Towards a More Realistic Evaluation: Testing the Ability to

Motivation

Temporal dimension is acquiring more importance in Recommender Systems. Many new algorithms try to incorporate temporal information (Ding & Li, 2005; Koren, 2009; Lathia, 2010).

An evaluation protocol taking into consideration evolution through time (Shani & Gunawardana, 2011), should be used to assess improvements of such models (Lathia et al., 2009).

However, it is notable that some evaluation settings used to test time-aware recommendation methods are not as realistic as should be expected.

And the usage of data made by Time Aware Matrix Factorization (Koren, 2009):

\[ \hat{r}_{u,i} = \mu + b_u(t) + b_i(t) + p_u(t)q_i \]  

(1)

\[ p_u(t) = p_{u,i} + \alpha_{u,i} \cdot \text{sign}(t - t_i)[t - t_i]^{\beta} + p_{u,i}t \]  

(2)

\[ b_u(t) = b_u + \alpha_u \cdot \text{sign}(t - t_u)[t - t_u]^{\beta} + b_{u,t} \]  

(3)

Red terms rely on the prediction date.

Note that such terms require overlapping training/test data.

Would the Time Aware Matrix Factorization model outperform a basic Matrix Factorization model if no overlapping training/test data (from a temporal point of view) were available?

Evaluation Protocol

Time aware training/validation/test data division.
Selection of 500 users with longest rating timespan.
Generation of 5 overlapping splits of 100 randomly assigned users each.
Evaluation on 21 separated temporal test intervals of 30 days each.

Results

Changes in test ratings distribution affect algorithms almost equally.

MF achieve best results on both metrics when using Time Aware protocol.

When using Last Ratings protocol, lowest RMSE is achieved using TimeMF_Future (as expected); best precision is achieved by MF.

Conclusions

Recent or ongoing work show that:

On evaluated tasks, simpler is better, i.e., using a basic MF model yields to better results than more complex model (due to overfitting?).

Using a more strict time-aware evaluation may lead to important differences in results, when compared with a less restrictive scheme.

Future Work:

Replication on larger, up-to-date datasets is required. Analyze other time-aware approaches.