Brodt, T., & Hopfgartner, F.

Shedding light on a living lab: the CLEF NEWSREEL open recommendation platform

Brodt, T., & Hopfgartner, F. (2014). Shedding light on a living lab. In Proceedings of the 5th Information Interaction in Context Symposium on - IIiX '14. ACM Press. https://doi.org/10.1145/2637002.2637028

Shedding Light on a Living Lab: The CLEF NEWSREEL Open Recommendation Platform

Torben Brodt
plista GmbH
10119 Berlin, Germany
tb@plista.com

Frank Hopfgartner
Technische Universität Berlin
10587 Berlin, Germany
frank.hopfgartner@tu-berlin.de

ABSTRACT

In the CLEF NEWSREEL lab, participants are invited to evaluate news recommendation techniques in real-time by providing news recommendations to actual users that visit commercial news portals to satisfy their information needs. A central role within this lab is the communication between participants and the users. This is enabled by The Open Recommendation Platform (ORP), a web-based platform which distributes users' impressions of news articles to the participants and returns their recommendations to the readers. In this demo, we illustrate the platform and show how requests are handled to provide relevant news articles in real-time.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software – performance evaluation (efficiency and effectiveness)

General Terms

Experimentation, Measurement, Performance

Keywords

real-time news recommendation, evaluation, living-lab

1. INTRODUCTION

One of the biggest challenges in the evaluation of information access systems is the limited access to real user interactions that would allow testing research hypotheses in a large scale. In order to address this issue, the idea of a living lab has been proposed (e.g., [15, 3, 4]), i.e., the provision of a shared experimental environment and real users to test information access approaches.

An example living lab is CLEF NEWSREEL[1] [13], a campaign-style evaluation lab on news recommendation in real-time which is organized as part of CLEF 2014. Within this lab, researchers can benchmark news recommendation techniques in real-time by recommending news articles to actual users that visit commercial news portals to satisfy their individual information needs, i.e., participants are facing real users in a living lab environment. Participants are given the opportunity to develop news recommendation algorithms and have them tested by potentially millions of users over the period of one year, or even longer. In this demo paper, we describe the infrastructure that is provided to enable this use case. We argue that this demo can shed light on the successful implementation of living labs.

The paper is structured as follows. In Section 2, we provide an overview of related work. Section 3 briefly introduce the CLEF NEWSREEL living lab scenario. Section 4 describes the open recommendation platform. In Section 5, we discuss the application of this platform.

2. RELATED WORK

Recommender systems are specific types of information access systems that focus on *proactively* assisting users in finding items (e.g., books, music, videos) that they were not looking for. This proactive element sets them apart from traditional information retrieval (IR) systems that *passively* provide information based on users' search queries. Despite this difference, the evaluation of both IR and recommender systems is based on protocols developed for the evaluation of digital libraries. Robertson and Hancock-Beaulieu [19] distinguish between two evaluation methodologies for digital libraries: system-driven evaluation and user-oriented evaluation. In the context of recommender systems, Gunawardana and Shani [12] refer to offline and online evaluation.

System-driven evaluation is based on the work of Cleverdon et al. [8], who promoted the use of test collections and closed laboratory environments to test parameters and algorithms. Given the success of the Text REtrieval Conference (TREC) [23], system-driven evaluation is the most dominant methodology to evaluate IR systems. In the recommender systems domain, various datasets have been released (e.g., [6, 9, 11, 24]) that allow system-driven – or offline – evaluation. Here, the most commonly used dataset consists of movie ratings which were released in the context of the Netflix Challenge [5].

User-oriented, or online evaluation, focuses more on evaluation in a realistic and operational environment. Differing from system-centered evaluation, no experimental restrictions with respect to users' search tasks, relevance assessments or other constraints are applied, i.e., users are using

[1] `http://clef-newsreel.org/`

information access systems to satisfy their own information need. Shani and Gunawardana [21] argue that online evaluation provides the strongest evidence on how well a recommender systems performs. Consequently, user-oriented evaluation plays an important role in the evaluation of recommender systems (e.g., [17]). A protocol for user-oriented evaluation, referred to as A/B testing, is outlined by Amatriain [2]. For an A/B test, users are split into different groups, each group interacting with a variant of a recommender system. By observing their behavior with these systems, conclusions about the quality of the recommendations can be drawn. Although the protocol is rather simple, it comes with a major drawback. In order to get meaningful results, a large user base is required. While this is no problem for commercial providers, the lack of access to actual users hinders research at universities significantly.

Addressing this issue, Kelly et al. [15] argue for the implementation of a living lab that brings together researchers and users "to facilitate ISSS [Information-Seeking Support System] evaluation". Pirolli [18] argues that such living labs could attract researchers from many different domains. Although the idea has been discussed on multiple occasions (e.g., [14, 1, 4]), living labs for the evaluation of information retrieval evaluation have not been established yet.

A first proposal for a living lab for information retrieval research is proposed by Azzopardi and Balog [3] who define an infrastructure that allows different parties to interact with each other. Moreover, they illustrate how this infrastructure can be used in a specific use case. Although their work can be considered to be a key contribution for the definition of living labs, their work remains theoretical. In this paper, we introduce the application of a living lab for the benchmarking of news recommendation algorithms in real time. Within this living lab, different parties interact with each other using a shared infrastructure: Users visit news portals of commercial providers, these visits are reported to researchers whose task is to identify other news articles of this provider which are then recommended to the user for further reading. To the best of our knowledge, it is the first living lab for the evaluation of information access systems. An overview of this living lab is provided in the next section.

3. THE CLEF NEWSREEL LIVING LAB

CLEF NEWSREEL [13] is a campaign-style living lab that is organized as part of CLEF 2014. It is a continuation of the News Recommender Systems challenge [22] which was organized in conjunction with ACM RecSys 2013.

In the NEWSREEL scenario, users visit a commercial news portal and read an online article. On the bottom of the page, they find a small widget box labelled "You might also be interested in", "Recommended articles", or similar where they can find a list of recommended news articles. Dependent on the actual content provider, these recommendations often consist of a small picture and accompanying text snippets.

While some publishers provide their own recommendations, more and more providers rely on the expertise of external companies such as plista[2] who do provide such recommendation services. plista is a company that runs a content and ad recommendation service on thousands of premium websites (e.g., news portals, entertainment portals). In order to outsource this recommendation task to plista, the publishers firstly have to inform them about newly created articles and updates on already existing articles on their news portal. In addition, whenever a user visits one of these online articles, the content provider forwards this request to plista. These clicks on articles are also referred to as impressions. Plista then determines related articles which are then forwarded to the user and displayed in above mentioned widget box as recommendations. Having a large customer base, plista processes millions of user visits in real time on a daily basis.

In the context of the NEWSREEL lab, plista grants participating research teams access to a certain amount of these requests. The lab consists of two tasks. In the first task, participants have to predict clicks in a comprehensive dataset [16] that has been recorded in June 2013. The dataset includes both user and item features along with interactions in between them. Interactions are characterized as either clicks (a user clicked on a recommended article) or impressions (a user reads an article). Figure 1 visualizes the number of impressions over time for an exemplary news domain. A preliminary analysis of the dataset is provided by Esiyok et al. [10].

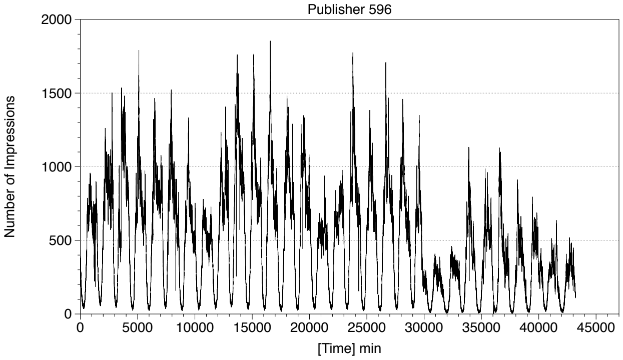


Figure 1: Recorded impressions over time for an exemplary news domain.

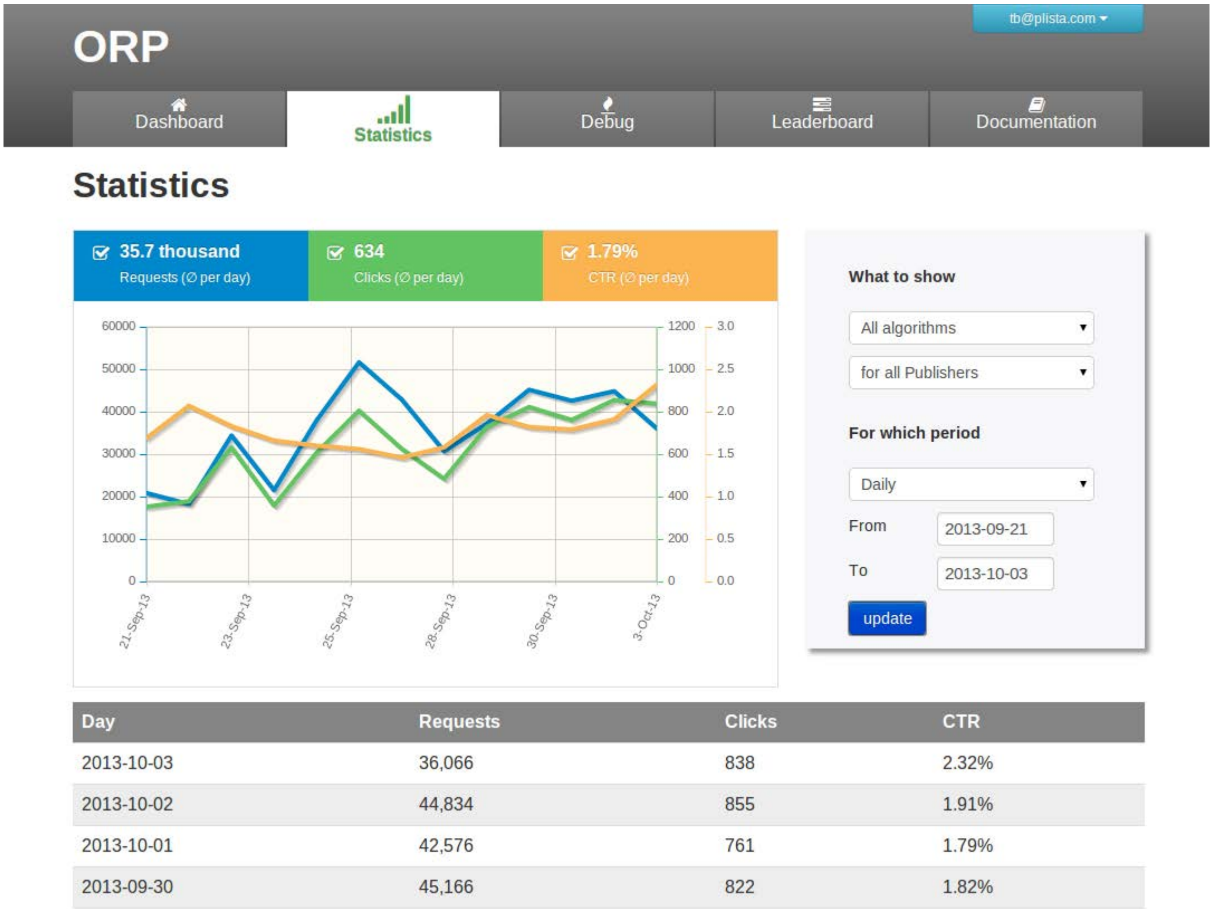
In the second task, participants are asked to provide recommendations in real-time for actual users, i.e., the list of related articles is not determined by plista, but by the participating research teams. The communication between the participants and plista, as well as the monitoring and evaluation is handled by the Open Recommendation Platform (ORP) which is outlined in the next section.

4. THE OPEN RECOMMENDATION PLATFORM

The distributed Open Recommendation Platform[3] (ORP) is capable of delivering different recommendation implementations and to take track of the recommender results. The platform was opened either to allow researchers to try out their ideas in a real world scenario or to profit from the research community by connecting their expertise. The best algorithms are chosen in real-time using a multi-armed bayesian bandit [20]. The different results of many recommenders might be blended together in the future. There are a dozen

[2] http://www.plista.com/

[3] http://orp.plista.com/



Statistics

35.7 thousand Requests (∅ per day)

634 Clicks (∅ per day)

1.79% CTR (∅ per day)

What to show

All algorithms

for all Publishers

For which period

Daily

From 2013-09-21

To 2013-10-03

Day	Requests	Clicks	CTR
2013-10-03	36,066	838	2.32%
2013-10-02	44,834	855	1.91%
2013-10-01	42,576	761	1.79%
2013-09-30	45,166	822	1.82%

Figure 2: Screenshot of The Open Recommendation Platform.

of premium publishers already activated for researchers. After researchers register for the service, they need to provide a server address on which their implementation of a recommender is running.

4.1 Requests

ORP will send requests using HTTP POST requests including item updates, event notifications and recommendation requests. Event notifications are the actual user interactions, i.e., users' visits, referred to as impressions, to one of the news portals that rely on the plista service, or clicks to one of the recommended articles. The item updates include information about the creation of new pages on the content providers' server and it allows participants to provide content-based recommendations. Recommender algorithms and evaluation models can also be build on top of the context, which includes the user id provided by a cookie, publisher id, browser, device, operating system and more, either from the http context or additionally being enhanced by plista using categorization heuristics and classifiers. Expected responses to the recommendation requests are related news article from the same content provider, which are then provided as recommendations to the visitors of the page. These recommendations are usually displayed at the end of an article.

Since recommendations need to be provided in real-time, the expected response has to be send within 100ms, i.e., recommenders have to be quick. If too much time is lost due to network latency (e.g., when the participant has a slow internet connection or is physically remote from the ORP server), the algorithms can also be installed on a server provided by plista.

4.2 Data Format

The API uses JSON for data encoding. The contextual data in the ORP is represented through vectors. They are identified by numeric IDs and are associated with elements of various data types. Vectors allow to describe an object by layering attributes. All vectors belong to one of two classes: input vectors or output vectors. Input vectors describe the context of events and messages and may be used by a partner for contextual optimization. Input vectors are static and can not be modified. Output vectors are used to convey information about calculation results. During transmission, vectors are grouped together by their type and packaged in a map where the key is the vector ID and the value is the associated value of the vector (depending on its type). The vectors group maps are again grouped together depending on their class. For further details about the data format, the reader is referred to [7].

4.3 Graphical User Interface

When logging in, the participants can see their algorithms' performance over time. Performance is measured in impressions, clicks and click-through rate (CTR) per day. An impression record is created whenever a user reads an article and the participant received a request to provide recommendations for this visit. Clicks represent users following links to articles that have been recommended while reading a news article. CTR is defined as the ratio of clicks over impressions. A screenshot of the ORP GUI is shown in Figure 2.

5. DISCUSSION AND CONCLUSION

In this demo, we showcase the underlying technology that allows us to run a living lab on real-time news recommendation. The platform has been used to run a news recommender challenge, co-located with ACM RecSys'13. To reduce the efforts for the campaign-style living lab that is part of CLEF'14 a basis implementation on top the Apache Mahout framework is available[4]. We argue that the underlying infrastructure can serve as a prototypical implementation of other living labs that support further research in the evaluation of information access systems.

Acknowledgement

The work leading to these results has received funding (or partial funding) from the Central Innovation Programme for SMEs of the German Federal Ministry for Economic Affairs and Energy, as well as from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 610594.

6. REFERENCES

[1] M. Agosti, N. Fuhr, E. Toms, and P. Vakkari. Evaluation Methodologies in Information Retrieval (Dagstuhl Seminar 13441). *Dagstuhl Reports*, 3(10):92–126, 2014.

[2] X. Amatriain. Mining large streams of user data for personalized recommendations. *ACM SIGKDD Explorations Newsletter*, 14(2):37, Apr. 2013.

[3] L. Azzopardi and K. Balog. Towards a living lab for information retrieval research and development - a proposal for a living lab for product search tasks. In *CLEF*, pages 26–37, 2011.

[4] K. Balog, D. Elsweiler, E. Kanoulas, L. Kelly, and M. Smucker. Report on the CIKM Workshop on Living Labs for Information Retrieval Evaluation. *SIGIR Forum*, 48(1), 2014.

[5] R. M. Bell, Y. Koren, and C. Volinsky. All Together Now: A Perspective on the Netflix Prize. *Chance*, 23(1):24–29, 2010.

[6] J. Bennett and S. Lanning. The netflix prize. In *KDDCup'07*, 2007.

[7] T. Brodt, T. Heintz, A. Bucko, and A. Palamarchuk. ORP protocol. `http://orp.plista.com/documentation/download.pdf`, 2013.

[8] C. Cleverdon, J. Mills, and M. Keen. Factors determining the performance of indexing systems. Technical report, ASLIB Cranfield project, Cranfield, 1966.

[9] G. Dror, N. Koenigstein, Y. Koren, and M. Weimer. The Yahoo! Music Dataset and KDD-Cup. In *JMLR: Workshop and Conference Proceedings*, pages 3–18, 2012.

[10] C. Esiyok, B. Kille, B. J. Jain, F. Hopfgartner, and S. Albayrak. Users' reading habits in online news portals. In *IIiX'14*. ACM, 08 2014. to appear.

[11] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[12] A. Gunawardana and G. Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10:2935–2962, 2009.

[13] F. Hopfgartner, B. Kille, A. Lommatzsch, T. Plumbaum, T. Brodt, and T. Heintz. Benchmarking news recommendations in a living lab. In *CLEF'14: Proceedings of 5th International Conference of the Cross-Language Evaluation Forum*, CLEF'14. Springer-Verlag, 2014. to appear.

[14] J. Kamps, S. Geva, C. Peters, T. Sakai, A. Trotman, and E. M. Voorhees. Report on the sigir 2009 workshop on the future of ir evaluation. *SIGIR Forum*, 43(2):13–23, 2009.

[15] D. Kelly, S. T. Dumais, and J. O. Pedersen. Evaluation challenges and directions for information-seeking support systems. *IEEE Computer*, 42(3):60–66, 2009.

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

[17] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, Oct. 2012.

[18] P. Pirolli. Powers of 10: Modeling complex information-seeking systems at multiple scales. *IEEE Computer*, 42(3):33–40, 2009.

[19] S. E. Robertson and M. Hancock-Beaulieu. On the Evaluation of IR Systems. *Information Processing and Management*, 28(4):457–466, 1992.

[20] J. Seiler, A. Bucko, and T. Heintz. Bayesian model averaging for online ensemble learning in news article recommendations. In *NRS'13: Working Notes of the International Challenge on News Recommender Systems*, 10 2013.

[21] G. Shani and A. Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.

[22] M. Tavakolifard, J. A. Gulla, K. C. Almeroth, F. Hopfgartner, B. Kille, T. Plumbaum, A. Lommatzsch, T. Brodt, A. Bucko, and T. Heintz. Workshop and challenge on news recommender systems. In *RecSys'13: Proceedings of the International ACM Conference on Recommender Systems*. ACM, 10 2013.

[23] E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, USA, 1 edition, 2005.

[24] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *WWW'05*, pages 22–32. ACM, 2005.

[4] `https://github.com/plista/orp-kornakapi-mahout`