Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Track 1 – Matrix Factorization

Implementation Ideas

Collaborative Filtering

Markus Freitag, Jan-Felix Schwarz

12 May 2011

Data Processing

- **MySQL** database
- **Talend** to load training and validation data
- Partition training and validation sets in separate sets for different item types → **Views**

Feature Training

- Implemented the **Stochastic Gradient Descent** algorithm
  - Trains one feature
  - Approximation procedure
  - Each iteration processes the whole rating set

- **Todo**
  - Identify causes for anomalies
  - Performance optimization: cache as much as possible
    - Optimally use complete available memory
    - Improve memory efficiency

How many Features to Learn?

Regularization

Idea: penalize complexity

$$\min_{q^*,p^*} \sum_{(u,i)\in\kappa} (r_{ui} - q_i^T p_u)^2 + \lambda(|| q_i ||^2 + || p_u ||^2)$$

- Implement a second approximation algorithm that calls *trainNewFeature()* until the minimization goal is reached

Item Relations

- Use explicit item relations (hierarchy) to improve predictions
- Come up with hypothesis and validate them
  - E.g.: “user’s genre rating affects the user’s predicted ratings of artists of that genre”
  - Anything lowering the RMSE is good
- Relearn feature vectors?
  - In which order?

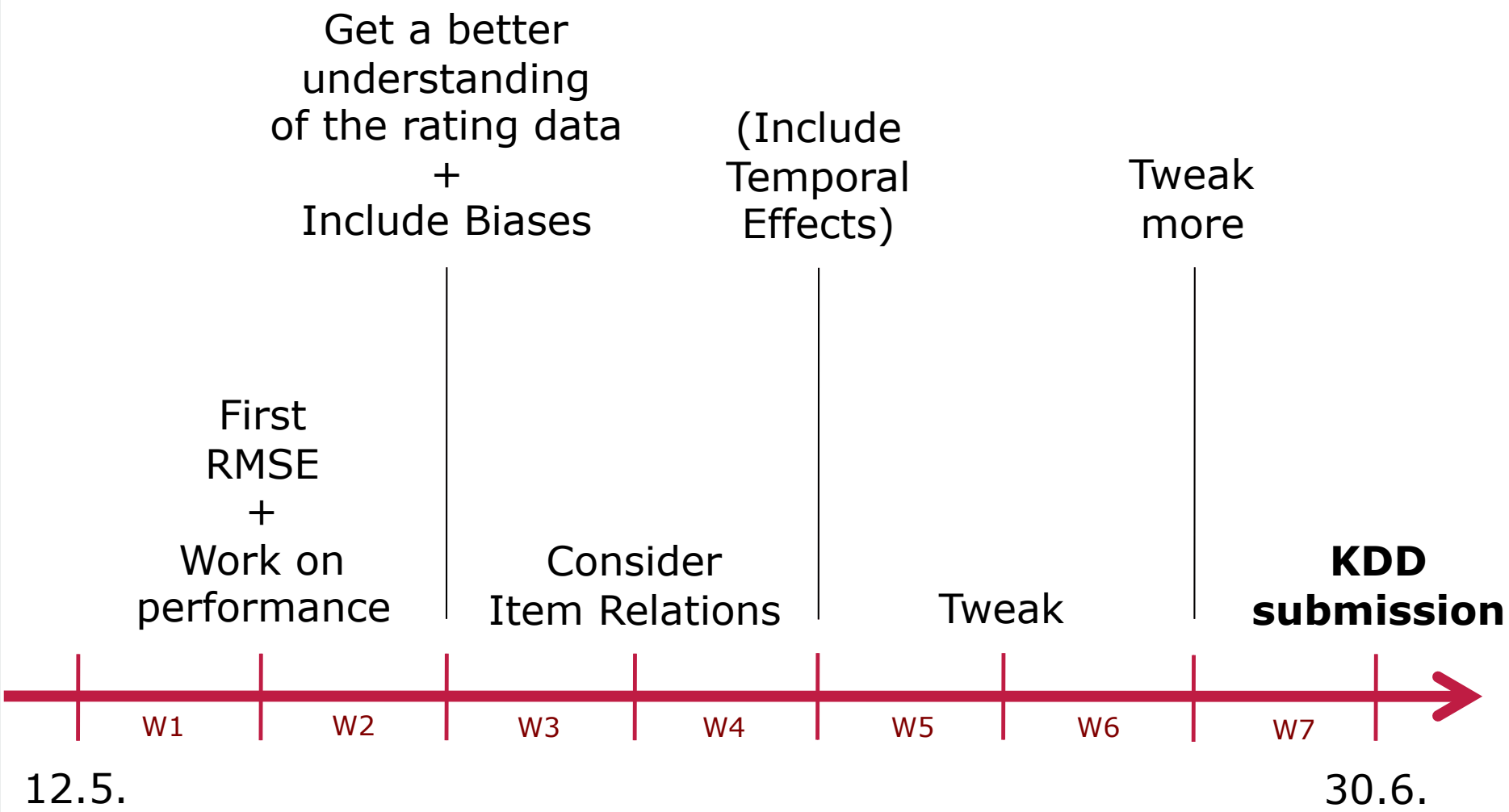
Biases and Temporal Effects

- Proposed by Bell et. al to improve predictions
- Calculate biases (global average and deviations for items and users)
- Identify temporal dynamics
  - Express biases as a function over time
  - (Express user vectors as a function over time)
- Relearn feature vectors

Readjust Predictions

- Not all values in the interval [0;100] are equally frequently used
  - Gather statistics (histograms)
  - Round predictions
- Many ratings of the same user at the same point in time
  - Many of these have the same value
- Analyze and improve understanding of the rating data
  - Try to understand underlying causes
  - Incorporate knowledge to readjust predictions

Roadmap


Get a better understanding of the rating data + Include Biases
(Include Temporal Effects)
Tweak more
First RMSE + Work on performance
Consider Item Relations
Tweak
KDD submission
W1
W2
W3
W4
W5
W6
W7
12.5.
30.6.