Evaluating prediction-based recommenders for kids

Michael Green, Oghenemaro Anuyah, Devan Karsann
People and Research Information Team
Dept. of Computer Science
Boise State University
Boise, Idaho
{michaelgreen,oghenemaroanuyah, devankarsann}@u.boisestate.edu

Maria Soledad Pera
People and Research Information Team (PIReT)
Dept. of Computer Science
Boise State University
Boise, Idaho
solepera@boisestate.edu

ABSTRACT

In this position paper, we highlight a number of issues that exist with the use of traditional metrics in evaluating recommender systems when children are the target audience. Our focus is on discrepancies that arise as a result of the differing rating behaviour of adult users when compared to children, and how these differences can warrant a reconsideration of existing assessment metrics and their validity in this context.

KEYWORDS

recommender systems, evaluation, children

ACM Reference Format:
Michael Green, Oghenemaro Anuyah, Devan Karsann and Maria Soledad Pera. 2019. Evaluating prediction-based recommenders for kids. In *KidRec '19: Workshop in International and Interdisciplinary Perspectives on Children & Recommender and Information Retrieval Systems, Co-located with ACM IDC, June 15, 2019, Boise, ID.* ACM, New York, NY, USA, 3 pages.

KidRec '19, June 15, 2019, Boise, ID

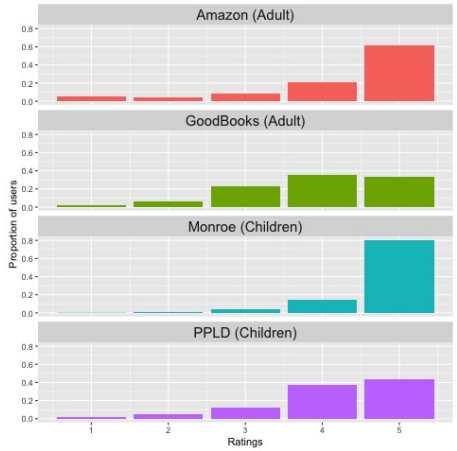

Figure 1: Rating distribution across children and adult datasets.

[1]We partnered with Pikes Peak Library District and Monroe County Public Library in order to obtain this data for research purposes

RECOMMENDER SYSTEMS & EVALUATION METRICS

Recommender systems offer users items of interest (e.g., songs, movies, or books) [4], and in generating recommendations these systems often depend on the availability of user historical data in the form of explicit (e.g., ratings) or implicit (e.g., items clicked) feedback. This user feedback constitutes the main building block in evaluating the correctness of recommender systems [3]. Commonly, recommender systems are evaluated in terms of Root Mean Square Error (**RMSE**) and Mean Absolute Error (**MAE**). Both of these metrics quantify system performance by comparing differences among predicted ratings and actual (i.e., user defined) ratings. The lower the difference the better the system. Given that (i) popular frameworks used for determining the correctness of recommender systems are directly influenced by behaviour of the target audience (i.e., users for whom recommendations are made and who directly provide ratings for offline/online experiments) and (ii) user rating behaviour across adult and younger audiences may not be similar; we argue for the need to look further into the degree which metrics such as MAE and RMSE fail to properly capture user preferences and system performance in recommender systems tailored towards children. For the rest of this paper, we report on an initial analysis conducted to examine the differences in children's rating behaviors and discuss how utilizing the aforementioned metrics for evaluating performance of recommender system algorithms on children's datasets could affect the overall results. In particular, we focus on *book* recommendations as these types of recommendations can help children develop reading skills and habits, provided the recommended material is suitable for and interesting to them [5].

EVALUATION METRICS & TARGET AUDIENCES

While RMSE and MAE are widely used metrics that represent recommender system performance, it has been called into question if they perform well with systems that use data collected directly from children [2]. To address some of the limitations in using children data with recommenders, the authors in [1] proposed leveraging publicly available datasets for adults to enhance the few available children's datasets while still retaining RMSE as the measure of performance. The results demonstrated that (a) children's datasets employing User to User and Matrix Factorization recommendation algorithms tended to have lower RMSE than adult datasets, and (b) using adult datasets to cross learn to enhance the recommendation process instead caused an increase in RMSE. It was also discovered that children rarely take advantage of the full rating scale, and rated most of the items on the high side (e.g., 4 or 5), which we believe contributed to low performance results. We explore this discovery further by examining two well-known datasets created based on adult ratings: **GoodBooks** [6] and **Amazon** Book Reviews; as well as two datasets based on children ratings curated from public libraries: the **PPLD** (Pikes Peak Library District) and **Monroe** datasets[1]. Statistics describing each of these datasets can be found on Table 1. As illustrated in Figure 1, while Amazon does have a similar distribution

Table 1: Statistics summarizing datasets considered in our analysis. *A* indicates that the target audience for the dataset are adults, while *C* indicates that the dataset is oriented to children.

Dataset	Target	Users	Ratings	Avg. ratings per user	S.D. ratings per user
Amazon	A	~ 8M	~ 22M	2.8	23.02
GoodBooks	A	~ 53K	~ 5M	111.9	26.07
Monroe	C	211	284	1.4	0.98
PPLD	C	335	1094	3.3	5.79

as Monroe, we can see notable differences in the standard deviation of average number of ratings between the adult and children datasets. There is also a dramatic difference in number of users and number of ratings. We infer that these differences can have a substantial effect on the use of distance metrics like RMSE and MAE when assessing the degree to which a recommender systems based on children ratings is "good", or not.

RMSE, CHILDREN DATA, & NEXT STEPS

Since rating distributions are heavily centered around 4/5 values in the Monroe and PPLD datasets (see Figure 1), there is a low chance of any given recommender predicting a rating score below a 4. This translates into a non-significant penalization in the computation of RMSEs. This does not hold true in the presence of a wider rating distribution and a greater number of users/ratings. In these cases, RMSEs will be much more heavily penalized for mistaken prediction ratings across the entire 0 to 5 scale. RMSE has been considered one of the go-to metrics for offline evaluations of recommender systems [4], however, we believe that with the unique attributes of children's datasets, a different approach may be required to evaluate these systems correctly. One idea is to consider the normalization of children's ratings. Normalization is the method of subtracting the average score of a user's rating from all of their ratings to provide a new distribution of ratings. Can we alter this methodology to allow children's datasets' distributions to match those of adult datasets? Is there an approximation which may not map one-to-one in distribution proportions but still provide us with better RMSE results? Can incorporating written reviews into the normalization process help us reach these goals? Should we do so, and if we do, what will be the cost to capturing true user preference?

ACKNOWLEDGMENTS

Work partially funded by NSF Award 1565937.

REFERENCES

[1] Ion Madrazo Azpiazu, Michael Green, Oghenemaro Anuyah, and Maria Soledad Pera. 2018. Can we leverage rating patterns from traditional users to enhance recommendations for children? *arXiv preprint arXiv:1808.08274* (2018).

[2] Michael D Ekstrand. 2017. Challenges in evaluating recommendations for children. In *International Workshop on Children & Recommender Systems. Available at: https://drive. google. com/file/d/0B8S_mL8gIjZwVzZSLVZKN3Q4RG8/view*.

[3] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on Fairness, Accountability and Transparency*. 172–186.

[4] Kim Falk. 2019. *Practical Recommender Systems*.

[5] Maria Soledad Pera and Yiu-Kai Ng. 2014. Automating readers' advisory to make book recommendations for k-12 readers. In *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 9–16.

[6] Zygmunt Zajac. 2017. Goodbooks-10k: a new dataset for book recommendations. http://fastml.com/goodbooks-10k. *FastML* (2017).