Comparing offline and online recommender system evaluations on long-tail distributions

Gabriel S. P. Moreira
CI&T, Campinas, SP, Brazil
gabrielpm@ciandt.com

Gilmar Souza
CI&T, Campinas, SP, Brazil
gilmarj@ciandt.com

Adilson M. da Cunha
ITA, Sao Jose dos Campos, SP, Brazil
cunha@ita.br

ABSTRACT

In this investigation, we conduct a comparison between offline and online accuracy evaluation of different algorithms and settings in a real-world content recommender system. By focusing on recommendations of long-tail items, which are usually more interesting for users, it was possible to reduce the bias caused by extremely popular items and to observe a better alignment of accuracy results in offline and online evaluations.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - information filtering.

Keywords

Recommender systems, offline evaluation, online evaluation, click-through rate, accuracy metrics, long-tail.

1. EVALUATION METHODOLOGY

This investigation focuses in a comparison between offline and online evaluation results in a recommender system implemented in Smart Canvas®, a platform that delivers web and mobile user experiences through curation algorithms. Smart Canvas features a mixed hybrid recommender system, in which items recommended by all available algorithms are aggregated and presented to users.

It was conducted in one production environment, which consists in the website of a large shopping mall. The accuracy of different recommender algorithms and variations of their settings were assessed in offline evaluation and further compared to online measures with real users (A/B testing).

In this investigation, three experiments were conducted, each of them varying only one setting at a time, in both offline and online evaluations. They involve two algorithms implemented in Smart Canvas: Content-Based Filtering (based on TF-IDF and cosine distance) and Item-Item Frequency (a model-based algorithm based on co-frequency of items interactions in user sessions).

For all experiments, accuracy was evaluated under two perspectives considering (1) all recommended items and (2) only long-tail items. The main reasons for this two-fold analysis is that recommendations of non-popular items matching users interests might be more relevant to them. Popular items may also bias the evaluation of recommenders accuracy.

1.1 Offline Evaluation

Offline evaluation is usually done by recording the items users have interacted with, hiding some of this user-item interactions (test set) and training algorithms on the remaining information (train set) to assess the accuracy.

A time-based approach [3] was used to split train and test sets. User interactions occurred during the period before the split date were used as train set (20 days), and the period after composed the test set (8 days), as shown in Figure 1.

Figure 1: The Offline Evaluation Dataset Split

It simulates the production scenario, where the known user preferences until that date are used to produce recommendations for the near future. Test set comprised 342 users in common with train set, with a total of 636 interactions during test set.

This investigation uses an offline evaluation methodology named as One-Plus-Random or RelPlusN [3], in which for each user the recommender is requested to rank a list with relevant items (those that the user has interacted with in the test set) and a set of N non-relevant items (random items, which the user has never interacted with).

The final performance are averaged over Click-Through Rate (CTR), a common metric for recommender and advertising systems, here referred as Offline CTR. It was calculated as a ratio between the top recommended items, which the users in fact interacted in test set, and the total number of simulated recommendations.

1.2 Online Evaluation

For online evaluation, an engine was developed to randomly split users traffic and assign to one of the experiments of the hybrid recommender system (A/B testing), each varying only one setting of the two component algorithms. The online evaluation involved 402 distinct items, 45,000 users, 5,850 recommendations, and 183 interactions.

The Click-Through Rate (CTR) metric was also used to measure online accuracy of recommendations. Online CTR was the ratio of interactions on recommended items and the total of recommended items viewed by users during their sessions.

2. RESULTS

Three experiments were performed in both offline and online evaluations. In Experiments #1 and #2, Content-Based Filtering settings named MinSimilarity and ItemDaysAgeLimit were assessed individually with different values. In Experiment #3, an Item-Item Frequency setting named LastXInteractedItems were varied.

Accuracy (CTR) was evaluated under two perspectives considering: (1) all recommended items, including the very popular ones and (2) only long-tail items.

The ideal scenario would be offline metrics varying in the same direction of the CTR measures. That behavior would indicate that offline evaluation could be used to cost-effectively identify the best setting values for recommender algorithms before involving users in online evaluation.

However, Online and Offline CTR behaviour did not align in perspective (1), considering all recommended items, as can be seen in Figure 2.

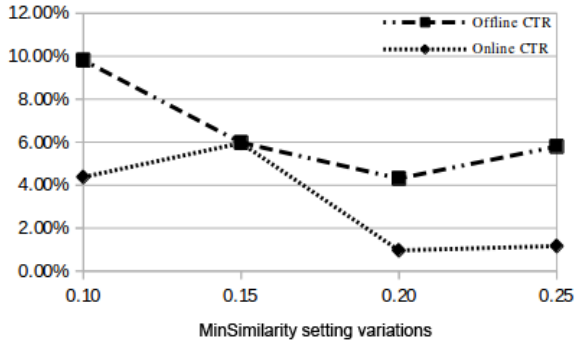

Figure 2: Experiment #1 (including popular items) - CTR for Content-Based algorithm - MinSimilarity

This investigation went further for better understanding of the misalignment between offline and online evaluations in this context. It was assessed whether the very popular items could introduce a bias in recommender accuracy analysis, ignoring extremely popular items and considering only long-tail items in perspective (2).

For offline evaluation, the top 1.1% items concentrated 22% of the interactions and were further ignored. For online experiments, it was also ignored the 1.5% most popular items, responsible for 41% of the interactions in the website.

Considering only the long-tail items in Experiment #1, the Offline and Online CTR turned out to be nicely aligned, as shown in Figure 3. The best setting value for the MinSimilarity threshold was 0.1, following the same trend for both CTR metrics.

In Experiment #2 for long-tail items, the metric variations were very similar to the results considering popular items, so there was no prediction gain by removing very popular items from the analysis.

In Experiment #3, the CTR metrics variation were yet more aligned by keeping only long-tail items (charts omitted due to space reasons).

In Experiments #1 and #3, considering only long-tail items, offline evaluation was an adequate predictor of the online accuracy as a function of their setting thresholds.

The observed bias of popular items over evaluation accuracy metrics are aligned to recent studies like [1] and [2].

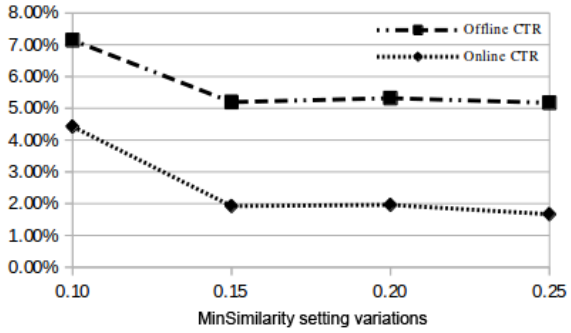

Figure 3: Experiment #1 (long-tail) - CTR for Content-Based algorithm - MinSimilarity

3. CONCLUSION

In this study, Offline and Online experiments were performed and compared in a real production environment of a hybrid recommender system. The results did not correlate for most experiments, but when focusing on long-tail items, it was possible to observe how popular items can bias the accuracy evaluation. Two out of three experiments on long-tail items had Offline CTR very aligned to Online CTR.

The evaluation of long-tail items may be a candidate for deeper investigation in future studies, aiming to increase confidence in offline evaluation results. Furthermore, focusing on accuracy optimization for long-tail items, algorithms may bring to the users a clear perception of the ability of the system to recommend non-trivial relevant items.

This study is still ongoing to provide a better understanding of the relationship between offline and online evaluation results. Besides accuracy, it is suggested a similar investigation of other properties like coverage and more long-term metrics, related to users engagement.

4. ACKNOWLEDGEMENTS

Our thanks to CI&T for supporting the development of Smart Canvas® recommender system evaluation framework and to the ITA for providing the research environment.

5. REFERENCES

[1] J. Beel, M. Genzmehr, S. Langer, A. Nürnberger, and B. Gipp. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In *Proc. Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, pages 7–14. ACM, 2013.

[2] F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber. Offline and online evaluation of news recommender systems at swissinfo. ch. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 169–176. ACM, 2014.

[3] A. Said and A. Bellogín. Comparative recommender system evaluation: benchmarking recommendation frameworks. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 129–136. ACM, 2014.