

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a preprint version which may differ from the publisher's version.

For additional information about this publication click this link.

<http://hdl.handle.net/2066/176354>

Please be advised that this information was generated on 2019-01-23 and may be subject to change.

CLEF 2017 NewsREEL Overview: Offline and Online Evaluation of Stream-based News Recommender Systems

Benjamin Kille¹ (✉), Andreas Lommatzsch¹, Frank Hopfgartner², Martha Larson³, and Torben Brodt⁴

¹ TU Berlin, Berlin, Germany

{benjamin.kille, andreas.lommatzsch}@dai-labor.de

² University of Glasgow, Glasgow, UK

frank.hopfgartner@glasgow.ac.uk

³ Radboud University, Nijmegen, and TU Delft, Delft, The Netherlands

m.a.larson@tudelft.nl

⁴ Plista GmbH, Berlin, Germany

torben.brodt@plista.com

Abstract. The CLEF NewsREEL challenge allows researchers to evaluate news recommendation algorithms both online (NewsREEL Live) and offline (NewsREEL Replay). Compared with the previous year NewsREEL challenged participants with a higher volume of messages and new news portals. In the 2017 edition of the CLEF NewsREEL challenge a wide variety of new approaches have been implemented ranging from the use of existing machine learning frameworks, to ensemble methods to the use of deep neural networks. This paper gives an overview over the implemented approaches and discusses the evaluation results. In addition, the main results of Living Lab and the Replay task are explained.

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

1 Introduction

The development of recommender services based on stream data is a challenging task. Systems optimized for handling streams must be able to ensure highly precise recommendations taking into account the continuous changes in the stream as well as changes in the user preferences. In addition to technical complexity of the algorithms must be considered ensuring the seamless integration of recommendations into existing applications as well as ensuring the scalability of the system.

Researchers in Academia often focus on the development of algorithms only tested based on static datasets due to the lack of access to live data. CLEF NewsREEL [5] provides the opportunity to evaluate algorithms both based on live data (NewsREEL Live Task) and offline simulated streams (NewsREEL Replay Task). The benchmarking of the algorithms considers both the recommendation precision (measured by the Click-Through-Rate) and technical aspects (measured by reliability and response time). The Replay Task gives new participants and students an easy access to the NewsREEL

challenge due to the fact the task can be run on standalone hardware without online access and the necessity to fulfill specific time constraints. In addition, the Replay task simplifies the debugging and the simulation of streams. Algorithms shown to be working offline can then be evaluated in the NewsREEL Live task without any changes.

In the 2017's edition of CLEF NewsREEL participants have implemented and evaluated a wide spectrum of algorithms. Most teams participated in both online and the offline evaluation. In this paper we give an overview over the implemented approaches and discuss the evaluation results. 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 and summarized in Sections 4 and 5 and discussed in Section 6.

2 Scenario and Lab Setup

In 2017, NewsREEL has continued the quest to bridge the worlds of data-driven offline evaluation and user-centric live experience. NewsREEL has offered two tasks: NewsREEL Live and NewsREEL Replay. We describe both tasks and conclude the section with a discussion of meta-challenges for participants.

2.1 NewsREEL Live: Benchmarking News Recommendations in a Living Lab

NewsREEL continues to provide participants the unique opportunity to explore how their ideas affect news readers. Participants deploy their recommendation algorithms and connect it to the Open Recommendation Platform (ORP) [3]. Subsequently, their system receives different types of messages initiating on a selection of news publishers. Messages of type *item update* inform about changes to the set of news articles. New articles may be added and existing articles may be updated. Messages of type *event notification* convey happenings on the publishers platform caused by readers' actions. Readers may access news articles or click on recommendations. Error messages notify participants about system malfunctions. These include delayed responses and invalid items. Finally, messages of type *recommendation request* expect a list of news articles in return. The recommendations will be displayed to the reader if the participant is randomly selected among all active systems.

ORP keeps track of readers' reaction to recommendations. For each participant, it counts the clicks as well as requests. Requests can be considered on two levels. On the one hand, we may consider a request for recommendations as a single entity. On the other hand, we may consider a request for each individual item being recommended. Readers clicking on a recommended article will typically trigger the page being reloaded. As a result, the former way to count requests will yield a lower number than the latter. We challenged participants to find the configuration which minimizes the click through rate (CTR). Herein, the CTR refers to the number of clicks divided by the number of requests counting lists instead of individual items.

Our partners at Plista have revised ORP for NewsREEL 2017. Changes concerned both front-end and back-end. The former user interface has been deprecated in favor of a replacement which is under ongoing development. Figure 1 and Figure 2 show

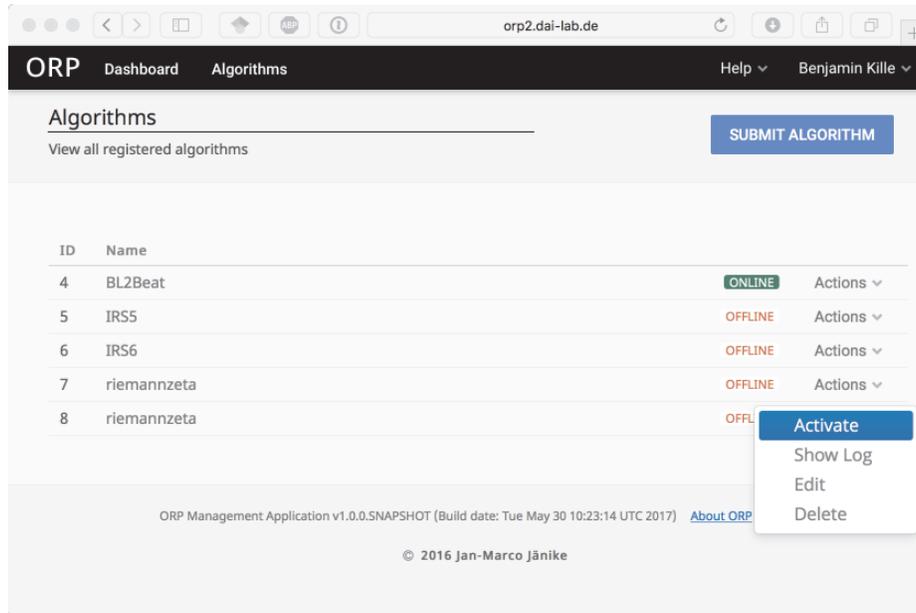


Fig. 1. The screenshot shows the new webpage for controlling the recommender algorithms. The screen allows participants to register, activate, and edit recommendation algorithms.

the currently available state of the new user interface. The new interface comes with a flexible way to create dashboards. Participants can arrange their favorite information as they please. Note, that the user interface has not yet been available in NewsREEL 2017. Participants received a tutorial illustrating how to use ORP by means of API calls until the new user interface would be available. Plista migrated ORP to a new server and extended its API. Calling the API, participants can control the communication with ORP programmatically. The data format has been kept to reduce the efforts to update existing implementations for NewsREEL’s participants.

2.2 NewsREEL Replay: Benchmarking Stream-based News Recommendations Offline

The NewsREEL Replay task allows participants to evaluate stream-based news recommender algorithms offline. As described by Scriminaci [18], offline evaluation ensures the exact reproducibility of experiments as well as the fine-grained analysis and optimization of algorithms. Participants can simulate different load scenarios as well as check the reliability of new approaches. Teams can optimize parameters before deploying algorithms to NewsREEL Live.

For the NewsREEL Replay task a data set and software components for simulating the data stream are provided. Participants have access to a data set comprising a collection of messages analogous to NewsREEL Live. The messages are chronologically ordered and cover the period of four weeks starting on February 1, 2016. For further details

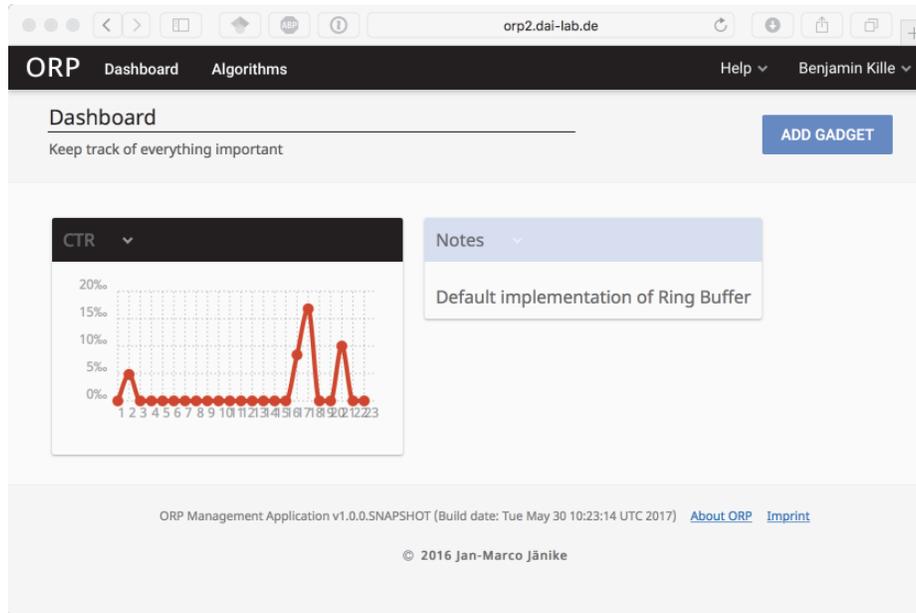


Fig. 2. The screenshot shows the dashboard of the web portal for controlling the recommender algorithms. The dashboard shows different gadgets visualizing participants' notes as well as the recommender precision of active algorithms.

about the nature of the data set, we refer to [8]. Plista's customer base changes over time which is why some publishers are currently accessible through ORP but are not included in the data set. Besides the data set, participants receive software to conduct offline evaluations. The software simulates two systems. On the one hand, the software emulates ORP's functionality. This part sends recommendation requests, compares recommendations to logs with later timestamps, and records the time taken until the recommendations arrive. On the other hand, the software emulates the recommendation engine. Participants include their own implementation in this part and modify it to determine how changes affect the performance. The software produces estimated click through rates and response time distributions. The estimated CTR relates to impressions rather than clicks. For further details about this evaluation resource, we refer to [10].

Analogous to the online evaluation, participants ought to find the configuration with the highest CTR. Simultaneously, the offline evaluation reveals how changing recommendation algorithms affect the response time. Participants generate insights on which configurations accomplish a reasonable trade-off amid prediction accuracy and response time.

In 2017, we have released a new data set. The data set covers a four week period and adheres to the format used in previous editions of NewsREEL. The software used to conduct offline experiments facilitates re-using existing implementations. Hence, participants experience minimal requirements to start their experiments.

2.3 Discussion

NewsREEL 2017 constitutes a major revision with changes to main resources. Plista revised and migrated ORP to achieve better stability, maintainability, and flexibility. We released a new data set with more recent interactions. We moved from Idomaar to a new evaluator. It took time to update support materials such as tutorial, descriptions, and references. Having previous materials available, some participants reported confusion. The new user interface has not been finished in time to engage participants. The scale of the data set has been challenging for participants. NewsREEL has been used within the scope of university lectures. This confirms the interest of academic institutions to provide students with more realistic problems.

3 Participation

In the 2017 edition of NewsREEL, 87 participants have registered. Both tasks attracted similarly many participants with NewsREEL Replay slightly ahead with 79 registrations compared to NewsREEL Live with 64 participants. Participants deployed 27 recommendation services in Task 1. We received a total of six working notes describing participants' approaches. For a more detailed analysis of peoples' motivation to participate, we refer to [16].

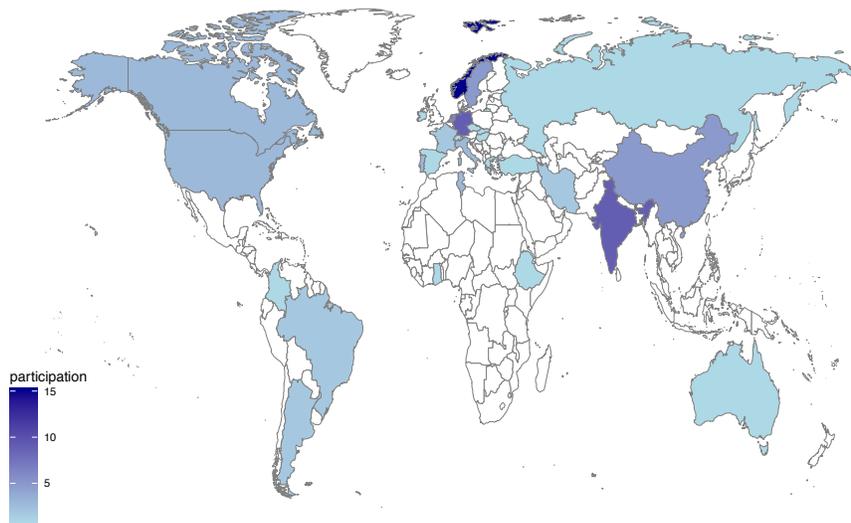


Fig. 3. Worldwide participation in NewsREEL 2017. Color shade represents the number of participants per country.

Figure 3 illustrates how registrations are distributed across the globe. All continents (except Antarctica) have at least one participant who registered for NewsREEL 2017. Norway (15), Germany (9), and the Netherlands (6) had most registrants.

4 Results

This section presents the results for both tasks of NewsREEL 2017. First, Section 4.1 introduces participants' achievements in NewsREEL Live. Second, Section 4.2 illustrates participants' results in NewsREEL Replay. For results of the previous campaigns, we refer to [7, 12, 11].

4.1 NewsREEL Live Results

ORP has undergone revisions until March, 2017. Plista created accounts for all registered participants on March 23, 2017. From this time on, they could establish communication with ORP to initiate evaluations. This allowed participants to explore parameter space to find the optimal configuration of their algorithms. Setting up their systems has been a challenging endeavor. They had to implement, deploy, and maintain their systems. Nineteen systems have been active in the evaluation period starting on April 23 and ending with May 7, 2017. An error in ORP's internal logging occurred on April 28, 2017. Unfortunately, no information is available for this day. Table 1 lists our observations for all nineteen systems. Systems are presented in alphabetical order. The system "BL2Beat" refers to the organizers' baseline implementation. Participants registered from as few as two up to as many as 1268 clicks in the fourteen day period. The number of impressions refers to how often the recommendations of a system have been shown to readers. We observe a considerable variance from 349 to 81 245 impressions. The variance emerges as some participants had their systems connect for longer periods than some competitors. Some participants were connected to ORP for 289 h, whereas other systems remained disconnected for most of the time. Table 1 includes the average number of clicks and impressions per hour. These values reveal whether participants experienced similar conditions. On average, participants received 203.6 (mean) or 224.0 (median) impressions per hour. Incidentally, the median value refers precisely to the baseline implementation. On average, participants registered 2.6 (mean) or 2.8 (median) clicks per hour.

Figure 4 illustrates the performance in more detail. Each triangle corresponds to an algorithm which served recommendations to ORP. The x -value refers to the total number of impressions. The y -value refers to the total number of clicks. Consequently, the triangles' positions indicate the average CTR per day. Two colored areas highlight ranges of the CTR. The blue area refers to CTR below 1 %, whereas the brown area refers to CTR above 2 %. A majority of participants finds itself in between both areas. A few participants have been active for a relatively short period. They received few impressions and clicks and thus clutter close to the origin.

Table 1. Results of NewsREEL Live for the time from April 23 to May 7, 2017 except April 28. On April 28, the logging system failed such that no results have been recorded. The table lists the total number of clicks, the total number of impressions, how many hours the system has been activated, the average number of clicks per hour ($\langle Ch^{-1} \rangle$), the average number of impressions per hour ($\langle Ih^{-1} \rangle$), and the click through rate.

Recommender	Clicks (C)	Impressions (I)	Hours	$\langle Ch^{-1} \rangle$	$\langle Ih^{-1} \rangle$	CTR
BL2Beat	726	62 052	277	2.62	224.01	0.0117
eins	817	61 524	289	2.82	212.88	0.0132
Has logs	6	816	32	0.18	25.50	0.0073
IL	813	79 120	289	2.81	273.77	0.0102
IRS5	58	3708	15	3.86	247.20	0.0156
IT5	925	68 582	289	3.20	237.30	0.0134
ody0	166	23 023	138	1.20	166.83	0.0072
ody1	810	68 768	289	2.80	237.95	0.0117
ody2	875	63 950	289	3.02	221.28	0.0136
ody3	813	59 227	288	2.82	205.64	0.0137
ody4	1139	72 601	289	3.94	251.21	0.0156
ody5	1268	81 245	289	4.38	281.12	0.0156
ORLY_KS	896	42 786	276	3.24	155.02	0.0209
RIADI_hyb	2	349	5	0.40	69.80	0.0057
RIADI_nehyb	764	75 535	249	3.06	303.35	0.0101
Riadi_NV_01	12	443	53	0.22	8.35	0.0270
RIADI_pn	879	77 723	274	3.20	283.66	0.0113
WIRG	600	49 830	236	2.54	211.14	0.0120
yl-2	747	60 814	241	3.09	252.34	0.0122

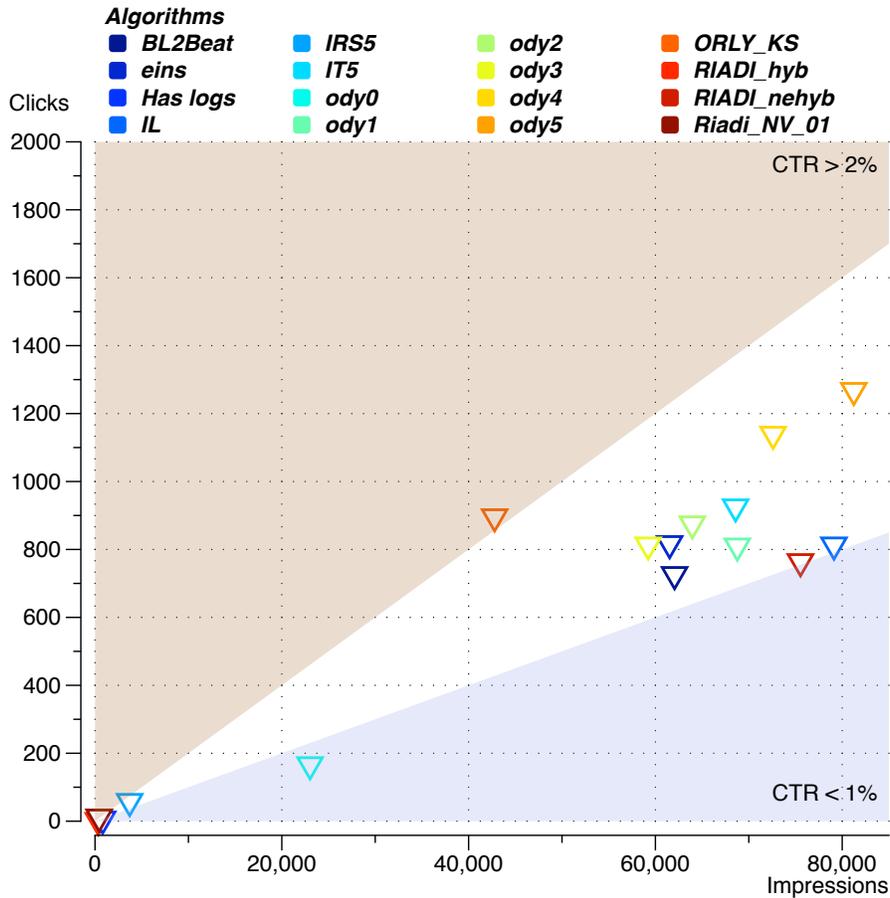


Fig. 4. Participants responded to recommendation requests and ORP tracked whenever users clicked on the suggestions. For each algorithm we computed the mean number of requests and clicks per day. The figure illustrates the CTR for each algorithm. The size of the symbol reflects how many days the algorithm was available.

4.2 NewsREEL Replay Results

The Offline Evaluation task has attracted several teams. The teams engaged in the NewsREEL Replay mainly focused on testing new recommender approaches (e.g. deep neuronal networks [14]), the efficiently optimization of parameter configuration (e.g. finding similarity metrics for Collaborative Filtering [1]), and on studying the technical complexity of algorithms. NewsREEL Replay does not require a permanent internet connection. This ensures a low barrier to participate in the NewsREEL challenge and motivates participant to test new ideas and algorithms.

Testing new Approaches Applying innovative ideas in a recommendation scenario typically requires extended testing and debugging. Before setting up a stable running live system, algorithms are prototypically implemented in order to proof that the new approach is suitable for the scenario. The NewsREEL Replay task provides such a testing environment. Participants can simulate the stream on local hardware and study the strength and weaknesses of new algorithms. The offline tests can control the load (by defining the number of concurrent messages sent by the offline simulation environment) and debug the functionality of the implemented solution. In the NewsREEL challenge 2017 most new teams tested the algorithms first offline before participating in the NewsREEL Live task. New recommender approaches based on Contextual Bandits and Deep Neural Networks have been evaluated offline.

Parameter Optimization In addition to the testing of new approaches, the optimization of suitable parameter configuration is an important task. The parameter configuration requires a sufficiently large data stream in order to ensure significant optimization results. For speeding up the parameter optimization a parallelization of the optimization should be supported. The NewsREEL Replay task addresses this need. The provided dataset and the offline stream simulation components allow participants to simulate the data stream in parallel on different machines and with different hardware configuration. In addition, the simulated stream can be replayed faster in order to accelerate the optimization process. The offline stream simulation ensures reproducible evaluation results as well as the comparability of the results obtained in different evaluation runs. This aspect of the NewsREEL Replay task has been extensively used by several teams (e.g. by Beck et al. [1]).

Technical aspects The tight time constraints in the NewsREEL Live task and the continuous changes in the number of messages make it difficult to analyze the technical aspects of implemented algorithms. In the NewsREEL Live task several peaks in the number of messages can be observed. Algorithms running in NewsREEL Live must be able to handle such load peaks. The offline stream simulation component allows participants to analyze load peaks by defining the numbers of messages concurrently sent to the recommender. This helps participants to identify bottlenecks and to study the handling of concurrent messages. The analysis of the response time has been conducted by several teams by plotting histograms describing the frequency of different response times. This is of special interest in ensemble-based approaches integrating different algorithms with varying technical complexities.

Discussion The NewsREEL Replay task enables the fine-grained analysis of new algorithms and allows participants the efficient optimization of parameters. As NewsREEL Replay can be run offline without considering response time constraints, it is a good starting point for new participants to evaluate new ideas and algorithms. NewsREEL Replay has been used by most participants for optimizing the algorithms with respect to both recommendation precision and technical complexity.

5 Working Notes Summary

In NewsREEL 2017 the participants have evaluated a broad spectrum of recommender approaches ranging from using existing frameworks and tools to ensemble methods to the use of deep neural networks.

Bons et al. [2] developed a graph-based recommender algorithm. The graph consists of nodes representing the items and directed edges describing the frequency that the two connected news items are read in the specific sequence. Recommendation requests are answered by computing the strongest item sequence containing the itemID given in the recommendation request. The graph is managed in a Neo4j graph database. Recommendations are computed based on a database query. If the itemID in the recommendation request does not exist in the graph or the node is not yet connected with the graph, the most recently created news items are return. The evaluation of the strategy shows that the implemented graph-based recommender reaches a high CTR in the Living Lab scenario. The implementation works efficiently ensuring that the time-constraints with respect to response time are reliably fulfilled.

Golian and Kuchar [4] 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.

Ludmann [17] focuses on managing streams. His system relies on *Odysseus*, a data stream management systems. Therein, he defines a set of queries which take part of the data stream and determine the most popular articles. The selection entails the length of the data stream segment as essential parameter. The working notes presents observations in NewsREEL Live with a variety of parameter configurations. Results suggest that considering successful recommendations improves the click through rates.

Beck et al. [1] developed a hybrid recommender system combining item-based Collaborative Filtering algorithms with a most popular recommender. The system is implemented using the Apache Mahout framework. The message stream is processed in split into batches of equal size. Having collected the required number of messages for a batch, the system builds a recommender model for this batch. When the model building is completed, the new model replaces the old model. In order to ensure that for every request recommendations are provided, a most popular item recommender runs a backup recommender. If the Collaborative Filtering-based recommender fails or does not provide a sufficient number of results, the recommendation result is completed by the backup recommender. The evaluation of the recommender shows that the implemented solution provides highly precise results and fulfills the technical requirements with respect to response time and scalability.

Liang et al. [15] discuss how contextual bandits can be used to compute recommendations. The authors define a list of recommendation models considering recency, categories, and reading sequences among other factors. Their contextual bandit approach seeks to determine a strategy mapping models to contexts in order to maximize the expected rewards. They apply their contextual bandit both in NewsREEL Live and

NewsREEL Replay. The working note reports that performances vary by the domain under consideration.

Kumar et al. [14] present d a hybrid recommender system for news. They combine collaborative filtering with content-based filtering using a neural net architecture. Part of the architecture models the relation amid users and items. The other part of the architecture maps articles' text onto a common latent space. The authors conduct an offline experiment which compares their proposed method to three baselines. The experiment focuses on readers who had previously read ten to fifteen articles. Their results favor their approach over the baselines in terms of hitrate and normalized discounted cumulative gain.

The variety of methods used to address NewsREEL's tasks indicate a large number of connected research questions for the future. Most approaches achieved results superior to the baseline and still yield the potential for further optimization.

6 Discussion

Similar to the past few iterations of CLEF NewsREEL [6, 13, 9], we were pleased to see that participants trialled very diverse approaches to provide news recommendations. We argue that this is due to the opportunity to evaluate recommendation algorithms in an industry setting.

Of both tasks, the evaluation in an online setting, referred to as NewsREEL Live throughout the campaign, appears to be more attractive amongst participants. This is also similar to previous years where we saw more teams evaluating their algorithms using the Open Recommendation Platform run by plista. At the same time, this year, an increasing number of participants also tested their algorithms in the offline setting, referred to as NewsREEL Replay. One of the advantages of offline evaluation is that it allows to benchmark algorithms that might not be suitable (yet) to be operated in an online setting.

Although the performance of the algorithms presented are promising, we argue that there is still space for improvement with the aim of increasing the overall click-through rate. We therefore would like to encourage researchers to perform more studies using the data and infrastructure that has been provided as part of CLEF NewsREEL.

References

1. P. D. Beck, M. Blaser, A. Michalke, and A. Lommatzsch. A System for Online News Recommendations in Real-Time with Apache Mahout. In *Working Notes of the 8th International Conference of the CLEF Initiative, Dublin, Ireland*. CEUR Workshop Proceedings, 2017.
2. P. Bons, N. Evans, P. Kampstra, and T. van Kessel. A News Recommender Engine with a Killer Sequence. In *Working Notes of the 8th International Conference of the CLEF Initiative, Dublin, Ireland*. CEUR Workshop Proceedings, 2017.
3. T. Brodt and F. Hopfgartner. Shedding light on a living lab: the CLEF NewsREEL open recommendation platform. In *Fifth Information Interaction in Context Symposium, IiX '14, Regensburg, Germany, August 26-29, 2014*, pages 223–226, 2014.

4. C. Golian and J. Kuchar. News Recommender System based on Association Rules at CLEF NewsREEL 2017. In *Working Notes of the 8th International Conference of the CLEF Initiative, Dublin, Ireland*. CEUR Workshop Proceedings, 2017.
5. 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, 2015.
6. F. Hopfgartner, B. Kille, A. Lommatzsch, T. Plumbaum, T. Brodt, and T. Heintz. 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*, pages 250–267, 2014.
7. B. Kille, T. Brodt, T. Heintz, F. Hopfgartner, A. Lommatzsch, and J. Seiler. NewsREEL 2014: Summary of the news recommendation evaluation lab. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, pages 790–801, 2014.
8. 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.
9. B. Kille, A. Lommatzsch, G. G. Gebremeskel, F. Hopfgartner, M. Larson, J. Seiler, D. Malagoli, A. Serény, T. Brodt, and A. P. de Vries. Overview of NewsREEL'16: Multi-dimensional evaluation of real-time stream-recommendation algorithms. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, pages 311–331, 2016.
10. B. Kille, A. Lommatzsch, F. Hopfgartner, M. Larson, and A. P. de Vries. A stream-based resource for multi-dimensional evaluation of recommender algorithms. In *The 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*. ACM, 2017.
11. B. Kille, A. Lommatzsch, F. Hopfgartner, M. Larson, J. Seiler, D. Malagoli, A. Serény, and T. Brodt. CLEF NewsREEL 2016: Comparing multi-dimensional offline and online evaluation of news recommender systems. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 593–605, 2016.
12. B. Kille, A. Lommatzsch, R. Turrin, A. Serény, M. Larson, T. Brodt, J. Seiler, and F. Hopfgartner. Overview of CLEF NewsREEL 2015: News recommendation evaluation lab. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*, 2015.
13. B. Kille, A. Lommatzsch, R. Turrin, A. Serény, M. Larson, T. Brodt, J. Seiler, and F. Hopfgartner. Stream-based recommendations: Online and offline evaluation as a service. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings*, pages 497–517, 2015.
14. V. Kumar, D. Khattar, S. Gupta, M. Gupta, and V. Varma. Deep Neural Architecture for News Recommendation. In *Working Notes of the 8th International Conference of the CLEF Initiative, Dublin, Ireland*. CEUR Workshop Proceedings, 2017.
15. Y. Liang, B. Loni, and M. Larson. CLEF NewsREEL 2017: Contextual Bandit News Recommendation. In *Working Notes of the 8th International Conference of the CLEF Initiative, Dublin, Ireland*. CEUR Workshop Proceedings, 2017.
16. A. Lommatzsch, B. Kille, F. Hopfgartner, M. Larson, T. Brodt, J. Seiler, and Ö. Özgöbek. CLEF 2017 NewsREEL overview: A stream-based recommender task for evaluation and education. In *8th International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2017)*. Springer, 2017.

17. C. Ludmann. Recommending News Articles in the CLEF News Recommendation Evaluation Lab with the Data Stream Management System Odysseus. In *Working Notes of the 8th International Conference of the CLEF Initiative, Dublin, Ireland*. CEUR Workshop Proceedings, 2017.
18. M. Scriminaci, A. Lommatzsch, B. Kille, F. Hopfgartner, M. Larson, D. Malagoli, A. Serény, and T. Plumbaum. Idomaar: A framework for multi-dimensional benchmarking of recommender algorithms. In *Proceedings of the Poster Track of the 10th ACM Conference on Recommender Systems (RecSys 2016), Boston, USA, September 17, 2016.*, 2016.