imed Zitouni
Microsoft
Redmond, USA
izitouni@microsoft.com

Paul A. Crook†
Facebook
Seattle, USA
pacrook@fb.com

ABSTRACT

Task and session identification is a key element of system evaluation and user behavior modeling in Intelligent Assistant (IA) systems. However, identifying task and sessions for IAs is challenging due to the multi-task nature of IAs and the differences in the ways they are used on different platforms, such as smart-phones, cars, and smart speakers. Furthermore, usage behavior may differ among users depending on their expertise with the system and the tasks they are interested in performing. In this study, we investigate how to identify tasks and sessions in IAs given these differences. To do this, we analyze data based on the interaction logs of two IAs integrated with smart-speakers. We fit Gaussian Mixture Models to estimate task and session boundaries and show how a model with 3 components models user interactivity time better than a model with 2 components. We then show how session boundaries differ for users depending on whether they are in a learning-phase or not. Finally, we study how user inter-activity times differs depending on the task that the user is trying to perform. Our findings show that there is no single task or session boundary that can be used for IA evaluation. Instead, these boundaries are influenced by the experience of the user and the task they are trying to perform. Our findings have implications for the study and evaluation of Intelligent Agent Systems.

ACM Reference Format:
Seyyed Hadi Hashemi, Kyle Williams, Ahmed El Kholy, Imed Zitouni, and Paul A. Crook. 2018. Impact of Domain and User's Learning Phase on Task and, Session Identification in Smart Speaker Intelligent Assistants. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3269206.3271803

Keywords: intelligent assistants; behavioral dynamics; user sessions; mixture models

1 INTRODUCTION

There is a growing interest in integrating Intelligent Assistant (IA) Systems in different devices with an aim of providing enriched experiences for users [3]. For instance, IAs such as Apple Siri, Google Now, Microsoft Cortana and Amazon Alexa have been integrated with Desktop computers, smart phones, and smart speakers. However, user behavior varies in different contexts [10, 11, 17, 34], like platform, input method, etc. For example, users can click on IA responses and change their view-port in interacting with an IA on smart-phones or desktops [19, 38], which is not available in smart speakers. Therefore, due to behavioral dynamics in interacting with IAs, their evaluation on different platforms is challenging, suggesting that different means of evaluation for different platforms may be necessary.

Understanding user behavior and evaluating user satisfaction in interacting with IAs on mobile phones and Desktop computers has previously been studied [15, 18–20, 25, 38, 39]; however, to our knowledge, there have been no studies investigating user satisfaction and IA effectiveness for smart speakers, which are becoming increasingly popular. For instance, one study found that there was a 128.9% increase in the number of smart speaker users in the United States in 2017 compared to 2016[1]. In this paper, we use the phrase smart speaker to refer to a wireless speaker device that integrates an intelligent assistant. For the purpose of this study, we focus on devices that have no screen and where the only method of communicating with the device is via voice.

Smart speakers can be used for many tasks, such as arranging meetings and controlling home devices via home-automation. This multi-task nature of smart speakers creates a multi-task experience for users, where a task refers to a single goal or information need that the user wishes to satisfy [14]. Furthermore, a series of tasks can be composed to form a session, which refers to a short period of contiguous time spent to fulfill one or multiple tasks [16]. Evaluating the satisfaction of users for tasks and sessions is a critical component of IA evaluation; however, it is not obvious how one should define task and session boundaries for IAs.

Identifying sessions based on user inactivity thresholds as a session timeout is the most common session identification approach in Information Retrieval (IR) [5, 8, 25, 33]. The basic idea is to define an inactivity window that can be used to separate sessions. The idea was first proposed by Catledge and Pitkow [4], in which they use client-side tracking to examine browsing behavior. They reported the mean time between logged events is 9.3 minutes and, by choosing to add 1.5 standard deviation to the mean, they proposed

*Work done while interning at Microsoft.
†Work done while at Microsoft.

[1] https://www.emarketer.com/Article/Alexa-Say-What-Voice-Enabled-Speaker-Usage-Grow-Nearly-130-This-Year/1015812

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3271803

a 25.5 minutes inactivity threshold. Over time, this threshold has smoothed out to 30 minutes. Recently, Halfaker et al. [9] proposed a session identification approach by fitting a mixture of Gaussians and reported 1 hour as an inter-activity time threshold as session boundary being appropriate for most user initiated actions. User inter-activity time is the time difference between two consequent user actions in interacting with an information system. An extension of this work for IAs was presented in [23], where it was shown that the session boundary for an IA on a Desktop Computer was 2 minutes. The experiment was also repeated for Web search and shown to be 24.1 minutes. The differences between these three studies suggests that there is no single session boundary that is applicable across platforms.

Furthermore, previous research has considered the session boundary as a fixed threshold for all IA users. However, in this study, we show that there is no single approach to modeling task and session boundaries. Instead, task and session boundaries are affected by contextual factors such as a user's experience in using the system and the task they are trying to accomplish. Specifically, the multi-task nature of IAs leads to different types of user experiences compared to traditional IR systems. Furthermore, there is often a learning curve associated with being new to an IA. In addition to this, tasks related to some IA domains require a longer time to be fulfilled compared to other domains. Therefore, using a single task and session boundary cut-off over all domains and users expertise levels is not ideal for evaluation.

In this paper, we study the impact that learning curves and usage domains have on task and session boundary cutoffs. Specifically, we jointly identify task and session boundary by fitting a 3-component Gaussian Mixture Model (GMM) on users inter-activity times in interacting with smart speakers. We focus on smart speakers as they have not been studied before and, as previously mentioned, it is expected that user behavior will differ from that of other platforms. However, our findings are applicable to other platforms as well.

In particular, our main aim is to study the question: **What is the impact of the learning curve and task domain on task and session boundaries when interacting with intelligent assistants?** Specifically, we answer the following research questions:

(1) *How does one effectively measure task and session boundary cut offs in intelligent assistant systems?*

(2) *Do user learning curves have an impact on session boundary cut-offs?*

(3) *What is the impact of the domain on task and session boundary cut-offs?*

Our contributions include: (1) applying an unsupervised approach using a Gaussian Mixture Model (GMM) with 3 components to jointly identify task and session boundary cut-offs; (2) a detailed study of the impact of the learning curve on task and session boundary cut-offs; (3) an analysis of the impact of usage domain on inactivity thresholds for task and session identifications.

In making these contributions, the rest of the paper is organized as follows. In Section 2, we review related work on task and session boundary identification. The session boundary cut-off estimation based on a GMM is described in Section 3. Then, we thoroughly analyze the impact of the learning curve and domain on task and session boundary cut-off in Section 4 and 5. Finally, we present conclusions and future work in Section 6.

2 RELATED WORK

User session have been extensively used in IR to develop metrics for web analytics and user behavioral understanding. To create sessions, three main group of approaches have been used in the literature, namely, navigation-oriented, query-refinement oriented and time-oriented approaches.

Navigation-oriented approaches take advantage of browsing patterns based on HTTP referrers and URLs associated with each request. Cooley et al. [5] proposed an approach to identify sessions, which is based on detecting the start and end of a session based on navigation behavior of users. The beginning of a navigation behavior (without a referrer) shows the start of a session and the end of a session is a point that the navigational trail can not be traced to a previous request.

Although navigation-oriented approaches are effective in identifying task (addressing a single information need) [29], the complexity of this approach and its developmental focus on tasks over sessions makes them inadequate for session identification [9].

Session identification based on query refinements has also been shown to be only effective in identifying single information need sessions (i.e., task in our definition) [1, 14, 30–32]. Specifically, Jansen et al. [14] defined a session as "a series of interactions by the user toward addressing a single information need", which is very similar to the definition of task in our study, which we discuss in more detail in Section 3. Jansen et al. showed that the query content is a better signal in identifying tasks compared to a session boundary based on a time-oriented approach.

He et al. [1] and Ozmutlu et al. [30, 31] proposed a task identification approach based on detecting topic shifts using lexical query reformulations. Moreover, Radlinski and Joachims [32] proposed an approach to identify the topic relevance of a sequence of queries, which is effective for task identification. In a same line of research, Li et al.[22] combined topic models with Hawkes processes to identify search tasks. Furthermore, a Bayesian model for extracting hierarchies of search tasks and sub-tasks is proposed [24]. However, their extensive focus on tracing user queries in order to determine if they address a single information need limits the use in identifying sessions.

Time-oriented session identification approaches are based on estimating an inactivity threshold between logged user interactions. If there is a long period of inactivity between a user's activities, it is likely the user is no longer active, which leads to ending the session and creating a new session when the user returns. The time-oriented session identification was first proposed by Catledge and Pitkow [4], in which they use client-side tracking to examine browsing behavior. They reported 25.5 minutes inactivity threshold as the session boundary, which has been smoothed out to 30 minutes over time and is the value commonly used in the literature [8, 33].

Although the time-oriented approach has been widely used for session identification, some studies have criticized the effectiveness of the time-oriented approach in identifying sessions [16, 26, 28]. Jones and Klinker [16] proposed a supervised approach for automated segmentation of users' query streams into hierarchical units of search goals and missions and reported that the 25.5 minutes threshold is not effective and performs "no better than random" in identifying search tasks. However, they also reported that the time-oriented approach is more effective for session identification compared to task identification.

On the other hand, Halfaker et al. [9] proposed a session identification approach based on a GMM modeled to fit the within-session and between-session user inter-activity times. In contrast to Jones and Klinker [16], Halfaker et al. [9] showed that the global inactivity threshold is an effective session identification approach and reported 1 hour as an inter-activity time threshold, which is appropriate for most user initiated actions. The main disagreement between these two studies is on task identification, for which Jones and Klinker [16] criticize time-oriented approaches as being ineffective, but not session identification. We adopt an approach similar to Halfaker et al. [9] to jointly estimate task and session boundaries using a mixture of Gaussians fit on users inter-activity times.

Recently, Mehrotra et al. [23] applied a 2-component GMM to estimate session boundary in IAs. The authors showed that the session boundary in Microsoft Cortana on Desktop is much shorter than the common 30 minutes session boundary cut-off in traditional search engines. Our work is similar to the cited work in that we also fit a GMM; however, we show that there is no single appropriate fixed session boundary cut-off for IAs and that the session boundary is dependent on contextual factors, such as user expertise.

The research presented in this study is different from the other time-oriented session identification studies as it empirically shows that the task and session boundary cut-off is not static and fixed for all users. Specifically, in Section 4 and 5, we show how the user learning curve and task domains impact task and session boundary cut-offs in IAs.

3 SESSION BOUNDARY CUTOFF ESTIMATION

This section presents an unsupervised approach for task and session identification using GMMs in order to answer our first research question: *How does one effectively measure task and session boundary cut offs in intelligent assistant systems?*

3.1 Definitions

In IR, there are three common ways of defining sessions. A session may refer to: "(1) a set of queries to satisfy a single information need; (2) a series of successive queries; or (3) a short period of contiguous time spent querying and examining results." [9, 16] However, in search engine log analysis literature, it is common to use definition (1) as a task definition, in which a user performs a series of interactions to address a single information need [8, 14].

In IAs, users usually take a sequence of steps with an aim of achieving a goal to solve one or more tasks [19]. Since IAs have the ability to keep context from previous queries, this allows for task chaining where the context of one task can be used as input to the next. Considering the multi-task nature of the IA usage, we therefore define tasks and sessions as follows:

- **Task** is a single information need that can be satisfied by at least one query and one IA generated response.
- **Session** is a short period of contiguous time spent to fulfill one or multiple tasks.

Note, that our session definition is similar to the definition of sessions in [16]. Table 1 shows an example of a sequence of user's interactions to fulfill three tasks over the course of two sessions. In this example, using common time-based session boundaries in the IR literature [4, 9, 23] lead to 2 sessions. In the first session, the user is trying to complete two tasks: one for setting an appointment and one for sending a text message to someone. After completing these two tasks, the user had about 1 hour of inactivity before querying the IA for fulfilling the third task on controlling media, which leads to creating the second session.

3.2 Fitting Mixture of Gaussians

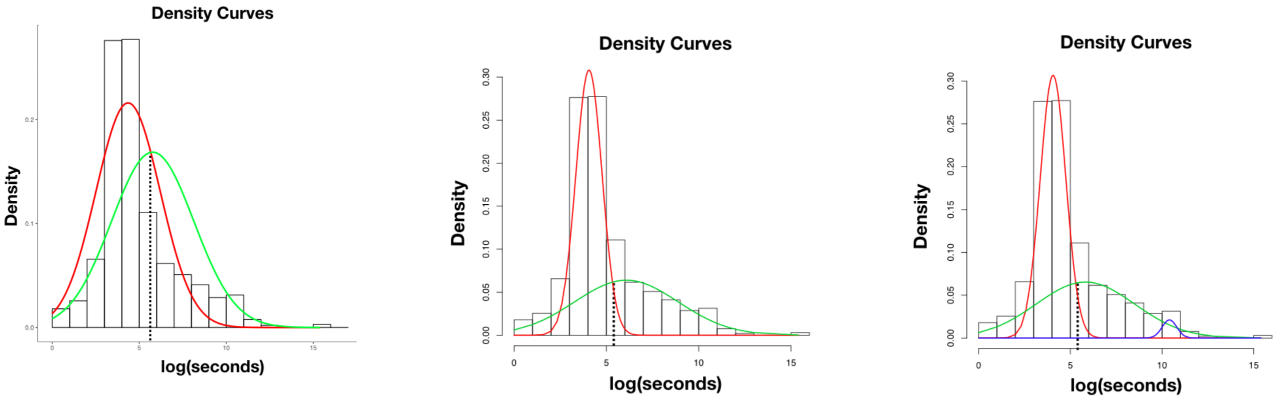
As previously mentioned, using a time-based threshold has been the most common approach in identifying search sessions [8, 33]. Halfaker et al. [9] proposed a methodology based on GMMs to identify clusters of user activities and argue that the regularity with which these activity clusters appear provide a good estimate of inactivity thresholds for defining sessions. More recently, Mehrotra et al. [23] showed that the 2-component GMM is an effective approach to identify sessions in interactions with IA. They also show that using a 2-component GMM leads to finding a 24.1 minutes session boundary for a well-known commercial search engine, which is close to findings of previous studies [8, 33]

Table 1: An example of a user's sessions in a smart speaker.

User Utterance and System Response	Timestamp
Session 1 Task 1: Calendar review and setting an appointment	
U1: What does my day look like tomorrow?	7/14/2017 7:28:45 PM
S1: You don't have anything scheduled for tomorrow.	
U2: Set an appointment with Mary at 14:30 for tomorrow.	7/14/2017 7:28:53 PM
S2: Alright, I can add appointment with Mary to your calendar for tomorrow at 14:30. Sound good?	
U3: Yes, thanks!	7/14/2017 7:28:58 PM
Task 2: Sending a text	
U4: Text Alex and ask if he is available for a short meeting tomorrow	7/14/2017 7:30:37 PM
S4: Sorry, I can't send messages here. Try the app on your phone or PC.	
U5: Send a message on skype to Alex and ask if he is available for a short meeting tomorrow	7/14/2017 7:31:03 PM
S5: Sending a message to Alex, skype	
Session 2 Task 3: Playing music and controlling media	
U6: Play jazz music on Spotify.	7/14/2017 8:38:05 PM
S6: And now, the playlist Coffee Table Jazz on Spotify.	
U7: Set the volume to 40%.	7/14/2017 8:38:17 PM
S7: I've set the volume to 40%.	

In this paper, we follow the same methodology that is based on GMMs [9, 23]. However, we focus on jointly identifying tasks and sessions by estimating task and session boundary cut-offs using a 3-component GMM. Jointly identifying task and session boundaries helps having a more accurate Gaussian fits on the inter-activity times of user interactions with the IA, and thus having a more accurate task and session boundary identification. We will detail 2- and 3-component GMM methodologies in the rest of this section.

In order to apply the GMM model to identify the inter-activity type component clusters, we pre-process the users interaction logs of IA usage in order to obtain per-user inter-activity times, which is essential to apply the GMM for identifying tasks and sessions. We plot a histogram based on logarithmically scaled inter-query times in seconds and look for evidence of one or two valleys. We follow Halfaker et al. [9] in using the visual inspection method to set the number of component clusters in the GMM. They proposed a visual inspection based on the number of observed valleys in the users inter-activity times as a better approach to define number of clusters compared to other statistical cluster separation measures like Davies-Bouldin Index (DBI) [6]. In Figure 2c, an example of observed valleys in users' inter-activity times histogram is shown by black arrows. After identifying the number of clusters, we fit a K-component GMM [2] on the logarithmically scaled inter-query times via Expectation Maximization. We fit both 2- and 3- component GMMs depending on what we observe in the histogram of inter-query times. In the next section, we describe the use of a 2-component GMM and then follow that with a discussion of when it is appropriate to use a 3-component GMM.



(a) Actual Gaussian distributions of users within-task and between-task inter-activity time

(b) EM fit of a 2-component GMM

(c) EM fit of a 3-component GMM

Figure 1: Task boundary cut off evaluation based on task boundary crowdsourced labels.

3.2.1 Fitting Mixture of Two Gaussians. The main assumption behind fitting a 2-component GMM for identifying sessions is that the inter-activity times of user interactions contains two component clusters: (1) within session inter-activity times (e.g., the time difference between user query U1 and U2 in Table 1); and (2) between session inter-activity times (e.g., the time difference between user query U5 and U6 in Table 1). If two clusters have been visually inspected, we fit the 2-component GMM on the logarithmically scaled inter-query times using following Expectation Maximization:

$$f(x, \theta) = \sum_{k=1}^{K} p_k N(x; m_k \sigma_k),$$

in which, $K = 2$ for a 2-component GMM and $N(x; m_k \sigma_k)$ is a Gaussian distribution with mean m_k and standard deviation σ_k. We follow [2, 9, 23] in parameter estimation using Expectation Maximization, where the goal is to maximize the likelihood function with respect to the mixing coefficients, the means and the covariances of the components as the parameters.

E Step: compute the expected values of the posterior probabilities for given parameter values as follows:

$$p^i(k|n) = \frac{p_k^i N(x; m_k \sigma_k)}{\sum_{k=1}^{K} p_k^i N(x; m_k \sigma_k)}$$

M Step: re-estimate the parameters based on the current posterior probabilities:

$$m_k^{i+1} = \frac{\sum_{n=1}^{N} p^i(k|n) x_n}{\sum_{n=1}^{N} p^i(k|n)}$$

$$\sigma_k^{i+1} = \sqrt{\frac{1}{D} \frac{\sum_{n=1}^{N} p^i(k|n) ||x_n - m_k^{i+1}||^2}{\sum_{n=1}^{N} p^i(k|n)}}$$

As previously mentioned, the assumption in fitting a 2-component GMM is that the components represent two main aspects: **within-session** and **between-session** interactivity times. Therefore, if there is an effective fit of the bimodal components, a good estimate of inter-activity time threshold for identifying sessions is the point where the inter-activity time is equally likely to be within the first Gaussian fit (within-session) and the second Gaussian fit (between-session). The reflection of the point on the x-axis, which is the estimate of task or session boundary, is shown by dotted lines in Figure 1a and other figures of K-component GMMs.

3.2.2 Fitting Mixture of Three Gaussians. In fitting a 2-component GMM, we assume that user interactivity occurs either **within-session** or **between sessions**. However, the multi-task nature of IA lends itself to three interactivity time behaviors. For instance, a user of an IA may perform a sequence of interactions to complete a task followed by a brief pause, which we refer to as **between-task** inter-activity time. They may then complete another task followed by a long period of inactivity. Therefore, we have three inter-activity periods: (1) **within-task** inter-activity times (e.g., the time difference between user query U1 and U2 in Table 1); (2) **between-task** inter-activity times (e.g., the time difference between user query U3 and U4 in Table 1); and (3) **between-session** inter-activity times (e.g., the time difference between user query U5 and U6 in Table 1), noting that the combination of (1) and (2) represent within-session inter-activity time since a session is made up of multiple tasks. Table 1 shows an example of a user's sessions having all the above three inter-activity time behaviors. The multi-task behavior of users in interacting with IA motivates us to fit a 3-component GMM on users inter-activity time with an aim of both task and session identification by modeling all the above users inter-activity time behaviors. Jointly modeling both task and session boundary leads to a better fit on users' inter-activity times compared to 2-component GMM, and thus better estimation of task and session boundaries.

Given an effective fit of the three component GMM, we can deduce the following: (1) an estimate of the inter-activity time threshold for identifying tasks is the point where the inter-activity time is equally likely to be within the first Gaussian fit (**within-task**) and the second Gaussian fit (**between-tasks**); (2) an estimate of inter-activity time threshold for session identification is the point where the inter-activity time is equally likely to be within the second Gaussian fit (**between-tasks**) and the third Gaussian fit (**between-sessions**). The reflection of these points on the x-axis is shown by dotted lines in 3-component GMM diagrams, which are estimations of task and session boundaries.

In comparing our approach, the 3-component GMM has been applied in the literature for boundary identification but with a different perspective and application. Halfaker et al. [9] applied a

3-component GMM to model a low-frequency cluster, which represents an extended break corresponding to a life-event with a mode of around 2.5 months. They also fit the 3-component GMM on interactivity times to have a better fit on the Movielens[2] dataset. They report that in addition to the within-session and between-session interactivity times, they observed an additional component cluster at a high-frequency intervals. They argue that the high-frequency intervals is due to a rapid rating behavior that the Movielens interface allows for. They have observed a similar high-frequency intervals in Movielens searches, for which they stated "we are less sure on how to explain the high frequency component of MovieLens searches. It could be that, unlike when performing a web search (AOL) or reading encyclopedic content (Wikimedia), users' movie searches are more likely to benefit from more rapid iteration".

In contrast to Halfaker et al. [9] interpretation, we have shown that the 3-component GMM fitted on user inter-activities is not always about modeling an additional low-frequency or high-frequency clusters to better fit within-session and between-session inter-activity time distributions. Instead, we propose that fitting a 3-component GMM enables us to jointly identifying task and session boundaries. In the most recent work on session identification in IAs [23], a 2-component GMM was used to fit inter-activity times and therfore was not able to fit the three inter-activity clusters that appear in IAs. To the best of our knowledge, this is the first study to fit a 3-component GMM to model IA user inter-activity times and, as will be shown in the rest of this section, is often more effective than 2-components models.

3.3 Evaluation

In this section, we first evaluate the effectiveness of GMMs in task boundary identification based on a crowdsourced labeled data. Then, using system-generated task boundary labels, we evaluate effectiveness of the 3-component GMM in identifying within-task distributions. All the experimental results of this section are based on interactions of users with all expertise development level and all available domains.

3.3.1 Evaluation Based on Crowdsourced Labels. To evaluate the effectiveness of the GMM in identifying tasks, we use a dataset of tasks from an IA on desktop computers where the task boundaries were collected through crowdsourcing [25]. The target was identifying the boundary of each task within a session.

To collect the task identification labels for a user session, crowdsource workers judged if the user was trying to find the same information as the previous query by issuing the current query [25]. They could read the user's query or listen to the user's utterances, read or listen to system response, look at the original timestamp of queries, and see a screenshot of a search result page if landing on a search engine result page from the IA. In order to obtain a high-quality task boundary labels, at least 5 crowdsource workers judged each session and the final label is based on a majority vote. The dataset contains 600 IA Desktop sessions, which are divided by judges into around 2000 tasks. Using the crowdsourced labeled data, we can plot the actual Gaussian distribution of within-task and between-task inter-activity times. Therefore, the intersection of the within-task and between-task distributions is the point where the inter-activity time is equally likely to be in either component and is therefore taken as the task boundary.

Figure 1a shows the within-task and between-task inter-activity time distributions based on the crowdsourced labels, in which the task boundary is $2^{5.5} \sim 45$ seconds. Furthermore, Figures 1b and 1c show the user 2- and 3-component GMM fits of inter-activity times based only on the inter-activity times and not any labeled data.

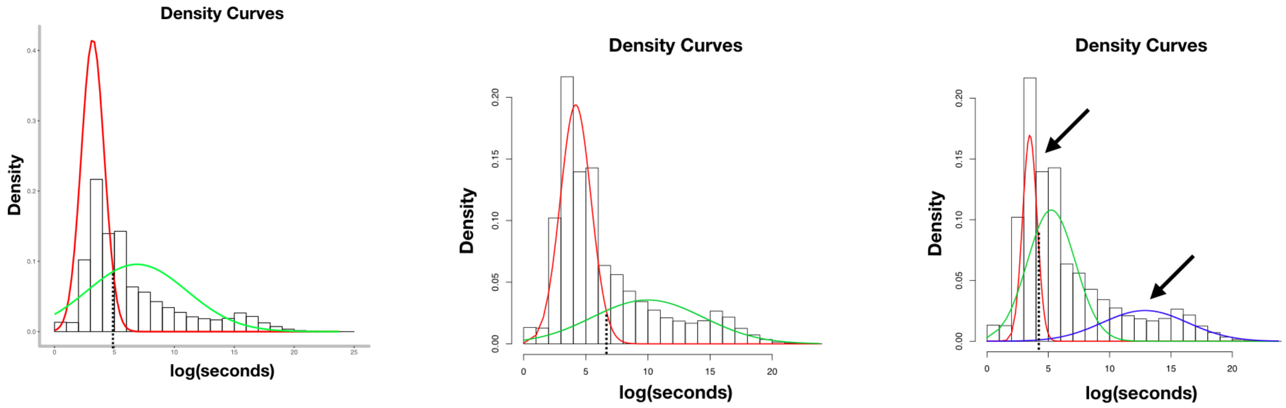
The experimental results show that the intersection point of within-task and between-task Gaussian fits for both the 2- and 3-components GMMs is $2^{5.4} \sim 42$ seconds. The task boundary estimation based on the GMMs is very close to the 45 seconds actual task boundary based on the labeled data. Therefore, in both the case of the 2- and 3-component GMMs, the data suggests that fitting a Gaussian provides a reasonable approach for modeling task boundaries. Note, the data used in this study was sampled at the session level [23]. Therefore we do not have multiple sessions per user, which makes it impossible to model between-session inter-activity times. This explains why the 2- and 3-component GMMs lead to the same task boundary since the GMMs only need to model the within-task and between-task components. The dataset used in the next section contains multiple sessions per user, which allows us to compare the effectiveness of the 2- and 3-component models.

3.3.2 Evaluation Based on System Task Boundary Labels. In the previous section, we showed the effectiveness of GMMs in identifying tasks. In this section, we evaluate the effectiveness of the 3-component GMM in fitting inter-activity times compared to the 2-component GMM. To achieve this, we make use of a high-quality task boundary classifier being used in a commercial IA. The commercial IA provides services for a variety of tasks and it can trace a user's interactions toward fulfilling a task, with the aim of identifying the task completion status, such as *completed*, *in-progress* and *canceled*. Using the task completion status, the IA can identify task boundaries of user interactions, which we call "system task boundary". As was the case with the crowdsourced data, we do not have access to a session completion status in the IA. However, we do have access to multiple sessions per user, which allows us to model the between-session inter-activity times. However, for our evaluation we only focus measuring within-task and between-task inter-activity times, which is available based on the system task boundary. Due to the fact that we do not have access to the system session boundary as it is defined in our paper, we do not evaluate the between-session inter-activity times.

To evaluate the 3-component GMM in modeling within-task and between-task inter-activity times, we sampled about 300K queries issued in a two months period of a commercial smart speaker usage. Figure 2a shows the IA identified within-task and between-task inter-activity time Gaussian distributions based on the system task boundary labels. The intersection of the within-task and between-tasks Gaussians based on the system task boundary leads to a $2^{5} \sim 30$ seconds boundary as the task boundary. Figures 2b and 2c show the fit of 2- and 3-component GMMs on the inter-activity times without using the system task boundary labels. The intersection point of the within-task and between-task distributions of the 2-components GMM leads to $2^{6.8} \sim 111$ seconds threshold as the task boundary. The intersection point of within-task and between-task distributions of the 3-components GMM estimates $2^{4.3} \sim 20$ seconds as task boundary cut-off. According to this result, the difference between 3-components GMM task boundary estimation and the system task boundary is 10 seconds, which shows the 3-component GMM based task boundary estimation is a more accurate approach to estimate task boundary cut-offs compared to the 2-component GMM based task boundary estimation with a 81 seconds difference.

We also measure the KL-divergence of the system task boundary labeled data distribution and the GMM fit via expectation maximization. KL-divergence is a similarity measure of two distributions and

[2] https://grouplens.org/datasets/movielens/



(a) Actual Gaussian distributions of users within-task and between-task inter-activity time

(b) EM fit of a 2-component GMM

(c) EM fit of a 3-component GMM

Figure 2: Task boundary cut off evaluation based on the system task boundary labels.

KL-divergence of two Gaussian distributions is given by [27]:

$$KL(p,q) = -\int p(x)\ log\ q(x)\ dx + \int p(x)\ log\ p(x)\ dx$$

$$= \frac{1}{2} log\ (2\pi\sigma_2^2) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}(1 + log 2\pi\sigma_1^2)$$

$$= log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2},$$

where, σ_1 and σ_2 are the standard deviations of the first and second Gaussian distributions, and μ_1 and μ_2 are the means of the first and seconds Gaussian distributions. A smaller KL-divergence value indicates a more similar Gaussian distributions. The KL-divergence of the system task boundary labeled within-task Gaussian from the within-task Gaussian fit of the 2- and 3-component mixture models are 0.4917 and 0.2544, respectively. According to this result, compared to the 2-component GMM, the 3-component GMM leads to a more effective and accurate within-task Gaussian distribution of user inter-activity times.

To summarize this section, we presented results of fitting 2- and 3-component GMM for task and session identification. The evaluation results show that the 3-component GMM leads to more accurate Gaussian fits of users inter-activity times and a more precise task boundary cut-offs compared to the 2-components GMM. Thus, in the rest of this paper, we use 3-component GMM to study the impact of domain and user learning-phase on the task and session boundaries. In the next section, we investigate the impact of learning curve on task and session boundaries by segmenting users by their levels of expertise.

4 IMPACT OF LEARNING-PHASE ON SESSION BOUNDARY CUTOFF

This section studies the impact of the learning-phase on session boundary cut-offs, aiming to answer our second research question: *Do user learning curves have an impact on session boundary cut-offs?* We begin by describing our data and then define the learning-phase in user behavior when interacting with an IA. We then discuss how session boundaries differ for the learning-phase and a so called normal usage phase.

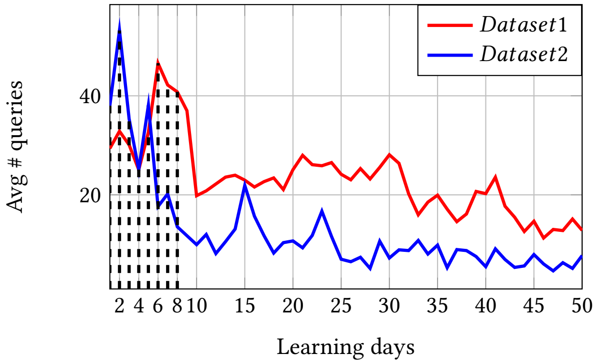


Figure 3: Impact of learning on average number of queries per day.

4.1 Experimental Data

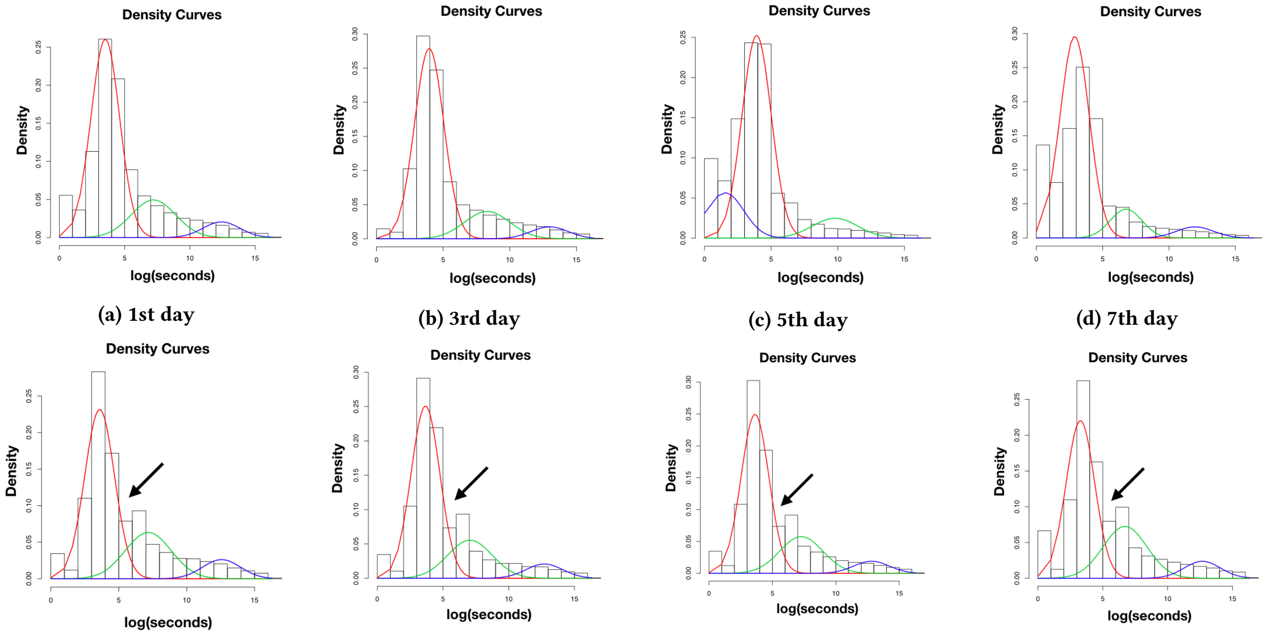
This study is based on two random samples of users interaction logs with two different commercial IAs being used on smart speakers. In the rest of the paper, we refer to them as Dataset 1 and Dataset 2. Both Dataset 1 and Dataset 2 are based on user interactions with the speaker from the first day they start using it. This enables us to evaluate the impact of learning-phase on session boundary cut-offs.

Dataset 1 consists of interaction logs of 2,087 users collected from March 2017 to September 2017. Users of Dataset 1 issued 731,128 queries in this period, which is 350 queries per user on average. The dataset has query timestamps and domain classifications of the queries.

Dataset 2 consists of interaction logs of 20 users with an average usage period of 264 days. The dataset includes 69,649 queries, which is 3,482 queries per user on average. Although the Dataset 2 has fewer users, the average number of queries per user is larger than Dataset 1. Furthermore, the queries span a larger time frame, which enables a longer term analysis.

4.2 learning-phase Definition

According to research in library search and search engines [7, 8, 13, 21, 36], domain expertise enhances search performance, and



(a) 1st day

(b) 3rd day

(c) 5th day

(d) 7th day

(e) 9th day

(f) 11th day

(g) 13th day

(h) 15th day

Figure 4: Dataset 1 users' interactivity time distribution from 1st day to 15th days of their usage.

the development of search expertise over time has been observed in prior studies [35–37]. Specifically, White et al [36] showed that non-expert domain expertise develop over time.

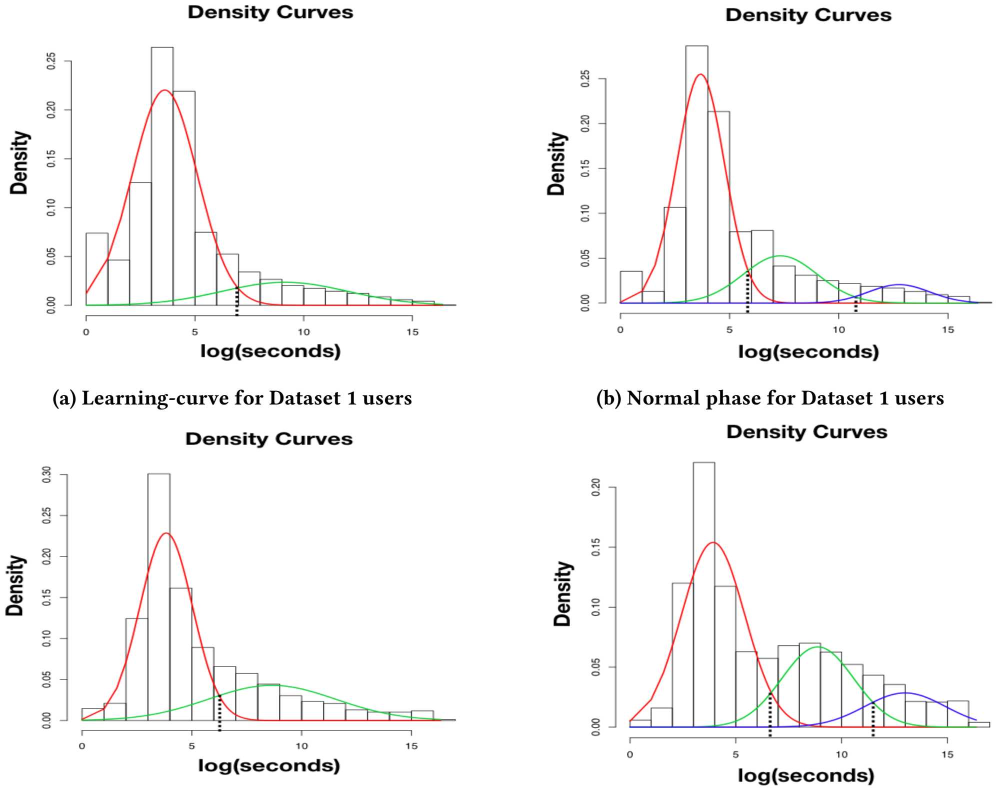
In this paper, we focus on the impact of the learning-phase on IA task and session identification. Smart speakers often provide a new experiences for users who are not familiar with them. Therefore, users who are new to using smart speakers are generally non-experts and curious to interact with the IA, which motivates them to query the smart speaker more frequently when they first start using it compared to the normal usage. We name this stage of the new user usage as the learning-phase, in which users try to learn the device functionality and satisfy their curiosity. In this paper, the period after the learning-phase is named normal-phase. The difference in users' behavior in learning-phase and normal phase has been also observed in Figure 3. Figure 3 shows impact of learning-phase on average number of queries issued per day. learning-phase is modeled by users' expertise based on number of usage days in this diagram.

Figure 3 indicates that the average number of queries issued by users drops after 8 usage days with no significant increase in average number of queries after the 8th day in both Dataset 1 and Dataset 2 interaction logs. For example, average number of queries drops from 42.18 and 40.79 in seventh and eighth days, respectively, to 36.99 and 19.83 average number of queries in the ninth and tenth days for Dataset 1 users. We observed a similar pattern, yet with less considerable drops, for Dataset 2. Specifically, average number of queries in Dataset 2 is 22.2, 13.55, 11.75, and 9.95, for days 7, 8, 9, and 10, respectively.

4.3 Identifying Session Boundary Cutoff in learning-phase

In addition to our observation in Figure 3, Figure 4 shows a per day basis analysis of users inter-activity times in Dataset 1. Plotting the inter-activity times of new users in their first fifteen days of using smart speakers leads to a histogram where there is no evidence of a trimodal Gaussian distributions in users' inter-activity times in first eight days. However, in this experiment, we have observed that after the eighth usage day, an additional valley appears in the histogram, which is shown by a black arrow. According to the observations in Figures 3 and 4, we chose to consider any user interactions logged in the first eight days of their usage as part of the learning-phase, in which users are curious and issue many queries. The rest we define as their normal phase. We know that using a 8 days learning-phase based on our observations is a strong assumption. For example, we could have chosen 7 or 9 days as the learning-phase. Identifying an accurate learning-phase is not the main focus of this experiment. In this experiment, we are interested in showing that the task and session boundaries are different in learning-phase compared to the normal-phase. In the rest of this section, we detail task and session boundary cut-offs on Dataset 1 and Dataset 2 in users learning-phase compared to their normal phase.

In Figure 5, we have plotted histograms of Dataset 1 and Dataset 2 users inter-activity times in their learning-phase as well as their normal phase. In Figures 5a and 5c, we fit a 2-component GMM on user inter-activity times in their learning-phase to estimate task boundary cut-off. The intersection point of the within-task and between-task Gaussian distributions (task boundary) of the learning-phases are $2^{6.9} \sim 119$ and $2^{6.3} \sim 79$ seconds for Dataset 1 and Dataset 2 users, respectively, which are similar to the 2 minutes task boundary cut-off of Microsoft Cortana on Desktop [23]. During the learning-phase, as we do not observe a tri-modal Gaussian distribution in user inter-activity times shown in Figures 5a and 5c, there is not any clear evidence of sessions in user inter-activity time distribution. One possible explanation of it is that users are



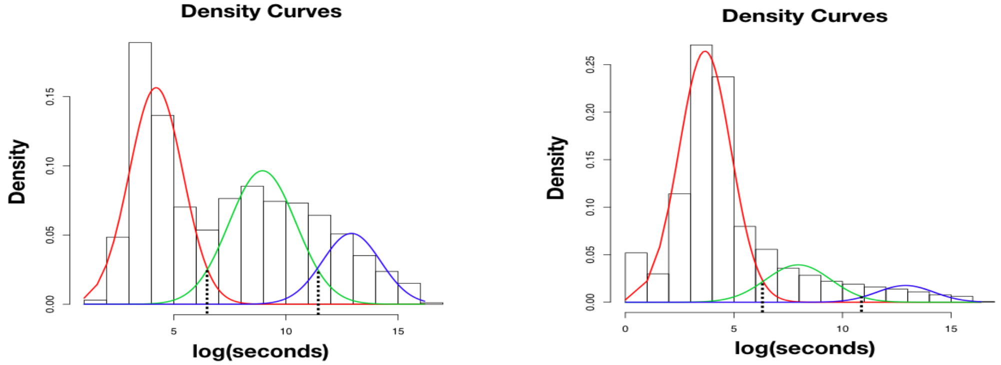
(a) Learning-curve for Dataset 1 users

(b) Normal phase for Dataset 1 users

(c) Learning-curve for Dataset 2 users

(d) Normal Phase for Dataset 2 users

Figure 5: Impact of learning-curve on task and session boundary. The difference between mean of the learning-curve and the normal-phase inter-activity times distributions is statistically significant based on t-test ($\rho < 0.05$).



(a) Control-media domain

(b) All domains

Figure 6: Impact of domains on session boundary. The difference between mean of the control-media and all domain inter-activity times distributions is statistically significant based on t-test ($\rho < 0.05$).

more curious to try the IA to learn its functionality in their learning-phase rather than querying the IA in a session-based scenario to fulfill one or more information needs.

Furthermore, Figures 5b and 5d show a fit of the 3-component GMM on the normal-phase inter-activity times. The intersection point of the within-task and between-task inter-activity times distribution (the task boundary cut-off) is $2^{5.8} \sim 56$ and $2^{6.5} \sim 91$ seconds for Dataset 1 and Dataset 2, respectively. In addition, we identify the session boundary cut-off based on the intersection point

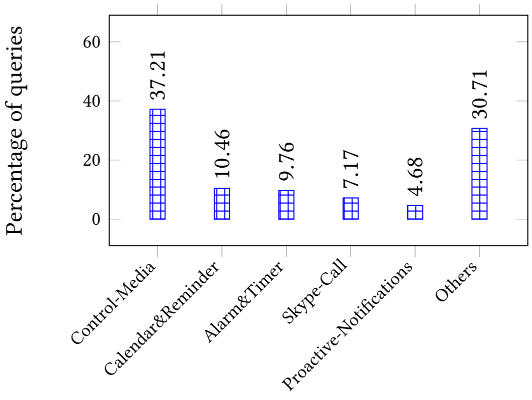


Figure 7: The top-5 popular generic domains in dataset1.

of the between-task and between-session inter-activity times distribution, which is $2^{10.8}$ *seconds* ∼ 30 *minutes* and $2^{11.5}$ *seconds* ∼ 48 *minutes* for Dataset 1 and Dataset 2, respectively.

Although the 48 minutes inactivity threshold to identify IA sessions seems long compared to the common 30 minutes session boundary cut-off in search engines, we suspect that the smart speakers domains of usage might contribute in this difference. Specifically, many of the queries are related to domains like controlling media and listening to music in smart speakers, which requires a longer session duration compared to the typical search engine sessions. Figure 7 shows the top-5 popular domains being used in Dataset 1. Apart from the control media, the rest of the top-5 popular domains are short-term tasks based on number of queries per task. That is one of the possible explanations why the task and session cut-offs in Dataset 1 are relatively short.

To summarize, in this section, we studied the impact of a learning-phase on session boundary cut-off and showed that there is not any evidence of sessions having multiple tasks during the learning-phase. By contrast, during the normal-phase, we observed evidence of both tasks and sessions in users inter-activity times.

5 IMPACT OF USAGE DOMAIN ON SESSION BOUNDARY CUTOFF

The final question we address in this paper is how different domains and tasks affect session boundaries. We investigate whether there is a dependency between session boundary and the domain of a user's information needs. This section answers our third research question: *What is the impact of the domain on task and session boundary cut-offs?*

As it is shown in Figure 7, about 37 % of the smart speaker usage in Dataset 1 is in control-media domain. Therefore, we focus on this domain and study how task and session boundaries in this domain differ from those of other domains. The control-media domain contains queries with intents, such as query media, play music and volume up. In this experiment, we preprocessed the inter-activity times for each day of data. Specifically, we analyzed user queries each day and, if all the queries of a user's interactions in a day is in control-media domain, we consider the inter-activity times of the user in the day as control-media inter-activity times. Otherwise, the user's inter-activity times in the day is in all domains category. The control-media inter-activity times is 11.6 % of the all user inter-activity times. All the available domain labels of users' queries available in Dataset 1 come from a high-quality commercial domain classifiers.

Figure 6 shows a fit of three Gaussians on users' inter-activity times of control-media domain against all domains together. As it is shown in Figure 6a, in control-media domain usages, the task boundary cut-off is $2^{6.5}$ ∼ 91 seconds and the session boundary cut-off is $2^{11.5}$ *seconds* ∼ 48*minutes*. In contrast, the 3-component GMM fit on all domains inter-activity times lead to $2^{6.3}$ ∼ 79 seconds task boundary cut-off and $2^{10.9}$ seconds ∼ 32 minutes session boundary cut-off estimations.

These results indicate that estimating task and session boundary cut-offs for all domains may be misleading. Instead the data suggest that task and session boundary cut-off identification should be done per domain or on a group of similar domains. Furthermore, we have observed that the difference between the mean of control-media inter-activity times distribution and the mean of all domain inter-activity times distribution is statistically significant based on t-test with $\rho < 0.05$, which supports our conclusion in this section.

6 DISCUSSION AND CONCLUSIONS

In this section, we will first discuss impact of our study in downstream applications and computational costs of our proposed task and session identification model. We then briefly detail conclusions.

6.1 Discussion

6.1.1 Impact on Downstream Applications. Task and session identification is a key element of many IR applications such as evaluation studies based on user interaction logs, user modeling, and personalization. Specifically, in user satisfaction prediction based on implicit signals from user interactions, which is an emerging metric to evaluate Web search engines [13] and IAs [12], task and session identification potentially has a direct impact on effectiveness of a user satisfaction classifier.

Table 1 shows an example of a user session in a smart speaker. In task 2 of this example, if a task identifier indicates that the task is terminated after system response to the user's fourth query (U4), an effective user satisfaction classifier would most likely classify the task as a dis-satisfactory (DSAT) task. However, if a task identifier indicates system response to the user's fifth query (U5) as the end of the task, the effective user satisfaction classifier would most likely classify the task as a satisfactory (SAT) task. Session identification would also have a similar impact on session-level user satisfaction prediction, for which we do not provide an example because of space limitation in this paper. Our main point is that the task and session identification could have a direct impact on user interaction log based studies such as user satisfaction prediction problem.

6.1.2 Cost. our proposed task and session identification model is a time-oriented approach and time-oriented approaches calculate task or session boundary once off-line then use it on-line without any re-estimation, it is computationally cost-effective compared to navigation-oriented and query-refinement oriented approaches. In fact, we just need to estimate task and session boundary once and then use it for a downstream application, which is almost as cheap as using the 30 minutes session boundary for session identification in search engine query logs.

6.1.3 Contextual Factor. In this paper, we have studied impact of contextual factors on task and session boundaries. We decided to focus on domain and learning-phase because our observations shows that they are important contextual factors to be aware of while modeling users' behavior. They were also easy to measure. Our main aim was to show that contextual factors do matter and these are just two examples of them. We agree that additional study is required to investigate other contextual factors such as location, which we leave for future work.

6.2 Conclusion

In this paper, we investigated the impact of the learning-phase and usage domain on task and session boundary cut-off for two different IAs being used in smart speakers. We experimented with an application of a 3-component Gaussian mixture model to fit user inter-activity times with the aim of jointly identifying both task and session boundary cut-offs in IA user interaction logs. Our main research question was: **What is the impact of the learning curve and task domain on task and session boundaries when interacting with intelligent assistants?** Specifically, we answer following research questions:

Our first research question was: *How does one effectively measure task and session boundary cut offs in intelligent assistant systems?* We evaluated 2- and 3-component GMMs in task and session boundary estimation based on crowdsourced task boundary cut-off labels and system generated task labels. Our results show that fitting a GMM on user inter-activity times is an effective approach to estimate task and session boundary cut-offs in IA usage on smart speakers. Furthermore, our experimental results show that using a 3-component GMM leads to a better estimation of task boundary cut-offs compared to a 2-component GMM.

Our second research question was: *Do user learning curves have an impact on session boundary cut-offs?* We showed how an additional inter-activity time cluster appears in normal phase, which is not available in learning-phase. We concluded that while using 2-component GMM leads to a reasonable fit of users inter-activity times in their learning-phase, fitting a 3-component GMM is more effective during the normal-phase. In fact, there is not significant evidence of sessions having multiple tasks in learning-phase, and users tend to accomplish tasks more frequently in learning-phase compared to normal-phase of their usage.

Our third research question was: *What is the impact of the domain on task and session boundary cut-offs?* According to the experimental results, task and session boundaries differ across domains and therefore the domain of a task should be considered when measuring these boundaries.

In summary, our general conclusion is that task and session boundary cut-offs are not static but are instead dependent on contextual factors like the user's learning curve and their usage domains.

REFERENCES

[1] 2002. Combining evidence for automatic Web session identification. *IPM* 38, 5 (2002), 727 – 742.

[2] Tatiana Benaglia, Didier Chauveau, David R. Hunter, and Derek S. Young. 2009. mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software* 32, 6 (2009), 1–29.

[3] F.A. Brown, M.G. Lawrence, and V.O.B. Morrison. 2017. Conversational virtual healthcare assistant. https://www.google.com/patents/US9536049 US Patent 9,536,049.

[4] Lara D. Catledge and James E. Pitkow. 1995. Characterizing browsing strategies in the World-Wide web. *Computer Networks and ISDN Systems* 27, 6 (1995), 1065 – 1073.

[5] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. 1999. Data Preparation for Mining World Wide Web Browsing Patterns. *KIS* 1, 1 (1999), 5–32.

[6] David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *PAMI* 2 (1979), 224–227.

[7] Carsten Eickhoff, Pieter Dekker, and Arjen P. de Vries. 2012. Supporting Children's Web Search in School Environments. In *IIiX*. 129–137.

[8] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the Journey: A Query Log Analysis of Within-session Learning. In *WSDM*. 223–232.

[9] Aaron Halfaker, Oliver Keyes, Daniel Kluver, Jacob Thebault-Spieker, Tien Nguyen, Kenneth Shores, Anuradha Uduwage, and Morten Warncke-Wang. 2015. User Session Identification Based on Strong Regularities in Inter-activity Time. In *WWW*. 410–418.

[10] Seyyed Hadi Hashemi, Charles L. A. Clarke, Jaap Kamps, Julia Kiseleva, and Ellen M. Voorhees. 2016. Overview of the TREC 2016 Contextual Suggestion Track. In *TREC*.

[11] Seyyed Hadi Hashemi and Jaap Kamps. 2017. Where To Go Next?: Exploiting Behavioral User Models in Smart Environments. In *UMAP*. ACM, 50–58.

[12] Seyyed Hadi Hashemi, Kyle Williams, Ahmed El Kholy, Imed Zitouni, and Paul Crook. 2018. Measuring User Satisfaction on Smart Speaker Intelligent Assistants Using Intent Sensitive Query Embeddings. In *CIKM*.

[13] Ahmed Hassan and Ryen W. White. 2013. Personalized Models of Search Satisfaction. In *CIKM*. 2009–2018.

[14] Bernard J. Jansen, Amanda Spink, Chris Blakely, and Sherry Koshman. 2007. Defining a Session on Web Search Engines: Research Articles. *JASIST* 58, 6 (2007), 862–871.

[15] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic Online Evaluation of Intelligent Assistants. In *WWW*. 506–516.

[16] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs. In *CIKM*. 699–708.

[17] Maryam Kamvar, Melanie Kellar, Rajan Patel, and Ya Xu. 2009. Computers and Iphones and Mobile Phones, Oh My!: A Logs-based Comparison of Search Users on Different Devices. In *WWW*. 801–810.

[18] Madian Khabsa, Aidan Crook, Ahmed Hassan Awadallah, Imed Zitouni, Tasos Anastasakos, and Kyle Williams. 2016. Learning to Account for Good Abandonment in Search Success Metrics. In *CIKM*. 1893–1896.

[19] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting user satisfaction with intelligent assistants. In *SIGIR*. 45–54.

[20] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *CHIIR*. 121–130.

[21] Ard W. Lazonder, Harm J.A. Biemans, and Iwan G.J.H. Wopereis. 2000. Differences between novice and experienced users in searching information on the World Wide Web. *JASIST* 51, 6 (2000), 576–581.

[22] Liangda Li, Hongbo Deng, Anlei Dong, Yi Chang, and Hongyuan Zha. 2014. Identifying and Labeling Search Tasks via Query-based Hawkes Processes. In *Proceedings of the KDD (KDD '14)*. 731–740.

[23] Rishabh Mehrotra, Ahmed El Kholy, Imez Zitouni, Milad Shokouhi, and Ahmed Hassan. 2017. Identifying User Sessions in Interactions with Intelligent Digital Assistants. In *WWW*. 821–822.

[24] Rishabh Mehrotra and Emine Yilmaz. 2017. Extracting Hierarchies of Search Tasks &#38; Subtasks via a Bayesian Nonparametric Approach. In *Proceedings of the SIGIR (SIGIR '17)*. 285–294.

[25] Rishabh Mehrotra, Imed Zitouni, Ahmed Hassan Awadallah, Ahmed El Kholy, and Madian Khabsa. 2017. User Interaction Sequences for Search Satisfaction Prediction. In *SIGIR*. 165–174.

[26] David Mehrzadi and Dror G. Feitelson. 2012. On Extracting Session Data from Activity Logs. In *SYSTOR*. 3:1–3:7.

[27] Tom Minka et al. 2005. *Divergence measures and message passing*. Technical Report. Technical report, Microsoft Research.

[28] Alan L Montgomery and Christos Faloutsos. 2001. Identifying web browsing trends and patterns. *Computer* 34, 7 (2001), 94–95.

[29] Mehran Nadjarbashi-Noghani and Ali A Ghorbani. 2004. Improving the referrer-based web log session reconstruction. In *Communication Networks and Services Research*. 286–292.

[30] H. Cenk Ozmutlu and Fatih ÃĞavdur. 2005. Application of automatic topic identification on Excite Web search engine data logs. *IPM* 41, 5 (2005), 1243 – 1262.

[31] Seda Ozmutlu. 2006. Automatic new topic identification using multiple linear regression. *IPM* 42, 4 (2006), 934 – 950.

[32] Filip Radlinski and Thorsten Joachims. 2005. Query Chains: Learning to Rank from Implicit Feedback. In *KDD*. 239–248.

[33] Milad Shokouhi, Umut Ozertem, and Nick Craswell. 2016. Did You Say U2 or YouTube?: Inferring Implicit Transcripts from Voice Search Logs. In *WWW*. 1215–1224.

[34] Jaime Teevan, Amy Karlson, Shahriyar Amini, A. J. Bernheim Brush, and John Krumm. 2011. Understanding the Importance of Location, Time, and People in Mobile Local Search Behavior. In *MobileHCI*. 77–80.

[35] Pertti Vakkari, Mikko Pennanen, and Sami Serola. 2003. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *IPM* 39, 3 (2003), 445 – 463.

[36] Ryen W. White, Susan T. Dumais, and Jaime Teevan. 2009. Characterizing the Influence of Domain Expertise on Web Search Behavior. In *WSDM*. 132–141.

[37] Barbara M. Wildemuth. 2004. The effects of domain knowledge on search tactic formulation. *JASIST* 55, 3 (2004), 246–258.

[38] Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Detecting good abandonment in mobile search. In *WWW*. 495–505.

[39] Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Is This Your Final Answer?: Evaluating the Effect of Answers on Good Abandonment in Mobile Search. In *SIGIR*. 889–892.