IT492: Recommendation Systems

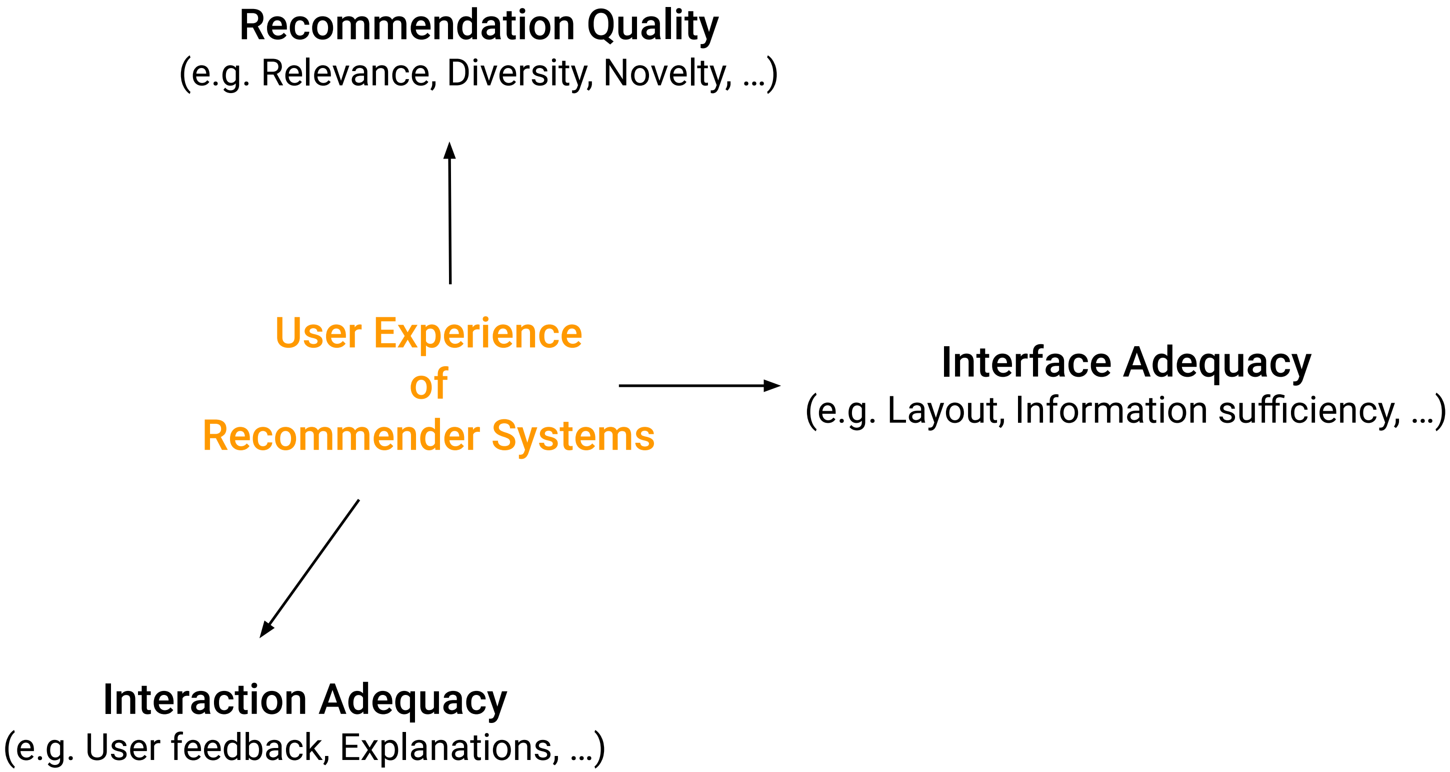
Lecture - 13–14

Evaluation of Recommendations

Arpit Rana

20th Mar 2022

User Experience: Three Dimensional View



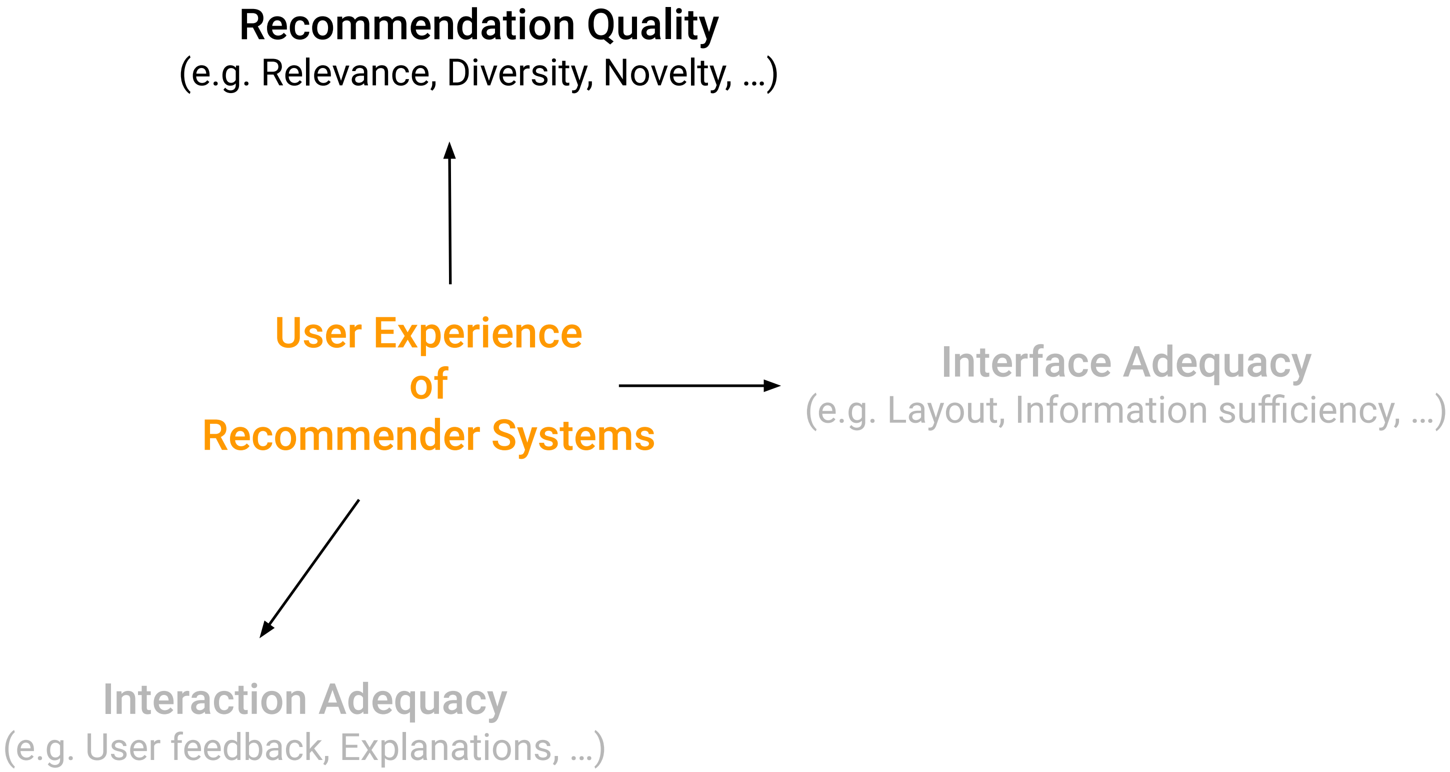
Recommendation Quality
(e.g. Relevance, Diversity, Novelty, …)

User Experience
of
Recommender Systems

Interface Adequacy
(e.g. Layout, Information sufficiency, …)

Interaction Adequacy
(e.g. User feedback, Explanations, …)

User Experience: Three Dimensional View



Recommendation Quality
(e.g. Relevance, Diversity, Novelty, …)

**User Experience
of
Recommender Systems**

Interface Adequacy
(e.g. Layout, Information sufficiency, …)

Interaction Adequacy
(e.g. User feedback, Explanations, …)

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 Suggestion

   - 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)

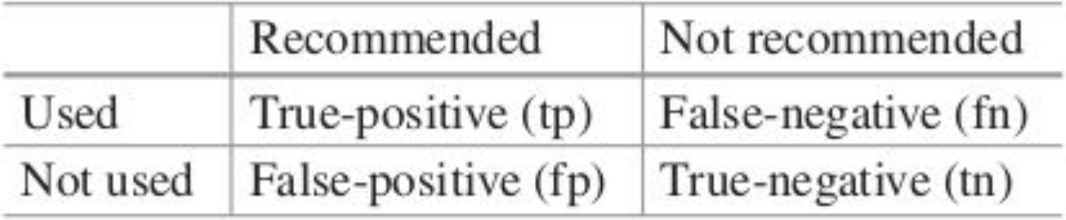
Relevance *(measure of "correctness")*

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

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

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)

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

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

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

Recommendation Quality: Relevance (Customer's View)

Relevance *(measure of "correctness")*

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

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

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

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

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

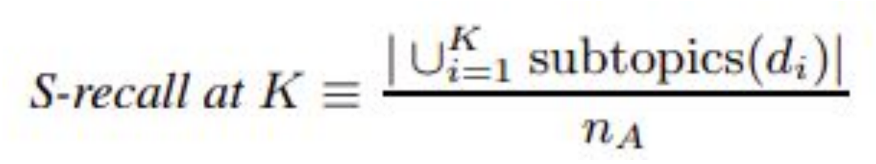
$$\frac{1}{|\mathbb{U}_T|}\sum_{u\in\mathbb{U}_T}\frac{1}{|R_u|(|R_u|-1)}\sum_{i\in R_u}\sum_{j\in R_u\setminus i} 1 - sim(F_i, F_j)$$

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



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

Recommendation Quality: Diversity

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

Aggregate/ Sales Diversity

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

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

- **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)*

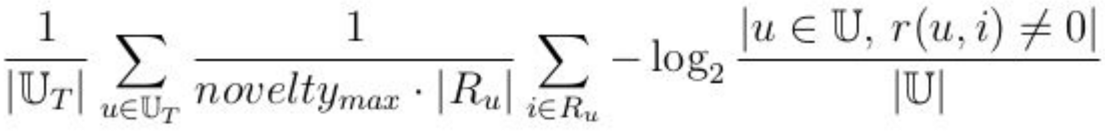
- **Unexpectedness *(measure of "surprise" to the user)***
  - Measured as dissimilarity of the recommended item from the items user typically consumes

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

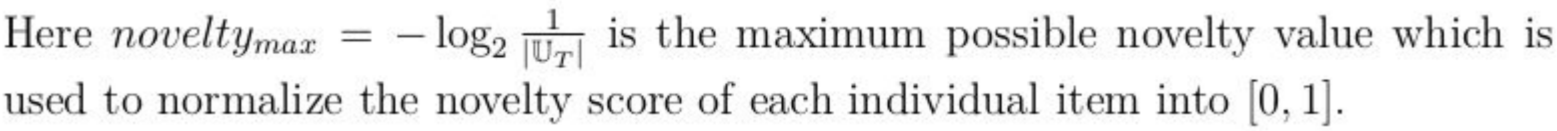
Recommendation Quality: Serendipity

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
  -



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



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)*

IT492: Recommendation Systems

Next lecture -

Evaluation of Recommendations