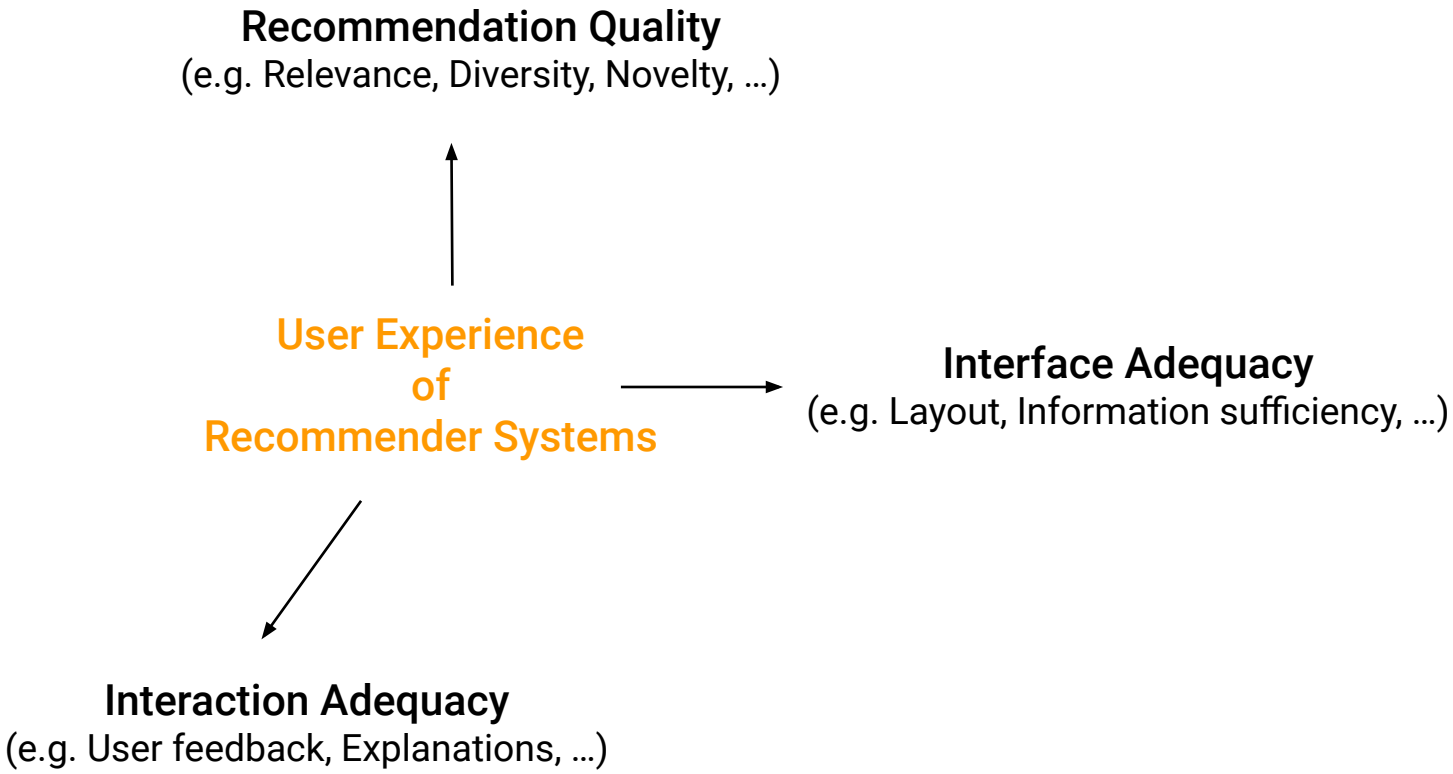
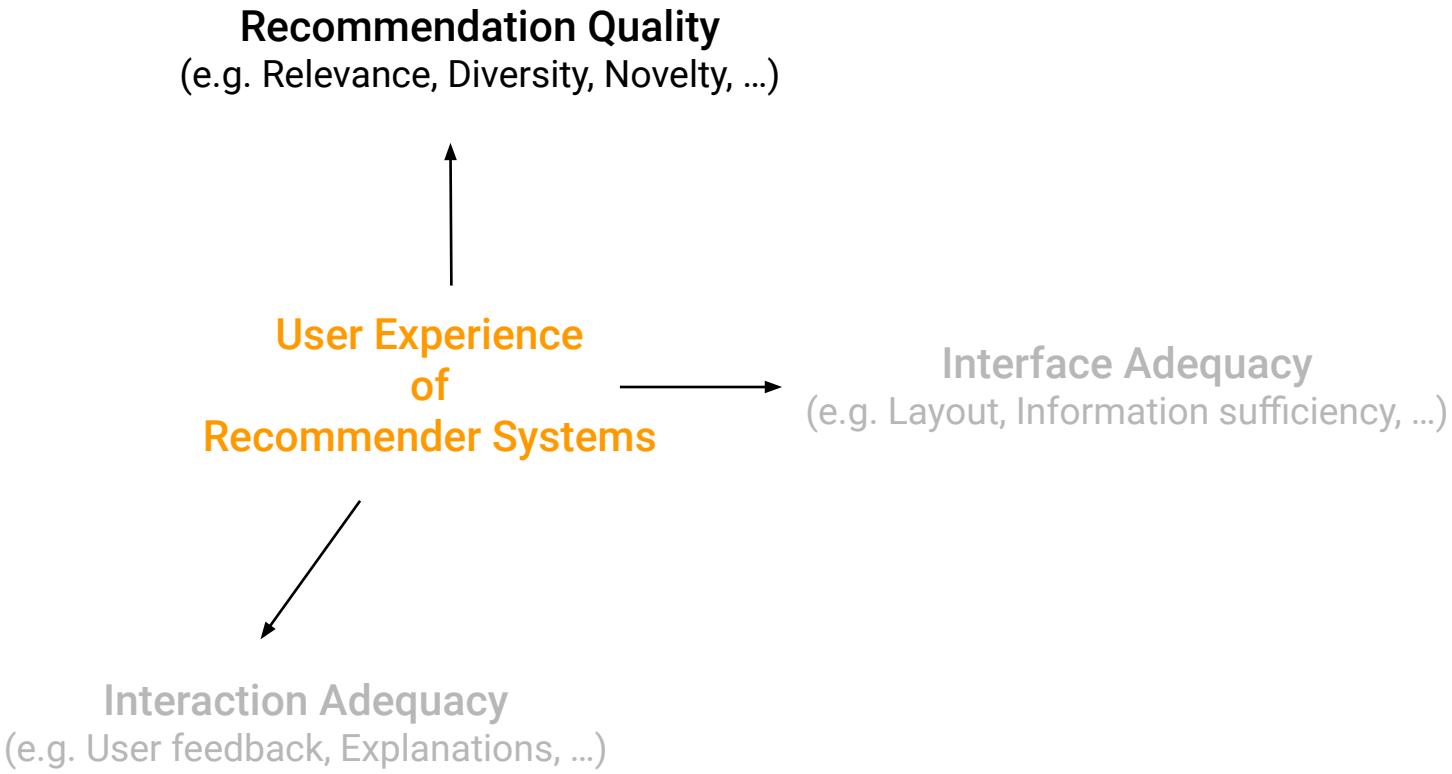
Lecture 09-10-11

- Evaluation of Recommendation Systems

User Experience: Three Dimensional View



User Experience: Three Dimensional View



Recommendation Quality: Relevance (Customer's View)

Relevance *(measure of "correctness")*

1. Recommendation as Rating Prediction
   - Correlation (rate/pred)
   - MAE, MSE, RMSE
2. Recommendation as a Set/ List of Suggestions
   - Precision@n
   - Recall@n
   - F-Measure
3. Recommendation as a rank-sensitive List
   - nDCG
   - MRR
4. Recommendation as a Search
   - Hit-rate
   - Rejection rate

Recommendation Quality: Relevance (Customer's View)

Relevance *(measure of "correctness")*

The system generates predicted ratings $\hat{r}_{ui}$ for a test set T of user-item pairs (u, i) for which the true ratings r_{ui} are known.

$$\text{MAE} = \frac{1}{|\mathcal{T}|} \sum_{(u,i) \in \mathcal{T}} |\hat{r}_{ui} - r_{ui}|$$

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{(u,i) \in \mathcal{T}} (\hat{r}_{ui} - r_{ui})^2}$$

Recommendation Quality: Relevance (Customer's View)

Relevance *(measure of "correctness")*

	Recommended	Not recommended
Used	True-positive (tp)	False-negative (fn)
Not used	False-positive (fp)	True-negative (tn)

$$\textbf{Precision} = \frac{\#tp}{\#tp + \#fp}$$

$$\textbf{Recall (True Positive Rate)} = \frac{\#tp}{\#tp + \#fn}$$

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Recommendation Quality: Relevance (Customer's View)

Relevance *(measure of "correctness")*

Assuming each user ***u*** has a "gain" $g_{u,i}$ from being recommended item ***i*** at position ***j*** in the list.

$$\mathrm{DCG} = \frac{1}{N} \sum_{u=1}^{N} \sum_{j=1}^{J} \frac{g_{u,i_j}}{\log_b(j+1)}$$

$$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{DCG}^*}$$

where DCG* is the ideal DCG.

Recommendation Quality: Relevance (Customer's View)

Relevance *(measure of "correctness")*

$$MRR = \frac{1}{|U_{all}|} \sum_{u=1}^{|U_{all}|} RR(u)$$

$$RR(u) = \sum_{i \leq L} \frac{relevance_i}{rank_i}$$

Recommendation Quality: Relevance (Business View)

Business Objective: Increase Revenue

- Increase sales,
- Increase profit,
- Increase the number of customers,
- Retain existing customers,
- Increase repeat visits, and so on.

Business Value *(measure of "effect on business")*

- Click-through rate
- Conversion rate
- Customer return/ retention rate
- Customer engagement

Recommendation Quality: Diversity

Diversity (measure of "variety" in the recommendation list)

Individual Diversity

- Intra-List Diversity (ILD): average pairwise distance between all the pairs of recommendation list
- Subtopic Recall (S-Recall): fraction of features covered in the recommendation list
- α-nDCG: redundancy-aware variant of nDCG

Recommendation Quality: Diversity

Diversity (measure of "variety" in the recommendation list)

Individual Diversity

- Intra-List Diversity (ILD): average pairwise distance between all the pairs of recommendation list

$$\frac{1}{|\mathbb{U}_T|} \sum_{u \in \mathbb{U}_T} \frac{1}{|R_u|(|R_u|-1)} \sum_{i \in R_u} \sum_{j \in R_u \setminus i} 1 - sim(F_i, F_j)$$

Recommendation Quality: Diversity

Diversity (measure of "variety" in the recommendation list)

Individual Diversity

- Subtopic Recall (S-Recall): fraction of features covered in the recommended list of items

$$\textit{S-recall at } K \equiv \frac{\left| \cup_{i=1}^{K} \text{subtopics}(d_i) \right|}{n_A}$$

Recommendation Quality: Diversity

Diversity (measure of "variety" in the recommendation list)

Aggregate/ Sales Diversity

- **Coverage (catalog):** fraction of items recommended at least once.

$$\frac{|\cup_{u \in \mathbb{U}_T} R_u|}{|\mathbb{I}|}$$

- **Distributional Inequality (Entropy/Gini -diversity):** degree of spread of recommendations across all candidate items

Recommendation Quality: Diversity

Diversity (measure of "variety" in the recommendation list)

Adaptive diversification

- **Propensity toward diversity:** user-profile spread over certain item features
- **Personalizing diversity:** user-level clustering based on their tolerance on diversification
- **Aspect-based diversification:** user-profile spread over standard item categories

Recommendation Quality: Diversity

Diversity (measure of "variety" in the recommendation list)

Challenges:

- *Diversity* and *Accuracy* are in trade-off
- *Objective* and *Subjective Diversity* may be different
- Adaptive Diversification may not work at the level of user-perception

Recommendation Quality: Serendipity

Serendipity *(measure of "delightful unexpectedness" of the recommendations)*

- Relevant + Unexpected + Novel

Recommendation Quality: Serendipity

Serendipity *(measure of "delightful unexpectedness" of the recommendations)*

- Unexpectedness *(measure of "surprise" to the user)*
  - Not expected to find item on her own

    OR

    Not expected to enjoy
  - Measured as dissimilarity of the recommended item from the items user typically consumes

Recommendation Quality: Serendipity

Serendipity *(measure of "delightful unexpectedness" of the recommendations)*

- Unexpectedness *(measure of "surprise" to the user)*
  - Measured as dissimilarity of the recommended item from the items user typically consumes

$$\frac{1}{|\mathbb{U}_T|} \sum_{u \in \mathbb{U}_T} \frac{1}{|R_u|} \sum_{i \in R_u} \min_{j \in P_u} 1 - sim(F_i, F_j)$$

Recommendation Quality: Serendipity

Serendipity *(measure of "delightful unexpectedness" of the recommendations)*

- Novelty *(measure of being "unknown" to the user)*
  - Measure of being "unknown"
  - Users don't prefer novel recommendations unless they trust the system
  - Measured as an inverse of popularity

$$\frac{1}{|\mathbb{U}_T|} \sum_{u \in \mathbb{U}_T} \frac{1}{novelty_{max} \cdot |R_u|} \sum_{i \in R_u} -\log_2 \frac{|u \in \mathbb{U},\, r(u,i) \neq 0|}{|\mathbb{U}|}$$

Here $novelty_{max} = -\log_2 \frac{1}{|\mathbb{U}_T|}$ is the maximum possible novelty value which is used to normalize the novelty score of each individual item into $[0, 1]$.

Recommendation Quality: Serendipity

Serendipity *(measure of "delightful unexpectedness" of the recommendations)*

- Relevant + Unexpected + Novel
- ***No consensus*** on the definition and the metric of serendipity in recommender systems
- The presence of ***emotional dimension***, not easy to quantify

Recommendation Quality: As a Search

Effectiveness (*maximize*)

Effectiveness is the degree to which the system helps the user to accomplish her task.

- e.g. finding a relevant recommendation or some broader measure of user satisfaction

Efficiency cost (*minimize*)

Efficiency cost is a measure of the effort involved in completing the task.

- e.g. In terms of total time elapsed, total number of user actions with the system's user interface, number of interaction cycles, or cognitive load

Recommendation Quality: As a Search

Effectiveness

- *Hit/ Rejection -rate* (on each interaction cycles)
- Similarity between the recommended item and the item of interest (on each interaction cycle)
- *Diversity* of Recommendations (in each interaction cycle)
- Average *Surprise* of Recommendations (in each interaction cycle)
- Overall task *success rate*
- *Decision accuracy, user's confidence and intention to return* (after task questionnaire)

Efficiency cost (*minimize*)

- Number of recommendation cycles
- Number of items viewed before the accepted item
- *Ease of use, Cognitive load (after task questionnaire)*

Recommendation As a Search: Offline Trials (Simulation) Protocols

Lorraine Mcginty and Barry Smyth. Adaptive selection: An analysis of critiquing and preference-based feedback in conversational recommender systems. International Journal of Electronic Commerce, 11(2):35–57, 2006.

Leave-one-out Methodology
Critiquing and preference-based forms| content-based settings | structured item descriptions

- **Base query:** randomly picked item
- **Set of queries:** random subsets of Base query's features (easy, moderate, difficult)
- **Target**: most similar to the Base query
- **Selection criteria (in each cycle)**: most similar to the target (critiques are the differences between the query and the selected item features)
- **End of conversation:** item of interest (Target) is found

Recommendation As a Search: Offline Trials (Simulation) Protocols

Kai Luo, Scott Sanner, Ga Wu, Hanze Li, and Hojin Yang. 2020. Latent Linear Critiquing for Conversational Recommender Systems. In Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.

Simulation Methodology
Critiquing feedback| latent collaborative settings | unstructured item descriptions

- **Target**: randomly picked liked item from user's test set
- **Critiquing criteria (in each cycle):** inconsistent keyphrases in order of their popularity in the dataset
- **End of conversation**: item of interest (Target) is found; otherwise, up to 10 cycles

Recommendation As a Search: Online Evaluation

Challenges in Online Trials/ Studies

- Developing User Interface for Evaluation
- Recruiting participants for the trial
- Making sure that the results are not biased
- Getting approval from Ethics committee is usually a longer process

Recommendation As a Search: Online User Trial Protocols

- Pu, Pearl Huan Z., and Pratyush Kumar. "Evaluating example-based search tools." Proceedings of the 5th ACM conference on Electronic commerce. 2004.
- Chen, Li, and Pearl Pu. "Evaluating critiquing-based recommender agents." AAAI. 2006.

User-trial Protocol
Critiquing feedback | content-based settings | structured item descriptions

- Scenario: Find an item that you would purchase if given the opportunity
- Base query: An item that user likes the most (from given)
- Critiquing criteria (in each cycle): Apply critiques as per the given constraints
- End of conversation: as the given tasks are over

Recommendation As a Search: Online User Trial Protocols

Rana, Arpit, and Derek Bridge. "Navigation-by-preference: a new conversational recommender with preference-based feedback." Proceedings of the 25th International Conference on Intelligent User Interfaces. 2020.

User-trial Protocol
Preference-based feedback | content-based settings | unstructured item descriptions

- **Scenario**: Find an item that you would enjoy watching with your putative companion
- **Base query**: Seed item from user's profile
- **Selection criteria (in each cycle)**: the one which user finds closer to the item of her interest
- **End of conversation**: User has to interact with the system up to 8 cycles

Next Lecture

- Collaborative Methods