Renewal: an online competition platform for news recommender systems

Julien Hay[a,b,c], **Erik M. Bray**[b], **Bich-Liên Doan**[b,c], **Fabrice Popineau**[b,c], **Anne-Catherine Letournel**[b], **and Ouassim Ait Elhara**[a]

[a]Octopeek SAS, 95880 Enghien-les-Bains, France; [b]Laboratoire de Recherche en Informatique, Paris-Saclay University, 91190 Gif-sur-Yvette, France; [c]CentraleSupélec, Paris-Saclay University, 91190 Gif-sur-Yvette, France

This paper presents the Renewal platform and the new features it will provide for the online evaluation of news recommender systems. The platform is currently under development. In future organized competitions, it will allow the collection of real-time feedback from users reading news stories through a dedicated mobile application in order to evaluate competing recommender systems in online settings. We discuss the benefits over offline settings and the differences with existing online platforms dedicated to this task. We also present the results of a simulation allowing us to determine under which conditions, notably in terms of number of users and number of evaluation days, it is possible to organize a Renewal competition.

News Recommendation | Online Evaluation | Recommender System

News recommendation

News recommender systems aim to reduce the information overload facing users by filtering news with regard to their interests. News recommendation is a specific task in the area of recommender systems because of the nature of the items of interest (e.g. dynamic popularity, rich textual content, large and continuously growing catalog, short lifespan (1)) and the users' consumption (e.g. the need for fresh news, long-term and short-term interests (2), the importance of the context such as the day of the week (3) and the location (4)). Less recent items are less relevant to a recommendation, implying the item cold-start problem, thus collaborative filtering methods are less used than content-based filtering (5). Recently, deep learning based methods, especially with attention layers, are shown to outperform standard recommendation algorithms (6, 7). Deep Reinforcement Learning was also applied to news recommendation (8).

In the research community of recommender systems, it is common practice to use offline experiments to evaluate recommender systems. These experiments often involve predicting the correct ranking of candidate items according to the recorded clicks of a set of users who have previously browsed an online newspaper. The accuracy objective often leads the over-personalization problem. Thus, ideally, recommendation lists must be *diverse* (i.e. containing non-redundant items) and *novel* with regards to the already read items (9). The list must contains relevant and surprising news items the user would not have discovered otherwise: this is called *serendipity* (10). Used datasets are, for instance, MIND (11), Adressa (12) and Plista (13).

However, these evaluations make strong assumptions that do not always correlate with user satisfaction (14). First, recorded clicks on items do not necessarily correspond to relevance of those items in an online recommendation scenario (15). Similarly, non-clicked items are not necessarily irrelevant items in an online recommendation scenario (16). Second, offline evaluation does not integrate user-system interactions: an action of the user can change their interests and the system's choices (17). Thus, the assessment of the usefulness of a system necessarily involves its online evaluation, with real user-system interactions in order to avoid evaluation biases of offline settings. Moreover, it has been proven that offline efficiency does not always correlate with online performances (18).

Not only do the majority of the authors conducting offline evaluation use their own data as pointed out in (5) ($> 80\%$ of reviewed papers), but also most of the online platforms are private: Google News (19), Yahoo! News (20), Tribune de Geneve (21), Forbes (22), swissinfo.ch (18), LePoint (22). CLEF NewsREEL (13) was the only competition platform (from 2015 to 2017) offering the research community the opportunity to conduct online evaluation of their recommender systems by connecting them to an online service. The authors intended to "close this gap between academia and industry" (13). The major advantage of this platform was that it allowed recommendations to be delivered to a large number of users in real time through several German online newspapers. This platform has led to the publication of numerous research works on the news recommendation task (23–25). However, it had a number of limitations that we intend to overcome by proposing the Renewal platform.

The Renewal platform

The limitations of NewsREEL (13) and proposed solutions are:

- Recommendations in NewsREEL involve multiple sources but were only performed on each newspaper independently (recommendations could not be made cross-newspaper). For Renewal challenges, we implemented a mobile application where users will directly read news coming from multiple newspaper and other online publication sources. A beta version of the mobile application is currently in closed testing on the Google Play Store.
- Recommendations were contextualized: news items were displayed below the content of a news item on which the user just clicked. In Renewal, the mobile application is a dedicated application: the home page displays a scrollable feed of news items that can be refreshed at the user's will.
- In NewsREEL, most users had a short history of actions because only browser sessions were considered. In Renewal, users are registered by the app (either anonymously, or by their e-mail address when linking their account between devices). Thus recommender systems have access to long-term interests through long-term browsing histories.

- The most used metric, the click-through rate (CTR), does not assess whether a user has read a news item or not (26), and less so whether the user found the item to be of interest. In Renewal, we consider a positive relevance feedback to be the combination of a click-through and a scrolling time greater than a predefined threshold: we call it the *click-and-read* feedback.

- With the agreement of users, we will provide contextual information to recommender systems (e.g. location, age, gender). We will also provide extra information such as the full content of news items which are semantically rich, since most of the data attached to the items in NewsREEL are the titles of the articles, and optionally their summary.

- For positive user experience, especially in news recommendation, it is important to respond to user requests in a short period of time (26). The NewsREEL platform imposes a response time of 100 milliseconds (13). To avoid this constraint, the next list of news item recommendations to a user is preloaded in their instance of the mobile app. Moreover, we provide baseline recommender systems which are intended to substitute for competitor systems in the case that they fail to respond. Competitor systems are allowed to pre-compute their recommendation lists. This alleviates the constraint of having to respond immediately at each request of users. Competitor systems can respond in advance and may refine their recommendation lists periodically. Finally, we do not mandate recommendation systems to model every registered user which can be resource intensive when there are a large number of users. Instead, we assign uniformly distributed slices of the userbase to each system and randomly re-assign them every n days, where n is adjustable.

For continuous evaluation of competitors' systems, we will use the interleaving method which consists of interleaving items from two competing systems. Then, it is only necessary to collect *click-and-read* feedback to decide whether system A recommends more relevant items than system B with methods such as *Team Draft* (27). This interleaving evaluation has proven to be more effective than the A/B testing used with statistical measures such as the click-through rate (CTR) (28, 29). Indeed, when using A/B testing, a user receives recommendations from only one system at the same time. The user's consumption habits, the location, the day of the week, and so on, will makes it necessary to perform a large number of tests to counteract these context biases. On the contrary, the interleaving method allows more than one system to share an evaluation context. After deciding which system (A or B) won according to users' feedback, we'll use an Elo-like rating system to generate the global ranking of systems (30) that we will display on the Renewal website in real time.

This project offers many other opportunities in the study of the news recommendation and its issues:

- crowd-labeling (e.g. fake news detection, quality, readability);

- solving the user/item cold start problem;

- recommending complementary news to the user when they finish reading an article.

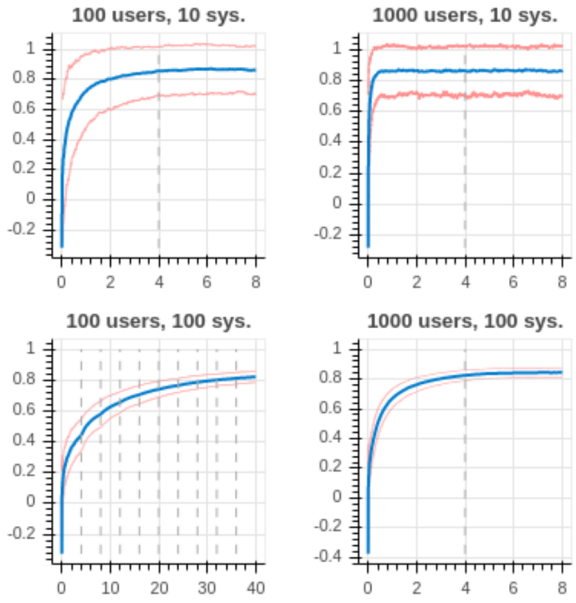

Fig. 1. Kendall's τ coefficient as a function of the number of days elapsed. Each graph is the average of 400 simulations. Red curves correspond to the average $\pm 2\sigma$ and comprise 95% of the values. Vertical lines are re-assignments of users to systems.

Predicting requirements by simulating a competition

Before organizing a competition, it is necessary to have an idea of how many users are needed to compute a ranking of the competitors in a reasonable time. For that purpose, we have conducted several simulations of competitions with different initial settings*. Figure 1 shows results of these simulations. We set different numbers of recommender systems and users. All systems have a random value (between 0.2 and 0.8) which corresponds to their probability to win against a random opponent. We did not simulate interleaving but only the outcomes of A versus B, which can lead to A wins, B wins, or a draw. Users are "active" users only, meaning they read news two times a day, which generated two competing outcomes of same systems. To assess the reliability of rankings obtained using an Elo-like rating system, we used the Kendall's τ coefficient (31) which compute the correlation between the predicted ranking and the ground truth ranking of the systems.

Results show that rankings become reliable (Kendall's τ of 0.82) in less than 4 days in cases where we have at least 10 times more users than systems. In the worst case (100 system and 100 users), rankings converge to the same Kendall's τ in 40 days. This indicates that a competition remains feasible with extremely unfavorable conditions.

References

1. JA Gulla, C Marco, AD Fidjestøl, JE Ingvaldsen, Ö Özgöbek, The intricacies of time in news recommendation in *Late-breaking Results, Posters, Demos, Doctoral Consortium and Workshops Proceedings of the 24th ACM Conference on User Modeling, Adaptation and Personalisation (UMAP 2016), Halifax, Canada, July 13-16, 2016.* (2016).
2. M An, et al., Neural news recommendation with long- and short-term user representations in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* (Association for Computational Linguistics, Florence, Italy), pp. 336–345 (2019).
3. A Lommatzsch, B Kille, S Albayrak, Incorporating context and trends in news recommender

*The source code is available at https://github.com/hayj/RenewalSimulator

systems in *Proceedings of the International Conference on Web Intelligence*, WI '17. (Association for Computing Machinery, New York, NY, USA), p. 1062–1068 (2017).
4. C Chen, X Meng, Z Xu, T Lukasiewicz, Location-aware personalized news recommendation with deep semantic analysis. *IEEE Access* **5**, 1624–1638 (2017).
5. S Raza, C Ding, A survey on news recommender system – dealing with timeliness, dynamic user interest and content quality, and effects of recommendation on news readers (2020).
6. H Wang, F Zhang, X Xie, M Guo, Dkn: Deep knowledge-aware network for news recommendation (2018).
7. C Wu, et al., NPA: neural news recommendation with personalized attention. *CoRR* **abs/1907.05559** (2019).
8. G Zheng, et al., Drn: A deep reinforcement learning framework for news recommendation in *Proceedings of the 2018 World Wide Web Conference*, WWW '18. (International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE), p. 167–176 (2018).
9. M Kaminskas, D Bridge, Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Trans. Interact. Intell. Syst.* **7** (2016).
10. D Kotkov, S Wang, J Veijalainen, A survey of serendipity in recommender systems. *Knowledge-Based Syst.* **111**, 180 – 192 (2016).
11. F Wu, et al., MIND: A large-scale dataset for news recommendation in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. (Association for Computational Linguistics, Online), pp. 3597–3606 (2020).
12. JA Gulla, L Zhang, P Liu, O Özgöbek, X Su, The adressa dataset for news recommendation in *Proceedings of the International Conference on Web Intelligence*, WI '17. (Association for Computing Machinery, New York, NY, USA), p. 1042–1048 (2017).
13. F Hopfgartner, et al., Benchmarking news recommendations: The clef newsreel use case. *SIGIR Forum* **49**, 129–136 (2016).
14. CN Ziegler, SM McNee, JA Konstan, G Lausen, Improving recommendation lists through topic diversification in *Proceedings of the 14th International Conference on World Wide Web*, WWW '05. (Association for Computing Machinery, New York, NY, USA), p. 22–32 (2005).
15. A Bellogín, A Said, Recommender systems evaluation in *Encyclopedia of Social Network Analysis and Mining :*. 2 edition, (2018).
16. J Beel, M Genzmehr, S Langer, A Nürnberger, B Gipp, A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation in *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, RepSys '13. (ACM, New York, NY, USA), pp. 7–14 (2013).
17. AD Myttenaere, B Golden, BL Grand, F Rossi, Study of a bias in the offline evaluation of a recommendation algorithm. *CoRR* **abs/1511.01280** (2015).
18. J Liu, P Dolan, ER Pedersen, Personalized news recommendation based on click behavior in *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10. (Association for Computing Machinery, New York, NY, USA), p. 31–40 (2010).
19. J Liu, P Dolan, ER Pedersen, Personalized news recommendation based on click behavior in *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10. (Association for Computing Machinery, New York, NY, USA), p. 31–40 (2010).
20. L Li, W Chu, J Langford, X Wang, Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11. (Association for Computing Machinery, New York, NY, USA), p. 297–306 (2011).
21. F Garcin, C Dimitrakakis, B Faltings, Personalized news recommendation with context trees in *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13. (Association for Computing Machinery, New York, NY, USA), p. 105–112 (2013).
22. E Kirshenbaum, G Forman, M Dugan, A live comparison of methods for personalized article recommendation at Forbes.com. *Lect. Notes Comput. Sci. (including subseries Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **7524 LNAI**, 51–66 (2012).
23. Y Liang, B Loni, MA Larson, Clef newsreel 2017: Contextual bandit news recommendation. in *CLEF (Working Notes)*. (2017).
24. PD Beck, M Blaser, A Michalke, A Lommatzsch, A system for online news recommendations in real-time with apache mahout. in *CLEF (Working Notes)*. (2017).
25. J Yuan, A Lommatzsch, B Kille, Clicks pattern analysis for online news recommendation systems. in *CLEF (Working Notes)*. pp. 679–690 (2016).
26. B Kille, F Hopfgartner, T Brodt, T Heintz, The plista dataset in *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge*, NRS '13. (Association for Computing Machinery, New York, NY, USA), p. 16–23 (2013).
27. F Radlinski, M Kurup, T Joachims, How does clickthrough data reflect retrieval quality? in *CIKM '08*. (2008).
28. A Schuth, K Hofmann, F Radlinski, Predicting search satisfaction metrics with interleaved comparisons in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15. (Association for Computing Machinery, New York, NY, USA), p. 463–472 (2015).
29. E Kharitonov, C MacDonald, P Serdyukov, I Ounis, Generalized team draft interleaving in *CIKM '15*. (2015).
30. R Herbrich, T Minka, T Graepel, Trueskill™: A bayesian skill rating system in *Advances in Neural Information Processing Systems 19*, eds. B Schölkopf, JC Platt, T Hoffman. (MIT Press), pp. 569–576 (2007).
31. MG Kendall, The treatment of ties in ranking problems. *Biometrika* **33**, 239–251 (1945).