A User Modeling Shared Challenge Proposal

Owen Conlan[1], Kieran Fraser[1], Liadh Kelly[2(✉)], and Bilal Yousuf[1]

[1] Adapt Centre, Trinity College Dublin, Dublin, Ireland
{owen.conlan,kieran.fraser,bilal.yousuf}@adaptcentre.ie
[2] Maynooth University, Kildare, Ireland
liadh.kelly@mu.ie

Abstract. Comparative evaluation in the areas of User Modeling, Adaptation and Personalization (UMAP) is significantly challenging. It has always been difficult to rigorously compare different approaches to personalization, as the function of the resulting systems is, by their nature, heavily influenced by the behavior of the users involved in trialing the systems. Developing comparative evaluations in this space would be a huge advancement as it would enable shared comparison across research. Here we present a proposal for a shared challenge generation in UMAP, focusing on user model generation using logged mobile phone data, with an assumed purpose of supporting mobile phone notification suggestion. The dataset, evaluation metrics, and challenge operation are described.

Keywords: Personalization · Evaluation · Shared task

1 Introduction

There is currently no established or standardized means for comparative evaluation of algorithms and systems developed by researchers in the User Modeling, Adaptation and Personalization (UMAP) space. The development of such methodologies has proven to be extremely difficult, but would be highly rewarding for the community. Privacy concerns, the challenges of working with interactive scenarios, and the individual differences in behavior between users all must be addressed in order to facilitate repeatable and comparable evaluation and to advance research in this domain. The EvalUMAP workshop series[1] [1,2] is a new concerted drive towards the establishment of shared challenges for comparative evaluation within the UMAP community. The first workshop in the series brought the community together to discuss challenges and potential solutions associated with generating shared evaluation challenges in the UMAP space.

[1] http://evalumap.adaptcentre.ie.

Authors listed alphabetically.

F. Crestani et al. (Eds.): CLEF 2019, LNCS 11696, pp. 171–177, 2019.
https://doi.org/10.1007/978-3-030-28577-7_13

Building on the success of the first edition of the workshop, the second edition made concrete steps towards identifying datasets and methods that could be exploited for shared UMAP evaluation challenges. It is intended that the third edition of the workshop, running at UMAP 2019, will further progress this move towards shared challenge generation. In this paper we present a proposed methodology for such a shared challenge in the community, including practical steps to implementation.

2 Related Work

Adaptive system evaluation has been a recurrent topic within the community over the years, for example [8,9,11]. However, a solution capable of delivering repeatable and comparable results that would become the standard method to evaluate UMAP research has yet to emerge.

Lessons can be learned here from progress in other domains in shared challenge generation. The nearest to our UMAP challenge being arguably that of the Information Retrieval (IR) community. In recent years the community has started to look more closely at bringing the user into the loop, exploring the creation of shared challenges that consider iterative search sessions (for example in initiatives such as [7]), providing profiles of individual users to aid search (for example, the new PIR-CLEF task[2]) and providing access to real users conducting real search tasks [6,10]. In working towards the possibility of shared challenges in the UMAP community we can learn from such initiatives. However, the types of algorithms and systems which the UMAP community seek to evaluate are of a distinct nature, and as such will require their own unique solution.

3 Proposed Shared Challenge Description

The use-case for the proposed challenge is personalized mobile phone notification generation with the intention of expanding to other use-cases and challenges in the future. Our previous work in this space [4] has explored intercepting incoming mobile notifications, mediating their delivery such that irrelevant or unnecessary notifications do not reach the end-user and generating synthetic notification datasets from real world usage data. The next step toward an improved notification experience is to generate personalised notifications in real-time, removing the need for interception and delivery mediation. Specifically, assuming individuals' interactions with their mobile phone have been logged, the challenge is to create an approach to generate personalized notifications on individuals' mobile phones, whereby such personalization would consist of deciding what events (SMS received, etc.) to show to the individual and when to show them. Given the number of steps associated with such personalization, the task proposed in this paper will focus on the first step in this process, that of user model generation using the logged mobile phone interactions. For this task a dataset consisting of several individuals' mobile phone interactions would be provided, described next.

[2] http://www.ir.disco.unimib.it/pir-clef2019/.

3.1 Challenge Dataset

The dataset associated with this proposed shared challenge is a simulated dataset that is based on mobile notifications gathered by the WeAreUs Android app. The dataset generation approach is described in [5]. The synthetic data provided in the challenge dataset is comprised of notification, engagement and contextual features. The notification features relate to the event: posting of notification to the user's device. The contextual features describe the user/device context at particular moments of interest such as when a notification is posted and when it is removed. The engagement features describe the reaction the user has to the notification. See Table 2 for an outline of the captured data features. Since this dataset consists of synthetic data, as opposed to real individuals data, the ethical and privacy concerns are negligible as the data cannot be combined or analysed to identify real individuals.

4 Challenge Operation

The challenge would operate with a campaign style format. Participants will be provided with a sample of data as described in the previous section, and will be required to create user models for the individuals described in the data.

As a means of steering user model creation toward a tangible goal, and hence toward evaluative metrics, two tasks in the domain of mobile notification management are proposed. Task 1 is an offline scenario where models are trained and then evaluated on a static test set. Task 2, in contrast, simulates a live interactive environment in which models must adapt on the fly. Participants can take part in one or both of these tasks to complete the challenge.

An OpenAI Gym environment, specifically Gym-push, is used for the challenge tasks. OpenAI Gym is an open source interface to reinforcement learning (RL) tasks. It provides environments for researchers to benchmark RL agents on simulations of real-world problems. Gym-push is a custom OpenAI Gym environment developed for this proposed challenge which simulates push-notifications arriving on a user's device, the context in which the user receives the notification and the subsequent reward received for engagements made by the user. Gym-push is the simulated environment which will be used to evaluate the performance of the challenge participants' user models. The participants will receive various context features from the environment which they can apply as input to their user models to generate personalized notifications. They can then pass these generated notifications to the environment for evaluation. Within the environment, an agent, acting as the user, will engage with the generated notifications and metrics measuring various facets of performance (discussed further in Sect. 5) which will be tracked. It is important therefore that the user models created conform to the requirements of the Gym-push environment to ensure evaluation can take place (implementation guidelines detailed in Sect. 4.3).

Following ACM's policy on *Artifact Review and Badging*[3] and to support best practice with regard to reproducibility [3], submitted participant models

[3] https://www.acm.org/publications/policies/artifact-review-badging.

will be stored in the Gym-push environment along with the version of data used to obtain their final performance results. Subsequently, the environment will be able to generate additional diverse notification datasets using these models. These additional datasets can also be utilised by other communities for various research purposes.

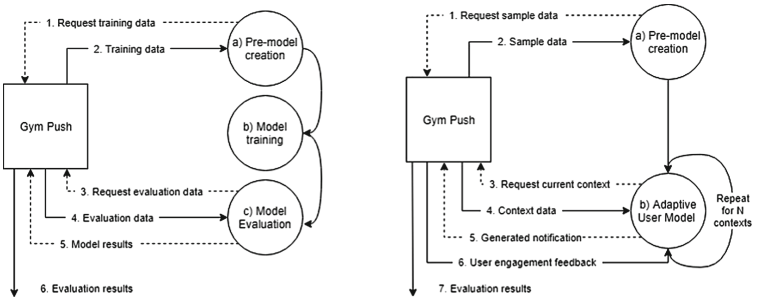

Fig. 1. Left: Task 1 operation flow; Right: Task 2 operation flow.

4.1 Task 1

Figure 1 (Left) illustrates the operation flow for participants partaking in Task 1. Participant can query Gym-push for 3 months of historical data, which takes a context (e.g. Time: 'morning', Place: 'airport', etc.) as input and outputs a personalized notification (e.g. App: 'news', Subject: 'weather', etc.) for the given context. Once the model is built, it can be evaluated, again using Gym-push. This is achieved by giving the participant an additional 3 months of contextual evaluation data with which to generate notifications. The resulting notifications are then returned to the environment where evaluation metrics are calculated.

4.2 Task 2

Figure 1 (Right) illustrates the differing operation flow for participants partaking in Task 2. Participants are asked to create a user model based on the same notification, context and engagement features but without historical notification data to train with (although, they can query the environment for sample data with which to create their model). In contrast to Task 1, this user model will need to query the Gym-push environment at each step to receive a current context feature and a previous user notification-engagement feature. As the environment steps through each context item and as engagement history becomes available, the user model can exploit this information to improve the generation of personalized notifications. The goal is to develop a model which adapts and learns how to generate personalized notifications in real-time, without prior history of the user (cold-start problem). Evaluation is continuous for this task and a summary of results is issued once all context features have been processed.

4.3 Model Guidelines

Task 1. The user model should take a context as input and produce a personalized notification as output. The context and notification should be strictly represented by the *Contextual* and *Notification* category features detailed in Table 2. More detail, including the set of values each feature can take, is available at the Gym-push repository.

Task 2. In addition to data available in Task 1, this model can also make use of notification-engagement data relating to a generated notification. The notification-engagements are represented by the *Engagement* category features noted in Table 2.

5 Evaluation Approach and Metrics

The Gym-push environment will evaluate the user models by deploying an agent to act as a user engaging with the generated personalized notifications. The agent will be trained on historical data of the user and decide, given the context, to open or dismiss the notification generated by the model. The following two metrics will be tracked in both Tasks 1 and 2:

Diversity - This metric will evaluate the diversity of generated personalized notifications which have been accepted by the agent over the 3 months. Notification sets which boast greater diversity will be scored higher.
Performance - This metric will track and compare engagements resulting from the generated personalized notifications with those of the actual notifications. Scenarios which improve end-user engagements are scored higher (see Table 1).

Table 1. Performance metric

For a given context		
Actual notification	Generated notification	Reward
Opened	Opened	+1
Dismissed	Dismissed	+0
Dismissed	Opened	+2
Opened	Dismissed	−1

Table 2. Dataset features

Category	Features
Notification	App, category, updates, subject, priority, ongoing, visibility
Contextual	Day, time, place, contact-significance, activity, noise, battery level, charging, headphones-in, light intensity, music-active, proximity, ringer-mode
Engagement	Time app last used, seen time, decision time, response time, action

Two additional metrics are tracked in Task 2:

Response Time - This metric evaluates the time it takes the user model to generate a notification once given the context by the environment. Shorter times are scored higher.
Learning Rate - This metric evaluates how quickly the performance metric (above) of the model improves over each time step (context item) of the environment.

6 Conclusions

While evaluation is an active topic of research within the UMAP community, to-date there are no shared evaluation challenges in the UMAP community which would allow comparison of developed systems in controlled environments similar to the evaluation labs offered in other research communities. Improved solutions for UMAP evaluation that have lower cost, are more repeatable, and more realistic are required. In this paper we propose one possible approach for shared challenge generation in the UMAP community, including use-case, data collection and evaluation methodology. This challenge focuses on user model generation using logged mobile phone data, with an assumed purpose of supporting mobile phone notification suggestion. It is not expected that this proposal is the only possible UMAP shared challenge, rather in putting forward this challenge proposal we seek to open further discussion and progress towards shared challenge generation for the UMAP community.

References

1. Conlan, O., Kelly, L., Koidl, K., Lawless, S., Levacher, K., Staikopoulos, A.: EvalUMAP 2016: towards comparative evaluation in the user modelling, adaptation and personalization space workshop. In: UMAP 2016 (2016)
2. Conlan, O., Kelly, L., Koidl, K., Lawless, S., Staikopoulos, A.: EvalUMAP 2017: towards comparative evaluation in the user modelling, adaptation and personalization space workshop. In: UMAP 2017 (2017)
3. Fehr, J., Heiland, J., Himpe, C., Saak, J.: Best practices for replicability, reproducibility and reusability of computer-based experiments exemplified by model reduction software. arXiv preprint arXiv:1607.01191 (2016)
4. Fraser, K., Yousuf, B., Conlan, O.: Synthesis and evaluation of a mobile notification dataset. In: Adjunct Publication of UMAP 2017 (2017)
5. Fraser, K., Yousuf, B., Conlan, O.: Scrutable and persuasive push-notifications. In: Oinas-Kukkonen, H., Win, K.T., Karapanos, E., Karppinen, P., Kyza, E. (eds.) PERSUASIVE 2019. LNCS, vol. 11433, pp. 67–73. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-17287-9_6
6. Hopfgartner, F., Kille, B., Lommatzsch, A., Plumbaum, T., Brodt, T., Heintz, T.: Benchmarking news recommendations in a living lab. In: Kanoulas, E., et al. (eds.) CLEF 2014. LNCS, vol. 8685, pp. 250–267. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11382-1_21

7. Hui Yang, G., Soboroff, I.: TREC 2016 dynamic domain track overview. In: TREC 2016 (2016)
8. Paramythis, A., Weibelzahl, S., Masthoff, J.: Layered evaluation of interactive adaptive systems: framework and formative methods. User Model. User-Adap. Interact. **20**(5), 383–453 (2010). https://doi.org/10.1007/s11257-010-9082-4
9. Park, K.S., Hwan Lim, C.: A structured methodology for comparative evaluation of user interface designs using usability criteria and measures. Int. J. Ind. Ergon. **23**(5–6), 379–389 (1999). https://doi.org/10.1016/S0169-8141(97)00059-0
10. Schuth, A., Balog, K., Kelly, L.: Overview of the Living Labs for Information Retrieval Evaluation (LL4IR) CLEF lab 2015. In: Mothe, J., et al. (eds.) CLEF 2015. LNCS, vol. 9283, pp. 484–496. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24027-5_47
11. Van Velsen, L., van der Geest, T., Klaassen, R., Steehouder, M.: User-centered evaluation of adaptive and adaptable systems: a literature review. Knowl. Eng. Rev. **23**(3), 261–281 (2008). https://doi.org/10.1017/S0269888908001379