EvalUMAP: Towards comparative evaluation in user modeling, adaptation and personalization

Owen Conlan[1], Liadh Kelly[2], Kevin Koidl[1], Séamus Lawless[1], and Athanasios Staikopoulos[1]

[1] ADAPT Centre, School of Computer Science and Statistics
Trinity College Dublin, Ireland
owen.conlan@scss.tcd.ie, kevin.koidl@scss.tcd.ie
seamus.lawless@scss.tcd.ie, athanasios.staikopoulos@scss.tcd.ie
[2] ADAPT Centre, Dublin City University, Ireland
liadh.kelly@dcu.ie

Abstract. There is currently no established or standardized means for the comparative evaluation of algorithms and systems developed by researchers in the User Modeling, Adaptation and Personalization (UMAP) space. The design and establishment of such methodologies has proven to be extremely difficult, but would be highly rewarding, as demonstrated by initiatives such as CLEF, TREC and NTCIR in the Information Retrieval domain. Privacy concerns, the challenges of working with interactive scenarios, and individual differences in behaviour between users must all be addressed in order to facilitate repeatable and comparable evaluation, and to advance research in this domain. In this paper we present EvalUMAP, a new concerted drive towards the establishment of shared challenges for comparative evaluation within the UMAP community.

1 Introduction

Research in the areas of User Modelling, Adaptation and Personalization (UMAP) faces a number of significant scientific challenges. One of the most significant of these challenges is the issue of comparative evaluation. 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 behaviour of the users involved in trialling the systems. To-date this topic has received relatively little attention when compared with other areas of Computer Science research, such as Information Retrieval (IR). Developing comparative evaluations in this space would be a huge advancement as it would enable shared comparison across research, which to-date has been very limited.

One of the significant challenges in establishing such an initiative, is that the UMAP community encompasses a broad range of research areas and technologies. An array of approaches to User Modeling exist, with no standardised approach to data capture, analysis or representation. Personalised and adaptive systems that utilise user models are also hugely varied in nature, from personalised IR systems which tailor the selection and ranking of content to an individual, to more complex personalised learning systems, which tailor interaction and learning offerings to an individual based upon their competency or performance in learning tasks.

Taking inspiration from communities such as IR and Machine Translation (MT), the EvalUMAP Workshop series[3] was established in 2016 with the ambitious goal of moving towards comparative evaluation in the UMAP community. A specific first goal was set, to propose and design one or more shared tasks to support the comparative evaluation of approaches to User Modelling, Adaptation and Personalization. The long term vision is the establishment of an annual shared challenge series, similar to TREC[4] and CLEF[5] in the IR space. The establishment of such shared tasks requires that appropriate models, content, metadata, user behaviours, etc. be available, in order to comprehensively compare how different approaches and systems perform. In addition, a number of metrics and observations would need to be outlined, that participants would be expected to perform in order to facilitate comparison. This is significant. To move towards this goal we, as a community need to greatly advance our understanding of, and methodology associated with UMAP evaluation. Including not only the technical challenges associated with design and implementation, but also privacy, ethics, legal and security issues, evaluation methodologies and metrics.

When compared with shared tasks in IR, EvalUMAP aims to develop tasks and test collections which are focused on variations in the user (represented by variations in the underlying user model) and the personalised decisions taken by the systems, rather than variations in the queries and/or relevance judgments provided. The ultimate goal is to have the users who are being modeled involved in judging the performance of the personalised systems, and thereby contributing to the iterative enhancement of the test collections used.

In the next section we provide background for the EvalUMAP initiative and then move on to provide an overview of progress to-date towards this goal. We conclude with a discussion of future directions.

2 Background

Despite research interest and progress being made in UMAP research, it is well understood within the community that progression has been limited by a lack of cross comparable evaluation methods [10] [13]. This problem was highlighted during the panel session at the UMAP 2015 conference. An outcome of which was the need for community exerted effort in developing cross comparable evaluation approaches in UMAP evaluation. The subsequent EvalUMAP 2016 workshop [7] began concrete discussion on this topic.

Currently, there are no established or standardized baselines or evaluation metrics, and no commonly available test collections. Privacy concerns, the chal-

[3] http://evalumap.adaptcentre.ie/
[4] http://trec.nist.gov/
[5] http://clef2017.clef-initiative.eu/

lenges of working with interactive scenarios, and the individual differences in behaviour between users all must be addressed in order to facilitate repeatable and comparable evaluation and to advance research in this domain. While overcoming these problems is a big challenge, there have been some notable efforts in the past from which to build on.

Park et. al [16] for example, propose a two phase evaluation model consisting of a qualitative pre-screening phase followed by quantitative user-based assessments (using objective measures) to compare various content alternatives. Weibelzahl [21] and Chin et. al [5] on the other hand focus more on the need for system-wide empirical evaluations determining which users were helped or hindered by user-adapted interactions. Van Velsen et. al [20] also attempt to produce a model summarising the variables being assessed at each stage of the process along with relevant methods to assess them.

As more methods evaluating the usefulness and accuracy of adaptive systems appeared, the need to evaluate these evaluation methods became evident. Klaassen et. al [19] for example evaluated three of the most common test methods used to detect usability problems in personalised systems. More recently, Paramythis et. al. [15] proposed to unify previous approaches presented in the literature by introducing the Layered Evaluation framework. This approach seeks to tackle the difficulties involved with evaluating adaptivity by decomposing such systems into independent layers of adaptivity (such as User Model, Content Model and Adaptive Decisions Logic). They also propose methods related to the various development lifecycle stages of interactive systems. However, this layered approach advocates evaluations that require a separation of concerns within the design process that is not always possible. The main advantage however is that it enables a clear identification of issues within elements of the design.

As can be seen, adaptive system evaluation has been a recurrent topic within the community over the years. Nevertheless, a solution capable of delivering repeatable and comparable results that would become the standard method to evaluate UMAP research has yet to emerge. Improved solutions for UMAP evaluation that have lower cost, are more repeatable, and more realistic are required.

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. This community has a long history in shared challenge generation, with multiple established shared challenge evaluation series running across the globe, namely TREC[6], NTCIR[7], FIRE[8], CLEF[9] and most recently MediaEval[10]. The evaluation methodology adopted by these shared challenges primarily involves the sharing of resources with participating teams to perform a task, for example data retrieval or annotation. Each participating team then uses their developed technique to perform the ad hoc task

[6] http://trec.nist.gov/

[7] http://research.nii.ac.jp/ntcir/index-en.html

[8] http://fire.irsi.res.in/fire/

[9] http://clef2017.clef-initiative.eu/

[10] http://www.multimediaeval.org/

using the provided data collection. Performance in the task is marked against an organizer provided gold standard.

These challenges traditionally considered the once off requirements of a typical or standard user of a system. 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 [9]), providing profiles of individual users to aid search (for example, the new PIR-CLEF task[11]) and providing access to real users conducting real search tasks [2][8]. Most recently, the Personalized Information Retrieval (PIR) task at CLEF 2017 introduces user profiles to personalize the retrieval process [12].

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 Outcomes of EvalUMAP 2016

The 1st International EvalUMAP 2016 Workshop brought together researchers from across the User Modelling, Adaptation and Personalization community to explore potential new ideas and approaches to support comparative evaluations. This area of research is identified as inherently complex, not only because of its focus on the user, but also because of the different and diverse domains involved. To date this has presented significant barriers to how research outcomes can be compared. In particular, the 1st EvalUMAP workshop investigated how the UMAP research community can facilitate the shared development of evaluation tasks and competitions. In the EvalUMAP 2016 workshop, 10 position and discussion papers were accepted, covering different evaluation aspects from potential frameworks and platforms to requirements and reference models as well as specific evaluation areas and metrics.

More specifically, the contributions of the papers were as follows: Koidl et. al. [12] alleged that researchers conducting evaluations in the fields of User Modelling and Personalisation face the challenge of missing continuing evaluation feedback and collaboration with the overall research community. As a result, the authors proposed a community-driven portal introduced as ECP (Evaluation Community Portal) specifically focused on evaluations within the UMAP community. ECP is inspired by work from the Cross-Language Evaluation Forum (CLEF)[13], and is based on the simplicity of Calls for Papers (CFP). As a starting point, the authors proposed the following key features: a) ability to post calls for participation in evaluations, b) ability to discuss approaches and findings in a forum manner, c) ability to upload and present data that can be

[11] http://www.ir.disco.unimib.it/pirclef2017/

[12] http://www.ir.disco.unimib.it/pirclef2017/

[13] http://www.clef-initiative.eu/

shared and used in other evaluations. Furthermore, an initial task force (community champions) leading these efforts needs to be identified, that will bootstrap the Portal and provide the initial community momentum.

Vahid et. al. [18] presented how related tasks were designed - involving gathering users profiles and objects of their interest, in well-known IR evaluation communities. The paper reviews and compares existing user tasks and describes task resource collection, methods and metrics. In particular, the paper describes two tasks a) the Contextual Suggestion Task of TREC (Text REtrieval Conference) and b) the Social Book Search Task of CLEF. The goal of the contextual suggestion task is to evaluate the search techniques for complex information needs of users with respect to context and their point of interest. This task investigates the development of systems that are able to make suggestions of sites with the goal to explore an unknown city based upon the user's personal interests in the user's home city. A set of user preferences, example suggestions and a set of contexts are given to participants as inputs. As evaluation metrics, Precision at Rank 5 (P@5), Mean Reciprocal Rank (MRR) and a modified version of Time-Biased Gain (TBG) are used to rank participants runs. The Social Book Search task investigates evaluation methodologies for a book search task using a combination of various aspects of retrieval and recommendation dealing with professional and user-generated meta-data. A set of book requests and a set of user profiles have been assumed as inputs of the task and a submitted ranked list of recommended books has been evaluated as the result of participant's systems. The official evaluation measure for this task is nDCG@10. It takes graded relevance values into account and is designed for evaluation based on the top retrieved results. In addition, P@10, MAP and MRR scores will also be reported, with the evaluation results.

Next, Pandit et. al. [14] emphasised the need to support the reproducibility of results in a systematic way. Reproducibility of results is a key element for the verification of scientific experiments and an important indicator of the quality of a published experiment. Therefore, it is vital to precisely and transparently share both the method and the data associated with an experiment. In particular, in their paper the authors explore how emerging linked data standards such as the P-PLAN, CSVW and DCAT ontologies can be applied to the description of the steps and data associated with a published adaptive or personalised experiment in a manner that can be easily located, linked, accessed and reused to repeat an experiment.

L. Kelly [11] proposed using the Living Labs methodology, an emerging evaluation paradigm in IR and recommender systems that provides a platform for supporting shared evaluation tasks. In this case, Living Labs will be adapted to allow for shared evaluation tasks in the UMAP community and, specifically, to support the individual requirements and differences, privacy concerns and the interactive nature of the space. In general, Living labs hold great promise for conducting realistic evaluations, with real users in natural task environments, and more importantly allowing for cross comparability (e.g. by providing a benchmarking platform, perform rankings) across research centres.

Adaji and Vassileva [1] introduced the Persuasive Systems Design (PSD) as a framework for developing and evaluating persuasive systems. Despite its extensive use as a guide for developing persuasive systems, its use as an evaluation tool for persuasive systems is yet to be exploited. The PSD framework is comprised of persuasive principles that could be used to develop and implement strategies that encourage personalisation and adaptation to user preferences. Using Netflix as a case study, the authors identified the implementation of the persuasive principles and the design of system features. This study can act as a guide for the development of evaluation metrics for persuasive related shared tasks.

Bogina and Kuflik [3] pointed out that User Adaptive Systems (UAS) do not intersect with software evaluations as commonly defined in Software Engineering domain. As a result, the authors suggested adopting the common software engineering practices and changing the community's practice and methods by integrating software testing as an integral part of the shared task evaluation process. This would result in more easily reproducible/reusable tasks and data for other members of the community.

Next, P.D. Bra [4] indicates that there is a strong focus on comparative evaluation in the research field of User Modeling, Adaptation and Personalization. In particular, the author discusses and argues when it is reasonable (makes sense) to perform such evaluation on adaptive systems and applications. For adaptive systems, the author argues that these types of comparisons have not gained much acceptance as being "evaluation" by the UMAP community. For applications, the author argues that it is difficult to perform a meaningful evaluation because it is hard to find something to compare the (use of the) application with. In both cases, having a common reference model and applying layered approaches among others may help the community get started and allow different systems and applications to be compared.

Staikopoulos and Conlan [17] focused on evaluating user-adaptive systems and pointed out that it is still a challenging research issue and a difficult task. This is because of the lack of widely accepted evaluation methods, data and the difficulty on generalizing the application areas (e.g. learning, information systems, business). In order to move towards comparative evaluations of user-adaptive systems and to ensure a scientific process the authors indicated some vital requirements. More specifically, the authors proposed developing a flexible common reference model and related metrics upon which user-adaptive systems and approaches could be evaluated and compared against each other both comprehensively (as a whole) as well as upon specific adaptation layers and aspects. The layers should encompass different aspects of a user-adaptive system such as a) inferring User/Group properties, b) identifying the user environment and context (e.g. location, affect), c) evaluating personalised content retrieval, d) evaluating the underlying decision making mechanisms, strategies and algorithms, e) evaluating the adaptation of content, navigation, presentation or user feedback and support, f) the user interaction and experience (e.g. evaluate usability, satisfaction) as well as, g) the system efficiency (e.g. scalability, responsiveness).

In addition, Yousuf and Conlan [22] proposed evaluating the usage of visual narratives as a way to indicate positive behavioural change in student engagement levels. In particular, their paper describes the VisEN framework, to provide visual narratives to students and motivate them with their level of engagement with their course.

Finally, based on the proven existing relationship between the language usage and the author's personality, Chin and Wright [6] proposed using specific evaluation metrics and statistic tests for inferring (predicting user features) and benchmarking user's personality from text. Such guidelines and metrics can then be used to design and evaluate related evaluation tasks. To do so, the authors reported the need for having an established corpora with detailed reporting requirements that will allow researchers to easily compare their algorithms for inferring personality from text. However, there will always be a need to extend the corpora to increase the coverage of different types of writing, time periods, and localities. As a result, the authors recommended having a series of corpora, with, perhaps, one added every few years to keep providing new data to the community.

4 Discussion

Developing means to conduct shared evaluation in the user modelling, adaptation and personalization (UMAP) space is inherently difficult. Not least because of privacy concerns, individual differences in behaviours between the users of systems and challenges associated with working in interactive scenarios. A further challenge is the underlying value of such datasets. Datasets which detail the actions or interests of authentic users are viewed as valuable, and in many cases, proprietry. This compounds the challenge of accessing this data. However, without this access the only resort that remains for reseachers to obtain up-to-date user data is via systems that were created within a research environment leading to potentially low numbers and data that is not up-to-date.

Overcoming these challenges will require greatly advancing our understanding of, and the methodology associated with UMAP evaluation. Challenges include:

- Defining tasks and scenarios for evaluation purposes
- Identification of potential corpora for shared tasks
- Interesting target tasks and explanations of their importance
- Combining existing evaluation metrics and methods
- Improving on previously suggested metrics and methods
- Proposing new evaluation metrics and methods
- Investiagting anonymization or decentralisation of user data to ensure proprietry value is not compromised
- Exploring potential partnerships with companies which hold user data to discuss how research can be conducted without the risk of privacy or value loss

The papers at the 1st EvalUMAP workshop (2016) covered both challenges and potential solutions associated with this area. It is envisaged that future workshops will build on these outcomes, creating a forum to present innovative datasets and shared challenges using these datasets to evaluate systems. The aim of this year's EvalUMAP workshop (2017) is to start scoping and designing a shared task(s). The resulting shared task(s) are to be accompanied by appropriate models, content, metadata, user behaviours, etc., and can be used to comprehensively compare how different approaches and systems perform. In addition, a number of metrics and observations will be outlined, that participants will be expected to perform to facilitate comparison.

To create a community around shared tasks for user modelling it is envisaged that the following aspects have to be addressed: (1) A clear understanding of the challenges and requirements related to the design of a shared task approach in the UMAP space; (2) the identification of suitable, publicly accessible datasets; and (3) an initial description for shared task evaluations using these identified and suitable datasets.

Establishing shared tasks that cover the many facets of a full personalisation system is challenging. With that in mind, the plan is to escalate the tasks year-on-year, starting with a user modelling challenge, then layering in some indicative personalisation decision making processes based on changes in user models, before identifying a mechanism to incorporate real users into the shared task. This escalation is necessary as the user is a complex element of the personalisation process; their actions heavily influence changes to the user model and subsequent personalisation decisions. In order to effectively compare different systems and approaches it is necessary to incorporate users in a meaningful and replicable manner. This of course presents an overhead in running shared tasks. All that being said, our plan is to run the first shared task in 2017-2018 as a static user modelling challenge based on historic social media data from the users and explicitly captured information about their expertise. The ADAPT Centre has committed to support these shared tasks.

5 Conclusions

The EvalUMAP workshop series considers the strengths and limitations of existing work in UMAP evaluation, while moving towards its ambitious goal of designing and establishing a forum for comparative evaluation in the UMAP space. The long term vision of EvalUMAP is the establishment of an annual shared challenge series. The workshop this year will focus on identifying candidate datasets that meet specific requirements (e.g. ownership, accessibility) and that could form the basis for designing shared task challenges and evaluations for the academic year 2017-18, which will be presented at an EvalUMAP 2018 forum. It is not intended that this is the only form of scholarly advancement, but through the shared tasks published and managed by the EvalUMAP Workshop a common baseline for comparison may be established.

Acknowledgements

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

1. Adaji, I., Vassileva, J.: Evaluating persuasive systems using the psd framework. EvalUMAP 16, UMAP 2016 Extended Proceedings 1618 (2016)
2. Balog, K., Kelly, L., Schuth, A.: Head first: Living labs for ad-hoc search evaluation (2014)
3. Bogina, V., Kuflik, T.: Building a bridge between user-adaptive systems evaluation and software testing. EvalUMAP 16, UMAP 2016 Extended Proceedings 1618 (2016)
4. Bra, P.D.: Evaluating adaptive systems and applications is often nonsense. EvalUMAP 16, UMAP 2016 Extended Proceedings 1618 (2016)
5. Chin, D., Chin, D.: Empirical evaluation of user models and user-adapted systems. User Modeling and User-Adapted Interaction 11(1), 327–337 (2001)
6. Chin, D.N., Wright, W.R.: Evaluation metrics for inferring personality from text. EvalUMAP 16, UMAP 2016 Extended Proceedings 1618 (2016)
7. Conlan, O., Kelly, L., Koidl, K., Lawless, S., Levacher, K., Staikopoulos, A.: Evalumap2016: Towards comparative evaluation in the user modelling, adaptation and personalization space workshop (2016)
8. Hopfgartner, F., Kille, B., Lommatzsch, A., Plumbaum, T., Brodt, T., Heintz, T.: Benchmarking news recommendations in a living lab (2014)
9. Hui Yang, G., Soboroff, I.: Trec 2016 dynamic domain track overview (2016)
10. Höök, K.: Steps to take before intelligent user interfaces become real. Interacting with Computers 12(4), 409–426 (2000)
11. Kelly, L.: Living labs for umap evaluation. EvalUMAP 16, UMAP 2016 Extended Proceedings 1618 (2016)
12. Koidl, K., Levacher, K., Conlan, O., Steichen, B.: Ecp: Evaluation community portal a portal for evaluation and collaboration in user modelling and personalisation research. vol. 1618 (2016)
13. Hernández del Olmo, F., Gaudioso, E.: Evaluation of recommender systems: A new approach. Expert Systems with Applications 35(3), 790–804 (2008)
14. Pandit, H., Hamed, R.G., Lawless, S., Lewis, D.: The use of open data to improve the repeatability of adaptivity and personalisation experiment. EvalUMAP 16, UMAP 2016 Extended Proceedings 1618 (2016)
15. Paramythis, A., Weibelzahl, S., Masthoff, J.: Layered evaluation of interactive adaptive systems: framework and formative methods. User Modeling and User-Adapted Interaction 20(5), 383–453 (2010)
16. Park, K.S., Hwan Lim, C.: A structured methodology for comparative evaluation of user interface designs using usability criteria and measures. International Journal of Industrial Ergonomics 23(5-6), 379–389 (1999)
17. Staikopoulos, A., Conlan, O.: Towards comparative evaluations of user-adaptive software systems. EvalUMAP 16, UMAP 2016 Extended Proceedings 1618 (2016)
18. Vahid, A.H., Hamed, R.G., Koidl, K.: A review of user-centred information retrieval tasks. EvalUMAP 16, UMAP 2016 Extended Proceedings 1618 (2016)

19. Van Velsen, L., van der Geest, T., Klaassen, R.: Identifying usability issues for personalization during formative evaluations: A comparison of three methods. International Journal of Human-Computer Interaction 27(7), 670–698 (2011)
20. Van Velsen, L., van der Geest, T., Klaassen, R., Steehouder, M.: User-centered evaluation of adaptive and adaptable systems: a literature review. The Knowledge Engineering Review 23(3), 261–281 (2008)
21. Weibelzahl, S., Weber, G.: Advantages, opportunities, and limits of empirical evaluations: Evaluating adaptive systems. Künstliche Intelligenz 3, 17–20 (2002)
22. Yousuf, B., Conlan, O.: Motivating behavioral change through personalized visual narratives. EvalUMAP 16, UMAP 2016 Extended Proceedings 1618 (2016)