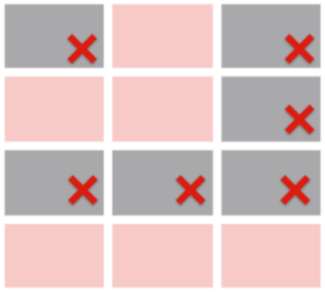
Effectiveness > Quantitative Evaluation

Predictive Quality Metrics

Predictive metrics assess if a system can make accurate predictions about item relevance.

- Precision@K
- Recall@K
- F1-score
- …

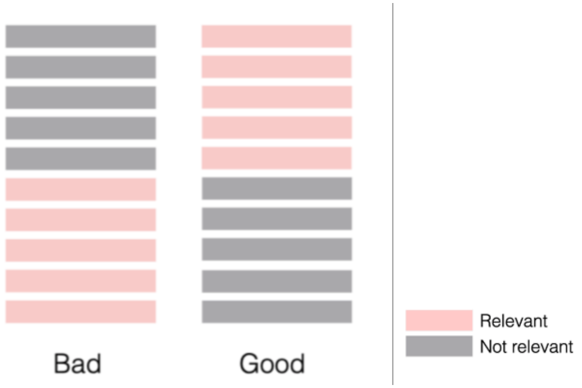


Ranking Quality Metrics

Ranking metrics assess the ability to order the items based on their relevance to the user or query.

In an ideal scenario, all the relevant items should appear ahead of the less relevant ones. Ranking metrics help measure how far you are from this.

- MRR
- Hit Rate@K
- NDCG
- …



Top-K recommendations

https://www.evidentlyai.com/ranking-metrics/evaluating-recommender-systems#ranking-quality-metrics

Top-K Recommendations

The Top-K Recommendation Metric is a commonly used metric for evaluating the **performance of a recommender system** in providing **a list of K items** that are supposed to be the **most relevant for a user**.

In short:

- The system **generates a list of K recommended items** for each user.
- The metric evaluates **how well this list matches the items the user actually found interesting** (e.g., items the user viewed, clicked on, or purchased).

The most common metrics:

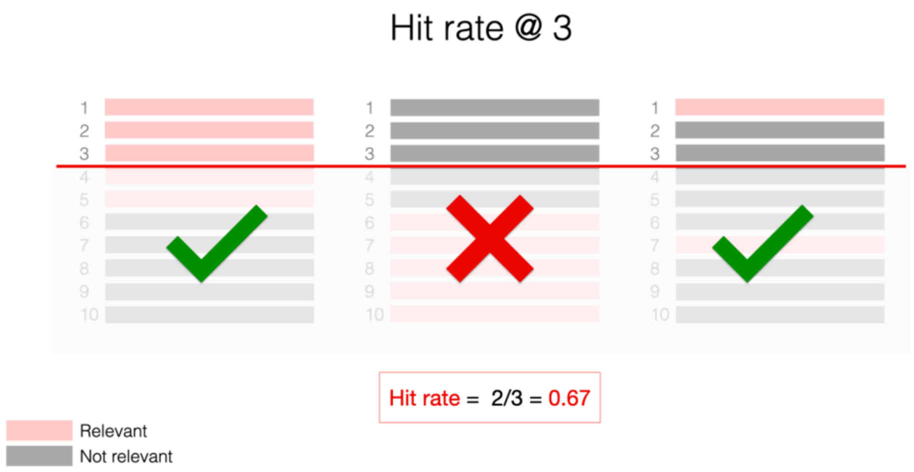
- Hit Rate@K
- Precision@K
- Recall@K
- NDCG@K
- MRR
- …

https://www.evidentlyai.com/ranking-metrics/evaluating-recommender-systems#ranking-quality-metrics

Hit Rate@K : calculates the share of users for which at **least one relevant item** is present in the ***K***.

$$\mathrm{HR@}K = \frac{\text{(Number of users with at least 1 hit)}}{\text{(Total number of users)}}$$

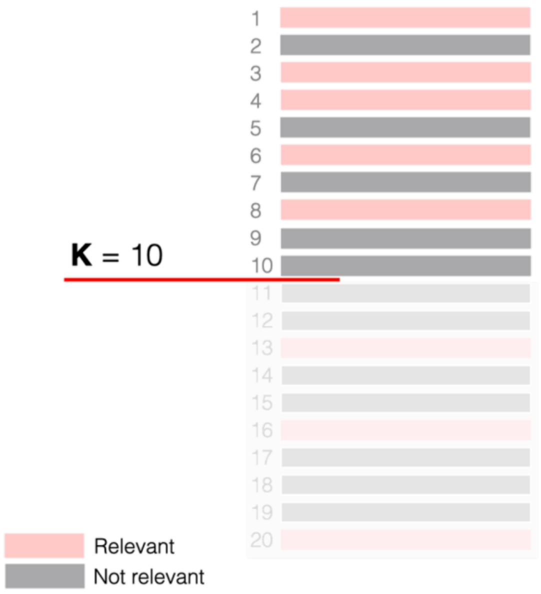
- A **hit** occurs when any of the ***K*** recommended items matches a known relevant item for that user (e.g., an item they clicked, bought, or liked).
- The metric ranges from 0 to 1. A higher value means better performance.



The recommender successfully included **at least one relevant item** in the top 3 for **67% of the users**.

https://www.evidentlyai.com/ranking-metrics/evaluating-recommender-systems#ranking-quality-metrics

Precision@K: measures the proportion of **relevant items** among the top **K** items.



$$\text{Precision@K} = \frac{(\text{N° of relevant items in top}-\text{K})}{(\text{Total number of items in K})}$$

$$\text{Precision@10} = \frac{(\text{5 relevant items among the top 10 recommended})}{(10)} = \frac{(5)}{(10)} = 0.5$$

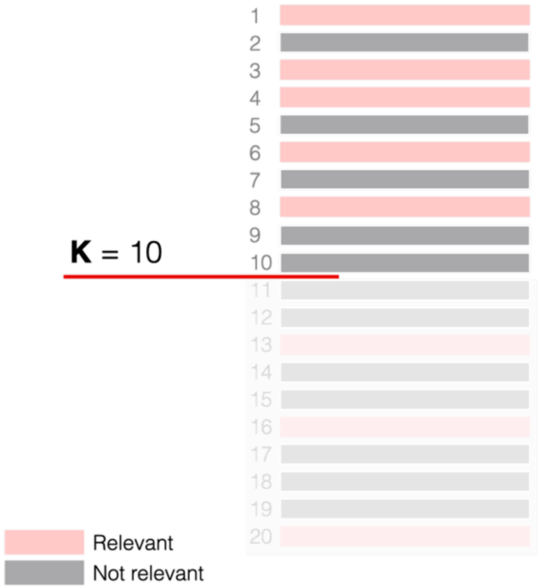
Out of the top-k items suggested, how many are actually relevant to the user?

→ The system achieved a 50% precision in its top 10 recommendations.

- The list shows **20 items ranked from 1 to 20**, with the **top 10** being evaluated (K = 10).
- The **red horizontal line** highlights the cut-off at K=10.

https://www.evidentlyai.com/ranking-metrics/evaluating-recommender-systems#ranking-quality-metrics

Recall@K: measures the percentage of **relevant items** correctly recommended in the top-K recommendations compared to **the total number of relevant items** in the dataset. It indicates how many of the relevant items you could successfully find.



$$\text{Recall@K} = \frac{(\text{N° of relevant items in top K})}{(\text{Total number of relevant items})}$$

$$\text{Recall@10} = \frac{(\text{5 relevant items found})}{(\text{8 total relevant items})} = \frac{(5)}{(8)} = 0.625$$

Out of all the relevant items in the dataset, how many could you successfully include in the top-K recommendations?

→ coverage of relevant items
→ The recommender system retrieved **62.5% of all relevant items** within its top-10 recommendations.

- The list shows **20 items ranked from 1 to 20**, with the **top 10** being evaluated (K = 10).
- The **red horizontal line** highlights the cut-off at K=10.
- In the top **10**, there are **5 relevant items**.
- Across all **20** items, there are **8 relevant items total**.

https://www.evidentlyai.com/ranking-metrics/evaluating-recommender-systems#ranking-quality-metrics

MRR (Mean Reciprocal Rank): is a metric used to evaluate **how high the first relevant item** appears in a ranked list of recommendations.

To calculate MRR, you first need to compute the **RR** (Reciprocal Rank) for each user. It is defined as:

$$\text{RR} = \frac{1}{\text{rank of the first relevant result}}$$

The **MRR** is the average of the reciprocal ranks across all users:

$$\text{MRR} = \frac{1}{\text{U}} \sum_{U=1}^{U} \frac{1}{\text{rank}_1}$$

U= Total number of users

Rank_i= position of the first relevant item for user *i*

MRR = (1 + 0.33 + 0.17 + 0.5) / 4 = 2 / 4 = 0.5

A **higher MRR** means users are seeing relevant results **earlier** in the list.

https://www.evidentlyai.com/ranking-metrics/evaluating-recommender-systems#ranking-quality-metrics

NDCG (Normalized Discounted Cumulative Gain): compares rankings to an **ideal order** where **all relevant items** are at the **top** of the list.

$$\text{NDCG@K} = \frac{DCG@K}{IDCG@K}$$

DCG = Discounted Cumulative Gain

IDCG = Ideal Discounted Cumulative Gain

To calculate NDCG@K:

- First, you measure how good your list is using **DCG** (Discounted Cumulative Gain). This gives more points to relevant items at the top of the list and fewer points to those lower down.

- Then, you compare that score to the **maximum score** you could get if the relevant ones were all **perfectly ranked (IDCG)**.

You can also average NDCG scores for all users to get an overall idea of your system's ranking quality.

NDCG values go from **0** to **1**:

- 1 means the list is perfectly ranked.
- Values closer to 0 mean the ranking is not very good.

Explainability

Qualitative Evaluation

Quantitative Evaluation

Markchom, Thanet, Huizhi Liang, and James Ferryman. "Review of Explainable Graph-Based Recommender Systems." arXiv preprint arXiv:2408.00166 (2024).

Explainability

Qualitative Evaluation

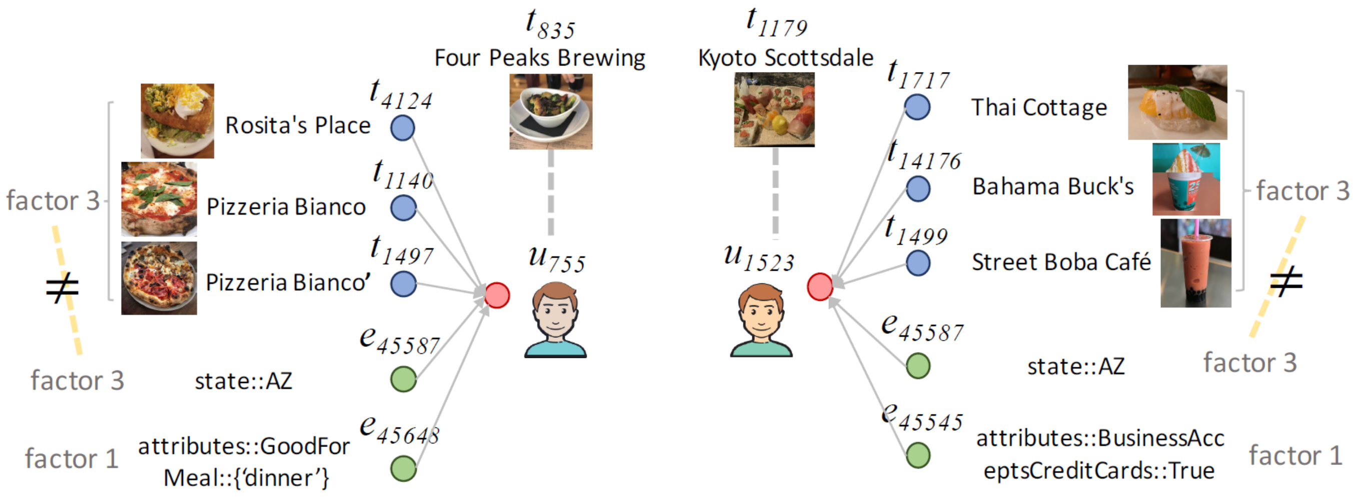
Visualization

Case Study

User Survey

Markchom, Thanet, Huizhi Liang, and James Ferryman. "Review of Explainable Graph-Based Recommender Systems." arXiv preprint arXiv:2408.00166 (2024).

- It is the representation of explanations and/or related information in a **visual** format (charts, graphs, or other visual aids).



Liu, Ninghao, et al. "Explainable recommender systems via resolving learning representations." Proceedings of the 29th ACM international conference on information & knowledge management. 2020.

Markchom, Thanet, Huizhi Liang, and James Ferryman. "Review of Explainable Graph-Based Recommender Systems." arXiv preprint arXiv:2408.00166 (2024).

Explainability

Qualitative Evaluation

Visualization

Case Study

User Survey

Markchom, Thanet, Huizhi Liang, and James Ferryman. "Review of Explainable Graph-Based Recommender Systems." arXiv preprint arXiv:2408.00166 (2024).

- It typically involves choosing at least one user and showing his/her recommendation obtained from the model along with its explanation.
- By comparing them with the user's personal profile or other evidence of the user's interests, it is possible to evaluate whether this explanation makes sense or is suitable for being used in real-world situations or not.
- Characteristics:
  - Usually involves a small number of hand-picked examples.
  - The analysis is narrative or descriptive.
  - Helpful for providing insight into the model's inner workings, particularly for debugging or explaining key behaviors.
- Limitations:
  - Not generalizable.
  - No user involvement — it's done from a developer's or researcher's perspective.

Markchom, Thanet, Huizhi Liang, and James Ferryman. "Review of Explainable Graph-Based Recommender Systems." arXiv preprint arXiv:2408.00166 (2024).

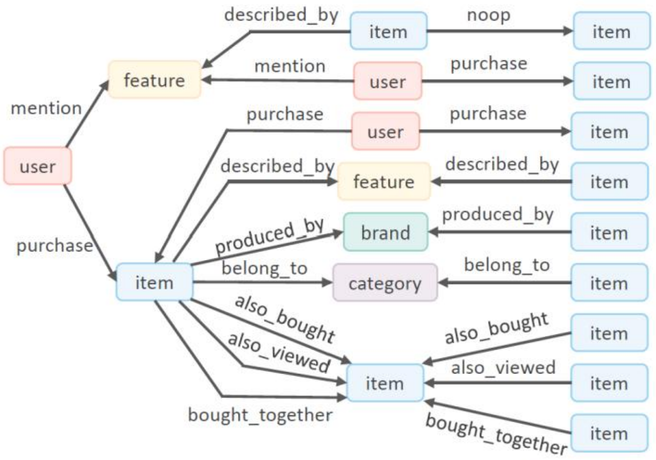


Figure 5: All 3-hop path patterns found in the results.

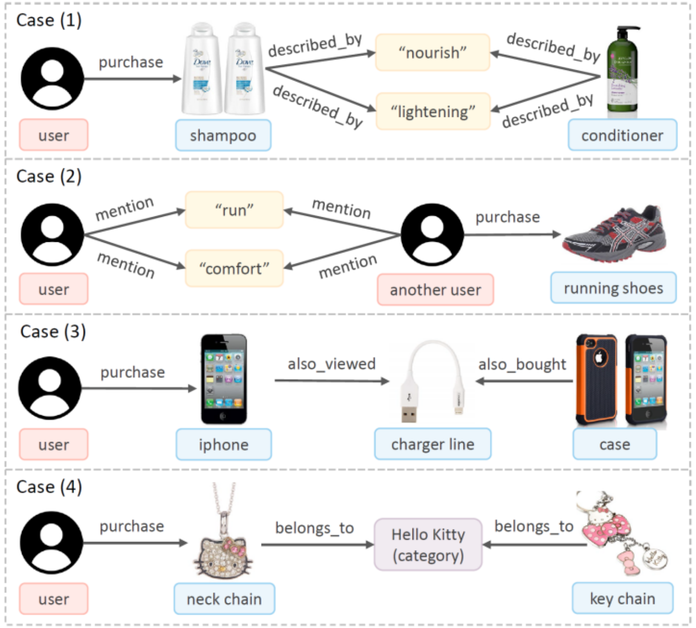


Figure 6: Real cases of recommendation reasoning paths.

Yikun et al. (2019) interpreted and evaluated the **explainability** of the PGPR model using a case study.

Yikun Xian, Zuohui Fu, S. Muthukrishnan, Gerard de Melo, and Yongfeng Zhang. 2019. Reinforcement Knowledge Graph Reasoning for Explainable Recommendation. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 285–294.

Explainability

Qualitative Evaluation

Visualization

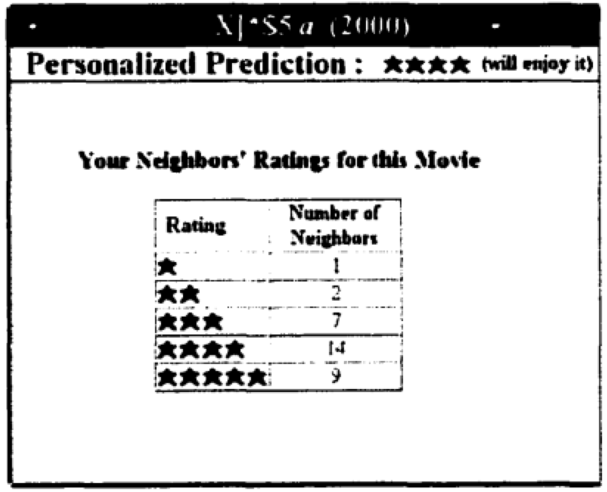
Case Study

User Survey

Markchom, Thanet, Huizhi Liang, and James Ferryman. "Review of Explainable Graph-Based Recommender Systems." arXiv preprint arXiv:2408.00166 (2024).

- It involves recruiting a group of people (which could be stakeholders or potential users of the systems)
- It uses **questionnaires/ A/B tests** to gain knowledge from these users regarding the explainability of recommender systems.
- Characteristics:
  - Typically involves a predefined task, such as rating the usefulness of explanations.
  - Can be quantitative (e.g., Likert scales, task performance) or qualitative (e.g., user feedback).
  - May include A/B testing or controlled experiments.
- Limitations:
  - Requires recruiting participants, making it more time- and resource-intensive.
  - May suffer from biases if not well-designed.

Markchom, Thanet, Huizhi Liang, and James Ferryman. "Review of Explainable Graph-Based Recommender Systems." arXiv preprint arXiv:2408.00166 (2024).



Herlocker et al. (2000).

- Study participants were presented with a **hypothetical** situation:
  - imagine that you are considering **going to the theater** to see a movie
  - you consult MovieLens for a **personalized movie recommendations**
  - MovieLens recommends one movie accompanied by a justification
- Each participant was:
  - provided with twenty-one **individual movie recommendations**, each with an **explanation component**
  - asked to rate on a **scale of one to seven** how likely they would be to **go and see the movie**.
- The **average responses on each explanation** were calculated

Herlocker, J. L., J. A. Konstan, and J. Riedl (2000). "Explaining collaborative filtering recommendations". In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work. ACM. 241–250.

Explainability

Qualitative Evaluation

Quantitative Evaluation

Markchom, Thanet, Huizhi Liang, and James Ferryman. "Review of Explainable Graph-Based Recommender Systems." arXiv preprint arXiv:2408.00166 (2024).

Explainability

Quantitative Evaluation

MEP (Mean Explainability Precision)

MER (Mean Explainability Recall)

xF-SCORE (Mean Explainability Score)

Fidelity

Performance Shift

Review matching

Markchom, Thanet, Huizhi Liang, and James Ferryman. "Review of Explainable Graph-Based Recommender Systems." arXiv preprint arXiv:2408.00166 (2024).

MEP (Mean Exaplianability Precision)

- It is defined as the proportion of **explainable items** in the top-n recommendation list, relative to the total number of recommended (top-n) items for **each user.**

$$\mathrm{MEP} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{|\mathcal{E}_u \cap \mathcal{Y}_u|}{|\mathcal{Y}_u|},$$

Where:

- $\mathcal{U}$: set of all users
- $\mathcal{E}_u$: set of **explainable items** for user u
- $\mathcal{Y}_u$: set of **recommended items** for user u
- $|\mathcal{E}_u \cap \mathcal{Y}_u|$: number of **explainable items actually recommended** to user u

Behnoush Abdollahi, Olfa Nasraoui: Using Explainability for Constrained Matrix Factorization. RecSys 2017: 79-83

MEP (Mean Exaplianability Precision)

It is defined as the proportion of **explainable items** in the top-n recommendation list, relative to the total number of recommended (top-n) items for **each user.**

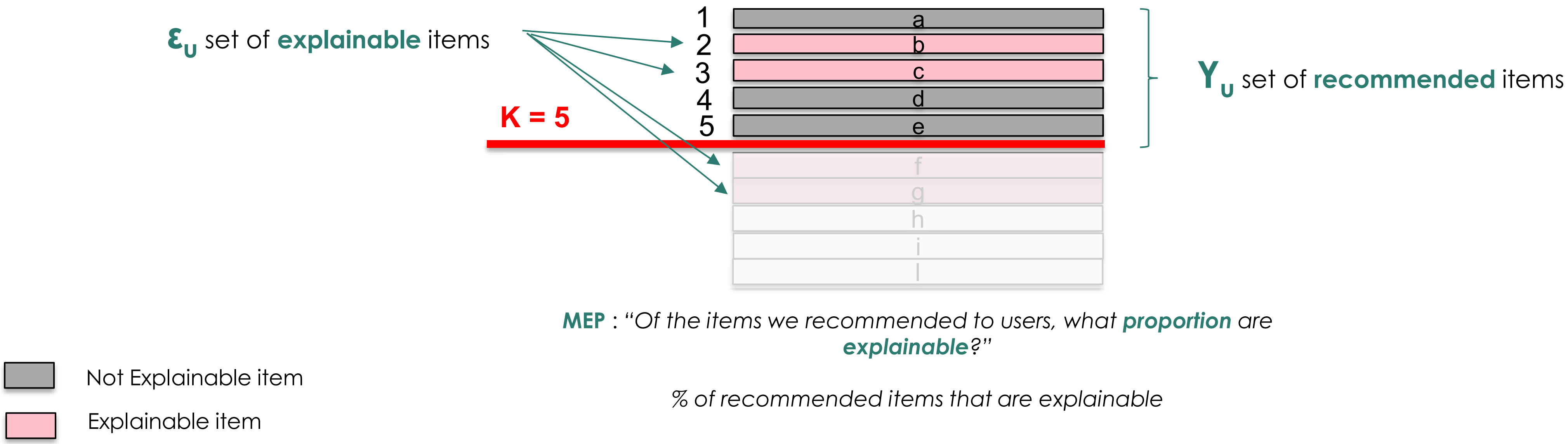
Suppose user u_1 has:

- **5** recommended items: $\mathcal{Y}_{u_1} = \{a, b, c, d, e\}$
- **4** explainable items: $\mathcal{E}_{u_1} = \{b, c, \mathrm{f}, \mathrm{g}\}$
- **N° of explainable items actually recommended** to user u:

  $|\mathcal{E}_{u_1} \cap \mathcal{Y}_{u_1}| = \{b, c\}$

Then for this user: $$\frac{|\mathcal{E}_{u_1} \cap \mathcal{Y}_{u_1}|}{|\mathcal{Y}_{u_1}|} = \frac{2}{5} = 0.4$$

40% of the recommended items are explainable.



MEP : *"Of the items we recommended to users, what* ***proportion*** *are* ***explainable****?"*

% of recommended items that are explainable

MER (Mean Explainability Recall)

- It is defined as the proportion of **explainable items** in the top-n recommendation list, relative to the total number of explainable items for a **given user**.

$$\mathrm{MER} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{|\mathcal{E}_u \cap \mathcal{Y}_u|}{|\mathcal{E}_u|},$$

$\mathcal{U}$ denotes the users set

$\mathcal{E}_u$ denotes the set of explainable items of user u

$\mathcal{Y}_u$ denotes the set of recommended items of user u.

Behnoush Abdollahi, Olfa Nasraoui: Using Explainability for Constrained Matrix Factorization. RecSys 2017: 79-83

MER (Mean Explainability Recall)

It is defined as the proportion of **explainable items** in the top-n recommendation list, relative to the total number of explainable items for a **given user**.

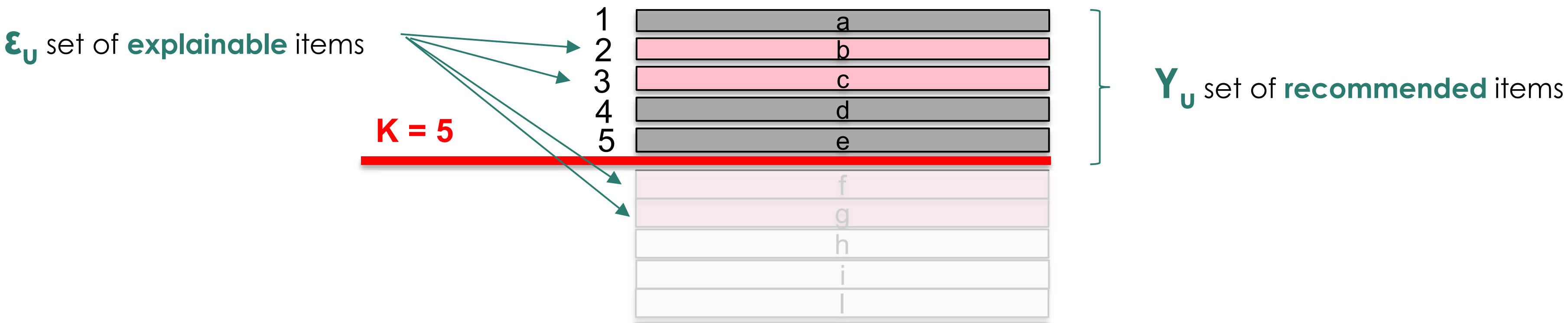
Suppose user u_1 has:

- **5** recommended items: $\mathcal{Y}_{u_1} = \{a, b, c, d, e\}$
- **4** explainable items: $\mathcal{E}_{u_1} = \{b, c, \mathrm{f}, \mathrm{g}\}$
- **N° of explainable items actually recommended** to user u:

  $|\mathcal{E}_{u_1} \cap \mathcal{Y}_{u_1}| = \{b, c\}$

Then for this user: $$\frac{|\mathcal{E}_{u_1} \cap \mathcal{Y}_{u_1}|}{|\mathcal{E}_{u_1}|} = \frac{2}{4} = 0.5$$

50% of the items that *could* have been explained to the user were actually recommended.



MER : *"Of all the items that we could have **explained** to the user, **how many did we actually recommend**?"*

% of explainable items that were recommended

Not Explainable item

Explainable item

xF-SCORE (Mean Explainability F-SCORE)

- The xF-score combines MEP and MER using the **harmonic mean**.

$$\mathrm{MEP} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{|\mathcal{E}_u \cap \mathcal{Y}_u|}{|\mathcal{Y}_u|},$$

$$\mathrm{MER} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{|\mathcal{E}_u \cap \mathcal{Y}_u|}{|\mathcal{E}_u|},$$

$$\text{xF-SCORE} = 2 \cdot \frac{\mathrm{MEP} \cdot \mathrm{MER}}{\mathrm{MEP} + \mathrm{MER}}$$

$\mathcal{U}$ denotes the users set

$\mathcal{E}_u$ denotes the set of explainable items of user u

$\mathcal{Y}_u$ denotes the set of recommended items of user u.

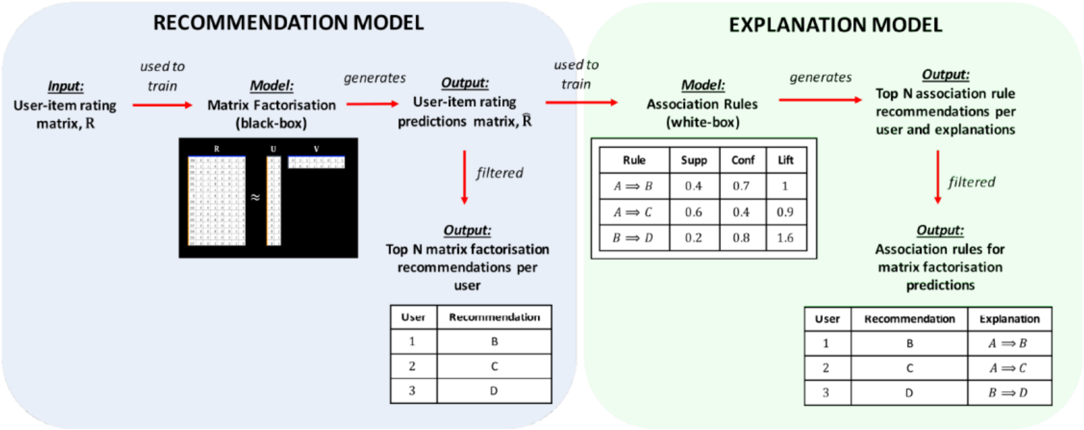
Behnoush Abdollahi, Olfa Nasraoui: Using Explainability for Constrained Matrix Factorization. RecSys 2017: 79-83

MEP, MER, and xF-SCORE

- **MEP** is about how *many of the recommended items* are explainable.
- **MER** is about how *many of the explainable items* were recommended.
- **xF-SCORE** gives a balanced measure of both, much like the F1-score in classification tasks.

Fidelity

This novel metric is defined as the **percentage of explainable items** in the recommended items



$$\text{Model Fidelity} = \frac{|\text{MF recommended items} \cap \text{AR retrieved items}|}{|\text{MF recommended items}|} = \frac{|\text{recommended items} \cap \text{explainable items}|}{|\text{MF recommended items}|}$$

MF (Matrix Factorisation) → black box model
AR (Association Rules) → white-box model

Georgina Peake, Jun Wang: Explanation Mining: Post Hoc Interpretability of Latent Factor Models for Recommendation Systems. KDD 2018: 2060-2069