Precision-oriented Evaluation of Recommender Systems: An Algorithmic Comparison

Alejandro Bellogín, Pablo Castells, Iván Cantador

Information Retrieval Group

Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain

{alejandro.bellogin, pablo.castells, ivan.cantador}@uam.es

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 a set L_u of target items for ranking:

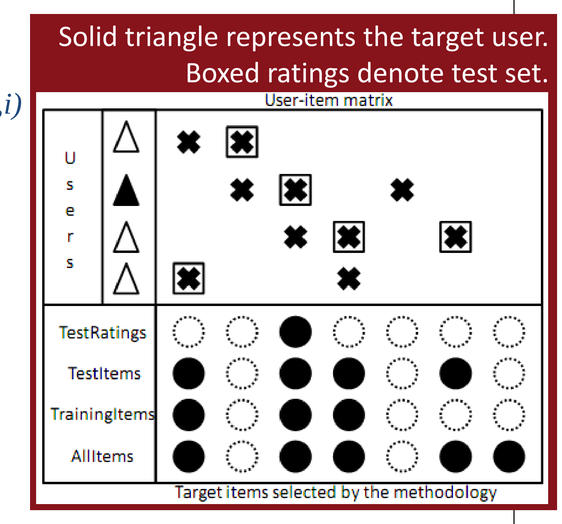
- For each user and item in the set, we request a rating prediction $r(u,i)$
- We sort the items by decreasing order of predicted rating value

Different authors have built the set L_u differently

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

(Notation: *Tr* and *Te* denote training and test sets)

- TestRatings (TR): $L_u = Te_u$. It needs a relevance threshold
- TestItems (TeI): $L_u = \cup_v Te_v \setminus Tr_u$
- TrainingItems (TrI): $L_u = \cup_{v \neq u} Tr_v$
- AllItems (AI): $L_u = \mathcal{I} \setminus Tr_v$
- One-Plus-Random (OPR): $L_{ui} = \{i\} \cup NR_u$, for i in $HR_u \subseteq Te_u$, $|NR_u| = 1000$

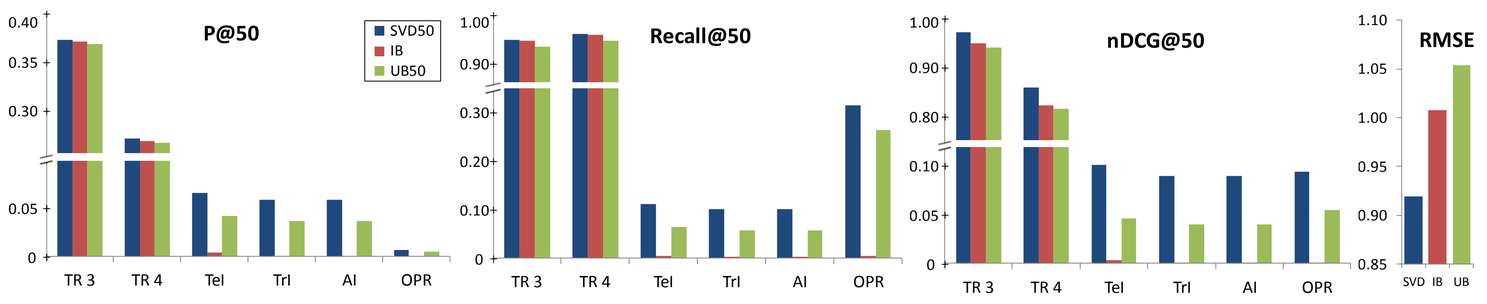


Empirical comparison

5-fold

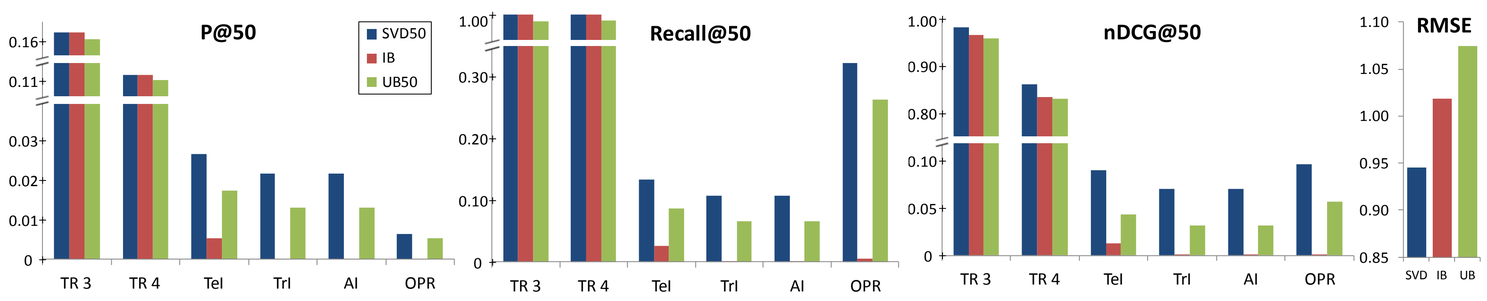
80% training

20% test



2 splits

10 ratings per user in test



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
- Recall@50: recall at 50
- nDCG@50: normalized discounted cumulative gain at 50
- RMSE: root mean square error

Discussion

- Comparative results with precision metrics are not the same as with error metrics (IB better than UB for RMSE, not for precision)
- TestRatings methodology only evaluates recommendations over known relevance → unrealistic situation.
- TestRatings' ranked list consists of top *rated* items, which may or may not be related with the recommended items the user would get in a real application
- Absolute performance values obtained by each methodology are very different
- TestItems obtains higher performance values than TrainingItems since non-relevant items for every user are omitted
- TrainingItems and AllItems are, as expected, completely equivalent
- The five methodologies are consistent for the two datasets, even though the test size for each user is different in each situation

$$P@50 = \frac{1}{|U|}\sum_{u \in U}\frac{|\mathrm{Rel}_u@50|}{50} \qquad \mathrm{nDCG@50} = \frac{1}{|U|}\sum_{u \in U}\frac{1}{\mathrm{IDCG}_u@50}\sum_{i=1}^{50}\frac{2^{rel(i_u)}-1}{\log(1+i_u)}$$

$$\mathrm{Recall@50} = \frac{1}{|U|}\sum_{u \in U}\frac{|\mathrm{Rel}_u|}{|\mathrm{Rel}_u@50|} \qquad \mathrm{RMSE} = \sqrt{\frac{1}{|Te|}\sum_{(u,i) \in Te}|\tilde{r}(u,i) - r(u,i)|}$$

Check the source code for the different methodologies: http://ir.ii.uam.es/evaluation/rs

References

(Bellogín et al 2011) A. Bellogín, J. Wang, P. Castells. Text Retrieval Methos Applied to Ranking Items in Collaborative Filtering. In ECIR 2011.

(Cremonesi et al 2010) P. Cremonesi, Y. Koren, R. Turrin. Performance of Recommender Algorithms on Top-N Recommendation Tasks. In RecSys 2010.

(Jambor & Wang 2010a) T. Jambor and J. Wang. Goal-driven Collaborative Filtering – A Directional Error Based Approach. In ECIR 2010.

(Jambor & Wang 2010b) T. Jambor and J. Wang. Optimizing Multiple Objectives in Collaborative Filtering. In Recsys 2010.

(Koren 2008) Y. Koren. Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model. In KDD 2008.

Methodology	Reference(s)
TestRatings	(Jambor & Wang 2010a) (Jambor & Wang 2010b)
TestItems	(Bellogín et al 2011)
OnePlusRandom	(Cremonesi et al 2010) (Koren 2008)

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

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

5th ACM Conference on Recommender Systems (RecSys2011)

Chicago, USA, 23-27 October 2011