Towards Unified Metrics for Accuracy and Diversity for Recommender Systems

Arij BOUBAKER, Zhe HUANG

23/02/2024

Summary

Motivation

- Recommender systems evaluation has evolved rapidly in recent years.
- Most research focused on tasks ranging from rating prediction to ranking metrics for top-n recommendation.
- Current RS metrics have been **accuracy-centric**, but there's a growing call within the community for incorporating diversity and novelty.
- Challenge of the **accuracy-diversity dilemma** in RSs : Users want to receive recommendations that are both relevant and diverse.
- The author proposed a unified metric that balances accuracy with diversity and novelty : $\alpha\beta$-**nDCG**.

Related Work

- NRBP extends RBP for diversification, combining intent-aware metrics and aspect diversity assessment.
- Expected Utility (EU) penalizes non-relevant document presence in ranking scenarios.
- Rank-Biased Utility (RBU) is informed by desired axioms for search result diversification.
- Normalized Discounted Cumulative Gain (nDCG) is a robust accuracy metric widely utilized in evaluation.
- α-nDCG adapts nDCG for aspect diversity evaluation, accounting for various aspects.

α-nDCG

- Focuses on evaluating the diversity of search results by considering the redundancy among different aspects of the items recommended.
- The probability of an item i being relevant to a user u is :

$$P(R = 1 \mid u, i) = 1 - \prod_{\phi=1}^{c} (1 - P(a_\phi \in u) \times P(a_\phi \in i))$$

where c is the number of different aspects considered in the recommendation and a_ϕ is an aspect of interest to the user.

- The probability that item i exhibits aspect a_ϕ, can be estimated as :

$$P(a_\phi \in i) = \begin{cases} 1 - \alpha & \text{if } a_\phi \notin i \\ \alpha & \text{if } a_\phi \in i \end{cases}$$

where the parameter α represents the uncertainty of an item exhibiting a particular aspect, typically within the range [0,1].

$\alpha\beta$-nDCG

- Extends α-nDCG by integrating user-specific preferences and distinguishing between missing and explicit ratings
- The same probability in a set of ranked items S up to position k :

$$P(R_k = 1 \mid u, i, S) = 1 - \prod_{\phi=1}^{c} (1 - P(a_\phi \mid u, i) \times P(a_\phi \mid u, S))$$

- The probability of aspect a_ϕ being in item i given user u is given by :

$$P(a_\phi|u,i) = \begin{cases} 0 & \text{if } a_\phi \notin i \\ \alpha(u,i) & \text{if } \nexists r_{u,i} \text{ and } a_\phi \in i \\ \beta(u, r_{u,i}) = \frac{r_{u,i}}{r_{\max}} \times \beta & \text{if } \exists r_{u,i} \text{ and } a_\phi \in i \end{cases}$$

- The probability of the user still being interested in aspect a_ϕ after seeing items in set S is given by :

$$P(a_\phi|u,S) = P(a_\phi|u) \prod_{i \in S} (1 - P(a_\phi|u,i))$$

From α-nDCG to $\alpha\beta$-nDCG

- The α parameter now accounts for the missing rating effect, instead of the confidence in the assessor's decision.
- The formulation of $P(a_\phi|u)$ estimates the aspect relevance for each user given his or her historical data rather than assuming all aspects are equally relevant
- The added β parameter accounts for both the confidence in the rating that the user assigned to each item, and how fast the user is satisfied with relevant items while exploring the ranking.
- α-nDCG is a generic treatment of aspect diversity while $\alpha\beta$-nDCG is a personalized treatment of aspect diversity

Axiomatic approach 1/2

- **Pri (Priority Inside Aspect) :** expresses the preference for the best-rated item when there is no difference in two items' aspects.

$$r_{u,i_{p+k}} > r_{u,i_p}, a_n \in i_p, i_{p+k} \wedge \forall_{\phi \neq n} a_\phi \notin i_p, i_{p+k} \Longrightarrow Q\left(\vec{i}_{p \leftrightarrow p+k}\right) > Q(\vec{i})$$

- **Deep (Deepness Inside Aspect) :** rewards placing highly relevant items in higher ranks within an aspect.

$$a_n \in i_p, i_q, i_{p+1}, i_{q+1} \wedge \forall_{\phi \neq n} a_\phi \notin i_p, i_q, i_{p+1}, i_{q+1},$$
$$p < q, r_{u,i_p} = r_{u,i_q} < r_{u,i_{p+1}} r_{u,i_{q+1}} \Longrightarrow Q\left(\vec{i}_{p \leftrightarrow p+1}\right) > Q\left(\vec{i}_{q \leftrightarrow q+1}\right)$$

- **NonPriSatAsp (Non Priority on Saturated Aspects) :** lowers the score when swapping items with saturated aspects.

$$\exists r_\Delta, s_\Delta \in \mathbb{R}^+ \mid r_{u,i_{p+k}} - r_{u,i_p} = r_\Delta, a_n \in i_p,$$
$$a_{n'} \in i_{p+k}, s\left(\vec{i}[0,\ldots,p], a_{n'}\right) - s\left(\vec{i}[0,\ldots,p], a_n\right) = s_\Delta$$
$$\Longrightarrow Q\left(\vec{i}_{p \leftrightarrow p+k}\right) < Q(\vec{i})$$

- **TopHeav (Top Heaviness Threshold) :** suggests that having relevant items at the top increases ranking quality.

$$n \in \mathbb{N}^+ \mid Q\left(i_1^r, i_2^0 \ldots i_{2n}^0\right) > Q\left(i_1^0, \ldots, i_n^0, i_{n+1}^r, \ldots, i_{2n}^r\right), r > 0$$

Axiomatic approach 2/2

- **TopHeavComp (Top Heaviness Threshold Complementary) :** indicates a preference for discovering relevant items without digging too deep. $\exists m \in \mathbb{N}^+ \mid Q\left(i_1^r, i_2^0 \dots i_{2m}^0\right) < Q\left(i_1^0, \dots, i_m^0, i_{m+1}^r, \dots, i_{2m}^r\right), r > 0$
- **AspRel (Aspect Relevance) :** favors items with aspects that are more relevant to the user's interests.

$$r_{u,j} = r_{u,j'} > 0, a_n \in j, a_{n'} \in j',$$
$$\forall_{\phi \neq n} a_\phi \notin j, \forall_{\phi \neq n'} a_\phi \notin j', w\left(u, a_n\right) > w\left(u, a'_n\right) \Longrightarrow Q\left(\vec{i}_{j \leftrightarrow j'}\right) < Q(\vec{i})$$

- **MoreAsp (More Aspect Contribution) :** indicates that the item with less aspect-level redundancy will get a higher score.

$$r_{u,j} = r_{u,j'}, \sum_{a_\phi \in j} w\left(u, a_\phi\right) - w\left(\vec{i}, a_\phi\right) < \sum_{a_\phi \in j'} w\left(u, a_\phi\right) - w\left(\vec{i}, a_\phi\right)$$
$$\Longrightarrow Q\left(\vec{i}_{j \leftrightarrow j'}\right) < Q(\vec{i})$$

- **MissOverNon (Missing over Non-Relevant) :** models the missing rating effect : the user would favour an unknown item rather than an item that he or she is known to dislike.

$$r_{u,j} = 0, \nexists r_{u,j'}, \exists a_\phi \in j' \mid s\left(u, a_\phi\right) < 1 \Longrightarrow Q\left(\vec{i}_{j \leftrightarrow j'}\right) > Q(\vec{i})$$

Evaluating Metrics Against Axioms

Metric	*Axioms*							
	PRI	DEEP	NONPRISATASP	TOPHEAV	TOPHEAVCOMP	ASPREL	MOREASP	MISSOVERNON
α-nDCG@k	●	●	●	●	●	○	○	○
S-Recall@k	○	○	○	○	○	○	○	○
S-RR@100%	○	○	○	○	○	○	○	○
NRBP	●	●	●	●	●	○	○	○
EU	●	●	●	●	●	●	●	○
RBU@k	●	●	●	●	●	●	●	○
αβ-nDCG@k	●	●	●	●	●	●	●	●

Figure – Comparison of different accuracy-diversity metrics against axioms

(● = axiom satisfied, ○ = axiom not satisfied)

Experiments and Result

- **Experiment Goals :**
  - Validate $\alpha\beta$-nDCG's robustness against established metrics.
  - Use real-world data to simulate RS challenges and assess metric performance.
- **Dataset and Methodology :**
  - Movielens 20M, featuring 19 genres.
  - ***Ideal ordering :*** Prioritizing items that maximize relevance and diversity.
  - **Perturbation techniques :**
    - Relevance swaps
    - Aspect redundancy
    - Aspect priority

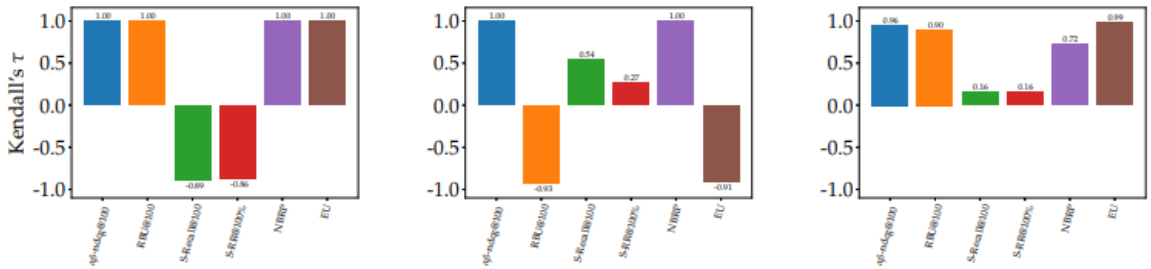

Figure – Experiment 1 : Rank Correlation with Ideal Ordering

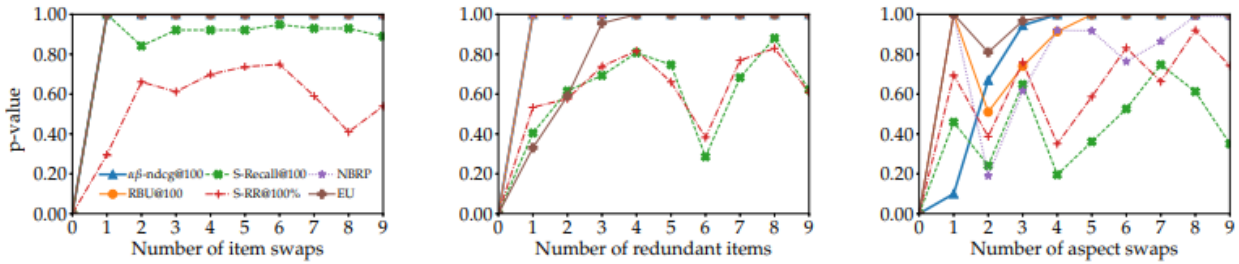

Figure – Experiment 2 : Discriminative power

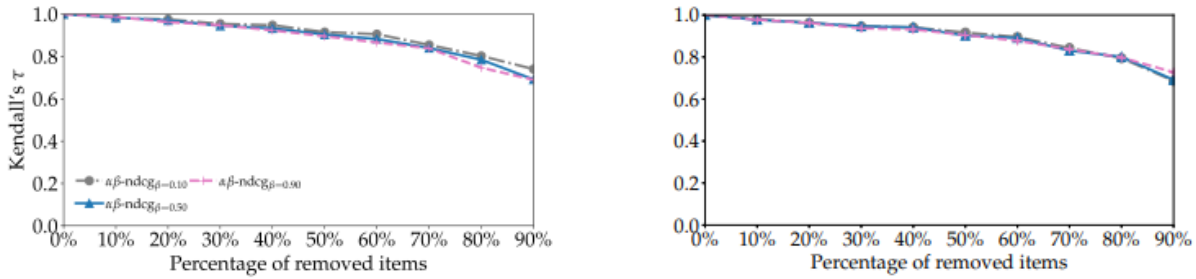

Figure – Experiment 3 : Robustness to Incompleteness

Limitations

- The paper primarily discusses metrics within the context of offline evaluation only.
- The proposed metrics and axioms are based on assumed user behavior models.
- The concept of an ideal ordering is simplified and might not capture all the nuances.

Bibliography

Javier Parapar and Filip Radlinski
Towards Unified Metrics for Accuracy and Diversity for Recommender Systems