Contrasting Offline and Online Results when Evaluating Recommendation Algorithms

Marco Rossetti
Department of Informatics, Systems and Communication
University of Milano-Bicocca, Italy
rossetti@disco.unimib.it

Fabio Stella
Department of Informatics, Systems and Communication
University of Milano-Bicocca, Italy
stella@disco.unimib.it

Markus Zanker
Faculty of Computer Science
Free University of Bozen-Bolzano, Italy
mzanker@unibz.it

ABSTRACT

Most evaluations of novel algorithmic contributions assess their accuracy in predicting what was withheld in an offline evaluation scenario. However, several doubts have been raised that standard offline evaluation practices are not appropriate to select the best algorithm for field deployment. The goal of this work is therefore to compare the offline and the online evaluation methodology with the same study participants, i.e. a within users experimental design. This paper presents empirical evidence that the ranking of algorithms based on offline accuracy measurements clearly contradicts the results from the online study with the same set of users. Thus the external validity of the most commonly applied evaluation methodology is not guaranteed.

Keywords

User study, Evaluation methodology, Experimental within users design

1. INTRODUCTION

From a methodological viewpoint research contributions in the field of recommender systems focus mainly on novel methods, that mainly come with the promise of more accurately identifying relevant content in a variety of application domains. For instance, the survey of Jannach et al. [7] gives evidence that validation of recommender systems research focuses mainly on offline evaluation scenarios. Thus, according to the paradigm of Machine Learning a predictive model is trained on a subset of the available data and it is optimized to correctly predict the withheld portions of the dataset. Due to the maturing of the field several works have been focusing on the methodological aspects of this evaluation approach in recommender systems research.

RecSys '16, September 15 - 19, 2016, Boston , MA, USA

ISBN 978-1-4503-4035-9/16/09...$15.00

DOI: http://dx.doi.org/10.1145/2959100.2959176

For instance, Said & Bellogín. [9] identified several inconsistencies when recommendation methods and offline evaluation practices are compared across different framework platforms. While this recent work focuses on the internal validity of offline evaluation practices other authors focus on the external validity. Research questions targeting the external validity are, for instance, if and how confirmed results from an offline evaluation can be generalized to a recommender system's encounter with real users. Therefore, in addition to offline evaluation on dead data many authors strongly advocate user-centric evaluation approaches [5, 11, 8]. Up to now several authors have reported results from comparing recommendation methods in an offline and an online evaluation setting [1, 4, 3] and did find evidence that algorithm performance in offline and online evaluation settings differs. However, none of the research works so far has been able to demonstrate in a *within users* experimental design statistically significant differences in algorithm performance between both evaluation settings. This paper's contribution therefore lies in presenting an innovative evaluation design that researches algorithms' precision and their ability to present novel and relevant items in offline and online evaluation settings.

2. STUDY DESIGN

2.1 Overview

The goal of this research is to assess the external validity of a comparative evaluation of different recommendation methods on offline data. We would therefore like to answer the following research questions:

1. Does the relative ranking of algorithms based on offline accuracy measurements predict the relative ranking according to an accuracy measurement in a user-centric evaluation?

2. Does the relative ranking of algorithms based on offline measurements of the predictive accuracy for long-tail items produce comparable results to a user-centric evaluation?

3. Do offline accuracy measurements allow to predict the utility of recommendations in a user-centric evaluation?

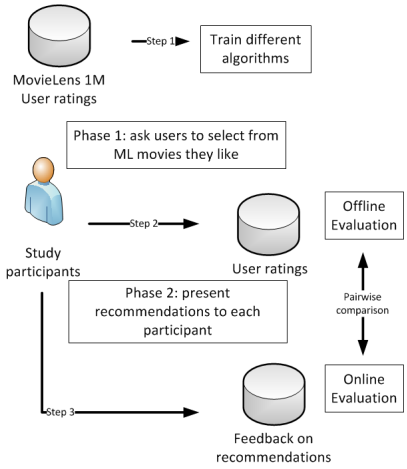

Figure 1: Big Picture

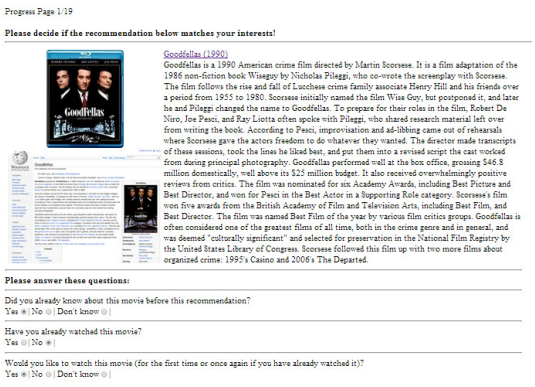

Figure 2: Screenshot of recommendations evaluation page.

For the purpose of answering these research questions we are employing a novel study design that compares offline and online measurements in a *within users* experimental design that is depicted in Figure 1. We are exploiting the MovieLens dataset in order to train different algorithms that will then be assessed according to the standard *offline* and *online* evaluation methodology by our study participants. In order to check for differences between offline and online methodology we selected four rather distinctive algorithms for this experimental setup:

1. Most-popular as a baseline algorithm, that is agnostic of the user's preferences and uniformly proposes the same very popular items to all users.

2. A matrix factorization model [12] with a 80 factor model that achieves highest offline accuracy based on a sensitivity analysis.

3. The same matrix factorization model, but with 400 factors, which is most accurate on long-tail items measured according to the proposition of [2].

4. Item-based nearest neighbour approach [10] as a simple memory-based recommendation mechanism, that recommends dissimilar sets of items compared with the other three algorithms.

Although there have been many different algorithms proposed recently the algorithm does not constitute a limitation of the findings in this study. This is because we focus on the different results of the same algorithms in an online and an offline setting and not on finding the most accurate algorithm according to a specific setting.

2.2 MovieLens Dataset

The dataset used to train the algorithms is the well-known *MovieLens 1M* (ML1M) dataset. This dataset contains 1,000,209 ratings from 6,040 users and 3,706 different movies. Ratings are integer values from an ordinal scale ranging from 1 to 5, where 5 is the best feedback. However, we wanted to ensure that efforts for our participants remain at acceptable levels and therefore decided to ask users only for unary preference statements. Therefore the ML data was transformed into unary user feedback. For each individual ML user record only 4 and 5 ratings that are greater than the user's average rating score were transformed into an unary *like* statement. Furthermore we considered only movies with an entry in the Freebase data repository in order to ensure the same data quality of movie descriptions when asking users for ratings or presenting recommendations. After this preprocessing the dataset consisted of 6,038 users and 3,086 movies. This data was then used to train the models for the four algorithms considered.

In the first phase we invited participants via our universities' mailing lists to browse our specifically developed movie portal and mark those items that they like. We designed and implemented a Django web application that offers users the possibility to participate at this user study from any location. In the first round users could access a portal that contained detailed information about all movies contained in the MovieLens dataset. The collected dataset contains positive feedback on items the user probably knows and likes and no feedback on items the user either dislikes or does not know. The algorithmic models trained on MovieLens could therefore be evaluated how accurate they were in predicting the liked items of our study participants in an offline evaluation scenario. The generation of the recommendation list for a user was performed using an all-but-one validation on the user data. In total 241 users have participated in the first phase that provided on average 137 unary ratings each.

2.3 Online evaluation

In the second round we invited first-round users to evaluate around twenty recommendations produced from these four different algorithms. Each algorithm computed 5 recommendations and we sorted them in a stratified way, such that the recommendations derived from each algorithm were equally distributed along the ranks. Since two or more algorithms can recommend the same item, the number of recommendations for each user went from a theoretical minimum of 5 to a maximum of 20 items, while the actual average was 17.37. Therefore users were asked to assess sequentially every recommended item and to answer a set of questions. For each recommended item a separate screen with information about the movie such as the title, the year of launch, the plot, the poster image extracted from Freebase and a snapshot of the related Wikipedia page was displayed (see Figure 2). The questionnaire items were the following:

- Did you already know about this movie before this recommendation? (Yes, No, Don't know)

- Have you already watched this movie? (Yes, No)

- Would you like to watch this movie (for the first time or once again if you have already watched it)? (Yes, No, Don't know)

With the first question we measure second order novelty [1], i.e. an item is considered novel if the user has never heard of it. We determine relevance of recommended items by asking about potential conversion, i.e. if the user considers to watch this movie. By combining answers from both questions we are able to identify *useful* recommendations, i.e. relevant items that the user did not yet know about.

In total 122 users returned to the second phase, out of which 100 provided their individual assessment to all recommended items. We selected these 100 users for our study and discarded data from all other users.

3. MEASUREMENT METHODOLOGY

One basic evaluation approach in recommender systems is borrowed from the field of Information Retrieval (IR). Ground truth is represented by a user × items matrix with binary rating values, where 1s indicate that the user liked the particular item and 0s indicate dislike or no rating. As users tend to give only positive ratings and ignore disliked items, the default assumption in an offline evaluation scenario is to treat unrated items as they were disliked. The standard IR measures are *Precision* and *Recall* where precision measures the share of True Positives in the list of recommended items while recall constitutes the share of True Positives among all items relevant for a specific user in ground truth. However, the semantics of these two basic IR measures are somewhat different in offline and online evaluation scenarios [6]:

- Offline scenario: True Positives might be underestimated, because users could like novel items that they did not know before and therefore no historic rating exists in ground truth. Correspondingly, False Positives might be overestimated. Due to this incompleteness of ground truth, many more items might be actually relevant for a specific user, but they are not recommended, thus also False Negatives might be underestimated.

- Online scenario: Only True and False Positives can be determined, if users are asked to rate all recommended items. Therefore, True Negatives and False Negatives remain unknown.

Based on these assumptions only Precision can be compared in our within users experiments in order to contrast online and offline evaluation results. Furthermore, we hypothesize that Precision in an offline evaluation scenario should be lower than Precision measured online due to mentioned underestimation of True Positives and the overestimation of False Positives in the offline setting. Furthermore, we also compute Precision for long tail items following the proposition of Cremonesi et al. [2], where the most popular items in the dataset are treated as negatives. The most popular items that could aggregate one third of all positive ratings in the dataset are considered as belonging to the short head.

Table 1: Accuracy statistics for all items, long tail items and useful recommendations

Setting		I2I	MF80	MF400	POP
All Items	Offline	0.438	0.504	0.454	0.34
All Items	Online	0.546	0.598	0.604	0.516
Long Tail	Offline	0.28	0.018	0.36	0
Long Tail	Online	0.356	0.054	0.528	0
Useful	Online	0.126	0.082	0.116	0.026

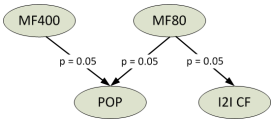

Figure 3: Precision offline

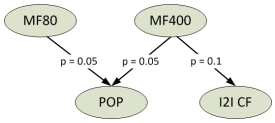

Figure 4: Precision online

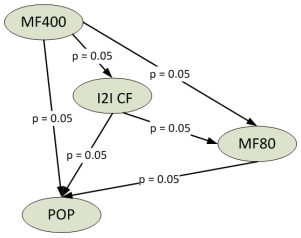

Figure 5: Domination graph for Precision on long tail items (offline and online)

4. RESULTS

Research question 1 is about the relative ranking of algorithms in offline and online accuracy assessments. In Table 1 the Precision measurements for all four algorithms and both evaluation settings are given. Furthermore, Figure 3 gives the domination graph based on pairwise comparison of a Wilcoxon signed rank test. An outgoing arc means that the method outperforms the method it is pointing to with a p-value below the significance level denoted on the arc. Both matrix factorization techniques perform better than the popularity baseline and the 80 factor model also clearly outperforms the I2I nearest neighbor method. However the online measurements clearly contradict this picture. As depicted in Figure 4 the 400 factor model outperforms the I2I at a 10% error rate, but the 80 factor model does not. Therefore this study gives empirical evidence that statistically significant differences measured in an offline evaluation scenario could not be confirmed in an online study.

Due to the strong bias of traditional accuracy measurements on popular items, Cremonesi et al. [2] proposed to exclude the most popular items from accuracy measurements (i.e. count correct predictions of very popular items as False Positives). Research Question 2 therefore asks if Precision on long tail items in offline measurements more reliably predicts the results that can be achieved in the online setting. Table 1 gives the actual figures and Figure 5 presents the domination graph. The Precision measurement of long tail items seems to better discriminate between methods and leads to exactly the same statistically significant ranking of algorithms in both evaluation settings. Obviously it clearly contradicts the offline Precision measurement, the 80 factor

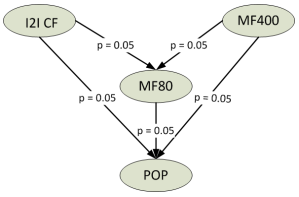

Figure 6: Domination graph based on Precision for items users liked and did not know (online)

model is by far worse in proposing non popular items than the 400 factor model or the I2I neighborhood model.

Finally the third research question explicitly explores the utility of recommendations. We simply define *useful recommendations* as propositions the user likes but did not yet know. We question if accuracy measurements on offline data provide an indication if an algorithm will be able to make more useful recommendations to users. In Table 1 Precision figures for useful recommendations are given, that are obviously much smaller than Precision on all and long tail items. When again testing for statistically significant differences the ranking of algorithms looks completely different from where we started in Figure 3. I2I collaborative filtering and the 400 factor model seem to be on par and both clearly outperform the two other methods. Neither traditional Precision measurements nor the proposed Precision of long tail items measure by [2] were able to predict this algorithm ranking by offline experiments. Although I2I CF seemed to recommend more popular items than the 400 factor model, the neighborhood method did still very good in identifying items that were novel to our study participants. Consequently, the presented empirical results clearly challenge the assumption of external validity of common evaluation practices of recommender systems and should therefore stimulate further methodological work into this direction.

5. CONCLUSIONS

The vast majority of the literature on recommender systems apply offline evaluation methodologies to assess the quality of the proposed approaches [7]. However, the external validity of such methodologies has been frequently questioned and recent works tried to compare offline and online evaluations. In this work we described a novel comparison approach between an offline evaluation protocol and an online user study. A dataset with *like* statements by 100 users on Movielens movies has been collected and accuracy statistics for models trained on the Movielens dataset have been evaluated on the collected dataset. These statistics have been compared with online statistics collected with a user study, where the same 100 users answered a set of questions on around 20 recommendations proposed to them. This work showed that offline precision underestimates online precision when both are considering all items and only long tail items. Furthermore, offline precision measurement does not provide the same ranking of algorithms as online precision does. Finally, the real utility for users of previously unknown but relevant recommended items determined on an online user study ranks algorithm in such a way that is not reproducible with offline evaluation metrics, neither on all items nor only on long tail items. Such results are a further clue that the external validity of commonly used offline evaluation protocols is not always guaranteed.

6. REFERENCES

[1] Paolo Cremonesi, Franca Garzotto, and Roberto Turrin. Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study. *ACM Trans. Interact. Intell. Syst.*, 2(2):11:1–11:41, June 2012.

[2] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 39–46, New York, NY, USA, 2010. ACM.

[3] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 161–168, New York, NY, USA, 2014. ACM.

[4] 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.

[5] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, January 2004.

[6] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender Systems An Introduction*. Cambridge University Press, 2010.

[7] D. Jannach, M. Zanker, M. Ge, and M. Groening. Recommender systems in computer science and information systems - a landscape of research. In *13th International Conference on Electronic Commerce and Web Technologies*, pages 76–87. Springer, 2012.

[8] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 157–164, New York, NY, USA, 2011. ACM.

[9] Alan Said and Alejandro Bellogín. Comparative recommender system evaluation: Benchmarking recommendation frameworks. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 129–136, New York, NY, USA, 2014. ACM.

[10] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA, 2001. ACM.

[11] Guy Shani and Asela Gunawardana. Evaluating recommendation systems. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 257–297. Springer US, 2011.

[12] Vikas Sindhwani, Serhat S. Bucak, Jianying Hu, and Aleksandra Mojsilovic. One-class matrix completion with low-density factorizations. In *Proceedings of the 2010 IEEE International Conference on Data Mining*.