

Evaluating Collaborative Filtering Recommender Systems

Jonathan L. Herlocker, Joseph A. Konstan, Loren g. Terveen, and John T. Riedel

Human Computer Interaction Group
École Polytechnique Fédérale de Lausanne (EPFL)

Rong Hu



Challenges

movielens
helping you find the right movies

Welcome dus@infovis.net
You've rated 48 movies.
You're the 24th visitor in the past hour.

★★★★★ = Must See
★★★★☆ = Will Enjoy
★★★☆☆ = It's OK
★★☆☆☆ = Fairly Bad
★☆☆☆☆ = Awful

Search Titles
Go!

Search by Genre/Date
All Genres | All Dates
Domain: All movies
Search Genre/Date!

Advanced Search

Select Buddies
Test Buddy
What are buddies?

You've searched for all titles.
Found 7233 movies, sorted by Prediction
Genres: All | Exclude Genres: None
Dates: All | Domain: All | Format: All | Language: All
Show Printer-Friendly Page | Download Results | Suggest a Title

Page 1 of 483 | Go to page: 1...96...192...288...384...480...last page 2>

Predictions for you	Your Ratings	Movie Information	Wish List
★★★★★	Not seen	Tainted (1998) info imdb Comedy, Thriller	<input type="checkbox"/>
★★★★★	Not seen	Friday Night Lights (2004) info imdb Action, Drama	<input type="checkbox"/>
★★★★★	Not seen	Harry Potter and the Prisoner of Azkaban (2004) info imdb Adventure, Children, Fantasy	<input type="checkbox"/>
★★★★★	Not seen	Spider-Man 2 (a.k.a. Spiderman 2) (2004) info imdb Action, Fantasy, Sci-Fi, Thriller	<input type="checkbox"/>
★★★★★	Not seen	Finding Nemo (2003) DVD, VHS, info imdb Adventure, Animation, Children, Comedy	<input type="checkbox"/>
★★★★★	Not seen	X-Men 2 (a.k.a. X2: X-Men United) (2003) DVD, VHS, info imdb Action, Adventure, Sci-Fi	<input type="checkbox"/>
★★★★★	Not seen	Oliver Twist (1948) info imdb Adventure, Crime, Drama	<input type="checkbox"/>
★★★★★	Not seen	Raiders of the Lost Ark (1981) DVD, info imdb Action, Adventure	<input type="checkbox"/>
★★★★★	Not seen	Indiana Jones and the Last Crusade (1989) DVD, info imdb Action, Adventure	<input type="checkbox"/>

MovieLens

vs.

NETFLIX
Justin Schneiderman | Your Account | Buy / Redeem Gift | Help

Browse DVDs | Watch Instantly | Your Queue | Movies You'll Love | Your Friends | DVD Sale \$5.99 +

Home | Genres | New Releases | Previews | Netflix Top 100 | Critics' Picks | Award Winners

Give Dad a gift he'll truly enjoy. Gifts from only \$4.99. Give

Movies For You

Justin, the following movies were chosen based on your interest in: Pan's Labyrinth, Music and Lyrics, Crash

OTHER MOVIES YOU MIGHT ENJOY

Peaceful Warrior | Blood Diamond | Hollywoodland

The Devil's Backbone
Twelve-year-old Carlos (Fernando Tielve) is the latest arrival at Santa Lucia School, an imposing stone building that shelters orphans of the ...
Read More

The Holiday
Stuck in a vicious cycle of dead-end relationships with two-timing men, Los Angeles resident Amanda (Cameron Diaz) and Londoner Iris (Kate Winslet) ...
Read More

Browse
Favorite Genres:
(Add Favorite Genres)
Other Genres:
All Genres
Action & Adventure
Anime & Animation
Blu-ray
Children & Family
Classics
Comedy
Documentary
Drama
Faith & Spirituality
Foreign
Gay & Lesbian
HD DVD
Horror
Independent
Music & Musicals
Romance
Sci-Fi & Fantasy
Special Interest
Sports & Fitness
Television

Netflix

- Evaluating recommender systems and their algorithms is **inherently difficult**
 - Different algorithms may be better or worse on different **data sets**
 - **Goals** for which an evaluation is performed may differ
 - Deciding what **combination** of measures to use in comparative evaluation is a significant challenge



Recommendation Algorithm Evaluation

- User Tasks
- Evaluation Methods
- Datasets
- Evaluation Metrics



- Annotation in Context
 - Filter through structure database to annotate messages in context using predictions
 - Examples:
 - Tapestry [Goldberg et al. 1992], GroupLens [Resnick et al. 1994]
- Find Good Items
 - Suggest specific items to users, and provide users with a ranked list of the recommended items
 - Examples:
 - Ringo [Shardanand and Maes 1995], Bellcore Video Recommender [Hill et al. 1995]



- Find All Good Items
 - Coverage becomes important in this task
- Recommend Sequence
 - The whole list should be pleasing as whole
- Just Browsing
 - Interface, ease of use, level and nature of information
- Find Credible Recommender
 - Useful recommendations that the user is sure to enjoy.

Users



- Key Decisions
 - Live User Experiment vs. Offline Analyses
 - Synthesized vs. Natural Data Sets
 - Properties of Data Sets



- Offline Analyses
 - Algorithm Evaluation
 - Predict certain withheld values from a dataset
- Advantage
 - **Quick and economical** to conduct large evaluations
- Weaknesses
 - **Natural sparsity of ratings data sets** limits the set of items that can be evaluated
 - They are limited to objective evaluation of prediction results



- Live User Experiments
 - Controlled experiments
 - Field studies
- Advantage
 - Evaluate user performance, satisfaction, participation, et
- Disadvantage
 - Expensive
 - The bias of users' responses



- Natural data sets
 - May imperfectly match the properties of the target domain and task
- Synthesized data sets
 - Specifically to match those properties
 - Appropriate for new domain
 - Unfair to other algorithms
 - › It fits their approach too well
 - › Drawing comparative conclusions is risky



- What properties should the dataset have in order to best model the tasks for which the recommender is being evaluated?
- Data Set Properties
 - Domain Features
 - e.g., user task, content domain
 - Inherent Features
 - e.g., implicit/explicit ratings, rating scales, item content attributes, demographic information
 - Sample Features
 - e.g., density, size, distribution





- Predictive Accuracy Metrics
- Classification Accuracy Metrics
- Rank Accuracy Metrics



- Measure how close the recommender system's predicted ratings are to the true user ratings
- Particular important for evaluating tasks where the predicting rating will be displayed to the user, e.g., *Annotation in Context*



- Mean Absolute Error (MAE)

$$|\bar{E}| = \frac{\sum_{i=1}^N |p_i - r_i|}{N}$$

- Advantages
 - Simple and easy to understand
 - Has well studied statistical properties
- Weakness
 - Less appropriate for the tasks returning a rank list, e.g., *Find Good Items*
 - Less appropriate when the granularity of true preference is small



- Mean Squared Error
 - Emphasis on large errors
- Root Mean Squared Error
 - Emphasis on large errors
- Normalized mean absolute error
 - Allow comparison between prediction runs on different datasets.





Classification Accuracy Metrics

- Measure the frequency with which a recommender system makes correct or incorrect decisions about whether an item is good
- Appropriate for tasks when users have true binary preferences, such *Find Good Items*



- Precision and Recall

	Selected	Not Selected	Total
Relevant	N_{rs}	N_{rn}	N_r
Irrelevant	N_{is}	N_{in}	N_I
Total	N_s	N_n	N

- **Precision:** the probability that a selected item is relevant

$$P = \frac{N_{rs}}{N_s}$$

- **Recall:** the probability that a relevant items will be selected

$$R = \frac{N_{rs}}{N_r}$$



- Precision and Recall
 - Advantages:
 - › More comprehensive
 - Disadvantages:
 - › Need to convert multi-scale ratings to binary scale
 - › Relevance is more inherently subjective in RS
 - › Recall and Precision should be considered together

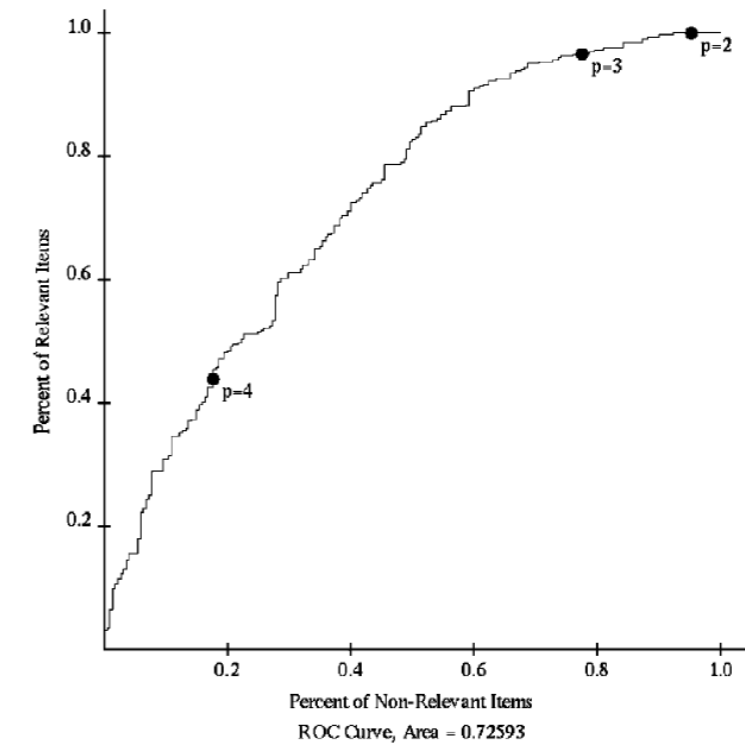
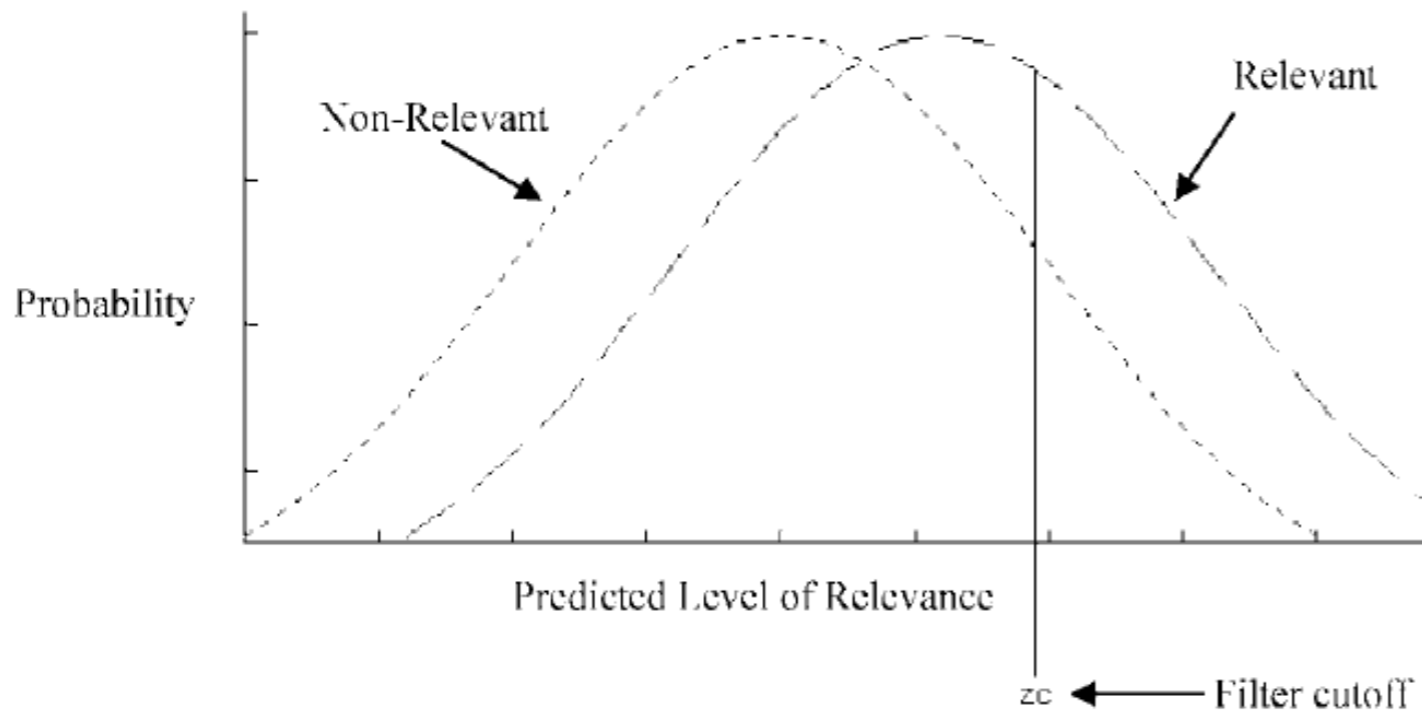
- **F1** metric combines precision and recall

$$F_1 = \frac{2PR}{P + R}$$



- ROC Curves

- ROC: Relative Operating Characteristic / Receiver Operating Characteristic



- ROC Curves

- Advantages

- › Developed from solid statistical decision theory
- › A single performance measure
- › Covers different recommendation lengths.

- Weakness

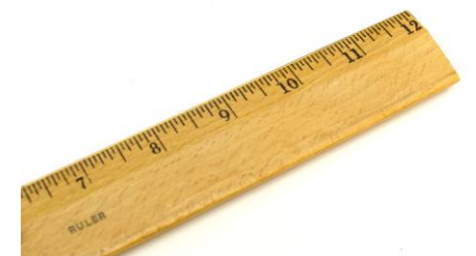
- › Ordering among relevant items has no impact.
- › A large set of potentially relevant items is needed
- › Users are only interested in one setting
- › Need a large number of data points to ensure good statistical power



- Measure the ability of a recommendation algorithm to produce a recommended ordering of items that matches how the user would have ordered the same items
- Appropriate for ranked list and in domains where users' preferences are non-binary



- Prediction-Rating Correlation
 - Pearson, Spearman's ρ , Kendall's Tau
 - Advantages
 - › Well understood by the scientific community
 - › Provide a single measurement score
 - Disadvantages
 - › Interchange weakness



- Half-life Utility Metric

- Evaluate the utility of a ranked list to the user
- Best for the tasks domain where there is an exponential drop in true utility as the search length increase.

$$R_a = \sum_j \frac{\max(r_{a,j} - d, 0)}{2^{(j-1)/(\alpha-1)}}$$

- NDPM Measure

- Normalized Distance-based Performance Measure
- Comparable among different datasets

$$NDPM = \frac{2C^- + C^+}{2C^l}$$



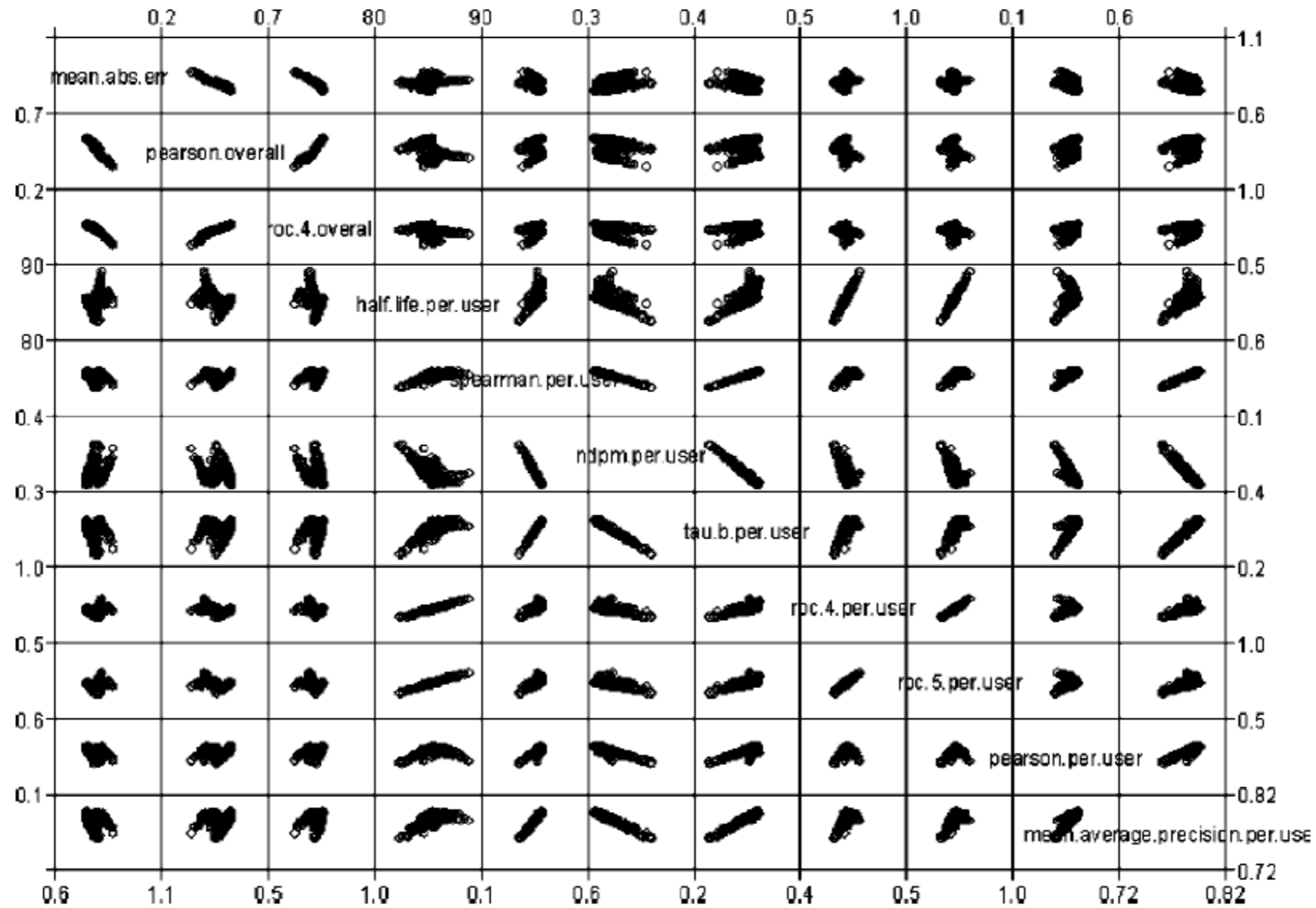


An Empirical Comparison of Evaluation Metrics

- Examine the extent to which the different evaluation metrics agreed or disagreed.
 - 432 different variants of the algorithm tested (Size of neighborhood, similarity, threshold, type of recommendation)
 - Examined evaluation metrics: MAE, Pearson, Spearman, ROC-4, ROC-5, half-life utility metric, mean average precision, NDPM metric
 - Per user/Overall
 - Dataset: MovieLens Dataset (100,000 movie ratings from 943 users on 1682 items)



An Empirical Comparison of Evaluation Metrics



- Coverage
 - Measures the percentage of a datasets that the recommender system is able to provide predictions
- Learning Rate
 - Measure how quickly an algorithm can produce good recommendations
- Novelty and Serendipity
 - Measure whether a recommendation is a novel possibility
 - Example: Bananas in a grocery store



- Confidence
 - Measure how sure is the recommender system that its recommendation is accurate



Personalized Prediction : ★★★★★ (will enjoy it)

Your Neighbors' Ratings for this Movie

Rating	Number of Neighbors
★	1
★★	2
★★★	7
★★★★	14
★★★★★	9

([Herlocker et al. 2000])

- User Evaluation
 - How to evaluate user reaction to a recommender system



- Effective and meaningful evaluation of recommender systems is challenging.
- Choose appropriate evaluation **datasets, methods** and **metrics**, taking into account **user tasks** and **contexts**.
- Accuracy alone is NOT enough.



Thanks!
Discussion