5-2

BREAKING DOWN USER-USER COLLABORATIVE FILTERING

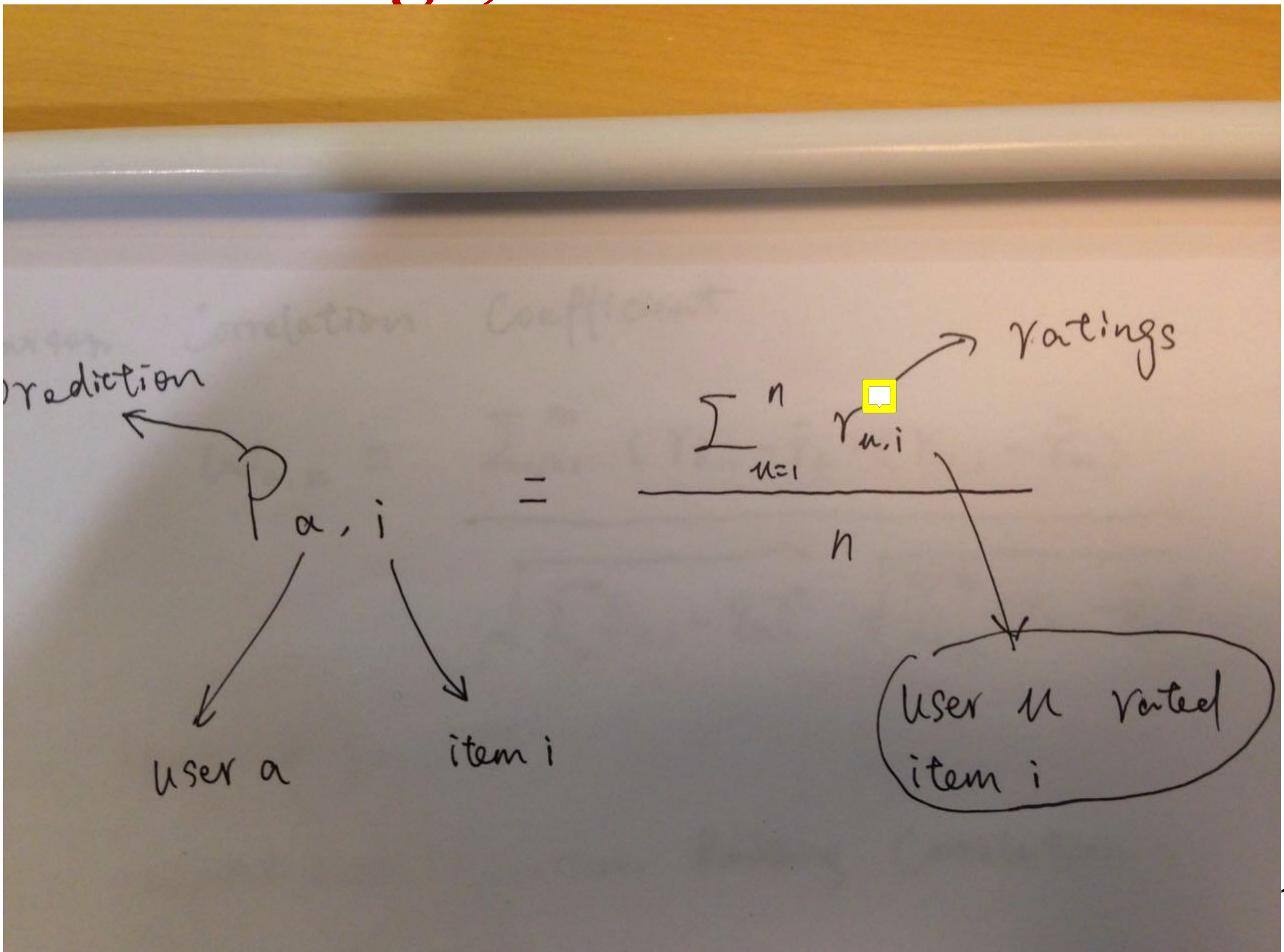
Key Reference

- **Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. CSCW, 1994**
- **An Algorithmic Framework for Collaborative Filtering**
  - **by Herlocker, Konstan, Borchers, Riedl**
  - **Proc. SIGIR 1999**

Rating Matrix

- Matrix R
  - R_{ui}: the rating from user u on item I
  - A very sparse matrix
- Question
  - To infer the values in the empty cells

Just Average, Non-Personalized



Rating Normalization, Non-Personalized

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{n} (r_{u,i} - \bar{r}_u)}{n}$$

- **May be out of the rating scale**

Rating Normalization, Personalized

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{n} (r_{u,i} - \bar{r}_u) \cdot w_{a,u}}{\sum_{u=1}^{n} w_{a,u}}$$

rating agreement

- **How to select the neighborhoods**
- **\bar{r}_u is the average value over all the ratings of u**
- **Remove the neighbors with negative agreement values**

Pearson Correlation Coefficient

Pearson Correlation Coefficient

$$w_{a,u} = \frac{\sum_{i=1}^{m} (r_{a,i} - \bar{r}_a)(r_{u,i} - \bar{r}_u)}{\sqrt{\sum_{i=1}^{m} (r_{a,i} - \bar{r}_a)^2} \sqrt{\sum_{i=1}^{m} (r_{u,i} - \bar{r}_a)^2}}$$

- range of [-1, 1]
- compared with Spearman Ranking Correlation
- m ratings in common

Here, \bar{r}_u is the average value over the ratings of u on the items both u and a have rated

Algorithm for U-U CF

- For a user u
  - Compute its similarity values to all the other users
  - Identify its nearest neighbors
- With the nearest neighbors, for each item i
  - Predict r_{ui} to the weighted sum of the ratings on item i from the neighbors

Issues on U-U CF

- **Low coverage**
  - **For an item, on which all the nearest neighbors have few ratings**

Implementation Issues

- **Given m users and n items**
  - **Computation can be a bottleneck**
    - Correlation between two users is O(n)
    - All correlations for a user is O(mn)
    - All pairwise correlations is O(m^2n)
  - **Lots of ways to make more practical**
    - More persistent neighborhoods
    - Cached or incremental correlations

User-User Variations and Tuning

- **Similarities**
- **Significance weighting**
- **Variance weighting**
  - **Considering the rating variance for an item**
- **Selecting neighborhoods**
- **Normalizing ratings**

Computing Similarities

- **Pearson correlation**
- **Spearman rank correlation**
  - **Hasn't been found to work as well here**
- **Cosine Similarity**

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Significance Weighting

- **Consider the number of co-rated items**
  - **multiply by min(n,50)/50**
    - n is the number of common ratings
    - 50 is the cutoff number

Considering the Rating Variance for an Item

- **Variance weighting**

$$w_{a,u} = \frac{\sum_{i=1}^{m} z_{a,i} * z_{u,i}}{m}$$

Z-score based

$$w_{a,u} = \frac{\sum_{i=1}^{m} v_i * z_{a,i} * z_{u,i}}{\sum_{i=1}^{m} v_i} \qquad (5)$$

We computed an item variance weight as $v_i = \frac{var_i - var_{min}}{var_{max}}$ where $var_i = \frac{\sum_{u=1}^{n} (r_{u,i} - \overline{r_i})^2}{n-1}$, and var_{min} and var_{max} respectively are the minimum and maximum variances over all items. Contrary to our initial hypothesis, applying vari-

Normalizing Ratings, Why?

- **Users rate differently**
  - **Some rate high, others low**
- **Averaging ignores these differences**
- **Normalization compensates for them**

Rating Normalization: Mean-centering

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{n} (r_{u,i} - \bar{r}_u)}{n}$$

- **May be out of the rating scale**

Rating Normalization: z-score normalization

$$P_{a,i} = \frac{\sum_{u=1}^{n} z_{u,i}}{n} \cdot \sigma_a + \bar{r}_a$$

- σ_a : the standard deviation of the ratings of User a
- $\bar{r}_a$: average rating of user a
- $z_{u,i} = \dfrac{r_{u,i} - \bar{r}_u}{\sigma_u}$

Selecting Neighborhoods

- **Threshold similarity**
- **Top-N neighbors by similarity**
- **Combined**

How Many Neighbors?

- In theory, the more the better
  - If we have a good similarity measure
- In practice, noise from dissimilar neighbors decreases usefulness
- Between 25 and 100 is often used
- Fewer neighbors → lower coverage
  - Use the same group of neighbors for different items
  - Give up personalized recommendation if the neighbors do not have enough ratings on the target item

Good Configurations

- **Similarities**
  - **Pearson correlation, Spearman ranking correlation**
- **Significance weighting**
  - **Needed**
- **Variance weighting**
  - **Does not work**
- **Selecting neighborhoods**
  - **Top 30**
- **Normalizing ratings**
  - **Needed**

Revisit to Key Reference

- **An Algorithmic Framework for Collaborative Filtering**
  - **by Herlocker, Konstan, Borchers, Riedl**
  - **Proc. SIGIR 1999**