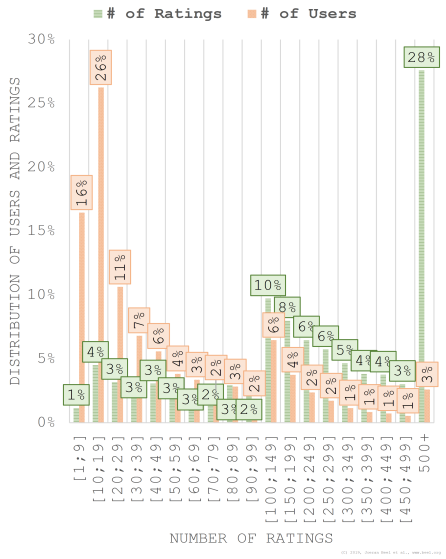


Figure 1: Distribution of users and ratings in the MovieLens dataset. 16% of users rated less than 10 movies, 26% rated between 10 and 19 movies. In most MovieLens releases (100k, 1m, 10m, ...), these 42% of users and their ratings are not included. 3% of users have 500 or more ratings and contribute 28% of all ratings in MovieLens.

*This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 13/RC/2106 and funding from the European Union and Enterprise Ireland under Grant Number CF 2017 0303-1.

[1]The 'MovieLens Latest Full' dataset is not pruned. All other variations of MovieLens are pruned (100k, 1m, 10m, 20m, and 'MovieLens Latest Small'). However, the 'MovieLens Latest Full' dataset is not recommended for research as it is changing over time.

Data Pruning in Recommender Systems Research: Best-Practice or Malpractice?

Joeran Beel*
Trinity College Dublin, School of Computer Science & Statistics, ADAPT Centre
Dublin, Ireland
beelj@tcd.ie

Victor Brunel
Polytech Clermont-Ferrand, Department of Mathematical Engineering and Modeling
Clermont-Ferrand, France
victor.brunel@etu.uca.fr

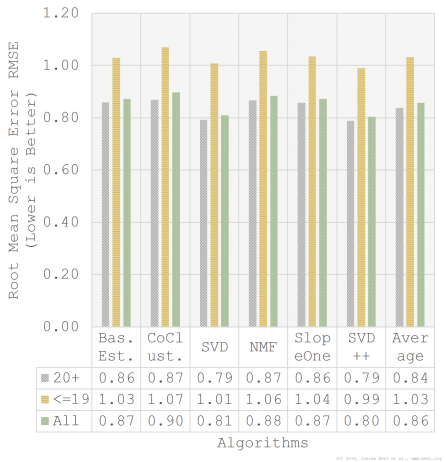
ABSTRACT

Many recommender-system datasets are pruned, i.e. some data is removed that wouldn't be removed in a production recommender-system. For instance, *MovieLens* contains only data from users who rated 20 or more movies.[1] Similarly, some researchers prune data themselves, and conduct experiments only on subsets of the original data, sometimes as little as 0.58%. We conduct a study on data pruning, and find that 48% of researchers used pruned datasets. MovieLens was the most used dataset (40%) and can be considered as a defacto standard dataset. Based on MovieLens, we find that removing users with less than 20 ratings is equivalent to removing 5% of ratings and 42% of users. Ignoring these users may not be ideal as users with less than 20 ratings have an RMSE of 1.03 on average, i.e. 23% worse than users with 20+ ratings (0.84). We discuss the results and conclude that pruning should be avoided, if possible, though more discussion in the community is needed.

KEYWORDS

Recommender Systems, Datasets, Pruning, Data Pruning, Evaluation

ACM RecSys 2019 Late-breaking Results, 16th-20th September 2019, Copenhagen, Denmark



	Bas. Est.	CoClust.	SVD	NMF	SlopeOne	SVD++	Average
20+	0.86	0.87	0.79	0.87	0.86	0.79	0.84
<=19	1.03	1.07	1.01	1.06	1.04	0.99	1.03
All	0.87	0.90	0.81	0.88	0.87	0.80	0.86

Figure 2: RMSE of six collaborative filtering algorithms, and the overall average of all algorithms, for the three data splits. For users with 1-19 ratings, RMSE is 1.03, compared to 0.84 (18% difference) for users with 20 and more ratings. The ranking of the algorithms is the same on all three data splits (SVD++ performs best, SVD second best... Co-Clustering worst).

.

[2]Quote from our email exchange with Joseph A. Konstan, one of the MovieLens founders. Permission to quote from the email was granted.

INTRODUCTION

'Data pruning' is common practice in recommender-systems research. We define data pruning as the removal of instances from a dataset that would not be removed in the real-world, i.e. when used by recommender systems in production environments. Reasons to prune datasets are manifold and include user interests when publishing data (e.g. data privacy) or business interests. 'Data pruning' differs from 'data cleaning' as such that data cleaning (e.g. outlier removal) is typically a prerequisite for the effective training of recommender-system and machine-learning algorithms, whereas data pruning is not affecting the algorithm performance in itself.

A prominent example is MovieLens in most of its variations[1] [3]. MovieLens contains information about how users of *MovieLens.org* rated movies, but the MovieLens team decided to exclude ratings of users who rated less than 20 movies.[1] The reasoning was as follows:[2] "*(1) [researchers] needed enough ratings to evaluate algorithms, since most studies needed a mix of training and test data, and it is always possible to use a subset of data when you want to study low-rating cases; and (2) the movies receiving the first ratings for users during most of MovieLens' history are biased based on whatever algorithm was in place for new-user startup (for most of the site's life, that was a mix of popularity and entropy), hence the MovieLens team didn't want to include users who hadn't gotten past the 'start-up' stage [...].*"

Not only the creators of datasets may prune data, but also individual researchers may do so. For instance, Caragea et al. pruned the CiteSeer corpus for their research [2]. The corpus contains a large number of research articles and their citations. Caragea et al. removed research papers with fewer than ten and more than 100 citations as well as papers citing fewer than 15 and more than 50 research papers. From originally 1.3 million papers in the corpus around 16,000 remained (1.2%). Similarly, Pennock et al. removed many documents so that only 0.58% remained for their research [5].

We criticized the practice of data pruning previously, particularly when only a fraction of the original data remains [1]. We argued that evaluations based on a small fraction of the original data are of little significance. For instance, knowing that an algorithm performs well for 0.58% of users is of little significance if it remains unknown how the algorithm performs for the remaining 99.42%. Also, it is well known that collaborative filtering tends to perform poorly for users with few ratings [4]. Hence, when evaluating collaborative filtering algorithms, we would consider it crucial to not ignore users with few ratings, i.e. those users for whom the algorithms presumably perform poorly.

Our criticism was based more on 'gut feeling' than scientific evidence. To the best of our knowledge, no empirical data exists on how widely data pruning is applied, and how pruning affects recommender-systems evaluations. Also, the recsys community has not discussed if, and to what extent a) datasets should be pruned by their creators and b) whether individual researchers should prune data. We conduct the first steps towards answering these questions. Ultimately, we hope to stimulate a discussion that leads to widely accepted guidelines on pruning data for recommender-systems research.

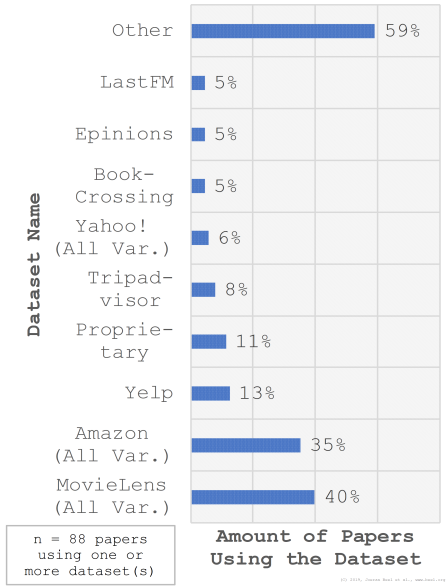


Figure 3: Most popular recommender system datasets being used by the authors of 88 full- and short papers at the ACM Conference on Recommender Systems (RecSys) 2017 & 2018.

.

[3]We might have missed some relevant information on pruning if that information was provided in a section other than the Methodology section.

[4]For users with only 1 rating, 'Surprise' uses a special technique for evaluations. Please also note that using a random sample is not an example of data pruning.

METHODOLOGY

To identify how widespread data pruning is, we analyzed all 112 full- and short papers published at the ACM Conference on Recommender Systems 2017 and 2018. 88 papers (79%) used offline datasets, the remaining 24 papers (21%) conducted e.g. user studies or online evaluations. For the 88 papers, we analyzed, which datasets the authors used, whether the datasets were pruned by the original creators and whether authors conducted pruning. To identify the latter part, we read the Methodology sections of the manuscripts or similarly named sections.[3] This analysis was done by a single person rather quickly. Consequently, the reported numbers should be seen as ballpark figures.

To identify the effect of data pruning on recommender-system evaluations, we run six collaborative filtering algorithms from the *Surprise* library, namely SVD, SVD++, NMF, Slope One, Co-Clustering, and the Baseline Estimator. We use the unpruned 'MovieLens Latest Full' dataset (Sept. 2018), which contains 27 million ratings by 280,000 users including data from users with less than 20 ratings. Due to computing constraints, we use a random sample with 6,962,757 ratings made by 70,807 users.[4]

We run the six algorithms on three sub-sets of the dataset, i.e. for a) the entire unpruned dataset, b) the data that would be included in a 'normal' version of MovieLens (users with 20+ ratings) c) the data that would be 'normally' not included in the MovieLens dataset (users with less than 20 ratings). We compare how algorithms perform on these different sets, and measure the performance of algorithms by Mean Absolute Error (**MAE**) and Root Mean Square Error (**RMSE**). As the two metrics led to almost identical results, we only report RMSE. Our source code and analysis of the manuscripts is available at https://github.com/BeelGroup/recsys-dataset-pruning.

RESULTS

Popularity of (Pruned) Recommender-Systems Datasets

The authors of the 88 papers used a total of 64 unique datasets, whereas we counted different variations of MovieLens, Amazon and Yahoo! as the same dataset. Our analysis empirically confirms what is common wisdom in the recommender-system community already: MovieLens is the de-facto standard dataset in recommender-systems research. 40% of the full- and short papers at RecSys 2017 and 2018 used the MovieLens dataset in at least one of its variations (Figure 3). The second most popular dataset is Amazon, which was used by 35% of all authors. Other popular datasets are shown in Figure 3 and include Yelp (13%), Tripadvisor (8%), Yahoo! (6%), BookCrossing (5%), Epinions (5%), and LastFM (5%). 11% of all researchers used a proprietary dataset, and 2% used a synthetic dataset.

50% of the authors conducted research with a single dataset, 31% used two datasets, and only 2% used six or more datasets (Figure 4). The highest number of datasets being used was 7. On average, researchers used 1.88 datasets. 40% of the authors used a pruned dataset, and 15% pruned data themselves. In total, 48% of all authors conducted research at least partially with pruned data.

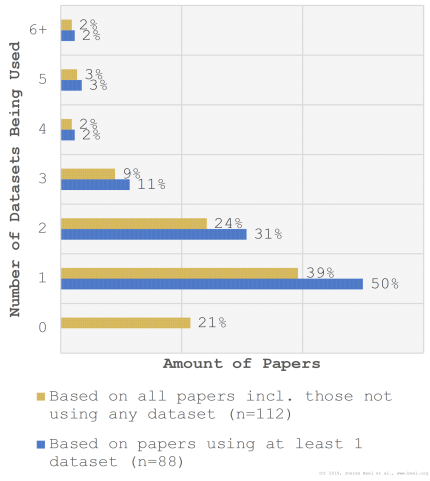


Figure 4: Number of datasets being used by authors of the ACM Conference on Recommender Systems. 50% of authors used just one dataset, 31% used two datasets, and 18% used three or more datasets.

[5]The actual numbers in the pruned MovieLens versions may somewhat differ given that we just used a sample, and the different MovieLens versions differ due to the fact that they include data from different time periods.

[6]It is actually not trivial to decide how to split the data in this case. Follow-up research is needed to confirm the numbers, and investigate different options.

[7]The authors evaluated different algorithms. For users with few ratings (<8), other algorithms performed best than for users with more ratings.

The Effect of Data Pruning

The user-rating distribution in MovieLens follows a long-tail distribution (Figure 1). 42% of the users have less than 20 ratings, and these users contribute 5% of all ratings. The remaining 58% of users, with 20+ ratings, contribute 95% of all ratings. The top 3% of users – those with 500+ ratings – contribute 28% of all ratings in the dataset. Consequently, using a pruned MovieLens variation (100k, 1m, ...) is equivalent to ignoring around 5% of the ratings and 42% of users.[5]

There are notable differences for the three data splits in terms of algorithm performance (Figure 2). Over the entire unpruned data, RMSE of the six algorithms is 0.86 on average, with the best algorithm being SVD++ (0.80), closely followed by SVD (0.81). The worst performing algorithm is Co-Clustering (0.90). For the subset of ratings from users with 20+ ratings – that equals a 'normal' MovieLens dataset – RMSE over all algorithms is 0.84 on average (2.12% lower, i.e. better). In other words, using a pruned version of MovieLens will lead, on average, to a 2.12% better RMSE compared to using the unpruned data. But, to make this clear, the algorithms do not actually perform 2.12% better. The results only appear to be better because data for which the algorithms tend to perform poorly was excluded in the evaluation. The ranking of the algorithms remains the same when comparing the pruned with the unpruned data (SVD++ performs best, followed by SVD, and Co-Clustering performs worst).

We also looked at the users grouped by the number of ratings per user[6]. Figure 5 shows the RMSE for users with 1-9 ratings, 10-19 ratings, ... 500+ ratings. There is a constant improvement (i.e. decrease) in RMSE the more ratings user have. On average, the six algorithms achieve an RMSE of 1.03 for users with less than 20 ratings (1.07 for users with <=9 ratings; 1.02 for users with 10–19 ratings). This contrasts an average RMSE of 0.84 for users with 20+ ratings. In other words, RMSE for users in a pruned MovieLens dataset is 23% better than RMSE for the excluded users. For SVD and SVD++, the best performing algorithms, this effect is even stronger (+27% for SVD; +25% for SVD++).

DISCUSSION & FUTURE WORK

Data Pruning is a widespread phenomenon with 48% of short- and full papers at RecSys 2017/2018 being based at least partially on pruned data. *MovieLens* nicely illustrates an issue that probably applies to many datasets with user-rating data. In the pruned MovieLens datasets, the number of removed ratings is rather small (5%). However, these 5% ratings were made by 42% of the users. For researchers focusing on how well individual ratings can be predicted, the removed data has probably little impact. For researchers who focus on user-specific issues, ignoring 42% of users is probably not ideal, particularly as their RMSE is 23% worse than the RMSE of the users with 20+ ratings.

When discussing data pruning, the probably most important question is whether pruning changes the ranking of the evaluated algorithms. Other research has already shown that the ranking may change, though that research was not conducted in the context of data pruning [4].[7] In our study,

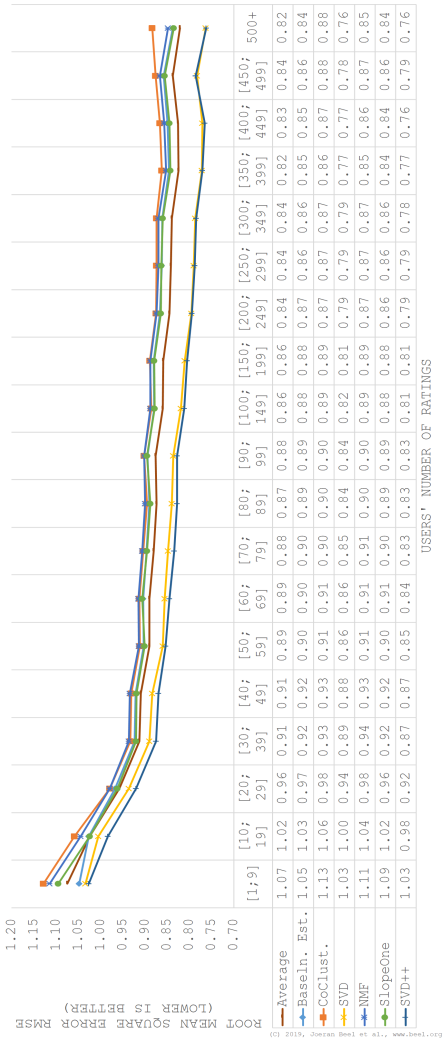


Figure 5: RMSE of the six collaborative filtering algorithms and the overall average, by users' number of ratings. Algorithms applied to users with 1-9 ratings achieve an RMSE of 1.07 on average. Algorithms applied to users with 500+ ratings achieve an RMSE of 0.82 on average (37% better).

the ranking of algorithms did not change. The algorithm that was best (second best...) on the pruned MovieLens data was also best (second best...) on the unpruned data. However, it seems likely to us that rankings may change if more diverse algorithms are compared, e.g. collaborative filtering vs. content-based filtering. Also, the MovieLens dataset is relatively moderately pruned. We consider it likely that heavy pruning, where only a small fraction remains (e.g. Pennock et al. [5]), might lead to a change in the ranking of algorithms. More research with more diverse algorithms, different datasets, and different degrees of pruning is needed to confirm or reject that assumption. A qualitative study could help to identify details on the motivation of dataset creators and researchers to prune data.

Given the current results, we propose that data pruning should be applied with great care, or, if possible, be avoided. We would not generally consider pruning as a malpractice, but certainly not a best-practice either. In some cases, especially when large parts of data are removed, data pruning may become a malpractice, though the community yet has to determine how much removed data is too much. As a starting point, we would recommend the following guidelines, though this is certainly not a definite recommendation, and more discussion in the community is needed:

(1) Publishers of datasets should avoid pruning – if possible. If there are compelling reasons to prune data (e.g. ensuring privacy), these should be clearly communicated in the documentation.
(2) Researchers using pruned datasets should discuss the implications in their manuscript.
(3) Individual researchers should not prune data. If they feel that an algorithm may perform well on a subset of the data, they should report performance for both the entire dataset and the subset. If researchers conduct pruning anyway, they should clearly indicate this in the manuscript, provide reasoning, and discuss the implications.

It may not always be obvious where data cleaning ends, legitimate data pruning begins, and when data pruning becomes a malpractice. The community certainly needs more discussion about this issue. We are confident that with widely agreed guidelines on data pruning, recommender-systems research will become more reproducible, more comparable and more representative of 'the real world'.

REFERENCES

[1] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Research Paper Recommender Systems: A Literature Survey. *International Journal on Digital Libraries* 4 (2016), 305–338. https://doi.org/10.1007/s00799-015-0156-0
[2] Cornelia Caragea, Adrian Silvescu, Prasenjit Mitra, and C Lee Giles. 2013. Can't See the Forest for the Trees? A Citation Recommendation System. In *iConference*. 849–851. https://doi.org/10.9776/13434
[3] F Maxwell Harper and Joseph A Konstan. 2016. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4 (2016), 19.
[4] Daniel Kluver and Joseph A Konstan. 2014. Evaluating recommender behavior for new users. In *Proceedings of the 8th ACM Conference on Recommender Systems*. ACM, 121–128.
[5] David M Pennock, Eric Horvitz, Steve Lawrence, and C Lee Giles. 2000. Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach. In *Sixteenth Conference on Uncertainty in Artificial Intelligence*. 473–480.