International Journal of Mathematical, Engineering, Biological and Applied Computing, 2022, 1, 49-51
www.scipublications.org/journal/index.php/ijmebac
DOI: 10.31586/ijmebac.2022.350

Mini Review

Open-Source Datasets for Recommender Systems Analysis

Raja Marappan *

School of Computing, SASTRA Deemed University, Thanjavur, India

*Correspondence: Raja Marappan (professor.m.raja@gmail.com)

Abstract: There are different traditional and nontraditional datasets available to investigate the performance of recommender systems. This article focuses on the different datasets required for the investigation of recommender systems.

Keywords: Recommender Systems; Traditional Datasets; Nontraditional Datasets; Systems Analysis

1. Introduction

The following terms are used to define the recommendation systems: items, user, and ratings as sketched in Table 1 [1-5].

2. Datasets

This section explores the different datasets required to investigate the recommendation systems. The specification and availability of different datasets are sketched in Table 2 [6-9].

The comparison of datasets using different metrics – users, items, ratings, density, and rating scale is sketched in Table 3 [10-15].

3. Conclusions & Future Work

This article explained the datasets required for the investigation of recommender systems. These datasets are also compared using the metrics such as users, items, ratings, density, and rating scale. The recommender systems can be developed using several soft computing models in the future [16-20].

How to cite this paper:
Marappan, R. (2022). Open-Source Datasets for Recommender Systems Analysis. *International Journal of Mathematical, Engineering, Biological and Applied Computing*, *1*(2), 49–51. Retrieved from https://www.scipublications.com/journal/index.php/ijmebac/article/view/350

Received: May 7, 2022
Accepted: June 24, 2022
Published: June 26, 2022

Table 1. Terms for recommendation systems

Term	Definition
Item	This defines what is to be recommended. For example, this refers – information, movies, products etc.
User	The user rates products or items and receives new items recommendations.
Rating	The user's choice or preference is defined as a rating. For example, rating defines – dislike or like, 1 star to 5 stars, integer or floating representation, etc.

Table 2. Datasets – specification and availability.

Dataset	Specification	Availability
MovieLens	Collection of movie ratings with 27000 movies & 140000 users	https://grouplens.org/datasets/movielens/
Jester	The joke rating system consists of 6 million ratings of 150 jokes	http://eigentaste.berkeley.edu/
Book-Crossings	Book rating dataset with 1.1 M ratings (90000 users & 270000 books)	http://www2.informatik.uni-freiburg.de/~cziegler/BX/
Last.fm	Music recommendations dataset with aggregated data.	https://grouplens.org/datasets/hetrec-2011/
Wikipedia	Collaborative encyclopedia used for general applications.	https://en.wikipedia.org/wiki/Wikipedia:Database_download#English-language_Wikipedia
OpenStreetMap	Collaborative project for maps.	https://planet.openstreetmap.org/planet/full-history/
Python Git Repositories	Git repositories Python code.	https://github.com/python
MovieLens 25M	Collection of movie ratings with 62423 movies & 25000095 ratings.	https://grouplens.org/datasets/movielens/25m/
Social Network Influencer	Learning task preferences.	https://www.kaggle.com/c/predict-who-is-more-influential-in-a-social-network/data
Million Song	Audio features for music tracks.	http://millionsongdataset.com/
Free Music Archive	Music analysis audio downloads.	https://github.com/mdeff/fma
Netflix Prize	Applied in the competition of Netflix Prize.	https://academictorrents.com/details/9b13183dc4d60676b773c9e2cd6de5e5542cee9a
Amazon Review	The reviews collection.	https://nijianmo.github.io/amazon/index.html
Yahoo! Music User Ratings	Musical artists collection preferences.	https://webscope.sandbox.yahoo.com/catalog.php?datatype=r&guce_referrer=aHR0cHM6Ly9naXRod-WIuY29tL2Nhc2VyZWMvRGF0YXNld-HMtZm9yLVJlY29tbWVuZGVyLVN5c3RlbXM&guce_referrer_sig=AQAAAAkyH74jyiIv4JxPjvejltL1_Sk-yDNt-NAbIpHn2YfUnG1v-2mxj_XOD-qtpvdqg-aoNtTk9pzWVkYzz3ZbvN5C2_RrjVAowWPR7lmx-Gid-aMerX8qOzosJayRViVuW2IEoTjMAeZ8xJlIoK38-6GQA-JOwZjFsSv0AyQNj4oagqX&guccounter=2
Steam Video Games	Collection of the behaviors of users.	https://www.kaggle.com/datasets/tamber/steam-video-games

Table 3. Performance comparison of datasets using different metrics

Dataset	Items	Users	Density	Ratings	Rating Scale
Book-Crossing	271379	92107	0.0041%	1031175	[1, 10], and implicit
Wikipedia	4936761	5583724	0.0015%	417996366	Interactions
Git	1757	790	0.95%	13165	Interactions
Jester	150	124113	31.50%	5865235	[-10, 10]
Last.fm	17632	1892	0.28%	92834	Play Counts
OpenStreetMap	108330	231	0.82%	205774	Interactions
Movielens 1M	3883	6040	4.26%	1000209	[1-5]
Movielens 10M	10681	69878	1.33%	10000054	[0.5-5]
Movielens 20M	27278	138493	0.52%	20000263	[0.5-5]

References

[1] G. Adomavicius, A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions IEEE Trans. Knowl. Data Eng. (2005), 10.1109/TKDE.2005.99

[2] J. Chen, X. Wang, S. Zhao, F. Qian, Y. Zhang. Deep attention user-based collaborative filtering for recommendation Neurocomputing, 383 (2020), 10.1016/j.neucom.2019.09.050

[3] A. Da'u, N. Salim, I. Rabiu, A. Osman. Recommendation system exploiting aspect-based opinion mining with deep learning method. Inf. Sci., 512 (2020), 10.1016/j.ins.2019.10.038

[4] Lu J., Zhang Q., Zhang G. Recommender Systems: Advanced Developments World Scientific (2020)

[5] Zhang S., Yao L., Sun A., Tay Y. Deep learning based recommender system: A survey and new perspectives ACM Comput. Surv. (2019)

[6] Zhongying Zhao, Xuejian Zhang, Hui Zhou, Chao Li, Maoguo Gong, Yongqing Wang Hetnerec: heterogeneous network embedding based recommendation Knowl. Base Syst., 204 (2020), Article 106218

[7] Liao W., Zhang Q., Yuan B., Zhang G., Lu J. Heterogeneous multidomain recommender system through adversarial learning IEEE Trans. Neural Netw. Learn. Syst. (2022)

[8] Zhang Q., Liao W., Zhang G., Yuan B., Lu J. A deep dual adversarial network for cross-domain recommendation IEEE Trans. Knowl. Data Eng. (2021)

[9] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, Qing He A Survey on Knowledge Graph-Based Recommender Systems IEEE Transactions on Knowledge and Data Engineering (2020)

[10] Zhang Y., Chen X. Explainable recommendation: A survey and new perspectives (2020) arXiv preprint arXiv:1804.11192

[11] Bhaskaran, S.; Marappan, R.; Santhi, B. Design and Comparative Analysis of New Personalized Recommender Algorithms with Specific Features for Large Scale Datasets. Mathematics 2020, 8, 1106. https://doi.org/10.3390/math8071106

[12] Bhaskaran, S.; Marappan, R.; Santhi, B. Design and Analysis of a Cluster-Based Intelligent Hybrid Recommendation System for E-Learning Applications. Mathematics 2021, 9, 197. https://doi.org/10.3390/math9020197

[13] Marappan, R. (2022). Classification and Analysis of Recommender Systems. International Journal of Mathematical, Engineering, Biological and Applied Computing, 1(1), 17–21. DOI: 10.31586/ijmebac.2022.331

[14] Marappan, R., & Bhaskaran, S. (2022). Movie Recommendation System Modeling Using Machine Learning. International Journal of Mathematical, Engineering, Biological and Applied Computing 2022, 1(1), 12-16. DOI: 10.31586/ijmebac.2022.291

[15] Marappan, R., & Bhaskaran, S. (2022). Analysis of Network Modeling for Real-world Recommender Systems. International Journal of Mathematical, Engineering, Biological and Applied Computing, 1(1), 1–7. DOI: 10.31586/ijmebac.2022.283

[16] Marappan, R.; Sethumadhavan, G. Complexity Analysis and Stochastic Convergence of Some Well-known Evolutionary Operators for Solving Graph Coloring Problem. Mathematics 2020, 8, 303. https://doi.org/10.3390/math8030303

[17] Marappan, R., Sethumadhavan, G. Solution to Graph Coloring Using Genetic and Tabu Search Procedures. Arab J Sci Eng 43, 525–542 (2018). https://doi.org/10.1007/s13369-017-2686-9

[18] Raja Marappan: A New Multi-Objective Optimization in Solving Graph Coloring and Wireless Networks Channels Allocation Problems. Int. J. Advanced Networking and Applications Volume: 13 Issue: 02 Pages: 4891-4895 (2021)

[19] R. Marappan and G. Sethumadhavan, "A New Genetic Algorithm for Graph Coloring," 2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation, 2013, pp. 49-54, doi: 10.1109/CIMSim.2013.17.

[20] Raja Marappan, S. Bhaskaran. (2022). Analysis of Recent Trends in E-Learning Personalization Techniques. The Educational Review, USA, 6(5), 167-170. DOI: http://dx.doi.org/10.26855/er.2022.05.003