RUCAIBox / RecSysDatasets Public

Code Issues 7 Pull requests Actions Projects Wiki Security Insights

master

RecSysDatasets / README.md

ShanleiMu rename Baidu Yun to Baidu Wangpan

10 contributors

189 lines (154 sloc) | 15.7 KB

# Datasets For Recommender Systems

This is a repository of public data sources for Recommender Systems (RS).

All of these recommendation datasets can convert to the atomic files defined in RecBole which is a unified, comprehensive and efficient recommendation library.

After converting to the atomic files, you can use RecBole to test the performance of different recommender models on these datasets easily. For more information about RecBole, please refer to RecBole.

## Usage

In order to use RecBole, you need to convert these original datasets to the atomic file which is a kind of data format defined by RecBole.

We provide two ways to convert these datasets into atomic files:

1. Download the raw dataset and process it with conversion tools we provide in this repository. Please refer to conversion tools.
2. Directly download the processed atomic files. Baidu Wangpan (Password: e272), Google Drive.

## Datasets link and brief introduction

## Shopping

- Amazon: This dataset contains product reviews, only-rating data (ratings) and metadata(descriptions, category information, price, brand, and image features) from Amazon, including 142.8 million reviews spanning May 1996 - July 2014.
- Epinions: This dataset was collected from Epinions.com, a popular online consumer review website. It contains trust relationships amongst users and spans more than a decade, from January 2001 to November 2013.
- Yelp: This dataset was collected from Yelp.com. The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes.
- Tmall: This dataset is provided by Ant Financial Services, using in the IJCAI16 contest.
- DIGINETICA: The dataset includes user sessions extracted from an e-commerce search engine logs, with anonymized user ids, hashed queries, hashed query terms, hashed product descriptions and meta-data, log-scaled prices, clicks, and purchases.
- YOOCHOOSE: This dataset was constructed by YOOCHOOSE GmbH to support participants in the RecSys Challenge 2015.
- Retailrocket: 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.
- Ta Feng: The dataset contains a Chinese grocery store transaction data from November 2000 to February 2001.

## Advertising

- Criteo: This dataset was collected from Criteo, which consists of a portion of Criteo's traffic over a period of several days.
- Avazu: This dataset is used in Avazu CTR prediction contest.
- iPinYou: This dataset was provided by iPinYou, which contains all training datasets and leaderboard testing datasets of the three seasons iPinYou Global RTB(Real-Time Bidding) Bidding Algorithm Competition.

## Check-in

- Foursquare: This dataset contains check-ins in NYC and Tokyo collected for about 10 month. Each check-in is associated with its time stamp, its GPS coordinates and its semantic meaning.
- Gowalla: This dataset is from a location-based social networking website where users share their locations by checking-in, and contains a total of 6,442,890 check-ins of these users over the period of Feb. 2009 - Oct. 2010.

### Movies

- MovieLens: GroupLens Research has collected and made available rating datasets from their movie web site.
- Netflix: This is the official data set used in the Netflix Prize competition.
- Douban: Douban Movie is a Chinese website that allows Internet users to share their comments and viewpoints about movies. This dataset contains more than 2 million short comments of 28 movies in Douban Movie website.

### Music

- Last.FM: This dataset contains social networking, tagging, and music artist listening information from a set of 2K users from Last.fm online music system.
- LFM-1b: This dataset contains more than one billion music listening events created by more than 120,000 users of Last.FM. Each listening event is characterized by artist, album, and track name, and includes a timestamp.
- Yahoo Music: This dataset represents a snapshot of the Yahoo! Music community's preferences for various musical artists.

### Books

- Book-Crossing: This dataset was collected by Cai-Nicolas Ziegler in a 4-week crawl (August / September 2004) from the Book-Crossing community with kind permission from Ron Hornbaker, CTO of Humankind Systems. It contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.

### Games

- Steam: This dataset is reviews and game information from Steam, which contains 7,793,069 reviews, 2,567,538 users, and 32,135 games. In addition to the review text, the data also includes the users' play hours in each review.

### Anime

- Anime: This dataset contains information on user preference data from myanimelist.net. Each user is able to add anime to their completed list and give it a rating and this dataset is a compilation of those ratings.

### Pictures

- Pinterest: This dataset is originally constructed by paper Learning image and user features for recommendations in social networks for evaluating content-based image recommendation, and processed by paper Neural Collaborative Filtering.

### Jokes

- Jester: This dataset contains anonymous ratings of jokes by users of the Jester Joke Recommender System.

### Exercises

- KDD2010: This dataset was released in KDD Cup 2010 Educational Data Mining Challenge, which contains the situations of students submitting exercises on the systems.

### Websites

- Phishing Websites: This dataset contains 30 kinds of features of 11,055 websites and labels of whether they are phishing websites or not. The websites' features includes 12 address-bar based features, 6 abnormal based features, 5 HTML-and-JavaScript based features and 7 domain based features.

### Adult

- Adult: This dataset is extracted by Barry Becker from the 1994 Census database, which consists of a list of people's attributes and whether they make over 50k a year.

### News

- MIND This dataset is a large-scale dataset for news recommendation research. It was collected from anonymized behavior logs of Microsoft News website. MIND contains about 160k English news articles and more than 15 million impression logs generated by 1 million users.

## Datasets information statistics

### General Datasets

| SN | Dataset | #User | #Item | #Inteaction | Sparsity | |
|---|---|---|---|---|---|---|
| 1 | MovieLens | - | - | - | - | Ratin |
| 2 | Anime | 73,515 | 11,200 | 7,813,737 | 99.05% | Ratin<br>[-1, 1 |
| 3 | Epinions | 116,260 | 41,269 | 188,478 | 99.99% | Ratin<br>[1-5] |

| SN | Dataset | #User | #Item | #Inteaction | Sparsity | |
|---|---|---|---|---|---|---|
| 4 | Yelp | 1,968,703 | 209,393 | 8,021,122 | 99.99% | Ratin<br>[1-5] |
| 5 | Netflix | 480,189 | 17,770 | 100,480,507 | 98.82% | Ratin<br>[1-5] |
| 6 | Book-Crossing | 105,284 | 340,557 | 1,149,780 | 99.99% | Ratin<br>[0-10 |
| 7 | Jester | 73,421 | 101 | 4,136,360 | 44.22% | Ratin<br>[-10, |
| 8 | Douban | 738,701 | 28 | 2,125,056 | 89.73% | Ratin<br>[0,5] |
| 9 | Yahoo Music | 1,948,882 | 98,211 | 11,557,943 | 99.99% | Ratin<br>[0, 10 |
| 10 | KDD2010 | - | - | - | - | Ratin |
| 11 | Amazon | - | - | - | - | Ratin |
| 12 | Pinterest | 55,187 | 9,911 | 1,445,622 | 99.74% | - |
| 13 | Gowalla | 107,092 | 1,280,969 | 6,442,892 | 99.99% | Chec |
| 14 | Last.FM | 1,892 | 17,632 | 92,834 | 99.72% | Click |
| 15 | DIGINETICA | 204,789 | 184,047 | 993,483 | 99.99% | Click |
| 16 | Steam | 2,567,538 | 32,135 | 7,793,069 | 99.99% | Buy |
| 17 | Ta Feng | 32,266 | 23,812 | 817,741 | 99.89% | Click |
| 18 | Foursquare | - | - | - | - | Chec |
| 19 | Tmall | 963,923 | 2,353,207 | 44,528,127 | 99.99% | Click/ |
| 20 | YOOCHOOSE | 9,249,729 | 52,739 | 34,154,697 | 99.99% | Click/ |
| 21 | Retailrocket | 1,407,580 | 247,085 | 2,756,101 | 99.99% | View, |
| 22 | LFM-1b | 120,322 | 3,123,496 | 1,088,161,692 | 99.71% | Click |
| 23 | MIND | - | - | - | - | Click |

## CTR Datasets

| SN | Dataset | #User | #Item | #Inteaction | Sparsity | Interaction Type |
|---|---|---|---|---|---|---|
| 1 | Criteo | - | - | 45,850,617 | - | Click |
| 2 | Avazu | - | - | 40,428,967 | - | Click<br>[0, 1] |
| 3 | iPinYou | 19,731,660 | 163 | 24,637,657 | 99.23% | View/Click |
| 4 | Phishing websites | - | - | 11,055 | - | |
| 5 | Adult | - | - | 32,561 | - | income>=50k<br>[0, 1] |

## Knowledge-aware Datasets

These knowledge-aware recommender datasets are based on KB4Rec, which associate items from recommender systems with entities from Freebase.

Raw datasets information

| SN | Dataset | #Items | #Linked-Items | #Users | #Interactions |
|---|---|---|---|---|---|
| 1 | MovieLens | 27,278 | 25,503 | 138,493 | 20,000,263 |
| 2 | Amazon-book | 2,370,605 | 108,515 | 8,026,324 | 22,507,155 |
| 3 | LFM-1b (tracks) | 31,634,450 | 1,254,923 | 120,322 | 319,951,294 |

After filtering by 5-core (And filter out the tracks that are listened to less than 10 times in LFM-1b)

| SN | Dataset | #Items | #Linked-Items | #Users | #Interactions |
|---|---|---|---|---|---|
| 1 | MovieLens | 18,345 | 18,057 | 138,493 | 19,984,024 |
| 2 | Amazon-book | 367,982 | 34,476 | 603,668 | 8,898,041 |
| 3 | LFM-1b (tracks) | 615,823 | 337,349 | 79,133 | 15,765,756 |