# Join the Living Lab: Evaluating News Recommendations in Real-Time

Frank Hopfgartner[1] and Torben Brodt[2]

[1] University of Glasgow, Glasgow, United Kingdom
frank.hopfgartner@glasgow.ac.uk
[2] plista GmbH, Berlin, Germany
tb@plista.com

**Abstract.** Participants of this tutorial learnt how to participate in CLEF NEWSREEL, a living lab for the evaluation of news recommender algorithms. Various research challenges can be addressed within NEWSREEL, such as the development and evaluation of collaborative filtering or content-based filtering strategies. Satisfying information needs by techniques including preference elicitation, pattern recognition, and prediction, recommender systems connect the research areas information retrieval and machine learning.

**Keywords:** recommender systems; living lab; user-centric evaluation; large-scale evaluation

## 1 Introduction

Thanks to de-facto standard evaluation measures, frameworks, and datasets, we are able to evaluate the performance of various aspects of information retrieval and recommender systems – also known as information access systems – and compare them to state-of-the-art approaches. In the context of information retrieval evaluation, benchmarking campaigns such as CLEF, TREC, or FIRE played an important role in establishing these evaluation standards. For recommender systems evaluation, the release of the Netflix Challenge dataset had a similar impact as it triggered further research in the field.

One of the main strengths of these benchmarking campaigns is the release of common datasets (e.g., [1]). On the one hand, the use of shared datasets has shown to be of great benefit for studying various aspects of information access systems as they can be used to fine-tune algorithms or models to increase standard evaluation metrics such as precision and recall. On the other hand, data-centric studies often ignore the role that the user plays in an information retrieval or recommendation scenario. It is the user's information need that needs to be satisfied and it is the user's personal interests that need to be considered when adapting retrieval results or when providing good recommendations. In particular, user-centric evaluation of information access systems (e.g., [2]) is essential in order to evaluate the full performance of adaptive (or personalised) approaches. Unfortunately though, most user studies lack of a large user base

which would be required to confirm research hypotheses. Hence, addressing this shortcoming, various methodologies have been suggested such as user simulation [3] or the evaluation of systems in a playful scenario [4]. Although these approaches can be used for "fine-tuning" of algorithms [5] or evaluation in a competitive environment, the artificial nature of this experimental setup casts some doubt on to which degree these findings can be generalised. User limitations are often not an issue for commercial providers of information access systems who have access to large user bases. Therefore, large-scale user-centric online evaluation, also referred to as A/B testing, is the first choice for the evaluation of commercial information systems.

Addressing this difference between academic and industry-based evaluation potentials, the application of a *living lab* has been proposed (e.g., [6, 7]) that grant researchers access to real users who follow their own information seeking tasks in a natural and thus realistic contextual setting. For user-centric research on information access systems, realistic context is essential since it is a requirement for a fair and unbiased evaluation. Kelly et al. [8] argue that "a living laboratory on the Web that brings researchers and searchers together is needed to facilitate ISSS [Information-Seeking Support System] evaluation. Such as lab might contain resources and tools for evaluation as well as infrastructure for collaborative studies. It might also function as a point of contact with those interested in participating in ISSS studies." Although the idea of such industry-academia partnership is not new, it was not until recently that the first living labs emerged that allow research in the fields. So far, two living labs have been established that focus on the evaluation of information retrieval (LL4IR) and recommender systems (NEWSREEL) algorithms, respectively. In this tutorial, the participants learnt how to participate in NEWSREEL, a living lab for the evaluation of news recommendations in real-time. The remainder of this paper is organised as follows. In Section 2, we introduce the news recommendation use case. Section 3 introduces the target audience. The format of the tutorial is outlined in Section 4.

## 2 News Recommendation Use Case

The living lab infrastucture that was introduced within this tutorial is provided by plista GmbH[3], a data-driven media company which provides content and advertising recommendations for thousands of websites (e.g., entertainment portals, news content pages). So far, the infrastructure has been used in the News Recommendation Challenge (NRS'13), held in conjunction with ACM RecSys 2013 and in NEWSREEL, a campaign-style evaluation lab that is organised as part of CLEF 2014 and 2015. In the remainder of this section, we briefly outline the new recommendation use case. For a more detailed description of the recommendation scenario, the provided content and its users, the reader is referred to [9]. An overview of the approaches of last year's participants of NEWSREEL 2014 is provided in [10].

[3] http://plista.com/

The use case centres around users who visit selected news portals. As described in [9], the vast majority of these users come from one of the German-speaking countries (Germany, Austria, Switzerland) in Central Europe. Whenever the users visit one of the selective news portals, a small widget box is displayed at the bottom or the side of the page labelled "Recommended articles" or "You might also be interested in". Within this box, the users can find a list of recommended news articles in the form of text snippets and small pictures. These recommendations are usually provided by plista. In the context of this living lab evaluation, plista provides an API that allows researchers to determine news articles that may be relevant for users who visited the page. Having a large customer base, plista processes millions of user visits on a daily basis. By providing the infrastructure of this living lab, they hence allow researchers to test and benchmark news recommendation algorithms in real-time by a large number of users.

## 3 Target Audience

Target audience were researchers in the field of information access systems with programming skills who are interested in evaluating recommender algorithms in a large scale by a large number of users. Focusing on above mentioned scenario, participants of this tutorial learnt how to implement their own recommendation algorithms and to benchmark them using the Open Recommendation Platform [11] which is the underlying platform of plista's living lab on real-time news recommendation.

## 4 Format of the Tutorial

The tutorial touched on two main research areas: (1) The development of web-based recommendation algorithms and (2) the evaluation of these techniques in real-time using real users in a large scale.

First, we introduced recommender systems from an academic point of view. We outlined central paradigms, state-of-the-art techniques, and existing evaluation protocols. Second, we presented the context of news recommendation. News recommendation entails a number of additional requirements. In particular, news recommender systems have to obey response time limitations. Further, we introduced the *Open Recommendation Platform* (ORP) [11] operated by plista. Participants learnt about its data structures, system components, and evaluation criteria. Finally, we immersed into existing implementations which participants can use to build their own news recommendation systems connected to ORP. Different APIs and SDKs were presented.

## References

1. Kille, B., Hopfgartner, F., Brodt, T., Heintz, T.: The plista dataset. In: NRS'13: Proceedings of the International Workshop and Challenge on News Recommender Systems, ACM (10 2013) 14–21

2. Hopfgartner, F., Jose, J.M.: Semantic user modelling for personal news video retrieval. In: Advances in Multimedia Modeling, 16th International Multimedia Modeling Conference, MMM 2010, Chongqing, China, January 6-8, 2010. Proceedings. (2010) 336–346
3. Hopfgartner, F., Urban, J., Villa, R., Jose, J.M.: Simulated testing of an adaptive multimedia information retrieval system. In: International Workshop on Content-Based Multimedia Indexing, CBMI '07, Bordeaux, France, June 25-27, 2007. (2007) 328–335
4. Schoeffmann, K., Ahlström, D., Bailer, W., Cobârzan, C., Hopfgartner, F., McGuinness, K., Gurrin, C., Frisson, C., Le, D., del Fabro, M., Bai, H., Weiss, W.: The video browser showdown: a live evaluation of interactive video search tools. IJMIR **3**(2) (2014) 113–127
5. White, R.W., Ruthven, I., Jose, J.M., van Rijsbergen, C.J.: Evaluating implicit feedback models using searcher simulations. ACM Trans. Inf. Syst. **23**(3) (2005) 325–361
6. Kamps, J., Geva, S., Peters, C., Sakai, T., Trotman, A., Voorhees, E.M.: Report on the SIGIR 2009 workshop on the future of IR evaluation. SIGIR Forum **43**(2) (2009) 13–23
7. Azzopardi, L., Balog, K.: Towards a living lab for information retrieval research and development - a proposal for a living lab for product search tasks. In: CLEF. (2011) 26–37
8. Kelly, D., Dumais, S.T., Pedersen, J.O.: Evaluation challenges and directions for information-seeking support systems. IEEE Computer **42**(3) (2009) 60–66
9. Hopfgartner, F., Kille, B., Lommatzsch, A., Plumbaum, T., Brodt, T., Heintz, T.: Benchmarking news recommendations in a living lab. In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings. (2014) 250–267
10. Kille, B., Brodt, T., Heintz, T., Hopfgartner, F., Lommatzsch, A., Seiler, J.: NEWSREEL 2014: Summary of the news recommendation evaluation lab. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. (2014) 790–801
11. Brodt, T., Hopfgartner, F.: Shedding Light on a Living Lab: The CLEF NEWSREEL Open Recommendation Platform. In: Proceedings of the Information Interaction in Context conference. IIiX'14, Springer-Verlag (2014) 223–226