CLEF NewsREEL 2016: Comparing Multi-dimensional Offline and Online Evaluation of News Recommender Systems

Benjamin Kille[1], Andreas Lommatzsch[1], Frank Hopfgartner[2], Martha Larson[3], Jonas Seiler[4], Davide Malagoli[5], András Serény[6], and Torben Brodt[4]

[1] TU Berlin, Berlin, Germany
{benjamin.kille,andreas.lommatzsch}@dai-labor.de
[2] University of Glasgow, Glasgow, UK
frank.hopfgartner@glasgow.ac.uk
[3] TU Delft, Delft, The Netherlands
m.a.larson@tudelft.nl
[4] Plista GmbH, Berlin, Germany
{torben.brodt,jonas.seiler}@plista.com
[5] ContentWise R&D — Moviri, Milan, Italy
davide.malagoli@moviri.com
[6] Gravity R&D, Budapest, Hungary
sereny.andras@gravityrd.com

Abstract. Running in its third year at CLEF, NewsREEL challenged participants to develop news recommendation algorithms and have them benchmarked in an online (Task 1) and offline setting (Task 2), respectively. This paper provides an overview of the NewsREEL scenario, outlines this year's campaign, presents results of both tasks, and discusses the approaches of participating teams. Moreover, it overviews ideas on living lab evaluation that have been presented as part of a "New Ideas" track at the conference in Portugal. Presented results illustrate potentials for multi-dimensional evaluation of recommendation algorithms in a living lab and simulation based evaluation setting.

Keywords: recommender systems · news · multi-dimensional evaluation · living lab · stream-based recommender

1 Introduction

Businesses continuously optimize their recommender systems to provide a satisfying user experience. Their efforts concern different aspects including the degree to which they are able to match users' preference and system availability. Gomez-Uribe and Hunt [6] describe how Netflix compares recommendation algorithms by A/B testing and offline experiments. Researchers in academia lack access to operating recommender systems. Instead, they rely on data sets comprising recorded interactions between users and items. Evaluation with data sets neglects some key practical challenges that recommender systems face. First, recommender systems have to handle a continuous stream of requests. They cannot afford to let users wait until recommendations are computed. Depending on context, recommendations have to be provided instantaneously. Second, the collections of users and items fluctuate. Users sign up or quit using the service. Items emerge or become obsolete. Data sets fail to reflect these dynamics. CLEF NewsREEL represents a unique opportunity for researchers to experience the evaluation of recommender systems in real-world settings. Participants have access to a large-scale data set that can be used for a "traditional" offline evaluation. In addition, participants can connect to a recommender system that provides suggestions to readers of online news articles from several publishers. The system tracks how well the participants' algorithms perform in terms of clicks and response rates. This year's edition of NewsREEL emphasizes the multi-dimensional evaluation paradigm. Participants are asked to measure their offline success rate as well as response time. In this way, we can compare functional as well as non-functional performance criteria offline and online.

Extending [12], in which we outline specifics of the NewsREEL 2016 campaign, in this paper focuses on presenting the results of the campaign. The paper is structured as follows. In Section 2, we briefly outline the recommendation scenario that is addressed by NewsREEL. In Section 3, we provide an overview of the teams that registered to participate. Results are presented in Section 4 and discussed in Section 5. In order to discuss further steps in the realization of the living lab evaluation paradigm, we jointly organized a "New Ideas" track with the organizers of CLEF LL4IR at this year's CLEF conference. An overview of the papers that have been accepted in this track is given in Section 6. The paper concludes in Section 7.

2 Lab Set Up

NewsREEL features two tasks related to news recommender systems. Task 1 addresses online evaluation protocols. Task 2 asks participants to conduct experiments on previously recorded interactions. Participants engaging in both tasks observe functional as well as non-function characteristics of their algorithms. They determine how accurately their methods predict users' preferences. Additionally, they assess how well their systems handle realistic conditions. In Task 1, their systems need to reply in less than 100 ms. In Task 2, they process a large amount of data. In the remainder of this section, we provide a brief overview of the two tasks. For a more detailed description of the NewsREEL scenarios, we refer to [8].

2.1 Task 1: Benchmarking News Recommendation in a Living Lab

In Task 1 of the campaign, participants had the opportunity to benchmark news recommendation algorithms in a living lab setting. The scenario has been described in detail in [9]. In order to participate, participants had to connect a recommendation service to the Open Recommendation Platform (ORP) [1]. ORP orchestrates the workflow between recommendation services and visitors of online news websites. Visitors trigger recommendation requests as they access news articles. ORP receives the requests and randomly forwards them to a recommendation services. Having received a list of suggested news items, ORP delivers these to the visitors. This setting lets participants experience the requirements that operating news recommender systems must fulfill. ORP keeps track of participants' performance in terms of click-through-rate (CTR) and response rate. CTR represents the proportion of recommendations that users clicked. Response rate represents the proportion of requests recommendation services successfully responded to.

2.2 Task 2: Benchmarking News Recommendations in a Simulated Environment

Task 2 focused on benchmarking news recommendations in a simulated environment. For further details on the experimental setup, we refer to [13]. This year, participants received a large-scale data set comprising of more than 100 million interactions. All interactions are associated with a timestmap. For this reason, participants can replay events in chronological order. In this way, they simulate a realistic setting for news recommender systems. In contrast the case of Task 1, participants cannot distinguish requests from notifications. They respond to all interactions as if they were recommendation requests. We determine performance in terms of correct predictions and response time distribution. Whenever we detect interaction between a user and a suggested item within a future window of 5 min, we consider the recommendation successful. We compute the proportion of successful recommendation to the total number of recommendation obtaining a quantity similar to the CTR. In addition, we measure how much time elapses until the recommendation service provides suggestions. Consequently, we consider the distribution of response times to estimate how well systems scale to heavy loads.

3 Participation

In the 2016 edition of NewsREEL, 48 participants registered. Both tasks attracted similarly many participants with Task 1 slightly ahead with 46 registrations compared to Task 2 with 44 participants. Participants deployed 73 recommendation services in Task 1. We received a total of seven working notes describing participants' approaches.

Figure 1 illustrates how registrations are distributed across the globe. All continents (except Antarctica) have at least one participant who registered for NewsREEL 2016. China, Europe, India, and Turkey appear particularly interested in NewsREEL.

4 Results

In the remainder of this section, we first, in Section 4.1, outline the evaluation periods chosen for the evaluation of algorithms in Task 1 and present the results of these periods. Then, we present the results for Task 2 in Section 4.2. Finally, we summarize the participants' approaches in Section 4.3.

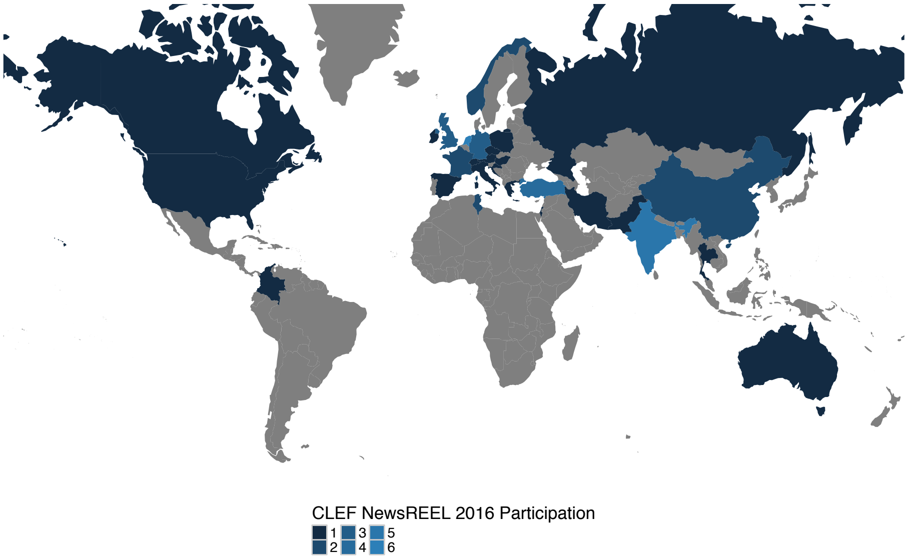

Fig. 1. Worldwide participation in NewsREEL 2016. Color shade represents the number of participants per county.

4.1 Task 1

Participants had access to ORP during the entire CLEF cycle 2015-10-30 to 2016-05-25. This facilitated exploring a variety of algorithms, optimizing parameters, and testing hardware settings. Using ORP may initially challenge participants. They have to maintain the system. This involves monitoring, updating, and troubleshooting. The effort exceeds experimentally determining the best ranking. Participants received detailed feedback for three periods dedicated for testing:

- 06–12 February 2016
- 20–26 February 2016
- 05–11 March 2016

The winner was intended to be determined in a four-week period scheduled for 2–29 April 2016. Unfortunately, ORP exhibited a malfunction in early April. Participants received only part of the information in form of requests but lacked notification messages. We cannot eliminate performance differences that arise by chance unless we observe them over sufficiently long time. For this reason, we extended the evaluation period until 20 May 2016. Table 1 summarizes our observations. We notice that the level of engagement varied among participants. Some settled for individual strategies while others enter up to 14 different methods. Some competed for 49 days. Others had systems turned off for some days. The average click-through-rate varied from 0.42 % to 1.23 %. Some participants managed to achieve a response rate close to 100 %.

Figure 2 illustrates the performance in more detail. Each triangle corresponds to an algorithm which served recommendations to ORP. The triangles' sizes reflect the number of days which the algorithm has been active. The larger the triangle the more days the algorithm has been active. The triangles' positions indicate the average CTR per day. We observe that most algorithms achieve CTR between 0.5 % to 1.0 %. Algorithms with fewer requests manage to exceed the 1 per cent mark. We need to be careful as we interpret these performances given that these algorithms served noticeably fewer requests.

Figure 3 depicts the relation between response rate and number of requests. Each triangle represents an algorithm's response rate and number of requests for the same day. We notice that response rates stretch almost the entire interval $[0, 1]$. An exponential fit shows that higher response rates promise more requests. In addition, we observe a cluster of measurements with high response rate yet relatively low numbers of requests. These systems could perform well yet be switched off for parts of the day.

4.2 Task 2

The Offline Evaluation task has attracted several teams. The participating teams focused on two main research questions. First, teams mined the provided offline log files in order to find characteristic patterns and features useful for predicting the relevance of an article. Second, the teams analyzed the technical complexity of the algorithms by examining how the number of concurrent requests influences the response time and the throughput.

Table 1. In Task 1, we observe 17 participants competing in the time frame starting on 2016-04-02 and ending on 2016-05-20. They explored between 1 to 14 algorithms individually amounting to a total of 74 algorithms (cf. column A). Their activity stretched from 11 to 49 days (cf. column D). Team "is@uniol" registered most clicks and requests.

Team	A	D	Requests	Responses	CTR	RR
abc	3	49	446 372	3704	0.83%	99.86%
artificial intelligence	1	48	356 504	3175	0.89%	97.39%
baseline	1	37	123 251	783	0.64%	70.53%
berlin	2	49	331 729	2514	0.76%	95.40%
breaking-news	10	23	26 523	111	0.42%	30.63%
cwi	4	48	1 411 916	11 418	0.81%	96.72%
fc_tudelft	2	49	394 419	3049	0.77%	94.06%
flumingsparkteam	1	33	249 908	1465	0.59%	99.75%
is@uniol	10	49	2 316 399	21 044	0.91%	91.05%
moldawien-madness	1	12	67 103	471	0.70%	99.25%
news_ctr	2	28	79 261	803	1.01%	63.26%
riadi-gdl	6	46	470 484	3402	0.72%	76.98%
riadi-gdl-2	5	18	162 253	1321	0.81%	71.38%
sparkdev	1	42	20 851	102	0.49%	79.05%
wise	2	11	105 173	843	0.80%	99.40%
xyz	8	47	444 101	3838	0.86%	67.27%
xyz-2.0	14	47	320 689	3956	1.23%	56.56%

Studying the Lifecycle of very Popular News Items Since most popular recommender algorithms have been shown to be very successful in the NewsREEL challenge, several teams analyze how the popularity of items changes over time. The conducted experiments show that popular items receive a lot of intention in the first hour after the release date. Surprisingly, the very popular items stay popular for 2–4 days. This means, based on the number of impressions these article are still popular several days after the release. For several very popular articles, it was shown that the news article text has been updated. These news article updates seem to increase the interest in older news items. The lifecycle of the most popular articles depends on the specific publisher. Several teams showed that the parameters of the recommender algorithms must be optimized in order to meet the characteristics of the different news portals for that recommendations must be provided in the NewsREEL challenge.

Technical aspects The scalability has been analyzed by several teams in NewsREEL Task 2 ("online simulated stream"). The experiments show that the typical load ($\approx 10 - 50$ concurrent requests) is successfully handled by all the recommender approaches that were implemented. This is underlined by the low error rate in the Living Lab Evaluation. For handling extreme load peaks the model building and the computation of recommendations should be separated. The separation enables the use of sophisticated model building algorithms without slowing down the provision time of recommendation results. Big data frameworks (such as Apache Spark and Apache Flink) are powerful tools for implementing machine learning algorithm and for build-

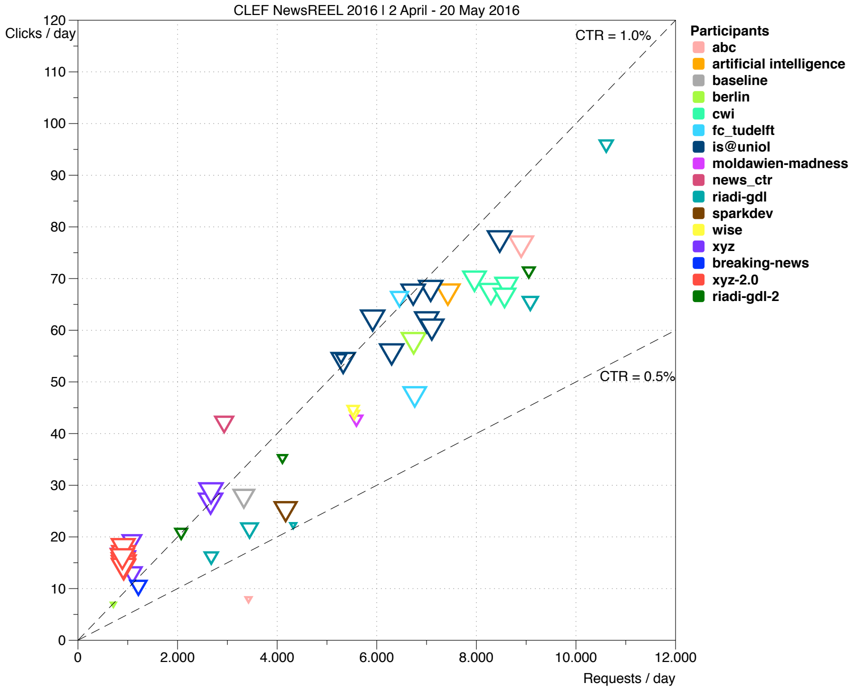

Fig. 2. Participants responded to recommendation requests and ORP tracked whenever users clicked on the suggestions. For each algorithm we computed the mean number of requests and clicks per day. The figure illustrates the CTR for each algorithm. The size of the symbol reflects how many days the algorithm was available.

ing recommender models. These frameworks are optimized for the processing of huge data collections, but cannot fulfill the real-time requirements. In the NEWSREEL model, updates in real-time are not needed; it is sufficient to update the recommender models on a minute=by-minute basis. For this reason, the big data frameworks are suitable frameworks for processing huge data streams; the computation of recommendations based on the models created should be done by a separate component optimized for providing results in real time.

Comparison Online/Offline The comparison of online and offline Evaluation has been analyzed by different teams. In order to ensure similar conditions in the online and the offline evaluation, the teams used in the offline evaluation the log data for timeframes detailed online evaluation data have been provided by PLISTA.

The evaluation of the Click-Through-Rate shows that there is no direct correlation between the online CTR and the offline CTR. This can be explained by the fact, that the offline CTR is computed based on the impression stream abstracting from the placement of recommendation results. A high CTR in the offline evaluation does not mean that the recommender reaches a high CTR in the Living Lab evaluation. A high

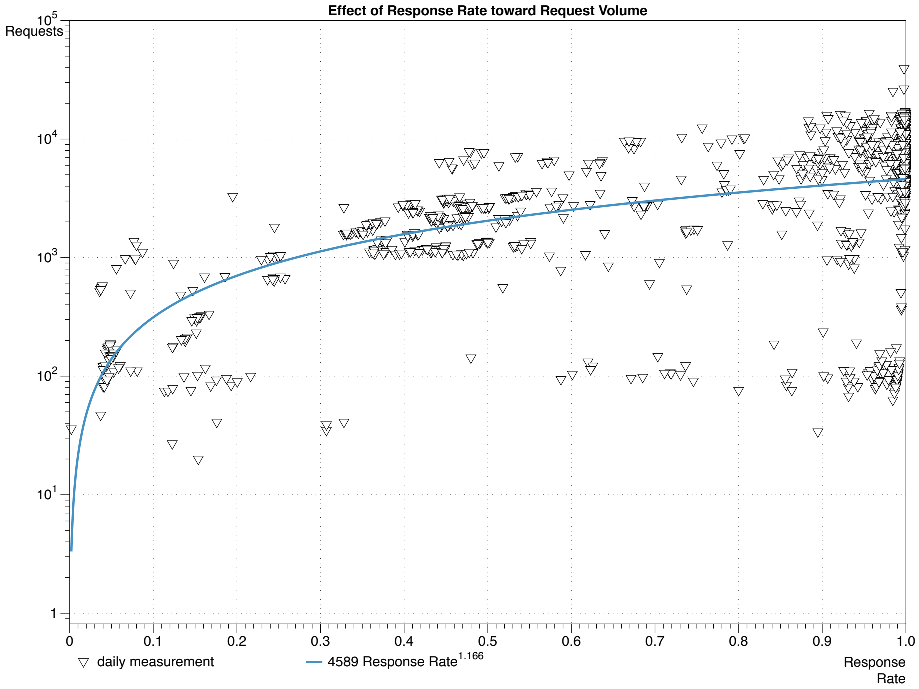

Fig. 3. Participants received recommendation requests from ORP. The figure illustrates the relation between the response rate and the number of requests. The fitted line indicates a noticeably reduced numbers of requests at low response rates.

CTR in the offline evaluation shows that the recommender works correctly and can be further analyzed based on live user feedback.

4.3 Working Notes Summary

Domann et al. [4] describe how they used APACHE SPARK. They argue that scalability manifests a major concern for real-world news recommender systems. APACHE SPARK fosters scalability by distributing computation among connected workers. The evaluation reports superior offline performance to the baseline in terms of CTR. Conversely, the authors find that their system achieves high availability with a response rate of more than 99 per cent.

Lommatzsch et al. [14] analyze most-popular strategies for news recommendation. The manuscript introduces a tool to monitor how articles' popularity develops over time. The authors propose a popularity-driven recommender, which adjusts its parameter for publishers individually. The evaluation indicates that the proposed method achieve a CTR superior to the baseline. Simultaneously, the proposed method scales well with increasing numbers of requests and manages to keep to response time limits.

Ciobanu and Lommatzsch [2] discuss how APACHE FLINK can be used to build a most-popular news recommender system. The authors stress that real-world news recommender systems require scalable algorithms. They introduce APACHE FLINK as a framework supporting scalable stream processing. The authors evaluate how changing time interval and model updating affects the recommender's accuracy. Moreover, they report that by using APACHE FLINK, their system achieved a more than 95 per cent response rate.

Yuan and Lommatzsch [17] analyze click patterns in time series from NewsREEL 2016. They show that a limited set of news items attract a majority of clicks, and that they continue to dominate for longer times than expected. The manuscript presents a series of experiments in the context of online news recommender system evaluation. The authors report that content-based methods achieve considerably lesser click-through-rates than popularity-based methods.

Probst et al. [15] describe how they used the AKKA framework to implement a news recommendation strategy. First, they focus on a popularity-based recommender. Second, they implement a delegation strategy including a popularity-based, recency-based, and category-based recommendation algorithm. They report superior CTR compared to the baseline without having to sacrifice scalability. Additionally, they conduct an A/A-test indicating marginal variance between four instances of the same algorithm.

Corsini and Larson [3] discuss how images affect users' response to recommendations. They argue that selecting promising images increases the likelihood of clicks. They introduce an image processing pipeline. The pipeline detects faces and image salience. A binary classifier subsequently decides whether an image is interesting or not. The authors evaluate the approach offline and online. They report improvements in the offline case. Further work is necessary to achieve reliable online evaluation results.

Gebremeskel and de Vries [5] are interested in achieving a deeper understanding of the source of the difference in performance between online and offline evaluation. They investigate differences between different streams sent by ORP to the challenge participants, with the eye to determining whether they can be explained by chance, or whether there are more systematic effects that might cause differences between stream. Further, they compare performance measurements made in 2015 and 2016. Their results point to a persistent gap between online and offline evaluation: one does not predict the other. Further, the relative performance of the set of basic algorithms that they investigates is different between 2015 and 2016, suggesting that more work is necessary to understand the reproducibility of online evaluation.

5 Discussion

News recommender systems ought to accurately predict which news articles visitors would like to read. NewsREEL offers participants the chance to evaluate how well their algorithms perform compared to competitors. In addition, participants get to experience a realistic scenario for operating news recommender systems. They can assess non-functional aspects such as scalability.

In Task 1, we observed how participants' recommender systems performed in terms of CTR and response rate. A malfunction of ORP rendered part of the evaluation period incomparable. Participants only received a subset of the information that is typically available. We noticed that some participating systems received considerable more requests than others. According to our expectations, all systems should receive similar numbers of requests as long as they are similarly available. Deviating numbers of requests emerge when systems are disabled or produce invalid or delayed suggestions. We observe both phenomena. Figure 3 shows that systems producing more errors received a smaller number of requests. In addition, some systems achieved high response rates but still received fewer requests. This can be caused when participants temporarily disable their systems. ORP only sends requests to enabled systems which is why disabled systems receive fewer requests.

The question of whether systems with different numbers of requests can be meaningfully compared draw the attention of several participants (cf. [**?**] and [15]). They conducted experiments using identical instances of an algorithm. They expected to observe similar results for those systems. They noticed some variations that could be caused by the fact that they served different users. Still, the individual CTR of identical methods converged over time hardly distinguishable values. The amount of time needed to ensure convergence is still a question open to investigation.

Participants used various frameworks to ensure adequate scalability. APACHE SPARK, APACHE FLINK, and AKKA achieved response rates close to 100 per cent. At the moment, we cannot compare these frameworks since each has been used by different individual participants. In a future edition, it would be interesting to further investigate the advantages and disadvantages to which each framework subjects its users.

6 New Ideas Track

This year's CLEF NewsREEL lab featured a New Ideas Track. The track was a joint initiative between CLEF LL4IR and CLEF NEWSREEL. The track called for people's ideas on how the living labs methodology might be further developed into the future. This included challenge discussion, people's ideas related to existing living labs tasks and also ideas for new living labs tasks describing new use-cases or design ideas related to living labs. These ideas could be speculative, opening up new challenges in the space or more concrete task proposals for future living labs tasks the authors would like to run. The following three position papers were selected for presentation at the CLEF conference in Portugal:

Kelly [11] put forward using a research-centers focused living labs approach for shared task generation for the user modeling, adaptation and personalization (UMAP) space. The basic idea of this approach is to distribute goal-specific tools and protocols, along with challenge participants developed techniques, to research centers participating in the shared task. The research centers then recruit experiment subjects locally to test the techniques on their local PCs, as they go about their normal activities, using the provided tools and protocols.

Jabeur et al. [10] proposed to extend the LL4IR framework with a 'Living Ranking' component with the aim of providing an evaluation framework for real-time ranking. The proposed extension enables the delivery of a real-time ranking for fresh queries while maintaining the simplicity of the LL4IR framework. In particular, participants need to provide an algorithm that will be executed in real time.

Schaer and Tavakolpoursaleh [16] introduced the idea of developing a standard extension for common search engines and repository systems. They argue that this would not only increase the number of possible living labs participants on the site level but would additionally bring some other benefits like common standards and practices. Moreover, they summarize their experience of including the LL4IR API into a operating search engine.

7 Conclusion

In this paper, we provided an overview of CLEF NewsREEL 2016. Similar to last year's campaign, the 2016 lab focused on two different evaluation paradigms, namely online and offline benchmarking of news recommendation algorithms. Differing from last year where most participants focused on the online evaluation scenario only that was addressed in Task 1, this year, various teams explored both tasks. Although the main evaluation metric used at NewsREEL focused on the quality and suitability of the recommendations, measured using the users' clicking behavior, participants focused in their work on different aspects of news recommendation performance. While some focused on scalability issues of their methods or frameworks, others focused on other aspects such as monitoring popularity of news articles over time or the potentials of the NewsREEL setup for the comparison of A/B testing and offline evaluation.

Concluding from this year's lab and submissions, we argue that NewsREEL provides an excellent opportunity for exploring novel evaluation paradigms in the field of information access research. In particular, we argue that NewsREEL can serve as a prime example of the idea of providing evaluation as a service as discussed by Hanbury et al. [7]. Moreover, as preliminary results submitted by the participants suggest, the lab provides appropriate facilities to compare online and offline evaluation of recommendation algorithms. This also includes further research on multi-dimensional evaluation of recommender systems.

This year has opened several specific topics, which we have identified as particularly relevant moving forward. First, we have uncovered radical imbalances in the click distribution over new items. A limited number news items attract a majority of clicks. The popularity of a few items cannot be attributed to an emphemeral trend, since it endures over a period of several days. Second, we have established that it is worthwhile to attempt to exploit the characteristics of images when making recommendation. Simple experiments involving automatic face detection and generation of salience scores hint that even more dramatic improvements could be achieved by investing additional efforts in approaches that make use of image processing techniques. Finally, we note that reproducibility appears to be a very elusive goal: differences between online and offline evaluation must be better understood. In particular, factors that have been previously ignored, such as the display position of results on webpages, might actually have more impact than we realize. Also, further attention is necessary to fully understand the conditions under the results of different online tests can be considered fairly comparable.

Acknowledgments

The research leading to these results was performed in the CrowdRec project, which has received funding from the European Union Seventh Framework Program FP7/2007–2013 under grant agreement No. 610594. The “New Ideas” track was jointly organized with the organizers of CLEF LL4IR.

References

1. T. Brodt and F. Hopfgartner. Shedding light on a living lab: the CLEF NEWSREEL open recommendation platform. In *IIiX 2014 : Fifth Information Interaction in Context Symposium*, pages 223–226. ACM, August 2014.
2. A. Ciobanu and A. Lommatzsch. Development of a News Recommender System based on Apache Flink. In *Working Notes of the 7th International Conference of the CLEF Initiative, Evora, Portugal*. CEUR Workshop Proceedings, 2016.
3. F. Corsini and M. Larson. CLEF NewsREEL 2016: Image based Recommendation. In *Working Notes of the 7th International Conference of the CLEF Initiative, Evora, Portugal*. CEUR Workshop Proceedings, 2016.
4. J. Domann, J. Meiners, L. Helmers, and A. Lommatzsch. Real-Time News Recommendations Using Apache Spark. In *Working Notes of the 7th International Conference of the CLEF Initiative, Evora, Portugal*. CEUR Workshop Proceedings, 2016.
5. G. Gebremeskel and A. de Vries. Recommender Systems Evaluations: Offline, Online, Time and A/A Test. In *Working Notes of the 7th International Conference of the CLEF Initiative, Evora, Portugal*. CEUR Workshop Proceedings, 2016.
6. C. A. Gomez-Uribe and N. Hunt. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Transactions on Management Information Systems*, 6(4):13:1–13:19, 2015.
7. A. Hanbury, H. Müller, K. Balog, T. Brodt, G. V. Cormack, I. Eggel, T. Gollub, F. Hopfgartner, J. Kalpathy-Cramer, N. Kando, A. Krithara, J. J. Lin, S. Mercer, and M. Potthast. Evaluation-as-a-service: Overview and outlook. *CoRR*, abs/1512.07454, 2015.
8. F. Hopfgartner, T. Brodt, J. Seiler, B. Kille, A. Lommatzsch, M. Larson, R. Turrin, and A. Serény. Benchmarking news recommendations: the CLEF NewsREEL use case. *SIGIR Forum*, 49(2):129–136, December 2015.
9. 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 the 5th International Conference of the CLEF Initiative*, pages 250–267. Springer Verlag, September 2014.
10. L. B. Jabeur, L. Soulier, and L. Tamine. Living Ranking: from online to real-time information retrieval evaluation. In *Working Notes of the 7th International Conference of the CLEF Initiative, Evora, Portugal*. CEUR Workshop Proceedings, 2016.
11. L. Kelly. Research-Centres Centred Living Labs. In *Working Notes of the 7th International Conference of the CLEF Initiative, Evora, Portugal*. CEUR Workshop Proceedings, 2016.

12. B. Kille, A. Lommatzsch, G. Gebremeskel, F. Hopfgartner, M. Larson, T. Brodt, J. Seiler, D. Malagoli, A. Sereny, and A. D. Vries. Overview of NewsREEL'16: Multi-dimensional Evaluation of Real-Time Stream-Recommendation Algorithms. In *CLEF 2016: 7th Conference and Labs of the Evaluation Forum*, June 2016.
13. B. Kille, A. Lommatzsch, R. Turrin, A. Sereny, M. Larson, T. Brodt, J. Seiler, and F. Hopfgartner. Stream-Based Recommendations: Online and Offline Evaluation as a Service. In *6th International Conference of the CLEF Initiative*, CLEF'15, pages 487–507, 2015.
14. A. Lommatzsch, N. Johannes, J. Meiners, L. Helmers, and J. Domann. Recommender Ensembles for News Articles based on Most-Popular Strategies. In *Working Notes of the 7th International Conference of the CLEF Initiative, Evora, Portugal*. CEUR Workshop Proceedings, 2016.
15. P. Probst and A. Lommatzsch. Optimizing a Scalable News Recommender System. In *Working Notes of the 7th International Conference of the CLEF Initiative, Evora, Portugal*. CEUR Workshop Proceedings, 2016.
16. P. Schaer and N. Tavakolpoursaleh. Ideas for a Standard LL4IR Extension - Living Labs from a System Operator's Perspective. In *Working Notes of the 7th International Conference of the CLEF Initiative, Evora, Portugal*. CEUR Workshop Proceedings, 2016.
17. J. Yuan and A. Lommatzsch. Clicks Pattern Analysis in Real Stream News Recommender Systems. In *Working Notes of the 7th International Conference of the CLEF Initiative, Evora, Portugal*. CEUR Workshop Proceedings, 2016.