Visualizing Ratings in Recommender System Datasets

Diego Monti
Politecnico di Torino
diego.monti@polito.it

Giuseppe Rizzo
LINKS Foundation
giuseppe.rizzo@linksfoundation.com

Maurizio Morisio
Politecnico di Torino
maurizio.morisio@polito.it

ABSTRACT

The numerical outcome of an offline experiment involving different recommender systems should be interpreted also considering the main characteristics of the available rating datasets. However, existing metrics usually exploited for comparing such datasets like sparsity and entropy are not enough informative for reliably understanding all their peculiarities. In this paper, we propose a qualitative approach for visualizing different collections of user ratings in an intuitive and comprehensible way, independently from a specific recommendation algorithm. Thanks to graphical summaries of the training data, it is possible to better understand the behaviour of different recommender systems exploiting a given dataset. Furthermore, we introduce RS-viz, a Web-based tool that implements the described method and that can easily create an interactive 3D scatter plot starting from any collection of user ratings. We compared the results obtained during an offline evaluation campaign with the corresponding visualizations generated from the HetRec LastFM dataset for validating the effectiveness of the proposed approach.

KEYWORDS

Visualization Tool, Rating Dataset, Offline Evaluation

1 INTRODUCTION

Being able to correctly interpreting the results obtained during an offline evaluation of different recommender systems is of paramount importance for understanding the quality of the suggested items [5]. However, this task is particularly difficult as it requires knowing several details regarding the evaluation protocol and the rating dataset exploited for conducting the experiments [9]. For example, sparse datasets usually yield to lower evaluation scores with respect to more dense datasets [3]. On the other hand, datasets with many popular items tend to advantage systems that create less diverse suggestions [10], like the most popular baseline. There are also some subtle differences among rating datasets related to the application domain or the collection protocol that could affect the choice of the most appropriate recommender system.

Different metrics have been proposed in literature to summarize the main characteristic of a rating dataset, i.e. sparsity or entropy. However, we argue that such metrics are not sufficient for comparing datasets in a reliable way, as many other facets should be taken into account. For example, it is not possible to understand the rating behaviors of specific groups of users nor the popularity of the most rated items by only looking at some general statistics computed on the whole dataset.

IntRS '19: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, 19 Sept 2019, Copenhagen, DK.

A possible solution to this problem could be represented by data visualization techniques [7]. However, most of the methods available in literature are designed to display the output of a recommendation model and not the original dataset [1, 6]. In contrast, we argue that it is necessary to visually explore a rating dataset even before it is used to train a recommender system, for understanding how the input data will influence the outputs under analysis.

In this paper, we propose a novel qualitative approach based on data visualization for creating a graphical summary of any collection of user preferences. This method is useful for visually identifying similarities and differences among the available datasets. In fact, we argue that if two datasets result in similar visualizations, the behavior of different recommender systems relying on them will be consistent. Furthermore, we present a Web-based tool, named RS-viz, for easily constructing the proposed visualization and comparing rating datasets in an intuitive way. RS-viz is freely available on GitHub[1] and its usage is described in an introductory video.[2]

Differently from the plotting capabilities already available in specialized software like Matlab or Scilab, our approach is more general, as it can be applied in a consistent way by different users on any dataset and it can be exploited on many devices without the need of installing specific tools.

The remainder of this paper is structured as follows: in Section 2 we review existing visualization techniques in the context of rating datasets, while in Section 3 we present the approach used to construct the scatter plot and we describe the implementation details of the Web-based tool RS-viz. In Section 4, we comment on the outcome of an evaluation campaign designed to validate the proposed method. Finally, in Section 5, we provide the conclusions and we outline possible future works.

2 RELATED WORK

Different authors have proposed to create interactive visualizations for qualitative evaluating the goodness of the recommended items or helping the users to identify the most relevant suggestions.

For example, Kunkel et al. [7] created a 3D map-based visualization that represents the preferences of a user on the entire space of items. The user can inspect the profile created by the recommender and also manually modify it, if necessary.

Çoba et al. [2] extended the *rrecsys* library by adding to it graphical capabilities for performing an offline visual evaluation of different recommendation approaches with respect to the popularity of the suggested items.

Gil et al. [6] introduced VisualRS, a tool capable of creating tree graph structures for exploring the most important relationships between items or users. The graph-based visualization is useful for comparing the results of different recommendation approaches and selecting the most appropriate one for a given task.

[1] https://github.com/D2KLab/rs-viz
[2] https://doi.org/10.6084/m9.figshare.8197706

In contrast, Cardoso et al. [1] proposed to combine the output of different recommender systems with human-generated data to allow users to explore the suggested items in an effective way. This method could also be exploited to compare the results of different recommender systems in a qualitative way.

All the reviewed approaches are based on popular recommender systems. To the best of our knowledge, this paper represents the first formal attempt to visualize the ratings available in an offline dataset independently from a specific recommender system.

3 VISUALIZATION APPROACH

In this section, we first describe the algorithm that we devised for creating a scatter plot that represents a rating dataset (Section 3.1), then we introduce the implementation details of RS-viz (Section 3.2).

3.1 Scatter plot construction

In order to visually represent the rating matrix associated with a generic dataset we opted for a 3D scatter plot. The rationale behind this choice is that each point in the visualization could intuitively represent a single rating from the dataset: the value of the x-axis is the identifier of the user, the value of the y-axis is the identifier of the item, while the value of the z-axis is the rating itself, if it is expressed on a numerical scale.

However, it is easy to foresee that this approach cannot handle complex datasets with many preferences, as it requires one point for each rating. If the ratings available are only binary, a traditional scatter plot would suffice.

For these reasons, we decided to create a more compact representation of the rating matrix before visualizing it. In details, we first associated the users and the items with internal numerical identifiers according to their frequency of appearance in the dataset. For example, we associated the most rated item with the value of 1, and the second most rated item with the value of 2. The same approach was followed for ordering the identifiers of the users according to the number of ratings that they expressed.

Then, we linearly normalized such identifiers within an interval ranging from 0 to a user provided value, which represents the size of a squared rating matrix in a transformed space. Finally, we binarized the ratings from the original dataset according to a user provided threshold and we counted, for each cell of the transformed matrix, the number of positive preferences associated with that cell.

For example, if the user 40 expressed a preference for the item 360 in a dataset where the number of users is 941, the number of items is 1446, and the number of normalized users and items is equal to 100, that rating would be associated with the cell $(4, 24)$ because $\lfloor 40 \div 941 \times 100 \rfloor = 4$ and $\lfloor 360 \div 1446 \times 100 \rfloor = 24$.

Therefore, the value of the z-axis represents the number of positive ratings associated with a sub-matrix of the original dataset, sorted by item popularity and user activity. In order to enhance the readability of the visualization, we also represented the value of the z-axis using a logarithmic color scale.

As an example of the proposed method, we report in Figure 1 and Figure 2 the scatter plots obtained from the MovieLens 100K and MovieLens 1M datasets, when the rating threshold is equal to 3, and the number of normalized users and items is equal to 100.

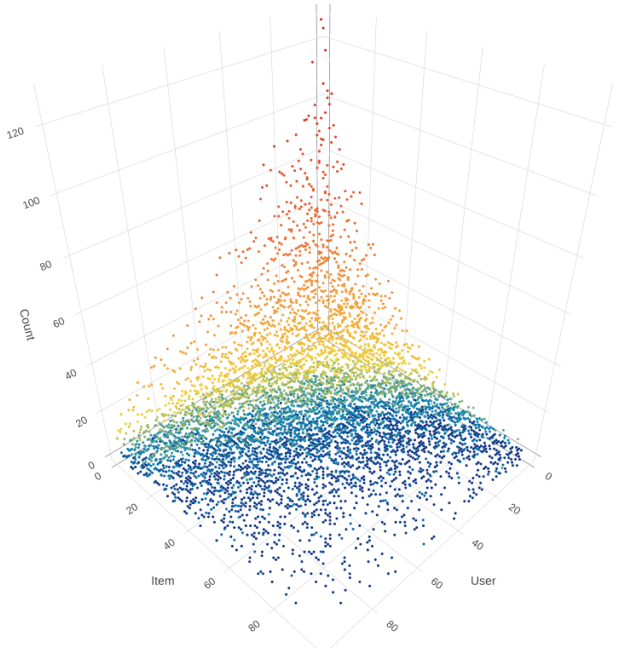


Figure 1: The MovieLens 100K dataset.

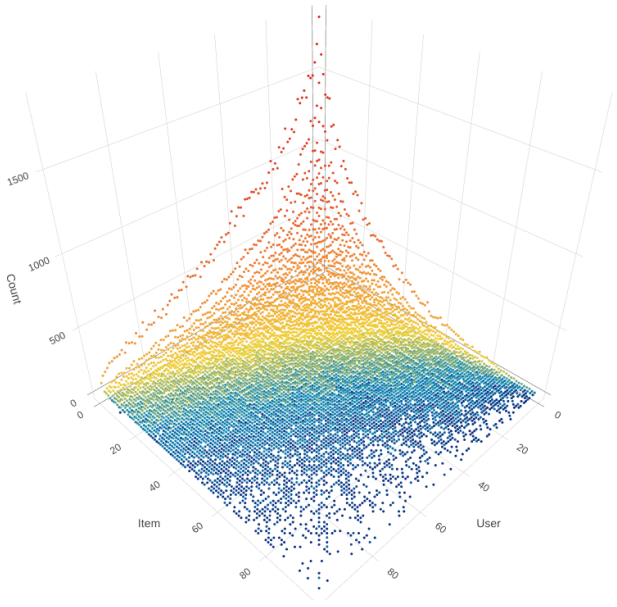


Figure 2: The MovieLens 1M dataset.

By looking at the values of the z-axis, it is possible to observe in an intuitive way that MovieLens 1M contains a higher number of popular items and of very active users. This conclusion is consistent with the findings of other works that analyze the main characteristics of the MovieLens datasets [3].

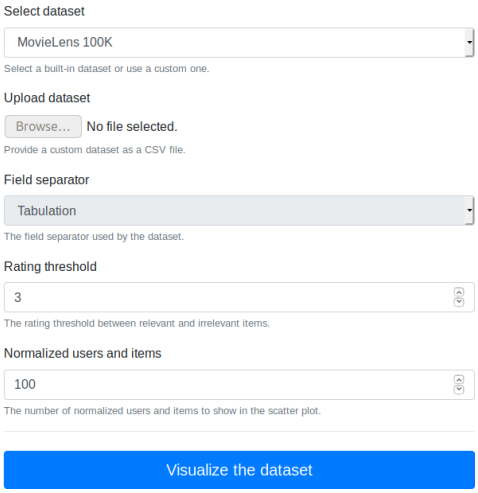


Figure 3: The configuration parameters of RS-viz.

3.2 Software implementation

We realized a software implementation of the proposed approach as a Web-based tool, called RS-viz, which is freely available. Our visualization framework has been developed using the JavaScript programming language and it runs entirely in a user's browser. For this reason, it can also be exploited for analyzing private datasets, as no information about them is sent to remote servers.

The user needs to visit the Web-page of RS-viz[3] and select one of the built-in datasets or provide her own dataset as a CSV file. Then, she needs to specify the threshold between positive and negative ratings and the number of normalized users and items, which should be selected also considering the rating scale of the input dataset and the desired visualization density. A screenshot of the form containing the configuration parameters of RS-viz is reported in Figure 3.

After a few seconds, an interactive 3D scatter plot is constructed on the right side of the page. The user can inspect the plot by rotating the camera and finally save the result as a PNG file.

4 EVALUATION CAMPAIGN

In the following, we report the numerical outcomes of an evaluation campaign conducted on the HetRec LastFM dataset using different recommendation approaches with the purpose of understanding if our visualization technique is capable of capturing the different characteristics of a rating dataset and to what extend they influence the recommendation coverage and accuracy.

4.1 Experimental setup

We performed two different experiments with the HetRec LastFM dataset and our evaluation framework RecLab [8].[4]

In the first one, we set the rating threshold equal to 0, while in the second one, we set it equal to 1,000. For the other parameters, we used the default values of the framework: we selected a random splitting protocol, the test set size as the 20% of the dataset, and the length k of the recommended lists equal to 10.

We considered different recommendation approaches, namely the most popular and random baselines and the MyMediaLite [4] implementations of the Item KNN, User KNN, BPRMF, and WRMF recommender systems using their default settings.

We computed the metrics of coverage, precision, recall, and NDCG. The results of these experiments are reported in Table 1. The same datasets obtained from HetRec LastFM by varying the rating threshold were exploited for creating two scatter plots using RS-viz, as displayed in Figure 4.

4.2 Discussion

From the visualization provided in Figure 4a, we can observe that the HetRec LastFM dataset has a very different structure from the one of the MovieLens datasets. In fact, a limited number of items are associated with the preferences of almost all users, as it can be deduced by considering only the ratings expressed for popular items, that is the ones with low identifiers. Please note that such ratings seem not related to the identifier of the user, resulting in a scatter plot that resembles the shape of a half cylinder.

Furthermore, less popular items seem to be liked by less active users. This behavior can be observed by looking at the lower part of Figure 4a. Users with a high identifier have rated a more widespread set of items, while users with a low identifier have rated popular items more frequently.

These differences can be easily explained if we consider the collection protocol and the domain of the dataset under analysis. The ratings in the LastFM datasets represent the number of times a user listened to a particular artist: therefore, they were collected in an implicit way and their values range from one to tens of thousands.

Also the strange area in the plot with almost no preferences is a direct result of the collection protocol, which relied on the LastFM website to obtain the top artists for a set of users. In fact, the list of artists available in the dataset is limited to 50 items for each user.

If we increase the value of the rating threshold, we can observe that the resulting scatter plot represented in Figure 4b is more similar to the ones of the MovieLens datasets, resulting in a very typical long tail distribution with respect to both the items and the users. This outcome is due to the fact that we removed ratings produced by more casual listeners.

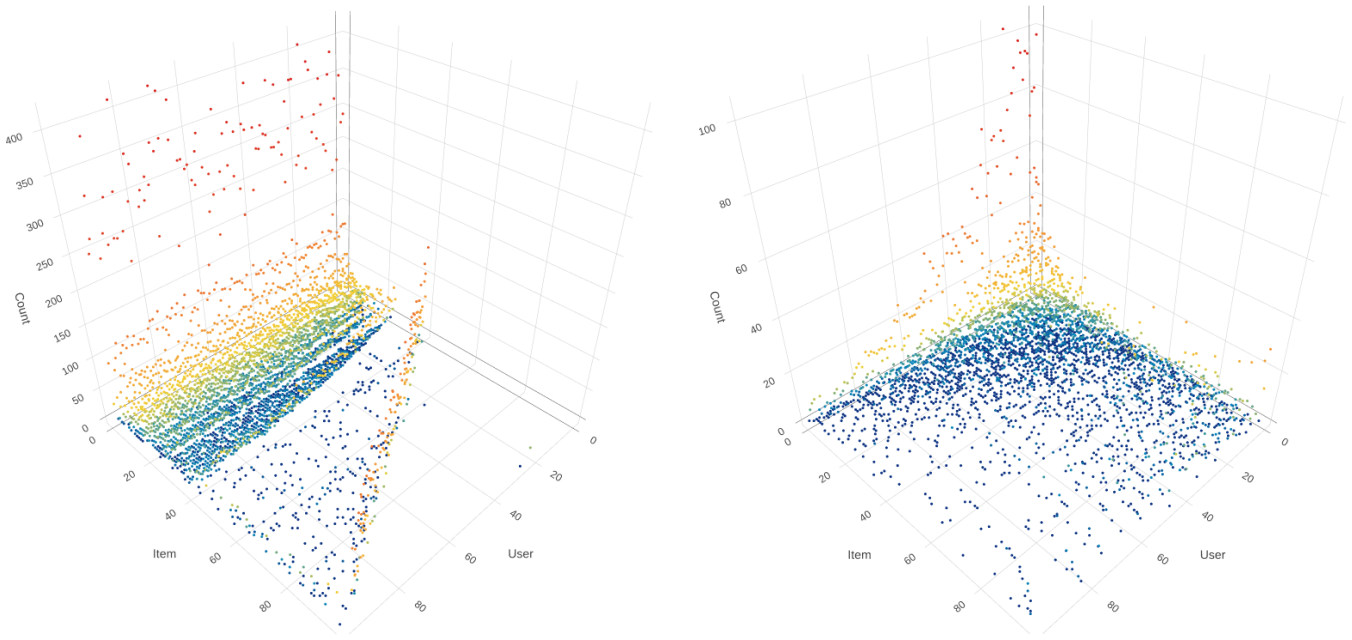
From the numerical outcomes of the experiments, we can deduce that the User KNN and WRMF algorithms are the most appropriate ones with both the different rating thresholds. In general, all the recommenders available perform worse with an higher threshold. In fact, from the visualizations it is clear that the number of available preferences is much lower with respect to the MovieLens 100K dataset, as the scatter plot represented in Figure 4b is sparser than the one available in Figure 1. Because user preferences are more limited in number and fragmented, the task of any recommender system is necessarily harder.

Interestingly, the Item KNN, differently from the User KNN, experienced a dramatic drop in all the metrics considered. This result may have been caused by the fact that a very low number of users is available for each item of the dataset. Also this characteristic can be observed from the generated scatter plot by looking at the

[3] http://datascience.ismb.it/rs-viz/
[4] http://datascience.ismb.it/reclab/

Table 1: The numerical results of the experimental comparison using the HetRec LastFM dataset.

	Rating threshold = 0				Rating threshold = 1,000			
Algorithm	Coverage	Precision	Recall	NDCG	Coverage	Precision	Recall	NDCG
Random	0.706679	0.000798	0.000745	0.000858	0.705562	0.000107	0.000622	0.000133
Most Popular	0.001692	0.071170	0.071480	0.079673	0.001684	0.022122	0.090233	0.027437
Item KNN	0.235321	0.129362	0.131967	0.145258	0.107233	0.002878	0.013012	0.002686
User KNN	0.030074	0.157234	0.160353	0.193121	0.049343	0.040672	0.160767	0.055013
BPRMF	0.022979	0.081277	0.082248	0.094737	0.003756	0.021695	0.088211	0.024366
WRMF	0.015558	0.159947	0.162332	0.195107	0.012886	0.039606	0.157484	0.053148



(a) Rating threshold = 0 **(b) Rating threshold = 1,000**

Figure 4: The 3D scatter plots obtained using the HetRec LastFM dataset with different rating thresholds.

lower part of Figure 4b. The white horizontal stripes denote groups of items that have been rated by only a few very active users.

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed a method for creating graphical summaries of any rating dataset for the purpose of enabling researchers and practitioners to better interpret the results of an offline evaluation campaign. Furthermore, we introduced RS-viz, a Web-based tool capable of creating an interactive 3D scatter plot according to the aforementioned approach starting from a user provided CSV dataset or a built-in collection of ratings.

We validated the capabilities of such visualizations to reveal useful information by comparing the graphical representations of the HetRec LastFM dataset constructed with different rating thresholds with the numerical outcomes of two offline experiments involving various recommendation techniques.

As future work, we would like to quantitatively characterize rating datasets according to different dimensions and place them in various categories, for example by analyzing the diversity of user preferences or the tendency to rate popular items only. This empirical categorization would enable the users of our tool to better understand the ratings available and to select the most appropriate recommendation approach according to such proprieties.

Furthermore, we would like to improve RS-viz by developing other visualization methods to enable more comprehensive analysis and to evaluate its effectiveness by checking if researchers and practitioners are able to correctly use it to explain the performance of different recommender systems on a particular dataset.

Finally, additional studies are needed to better understand how the proposed approach could be extended for also visualizing non-conventional datasets, for example the ones enhanced with context-aware information like spatial and temporal data.

REFERENCES

[1] Bruno Cardoso, Gayane Sedrakyan, Francisco Gutiérrez, Denis Parra, Peter Brusilovsky, and Katrien Verbert. 2019. IntersectionExplorer, a multi-perspective approach for exploring recommendations. *International Journal of Human-Computer Studies* 121 (2019), 73–92. https://doi.org/10.1016/j.ijhcs.2018.04.008

[2] Ludovik Çoba, Panagiotis Symeonidis, and Markus Zanker. 2017. Visual Analysis of Recommendation Performance. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*. ACM, New York, NY, USA, 362–363. https://doi.org/10.1145/3109859.3109982

[3] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of Recommender Algorithms on Top-n Recommendation Tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys '10)*. ACM, New York, NY, USA, 39–46. https://doi.org/10.1145/1864708.1864721

[4] Zeno Gantner, Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2011. MyMediaLite: A Free Recommender System Library. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11)*. ACM, New York, NY, USA, 305–308. https://doi.org/10.1145/2043932.2043989

[5] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys '10)*. ACM, New York, NY, USA, 257–260. https://doi.org/10.1145/1864708.1864761

[6] Stephanie Gil, Jesús Bobadilla, Fernando Ortega, and Bo O. Zhu. 2018. VisualRS : Java framework for visualization of recommender systems information. *Knowledge-Based Systems* 155 (2018), 66–70. https://doi.org/10.1016/j.knosys.2018.04.028

[7] Johannes Kunkel, Benedikt Loepp, and Jürgen Ziegler. 2017. A 3D Item Space Visualization for Presenting and Manipulating User Preferences in Collaborative Filtering. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17)*. ACM, New York, NY, USA, 3–15. https://doi.org/10.1145/3025171.3025189

[8] Diego Monti, Giuseppe Rizzo, and Maurizio Morisio. 2018. A Distributed and Accountable Approach to Offline Recommender Systems Evaluation. In *Proceedings of the Workshop on Offline Evaluation for Recommender Systems at the 12th ACM Conference on Recommender Systems*. REVEAL 2018, Vancouver, BC, Canada, Article 6, 5 pages. https://arxiv.org/abs/1810.04957

[9] Alan Said and Alejandro Bellogín. 2014. Comparative Recommender System Evaluation: Benchmarking Recommendation Frameworks. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 129–136. https://doi.org/10.1145/2645710.2645746

[10] Saúl Vargas and Pablo Castells. 2011. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. ACM, New York, NY, USA, 109–116. https://doi.org/10.1145/2043932.2043955