A Stream-based Resource for Multi-Dimensional Evaluation of Recommender Algorithms

Benjamin Kille
Andreas Lommatzsch
benjamin.kille@tu-berlin.de
andreas.lommatzsch@dai-labor.de
Technische Universität Berlin –
DAI-Lab, Berlin, Germany

Frank Hopfgartner
frank.hopfgartner@glasgow.ac.uk
University of Glasgow, Glasgow, UK

Martha Larson
Arjen P. de Vries
m.larson@science.ru.nl
arjen@acm.org
Radboud University, Nijmegen,
The Netherlands

ABSTRACT

Recommender System research has evolved to focus on developing algorithms capable of high performance in online systems. This development calls for a new evaluation infrastructure that supports multi-dimensional evaluation of recommender systems. Today's researchers should analyze algorithms with respect to a variety of aspects including predictive performance and scalability. Researchers need to subject algorithms to realistic conditions in online A/B tests. We introduce two resources supporting such evaluation methodologies: the new data set of stream recommendation interactions released for CLEF NewsREEL 2017, and the new Open Recommendation Platform (ORP). The data set allows researchers to study a stream recommendation problem closely by "replaying" it locally, and ORP makes it possible to take this evaluation "live" in a living lab scenario. Specifically, ORP allows researchers to deploy their algorithms in a live stream to carry out A/B tests. To our knowledge, NewsREEL is the first online news recommender system resource to be put at the disposal of the research community. In order to encourage others to develop comparable resources for a wide range of domains, we present a list of practical lessons learned in the development of the dataset and ORP.

CCS CONCEPTS

•**Information systems → Recommender systems; Test collections;** *Relevance assessment;*

KEYWORDS

streams; recommender system; multi-dimensional benchmarking

ACM Reference format:
Benjamin Kille, Andreas Lommatzsch, Frank Hopfgartner, Martha Larson, and Arjen P. de Vries. 2017. A Stream-based Resource for Multi-Dimensional Evaluation of Recommender Algorithms. In *Proceedings of SIGIR '17, Shinjuku, Tokyo, Japan, August 7–11, 2017,* 4 pages.
DOI: 10.1145/3077136.3080726

SIGIR '17, Shinjuku, Tokyo, Japan

978-1-4503-5022-8/17/08…$15.00.
DOI: 10.1145/3077136.3080726

1 INTRODUCTION

Many news publishers put a small widget box at the bottom of their articles. The box presents a list of recommended news articles. Publishers have to consider a variety of quality criteria to provide adequate services to their readers. This mandates a *multi-dimensional evaluation*. This paper presents resources that allow researchers to evaluate stream-based algorithms with respect to multiple environments. The resources are released by REcommendation Evaluation Lab (NewsREEL) [5], a campaign-style evaluation lab of the CLEF conference. The resources correspond to the two subtasks of this evaluation lab, NewsREEL Live and NewsREEL Replay. NewsREEL Live implements online evaluation, *i.e.*, researchers gain access to the infrastructure of a company allowing them to evaluate different recommendation techniques using A/B testing. NewsREEL Live provide researchers an infrastructure best described as a living lab. They supply recommendations which are subsequently displayed to actual users. The users' feedback constitutes the basis for an authentic evaluation. The second subtask of NewsREEL focuses on simulating a static data stream as provided by the living laboratory platform. For this, we introduce a new data set comprising interactions between users and various news portals in a one-month time span.

This paper is structured as follows. Section 2 provides a brief overview of research challenges that can be addressed using the NewsREEL resources. Section 3 explains the resources in detail. We give an overview of the recommendation platform that serves as the key component of the living laboratory scenario addressed by NewsREEL. In addition, we outline the new data set that we released to enable further research in an offline setting. Section 4 concludes the paper and provides recommendations for how to create an online evaluation resource useful for the research community.

2 RESEARCH CHALLENGES

Resources provided by CLEF NewsREEL 2017 allow participants to employ two evaluation settings. First, participants can engage with actual users reading news and monitor their feedback by means of a living lab platform. Second, participants can use recordings of such requests and feedback by replaying these. Figure 1 provides an overview of the different options.

From a technical point of view, the news recommendation scenario addressed by NewsREEL is quite challenging. Recommendations must be generated in real-time whenever a visitor accesses a news article on one of the news portals. Instead of computing

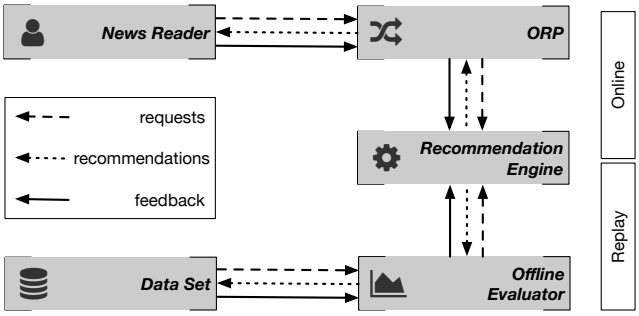

Figure 1: CLEF NewsREEL 2017 overview.

recommendations based on a static set of users and items, the challenge here is to provide recommendations for a news article stream characterized by a continuously changing set of users and items. News content publishers constantly update their existing news articles, or add new content. The short life cycle of items and the strict time-constraints for recommending news articles limit the practical feasibility of many current approaches to recommendation and should be considered to ensure the impact of research findings. In a stream-based scenario, the recommender algorithms must be able to cope with a large number of newly created articles and should be able to discard old articles, since recommended news articles should be "new". For these reasons, recommender algorithms must be continuously adapted to meet the special requirements of the news recommendation scenario.

Analyzing large numbers of papers that address research challenges as mentioned above, Sakai [8] highlights the importance of appropriate sample sizes. Industry addresses the need for large samples by performing A/B testing, which benchmarks variants of a recommender system using large groups of users. A/B testing is currently being increasingly adopted for the evaluation of commercial systems with a large user base as it provides the advantage of observing the efficiency and effectiveness of recommendation algorithms under real conditions. Joachims and Swaminathan summarize challenges arising in this evaluation methodology in their SIGIR tutorial on "Counterfactual Evaluation and Learning for Search, Recommendation and Ad Placement" [6].

Unfortunately, university-based researchers struggle unless they closely collaborate with industry (*e.g.*, [7]) or develop their own infrastructure and user base (*e.g.*, [1]). Without online testing opportunities open to the research communities, they cannot employ online evaluation on a larger scale, which is the de-facto standard evaluation methodology in industry. NewsREEL addresses this issue by providing access to a living lab environment allowing university-based researchers to perform A/B testing. This paper presents the new and improved infrastructure, deployed early 2017, which builds on the first version introduced in [3].

Evaluating recommender algorithms online impedes replicability as users and items continuously change. This limitation hampers detailed evaluation and, consequently, the optimization of algorithms, since different algorithms or different parameter settings cannot be tested in a procedure that can be repeated exactly. For this reason, there is a need for the use of offline data sets that allow fine-tuning of algorithms and enable us to replicate results. Our second contribution addresses this issue: NewsREEL has released a new data set that can be used for simulating an online data stream of user requesting recommendations. Both online and offline resources are introduced in the remainder of this paper.

3 RESOURCES

In this section, we discuss in detail the resources released by NewsREEL, and the way in which they support both online as well as offline evaluation. For online evaluation, we allow researchers to connect their algorithms to live streams of events to conduct A/B testing experiments. For offline evaluation, NewsREEL provides a data set and software components for replaying the data set as a stream. The replay component ensures that recommender algorithms tailored for the online evaluation scenario can also be evaluated offline. In the following, we cover the data, the framework, data statistics, and evaluation metrics.

3.1 Architecture

The evaluation architecture consists of three components for both the online and offline setting. First, a data source provides streams of events. In the online case, live interactions emerge from publishers' websites connected to ORP. In the offline case, recorded streams are provided in form of a static data set. Second, a recommendation engine provides lists of recommendations upon request. Using the identical recommendation engine online and offline provides useful insights in the operational characteristics of the algorithm. Finally, a component orchestrates messages and keeps track of feedback on items recommended by the recommendation engine. In the online case, the Open Recommendation Platform (ORP) fulfills this role. In the offline case, we provide an Offline Evaluator to evaluate recommendations for streams generated from the data set. Figure 1 illustrates the architecture, and highlights the symmetry between the two settings.

3.2 Data

The conditions for online and offline evaluation are comparable. Researchers have access to similar information in both settings. Messages adhere to the same data format facilitating the implementation of modular recommendation engines. Streams convey content-based data (the news items' titles and abstracts), interaction data (user-item interactions — "impressions" and "clicks" used as implicit feedback) as well as contextual information (*e. g.* geolocation, user-agent data describing browsers and devices types).

Content-data For each item the data set provides its title ($\approx$ 60 characters per item) and the abstract ($\approx$ 300 characters per item). Title and abstract data can be used for content-based recommendation and clustering approaches.

Interaction Data The majority of messages describes interactions between users and items. The data enable an analysis of user preferences and characteristic user behavior. They capture how topics evolve and how much attention they manage to attract.

Context Messages convey a variety of contextual information describing the circumstances in which users read articles. These include information related to the browsing setup, time, and location. For instance, references to the device type let participants create

recommendations tailored specifically to mobile or non-mobile conditions.

3.3 Online Evaluation Framework

The Open Recommendation Platform (ORP) was initially introduced in [3]. ORP has been updated and re-released for NewsREEL 2017. Besides a novel graphical user interface, participants can now use a RESTful API. Functions can be executed programmatically without the need to manually log in. Thereby, participants can design their systems to flexibly react to changing conditions. For instance, the system can reactivate communication.

3.4 Online Data Statistics

External factors affect the message stream emitted by ORP. Both visitors' interests and publishers' configurations change dynamically. Hence, we cannot reliably predict the number of messages participants will receive. We distinguish three types of messages: event notifications, item updates, and recommendation requests. Notifications inform participants about visitor actions such as reading articles or clicking suggestions. Item updates inform participants about changes in the article collection. Changes include adjustments to existing articles and addition of new ones. Request messages expect a list of suggested articles in return. We observed the number of events for each type over the period 8–27 February 2017 in order to obtain a recent sample, reporting a median of ≈ 2.1 million notifications, ≈ 10.8 thousand item updates, and ≈ 3.7 thousand requests per day. Having a recommendation engine available over the course of a month, participants ought to be able to collect about 110 000 requests.

3.5 Online Metrics

ORP keeps track of the recommendations delivered and whether users click on them. Participants can query the number of requests and clicks for arbitrary periods. The fraction of clicks to requests represents the click-through rate. In addition, ORP records the number of errors produced by participants. Errors occur when participants' recommendation engines take too long to respond or provide invalid items. The fraction of errors to requests represents the error rate. Ideally, participants achieve a high click-through rate and a low error rate.

3.6 Offline Data Statistics

The offline data set has been created by recording all messages in the online evaluation in February 2016 (1–28 February 2016). A timestamp has been added to each message describing when the message had occurred. Within the time window of the data set, five different news portals were connected to ORP. The data set comprises 170 million messages describing user-item interactions and 60 thousand describing creation or update of items.

3.7 Offline Metrics

The offline data set and the framework for creating the streams enable the evaluation with respect to (i) recommendation precision, and (ii) technical complexity.

(i) The recommendation precision is computed in terms of offline CTR. A recommendation for a request from user u is counted as successful, if u accesses the recommended news item in the near future (within the next 5 min). This metric can be calculated directly based on the data set and does not require explicit user feedback in form of ratings. In NewsREEL Replay, there will be an inherent bias toward the recommender system deployed during the period when the data set had been recorded. We consider a recommendation successful, if we observe the user reading any recommended article included in the recommendations. We ignore whether the reading followed the user clicking the recommendation or through navigating the news portal. (ii) The technical complexity is analyzed by measuring the recommender's distribution of response time. By simulating different load levels (replaying the stream faster than the original stream) the influence of the load and the hardware resources used by the recommender on the response time can be analyzed.

3.8 Offline Evaluation Framework

One of the most important characteristics of the NewsREEL scenario is that the data are represented as a stream. In order to preserve this characteristic, the NewsREEL resources include a component for replaying the recorded stream. This enables an offline evaluation similar to the online evaluation scenario. The replaying component lets participants control the load by regulating the rate at which requests are sent. It emulates the server used in the online scenario. The component imports the offline data sets and sends it message-wise to the recommender. The provided recommendations are evaluated with respect to offline CTR. In addition, a response time statistic is created.

3.9 Discussion

Both online and offline resources are inherently based on stream data. Both parts of the resource provide the same information and same data characteristics (short item life cycle, highly context-dependent user preferences, variance in the user behavior). The evaluation metrics differ between the online and the offline scenario: The online scenario uses real-world user feedback to compute the CTR. In the offline scenario, the recommendation accuracy is calculated based on the overlap between the future impressions and the recommendations. The offline evaluation enables the exact repeatability and a detailed analysis of the computational complexity. The combination of online and offline stream resources allows an comprehensive analysis and optimization of recommender algorithms combining industry-style A/B testing and the academic evaluation based on a recorded data set. As a result, NewsREEL provides the resources to evaluate recommendation systems in a way that is more representative than previous data sets in the recommendation domain (such as Netflix [2] and MovieLens [4]).

4 CONCLUSION AND DISCUSSION

In summary, the resources described in this paper provide researchers outside of industry with a unique opportunity to evaluate recommender algorithms. The accompanying data set allows us to repeat experiments exactly and tune parameters *a priori* for the online evaluation. The data set and the online platform share similar characteristics with respect to fluctuations of the article collection and user set. Researchers can explore content, similar usage

patterns, and context as information sources to optimize recommendations. Besides accuracy, researchers can calculate other quality measures such as response rates, error rates, and throughput.

Beyond recommender systems, the stream-based nature of the data is representative of many web-based application scenarios. The text and image content of the news articles provide an interesting opportunity for testing classic information retrieval algorithms. Further, the data set supports the analysis of behavioral patterns and trends in a news ecosystem. The data set spans four weeks with millions of logged events; we expect that there is much yet hidden in the event stream remaining to be discovered.

4.1 Challenges of Creating Stream Resources

In this section, we discuss some challenges of creating stream-based resources useful to others interested in developing similar living labs. Not all online systems are suitable as living labs, and the choice should be made carefully. The stream information should be rich, but should not raise privacy concerns. It should not be possible for an algorithm to generate recommendations that would disturb or upset users. Response time seriously affects user experience, and recommender algorithms that cannot maintain required response times must be automatically disabled.

Creating stream-based resources requires the implementation of a living lab platform, or the adaptation of an existing platform to communicate with a given data source. Once deployed, the platform must be maintained, a consideration easily underestimated by those accustomed to only producing static data sets. ORP receives a fraction of a stream diverted from the main system. This data stream must be carefully monitored. The stream can suddenly undergo radical changes. These changes affect the amount of time necessary to gain meaningful results. Controlling the volume of messages can extend or reduce this time. Close communication with the teams in charge of the front-end and the back-end of the main system is necessary in order to keep the living lab running smoothly.

4.2 Challenges of Supporting Researchers

In this section, we discuss how to support researchers in order to ensure successful uptake of the stream-based evaluation resources. First, the living lab platform should provide researchers with as much information as possible on the status and performance of their algorithms. The newly released version of ORP pays particular attention to this point. It allows researchers to monitor their systems' status, but also to activate multiple request streams in order to experimentally confirm the random nature of the assignment of requests to streams.

Second, researchers need support in acquiring the broad skills necessary to succeed in stream-based evaluation. The high volume of data combined with challenging response time bounds require to pay close attention to efficiency and reliability. Still, machine learning courses generally fail to teach the specific skills needed to implement and maintain a recommender system. The dynamic nature and high variability of the data stream make it impossible to draw conclusions based on short-term observations. Researchers need well develop statistical and analytical skills to come to meaningful results.

We have learned that providing detailed tutorials and other informational materials in advance helps users getting started. In addition, we have noticed that providing a baseline implementation for the online evaluation as well as an evaluator for the offline evaluation lowers the threshold for researchers to start evaluating. Finally, we recognized that interactive communication is the key to avoid ambiguity. Responding promptly and comprehensively to inquiries helps participants to avoid frustration and to move forward in evaluation.

4.3 Outlook

Until now, academic researchers have focused primarily on evaluation involving data sets and user studies. However, a method that makes correct predictions but takes minutes to compute is of little value to businesses. Moving forward, research should consider non-functional requirements such as maintenance, reliability, and response time.

The academic approach is designed to support repeatability. The industrial approach seeks to optimize business success. The search for the best method to provide relevant information to users relates to both objectives. However promising the (replicated) results of a new recommendation algorithm, it will not be applied in practice unless customers react positively to its output. Conversely, academic research is problematic if meaningful results can only be reported based upon industrial evaluation that cannot be replicated. NewsREEL is designed to unite both worlds by providing resources that allow both online and offline evaluation.

REFERENCES

[1] J. Beel, C. Breitinger, S. Langer, A. Lommatzsch, and B. Gipp. Towards reproducibility in recommender-systems research. *User Model. User-Adapt. Interact.*, 26(1):69–101, 2016.
[2] J. Bennett and S. Lanning. The Netflix prize. In *In KDD Cup and Workshop in conjunction with KDD*, 2007.
[3] T. Brodt and F. Hopfgartner. Shedding light on a living lab: the CLEF NewsREEL open recommendation platform. In *IIiX '14*, pages 223–226, 2014.
[4] F. M. Harper and J. A. Konstan. The MovieLens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19, Dec. 2015.
[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] T. Joachims and A. Swaminathan. Counterfactual evaluation and learning for search, recommendation and ad placement. In *SIGIR'16*, pages 1199–1201, 2016.
[7] A. Maksai, F. Garcin, and B. Faltings. Predicting online performance of news recommender systems through richer evaluation metrics. In *RecSys'15*, pages 179–186, 2015.
[8] T. Sakai. Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006-2015. In *SIGIR '16*, pages 5–14. ACM, 2016.