Datasets

for “Graph-based recommender systems with explicit negative feedback encoding” research project.

Short-video/movie recommendations

VK RecSys

Link: https://ods.ai/competitions/aivkchallenge
We have been living in the world of social networks for quite a long time, and the topic of likes has long been clear (and dislikes too). AIVK (the direction that develops AI technologies in the company's key products) regularly studies and predicts feedback from VK Clips users to analyze data and improve the recommendation system.

Advantages

1. A new dataset with explicit negative feedback in the form of dislikes.
2. Competition.
3. There is relevant literature on the predicate for short videos.
4. Since the competition is new -> active community.

Disadvantages

1. A question about the possibility of using a dataset for the purpose of research.The possibility of using it should be agreed with the organizers.
2. Has not participated in international competitions and is not a standard for the studied research in this field.

KuaiRec

Link: https://kuairec.com
KuaiRec is a real-world dataset collected from the recommendation logs of the video-sharing mobile app Kuaishou. For now, it is the first dataset that contains a fully observed user-item interaction matrix. For the term “fully observed”, we mean there are almost no missing values in the user-item matrix, i.e., each user has viewed each video and then left feedback.

Advantages

1. There are features for users and items.
2. Fully observed user-item interaction matrix ensures comprehensive data coverage with no missing values.
3. With all user preferences known, KuaiRec can used in offline evaluation (i.e., offline A/B test) for recommendation models.

Disadvantages

1. High computational resources may be required due to the dense nature of the dataset.
2. Less applicable for studying sparsity-focused recommendation methods.
3. Implicit feedback (duration of viewing).

KuaiRand

Link: https://kuairand.com/
KuaiRand is an unbiased sequential recommendation dataset collected from the recommendation logs of the video-sharing mobile app, Kuaishou.

Advantages

1. Provides unbiased sequential recommendation data, essential for sequence-based modeling.
2. There are features for users and items.
3. We introduce 12 feedback signals (for example, click, like and view time) for each interaction to describe the complex feedback from the user.
4. It has the most comprehensive additional information, including explicit user IDs, interaction timestamps, and advanced features for users and products.

Disadvantages

1. High computational resources may be required due to the dense nature of the dataset.

KuaiSAR

Link: https://kuaisar.github.io/
KuaiSAR is a unified search and recommendation dataset containing the genuine user behavior logs collected from the short-video mobile app, Kuaishou, a leading short-video app in China with over 300 million daily active users.

Advantages

1. Large-scale, real-world dataset suitable for hybrid system research.
2. Covers a broad spectrum of user behavior with detailed interaction logs.
3. There are features for users and items.
4. Logs users' authentic interactions, including both positive and negative feedback.

Disadvantages

1. Might pose challenges for models not designed for hybrid systems.
2. High computational resources may be required due to the dense nature of the dataset.

MovieLens [all]

Link: https://grouplens.org/datasets/movielens/
GroupLens Research has collected and made available rating data sets from the MovieLens web site. The data sets were collected over various periods of time, depending on the size of the set. Before using these data sets, please review their README files for the usage licenses and other details.

Advantages

1. Widely used, benchmark dataset with consistent updates and community support.
2. Varied sizes (small to large) make it adaptable for different computational capacities.
3. Rich metadata and ratings provide opportunities for diverse recommendation tasks.

Disadvantages

1. Smaller datasets may not reflect real-world sparsity.

Yahoo Movies

Link: large link
This dataset contains a small sample of the Yahoo! Movies community's preferences for various movies, rated on a scale from A+ to F. Users are represented as meaningless anonymous numbers so that no identifying information is revealed. The dataset also contains a large amount of descriptive information about many movies released prior to November 2003, including cast, crew, synopsis, genre, average ratings, awards, etc.

Advantages

1. Contains rich descriptive information about movies (e.g., cast, crew, genre).
2. Includes user ratings and preferences for nuanced recommendation tasks.

Disadvantages

1. Limited to data collected before 2003, possibly outdated for modern trends.
2. Smaller size compared to contemporary datasets, potentially limiting scalability.

Netflix Prize

Link: https://academictorrents.com/details/9b13183dc4d60676b773c9e2cd6de5e5542cee9a
This is the official data set used in the Netflix Prize competition. The data consists of about 100 million movie ratings, and the goal is to predict missing entries in the movie-user rating matrix.

Advantages

1. Large-scale dataset with around 100 million ratings, suitable for robust modeling.
2. Benchmark for collaborative filtering and sparsity-focused research.

Disadvantages

1. Only contains ratings without other interaction types or contextual metadata.
2. Limited to data collected before 2009, possibly outdated for modern trends.

Music recommendations

Last-FM

Link: http://www.millionsongdataset.com/lastfm/index.html
The official song tag and song similarity dataset of the Million Song Dataset. The MSD team is proud to partner with Last.fm in order to bring you the largest research collection of song-level tags and precomputed song-level similarity. All the data is associated with MSD tracks, which makes it easy to link it to other MSD resources: audio features, artist data, lyrics, etc.

Advantages

1. A popular large dataset in the music domain.
2. There are statistics on 2 articles with negative feedback and can be found without using negative feedback.
3. Information about music is available: genres, titles, authors, you can work with the encoder for items.

Disadvantages

1. There are a lot of interactions, large computing resources are required.

Spotify

Link:
https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge/dataset_files
The Spotify Million Playlist Dataset Challenge consists of a dataset and evaluation to enable research in music recommendations. The dataset contains 1,000,000 playlists, including playlist titles and track titles, created by users on the Spotify platform between January 2010 and October 2017. The evaluation task is automatic playlist continuation: given a seed playlist title and/or initial set of tracks in a playlist, to predict the subsequent tracks in that playlist.

Advantages

1. A popular large dataset in the music domain.
2. There are statistics on 3 articles with negative feedback and can be found without using negative feedback.
3. Information about music is available: genres, titles, authors, you can work with the encoder for items.

Disadvantages

1. There are a lot of interactions, large computing resources are required.

Zvuk

Link: https://www.kaggle.com/datasets/alexxl/zvuk-dataset
The dataset was presented in the paper "From Variability to Stability: Advancing RecSys Benchmarking Practices" on the conference KDD'24

The dataset of listening to music tracks from the Zvuk service for a period of 4 months contains information about user sessions and is suitable for solving sequence recommendation problems.

Advantages

1. A new and medium-sized dataset.
2. A hackathon was recently held, you can take the results for research from there.
3. The dataset of listening to music tracks from the Zvuk service for a period of 4 months.

Disadvantages

1. Implicit feedback - listening length.
2. There are a lot of interactions, large computing resources are required.

Piki Music

Link: https://github.com/sstoikov/piki-music-dataset
The Piki Music dataset currently consists of 8896 anonymized users, 246,450 anonymized songs and 1,762,502 ratings and the data collection is still on-going.

Advantages

1. Strong feedback indicator:
   a. 2 if the song is super liked;
   b. 1 if the song is liked;
   c. 0 if the song is disliked;

Disadvantages

1. Small size of dataset.
2. Explicit feedback is expressed through implicit encoding.

Yahoo Music

Link: Yahoo Music
This dataset represents a snapshot of the Yahoo! Music community's preferences for various songs. The dataset contains over 717 million ratings of 136 thousand songs given by 1.8 million users of Yahoo! Music services. The data was collected between 2002 and 2006.

Advantages

1. A large collection of interactions.

2. Artist, Album, and Genre Meta Information.

Disadvantages

1. The dataset is quite old and may have completely different behavioral patterns that differ from modern ones.
2. We did not find statistics in studies correlating with our topic.

Amazon-Music

Link: https://www.kaggle.com/datasets/deovcs/amazon-dataset
This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).

Advantages

1. Here is an approach with evaluation and additionally with text feedback from users.
2. There are timestamps of interactions.

Disadvantages

1. A small dataset.
2. It was not often used in research, we have 1 research paper in the chosen domain.

E-commerce recommendations

Taobao

Link: https://tianchi.aliyun.com/dataset/649
User Behavior is a dataset of user behaviors from Taobao, for recommendation problems with implicit feedback. The dataset is offered by Alibaba.

Advantages

1. A constantly updated dataset from 2012 (although the last update was in 2017).
2. A standard for the studied research in this field.
3. Open data - frequently encountered in relevant studies.

Disadvantages

1. Implicit feedback.
2. Short time span: The data spans from November 25 to December 3, 2017, which is only 9 days.
3. Lack of contextual information.

MegaMarket

Link: https://www.kaggle.com/datasets/alexxl/megamarket/data
The dataset on purchases on the MegaMarket marketplace for a period of 4 months, there are 2,730,776 users and 3,562,321 items

Advantages

1. New dataset (2023)
2. Long time span: The data spans for 4 months.

Disadvantages

1. There are just 3 type of event: 0 - click, 1 - add to cart, 2 - purchase
2. There are a lot of interactions, large computing resources are required. (200.000.000)

Amazon-Electronics (all directions)

Link: https://nijianmo.github.io/amazon/index.html
This Dataset is an updated version of the Amazon review dataset released in 2014. As in the previous version, this dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs) (2018)

Advantages

1. Very extensive description - there are reviews (text), and ratings, and price lists etc
2. Long time span (1996-2018)
3. There are all information about product and review from user (and rating feedback)
4. A popular dataset (but using by partition)

Disadvantages

1. There are a lot of interactions, large computing resources are required (the most extensive dataset than other)

Retailrocket

Link: https://www.kaggle.com/retailrocket/ecommerce-dataset

The dataset consists of three files: a file with behaviour data (events.csv), a file with item properties (item_properties.csv) and a file, which describes category tree (category_tree.csv). The data has been collected from a real-world ecommerce website. It is raw data, i.e. without any content transformations, however, all values are hashed due to confidential issues. The purpose of publishing is to motivate researches in the field of recommender systems with implicit feedback.

Advantages

1. implicit feedback
2. there is use in other research studies

Disadvantages

1. Unknown source
2. the data was output only once, never supplemented or improved

Yelp (2022)

Link:
https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset/download?datasetVersionNumber=4
It is the latest version of Yelp dataset, which contains 908,915 tips by 1,987,897 users over 1.2 million business attributes like hours, parking, availability, and ambience aggregated check-ins over time for each of the 131,930 businesses. Yelp Dataset is a dataset provided by the Yelp platform that primarily focuses on local businesses and user interactions with establishments such as restaurants, stores, beauty salons, etc.

Advantages

1. User reviews (with text and ratings).
2. Establishment information (categories, location, hours).
3. Social interactions such as friends and recommendations.

4. Feedback - Reviews include explicit feedback in the form of ratings (1-5 stars) and text reviews, making Yelp useful for sentiment analysis and recommendation systems.

Disadvantages

1. Data does not include products or transactions.
2. Limited scope (businesses instead of product sales).

Diginetica

Link: https://competitions.codalab.org/competitions/11161
Dataset provided by DIGINETICA and its partners containing anonymized search and browsing logs, product data, anonymized transactions, and a large data set of product images.

Advantages

1. Session focus - the dataset is organized by user sessions, making it ideal for analyzing the sequence of user actions.
2. Interaction logs, including clicks, search queries, views, and purchases. Product metadata: categories, descriptions, prices.
3. The data contains explicit (purchases) and implicit (clicks, views) feedback. Used to build sequential recommendation models (e.g., session-based recommendation systems).
4. Well suited for temporal behavior analysis.

Disadvantages

1. No information about users' social connections.
2. Does not provide geographic information or text reviews.