01

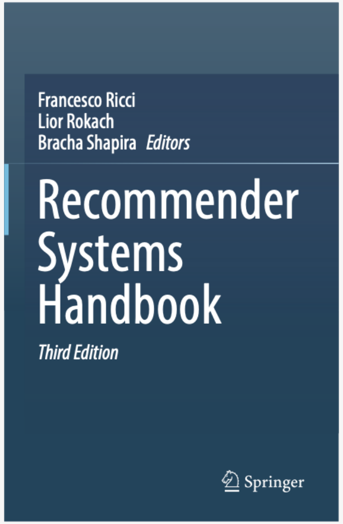
Introduction

A recommender system is typically designed and trained under the assumption that it can learn ***user preference patterns*** *and predict what users are likely to consume in the near future.*

To validate this assumption, an appropriate ***evaluation procedure*** *is required.*

The **RecSys Handbook** [7] distinguishes three experimental settings:

OFFLINE EXPERIMENTS
USER STUDIES
ONLINE EXPERIMENTS



*Offline experiments are the **most accessible** as they do not require to deal with real-time feedback.*

[7] Ricci, Rokach, Shapira, Recommender Systems Handbook, 2022

Offline evaluation *represents **the most used approach** in scientific research.*

Exploring the Landscape of Recommender Systems Evaluation: Practices and Perspectives

CHRISTINE BAUER, Paris Lodron University Salzburg, Austria
EVA ZANGERLE, University of Innsbruck, Austria
ALAN SAID, University of Gothenburg, Sweden

Survey [8]: adoption rates

72% Use offline evaluation

60% Use **only** offline evaluation

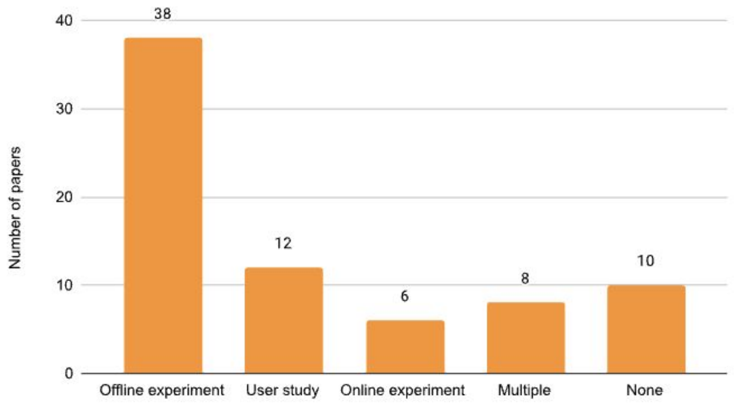
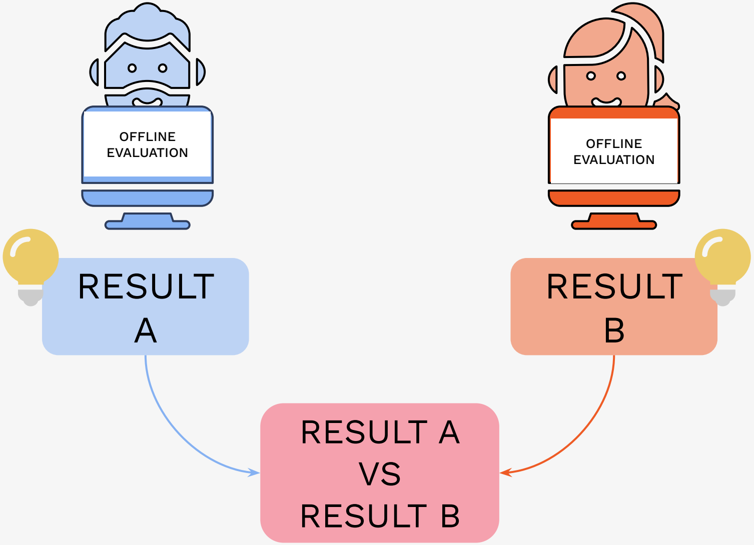


Fig. 5. Number of papers per experiment type.

[8] Bauer et al., TORS 2024, “Exploring the Landscape of Recommender Systems Evaluation: Practices and Perspectives”

Offline evaluation involves training and testing a recommender system on historical user feedback within a defined time period.

Offline evaluation also enhances comparability: different studies sharing **the same experimental setting** can be more fairly compared, thus accelerating scientific progress.



Publicly available datasets are a cornerstone of recommender systems research.

MovieLens [9] released for the first time in 1998 is the most used.

grouplens

The Netflix challenge [10] accelerated research progress with a 100M interactions dataset.

Amazon Reviews [11] enables large-scale multimodal RS.

January 2025: Yandex releases Yambda-5B [12].

[9] https://movielens.org/
[10] Bell & Koren, SIGKDD Exp. 2007, Lessons from the Netflix …
[11] Hou et al., CoRR 2024, Bridging Language and Items for Retrieval …
[12] Ploshkin et al., CoRR 2025, Yambda-5B – A Large-Scale Multi-modal …

This is how the Movielens dataset is publicly distributed.

A ZIP FILE

A CHECKSUM

recommended for new research

MovieLens 32M

MovieLens 32M movie ratings. Stable benchmark dataset. 32 million ratings and two million tag applications applied to 87,585 movies by 200,948 users. Collected 10/2023 Released 05/2024

- README.txt
- ml-32m.zip (size: 239 MB, checksum)

Permalink: https://grouplens.org/datasets/movielens/32m/

How the authors refer to it in their papers?

d the other baselines on three well-known
s, namely *MovieLens 1M*[1], *Yahoo! Movies*

[1] MovieLens 1M Dataset: https://grouplens.org/datasets/movielens/1m/

Reference to the paper or to the public source.

Name
..
amazon.rar
ml10m.part01.rar
ml10m.part02.rar
ml10m.part03.rar
ml10m.part04.rar
ml10m.part05.rar

No reference or link. Only a copy of the dataset.

accuracy of DHCRS with the state-of-the-art methods in terms of HR and NDCG. The experiments are conducted on four datasets: Movielens (ML) 1M, 10M, and 20M, and Netflix. We evaluate the results on each user and then report the aver-

No link, no reference, no dataset copy.

The Gowalla Dataset

The Gowalla Dataset was published by the SNAP research group from Stanford in 2011 [13].

Gowalla is a location-based social network from which **6+ million check-ins** have been collected between Feb. 2009 and Oct. 2010.

It consists of two datasets: a friendship network and a collection of users' check-ins.

GOWALLA

[13] Cho et al., KDD 2011, Friendship and mobility: user movement in location-based social networks

Its popularity stems partly from the LGCN paper [14].

LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation

Xiangnan He
University of Science and Technology of China
xiangnanhe@gmail.com

Kuan Deng
University of Science and Technology of China
dengkuan@mail.ustc.edu.cn

Xiang Wang
National University of Singapore
xiangwang@u.nus.edu

Yan Li
Beijing Kuaishou Technology Co., Ltd.
liyan@kuaishou.com

Yongdong Zhang
University of Science and Technology of China
zhyd73@ustc.edu.cn

Meng Wang*
Hefei University of Technology
eric.mengwang@gmail.com

LGCN - SIGIR 2020 (5k+ cit)

Let’s suppose you want to reproduce that experimental setting [15].

Neural Graph Collaborative Filtering

Xiang Wang
National University of Singapore
xiangwang@u.nus.edu

Xiangnan He*
University of Science and Technology of China
xiangnanhe@gmail.com

Meng Wang
Hefei University of Technology
eric.mengwang@gmail.com

Fuli Feng
National University of Singapore
fulifeng93@gmail.com

Tat-Seng Chua
National University of Singapore
dcscts@nus.edu.sg

NGCF - SIGIR 2019 (4k+ cit)

To reduce the experiment workload and keep the comparison fair, we closely follow the settings of the NGCF work [39]. We request

[14] He et al., SIGIR 2020, LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation
[15] Wang et al., SIGIR 2019, Neural Graph Collaborative Filtering

LGCN - SIGIR 2020 (5k+ cit)

NGCF - SIGIR 2019 (4k+ cit)

Gowalla: This is the check-in dataset [21] obtained from Gowalla, where users share their locations by checking-in. To ensure the quality of the dataset, we use the 10-core setting [10], *i.e.*, retaining users and items with at least ten interactions.

ExpoMF [16] - WWW 2016 (450+ cit)

scientific articles data from arXiv[5]; 3) user bookmarks from Mendeley[6]; and 4) check-in data from the Gowalla dataset [4]. In more details:

Gowalla: contains user-venue checkins from a location-based social network. We pre-process the data such that all users and venues have a minimum of 20 check-ins. Furthermore, this dataset also contains locations

Friendship and mobility - KDD 2011

Reconstructing the origin of a dataset can be complicated without good practices.

It's especially true when papers adopt different filtering strategies.

[16] Liang et al., WWW 2016 — Modeling User Exposure in Recommendation

Gowalla - KDD 2011

Data Processing

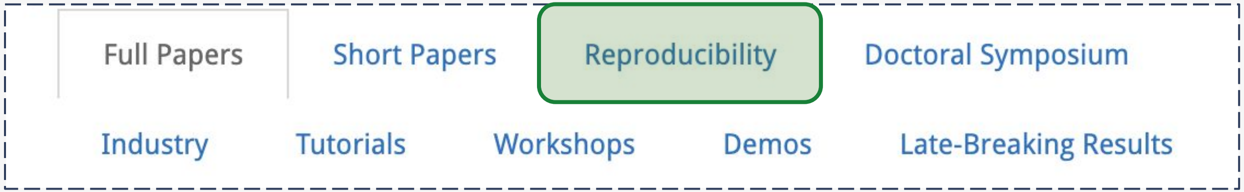
ExpoMF [16] - WWW 2016	20-core user and item
NGCF - SIGIR 2019	10-core user and item
LGCN - SIGIR 2020	The same as NGCF

Reproduce our test

Paper	#Users	#Items	#Ratings	Min. Rat. User
Gowalla	107,092	1,280,969	6,442,892	1
ExpoMF	57,629	47,198	2,3 M	20 (declared)
ExpoMF (Ours)	57,629	47,198	2,318,616	**1**
NGCF / NGCF	29,858	40,981	1,027,370	10 (declared)
NGCF (Ours)	**34,796**	**57,403**	**1,192,807**	**1**

- K-core order matters
- Not reproducible data processing

The reproducibility of scientific results in recommendation systems is a well-known problem [17, 18].



Data processing reproducibility is part of the problem.

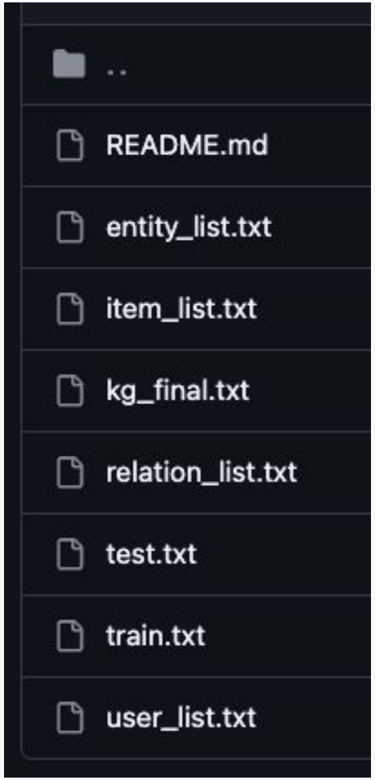
The standardization of data processing is the first step needed to promote reproducibility.

[17] Ferrari Dacrema et al., TOIS 2021, A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research
[18] Ferrari Dacrema et al., RecSys 2019, Are we really making much progress? A worrying analysis of recent neural recommendation ...

The need for standard procedures for data processing is particularly true when the recommendation pipelines deal with side information.

Multimodal, knowledge-aware, content-based, context-aware and cross-domain recommendation are fields where side information plays a key role.

Despite its crucial role, the importance of reliance on robust and reproducible procedures is under-discussed.



Knowledge Graph Datasets for Recommendation

Vincenzo Paparella[1,*], Alberto Carlo Maria Mancino[1,*], Antonio Ferrara[1], Claudio Pomo[1], Vito Walter Anelli[1] and Tommaso Di Noia[1]

[1]*Politecnico di Bari, Bari, Italy*

See the Movie, Hear the Song, Read the Book: Extending MovieLens-1M, Last.fm-2K, and DBbook with Multimodal Data

Giuseppe Spillo
University of Bari Aldo Moro
Bari, Italy
giuseppe.spillo@uniba.it

Elio Musacchio
University of Pisa
Pisa, Italy
elio.musacchio@phd.unipi.it

Cataldo Musto
University of Bari Aldo Moro
Bari, Italy
cataldo.musto@uniba.it

Marco de Gemmis
University of Bari Aldo Moro
Bari, Italy
marco.degemmis@uniba.it

Pasquale Lops
University of Bari Aldo Moro
Bari, Italy
pasquale.lops@uniba.it

Giovanni Semeraro
University of Bari Aldo Moro
Bari, Italy
giovanni.semeraro@uniba.it

[32] Paparella et al., KaRS@RecSys '23, Knowledge Graph Datasets for Recommendation
[33] Spillo et al., RecSys '25, See the Movie, Hear the Song, Read the Book: Extending MovieLens-1M, Last.fm-2K, and DBbook ...

Per-category files

Below are files for individual product categories, which have already had duplicate item reviews removed.

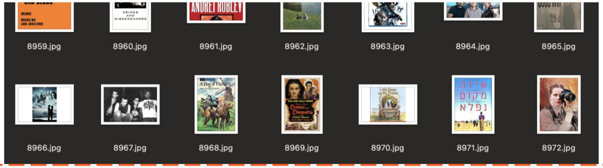
Books	reviews (22,507,155 reviews)	metadata (2,370,585 products)	image features
Electronics	reviews (7,824,482 reviews)	metadata (498,196 products)	image features
Movies and TV	reviews (4,607,047 reviews)	metadata (208,321 products)	image features
CDs and Vinyl	reviews (3,749,004 reviews)	metadata (492,799 products)	image features
Clothing, Shoes and Jewelry	reviews (5,748,920 reviews)	metadata (1,503,384 products)	image features
Home and Kitchen	reviews (4,253,926 reviews)	metadata (436,988 products)	image features
Kindle Store	reviews (3,205,467 reviews)	metadata (434,702 products)	image features
Sports and Outdoors	reviews (3,268,695 reviews)	metadata (532,197 products)	image features
Cell Phones and Accessories	reviews (3,447,249 reviews)	metadata (346,793 products)	image features
Health and Personal Care	reviews (2,982,326 reviews)	metadata (263,032 products)	image features
Toys and Games	reviews (2,252,771 reviews)	metadata (336,072 products)	image features
Video Games	reviews (1,324,753 reviews)	metadata (50,953 products)	image features
Tools and Home Improvement	reviews (1,926,047 reviews)	metadata (269,120 products)	image features
Beauty	reviews (2,023,070 reviews)	metadata (259,204 products)	image features
Apps for Android	reviews (2,638,173 reviews)	metadata (61,551 products)	image features
Office Products	reviews (1,243,186 reviews)	metadata (134,838 products)	image features
Pet Supplies	reviews (1,235,316 reviews)	metadata (110,707 products)	image features
Automotive	reviews (1,373,768 reviews)	metadata (331,090 products)	image features
Grocery and Gourmet Food	reviews (1,297,156 reviews)	metadata (171,760 products)	image features
Patio, Lawn and Garden	reviews (993,490 reviews)	metadata (109,094 products)	image features
Baby	reviews (915,446 reviews)	metadata (71,317 products)	image features
Digital Music	reviews (836,006 reviews)	metadata (279,899 products)	image features
Musical Instruments	reviews (500,176 reviews)	metadata (84,901 products)	image features
Amazon Instant Video	reviews (583,933 reviews)	metadata (30,648 products)	image features

Multi-modal Datasets

🌹🌹 Please cite our paper if you use the 'netflix' dataset~ ❤️
We collected a multi-modal dataset using the original Netflix Prize Data released on the Kaggle website. The data format is directly compatible with state-of-the-art multi-modal recommendation models like LLMRec, MMSSL, LATTICE, MICRO, and others, without requiring any additional data preprocessing.

`Textual Modality:` We have released the item information curated from the original dataset in the "item_attribute.csv" file. Additionally, we have incorporated textual information enhanced by LLM into the "augmented_item_attribute_agg.csv" file. (The following three images represent (1) information about Netflix as described on the Kaggle website, (2) textual information from the original Netflix Prize Data, and (3) textual information augmented by LLMs.)

`Visual Modality:` We have released the visual information obtained from web crawling in the "Netflix_Posters" folder. (The following image displays the poster acquired by web crawling using item information from the Netflix Prize Data.)



[34] He and McAuley, WWW '16, Ups and downs: Modeling the visual evolution of fashion trends with one-class …
[35] Wei et al., WWW '23, Multi-Modal Self-Supervised Learning for Recommendation

THE NEED FOR STANDARDIZATION

Without standardization:

- Different authors re-implement the same algorithms (reinventing the wheel)
- Different pipelines lead to different results (no reproducibility)
- No shared pipelines affect interoperability
- No shared pipelines cannot be customised for different experimental settings
- Pipeline validity can't be checked when not shared