Benchmarking News Recommendations: The CLEF NewsREEL Use Case

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

Torben Brodt, Jonas Seiler
plista GmbH, Germany
{*firstname.lastname*}*@plista.com*

Benjamin Kille, Andreas Lommatzsch
TU Berlin, Germany
{*firstname.lastname*}*@dai-labor.de*

Martha Larson
TU Delft, The Netherlands
m.a.larson@tudelft.nl

Roberto Turrin
ContentWise R&D, Italy
roberto.turrin@moviri.com

András Serény
Gravity R&D, Hungary
sereny.andras@gravityrd.com

October 7, 2015

Abstract

The CLEF NewsREEL challenge is a campaign-style evaluation lab allowing participants to evaluate and optimize news recommender algorithms. The goal is to create an algorithm that is able to generate news items that users would click, respecting a strict time constraint. The lab challenges participants to compete in either a "living lab" (Task 1) or perform an evaluation that replays recorded streams (Task 2). In this report, we discuss the objectives and challenges of the NewsREEL lab, summarize last year's campaign and outline the main research challenges that can be addressed by participating in NewsREEL 2016.

1 Introduction

Many online news portals display at the bottom of their articles a small widget box labelled, "You might also be interested in", "Recommended articles", or similarly where users can find a list of recommended news articles. From a technical point of view, this recommendation use case is rather challenging. First of all, recommendations are required in real-time whenever a visitor accesses a news article on one of these portals. Instead of computing 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. Moreover, news content publishers constantly update their existing news articles, or add new content. The short lifecycle of items and the strict time-constraints for recommending news

articles make great demands on the recommender strategies. 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. Thus, the recommender algorithms must be steadily adapted to meet the special requirements of the news recommendation scenario (e.g., [11]).

Since 2014, this recommendation scenario is addressed in the News REcommendation Evaluation Lab (NewsREEL)[1], a campaign-style evaluation lab of CLEF [7]. The lab focuses on support through (personalized) content selection in form of news recommendations. The NewsREEL challenge supports recommender system benchmarking in making a critical step towards wide-spread adoption of online benchmarking (i.e., "living lab evaluation" [14]). Further, the lab supports offline evaluation of stream recommendation, hence allowing multi-dimensional evaluation of stream-based recommender systems. Testing of stream-based algorithms is important for companies offering recommender systems services, or providing recommendations directly to their customers. However, until now, such testing has occurred in house. Consistent, open evaluation of algorithms across the board was frequently impossible. Because NewsREEL provides a comprehensive data set and enables reproducible evaluation of recommender system algorithms, it has the power to reveal underlying strengths and weaknesses of algorithms across the board. Such evaluation provides valuable insights which help to drive forward the state of the art.

This report outlines the objectives and challenges of NewsREEL, summarizes the 2015 lab and ends with an overview of the 2016 edition of the lab that will be organized as part of CLEF 2016.

2 CLEF NewsREEL

CLEF NewsREEL comprises two tasks evaluating news recommendation algorithms. Both tasks involve using streams of interactions between news portals and their readers. Task 1 provides access to an operating news recommender systems. Participants receive recommendation requests and can monitor how readers reacted to their suggestions. This scenario can be seen as an example of Evaluation-as-a-Service [10, 20] where participants access an API rather than receiving a data set. Task 2 offered a log file. Participants ought to use it as ground truth for a simulation-based evaluation. Thereby, they issue the same request to different algorithms and compare their performances. In addition, we can measure factors such as time and space complexity.

2.1 Task 1: Online Evaluation

In the first sub-task, the idea of living laboratories is implemented, i.e., researchers gain access to the resources of a company to evaluate different recommendation techniques using A/B testing. A/B testing aims to benchmark varieties of a recommender system by a larger group of users (e.g., [15, 23]). It is 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 [6, 19]. While online evaluation is the de-facto standard evaluation methodology in Industry, university-based researchers often lack access to either infrastructure or user base to perform online evaluation on a larger scale.

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

NewsREEL is the first living lab where researchers gain access to both infrastructure and user requests to benchmark algorithms for information access systems using A/B testing. The living lab is described in detail in [12]. A similar, somewhat more constrained, IR-centric set-up is implemented in the CLEF Living Labs for Information Retrieval lab [2].

Within NewsREEL, the infrastructure is provided by plista GmbH[2], a company that offers recommendation services for online publishers. Whenever a user requests an article from one of their customers' web portals, plista recommends further articles that the user might be interested in. In NewsREEL, plista outsources a subset of this recommendation task to interested researchers via their Open Recommendation Platform (ORP) [3]. Once a user visits a news web page assigned to the NewsREEL challenge, a recommendation request is sent to a randomly selected team who registered with ORP. For each request, the team then has to provide a list of up to six recommendations. Providing recommendations to real users, a time constraint of 100ms is set for completing the recommendation request.

2.2 Task 2: Offline Evaluation

The evaluation of recommender algorithms online in a living lab leads to results that are difficult to reproduce since the set of users and items as well as the user preferences change continuously. This hampers the evaluation and optimization of algorithms due to the fact that different algorithms or different parameter settings cannot be tested in an exactly repeatable procedure [13, 24]. Addressing this issue, the second sub-task of NewsREEL focuses on simulating a constant data stream as provided by ORP. In contrast to the first scenario, performing an offline evaluation allows us to issue the same request to different algorithms and subsequently compare them. Additionally, it allows to measure factors such as time and space complexity. The offline task is described in more detail in [18].

We provide a large data set comprising interactions between users and various news portals in a two-month time span. The data set is described in detail in [16]. Since these news portals publish articles in German, around 80% of all users come from one of the German-speaking countries in Central Europe. Figure 1 highlights the regions from where interactions are usually triggered.

Moreover, we employ the benchmarking framework Idomaar[3] that makes it possible to simulate data streams by "replaying" a recorded stream. The framework is being developed in the CrowdRec project and adopts open-source technologies widely known by the research community to allow handling of large-scale streams of data (e.g., Apache Kafka, Apache Spark, etc.). Idomaar allows us to execute and test the proposed news recommendation algorithms, independently of the execution framework and the language used for the development. Participants in this task had to predict users clicks on recommended news articles in simulated real-time. The proposed algorithms were evaluated with respect to both functional (i.e., recommendation quality) and non-functional (i.e., response time) metrics.

3 Summary: NewsREEL 2015

Forty-two teams registered for CLEF NewsREEL 2015. Of these teams, 38 teams expressed interest in both tasks. Participating teams distributed across the world including all conti-

[2] http://plista.com/

[3] http://rf.crowdrec.eu/

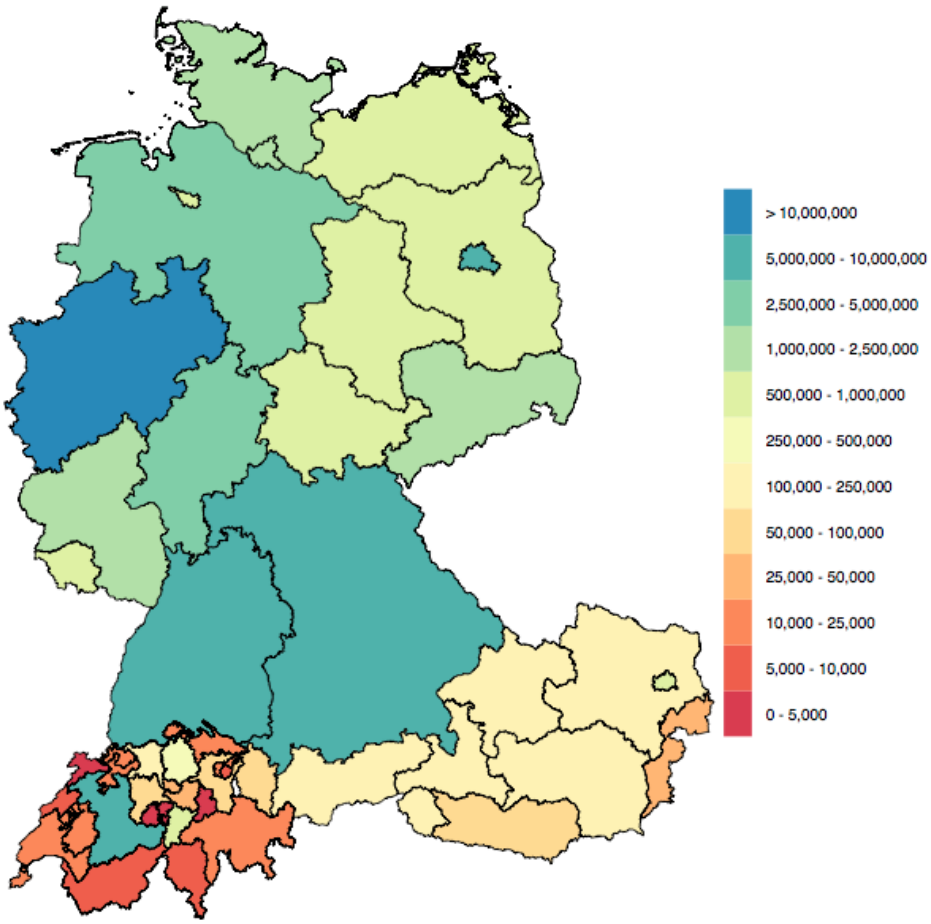

Figure 1: First-level and second-level NUTS in Germany, Austria, and Switzerland from were requests for articles were triggered. The scale indicates the number of requests during one month.

nents except Australia. In the remainder of this section, we provide a brief overview of the 2015 benchmarking campaign. For a detailed overview, we refer to [17].

3.1 Task 1

Nine teams actively competed in Task 1. Each team could operate several recommendation services. Plista provided five virtual machines to participants who were physically located far from their own infrastructure in Berlin, Germany. Without these machines participants would have faced issues with network latency. The participants could evaluate the performance of their algorithms throughout the whole campaign. However, in order to directly compare their performances, we defined three evaluation time frames (17–23 March, 7–13 April, and 5 May to 2 June 2015) during which we measured the performance based on the click-through-rate (CTR). We provided a baseline algorithm implementing a simple, but powerful recommendation strategy. The strategy recommended users the items most recently requested by other users. The idea behind this strategy is that items currently interesting to users might also be interesting for others. Thereby, the strategy assumes that users are able to determine relevant articles for others.

4 Task 2

The offline evaluation (based on a data set recorded between July and August 2015) enabled the reproducible evaluation of stream-based recommender algorithms. Having complete knowledge about the data set allowed us to implement new baseline strategies. In addition to the baseline recommender used in Task 1, we implemented the "optimal" recommender. The recommender searches in the data set the items that will be rewarded for the current request by the evaluation component. This strategy used knowledge about the future. Thus, the strategy is not a recommender algorithm; it only implements a data set look-up. Consequently, this strategy cannot be used in the online "live" evaluation. Nevertheless the measured CTR of the optimal recommender algorithm is interesting since the strategy allows us to measure the upper bound for the CTR in the analyzed setting.

5 Outlook: NewsREEL 2016

The lab has been accepted to run again as part of the next CLEF conference. Similar to 2015, the next round offers the same tasks that have been introduced in Section 2. We argue that these tasks provide a framework to tackle current research challenges that will be of interest for the IR and RecSys communities. In the remainder of this section, we outline some of the major research challenges that can be addressed by NewsREEL 2016.

First, the lab allows to study the relation between offline and online evaluation further. Prior studies (e.g., [8]) indicate that evaluation results with respect to recommendation precision differ significantly between online and offline evaluation, hence requiring additional research. By comparing results from both tasks, the lab has potential to address the research question in what extend the results obtained in the offline evaluation can be transferred into an online scenario. The offline evaluation guarantees reproducibility and allows us to compare algorithms on identical data. Online evaluation reflects algorithms' actual utility for users as their behavior is monitored. Our experiences from previous editions of NewsREEL indicate that the technical complexity of algorithms (e.g., scalability and response time) is similar in both NewsREEL tasks.

Second, the lab supports further research on algorithms' qualities beyond accuracy, a trending topic in recommender systems research (e.g., [1]). Besides, due to the time restrictions posed by the recommendation setting, a trade-off between accurate prediction and time-efficient algorithms needs to be found. In order to ensure the exact reproducibility of results and a fair comparison between different teams, standardized virtual machines are provided. This ensures that all teams use both the identical data set and exactly the same "virtual" hardware. This provides the basis for analyzing the technical complexity as well as the scalability of the algorithms. In order to hide the complexity of building the evaluation environment, we employ the Idomaar framework and facilitate getting started with it.

Third, we want to encourage participants to apply content-based approaches. Content-based techniques are suitable for tackling cold-start problems [22], a constant challenge within NewsREEL due to the steady changing set of items. As requested by various researchers, we intend to include English-speaking news portals in the NewsREEL challenge, hence allowing the participants to apply established text-processing and clustering algorithms.

Fourth, studies (e.g., [21]) suggest that the performance of recommender approaches highly depends on the context (e.g., the hour of the day). This means, that there is not the overall best recommender outperforming all other algorithms in all contexts. An open

research question is how to combine different recommender algorithms. A strategy is needed that selects the best recommender algorithm based on the user context and news situation. We expect the potential of hybrid approaches to be far from exhausted.

Finally, NewsREEL can serve as a test-bed to further study recommendation techniques for items that are provided in the form of a constant data stream. Streamed data triggers specific challenges for recommender systems (e.g., [4, 5]) as approaches that center around modeling recommendation as user-specific selection from static collections of items cannot easily be applied.

6 Conclusion

In this report, we introduced the campaign-style evaluation lab NewsREEL that focuses on benchmarking news recommendations in real-time. As we have shown, the lab addresses a number of open research challenges in the scope of information filtering. This includes in particular finding the trade-off between accuracy and speed of recommendation algorithms, the role and opportunities of context, hybrid, and content-based approaches, and innovative approaches to recommend items from a stream of data.

During the industry track of SIGIR15 in Chile, David Hawking mentioned [9] that we are now observing a significant balance shift between academic and industrial research presented at SIGIR. In fact, he reported an increase from 15% of industrial research papers presented at SIGIR'98 compared to 41% presented at SIGIR'15. He clarifies that the main reason for this tendency is the increasing need to evaluate research methods using large-scale data sets or user studies. By providing both large scale transaction data set and access to a large user base, we argue that NewsREEL can play an important role in closing this gap between academia and industry. A key challenge that we are facing right now is the connection between online and offline evaluation. As explained above, online evaluation, also referred to as A/B testing, is the standard evaluation methodology in industry. As A/B testing comes with its own requirements (e.g., the use of different evaluation metrics, reacting to user dynamics, scalability, to name a few), prior research performed in an offline setting can not easily be applied in an online context. Therefore, online testing still requires significant research efforts. By offering the same recommendation scenario in an online and offline setting, respectively, we argue that NewsREEL can play an important role to address this issue.

Concluding, we would like to explicitly invite interested researchers to take the opportunity to contribute significant knowledge to the field by tackling these challenges. Registration via the CLEF 2016 website[4] opened in November 2015.

Acknowledgment

The work leading to these results has received funding (or partial funding) from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 610594 (CrowdRec).

[4] `http://clef2016.clef-initiative.eu/`

References

[1] Xavier Amatriain, Pablo Castells, Arjen P. de Vries, Christian Posse, and Harald Steck, editors. *Proceedings of the Workshop on Recommendation Utility Evaluation: Beyond RMSE, RUE 2012, Dublin, Ireland, September 9, 2012*, volume 910 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.

[2] Krisztian Balog, Liadh Kelly, and Anne Schuth. Head first: Living labs for ad-hoc search evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1815–1818, New York, NY, USA, 2014. ACM.

[3] Torben Brodt and Frank Hopfgartner. Shedding Light on a Living Lab: The CLEF NewsREEL Open Recommendation Platform. In *Proceedings of the Information Interaction in Context conference*, IIiX'14, pages 223–226. Springer-Verlag, 2014.

[4] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: Experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1185–1194, New York, NY, USA, 2010. ACM.

[5] Ernesto Diaz-Aviles, Lucas Drumond, Lars Schmidt-Thieme, and Wolfgang Nejdl. Real-time top-n recommendation in social streams. In *Sixth ACM Conference on Recommender Systems, RecSys '12, Dublin, Ireland, September 9-13, 2012*, pages 59–66, 2012.

[6] Susan Dumais. Evaluating IR in situ. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, page 2, 2009.

[7] Nicola Ferro. CLEF 15th birthday: Past, present, and future. *SIGIR Forum*, 48(2):31–55, 2014.

[8] Florent Garcin, Boi Faltings, Olivier Donatsch, Ayar Alazzawi, Christophe Bruttin, and Amr Huber. Offline and online evaluation of news recommender systems at swissinfo.ch. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 169–176, New York, NY, USA, 2014. ACM.

[9] David Hawking. If SIGIR had an academic track, what would be in it? In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, page 1077, 2015.

[10] Frank Hopfgartner, Allan Hanbury, Henning Mueller, Noriko Kando, Simon Mercer, Jayashree Kalpathy-Cramer, Martin Potthast, Tim Gollup, Anastasia Krithara, Jimmy Lin, Krisztian Balog, and Ivan Eggel. Report of the Evaluation-as-a-Service (EaaS) Expert Workshop. *SIGIR Forum*, 49(1):57–65, 2015.

[11] Frank Hopfgartner and Joemon M. Jose. Semantic user modelling for personal news video retrieval. In *Advances in Multimedia Modeling, 16th International Multimedia Modeling Conference, MMM 2010, Chongqing, China, January 6-8, 2010. Proceedings*, pages 336–346, 2010.

[12] Frank Hopfgartner, Benjamin Kille, Andreas Lommatzsch, Till Plumbaum, Torben Brodt, and Tobias Heintz. Benchmarking news recommendations in a living lab. In *5th International Conference of the CLEF Initiative*, pages 250–267, 2014.

[13] Frank Hopfgartner, Jana Urban, Robert Villa, and Joemon M. Jose. Simulated testing of an adaptive multimedia information retrieval system. In *International Workshop on*

Content-Based Multimedia Indexing, CBMI '07, Bordeaux, France, June 25-27, 2007, pages 328–335, 2007.

[14] Diane Kelly, Susan T. Dumais, and Jan O. Pedersen. Evaluation challenges and directions for information-seeking support systems. *IEEE Computer*, 42(3):60–66, 2009.

[15] Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. Optimised scheduling of online experiments. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 453–462, 2015.

[16] Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias 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] Benjamin Kille, Andreas Lommatzsch, Roberto Turrin, András Serény, Martha Larson, Torben Brodt, Jonas Seiler, and Frank 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.

[18] Benjamin Kille, Andreas Lommatzsch, Roberto Turrin, András Serény, Martha Larson, Torben Brodt, Jonas Seiler, and Frank Hopfgartner. Stream-Based Recommendations: Online and Offline Evaluation as a Service. In *Proceedings of the Sixth International Conference of the CLEF Association, CLEF'15*, pages 497–517, 2015.

[19] Ron Kohavi. Online Controlled Experiments: Lessons from Running A/B/n Tests for 12 Years. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, page 1, 2015.

[20] Jimmy Lin and Miles Efron. Evaluation as a service for information retrieval. *SIGIR Forum*, 47(2):8–14, 2013.

[21] Andreas Lommatzsch. Real-time news recommendation using context-aware ensembles. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, pages 51–62, 2014.

[22] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 73–105. Springer US, 2011.

[23] Anne Schuth, Katja Hofmann, and Filip Radlinski. Predicting search satisfaction metrics with interleaved comparisons. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 463–472, 2015.

[24] Ryen W. White, Ian Ruthven, Joemon M. Jose, and C. J. van Rijsbergen. Evaluating implicit feedback models using searcher simulations. *ACM Trans. Inf. Syst.*, 23(3):325–361, 2005.