Outline

- Preprocessing
  - denoise
  - supplement
- Pairwise Training
  - Max-margin optimization problem
- Multifaceted Factorization Models
  - Extend the SVD++
- Context-aware Ensemble
  - Logistic Regression

Preprocess: Session analysis

- Negative : Positive = 92 : 8 ?
  - not all the negative ratings imply that the users rejected to follow the recommended items
- Eliminating these “omitted” records is necessary
  - These negative samples can not indicate users' interests

Preprocess: Session analysis

- Session slicing according to the time interval

$$\Delta t_s(u) = t_{s+1}(u) - t_s(u)$$

When $\Delta t_s(u) < 3600(s)$, we denote it as $\Delta \ddot{t}_s(u)$.

$$\tau(u) = \frac{1}{2} * \left(\tau_0 + \frac{\sum_{s=0}^{m} \Delta \ddot{t}_s(u)}{|\Delta \ddot{t}(u)|} \right)$$

If $\Delta t_s(u) > \tau(u)$, the training records on time $t_{s+1}(u)$ and $t_s(u)$ will be seperated to different subsets $\psi_k(u)$ and $\psi_{k+1}(u)$.

- Select the right samples from the right session:

$$0 < \frac{|\psi_k^+(u)|}{|\psi_k(u)|} \leq \epsilon = 0.86$$

$$\sigma_s - \sigma_- \leq \pi_- = 0$$

$$\sigma_+ - \sigma_s \leq \pi_+ = 3$$

Preprocess: Session analysis

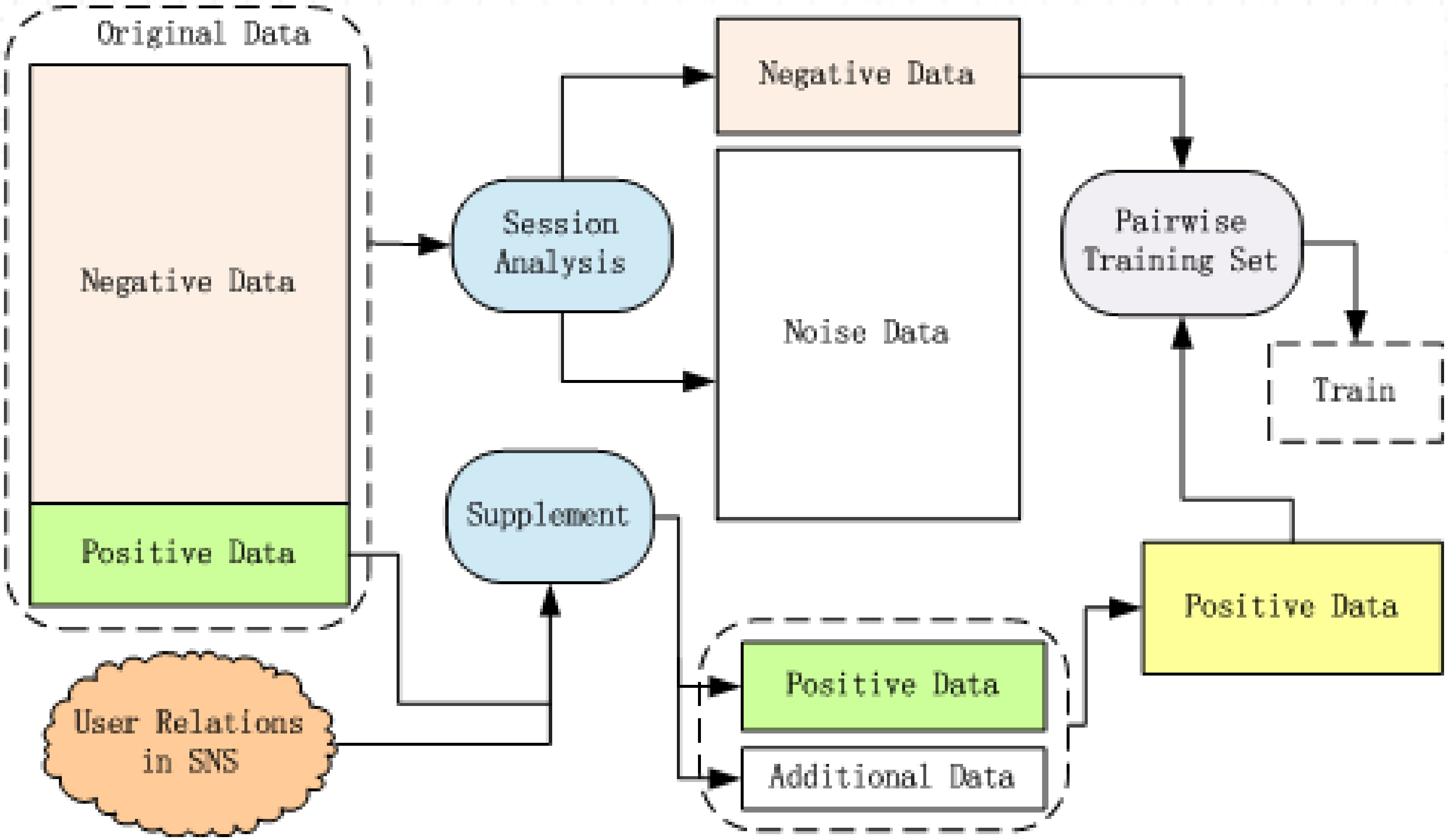
- Training dataset after preprocessing
  - Negative: 67,955,449 -> 7,594,443 (11.2%)
  - Positive: 5,253,828 ->4,999,118
- Benefits
  - improve precision (0.0037)
  - reduce computational complexity

Preprocess: Supply positive samples

- Lack of positive samples
  - An ideal pairwise training requires a good balance between the number of negative and positive samples
- Choose the users
  - users who have a far smaller number of positive samples than negative samples
- Generate the positive samples
  - Figure out from social graphs

$$\operatorname*{argmax}_{i \in S(u) \cap A(u)} \left(\xi_0 N_{at}(u,i) + \xi_1 N_{retweet}(u,i) + \xi_2 N_{comment}(u,i)\right)$$

The procedure of data preprocessing



Original Data
Negative Data
Positive Data
User Relations in SNS
Session Analysis
Supplement
Negative Data
Noise Data
Positive Data
Additional Data
Pairwise Training Set
Train
Positive Data