Motivation

Evaluation of Recommender Systems is still an area of active research.

Evaluation methodologies:
- Error-based (accuracy)
- Precision-oriented (ranking quality)

Realization that quality of the ranking is more important than accuracy in predicting rating values

Problem: difficult to compare results from different works

Precision-oriented metrics depend on
- Amount of relevant items
- Amount of non-relevant items

Different assumptions about the non-relevant set leads to biases in the measurements

Approach

A general methodology for evaluating ranked item lists

For each target user \( u \), we select \( L_u \) of target items for ranking:
- For each user and item in the set, we request a rating prediction \( r(u,i) \)
- We sort the items by decreasing order of predicted rating value

Different authors have built the set \( L_u \) differently

Different methodologies used in the state-of-the-art

(Notation: \( T_r \) and \( T_e \) denote training and test sets)
- TestRatings (TR): \( L_u = T_e \), It needs a relevance threshold
- TestItems (TeI): \( L_u = U \setminus T_e \)
- TrainingItems (TrI): \( L_u = U \cup T_r \)
- AllItems (AI): \( L_u = T \setminus T_r \)
- One-Plus-Random (OPR): \( L_u = \{i \mid |\forall \pi \in NR_i, \pi \in L_u\} \)

Recall@50: \( \sum \frac{|\pi \cap L_u|}{|\pi|} \)

nDCG@50: normalized discounted cumulative gain at 50

Reciprocal Rank (MRR)

\[ \sum_{u \in U} \frac{1}{|\pi_u|} \]

RMSE: root mean square error

Empirical comparison

Dataset: MovieLens 100K

- Recommenders
  - UB50: user-based recommender with 50 neighbors
  - IB: item-based recommender using adjusted cosine
  - SVD: matrix factorization technique using 50 factors

Metrics
- \( P@50 \): precision at 50
- \( R@50 \): recall at 50
- \( nDCG@50 \): normalized discounted cumulative gain at 50
- \( RMSE \): root mean square error

Future Work

Online experiment with real users’ feedback

Evaluate other metrics
- From IR: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR)
- From RS: Normalized Distance-based Performance Measure (NDPM), ROC curve

Alternative training / test generation
- E.g., temporal split

Conclusions

Four out of five methodologies are consistent with each other

The other methodology (TestRatings) has proved to overestimate performance values.

No direct equivalence found between results with error-based and precision-based metrics

Performance range of results depends on the methodology

References